Université de Montréal

Modélisation bayésienne avec des splines du comportement moyen d'un échantillon de courbes

par

James Merleau

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

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiae Doctor (Ph.D.)
en Statistique

janvier 2011

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée

Modélisation bayésienne avec des splines du comportement moyen d'un échantillon de courbes

présentée par

James Merleau

a été évaluée par un jury composé des personnes suivantes :

Pierre Duchesne

(président-rapporteur)

Jean-François Angers

(directeur de recherche)

Anne-Catherine Favre

(co-directeur)

Mylène Bédard

(membre du jury)

Roman Krzysztofowicz

(examinateur externe)

Marine Carrasco

(représentant du doyen de la FES)

Thèse acceptée le:

Février 2010

SOMMAIRE

Cette thèse porte sur l'analyse bayésienne de données fonctionnelles dans un contexte hydrologique. L'objectif principal est de modéliser des données d'écoulements d'eau d'une manière parcimonieuse tout en reproduisant adéquatement les caractéristiques statistiques de celles-ci. L'analyse de données fonctionnelles nous amène à considérer les séries chronologiques d'écoulements d'eau comme des fonctions à modéliser avec une méthode non paramétrique. Dans un premier temps, les fonctions sont rendues plus homogènes en les synchronisant. Ensuite, disposant d'un échantillon de courbes homogènes, nous procédons à la modélisation de leurs caractéristiques statistiques en faisant appel aux splines de régression bayésiennes dans un cadre probabiliste assez général. Plus spécifiquement, nous étudions une famille de distributions continues, qui inclut celles de la famille exponentielle, de laquelle les observations peuvent provenir. De plus, afin d'avoir un outil de modélisation non paramétrique flexible, nous traitons les noeuds intérieurs, qui définissent les éléments de la base des splines de régression, comme des quantités aléatoires. Nous utilisons alors le MCMC avec sauts réversibles afin d'explorer la distribution *a posteriori* des noeuds intérieurs. Afin de simplifier cette procédure dans notre contexte général de modélisation, nous considérons des approximations de la distribution marginale des observations, nommément une approximation basée sur le critère d'information de Schwarz et une autre qui fait appel à l'approximation de Laplace. En plus de modéliser la tendance centrale d'un échantillon de courbes, nous proposons aussi une méthodologie pour modéliser simultanément la tendance centrale et la dispersion de ces courbes, et ce dans notre cadre probabiliste général. Finalement, puisque nous étudions une diversité de distributions statistiques au niveau des observations, nous mettons

de l'avant une approche afin de déterminer les distributions les plus adéquates pour un échantillon de courbes donné.

Mots clés : Splines de régression bayésiennes, estimation non paramétrique de fonctions, MCMC avec sauts réversibles, critère d'information de Schwarz, approximation de Laplace, g-priors, synchronisation, modèle de dispersion, choix de modèle, modélisation d'hydrogrammes.

SUMMARY

This thesis is about Bayesian functional data analysis in hydrology. The main objective is to model water flow data in a parsimonious fashion while still reproducing the statistical features of the data. Functional data analysis leads us to consider the water flow time series as functions to be modelled with a nonparametric method. First, the functions are registered in order to make them more homogeneous. With a more homogeneous sample of curves, we proceed to model their statistical features by relying on Bayesian regression splines in a fairly broad probabilistic framework. More specifically, we study a family of continuous distributions, which include those of the exponential family, from which the data might have arisen. Furthermore, to have a flexible nonparametric modeling tool, we treat the interior knots, which define the basis elements of the regression splines, as random quantities. We then use MCMC with reversible jumps in order to explore the posterior distribution of the interior knots. In order to simplify the procedure in our general modeling context, we consider some approximations for the marginal distribution of the observations, namely one based on the Schwarz information criterion and another which relies on Laplace's approximation. In addition to modeling the central tendency of a sample of curves, we also propose a methodology to simultaneously model the central tendency and the dispersion of the curves in our general probabilistic framework. Finally, since we study several statistical distributions for the observations, we put forward an approach to determine the most adequate distributions for a given sample of curves.

Key words : Bayesian free-knot regression splines, nonparametric function estimation, MCMC reversible jump, Schwarz information criterion, Laplace approximation, g-priors, registration, dispersion models, hydrograph modelling.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Je tiens à sincèrement remercier Jean-François Angers pour sa disponibilité, son inspiration et tous ses conseils, techniques et autres. Un gros merci à Luc Perreault et à Anne-Catherine Favre, leur intérêt soutenu pour mes travaux et les nombreuses formes d'encouragement qu'ils m'ont manifestées au cours des dernières années ont été fort utiles.

INTRODUCTION

Le présent document est une étude statistique de la modélisation de données fonctionnelles qui proviennent du domaine de l'hydrologie. Les données considérées sont des séries chronologiques d'écoulements d'eau que nous cherchons à modéliser dans un contexte de modélisation bayésienne. Comme nous le verrons dans la suite de cette introduction, la modélisation des débits d'eau présente des défis importants et s'avère d'une grande importance au Québec puisque l'hydroélectricité représente la principale source d'énergie pour Hydro-Québec.

La thèse regroupe trois articles scientifiques qui sont présentés aux chapitres 2, 4 et 6. Les chapitres 1, 3 et 5 donnent de l'information complémentaire aux articles. Plus précisément, le chapitre 1 présente le problème hydrologique qui nous intéresse en illustrant le propos avec des données réelles ; ce faisant, nous discutons brièvement de certains modèles en hydrologie statistique qui sont reliés à nos travaux. Le chapitre 3 explicite la famille de distributions statistiques qui est utilisée au chapitre 4 (article 2) et donne le contexte général de modélisation de ce dernier. De plus, certains détails techniques concernant les estimateurs employés dans l'article sont aussi exposés. Finalement, le chapitre 5 indique comment la famille de distributions étudiée dans l'article 2 peut être utilisée pour modéliser simultanément la tendance centrale et la dispersion d'un jeu de données, ce qui est fait au chapitre 6 (article 3). Le chapitre 5 expose aussi les estimateurs auxquels l'article 3 fait appel.

La première section qui suit présente le contexte hydrologique dans lequel nos travaux s'inscrivent. La terminologie nécessaire pour la compréhension du problème scientifique y est tout d'abord introduite. Par la suite, un bref survol de certains modèles hydrologiques est effectué afin d'indiquer leurs limitations

par rapport aux problèmes qui nous intéressent. Certains modèles de l'hydrologie statistique, pertinents à notre propos, sont ensuite abordés pour faire un tour d'horizon des méthodes statistiques qui existent dans la littérature.

La deuxième section présente les solutions statistiques que nous proposons pour résoudre certains des problèmes discutés précédemment. Certains outils de modélisation utilisés dans les trois articles sont ensuite exposés. Les contributions scientifiques des trois articles, ainsi que la description des modèles employés dans chacun de ceux-ci, composent la troisième section.

0.1. Contexte hydrologique

En hydrologie, un des plus vieux problèmes est celui de l'étude des crues (voir par exemple Gumbel, 1958), où cette dernière est définie comme l'élévation du niveau d'un cours d'eau due à la fonte des neiges ou à des pluies abondantes ; nous parlons aussi d'une crue lorsqu'il y a une augmentation importante des débits d'eau. Une crue peut être un événement de courte durée lorsqu'elle est causée par des pluies intenses et brèves, mais elle peut aussi s'échelonner sur plusieurs jours ou semaines lorsqu'elle provient de la fonte des neiges. Au Québec, à cause de notre situation géographique et du contexte météorologique, deux principaux types de crue sont observés : la crue printanière et la crue automnale. Cette première crue, qui est habituellement beaucoup plus importante en intensité et en durée que la seconde pour les grands bassins versants, se produit au printemps lors de la fonte de la neige accumulée au sol ; la seconde crue survient à l'automne, ou à la fin de l'été, lors de fortes précipitations liquides.

Dans notre contexte, les crues printanières et automnales font référence aux événements hydrologiques qui débutent par un accroissement des débits d'eau et qui se terminent par une décroissance de ceux-ci, que nous désignons respectivement comme le début et la fin d'une crue. L'apport total en eau d'une crue, ou le volume de la crue, peut être calculé en connaissant son début et sa fin, puisque cette quantité est simplement la somme des apports de la crue. En général, il est possible de déterminer une valeur maximale des débits d'eau, ou du niveau, durant une crue et nous appelons cette valeur la pointe de la crue.

Le problème des crues s'est initialement manifesté dans le contexte des économies agraires puisqu'elles sont fortement dépendantes des événements hydrologiques (Gumbel, 1958). Depuis, dans les économies industrielles, la construction de barrages hydroélectriques a rendu l'étude des crues essentielle car les barrages doivent être en mesure de contenir l'eau provenant de celles-ci ; un état de fait qui concerne directement Hydro-Québec. Au Québec, les crues printanières constituent environ 50% du volume total des apports annuels pour la plupart des bassins versants. Ainsi, du point de vue de la gestion des ressources hydriques, les crues présentent un intérêt considérable. Par conséquent, la modélisation des événements de crue, que ce soit avec des modèles physiques ou statistiques, est fort importante pour Hydro-Québec puisque la sécurité des barrages, la génération d'hydroélectricité et la gestion des ressources hydriques en dépendent toutes.

En plus des crues, les débits d'eau à différents pas de temps se sont aussi avérés intéressants car ces derniers régissent la quantité d'énergie disponible pour les centrales hydroélectriques au cours du temps. Ainsi, une bonne connaissance des débits d'eau au pas de temps journalier, par exemple, permet une meilleure gestion quotidienne des ressources hydriques disponibles à des fins de production énergétique. Par conséquent, il existe aussi un intérêt pour la modélisation des séries chronologiques des apports en eau.

Puisque dans nos travaux, il est question des séries chronologiques annuelles de débits en eau à différents pas de temps, il est utile de définir un hydrogramme annuel comme étant une série chronologique de débits durant une année. Nous parlons alors d'un hydrogramme annuel à pas de temps quotidien pour désigner une série chronologique annuel de débits dont les mesures sont effectuées à chaque jour. Il est aussi possible de parler d'hydrogramme de crue qui désigne la série chronologique des apports en eau durant un événement de crue.

0.1.1. Modèles conceptuels et physiques

Bien que ce type de modèles ne soit pas traité dans ce qui suit, nous mentionnons tout de même l'existence de modèles hydrologiques construits afin de

reproduire les séries chronologiques des débits d'eau à partir de variables physiographiques et météorologiques. Deux modèles hydrologiques présentement utilisés par Hydro-Québec sont HSAMI (Fortin, 2000) et HYDROTEL (Fortin *et al.*, 2001a, 2001b). Ces modèles, basés sur des principes physiques et des relations empiriques, ont comme fonction de reproduire les écoulements d'eau étant donné des variables physiographiques telles le type de sol et l'occupation du sol, ainsi que des variables météorologiques telles la température et les précipitations.

Une fois que les paramètres des modèles physiques sont fixés ou calibrés, ils sont en mesure de prévoir temporellement les débits d'eau en utilisant les séries chronologiques des variables météorologiques. Les débits d'eau représentant la sortie finale, une partie de ces modèles transforme les écoulements en eau dans les différentes composantes du sol en débits simulés. Par exemple, HSAMI et HYDROTEL comportent trois composantes qui fournissent chacune leur propre contribution aux débits simulés. Ainsi, la forme des hydrogrammes produits par ces modèles est le résultat de ces trois composantes qui réagissent de façons différentes selon les conditions météorologiques à travers le temps.

Chacune des composantes possède ses propres caractéristiques hydrographiques. Ainsi, l'hydrogramme associé à une de ces composantes peut être par exemple un hydrogramme unitaire traditionnel (Sherman, 1932 ; Doodge, 1959 ; Chow, 1964 ; Chow *et al.*, 1988 ; Pilgrim et Cordery, 1993 ; Yue et Hashino, 2000) ou un hydrogramme unitaire synthétique (Snyder, 1938 ; U.S. Soil Conservation Service US-SCS, 1985) qui représentent deux méthodes pour modéliser la réponse en débits d'eau résultant des précipitations.

Les modèles hydrologiques sont donc principalement conçus pour reproduire les séries chronologiques des débits d'eau en faisant appel à des séries chronologiques météorologiques. Ainsi, ils ne sont pas nécessairement bien adaptés pour étudier des crues se produisant avec des faibles probabilités puisque pour ce faire, il faut être en mesure de générer des séries chronologiques météorologiques qui possèdent aussi des faibles probabilités d'occurrence, ce qui n'est pas une tâche facile.

À cause des défis importants associés à l'utilisation des modèles conceptuels et physiques pour étudier les phénomènes hydrologiques extrêmes, les hydrologues font appel à des méthodes statistiques.

0.1.2. Modèles en hydrologie statistique

0.1.2.1. *Modélisation des crues*

Au départ, l'hydrologie statistique s'est principalement développée autour des problèmes reliés aux crues. Plus précisément, l'analyse fréquentielle de crue étudie les pointes et les volumes de crue à l'aide de distributions statistiques pour ensuite faire de l'inférence au niveau des quantiles extrêmes de ces distributions. L'intérêt d'étudier les quantiles extrêmes, des crues printanières ou automnales, provient non seulement de la nécessité de se protéger contre les déversements lors de la construction de barrages, mais aussi du besoin d'optimiser la gestion des ressources hydriques.

En hydrologie, nous parlons de quantiles possédant une certaine période de retour, où cette dernière est définie comme étant l'inverse de la probabilité au dépassement. Par exemple, une période de retour de mille ans correspond à une probabilité au non-dépassement de 0.999. Ainsi, il est possible d'obtenir des quantiles, avec une certaine période de retour, pour les pointes et pour les volumes de crue d'un certain site et d'utiliser ces quantiles pour se protéger contre les événements extrêmes qui peuvent se produire avec une faible probabilité.

Les analyses fréquentielles des pointes et des volumes de crue se font habituellement de façon indépendante et pour chaque quantité, une distribution statistique adéquate doit être choisie. Les praticiens vont souvent vérifier l'adéquation de plusieurs distributions statistiques (voir Bobée et Ashkar, 1991 ; Rao et Hamed, 2000 ; Favre *et al.*, 2008) avant de faire leur choix. Cet exercice est souvent effectué par une inspection visuelle et l'expertise de l'hydrologue joue ici un rôle important. Par exemple, bien que la distribution généralisée des valeurs extrêmes (Coles, 2001) semble la loi de probabilité la plus appropriée pour modéliser les pointes, qui sont des valeurs maximales annuelles, d'autres distributions, telles la

loi log-Pearson type III, peuvent être recommandées par des comités d'experts (IACWD, 1982).

Plus récemment, des approches bivariées ont été mises de l'avant afin de modéliser conjointement les pointes et les volumes de crue ; par exemple, Yue *et al.* (1999) proposent une distribution Gumbel bivariée. Une autre méthodologie étudiée au cours des dernières années est l'utilisation des copules (Genest et Favre, 2007) qui permettent d'obtenir diverses formes de dépendance entre les variables d'intérêt, une fois les distributions marginales spécifiées.

Après avoir choisi des distributions statistiques et déterminé des quantiles avec certaines périodes de retour pour les pointes et les volumes de crue, l'hydrologue doit ensuite construire des hydrogrammes de crue qui possèdent ces caractéristiques afin de disposer de débits d'eau sous forme de séries chronologiques. La forme des hydrogrammes (de crue) est fort importante car c'est elle qui détermine la stratégie de gestion temporelle à adopter.

Bien qu'il existe diverses méthodes pour construire des hydrogrammes synthétiques possédant les propriétés des quantiles de crue, la méthode la plus utilisée à Hydro-Québec est la méthode de l'hydrogramme type (Nezhikhovsky, 1971 ; Sokolov *et al.* 1976). Cette approche consiste à prendre un hydrogramme observé dans le passé et à appliquer une transformation d'échelle afin que l'hydrogramme possède les propriétés de crue voulues. Certains auteurs suggèrent de modéliser les crues, ou plus précisément les crues transformées, en utilisant des densités de probabilité telles les lois gamma ou bêta. C'est cette dernière densité de probabilité que Yue *et al.* (2002) utilisent pour modéliser les crues printanières.

0.1.2.2. *Modélisation des séries chronologiques de débits*

À l'instar de l'étude des crues, la modélisation des séries chronologiques des débits d'eau représente un problème d'intérêt. Les séries chronologiques de débits affichent plusieurs caractéristiques particulières qui rendent leur modélisation à partir des méthodes usuelles inapplicables. Ces caractéristiques sont : la non-stationnarité des débits à travers le temps et la saisonnalité variable des événements importants d'une année à l'autre.

En effet, les débits ne sont pas stationnaires, que ce soit au niveau de la moyenne ou au niveau de la variabilité, car il existe différents régimes à travers l'année qui possèdent chacun leur propre tendance centrale et leur propre dispersion. Par exemple, un régime hivernal possède des débits faibles avec peu de dispersion, alors qu'un régime de crue affiche des débits élevés avec une forte dispersion. Ainsi, cette caractéristique rend l'utilisation des modèles de type ARIMA (Box *et al.*, 1994) presqu'impossible, sauf si les années sont divisées en régime et que chaque régime possède son propre modèle ARIMA ; cependant, un tel modèle doit aussi incorporer un sous-modèle pour les transitions entre les différents régimes (Anne-Catherine Favre, communication personnelle).

L'emploi des modèles de séries chronologiques qui prend en compte des effets saisonniers est aussi inhibé par la saisonnalité variable des événements de crue d'une année à l'autre. Cette saisonnalité variable est manifeste pour un événement comme la crue printanière qui se produit à différentes dates selon les années. En effet, l'avènement de cette crue dépend de facteurs météorologiques à la fin de l'hiver tels la température, la précipitation liquide, etc. Il est alors impossible d'utiliser directement des modèles SARIMA ou des modèles qui font appel à des séries de Fourier (Box *et al.*, 1994), puisque les cycles saisonniers ne correspondent pas à des périodes, ou des saisons, fixes.

Afin de s'attaquer à la complexité du processus stochastique qui régit les séries chronologiques de débits, des chercheurs (Salas *et al.*, 1980 ; Salas *et al.*, 1982 ; Vecchia *et al.*, 1983 ; Rasmussen *et al.*, 1996) ont proposé l'utilisation des modèles PARMA pour lesquels les paramètres diffèrent d'une période à l'autre. Il s'avère que ces modèles, bien qu'intéressants, doivent posséder un grand nombre de paramètres afin de reproduire les caractéristiques hydrographiques ; par exemple, pour modéliser des débits à pas de temps hebdomadaire, les utilisateurs vont souvent définir la période comme étant le pas de temps de la série. De plus, des transformations sont habituellement appliquées aux débits à chaque pas de temps pour être en mesure de respecter les hypothèses probabilistes sous-jacentes au modèle.

0.2. Les solutions proposées

À la section précédente, divers aspects de l'hydrologie, nécessaires à la compréhension du contexte dans lequel nos travaux s'inscrivent, ont été abordés. Ces aspects comprennent l'importance de la modélisation des crues, à cause des implications pratiques qui leur sont associées, mais aussi la pertinence de la modélisation des séries de débits d'eau à travers le temps. D'un point de vue statistique, nous avons vu les principaux défis reliés à ces types de modélisation. En ce qui concerne la difficulté de modéliser les caractéristiques des crues telles la pointe et le volume, le choix des distributions statistiques a été signalé ; il a aussi été indiqué que la modélisation de la forme des hydrogrammes de crue représente un défi lors d'études de crue. Pour sa part, la modélisation des séries chronologiques de débits d'eau est ardue à cause de la non-stationnarité des débits à travers le temps et de la périodicité variable des événements importants d'une année à l'autre.

L'approche que nous proposons est basée sur l'analyse des données fonctionnelles. Elle permet de résoudre les problèmes qui rendent l'emploi des approches usuelles des séries chronologiques inapplicables et d'obtenir un modèle pour représenter une forme type d'hydrogramme annuel en un certain site. De plus, le cadre probabiliste développé nous permet de considérer plusieurs distributions statistiques et ainsi, d'aborder le problème associé aux distributions statistiques qui sont derrière le processus qui régit les débits. Les prochaines sections présentent les outils de modélisation à la base des travaux effectués dans les trois articles.

0.2.1. La synchronisation

En présence d'un échantillon de courbes qui possèdent des caractéristiques structurelles similaires, au niveau de la forme, mais qui affichent des différences par rapport aux endroits du domaine où ces caractéristiques se produisent, il est souvent utile de synchroniser les courbes afin d'obtenir un profil représentatif de la structure fonctionnelle de l'échantillon. Ainsi, sous l'hypothèse qu'il existe un processus commun qui génère aléatoirement les courbes, à une transformation

temporelle près, la synchronisation peut être employée pour rendre les courbes plus homogènes temporellement. Du point de vue mathématique, une courbe observée est alors le résultat d'une fonction génératrice des formes de la courbe qui est appliquée à une fonction qui est à l'origine des variations temporelles. Sous cet angle, la synchronisation modélise l'inverse de la fonction génératrice de la variabilité temporelle.

Pour effectuer la synchronisation d'un échantillon de courbes, nous devons postuler un modèle pour les fonctions de synchronisation individuelles. Plusieurs méthodologies existent pour synchroniser des courbes et il est possible de faire appel à des méthodes déterministes ou statistiques. Les méthodes déterministes traitent les fonctions de synchronisation (ou leurs inverses) comme exactes, c'est-à-dire qu'elles ne sont pas considérées provenir d'un processus stochastique. Pour leur part, les méthodes statistiques font l'hypothèse que les fonctions génératrices de la variabilité temporelle sont issues d'un processus stochastique.

Parmi les approches de modélisation qui font appel à des fonctions de synchronisation déterministes, nous pouvons faire la distinction entre la synchronisation ponctuelle et la synchronisation fonctionnelle. La synchronisation ponctuelle fait appel à des points caractéristiques, ou des événements ponctuels, présents pour toutes les courbes de l'échantillon afin d'effectuer la synchronisation (Kneip et Gasser, 1992 ; Ramsay et Silverman, 2005), alors que la synchronisation fonctionnelle utilise toute l'information des courbes lors de la synchronisation. Plus précisément, les méthodes qui ont recours à ce dernier type de synchronisation font appel à des algorithmes d'optimisation qui prennent en compte tous les points des courbes lors de leur application. La synchronisation fonctionnelle est utilisée par les méthodes de modélisation de Kneip et Engel (1995), fondées sur une approche qui fait appel à l'invariance de la forme des courbes où chacune de celles-ci est une transformation affine d'un profil moyen ; de Ramsay et Li (1998), basée sur une famille de fonctions monotones et continues pour effectuer la synchronisation ; de Kneip *et al.* (2000), où les fonctions de synchronisation sont modélisées par la régression polynomiale locale ; de Gervini et Gasser (2004) qui utilisent une base de fonctions B-spline (voir section 0.2.2) pour les fonctions de synchronisation.

En ce qui concerne les méthodes de modélisation qui ont recours à des approches statistiques pour les fonctions de synchronisation, Ronn (2001) considère des translations temporelles aléatoires dans un contexte de maximisation de vraisemblance non paramétrique. Brumback et Lindstrom (2004) développent des fonctions de synchronisation aléatoires où celles-ci sont modélisées par une base de fonctions B-spline, alors que Liu et Muller (2004) proposent l'utilisation de fonctions de synchronisation monotones et stochastiques. Finalement, Telesca et Inoue (2008) mettent de l'avant un modèle bayésien hiérarchique, construit selon une méthodologie similaire à celle de Brumback et Lindstrom (2004), mais où les coefficients de la base des fonctions B-spline qui modélisent les fonctions de synchronisation sont aléatoires.

La synchronisation utilisée dans nos travaux est une synchronisation ponctuelle, c'est-à-dire qu'elle est effectuée à partir de points caractéristiques, ou d'événements ponctuels, qui sont présents pour toutes les courbes de l'échantillon. De plus, nous faisons appel à des fonctions de synchronisation qui sont linéaires par parties. Ce premier choix s'explique par le fait que les hydrogrammes annuels possèdent des points caractéristiques bien définis tels que la pointe de la crue printanière et celle de la crue automnale. Le second choix résulte d'une analyse de la conservation des volumes en apport calculés avant et après la synchronisation (Merleau *et al.*, 2005). En particulier, cette analyse indique que les volumes de crue printanière des hydrogrammes observés sont bien reproduits après une synchronisation effectuée avec des fonctions linéaires par parties. Par exemple, Ramsay et Li (1998), étudiant une forme de synchronisation fonctionnelle, donnent des exemples pour lesquels des fonctions de synchronisation possédant une courbure prononcée peuvent engendrer de sérieuses déformations des courbes observées, ce qui doit être évité dans notre contexte.

Notre méthode de synchronisation est exacte puisque chaque fonction linéaire de la fonction de synchronisation est complètement déterminée par le début et la fin des hydrogrammes, ainsi que les points caractéristiques utilisés, qui sont les temps associés à la pointe de la crue printanière et à celle de la crue automnale. Bien qu'il soit possible d'employer d'autres points caractéristiques tels

que le début et la fin des crues, cet exercice n'a pas été réalisé dans les présents travaux puisque les pointes des crues sont facilement identifiables comparativement aux débuts et aux fins des crues qui peuvent parfois causer des difficultés d'identifiabilité.

En utilisant les points caractéristiques discutés précédemment et les fonctions linéaires par parties, la condition de monotonicité des fonctions de synchronisation est respectée puisque les événements employés sont toujours dans le même ordre chronologique.

0.2.2. Les splines de régression

Afin de modéliser des courbes dans un contexte non paramétrique, un instrument de modélisation doit être choisi. Ici, nous utilisons le terme non paramétrique dans son acceptation usuelle, c'est-à-dire qu'une méthode non paramétrique ne présuppose pas une forme déterminée pour la fonction à modéliser. Bien que certaines méthodes que nous abordons dans ce qui suit sont parfois appelées semi-paramétriques, nous ne ferons pas cette distinction pour ne pas embrouiller l'exposé.

La modélisation non paramétrique a connu un essor fulgurant, durant les vingt dernières années, avec le développement de plusieurs méthodes telles que les splines de lissage (Hastie et Tibshirani, 1990 ; Wahba, 1990 ; Green et Silverman, 1994), les splines de régression (Smith et Khon, 1996 ; Denison *et al.*, 1998), les splines pénalisées de régression (Eilers et Marx, 1996 ; Ruppert *et al.*, 2003) et les ondelettes (Donoho et Johnston, 1994 ; Ogden, 1997). L'outil de modélisation que nous utilisons dans nos travaux sont les splines de régression puisqu'elles représentent un outil non paramétrique très efficace au niveau de la parcimonie, une propriété importante dans notre contexte. De plus, elles possèdent de bonnes propriétés mathématiques telles que la différentiabilité et l'intégrabilité. Avant de discuter plus spécifiquement des splines de régression, une mise en contexte de la modélisation avec les splines est présentée afin de faire ressortir les particularités des différentes approches.

L'évolution de la modélisation avec les splines s'est fait selon deux axes : les splines de lissage et les splines de régression. La notion de noeuds intérieurs, qui définissent les éléments d'une base de fonctions spline, se situe au coeur de l'utilisation des splines. En effet, ce sont le nombre de noeuds intérieurs et l'emplacement de ceux-ci qui déterminent la structure des éléments de la base et donc, la flexibilité du modèle. Bien qu'il soit possible d'utiliser le concept de multiplicité des noeuds intérieurs, c'est-à-dire des noeuds intérieurs répétés (voir Smith, 1979 ; de Boor, 1978, chapitre 7), afin de modéliser des discontinuités au niveau des dérivées d'une fonction cible, nous ne discutons pas de cet aspect puisque les courbes que nous cherchons à modéliser sont continues. Ainsi, nous nous concentrons sur le nombre et l'emplacement des noeuds intérieurs dans les paragraphes qui suivent.

Les splines d'interpolation, telles que définies usuellement (voir de Boor, 1978 ; Green et Silverman, 1994), représentent la solution à un problème d'optimisation qui est défini par la nécessité du modèle de passer par tous les points, de satisfaire des contraintes de continuité (au niveau des dérivés d'ordre 0, 1 et 2) et de minimiser un terme global de pénalité relié à la courbure de la modélisation, telle que mesurée par la seconde dérivée (voir chapitre 1 de Green et Silverman, 1994). La solution de ce problème est donnée par des splines cubiques naturelles pour lesquelles les noeuds intérieurs sont positionnés à tous les points de mesure.

De la même façon que les splines d'interpolation, les splines de lissage sont obtenues comme la solution à un problème d'optimisation avec les mêmes contraintes de continuité que celles des splines d'interpolation. Par contre, la modélisation avec les splines de lissage n'est pas requise de reproduire les observations, mais elle doit plutôt passer près de ces points selon une certaine mesure de distance, tout en minimisant un terme global de pénalité, relié à la courbure, mais qui dépend d'un paramètre de lissage. Pour une mesure de distance et un paramètre de lissage fixés, la solution à ce problème est aussi une base de modélisation donnée par les splines cubiques naturelles, pour laquelle les noeuds intérieurs sont situés à chacune des observations. Le paramètre de lissage détermine la magnitude des différents coefficients qui multiplient les éléments de la base de modélisation et

il joue donc un rôle crucial. Le choix de ce paramètre se fait habituellement en ayant recours à des méthodes telles la validation croisée, la validation croisée généralisée, etc. (voir par exemple Hastie et Tibshirani, 1990 ; Green et Silverman, 1994).

Dans le contexte de la modélisation avec les splines de régression, l'objectif est de déterminer le nombre de noeuds intérieurs et leur emplacement afin d'obtenir une base de fonctions spline qui est adéquate pour modéliser un jeu de données. Pour ce faire, plusieurs types de fonctions spline peuvent être utilisées. Les fonctions spline tronquée (de Boor, 1978 ; Denison *et al.*, 1998 ; Ruppert *et al.*, 2003), les fonctions B-spline (de Boor, 1978 ; He et Shi, 1998 ; DiMatteo *et al.*, 2001), les fonctions M-spline (Ramsay, 1988) constituent toutes des exemples de type de fonctions spline. Il existe des liens étroits entre ces différents types de fonctions. Par exemple, il est possible de passer de la base des fonctions spline tronquée à celle des fonctions B-spline par des transformations linéaires (voir de Boor, 1978 ; Ruppert *et al.*, 2003). En ce qui concerne les fonctions B-spline et M-spline, elles diffèrent par leur facteur de normalisation (Ramsay, 1988).

Eilers et Marx (1996), et plus récemment Ruppert *et al.* (2003), ont développé les splines pénalisées de régression. Les premiers auteurs utilisent la base des fonctions B-spline, alors que ces derniers utilisent la base des fonctions spline tronquée. L'idée sous-jacente aux splines pénalisées de régression est d'utiliser un bon nombre de noeuds intérieurs, mais cependant moins que dans le cas des splines de lissage (c'est-à-dire à tous les points observés), et de pénaliser un sous-ensemble des coefficients associés aux éléments de la base de modélisation. Ces méthodes constituent naturellement un compromis entre les splines de lissage et les splines de régression.

Nous choisissons de travailler avec les splines de régression avec une base de fonctions M-spline. En intégrant chacune des fonctions de la base M-spline, nous obtenons la base de fonctions I-spline qui sont des fonctions monotones croissantes (Ramsay, 1988). Cette propriété est intéressante dans notre contexte de modélisation car il existe aussi un intérêt pratique pour les hydrogrammes cumulatifs qui représentent l'apport total entre deux pas de temps donnés. En

ce qui concerne la détermination du nombre de noeuds et de leur emplacement, nous discutons de différentes méthodologies proposées dans la littérature et des approches que nous utilisons aux sections 0.3.1 et 0.3.2.

0.2.3. La famille de distributions statistiques

La modélisation statistique de données fonctionnelles est souvent effectuée en faisant l'hypothèse que les données proviennent de la loi normale (voir par exemple Ramsay et Silverman, 2005). Au niveau probabiliste, le cadre de modélisation est alors similaire à celui de la régression multiple ; cependant, la relation mathématique, ou la composante systématique, utilisée pour modéliser la tendance centrale des données est plus flexible, faisant appel à une des méthodes non paramétriques discutées à la section précédente. Dans un contexte de régression, la transformation de la variable dépendante, telle que proposée par Box et Cox (1964), permet de préserver l'hypothèse de la normalité pour une transformation adéquate. Par exemple, c'est cette approche que Smith et Khon (1996) utilisent pour leur modèle non paramétrique basé sur les splines de régression.

Les modèles linéaires généralisés (Nelder et Wedderburn, 1972 ; McCullagh et Nelder, 1989) constituent un prolongement du modèle de régression traditionnel. En effet, ceux-ci permettent la modélisation d'une variable aléatoire provenant d'une distribution statistique appartenant à la famille exponentielle. Plus précisément, le paramètre canonique, ou une fonction de celui-ci, qui définit la distribution statistique est modélisé par une fonction de variables auxiliaires. À cause des propriétés de différentiabilité de la famille exponentielle, il est possible de traiter d'une façon uniforme l'estimation des paramètres qui interviennent linéairement dans la composante systématique.

Dans un contexte non paramétrique, les modèles additifs généralisés (Hastie et Tibshirani, 1990) élargissent les possibilités offertes par les modèles linéaires généralisés au niveau des composantes systématiques. Ainsi, ce premier type de modèles permet de considérer des fonctions des variables auxiliaires qui ne sont pas nécessairement linéaires. Plus précisément, un modèle non paramétrique est

utilisé pour chacune des variables auxiliaires séparément et la somme de ces différents modèles donne la réponse globale pour reproduire la variable d'intérêt. Les splines de lissage (Hastie et Tibshirani, 1990 ; Green et Silverman, 1994), ainsi que les splines pénalisées de régression (Ruppert *et al.*, 2003) sont des outils non paramétriques employés avec ce type de modèles.

Tout comme dans le cas des modèles additifs généralisés, nous cherchons à modéliser un jeu de données avec une composante systématique non paramétrique dans un cadre probabiliste assez général. En effet, nous visons à étudier diverses hypothèses probabilistes au niveau des observations puisqu'il n'y a pas de consensus quant à la distribution statistique la plus adéquate pour modéliser les débits d'eau. Dans notre contexte, les données à modéliser sont continues et nous considérons une famille de distributions statistiques adaptée à ce type de données. Sous leur formulation originale, les modèles linéaires, ou additifs, généralisés incorporent trois distributions statistiques pour les variables continues, soient les distributions normale, gamma et inverse gaussienne. Afin d'explorer un plus grand éventail de distributions, nous considérons aussi des transformations des variables aléatoires issues de ces trois distributions ; par exemple, ceci nous permet de traiter les distributions lognormale et réciproque inverse gaussienne.

0.3. Contributions des différents articles

La présente section indique la contribution scientifique des travaux effectués pour chacun des trois articles rédigés dans le cadre de la thèse.

0.3.1. Article 1 : Bayesian modeling of hydrographs

Le premier article, paru dans la revue Water Resources Research (Merleau, J., Perreault, L., Angers, J.-F., et Favre, A.-C. (2007). Bayesian modeling of hydrographs, Water Resources Research 43, W10432, doi :10.1029/2006WR005376), propose une nouvelle approche pour modéliser les hydrogrammes annuels ou les débits d'eau au cours d'une année. En introduction, l'article présente les différentes méthodes existantes pour créer des hydrogrammes synthétiques de crue ou des séries synthétiques de débits d'eau. Une fois ce tour d'horizon effectué, la

méthodologie statistique que nous proposons est mise de l'avant. Elle repose sur l'analyse de données fonctionnelles (Ramsay et Silverman, 2005) dans un cadre bayésien (Robert, 2001). Les outils de l'analyse de données fonctionnelles que nous utilisons sont la synchronisation et les splines de régression (voir sections 0.2.1 et 0.2.2), alors que le cadre bayésien est utilisé au niveau de la structure probabiliste du modèle.

Dans le domaine de l'hydrologie statistique, la méthodologie exposée dans l'article est totalement nouvelle puisqu'aucun travail antérieur n'utilise les méthodes de l'analyse de données fonctionnelles afin de modéliser les hydrogrammes. En adoptant ce point de vue, nous sommes en mesure de considérer les hydrogrammes annuels comme un échantillon de courbes à modéliser statistiquement. Puisque les courbes annuelles affichent des caractéristiques typiques se produisant à différents moments selon les années, nous pouvons rendre ces courbes similaires au niveau temporel en faisant appel à la synchronisation des courbes. Après avoir synchronisé les courbes, nous possédons un échantillon plus homogène au niveau temporel et nous procédons à la modélisation du comportement moyen de cet échantillon d'hydrogrammes avec des splines de régression bayésiennes. Ceci nous permet d'obtenir un modèle parcimonieux, dans l'espace des coefficients des fonctions splines, qui représente la forme moyenne d'un hydrogramme provenant d'un site donné.

Le modèle probabiliste bayésien est construit en faisant l'hypothèse que les débits d'eau formant les hydrogrammes sont conditionnellement indépendants et que la distribution statistique qui les génère est une distribution normale multivariée. Pour les paramètres des fonctions spline et la variance, nous utilisons des distributions *a priori* conjuguées en nous basant sur les travaux en régression linéaire bayésienne (Raiffa et Schlaifer, 1961 ; Zellner, 1971). La distribution *a priori* des paramètres est une loi normale multivariée, alors que celle de la variance est une distribution inverse gamma. Les hyperparamètres de ces lois sont déterminés à partir d'information historique de moindre qualité. En ce qui concerne la matrice de covariance des coefficients des fonctions spline, nous employons la prescription

de Zellner (1986), pour laquelle la matrice est proportionnelle à la variance des observations et dépend de la matrice d'incidence du modèle de régression.

Comme mentionné à la section 0.2.2, une des difficultés avec l'utilisation des splines de régression concerne la détermination de l'emplacement des noeuds intérieurs qui définissent les éléments de la base de modélisation. Puisque les éléments de la base déterminent la modélisation et par conséquent, l'adéquation du modèle, l'emplacement des noeuds intérieurs est fort important. Dans un cadre de modélisation fréquentiste, des auteurs tels Friedman et Silverman (1989) et Stone *et al.* (1997) ont proposé des méthodes itératives analogues à la régression pas-à-pas. En ce qui concerne les approches bayésiennes, elles sont traitées à la section 0.3.2.

Dans cet article, nous faisons appel à une stratégie développée par He et Shi (1998) dans un contexte de modélisation de fonctions monotones croissantes avec des fonctions B-spline. Pour un nombre de noeuds fixé, les auteurs proposent d'utiliser le profil de la fonction monotone croissante cible pour effectuer une projection des quantiles de la variable dépendante sur l'axe de la variable indépendante. Puisque les noeuds intérieurs doivent être positionnés à des endroits de la courbe à modéliser où les variations sont les plus importantes, cette méthodologie permet d'atteindre cet objectif.

Dans notre contexte, puisque les hydrogrammes sont des fonctions strictement positives, chaque hydrogramme cumulatif est par conséquent une fonction monotone croissante. Ainsi, il nous est possible d'utiliser l'approche de He et Shi (1998) à l'hydrogramme cumulatif moyen pour déterminer l'emplacement des noeuds lorsque le nombre de noeuds est fixé. Bien que le nombre de noeuds intérieurs et leur emplacement ne soient pas considérés comme des quantiés aléatoires, nous explorons tout de même différentes configurations nodales avec la méthodologie exposée au paragraphe précédent. Afin de déterminer une configuration «optimale» parmi celles explorées, nous utilisons les facteurs de Bayes (Jeffreys, 1961 ; Kass et Raftery, 1995) qui constituent une méthode de sélection qui découle directement des hypothèses probabilistes du modèle bayésien.

L'approche proposée est appliquée à deux jeux de données, soient les hydrogrammes de deux bassins versants : Churchill Falls qui se trouve dans le nord du Québec et Gouin qui se trouve plus au sud. Globalement, la méthodologie fonctionne bien puisque nous sommes en mesure d'obtenir des hydrogrammes de référence adéquats avec la synchronisation et de les modéliser ensuite avec le modèle bayésien proposé. Par contre, certains éléments demeurent problématiques. Les intervalles de confiance résultant du modèle bayésien semblent trop larges pour certaines parties du domaine temporel, alors qu'ils sont possiblement trop étroits pour la crue printanière. Cet effet provient de l'hypothèse concernant la variance constante à travers l'année, c'est-à-dire l'homoscédasticité du processus. De plus, bien que la méthode pour étudier des configurations de noeuds intérieurs permette d'obtenir une bonne adéquation entre le modèle et les hydrogrammes de référence au niveau de la crue printanière, la crue automnale n'est pas aussi bien approchée par le modèle. Ce problème vient du fait que l'approche de détermination des noeuds ne dépend réellement que du nombre de noeuds et du profil de l'hydrogramme cumulatif moyen, ce qui limite l'exploration des configurations nodales.

0.3.2. Article 2 : Modelling the average behaviour of a sample of curves with Bayesian regression splines

Le deuxième article, soumis à la revue canadienne de statistique au mois de juin 2009, met de l'avant un contexte de modélisation statistique plus général que celui présenté dans le premier article. Plus précisément, nous étudions une famille de distributions continues au niveau des observations qui inclut les distributions de la famille exponentielle et par conséquent, la distribution normale utilisée dans l'article 1. De plus, nous traitons maintenant le nombre de noeuds intérieurs et leur emplacement comme des quantités aléatoires à l'intérieur du modèle bayésien. Ces deux prolongements au modèle développé dans le premier article nous permettent d'adresser certaines limitations de ce dernier. Dans un premier temps, les distributions continues de la famille étudiée peuvent posséder une dépendance entre la variance et la moyenne, ainsi des intervalles de confiance

plus réalistes peuvent être obtenus. Dans un second temps, pour ajouter de la flexibilité à la modélisation de la composante systématique, les noeuds intérieurs sont considérés comme des quantités aléatoires, ce qui améliore la modélisation de l'hydrogramme de référence.

Dans le domaine de la modélisation statistique, la principale contribution de l'article est l'utilisation d'une famille assez générale de distributions continues dans le contexte de modélisation avec des splines de régression où le nombre de noeuds intérieurs et l'emplacement de ceux-ci sont considérés comme aléatoires. Traitant les noeuds intérieurs comme des quantités aléatoires avec un support à dimension variable, nous faisons appel au MCMC avec sauts réversibles (Green, 1995) mais en simplifiant cette procédure afin d'éviter les difficultés reliées aux changements de dimension. Bien que les articles de Denison *et al.* (1998) et de DiMatteo *et al.* (2001) proposent des méthodologies similaires dans le cas d'observations provenant de la distribution normale et de celle de Poisson, nous prolongeons ces approches à notre contexte de modélisation de données hydrologiques continues.

La famille de distributions étudiée possède la forme de la famille exponentielle mais elle permet la possibilité de transformer la variable d'intérêt. Ainsi, dans notre contexte de modélisation hydrologique, nous sommes en mesure de traiter d'une façon uniforme les distributions suivantes : normale ($\mathcal{N}$), gamma ($\mathcal{G}$), inverse gaussienne (IG), lognormale ($\mathcal{LN}$) et réciproque inverse gaussienne (RIG). Les trois premières distributions appartiennent à la famille exponentielle, alors que les deux dernières sont obtenues par une transformation d'une variable provenant d'une distribution appartenant à la famille exponentielle. En hydrologie, la plupart de ces distributions sont utilisées pour modéliser des débits d'eau (voir Gumbel, 1958 ; Bobée et Ashkar, 1991 ; Rao et Hamed, 2000) et de plus, chacune possède une dépendance spécifique entre la moyenne et la variance du processus (McCullagh et Nelder, 1989 ; Jorgensen, 1997). Nous cherchons à étudier plusieurs distributions dans un contexte de modélisation uniforme afin de pouvoir déterminer les distributions les plus adéquates pour modéliser les débits d'eau pour un bassin versant donné (voir section 0.1.2).

Le modèle probabiliste bayésien postule que les observations sont conditionnellement indépendantes et qu'elles sont issues d'une distribution appartenant à la famille décrite au paragraphe précédent. Pour les coefficients des fonctions spline, nous employons une distribution *a priori* normale multivariée avec la même matrice de covariance que dans le premier article (Zellner, 1986), mais ici elle est proportionnelle à l'inverse du paramètre de précision des observations. Nous explorons aussi une matrice de covariance identitaire afin d'étudier l'effet de la matrice de covariance sur la modélisation. Pour le paramètre de précision, une distribution *a priori* gamma est utilisée. De l'information historique de moindre qualité est employée pour fixer les hyperparamètres des distributions *a priori*. Ces distributions *a priori* sont toutes les deux conditionnelles au nombre de noeuds intérieurs et à leur emplacement.

Il est possible, sans faire appel au MCMC avec sauts réversibles, de traiter les configurations nodales d'une façon aléatoire dans un cadre bayésien. Puisque la détermination de l'emplacement des noeuds intérieurs et de leur nombre est un problème similaire à la sélection de variables dans un modèle de régression linéaire, les méthodes bayésiennes pour la sélection de variables automatique via les méthodes MCMC (George et McCulloch, 1993, 1997) peuvent être utilisées pour effectuer la sélection des noeuds intérieurs. C'est cette approche que Smith et Kohn (1996) étudient en employant un paramètre vectoriel dont chacune des composantes binaires indique l'inclusion ou non d'un noeud intérieur.

Notre approche, pour traiter le nombre de noeuds intérieurs et leur emplacement comme des quantités aléatoires, est inspirée par celles développées par Denison *et al.* (1998) et DiMatteo *et al.* (2001). En se basant sur les travaux de Green (1995), Denison *et al.* (1998) indiquent comment le MCMC avec sauts réversibles peut être utilisé dans le contexte des splines de régression. Cependant, ils ne considèrent pas un modèle bayésien complet afin d'éviter les difficultés associées au changement de dimension au niveau des coefficients des fonctions spline. Bien que certains auteurs tels Lindstrom (2002) abordent ces difficultés

en incluant une étape de mise-à-jour des coefficients dans la procédure, nous préconisons la méthodologie proposée par DiMatteo *et al.* (2001) pour sa simplicité conceptuelle et algorithmique.

L'idée derrière cette dernière approche est d'intégrer les coefficients des fonctions spline à chaque étape du MCMC et ainsi, de seulement explorer les configurations des noeuds intérieurs à travers le MCMC. Les deux principaux avantages d'effectuer cette étape analytiquement sont de simplifier la procédure par rapport aux changements de dimension et d'accélérer l'exploration de l'espace des configurations nodales puisque l'espace paramétrique global du problème est réduit. Dans le cadre de la famille exponentielle et des modèles linéaires généralisés, Biller (2004) présente une méthodologie similaire à celle proposée par Lindstrom (2002), où tous les paramètres du modèle sont simulés dans le MCMC.

Afin de simplifier et de rendre la procédure MCMC plus efficace, nous cherchons donc à intégrer la distribution conjointe des observations et des paramètres par rapport aux coefficients des fonctions spline et au paramètre de précision. Pour une série fixe de noeuds, ceci revient à calculer la marginale des observations étant donné les distributions *a priori* des coefficients et du paramètre de précision. Puisque la sélection de modèle en statistique bayésienne se fait généralement à l'aide des facteur de Bayes (Jeffreys, 1961 ; Kass et Raftery, 1995), qui sont des rapports de distributions marginales, beaucoup de travaux existent dans la littérature sur le calcul des marginales dans différents contextes de modélisation. Dans le cas du modèle linéaire pour lequel les observations sont distribuées selon une loi normale et que les distributions *a priori* sont conjuguées, l'expression pour la marginale est explicite (Raiffa et Schlaifer, 1961 ; Lempers, 1971 ; Zellner, 1971) ; c'est d'ailleurs cette expression que nous avons utilisée dans le cadre du premier article pour comparer les différentes configurations des noeuds intérieurs.

En ce qui concerne les modèles linéaires généralisés, les distributions marginales ne peuvent être calculées explicitement, et par conséquent, nous faisons appel à des approximations. Plus précisément, nous considérons une approximation basée sur le critère de Schwarz (1978), telle que suggérée par Kass et Wasserman (1995), Kass et Raftery (1995) et DiMatteo *et al.* (2001), et une approximation

qui fait appel à l'approximation de Laplace d'une intégrale (Tierney et Kadane, 1986 ; Tierney *et al.*, 1989 ; Shun et McCullagh, 1995). Nous notons que les travaux récents de Wang et George (2007) utilisent cette dernière approximation pour la sélection de variables dans un contexte de modèles linéaires généralisés, bien que leur approche diffère de la nôtre.

Une fois que les coefficients des fonctions spline sont intégrés, ou qu'une approximation est utilisée, il nous est possible d'effectuer l'intégrale par rapport au paramètre de précision. Ceci ne semble pas être fait habituellement puisque plusieurs auteurs emploient un estimateur du paramètre de précision, ou de dispersion, au lieu de procéder à l'intégration de celui-ci (voir par exemple Raftery, 1996 ; Wang et George, 2007). Ainsi, après avoir obtenu la distribution marginale qui ne dépend maintenant que de la configuration des noeuds intérieurs, celle-ci est utilisée à l'intérieur de la procédure MCMC avec sauts réversibles.

Ayant exploré la distribution *a posteriori* des configurations nodales, nous choisissons le mode de cette distribution comme la configuration la plus adéquate pour modéliser l'hydrogramme de référence. Bien qu'il soit possible d'utiliser toutes les configurations nodales visitées pour obtenir une représentation de l'hydrogramme de référence, nous adoptons l'approche basée sur le mode afin d'avoir une modélisation parcimonieuse au niveau des coefficients des fonctions spline. Pour une distribution statistique donnée, la modélisation de l'hydrogramme de référence correspond donc au mode de la distribution *a posteriori* des noeuds intérieurs.

Pour comparer l'adéquation des différentes distributions statistiques, nous travaillons conditionnellement au mode des configurations nodales. Ainsi, les distributions marginales, ou leurs approximations, discutées précédemment sont utilisées pour calculer des facteurs de Bayes. Nous sommes donc en mesure de discriminer numériquement entre les différentes hypothèses probabilistes au niveau des observations.

La méthodologie mise de l'avant est utilisée afin de modéliser les hydrogrammes du bassin versant Caniapiscau. Les distributions considérées sont les suivantes : normale ($\mathcal{N}$), gamma ($\mathcal{G}$), inverse gaussienne (IG), lognormale ($\mathcal{LN}$)

et réciproque inverse gaussienne (RIG). Pour chacune des distributions étudiées, conditionnellement à la meilleure fonction lien, la modélisation de l'hydrogramme de référence s'avère d'une assez bonne qualité. Cependant, les meilleures modélisations sont obtenues pour les distributions normale, gamma et lognormale. L'utilisation des facteurs de Bayes, telle que discutée ci-haut, révèle que les distributions gamma et lognormale sont les plus adéquates pour reproduire la forme de référence et la variabilité de l'échantillon des hydrogrammes. Au niveau de l'adéquation, il est intéressant de noter que les intervalles de crédibilité approximatifs confirment l'ordonnancement des distributions statistiques obtenu à partir des facteurs de Bayes.

0.3.3. Article 3 : Simultaneous modelling of the mean and dispersion functions of a sample of curves with Bayesian regression splines

Le troisième article, qui sera soumis à la revue Journal of the American Statistical Association au mois de septembre 2009, prolonge les travaux réalisés dans le cadre du deuxième article. Travaillant dans un contexte de modélisation similaire à celui de l'article 2, nous proposons de modéliser simultanément la courbe moyenne et la courbe de dispersion d'un échantillon de fonctions. En utilisant la même famille de distributions continues pour les observations, nous ne faisons plus l'hypothèse que le paramètre de dispersion (l'inverse du paramètre de précision), est le même à travers le domaine des fonctions. Ainsi, la flexibilité du modèle global est accrue et une mauvaise spécification distributionnelle peut être compensée par la modélisation de la courbe de dispersion. L'emplacement et le nombre de noeuds pour la courbe moyenne, ainsi que ceux pour la courbe de dispersion, sont traités comme des quantités aléatoires de la même façon que dans le deuxième article.

En statistique, la principale contribution de l'article est de modéliser simultanément les courbes moyenne et de dispersion d'un échantillon de données fonctionnelles dans un contexte non paramétrique, et ce, pour une famille assez générale

de distributions continues. Dans le cadre des modèles linéaires généralisés, la modélisation simultanée de la tendance centrale et de la dispersion autour de cette tendance centrale à l'aide de modèles linéaires est abordée par Nelder et Pregibon (1987) et au chapitre 10 de McCullagh et Nelder (1989). Efron (1986), de façon indépendante, met de l'avant la famille double exponentielle qui peut aussi être utilisée pour modéliser simultanément la tendance centrale et la dispersion par des modèles linéaires. Notre approche non paramétrique, basée sur les splines de régression, est plus près de la formulation des modèles linéaires généralisés au niveau de la distribution statistique des observations, bien que la famille que nous considérons est plus riche.

Dans un contexte non paramétrique, le chapitre 14 de Ruppert *et al.* (2003) traite de la modélisation simultanée de la tendance centrale et de la variance avec des splines pénalisées de régression sous l'hypothèse de la normalité des observations. Travaillant aussi avec des splines pénalisées de régression, mais avec la famille double exponentielle d'Efron (1986), Nott (2006) propose une méthode pour simultanément modéliser la tendance centrale et la dispersion d'un processus.

Le modèle bayésien utilisé fait l'hypothèse que les observations proviennent d'une distribution appartenant à la même famille de distributions que celle employée dans l'article 2 et qu'elles sont conditionnellement indépendantes. Puisque nous modélisons la tendance centrale et la dispersion avec des splines de régression, nous avons alors deux vecteurs de coefficients qui sont associés aux deux bases de modélisation. Conditionnellement aux deux séries de noeuds intérieurs qui définissent les bases de modélisation, les distributions *a priori* des deux vecteurs de coefficients sont des distributions normales multivariées. L'approche de Zellner (1986) pour spécifier les matrices de covariance de ces deux distributions est généralisée afin de prendre en compte la dispersion variable à travers le domaine des observations. Comme dans le cas des articles 1 et 2, de l'information historique de moindre qualité est employée afin de spécifier les hyperparamètres des distributions *a priori*.

Le traitement des noeuds intérieurs est effectué en utilisant une stratégie similaire à celle de l'article 2. Les deux séries de noeuds intérieurs, qui modélisent la courbe moyenne et la courbe de dispersion, sont considérées comme des quantités aléatoires. Nous utilisons encore le MCMC avec sauts réversibles (Green, 1995) qui fait appel aux distributions marginales des observations. Les distributions marginales sont maintenant obtenues en intégrant les deux vecteurs de coefficients associés aux deux bases de modélisation. Pour les mêmes raisons que celles explicitées dans le cadre de l'article 2, nous devons employer des approximations pour les distributions marginales. Nous considérons une approximation basée sur le critère de Schwarz et une autre fondée sur l'approximation de Laplace d'une intégrale.

Dans la procédure du MCMC avec sauts réversibles, les deux configurations nodales sont générées à chaque itération puisque la modélisation de la courbe moyenne et celle de la dispersion sont intrinsèquement liées. Après avoir exploré les distributions *a posteriori* des noeuds intérieurs, nous choisissons les modes comme la modélisation simultanée la plus adéquate pour reproduire la courbe de la tendance centrale et celle de la dispersion. L'utilisation de l'approximation de Laplace est équivalente à considérer que la distribution *a posteriori* des paramètres des deux composantes systématiques est une loi normale multivariée. Ainsi, pour le mode des configurations nodales, nous sommes en mesure de construire des intervalles de confiance pour l'échantillon de courbes et pour la courbe de dispersion. Finalement, en adoptant la même stratégie que pour l'article 2, nous pouvons comparer l'adéquation des diffŕentes distributions statistiques via le rapport des distributions marginales évaluées à leur mode respectif pour les noeuds intérieurs.

La méthode proposée est mise en oeuvre pour modéliser le même jeu de données que celui étudié dans l'article 2, c'est-à-dire les hydrogrammes du bassin versant Caniapiscau. Il est intéressant de noter qu'avec cette méthode, il est maintenant possible d'obtenir des intervalles de confiance adéquats pour l'échantillon d'hydrogrammes sous l'hypothèse des distributions normale ($\mathcal{N}$), inverse

gaussienne (IG) et réciproque inverse gaussienne (RIG). Cependant, les distributions lognormale ($\mathcal{LN}$) et gamma ($\mathcal{G}$) constituent les meilleures distributions selon les facteurs de Bayes.

Chapitre 1

CONTEXTE HYDROLOGIQUE ET HYDROLOGIE STATISTIQUE

Dans ce chapitre, nous revenons sur les concepts hydrologiques introduits à la section 0.1 et sur certaines des méthodes de modélisation de l'hydrologie statistique décrites à la section 0.1.2. Ce chapitre a pour but de mettre de l'avant la notation mathématique employée et d'expliciter certains modèles rencontrés en hydrologie statistique. Il est important de noter que ce chapitre est loin d'être exhaustif puisque l'hydrologie statistique constitue maintenant un domaine scientifique à part entière (voir par exemple, Rao et Hamed, 2000 ; Favre *et al.*, 2008). Il se veut plutôt une entrée en matière afin d'introduire certains concepts de base.

1.1. Les données hydrologiques étudiées

Un hydrogramme, tel que défini à la section 0.1, est une série chronologique qui représente le débit d'eau en un certain lieu. Dans nos travaux, nous nous concentrons sur les hydrogrammes annuels en un endroit déterminé, nous notons l'hydrogramme de l'année i par le vecteur $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in})'$, où $j(= 1, \ldots, n)$ représente l'indice de la mesure ; le débit y_{ij} possède le m^3/s comme unité de mesure. Nous supposons que les mesures sont effectuées à un pas de temps régulier et qu'il n'existe pas de données manquantes ; l'indice de mesure peut alors être utilisé comme l'incrément temporel. Nous notons la première mesure de l'année par $j = 1$ et elle correspond au premier jour julien, ou à la première semaine de l'année, selon que l'hydrogramme possède un pas de temps quotidien

ou hebdomadaire. Dans ce contexte, n représente la dernière mesure de l'année et détermine aussi le pas de temps de l'hydrogramme annuel ; par exemple, $n = 365$ pour un hydrogramme annuel à pas de temps quotidien. Finalement, nous avons à notre disposition plusieurs hydrogrammes annuels et le nombre d'années est donné par N ; nous avons alors $i = 1, \ldots, N$. Pour un certain site, nous écrivons l'ensemble des hydrogrammes annuels par le vecteur $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_i', \ldots, \boldsymbol{y}_N')'$.

Dans le contexte de la modélisation non paramétrique des hydrogrammes, nous considérons le temps comme une variable continue. Nous utilisons alors un vecteur temporel, défini par $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{in})'$, qui représente les instants de mesure associés à l'hydrogramme annuel $\boldsymbol{y}_i$. Puisque nous étudions seulement des hydrogrammes annuels avec des pas de temps réguliers qui sont les mêmes à travers les années, nous avons $\boldsymbol{x}_i = \boldsymbol{x} = (x_1, \ldots, x_j, \ldots, x_n)'$, $\forall i$.

La figure 1.1 illustre l'hydrogramme annuel à pas de temps hebdomadaire de l'année 1961 pour le site Caniapiscau situé dans le nord du Québec. Le premier temps de cette série représente la première semaine de janvier, alors que le dernier temps représente la dernière semaine de décembre. La partie supérieure de la figure 1.1 donne les principaux régimes observés au cours d'une année, c'est-à-dire la période hivernale, la crue printanière et la crue automnale. Ces régimes proviennent des conditions météorologiques qui existent au Québec.

Durant la période hivernale, le débit d'eau décroit à cause de l'accumulation des précipitations sous forme de neige et de glace. Lors de l'accroissement des températures à l'arrivée du printemps, une forte augmentation du débit se produit. Le débit demeure important jusqu'au moment où la fonte des neiges est complète. Cette période correspond à la crue printanière. Le débit demeure assez stable durant la période estivale sauf lors de précipitations importantes. Finalement, l'automne est habituellement caractérisé par une croissance du débit lors de fortes précipitations. Cette période représente la crue automnale. La figure 1.1 montre les caractéristiques importantes des crues. La pointe et le volume de la crue printanière y sont indiqués, ainsi que la pointe de la crue automnale. Finalement, le début et la fin de la crue printanière, qui déterminent la durée de celle-ci, sont également illustrés.

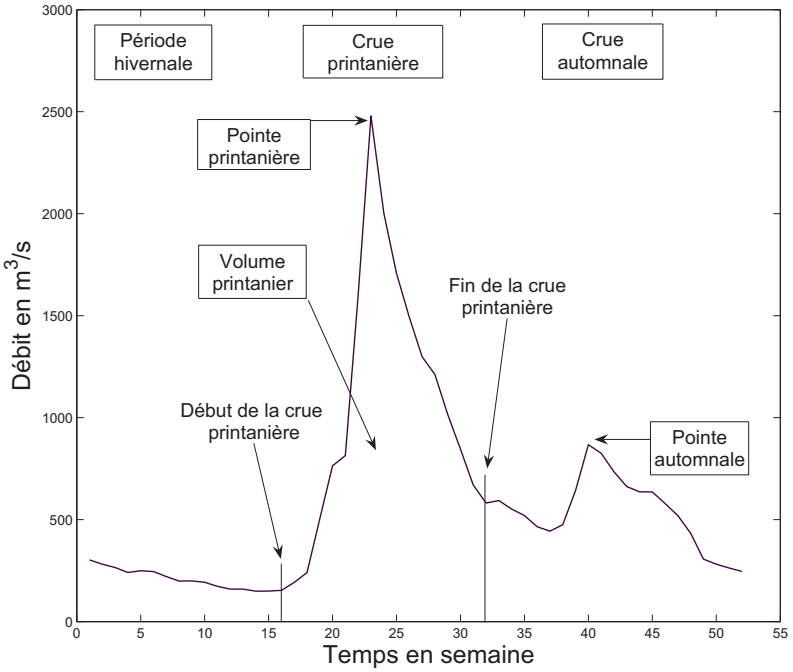


FIGURE 1.1. Exemple d'un hydrogramme annuel : l'année 1961 pour le site Caniapiscau. Les principales caractéristiques d'un hydrogramme annuel sont indiquées.

1.2. HYDROLOGIE STATISTIQUE

Maintenant que nous avons défini d'une façon plus pratique les caractéristiques des hydrogrammes et des crues, nous décrivons plus précisément certaines des méthodes utilisées en hydrologie statistique.

1.2.1. Modélisation des crues

À la section 0.1.2.1, la modélisation statistique des crues est discutée. En se concentrant sur les crues printanières, nous pouvons maintenant expliciter les quantités importantes qui définissent une crue, nommément la pointe de crue et le volume de crue. La pointe de crue est définie comme la valeur la plus élevée des débits durant la crue (voir figure 1.1), alors que le volume de crue correspond à la somme des débits durant la crue. En définissant j_i^d et j_i^f comme les indices correspondant aux débuts et à la fin de la crue printanière pour l'hydrogramme

annuel i, nous avons alors la pointe printanière

$$y_i^p = \max_j \{y_{ij}\}, \text{ où } j = j_i^d, \ldots, j_i^f, \tag{1.2.1}$$

et le volume de crue

$$y_i^v = \sum_{j=j_i^d}^{j_i^f} y_{ij}. \tag{1.2.2}$$

Ainsi, nous obtenons un vecteur de pointes printanières, $\boldsymbol{y}^p = (y_1^p, \ldots, y_i^p, \ldots, y_N^p)'$, et un vecteur de volumes de crue printanière, $\boldsymbol{y}^v = (y_1^v, \ldots, y_i^v, \ldots, y_N^v)'$, pour un site donné.

Une analyse fréquentielle de crue peut être effectuée à partir des deux vecteurs $\boldsymbol{y}^p$ et $\boldsymbol{y}^v$. Dans le cas où les pointes et les volumes sont traités comme des quantités indépendantes, chacune de ces quantités est modélisée par une loi de probabilité, soient les distributions L_p et L_v pour les pointes et les volumes respectivement. Il est alors possible de calculer des quantiles possédant certaines périodes de retour pour chacune de ces variables. En notant les fonctions de répartition des variables aléatoires y^p, la pointe, et y^v, le volume, par $L_p(q^p) = P(y^p \leq q^p)$ et $L_v(q^v) = P(y^v \leq q^v)$, les périodes de retour pour les quantiles q^p et q^v sont respectivement données par

$$T_{q^p} = \{1 - L_p(q^p)\}^{-1}, \tag{1.2.3}$$

$$T_{q^v} = \{1 - L_v(q^v)\}^{-1}. \tag{1.2.4}$$

Par conséquent, une fois que les lois de probabilité L_p et L_v sont déterminées, il est possible de déterminer des quantiles correspondant à des périodes de retour fixes. Par exemple, le quantile associé à une pointe possédant une période de retour de 10000 ans, $T_{q^p} = 10000$, peut être calculé à partir de L_p. Les quantiles avec des périodes de retour élevées sont ainsi calculés pour les pointes et les volumes d'un bassin versant donné.

Finalement, comme il a été indiqué à la section 0.1.2.1, il est ensuite nécessaire de construire des hydrogrammes synthétiques qui possèdent les propriétés des quantiles de crue. Dans le cas de la méthode de l'hydrogramme type, un hydrogramme observé est utilisé afin d'avoir une forme spécifique de référence ;

par exemple, l'hydrogramme de la figure 1.1 pourrait être employé comme hydrogramme type. Afin d'ajuster la crue printanière de cet hydrogramme pour qu'elle se conforme aux quantiles calculés, une transformation d'échelle est appliquée à celui-ci.

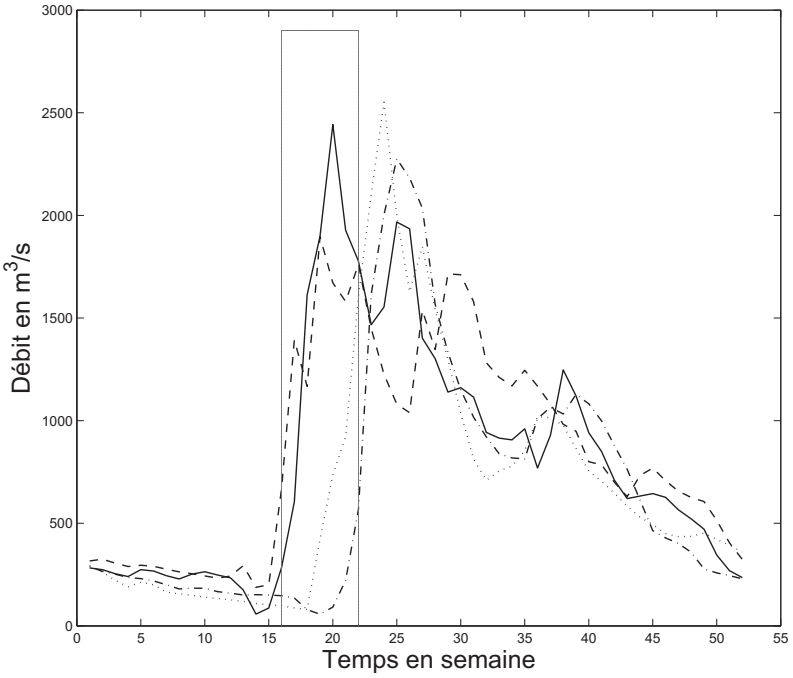


FIGURE 1.2. Saisonnalité variable des événements importants des hydrogrammes annuels. Les hydrogrammes des années 1972 (tiret-pointillé), 1982 (pointillé), 1986 (trait plein) et 1987 (tiret) pour le site Caniapiscau.

Le choix de l'hydrogramme type, c'est-à-dire d'un hydrogramme annuel particulier, demeure un choix arbitraire puisqu'une seule réalisation est considérée. À cause de la saisonnalité variable des événements importants d'une année à l'autre (voir section 0.1.2.2), il n'est pas possible de calculer directement une moyenne des hydrogrammes annuels observés afin d'obtenir un hydrogramme type. La figure 1.2 illustre le problème. Quatre hydrogrammes annuels pour le site Caniapiscau y sont présentés, soient les hydrogrammes pour les années 1972, 1982, 1986 et 1987. Nous voyons qu'en prenant la moyenne des quatre hydrogrammes à l'intérieur du rectangle, nous mélangeons des débits qui sont hétérogènes. Plus spécifiquement,

les hydrogrammes des années 1986 et 1987 sont en pleine crue printanière, alors que ceux des années 1972 et 1982 sont en début de crue printanière. Ainsi, à partir des hydrogrammes observés, il est impossible d'obtenir directement un hydrogramme moyen qui posséderait une forme type des hydrogrammes d'un bassin versant donné. C'est afin d'obtenir cette forme plus réaliste que nous utilisons la synchronisation.

1.2.2. Modélisation des séries chronologiques de débits

À la section 0.1.2.2, nous avons discuté des caractéristiques des débits d'eau qui rendent difficile la modélisation de ceux-ci avec les modèles usuels des séries chronologiques. Ces deux principales caractéristiques sont la saisonnalité variable des événements importants d'une année à l'autre (voir figure 1.2) et la non-stationnarité des débits durant l'année à cause de la présence de différents régimes.

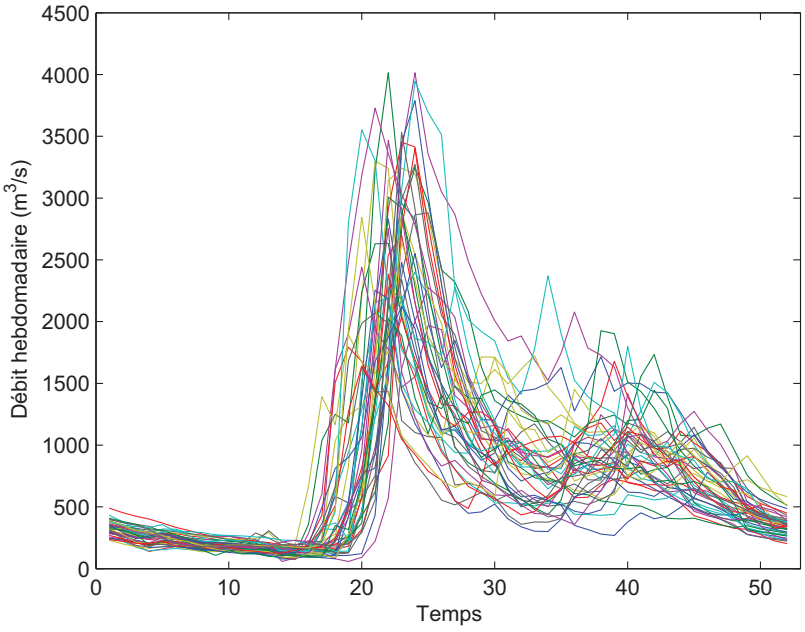


FIGURE 1.3. Hydrogrammes annuels observés au site Caniapiscau de 1961 à 2002.

La figure 1.3 présente les hydrogrammes annuels observés au site Caniapiscau de 1961 à 2002. Cette figure illustre bien les différentes caractéristiques des

hydrogrammes annuels. Dans un premier temps, nous voyons la présence de différents régimes au cours de l'année (voir aussi figure 1.1). La période hivernale affiche des faibles débits et la variabilité y est faible d'une année à l'autre. La période de la crue printanière est caractérisée par un accroissement important des débits et la variabilité entre les hydrogrammes annuels est grande durant cette période. Il doit aussi être noté que les crues printanières débutent et se terminent à différents moments selon les années. Après la crue printanière, il y a la période estivale-automnale qui présente des débits moins forts que ceux de la crue printanière mais plus élevés que ceux de la période hivernale. La variabilité pendant l'été et l'automne se situe aussi entre celles de l'hiver et de la crue printanière.

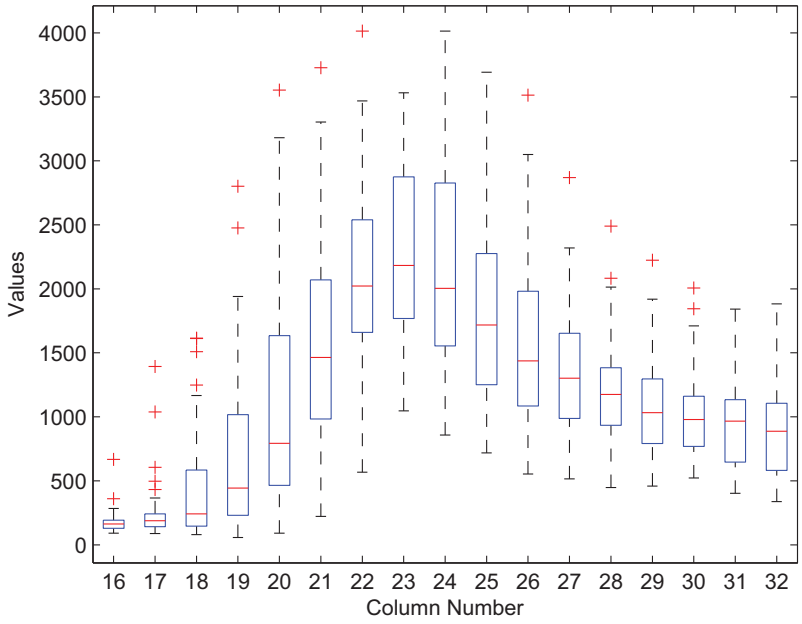


FIGURE 1.4. Graphiques en boîte des débits observés au site Caniapiscau de 1961 à 2002 pour les semaines 16 à 32.

La figure 1.4 confirme les changements de tendance centrale et de dispersion à travers l'année en présentant des graphiques en boîte, pour les semaines 16 à 32, des débits observés à Caniapiscau entre 1961 et 2002. Nous y voyons le changement de la tendance centrale, telle que mesurée par la médiane, à travers les semaines autour de la crue printanière, ainsi que le changement de la dispersion telle que mesurée par la distance interquartile.

Les caractéristiques illustrées aux figures 1.3 et 1.4 rendent les modèles ARIMA et SARIMA inapplicables aux séries chronologiques des débits. Comme il a été mentionné à la section 0.1.2.2, certains chercheurs utilisent les modèles PARMA afin de reproduire les séries de débits. Pour illustrer d'une façon sommaire ce type de modèles, nous considérons un modèle PAR(p). Ce modèle, bien que représentant un cas particulier, nous permet d'aborder différents aspects reliés aux modèles PARMA. Un modèle PAR(p) pour modéliser les débits est donné par

$$\left(\frac{z_{i,j}-\mu_j}{\sigma_j}\right)=\sum_{k=1}^{p}\phi_{k,j}\left(\frac{z_{i,j-k}-\mu_{j-k}}{\sigma_{j-k}}\right)+\varepsilon_{i,j} \tag{1.2.5}$$

où $z_{i,j}=z_j\left(y_{i,j}\right)$, une transformation bijective du débit $y_{i,j}$ qui peut dépendre de la semaine j, μ_j et σ_j représentent la moyenne et l'écart type du débit pour la semaine j, $\phi_{k,j}$ est le paramètre autorégressif entre les semaines j et $j-k$, et $\varepsilon_{i,j}$ est le terme d'erreur pour la semaine j de l'année i.

Plusieurs points doivent ici être notés. Dans un premier temps, la transformation bijective des débits des différentes années pour la semaine j peut dépendre de la semaine. Il est souvent nécessaire en pratique d'appliquer des transformations différentes selon les semaines puisqu'une transformation, qui rend les débits d'une semaine plus près de la normalité, ne le fera pas nécessairement pour une autre semaine (voir figure 1.4). Il est donc possible que chaque semaine possède sa propre transformation et dans le cas de débits hebdomadaires, il y aura alors 52 transformations. Toujours dans le cas de données hebdomadaires, il y aura 104 paramètres à évaluer afin de déterminer les moyennes et les écarts types. Finalement, les paramètres autorégressifs dépendent aussi du pas de temps de la série chronologique et donc pour des données hebdomadaires, nous avons alors $52\times p$ paramètres. En considérant des transformations de Box-Cox pour chaque semaine, qui dépendent d'un seul paramètre, le nombre de paramètres à évaluer pour un modèle PAR(p) appliqué à des données hebdomadaires sera donc de $52\times(3+p)$. Les modèles PARMA, tels qu'ils sont utilisés en pratique, sont donc loin d'être parcimonieux, bien qu'ils soient parfois en mesure de reproduire la variabilité des débits observés.

L'approche que nous proposons, basée sur l'analyse de données fonctionnelles, tente d'apporter des solutions parcimonieuses au problème de modélisation des débits. Nous croyons que la synchronisation des hydrogrammes permet d'obtenir un échantillon de courbes qui est plus homogène et qu'ainsi, il est possible de ne pas utiliser des transformations différentes à chaque pas de temps. De plus, une fois synchronisés, les hydrogrammes peuvent être employés pour obtenir un profil représentatif de l'échantillon. Finalement, une modélisation non paramétrique de cet hydrogramme de référence nous donne une représentation parcimonieuse de celui-ci dans un espace fonctionnel. Le chapitre suivant met ces idées en pratique.

Chapitre 2

BAYESIAN MODELING OF HYDROGRAPHS

Ce chapitre présente le premier article rédigé dans le cadre de cette thèse. L'article fut publié en 2007 comme l'indique la référence suivante

> Merleau, J., Perreault, L., Angers, J.-F., et Favre, A.-C. (2007). Bayesian modeling of hydrographs, Water Resources Research 43, W10432, doi :10.1029/2006WR005376.

Abstract

This article presents a new approach to model yearly hydrographs with daily or weekly streamflow measurements. The method considers yearly hydrographs as a sample of functions to be modelled nonparametrically in a Bayesian setting. The functional data analysis framework provides great flexibility to reproduce the features of yearly hydrographs, while the Bayesian probabilistic model ensures statistical coherence between the flood variables and the shapes of flood events. The proposed methodology is applied to two samples of hydrographs from two watersheds in the province of Québec.

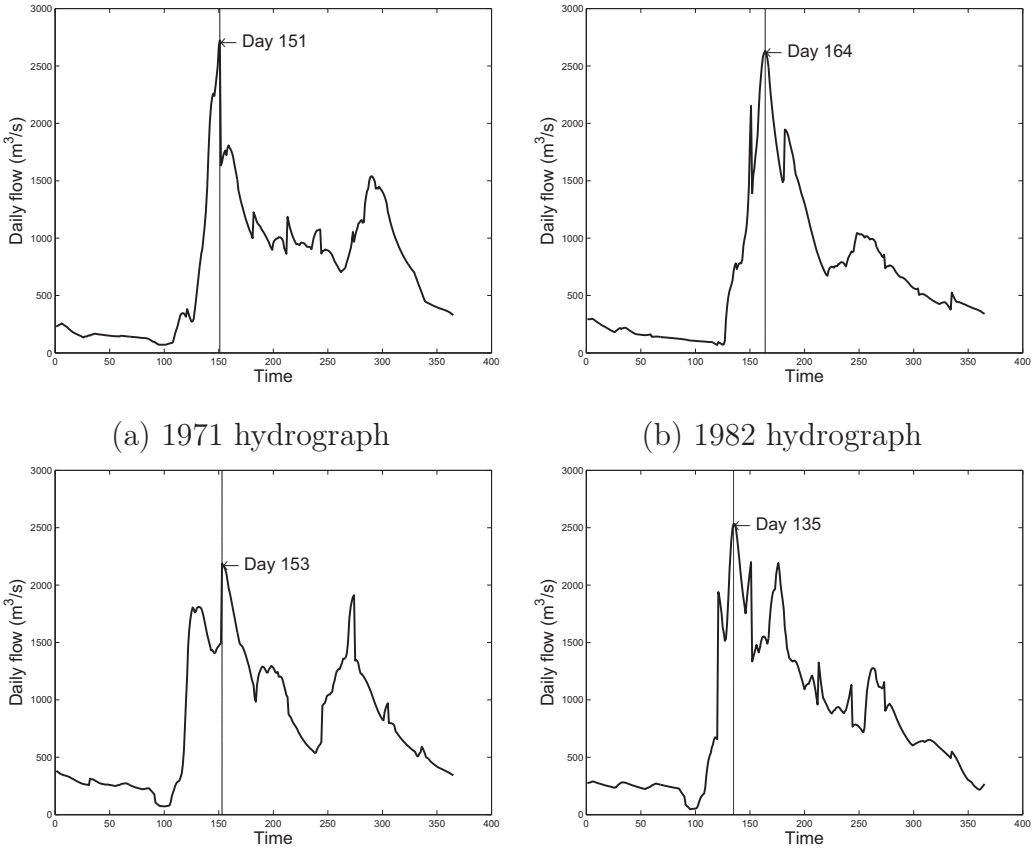
2.1. Introduction

Statistical modelling of hydrographs is important for many engineering purposes, in particular for energy planning and the design of power plants. Hydrographs are studied in these decision making contexts to ensure good water management and human population safety. For example, modelling of extreme hydrographs is necessary for the construction of dams which need to contain and

evacuate large quantities of water. In this context, synthetic hydrographs which preserve a realistic shape but simultaneously have extreme flood volumes and/or flood peaks are of interest for engineering planning. A good model to simulate extreme hydrographs thus needs to reproduce hydrographs with the aforementioned characteristics. In a water management context, hydrograph modelling has to be able to fulfill two main purposes. The first of these is to obtain a reference hydrograph for a given river, while the second consists in generating synthetic hydrographs that can occur with a given probability. It is difficult to construct a reference hydrograph since key features of different yearly hydrographs for a given river will happen at different times of the year and these features will often vary regarding their shapes (see Figures 2.1 and 2.2). For the purpose of generating hydrographs, a good model needs to be flexible enough to encompass a large variety of shapes which can be encountered in practice, since water management decisions depend heavily on these shapes. Several techniques have been set forth to model and simulate hydrographs. Some of these focus on flood events while others attempt to capture the stochastic process which governs water flow. The former methods usually model flood variables statistically, construct a design-flood hydrograph separately and combine the two levels of modelling to simulate hydrographs. The latter methods are based on time series analysis and are most often used to simulate a diversity of possible hydrographs for a given time horizon.

In the present paper, we propose a novel approach to model yearly hydrographs. Our method considers yearly hydrographs as a sample of functions to be modelled nonparametrically in a Bayesian setting. As will be shown, this functional data analysis framework offers the required flexibility to reproduce the characteristics of yearly hydrographs, but also provides a probabilistic model which ensures coherence between the flood variables and the shapes of flood events. Before exposing our new methodology, we will indicate the difficulties of conducting a statistical analysis of hydrographs and present the solutions that have been put forward in the literature.

Figure 2.1 illustrates 4 yearly hydrographs with daily measurements, while the same 4 yearly hydrographs with weekly measurements are shown in Figure 2.2.



(a) 1971 hydrograph

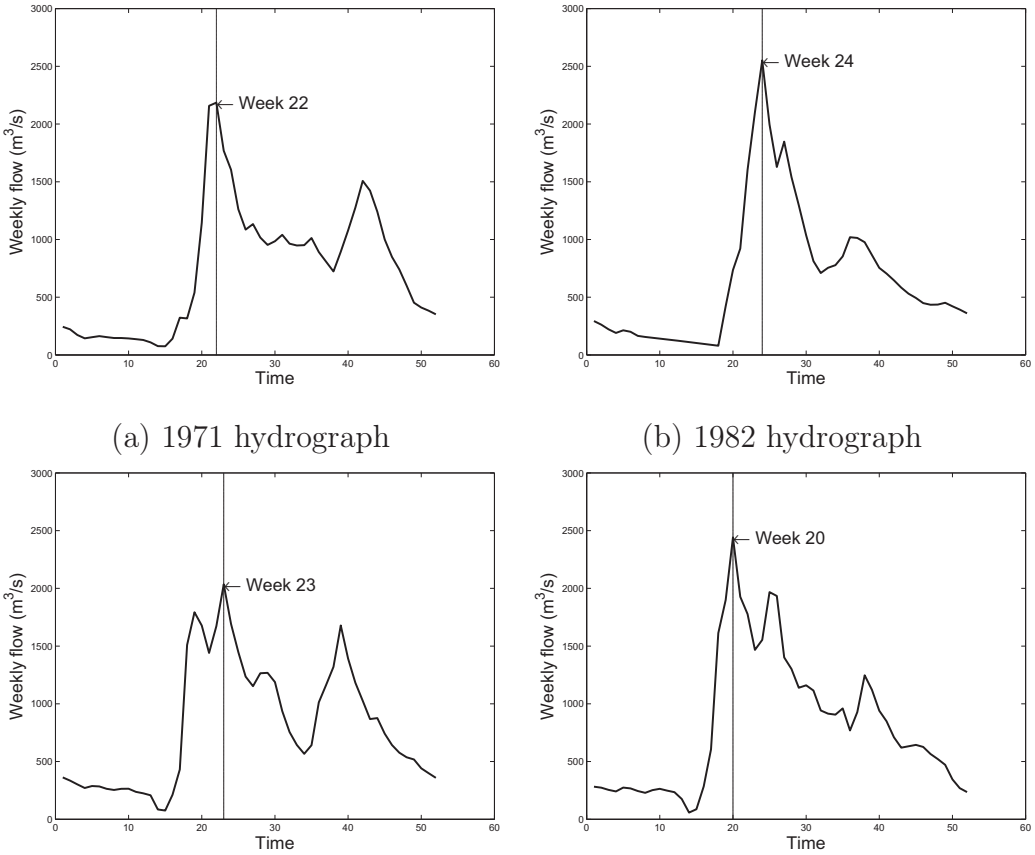
(b) 1982 hydrograph

(c) 1984 hydrograph

(d) 1986 hydrograph

FIGURE 2.1. Four yearly hydrographs with daily measurements. On each plot, the vertical line indicates the day at which the annual peak occured.

All these hydrographs come from the same basin in northern Québec. The first observation corresponds to the first measurement taken at the beginning of January, while the last observation corresponds to the last measurement at the end of December. The spring flood, mainly governed by snow melting, is present on each of the four hydrographs and starts roughly around the 100^{th} day of each year ; autumn floods, governed by heavy rainfall, are also present and occur roughly between days 250 and 325. The four spring floods show a wide variety of shapes, intensity and duration ; the time at which the flood peak happens, indicated by a vertical line, also varies between the different years. These differences are due to the climatic conditions and the amount of accumulated snow which vary from one year to the next ; the presence of late spring liquid precipitations also affects the



(a) 1971 hydrograph (b) 1982 hydrograph

(c) 1984 hydrograph (d) 1986 hydrograph

FIGURE 2.2. Four yearly hydrographs with weekly measurements. On each plot, the vertical line indicates the week at which the annual peak occured.

spring flood events and might cause secondary peaks. It is interesting to contrast the hydrographs of Figures 2.1 and 2.2 regarding some of their main characteristics. The hydrographs of Figure 2.2, with weekly time increments, are obviously smoother than the ones represented in Figure 2.1 (daily time increments), which causes the flood peaks to be flatter in Figure 2.2, especially for the hydrograph illustrated in (a). It is important to note that the main flood structures in Figure 2.1 can also be seen in Figure 2.2, although attenuated in certain cases. We thus see that complex structures are present for hydrographs with both daily and weekly time increments. An adequate hydrograph model therefore needs to capture these structures.

Flood peak, flood volume and flood duration are considered to be the main variables that summarize flood events and they are usually studied statistically through univariate or bivariate flood frequency analysis (Yue *et al.*, 1999 ; Javelle *et al.*, 2002). It is clear that studying the statistical distributions of these three variables is of major importance, but as pointed out by Yue *et al.* (2002), it is not enough to fully describe flood events because of the impact of their shapes in a water management situation. The approach often adopted in practice is to do a flood frequency analysis and proceed to calculate return periods for the different flood variables. Separately, one of the following construction methods is used to create a reference flood hydrograph and it is then adjusted to have the properties with the desired return periods.

A comprehensive review of the different methods to construct a design-flood hydrograph is given in Yue *et al.* (2002) and the interested reader is referred to the article for further details. Adopting the four categories listed in Yue *et al.* (2002), the construction methods are the traditional unit hydrograph (TUH) methods, the synthetic unit hydrograph (SUH) methods, the typical hydrograph (TH) methods and the statistical (S) methods. The TUH and SUH methods are based on hydrological principles. The TUH methods assume that the runoff response to rainfall is time invariant and that this response is linear as a function of rainfall. The SUH methods are based on empirical relationships that appear to exist between the parameters of a unit hydrograph and the physical characteristics of a drainage basin. A substantial number of articles have been devoted to the TUH methods (Sherman, 1932 ; Doodge, 1959 ; Chow 1964 ; Chow *et al.*, 1988 ; Pilgrim and Cordery, 1993 ; Yue and Hashino, 2000) and will not be discussed further here ; the same applies to the SUH methods (Snyder, 1938 ; U.S. Soil Conservation Service US-SCS, 1985). In fact, these approaches are not designed to produce realistic synthetic hydrographs for basins in northern regions like Québec where major floods are not the result of rainfall but mainly come from snow melting at the onset of spring. It is precisely for this reason that engineers in northern countries have relied on the TH methods.

The TH methods (Nezhikhovsky, 1971 ; Sokolov *et al.*, 1976) are widely used by practitioners. In this approach, a typical flood hydrograph, usually the one with the highest peak or the largest volume, is chosen from a river's sample of flood hydrographs. Each water flow value of the chosen flood hydrograph is then multiplied by a constant in order to get a flood peak and/or a flood volume corresponding to a given return period. This method considers the flood of the hydrograph as a function but it relies on a single historical realization. Therefore, it does not use all the information available in the sample of historical flood hydrographs.

The S methods, which include the approach put forward by Yue *et al.* (2002), consist of modelling the shape of each flood event by a probability density function, usually a gamma or a beta distribution. Yue *et al.* (2002) pursue this methodology further by studying shape variables of the adjusted beta distributions to the flood hydrographs. The shape variables, namely the shape mean and the shape standard deviation, are then considered as independent random variables and are each statistically modelled by a lognormal distribution. This enables the authors to consider return periods for the two shape variables. While incorporating a better probabilistic component to the problem by modelling the shape of flood events by two variables which are analyzed in a probabilistic framework, it seems to us that it is necessary to go further by considering hydrographs and their flood events as complex functions, and not restrict the shape of a flood to a model containing only two parameters.

Finally, modelling techniques based on time series are mostly used to generate a wide range of hydrographs which are considered to be statistically probable scenarios. Because of the complexity of the underlying processes, the time series models often need to include numerous parameters to capture the observed statistical properties of hydrographs. Periodic autoregressive moving average (PARMA) models (Salas *et al.*, 1980 ; Salas *et al.*, 1982 ; Vecchia *et al.*, 1983 ; Rasmussen *et al.*, 1996) or PARMAX models (Perreault and Latraverse, 2001 ; Ouhib, 2005), which include explanatory variables, seem to be able to reproduce observed properties of hydrographs. However, the period of these models is usually taken to be

the time increment of the series, therefore leading to an excessively large number of parameters for daily or weekly data. Furthermore, these methods cannot simulate hydrographs with fixed flood volumes and/or flood peaks because of their stochastic nature.

Statistical modelling of hydrographs is a complex multivariate problem since the objective is to reproduce the characteristics of a sample of functions. Yearly hydrographs, and their flood events, constitute complex functional data and should therefore be analyzed statistically in a functional data analysis framework (Ramsay and Silverman, 2005). For instance, it should be clear that the flood events illustrated in Figure 2.1 could not be reproduced by only one beta or gamma distribution since these distributions are unimodal functions. One could complexify the S methods by using a mixture of probability distribution functions (Titterington *et al.*, 1985) but even this approach seems unsatisfactory for the task at hand. Moreover, the S approach lacks cohesion since the flood characteristics such as the peak and volume are studied through flood frequency analysis, while the flood event shapes are modelled separately using a probability distribution function. The new method proposed in this paper brings forward an integrated approach in which hydrographs are modelled as functions in a probabilistic framework. This ensures statistical coherence between important characteristics of hydrographs, like flood peaks and flood volumes, and the shapes of the hydrographs.

In the next section, the tools of functional data analysis which we use in this study are put forward. We first describe an approach, based on landmark registration, to make the individual hydrographs of a given river similar on the time domain. We then set up a general nonparametric regression framework based on regression spline functions; this framework offers the modelling power and flexibility which are necessary to capture the different shapes of hydrographs. The Bayesian probabilistic model is exposed in section 2.3 and the methodology is applied to data in section 2.4.

2.2. Functional Data Analysis Context

Functional data analysis is often concerned with modelling longitudinal data, that is data formed by a collection of repeated measurements of a response variable on a certain experimental unit or individual. Longitudinal data are frequently encountered in the life sciences where it is often the case that a response variable is studied on several individuals through time. Some examples are growth curves, the effect of a treatment as a function of time on patients, etc. In analogy with longitudinal data, we consider each year as an experimental unit for which we have repeated water flow measurements.

We have, for each experimental unit i, the following observations :

$$(x_{i,1}, y_{i,1}), \ldots, (x_{i,j}, y_{i,j}), \ldots, (x_{i,n_i}, y_{i,n_i}),$$

where $x_{i,j}$ can be an explanatory variable or the time at which the response variable $y_{i,j}$ has been measured. We assume that $x_{i,j}$ is a deterministic variable, while $y_{i,j}$ is the random variable to be modelled. In our modelling context, $x_{i,j}$ is the time at which the water flow $y_{i,j}$ is measured for the year i ; furthermore, we have $x_{i,j} = x_j$ and $n_i = n$ for every i since the measurements in our case are always taken at the same time increments, either every day ($n = 365$) or every week ($n = 52$). Our data for year i is therefore of the following form :

$$(x_1, y_{i,1}), \ldots, (x_j, y_{i,j}), \ldots, (x_n, y_{i,n}),$$

where $i = 1, \ldots, N$, and N represents the number of yearly hydrographs in our sample.

As will be seen in section 2.2.2, each observed yearly hydrograph can be modelled with a nonparametric model. This is not the course we pursue in the present paper because we want to tackle another important issue, namely to obtain a hydrograph which is representative of a sample of hydrographs originating from a given river ; in other words, we seek to model the underlying average process of a sample of hydrographs, which we refer to as a representative or reference hydrograph.

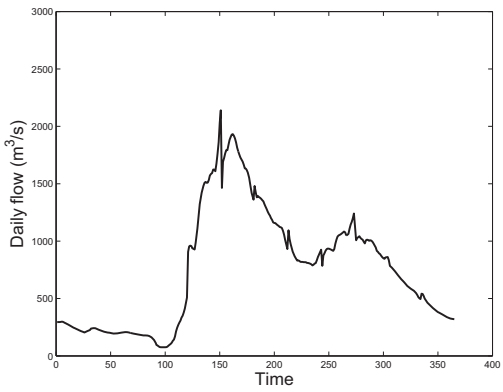
2.2.1. Landmark registration

The average of the four yearly hydrographs shown in Figure 2.1 (daily flow) is given in panel (a) of Figure 2.3, while panel (c) of the same figure gives the average of the yearly hydrographs of Figure 2.2 (weekly flow). It is clear that the mean hydrographs do not have flood events representative of those illustrated in Figures 2.1 and 2.2. For most rivers in northern Québec, the average of observed hydrographs, whether for daily or weekly measurements, cannot be used as a reference hydrograph. This average can be useful for volume analyses since it is indicative of the mean water flow during a certain period of the year, but it is not indicative of peak flows or of flood events shapes.

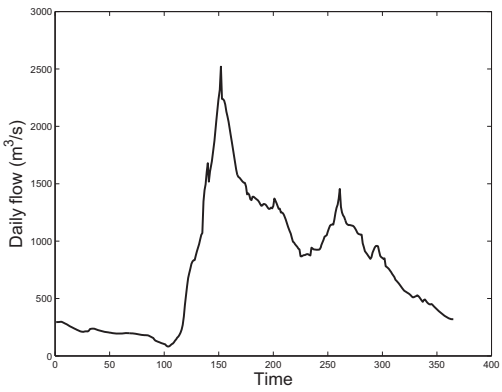
In order to model a reference hydrograph, we use landmark registration which has been studied by Kneip and Gasser (1988, 1992) in a statistical context. The key idea behind registration is to transform the independent variable, x in the present context, in order to make the yearly hydrographs similar on the domain of the transformed variable. For our purposes, this comes down to performing a time transformation such that the yearly hydrographs have important features occuring at simultaneous times ; for example, it is possible to perform time registration which makes all the flood peaks of the yearly hydrographs happen at the same time of the year. Specifically, landmark registration consists in identifying salient features of a sample of functions and using these landmarks to execute the registration. We want to go from the original time x to a registered time t, and therefore from the observations $(x_j, y_{i,j})$ to the registered observations $(t_{i,j}, y_{i,j})$, where $t_{i,j} = g_i(x_j)$ and $g_i(\cdot)$ is the registration function for year i. We note that the registration function should, at least intuitively, contain information on the climatic conditions of a given year i, a possibility which we are currently studying.

For the transformations to be one-to-one, the registration functions need to be monotonically increasing. Furthermore, we constrain the functions to transform the times at which important features happen to specified times. We thus have a sequence of constraints of the following form :

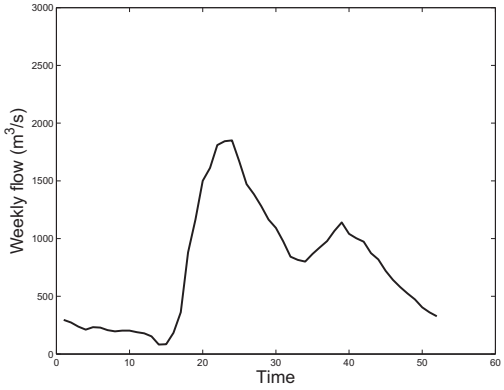
$$t_{i,\nu} = g_i(x_\nu) = \tau_\nu, \tag{2.2.1}$$



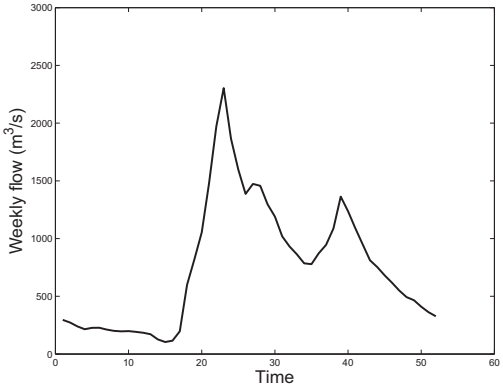
(a) Average of 4 observed hydrographs



(b) Average of 4 registered hydrographs



(c) Average of 4 observed hydrographs



(d) Average of 4 registered hydrographs

FIGURE 2.3. Daily measurements : (a) Average of 4 observed hydrographs, (b) average of the same 4 hydrographs after registration. Weekly measurements : (c) Average of 4 observed hydrographs, (d) average of the same 4 hydrographs after registration.

where x_ν represents the time at which the landmark ν occurs for year i and τ_ν is the specified time at which the landmark ν happens, for all years, in the transformed time domain.

The registration functions can be modelled by several methods : a Taylor expansion approximation (Angers *et al.*, 2005), interpolating splines (Kneip and Gasser, 1992), or an approach such as the one suggested by Ramsay and Li (1998). Here, we consider each function to be made up of linear parts $L_p(x)$ and we then have :

$$g_i(x) = \sum_{p=1}^{P} L_p(x) I_{D_p}(x) = \sum_{p=1}^{P} \left(c_{p,0} + c_{p,1} x\right) I_{D_p}(x), \tag{2.2.2}$$

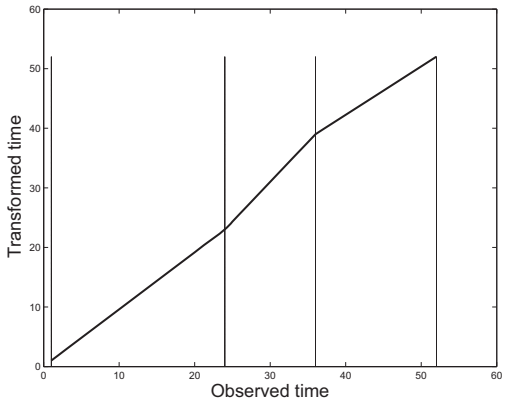
where D_p is the domain for which the linear function $L_p(x)$ is non zero, $I_{D_p}(x) = 1$ for $x \in D_p$ and 0 otherwise, and P represents the number of parts of the registration function. This simple model possesses an exact solution when the continuity of the registration function is imposed ; it also satisfies the monotonicity criterion as long as the landmarks are events which occur in the same sequence every year. Furthermore, this type of registration function generally performs well for preserving flood event volumes (Merleau *et al.*, 2005).

We will now illustrate the use of landmark registration to obtain a reference hydrograph for the hydrographs shown in Figures 2.1 and 2.2. We choose the four following events as landmarks : the first measurement of the year, the peak of the spring flood, the peak of the fall flood and the last measurement of the year. We then have the following constraints for the registration function of year i :

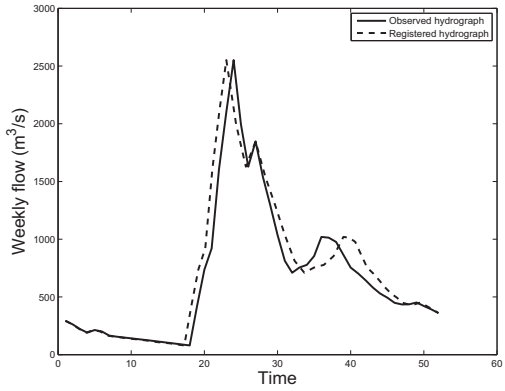
$$g_i(L_x) = L_x, g_i(x_s) = \tau_s, g_i(x_f) = \tau_f, g_i(U_x) = U_x, \tag{2.2.3}$$

where L_x and U_x are respectively the lower and upper bounds of the domain of x ; x_s and x_f are the times, for year i, at which the peak of the spring flood and the peak of the fall flood respectively happened ; τ_s and τ_f are the specified times at which the spring flood peak and the fall flood peak are fixed to occur in the domain of the synchronous time τ. We fix τ_s and τ_f to be the median values of the observed x_s and x_f. Panel (a) of Figure 2.4 shows the registration function for the yearly hydrograph given in Figure 2.2(b) (weekly flow). Panel (b) illustrates the effect of the registration function on the observed hydrograph. From the constraints given in equation (2.2.3), the registration function given in equation (2.2.2) is made up of 3 linear parts. The slope of a given part, $c_{p,1}$, determines if the corresponding section of the hydrograph is stretched ($c_{p,1} > 1$) or contracted ($c_{p,1} < 1$). In Figure 2.4, the middle section is stretched, while the first and last sections are contracted.

Panel (b) of Figure 2.3 shows the average obtained after registration for the hydrographs of Figure 2.1 (daily flow), and panel (d), the average of the registered hydrographs of Figure 2.2 (weekly flow). If we compare the average registered hydrographs with their observed counterparts, it is clear that registration makes the average hydrograph more representative of a sample of hydrographs. The



(a) Registration function



(b) Effect of registration on the observed hydrograph

FIGURE 2.4. (a) registration function for 1982 hydrograph; (b) effect of the registration function on the observed 1982 hydrograph. The vertical lines in (a) represent L_x, x_s, x_f and U_x (see equation (2.2.3)).

spring floods in panels (b) and (d) are much better defined and closer to the observed ones than those illustrated in panels (a) and (c). Furthermore, the peak value of the average spring floods, after registration, is the real average of the four observed hydrographs because of the way the registration is performed. We also notice the presence of secondary spring flood peaks in panels (b) and (d), which can also be seen in Figures 2.1 and 2.2.

2.2.2. Nonparametric regression with spline functions

In our functional data analysis context, we assume that

$$y_{i,j} = h_i(t_{i,j}) + \varepsilon_{i,j}, \tag{2.2.4}$$

where $h_i(t_{i,j})$ is a continuous function evaluated at $t_{i,j}$ and $\varepsilon_{i,j}$ is an error term. We therefore go from the data points $(t_{i,j}, y_{i,j})$ $(j = 1, \ldots, n)$ to a functional representation : $(t, h_i(t))$, for $t \in D_t = [L_t, U_t]$ where L_t and U_t represent respectively the lower and upper bounds of the t domain. In the present paper, we seek to model the average process which underlies yearly hydrographs and we therefore assume that $h_i(\cdot) = h(\cdot)$, for all i.

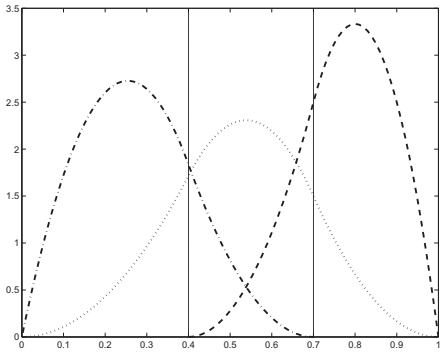
Several methods exist to estimate the function $h(\cdot)$: kernel methods (Hastie and Tibshirani, 1990), Fourier series, spline based methods (Ramsay and Silverman, 2005), wavelet methods (Ogden, 1997), etc. We choose to work with regression polynomial spline functions as a basis to evaluate the functions of interest because this type of basis possesses good mathematical properties such as differentiability and integrability, the latter property being useful in the present context as will be seen shortly. It also offers good flexibility and leads to parsimonious models when free-knots are used.

The function $h(t)$ is estimated by

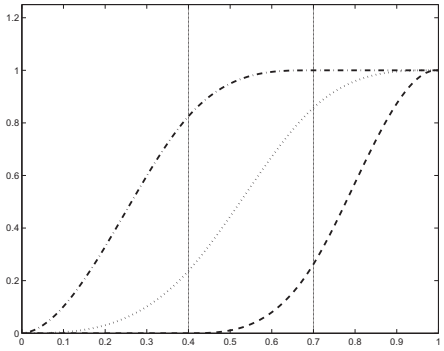
$$h_{\boldsymbol{\omega}}(t) = \sum_{k=1}^{K_{\boldsymbol{\omega}}} \beta_{k,\boldsymbol{\omega}} b_{k,\boldsymbol{\omega}}(t), \tag{2.2.5}$$

where $\beta_{k,\boldsymbol{\omega}}$ represents the coefficient of the basis element $b_{k,\boldsymbol{\omega}}(\cdot)$. Once the order l of the spline polynomial functions is fixed, the basis elements $b_{k,\boldsymbol{\omega}}(\cdot)$ are determined by the number of interior knots, m, and the ordered location of these knots, $\boldsymbol{r}^{(m)} = (r_1, ..., r_m)$; the number of elements in the basis is given by $K_{\boldsymbol{\omega}} = l + m$. We summarize this information in the model parameter $\boldsymbol{\omega} = (m, \boldsymbol{r}^{(m)})$. The determination of $\boldsymbol{\omega}$ is a model selection problem which is discussed in section 2.3.2. We note that this model can be understood as a linear model such as those encountered in linear regression. The model given in equation (2.2.4) can now be written as

$$\boldsymbol{y}_i = \boldsymbol{B}_{\boldsymbol{\omega}} \boldsymbol{\beta}_{\boldsymbol{\omega}} + \boldsymbol{\varepsilon}_i, \tag{2.2.6}$$



(a) M-spline functions



(b) I-spline functions

FIGURE 2.5. 3 M-spline and 3 I-spline functions for $\boldsymbol{\omega} = (2, (0.4, 0.7))$. The vertical lines indicate the positions of the interior knots.

where $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,n})'$ is a $n \times 1$ vector; $\boldsymbol{B}_{\boldsymbol{\omega}} = (\boldsymbol{b}_{\boldsymbol{\omega}}(t_{i,1}), \ldots, \boldsymbol{b}_{\boldsymbol{\omega}}(t_{i,n}))'$, a $n \times K_{\boldsymbol{\omega}}$ matrix, with $\boldsymbol{b}_{\boldsymbol{\omega}}(t_{i,j}) = (b_{1,\boldsymbol{\omega}}(t_{i,j}), \ldots, b_{K_{\boldsymbol{\omega}},\boldsymbol{\omega}}(t_{i,j}))'$, a $K_{\boldsymbol{\omega}} \times 1$ vector of the basis elements evaluated at $t_{i,j}$; $\boldsymbol{\beta}_{\boldsymbol{\omega}} = (\beta_{1,\boldsymbol{\omega}}, \ldots, \beta_{K_{\boldsymbol{\omega}},\boldsymbol{\omega}})'$, a $K_{\boldsymbol{\omega}} \times 1$ vector of parameters; $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n})$, a $n \times 1$ vector of error terms.

Figure 2.5 shows M-spline and I-spline functions for fixed order ($l = 3$) and fixed model parameter : $\boldsymbol{\omega} = (2, (0.4, 0.7))$, which corresponds to 2 interior knots positioned at 0.4 and 0.7. In this case there are 5 basis elements, *i.e.* $K_{\boldsymbol{\omega}} = 5$; for clarity, only 3 of these elements are illustrated in Figure 2.5. For the current knot configuration, panel (a) shows 3 M-spline functions, $b^M_{k,\boldsymbol{\omega}}(t)$, and the corresponding I-spline functions, $b^I_{k,\boldsymbol{\omega}}(t)$, are given in panel (b). Each M-spline function is made up of polynomial parts of order l, or degree $(l-1)$; in the current example, the polynomial parts are quadratic. Each I-spline function is an integrated M-spline function and constitutes a monotonically increasing function; therefore, I-spline

functions are well suited to model monotone functions, as indicated by Ramsay (1988). M-spline functions are closely related to B-spline functions which are widely used in statistics (He and Shi, 1998). We note that the M-spline functions integrate to 1 and are, in that respect, equivalent to probability distribution functions used in the S methods (see introduction). To pursue the parallel further, the S methods use only one basis element to model a flood event, while our approach uses several basis elements to give a representation of a yearly hydrograph.

As mentioned previously, I-spline functions form a good basis to model monotone functions. Since a yearly hydrograph is a positive function, the function

$$H(\tau) = \int_{L_t}^{\tau} h(t)\mathrm{d}t \tag{2.2.7}$$

is a monotone increasing function and it represents the integrated water flow from the beginning of the year to a certain time τ. This cumulative hydrograph is of particular interest for conducting volume analyses. If a hydrograph $h(t)$ is modelled with M-spline functions, $b^M_{k,\boldsymbol{\omega}}(t)$, in equation (2.2.5), then a model for $H(t)$ is obtained simultaneously and it is given by

$$H_{\boldsymbol{\omega}}(t) = \sum_{k=1}^{K_{\boldsymbol{\omega}}} \beta_{k,\boldsymbol{\omega}} b^I_{k,\boldsymbol{\omega}}(t), \tag{2.2.8}$$

where each coefficient $\beta_{k,\boldsymbol{\omega}}$ is the same for the two functional representations $h_{\boldsymbol{\omega}}(t)$ and $H_{\boldsymbol{\omega}}(t)$. This result follows from the fact that each coefficient $\beta_{k,\boldsymbol{\omega}}$ is independent of t and each I-spline function is an integrated M-spline.

2.3. Bayesian statistical model

We adopt the Bayesian paradigm instead of the frequentist approach for several reasons, some of which we now put forward. It enables the statistician to take into account, in a coherent probabilistic framework, the uncertainty related to all the parameters of the model and thus gives a more adequate representation of the uncertainty concerning a model ; frequentist approaches can often underestimate this uncertainty. As will be seen in section 2.3.2, it makes the model selection procedure formal in the sense that the selection follows directly from the initial assumptions about the probabilistic model ; therefore, it does not

rely on some *ad hoc* procedure. If we would choose to do so, it is easy in the Bayesian framework to include constraints on the parameter space through the *a priori* statistical distributions, thus making the implementation of constraints straightforward. Furthermore, the inclusion of expert opinion can also be included through the prior statistical distribution ; for example, it would be possible to consult hydrologists to obtain a given shape of a hydrograph and to transform this shape into the coefficient space of the spline functions. Although we have not done this in the paper, we are currently thinking of incorporating this aspect in our model.

In order to treat a certain function $h(t)$ as a random functional event, we can regard the parameters ($\beta_{k,\boldsymbol{\omega}}$) of equation (2.2.5) as random variables. In a Bayesian framework, the parameters, $\boldsymbol{\theta}$, of a given model are considered to be random variables which are drawn from a certain probability distribution. A Bayesian statistical model is made up of the following two elements (see, e.g., Lee, 1989 ; Bernardo et Smith, 1994) : a prior probability distribution for the model parameters, $\pi(\boldsymbol{\theta})$, and a probability distribution function, $f(\boldsymbol{y}|\boldsymbol{\theta})$, from which the observations arise. A prior distribution is a probabilistic formulation of the information available before observations are collected. From these two probability distributions, the posterior distribution associated with the model parameters can be obtained by applying Bayes' theorem :

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\mathrm{m}(\boldsymbol{y})}, \tag{2.3.1}$$

where $\mathrm{m}(\boldsymbol{y})$ is the marginal distribution of $\boldsymbol{y}$, *i.e.* the statistical distribution of $\boldsymbol{y}$ after the model parameters have been integrated and $\boldsymbol{\Theta}$ is the parameter space. All statistical inference concerning the parameters are made from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, which represents a probabilistic model for the parameters that has been updated by the empirical information.

2.3.1. Distributional hypotheses

We assume that the distribution of each vector $\boldsymbol{\varepsilon}_i$ is a multivariate normal distribution (see Appendix A for the definitions of the probability distributions

used in this section) :

$$\boldsymbol{\varepsilon}_i \sim \mathrm{N}_n\left(0, \sigma^2 \boldsymbol{\Sigma}_\varepsilon\right), \tag{2.3.2}$$

where σ^2 is a variance proportionality constant and $\boldsymbol{\Sigma}_\varepsilon$ is the covariance matrix which captures the covariance, or correlation, structure between the elements of $\boldsymbol{\varepsilon}_i$. We note that σ^2 and $\boldsymbol{\Sigma}_\varepsilon$ are taken to be the same for every year i. In the application section, we consider the elements of each $\boldsymbol{\varepsilon}_i$ to be independent and identically distributed; more formally, we have $\mathrm{Cov}\left(\varepsilon_{i,j}, \varepsilon_{i,j'}\right) = 0$ for $j \neq j'$ and $\varepsilon_{i,j} \sim \mathrm{N}_1\left(0, \sigma^2\right)$ for all j, where Cov represents the covariance operator and N_1 indicates a one-dimensional normal distribution. This leads to the following prescription (A1) : $\boldsymbol{\Sigma}_\varepsilon = \mathbb{I}_n$, where $\mathbb{I}_n$ is an n-dimensional unit diagonal matrix and (A1) indicates the first application assumption. It should be noted that under (A1), the variance is taken to be the same throughout the domain of the hydrograph, an assumption which is discussed further in the conclusion.

The probability distribution function of each vector of observations, $\boldsymbol{y}_i$, is then a multivariate normal distribution which we can write as :

$$\boldsymbol{y}_i \mid \boldsymbol{\beta}_{\boldsymbol{\omega}}, \sigma^2 \sim \mathrm{N}_n\left(\boldsymbol{B}_{\boldsymbol{\omega}} \boldsymbol{\beta}_{\boldsymbol{\omega}}, \sigma^2 \boldsymbol{\Sigma}_\varepsilon\right). \tag{2.3.3}$$

In the notation given above, we have the vector of observations $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$, and the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\boldsymbol{\omega}}', \sigma^2)'$ since we make the hypothesis that the design matrix $\boldsymbol{B}_{\boldsymbol{\omega}}$ is fixed and that the covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ is known. Since the yearly hydrographs are considered to be independent, the joint probability distribution of the observations is then given by : $f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(\boldsymbol{y}_i|\boldsymbol{\theta})$. Our interest lies in the vector of regression parameters $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ which represents the reference hydrograph in functional space.

We assume that the prior distribution can be written as

$$\pi\left(\boldsymbol{\beta}_{\boldsymbol{\omega}}, \sigma^2\right) = \pi\left(\boldsymbol{\beta}_{\boldsymbol{\omega}} | \sigma^2\right) \pi\left(\sigma^2\right) \tag{2.3.4}$$

and choose a conjugate model for the parameters, *i.e.* a probabilistic model with prior and posterior distributions from the same family of distributions. For normally distributed observations, we have the following conjugate prior distributions :

$$\boldsymbol{\beta}_{\boldsymbol{\omega}} \mid \sigma^2 \sim \mathrm{N}_{K_{\omega}} \left(\boldsymbol{\beta}^0_{\boldsymbol{\omega}}, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\omega}}\right), \tag{2.3.5}$$

$$\sigma^2 \sim \mathrm{I\Gamma} \left(\frac{\alpha_{\boldsymbol{\omega}}}{2}, \frac{\gamma_{\boldsymbol{\omega}}}{2}\right), \tag{2.3.6}$$

where $K_{\boldsymbol{\omega}}$ denotes the dimension of the spline basis (see equation (2.2.5)). Our *a priori* knowledge should be used to determine the hyperparameters : $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$, the mean of the multivariate normal distribution, $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$, the covariance structure between the components of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$, and the shape parameters of the inverse gamma distribution, $\alpha_{\boldsymbol{\omega}}$ and $\gamma_{\boldsymbol{\omega}}$.

Because of our lack of prior knowledge concerning the hyperparameters and because the first N_0 yearly hydrographs are data of lesser quality, because they have been reconstructed, than the rest of the data set for the province of Québec, we use this data to determine the prior distributions ; the N remaining historical hydrographs are considered the effective sample. The hydrographs of the first N_0 years are of lesser quality but they nonetheless contain information about hydrographs for a given site. Although we don't want to treat these hydrographs as part of the effective sample, it is reasonable to use this information in our model but to weigh it properly. For the prior distribution of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$, the N_0 years are used to determine the location vector $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$ (see section 2.4.2 for details). This is the second application assumption (A2). Concerning the covariance matrix, we assume that for a certain model defined by $\boldsymbol{\omega}$, we have (A3) : $\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \frac{1}{n_0}(\boldsymbol{B}'_{\boldsymbol{\omega}} \boldsymbol{\Sigma}_{\varepsilon}^{-1} \boldsymbol{B}_{\boldsymbol{\omega}})^{-1}$. This type of covariance structure was suggested by Zellner (1986) and can be justified in several ways (see Robert, 1994). From a practical point of view, it makes the implementation of the model fairly straightforward since only one parameter, n_0, needs to be specified, instead of a covariance matrix of dimension $K_{\boldsymbol{\omega}} \times K_{\boldsymbol{\omega}}$; furthermore, this type of structure is well adapted to take into account multicollinearity, which can be the case when spline functions are used as a basis,

since it allows for a large prior variance on the components affected by multicollinearity. As with the location vector $\boldsymbol{\beta}_{\boldsymbol{\omega}}^{0}$, the determinination of $\alpha_{\boldsymbol{\omega}}$ and $\gamma_{\boldsymbol{\omega}}$ is done with the first N_0 hydrographs and also depends on n_0 (see section 2.4.2 for details). This is the fourth application assumption (A4). The factor n_0 can be viewed as an indicator of our confidence in the prior information, therefore we have $n_0 = zN_0$, where $0 \leq z \leq 1$. If one chooses $z = 0$, then it is assumed that the prior information contains no information and the prior distributions are improper ; while if $z = 1$ is chosen, each of the N_0 yearly hydrographs contributes as much to the posterior distributions as one of the hydrographs in the effective sample. Since we know that the N_0 hydrographs are of lesser quality, z should lie somewhere between these two extreme cases. We have conducted tests, on real data, by varying the value of z and these tests have shown that its value does not have a serious impact on results. In the application section, we use $z = 1/N_0$ which leads to $n_0 = 1$ (A5) ; the prior information then contributes the equivalent of one hydrograph from the effective sample.

Our choice in prior distributions leads to a posterior distribution which can be written as

$$\pi\left(\boldsymbol{\beta}_{\boldsymbol{\omega}}, \sigma^2 | \boldsymbol{y}\right) = \pi\left(\boldsymbol{\beta}_{\boldsymbol{\omega}} | \sigma^2, \boldsymbol{y}\right) \pi\left(\sigma^2 | \boldsymbol{y}\right), \tag{2.3.7}$$

and by using standard Bayesian calculations for linear models (Robert, 1994), we have

$$\boldsymbol{\beta}_{\boldsymbol{\omega}} \mid \sigma^2, \boldsymbol{y} \sim \mathrm{N}_{K_\omega}\left(\boldsymbol{\beta}_{\boldsymbol{\omega}}^{*}, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{*}\right), \tag{2.3.8}$$

$$\sigma^2 \mid \boldsymbol{y} \sim \mathrm{I\Gamma}\left(\frac{\alpha_{\boldsymbol{\omega}}^{*}}{2}, \frac{\gamma_{\boldsymbol{\omega}}^{*}}{2}\right). \tag{2.3.9}$$

The explicit expressions for $\boldsymbol{\beta}_{\boldsymbol{\omega}}^{*}$, $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{*}$, $\alpha_{\boldsymbol{\omega}}^{*}$ and $\gamma_{\boldsymbol{\omega}}^{*}$ are given in Appendix B (section B.1).

By integrating out σ^2 in equation (2.3.8), the posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$, independent of σ^2, is given by a multivariate Student's t distribution :

$$\boldsymbol{\beta}_{\boldsymbol{\omega}} \mid \boldsymbol{y} \sim \mathrm{T}_{K_\omega}\left(\alpha_{\boldsymbol{\omega}}^{*}, \boldsymbol{\beta}_{\boldsymbol{\omega}}^{*}, \frac{\gamma_{\boldsymbol{\omega}}^{*}}{\alpha_{\boldsymbol{\omega}}^{*}} \boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{*}\right). \tag{2.3.10}$$

Our main interest lies in the posterior probability distributions given in equations (2.3.9) and (2.3.10) since any statistical inference proceeds from these distributions. In the present hydrological context, in which we want to simulate hydrographs, we can generate vectors of parameters from the posterior distribution (2.3.10). A simulated hydrograph will then correspond to a simulated vector of parameters since a hydrograph is fully determined by a vector $\boldsymbol{\beta}_{\boldsymbol{\omega}}$. Another statistical distribution that will prove useful in the following section is the marginal distribution of the observations, $\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega})$, which has been defined in equation (2.3.1) and is given explicitly in Appendix B (section B.2).

2.3.2. Model selection : determining the best $\boldsymbol{\omega}$

In this section, a method is given to determine the spline basis model parameter $\boldsymbol{\omega} = (m, \boldsymbol{r}^{(m)})$, where m is the number of interior knots and the vector of ordered positions of these knots is $\boldsymbol{r}^{(m)}$. As pointed out in section 2.2.2, once the order of the spline polynomial functions is fixed, the parameter $\boldsymbol{\omega}$ fully determines the spline basis ; therefore, the determination of this parameter is crucial to obtain a good fit to the data. Many methods have been suggested in the literature to determine this parameter. Some simple methods position interior knots at given quantiles of the independent variable, while more sophisticated algorithms rely on forward, backward and stepwise methods to determine the best $\boldsymbol{\omega}$ (Friedman and Silverman, 1989 ; Stone *et al.*, 1997). The different models obtained from the various parameters $\boldsymbol{\omega}$ are usually compared through a model fitting criterion such as AIC (Akaike, 1973), the Schwarz criterion (Schwarz, 1978), cross-validation (Hastie and Tibshirani, 1990), etc. These criteria are all essentially structured in the same manner in that they "reward" goodness of fit and "penalize" model complexity in order to obtain a model that fits the data well but is still parsimonious.

The method which we adopt to explore knot configurations, *i.e.* the support of $\boldsymbol{\omega}$, is based on an insight of He and Shi (1998) in their paper on modelling monotone functions with B-spline functions. Instead of positioning the knots at the quantiles of the independent variable, they use the quantiles of the monotone

increasing function to perform a projection on the independent variable axis. The advantage of this simple method is that it positions more interior knots in the regions where the monotone function is rapidly changing; more basis elements need to be present in these regions to give more modelling flexibility where the function fluctuates the most. In the present context, the cumulative hydrograph given in equation (2.2.7), $H(\cdot)$, is used to perform this operation. An illustration of this procedure is shown in Figure 2.6. The water flow axis is subdivided into 8 regular sections by 7 markers; these are projected on the time axis using the monotone function. The interior knots are then taken to be the resulting time coordinates.

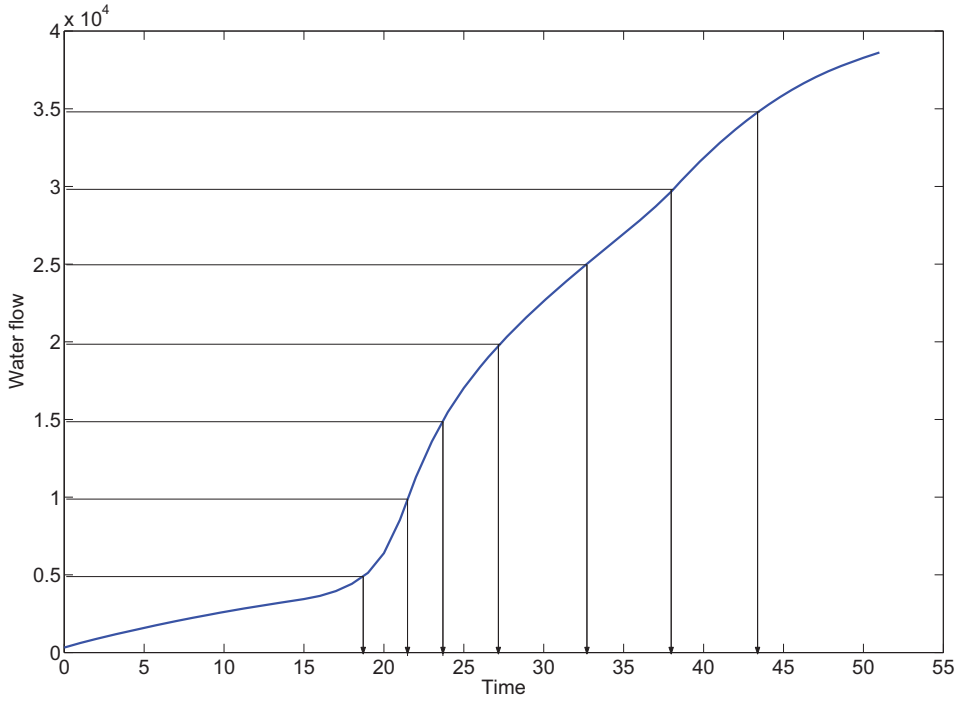


FIGURE 2.6. Illustration of He and Shi method for $m = 7$ interior knots.

A drawback with this technique is the fact that the vector $\boldsymbol{r}^{(m)}$ is solely determined by the number of interior knots m, which implies that a very limited number of knot configurations are explored. Furthermore, a fully Bayesian model would consider $\boldsymbol{\omega}$ to be a random quantity that follows a certain probability distribution. We are currently developing a fully Bayesian model similar to the

approach suggested by Denison *et al.* (1998). Nonetheless, the simple approach adopted here seems to work fairly well and it is also very effective regarding computational time.

Finally, a method needs to be adopted in order to discriminate between the different knot configurations. In a frequentist modelling context, one of the approaches mentioned previously would need to be chosen arbitrarily ; with the Bayesian approach, on the other hand, the model selection procedure follows directly from the hypotheses concerning the probabilistic model. The Bayesian model selection criterion, called the Bayes factor (Kass and Raftery, 1995), is given by the ratio of the marginal distributions of two competing models and indicates which of the models is more likely to be the best model. In the present context, let's say we are comparing model 1, $\boldsymbol{\omega}_1$, and model 2, $\boldsymbol{\omega}_2$, then the Bayes factor is given by

$$BF_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} = \frac{\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}_1)}{\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}_2)}. \tag{2.3.11}$$

Model 1 is more likely to be the best model when $BF_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} > 1$, while $BF_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} < 1$ indicates that model 2 has a higher probability to be the best model. For the current modelling context, an explicit expression for the Bayes factor is given in Appendix B (section B.3).

2.3.3. Bayesian estimator and confidence intervals

It is a well known result of Bayesian decision theory that the decision rule concerning a parameter under a squared error loss function is given by the expected value of this parameter (see Robert, 1994). Under squared error loss, the decision rule concerning $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ is to choose its expected value which is given by $\boldsymbol{\beta}^*_{\boldsymbol{\omega}}$. Once the best model $\boldsymbol{\omega}$ has been determined by the method discussed in the previous section, the Bayesian estimator of a representative hydrograph will therefore be :

$$\boldsymbol{h}^*_{\boldsymbol{\omega}} = \boldsymbol{B}_{\boldsymbol{\omega}}\boldsymbol{\beta}^*_{\boldsymbol{\omega}}. \tag{2.3.12}$$

Knowing that the posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ given in equation (2.3.10) is a multivariate Student's t distribution, it is possible to construct confidence intervals for linear combinations of the components of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ by using standard multivariate results (see Johnson and Wichern, 1992). The reader can refer to section B.4 (Appendix B) for the mathematical expression.

2.4. Application

2.4.1. Data

The streamflow data analyzed in this section are weekly net basin supplies for two basins situated in Québec and managed by Hydro-Québec, a public company that produces, transmits and distributes electricity throughout the province of Québec. Hydro-Québec currently operates 54 power plants supplied by 26 large reservoirs ; the major watersheds managed by Hydro-Québec are shown in Figure 2.7. In this paper, we focus on statistical modelling of yearly hydrographs from two different basins : Churchill Falls, which is located in northern Québec and has a basin area of 69 345 km^2, and Gouin, which is located in southern Québec and has a basin area of 9 376 km^2. These two watersheds serve as test-basins to explore the potential use of the approach proposed in the paper. For each watershed, a sample of weekly streamflows (in m^3/s) covering the period extending from 1950 to 2002 is considered (hydrologic data post 2002 being confidential).

Figure 2.8 shows five consecutive annual hydrographs with weekly streamflow for each watershed. Panel (a) shows the annual hydrographs observed at Churchill Falls during the 1989-1993 period. The sequence of hydrographs starts with the first week of January 1989 and ends with the last week of December 1993. Panel (b) illustrates 5 annual hydrographs at Gouin for the period extending from 1996 to 2000. In this case, the series starts with the first week of January 1996, and ends with the last week of December 2000. We notice that the two sequences possess some similarities and differences. They are similar in that the annual spring floods are more important than the autumn floods ; but they differ in their level of "smoothness". The yearly hydrographs of Churchill Falls, panel (a), possess very well defined spring and autumn floods which do not show strong variations ; the

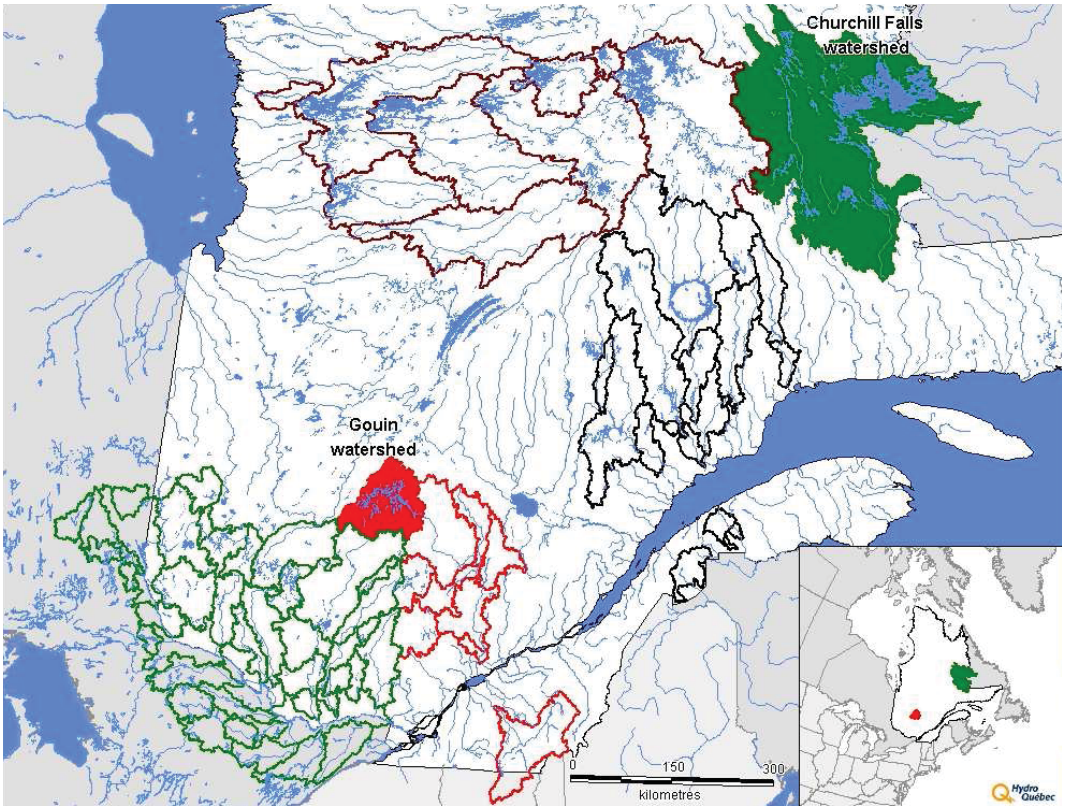
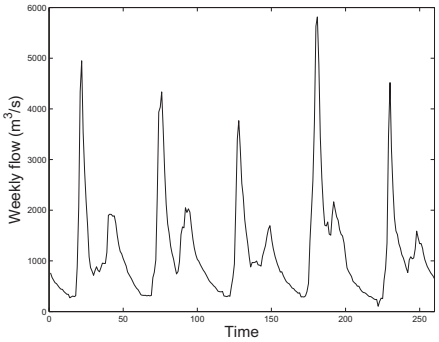


FIGURE 2.7. Location of major watersheds in the province of Québec.

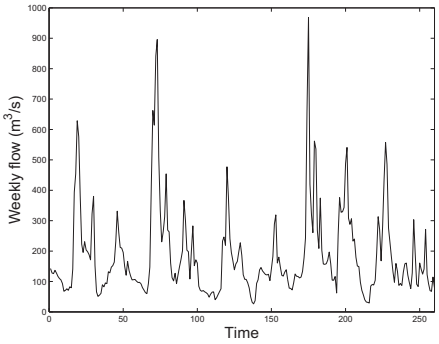
yearly hydrographs of Gouin, panel (b), exhibit more waterflow fluctuations. On average, the spring flood accounts for about 50% of the annual total volume and is composed of melted winter snowpack and spring precipitation. Thus, with respect to dam safety, hydropower generation, operation planning, and design of new power plants, a good model to simulate realistic hydrographs, particularly for this period, is certainly valuable to water resources planners and managers.

2.4.2. Model specifications

The order l of the spline functions which form the modelling basis needs to be specified in order to apply our method (see section 2.2.2). We could treat this quantity as a parameter to be estimated in the procedure, but we will consider it to be fixed as is usually done in practice. We choose to work with M-spline functions of order $l = 3$ which means that the basis elements are quadratic by parts.



(a) Churchill Falls



(b) Gouin

FIGURE 2.8. 5 consecutive yearly hydrographs for (a) Curchill Falls (1989-1993) and (b) Gouin (1996-2000).

The registration functions are modelled by linear parts as indicated in equation (2.2.2) and the constraints on these functions are taken to be the same as those given in equation (2.2.3). The landmarks which determine these constraints could be chosen differently, but we have found that this choice gives good results.

The hypotheses concerning the Bayesian probabilistic model are given in section 2.3.1. They are :

- (A1) $\boldsymbol{\Sigma}_\varepsilon = \mathbb{I}_n$ for each yearly hydrograph,
- (A2) $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$ determined by nonparametric least squares regression applied to the reference hydrograph for the $N_0 = 11$ historic yearly hydrographs covering the 1950 to 1960 period, leading to an effective sample formed by the $N = 42$ remaining hydrographs (1961-2002),
- (A3) $\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \frac{1}{n_0}(\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{B}_{\boldsymbol{\omega}})^{-1}$,

– (A4) $\alpha_{\boldsymbol{\omega}} = n_0 n$ and $\gamma_{\boldsymbol{\omega}} = n_0 S^0_{\boldsymbol{\omega}}$, where $S^0_{\boldsymbol{\omega}}$ is the average of the squared residuals of the regression in (A2),
– (A5) $n_0 = 1$.

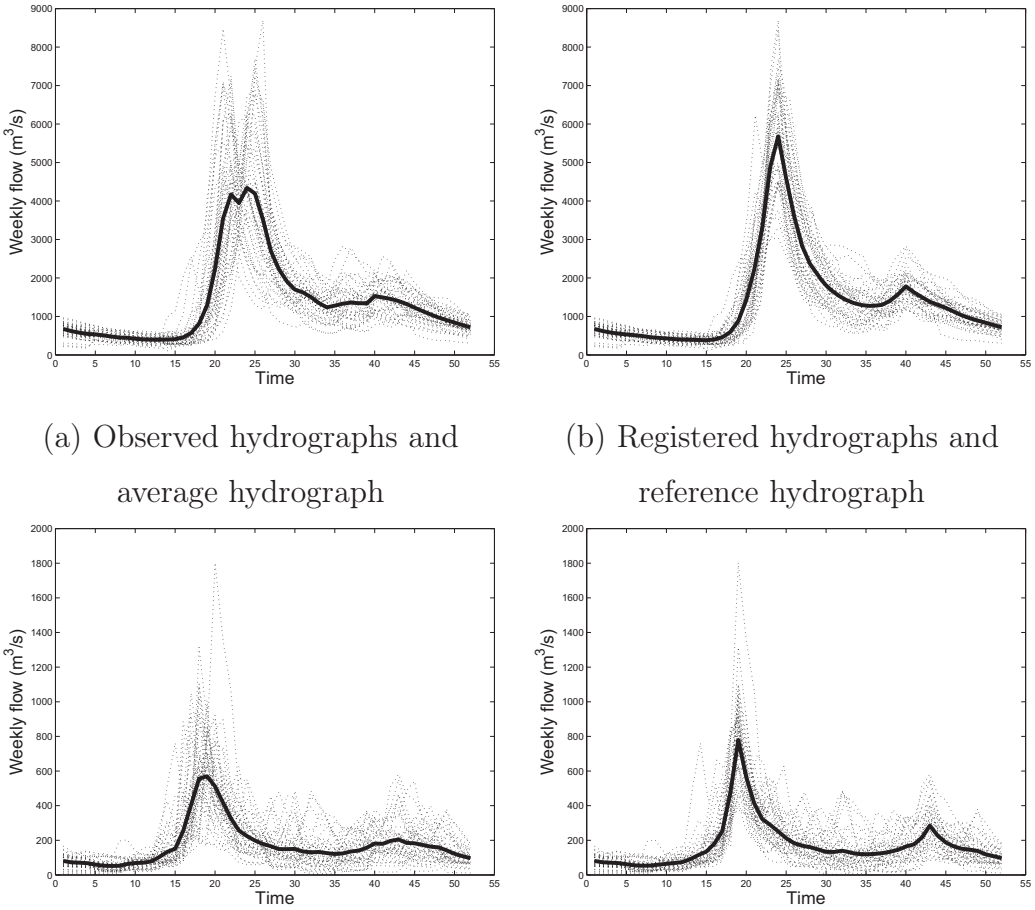
2.4.3. Results

2.4.3.1. *Registration and reference hydrographs*

We first apply landmark registration to each data set in order to be able to model the reference hydrographs of the two watersheds. Figure 2.9 illustrates the effect of registration on the two data sets. For Churchill Falls, panel (a) shows the observed hydrographs as dotted lines and the solid bold line represents their weekly average, *i.e.* the average hydrograph ; while panel (b) shows the registered hydrographs (dotted lines) and their average, *i.e.* the reference hydrograph, is illustrated as the solid bold line. In panels (c) and (d), the same exercise is performed for Gouin. For Churchill Falls, the average hydrograph has a flood event which is bimodal with the two modes being similar in magnitude (panel (a)). This is very rarely observed in practice and is a result of combining yearly hydrographs which are dissimilar regarding the moment at which the flood event occurs. For its part, the reference hydrograph for this site, shown in panel (b), possesses a well defined flood event. For Gouin, we notice that the reference hydrograph (panel (d)) has a higher peak flood than the average hydrograph (panel (c)) ; furthermore, the flood event is sharper and less spread out for the reference hydrograph than for the average hydrograph. The reference hydrographs therefore better capture the characteristics of the observed flood events since they possess a steep ascent, which is usually observed in practice, and a well defined peak. Furthermore, it needs to be noted that the peak flow of the reference hydrograph is equal to the true average of the observed peak flows by construction.

2.4.3.2. *Model for reference hydrographs*

Now that we have registered the hydrographs of each watershed, we can model the reference hydrographs with the probabilistic model exposed in section 2.3.1. In order to determine an adequate spline basis, the model selection approach



(a) Observed hydrographs and average hydrograph

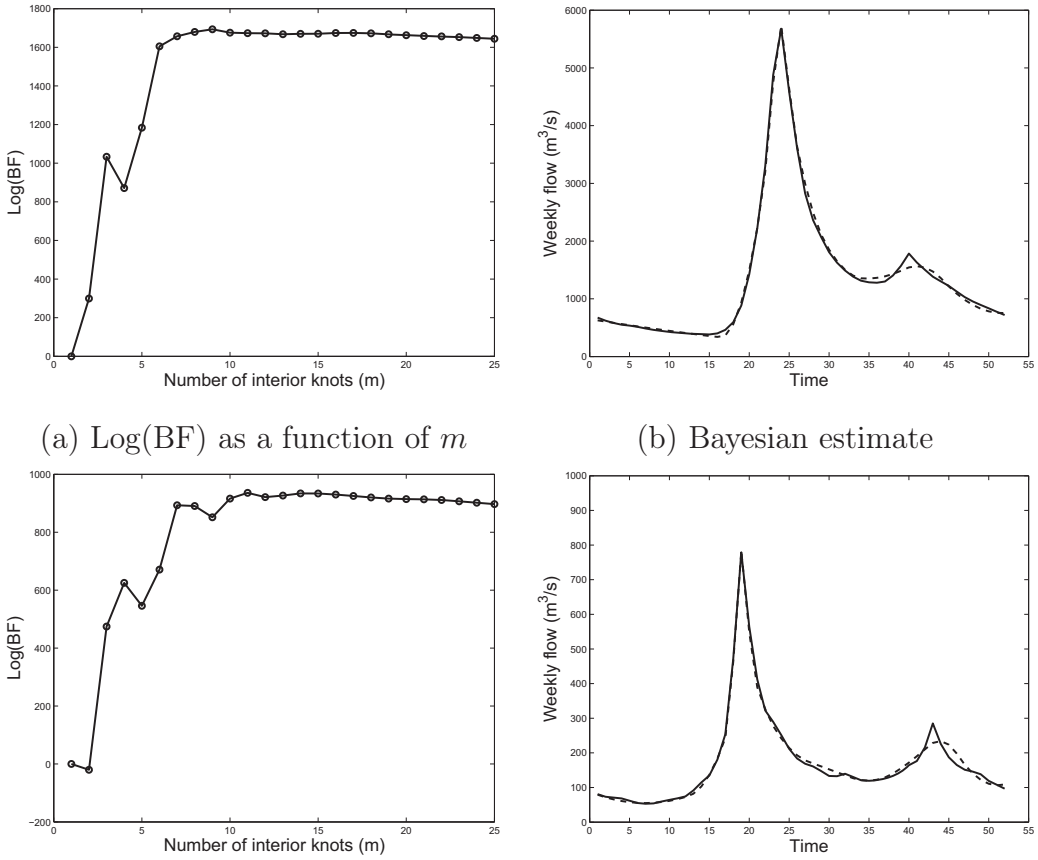
(b) Registered hydrographs and reference hydrograph

(c) Observed hydrographs and average hydrograph

(d) Registered hydrographs and reference hydrograph

FIGURE 2.9. Churchill Falls : (a) observed hydrographs (dotted lines) and their average (full bold line), (b) registered hydrographs (dotted lines) and their average (full bold line). Gouin : (c) observed hydrographs (dotted lines) and their average (full bold line), (d) registered hydrographs (dotted lines) and their average (full bold line).

described in section 2.3.2 is used. Figure 2.10, panels (a) and (c), shows the logarithm of the Bayes factors calculated with a reference model containing a single knot (the denominator of equation (2.3.11)). By writing a model which contains m interior knots as model $\boldsymbol{\omega}_m$, the Bayes factors shown in Figure 2.10 are given by $BF_{\boldsymbol{\omega}_m,\boldsymbol{\omega}_1}$, for $m = 1, ..., 25$. The maximum of $BF_{\boldsymbol{\omega}_m,\boldsymbol{\omega}_1}$ for Churchill Falls is obtained for 9 interior knots, *i.e.* a model basis containing 12 elements ; for



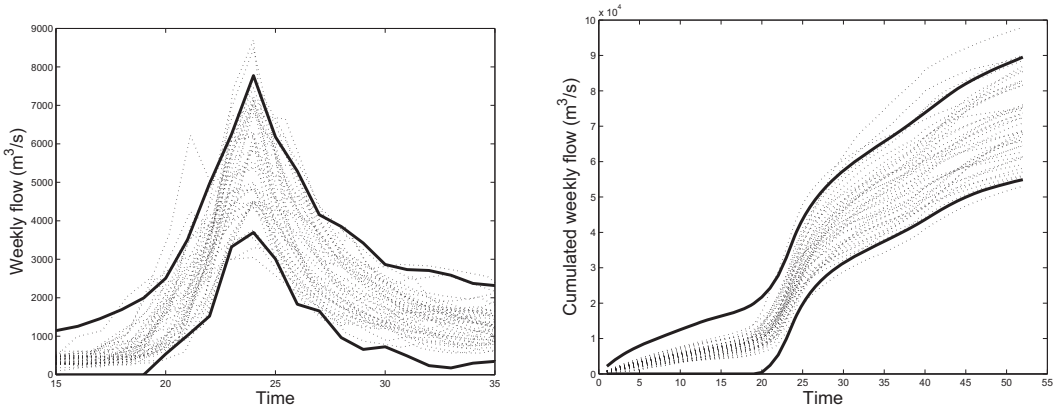
(a) Log(BF) as a function of m (b) Bayesian estimate

(c) Log(BF) as a function of m (d) Bayesian estimate

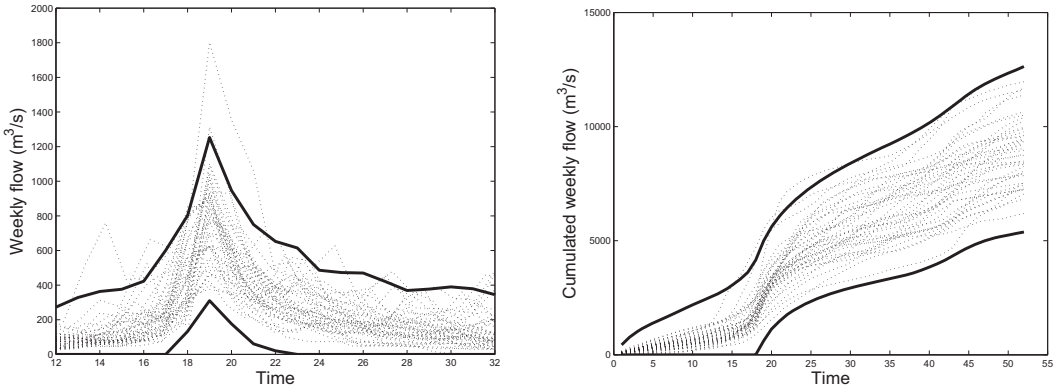
FIGURE 2.10. Churchill Falls : (a) logarithm of the Bayes factors (see equation (B.12)) : 9 interior knots give the best model ; (b) the reference hydrograph (solid line) of Figure 2.9(b) and the Bayesian estimate (dashed line). Gouin : (c) logarithm of the Bayes factors : 11 interior knots give the best model ; (d) the reference hydrograph (solid line) of Figure 2.9(d) and the Bayesian estimate (dashed line).

Gouin, the best model according to the Bayes factors contains 14 basis elements defined by 11 interior knots.

The Bayesian estimates given by equation (2.3.12) are illustrated in panels (b) and (d) of Figure 2.10. The reference hydrographs are shown as full lines, while dashed lines illustrate the estimates. We see that these estimates fit the reference hydrographs very accurately in the flood event region. Small discrepancies occur around the area of the autumn flood but globally the model performs



(a) Confidence interval for flood events (b) Confidence interval for cumulative hydrographs



(c) Confidence interval for flood events (d) Confidence interval for cumulative hydrographs

FIGURE 2.11. Churchill Falls : (a) 95% confidence interval for sample of flood events ; (b) 95% confidence interval for sample of cumulative hydrographs. Gouin : (a) 95% confidence interval for sample of flood events ; (b) 95% confidence interval for sample of cumulative hydrographs.

very well. It should be noted that the reference hydrographs, which are functions of dimension 52, have had their dimensions reduced considerably by their representation in functional space. The reference hydrograph of Churchill Falls is modelled by a parameter space of dimension 12, while the reference hydrograph of Gouin is modelled by a parameter space of dimension 14. As a comparative example, a PARMA(1,1) model with a period equal to the time increment of

the series would lead to a parameter space of dimension $208 = (52 \times 4)$ since for each time increment, there are 4 parameters to evaluate : the mean and the standard deviation of the observed hydrographs, as well as the two parameters of the ARMA process.

2.4.3.3. *Confidence intervals for the samples of registered yearly hydrographs*

It is possible, using equation (B.14) in section B.4, to construct simultaneous confidence intervals for the data points of a reference hydrograph and those of the cumulative reference hydrograph. Here, we consider 95% confidence intervals for the flood events and for the cumulative hydrographs. For the first confidence intervals, this is done by setting $\boldsymbol{a} = \boldsymbol{b}^M_{\boldsymbol{\omega}}(x_j)$ for a given data point j (see equation (2.2.5)). For the second confidence intervals, we set $\boldsymbol{a} = \boldsymbol{b}^I_{\boldsymbol{\omega}}(x_j)$ for a given data point j (see equation (2.2.8)).

The confidence intervals for the flood events of Churchill Falls are illustrated in panel (a) of Figure 2.11, along with the registered flood events. Panel (b) shows the confidence intervals for all the data points of the cumulative hydrograph, along with the yearly registered cumulative hydrographs. The same exercise is performed for Gouin and the results are shown in panels (c) and (d).

The effective sample has a size of $N = 42$ and there should thus be at most 2 or 3 yearly hydrographs outside a 95% confidence interval for a given data point. Panels (a) and (c) indicate that this is the case for most data points in the flood region. These confidence intervals are therefore well calibrated, or unbiased, since they seem to be able to capture the level of variability of the sample of flood events. The only data points which seem to behave differently are the ones located just before the flood peak for Churchill Falls. It should also be noted that the confidence intervals for the weeks preceding the beginning of the flood events are quite large. These two aspects are due to the constant variance assumption (A1) which causes the variance to be an average of the variances at each week. The level of variability is thus overestimated for the weeks preceding the flood event and underestimated for the weeks close to the flood peak.

For panels (b) and (d), there are at most 2 or 3 cumulative yearly hydrographs outside the confidence intervals for the majority of data points. Although the confidence intervals seem to reproduce the level of variability fairly well for most of the year, they appear to be unnecessarily wide in the first few weeks of the year ; this is also a result of the constant variance assumption.

2.5. Conclusion

In this paper, we have put forward a new approach to model the average properties of a sample of yearly hydrographs. We elaborated a methodology to obtain reference hydrographs representative of given samples and it was seen that the reference hydrographs reproduce adequately the flood events encountered in the samples. Furthermore, a previous study (Merleau *et al.*, 2005) has shown that the constructed reference hydrographs also preserve the average flood event volumes of hydrograph samples. We also exposed a nonparametric regression method in a Bayesian setting to model a reference hydrograph or any particular hydrograph in the sample. The approach was applied to two samples of yearly hydrographs with weekly streamflow in order to obtain a statistical representation of two reference hydrographs of different watersheds. Using the statistical model, confidence intervals were produced for the flood events and the cumulative streamflows of each watershed hydrographs. Although we didn't present an analysis of yearly hydrographs with daily measurements, we have found that the method performs just as well in this case.

The methodology proposed in this paper can be related to existing methods to model and simulate hydrographs. Using our model on a single typical hydrograph, our approach would be similar to the TH methods (see section 2.1). One major difference though is the fact that our model is statistically based and it can therefore be used to construct confidence intervals for example. As pointed out in section 2.2.2, the relation between the S methods and our approach is quite clear although the latter considers the hydrographs as random functions, while the former do not. Furthermore, our model is more general since it relies on a

functional representation based on spline functions, which can reproduce a wide variety of hydrograph shapes.

Finally, as mentioned in section 2.1, the modelling techniques based on time series are mainly used to simulate hydrographs corresponding to probable scenarios. In the Bayesian statistical context, the yearly hydrographs are treated as random functional events and it is thus also possible to simulate hydrographs in this setting. Vectors of parameters can be generated from the the posterior probability distribution (equation (2.3.10)) and to each vector of parameters corresponds a simulated hydrograph. If no constraints are placed on the parameters, the generated hydrographs represent random events and in this respect, our approach resembles the time series based methods. In our statistical framework though, it is also possible to simulate hydrographs which have fixed flood characteristics, with a certain probability of occurence, if constraints are put on the parameters.

We have seen that the method proposed in this paper performs very well globally. The aspects that still require attention are the constant variance assumption and the exploration of knot configurations. It was seen in section 2.4.3.3 that the constant variance throughout the year leads to confidence intervals which appear to be too wide in certain periods and too narrow in other periods. We are currently working on a method which models the variance components at each time increment ; this will correct the width of the confidence intervals. Although the knot determination method used in this paper performs well, it would be preferable to use a random knot selection procedure which explores a wide variety of knot configurations ; we are also working on a random knot procedure at the present time.

Acknowledgements

The authors would like to thank Frédéric Guay for providing the map shown in Figure 2.7 and the referees for their constructive comments, which improved the paper substantially.

Appendix A. Probability Distributions

Multivariate Normal distribution

If $\boldsymbol{X} \sim \mathrm{N}_q(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, then the variable $\boldsymbol{X}$ is distributed according to a q dimensional multivariate normal distribution which is given by :

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{q}{2}} \mid \boldsymbol{\Sigma} \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})\right\},$$

where $\boldsymbol{\theta}$ $(q \times 1)$ is the location vector of the distribution, $\boldsymbol{\Sigma}$ $(q \times q)$ is the covariance matrix associated with the components of $\boldsymbol{X}$ and $|\cdot|$ represents the determinant.

Inverse Gamma distribution

If $X \sim \mathrm{I}\Gamma(\alpha, \gamma)$, then the variable X is distributed according to an inverse gamma distribution which is given by :

$$f(x \mid \alpha, \gamma) = \frac{\gamma^{\alpha}}{\Gamma(\alpha)} \frac{1}{x^{\alpha+1}} \exp\left\{-\frac{\gamma}{x}\right\}.$$

Student's t distribution

If $\boldsymbol{X} \sim \mathrm{T}_q(\nu, \boldsymbol{\theta}, \boldsymbol{\Sigma})$, then the variable $\boldsymbol{X}$ is distributed according to a q dimensional Student's t distribution which is given by :

$$f(\boldsymbol{x} \mid \nu, \boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{\Gamma(\frac{\nu+q}{2})}{\mid \boldsymbol{\Sigma} \mid^{\frac{1}{2}} (\nu\pi)^{\frac{q}{2}} \Gamma(\frac{\nu}{2})} \left\{1 + \frac{(\boldsymbol{x}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})}{\nu}\right\}^{-\frac{\nu+q}{2}},$$

where $\boldsymbol{\theta}$ $(q \times 1)$ is the location vector of the distribution, $\boldsymbol{\Sigma}$ $(q \times q)$ is the covariance matrix associated with the components of $\boldsymbol{X}$, ν is the number of degrees of freedom and $|\cdot|$ represents the determinant.

Appendix B. Bayesian results

B.1. Parameters of the posterior statistical distributions

The parameters of the posterior statistical distributions are

$$\boldsymbol{\beta}^*_{\boldsymbol{\omega}} = (\boldsymbol{W}^G_{\boldsymbol{\omega}} + \boldsymbol{W}^0_{\boldsymbol{\omega}})^{-1}(\boldsymbol{W}^G_{\boldsymbol{\omega}}\boldsymbol{\beta}^G_{\boldsymbol{\omega}} + \boldsymbol{W}^0_{\boldsymbol{\omega}}\boldsymbol{\beta}^0_{\boldsymbol{\omega}}) \tag{B.1}$$

$$=^A \left(\frac{N}{N+n_0}\right)\boldsymbol{\beta}^L_{\boldsymbol{\omega}} + \left(\frac{n_0}{N+n_0}\right)\boldsymbol{\beta}^0_{\boldsymbol{\omega}}, \tag{B.2}$$

$$\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}} = (\boldsymbol{W}^G_{\boldsymbol{\omega}} + \boldsymbol{W}^0_{\boldsymbol{\omega}})^{-1} =^A \left(\frac{1}{N+n_0}\right)(\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{B}_{\boldsymbol{\omega}})^{-1}, \tag{B.3}$$

$$\alpha^*_{\boldsymbol{\omega}} = Nn + \alpha_{\boldsymbol{\omega}} =^A (N+n_0)\,n, \tag{B.4}$$

$$\gamma^*_{\boldsymbol{\omega}} = NS_{\boldsymbol{\omega}} + T_{\boldsymbol{\omega}} + \gamma_{\boldsymbol{\omega}} =^A NS_{\boldsymbol{\omega}} + T_{\boldsymbol{\omega}} + n_0 S^0_{\boldsymbol{\omega}}, \tag{B.5}$$

$$S_{\boldsymbol{\omega}} = \frac{\sum_i (\boldsymbol{y}_i - \boldsymbol{B}_{\boldsymbol{\omega}}\boldsymbol{\beta}^G_{\boldsymbol{\omega}})'\boldsymbol{\Sigma}^{-1}_{\varepsilon}(\boldsymbol{y}_i - \boldsymbol{B}_{\boldsymbol{\omega}}\boldsymbol{\beta}^G_{\boldsymbol{\omega}})}{N} \tag{B.6}$$

$$=^A \frac{\sum_i (\boldsymbol{y}_i - \boldsymbol{B}_{\boldsymbol{\omega}}\boldsymbol{\beta}^L_{\boldsymbol{\omega}})'(\boldsymbol{y}_i - \boldsymbol{B}_{\boldsymbol{\omega}}\boldsymbol{\beta}^L_{\boldsymbol{\omega}})}{N}, \tag{B.7}$$

$$T_{\boldsymbol{\omega}} = (\boldsymbol{\beta}^G_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}})' \left\{(N\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{\Sigma}^{-1}_{\varepsilon}\boldsymbol{B}_{\boldsymbol{\omega}})^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\omega}}\right\}^{-1}(\boldsymbol{\beta}^G_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}}) \tag{B.8}$$

$$=^A \left(\frac{n_0 N}{N+n_0}\right)(\boldsymbol{\beta}^L_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}})'\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{B}_{\boldsymbol{\omega}}(\boldsymbol{\beta}^L_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}}), \tag{B.9}$$

where $=^A$ indicates the given quantity evaluated under assumptions (A1), (A3), and (A4).

The posterior mean of the parameters, $\boldsymbol{\beta}^*_{\boldsymbol{\omega}}$, is given by a weighted average of the generalized least squares estimator $\boldsymbol{\beta}^G_{\boldsymbol{\omega}} = (\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{\Sigma}^{-1}_{\varepsilon}\boldsymbol{B}_{\boldsymbol{\omega}})^{-1}\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{\Sigma}^{-1}_{\varepsilon}\bar{\boldsymbol{y}}$ and of the prior location vector $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$. The weights are given by $\boldsymbol{W}^G_{\boldsymbol{\omega}} = N\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{\Sigma}^{-1}_{\varepsilon}\boldsymbol{B}_{\boldsymbol{\omega}}$ and $\boldsymbol{W}^0_{\boldsymbol{\omega}} = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\omega}}$. Under the application assumptions, $\boldsymbol{\beta}^*_{\boldsymbol{\omega}}$ simplifies to a weighted average of the ordinary least squares estimator $\boldsymbol{\beta}^L_{\boldsymbol{\omega}} = (\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{B}_{\boldsymbol{\omega}})^{-1}\boldsymbol{B}'_{\boldsymbol{\omega}}\bar{\boldsymbol{y}}$ and of the prior location vector, where the weights are proportional to the number of observations used to evaluate each of these quantities.

We also note that the posterior probability distribution of the variance parameter σ^2 depends on $S_{\boldsymbol{\omega}}$, which is proportional to the sum of squares of the residuals, and $T_{\boldsymbol{\omega}}$ which captures the discrepancy between the two vectors of parameters $\boldsymbol{\beta}^G_{\boldsymbol{\omega}}$ and $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$.

B.2. Marginal distribution

For the model exposed in section 2.3, the marginal distribution is given by

$$\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}) = \left(\frac{|\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}}|}{|\boldsymbol{\Sigma}_{\varepsilon}||\boldsymbol{\Sigma}_{\boldsymbol{\omega}}|}\right)^{1/2} \left(\frac{\Gamma(\alpha^*_{\boldsymbol{\omega}}/2)}{\pi^{Nn/2}\Gamma(\alpha_{\boldsymbol{\omega}}/2)}\right) \left(\frac{(\gamma_{\boldsymbol{\omega}})^{\alpha_{\omega}/2}}{(\gamma^*_{\boldsymbol{\omega}})^{\alpha^*_{\omega}/2}}\right) \tag{B.10}$$

$$=^A \left(\frac{n_0}{N+n_0}\right)^{K_{\omega}/2} \left(\frac{\Gamma(\alpha^*_{\boldsymbol{\omega}}/2)}{\pi^{Nn/2}\Gamma(\alpha_{\boldsymbol{\omega}}/2)}\right) \left(\frac{(\gamma_{\boldsymbol{\omega}})^{\alpha_{\omega}/2}}{(\gamma^*_{\boldsymbol{\omega}})^{\alpha^*_{\omega}/2}}\right), \tag{B.11}$$

where $|\cdot|$ represents the determinant.

B.3. Bayes factor

Using the marginal distribution given in equation (B.10), the Bayes factor for our model is

$$BF_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} = \left(\frac{|\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}_1}|}{|\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}_2}|}\frac{|\boldsymbol{\Sigma}_{\boldsymbol{\omega}_2}|}{|\boldsymbol{\Sigma}_{\boldsymbol{\omega}_1}|}\right)^{1/2} \left(\frac{\Gamma(\alpha^*_{\boldsymbol{\omega}_1}/2)}{\Gamma(\alpha^*_{\boldsymbol{\omega}_2}/2)}\frac{\Gamma(\alpha_{\boldsymbol{\omega}_2}/2)}{\Gamma(\alpha_{\boldsymbol{\omega}_1}/2)}\right)$$
$$\times \left(\frac{(\gamma_{\boldsymbol{\omega}_1})^{\alpha_{\omega_1}/2}}{(\gamma_{\boldsymbol{\omega}_2})^{\alpha_{\omega_2}/2}}\frac{(\gamma^*_{\boldsymbol{\omega}_2})^{\alpha^*_{\omega_2}/2}}{(\gamma^*_{\boldsymbol{\omega}_1})^{\alpha^*_{\omega_1}/2}}\right) \tag{B.12}$$

$$=^A \left(\frac{n_0}{N+n_0}\right)^{(K_{\omega_1}-K_{\omega_2})/2} \left(\frac{\Gamma(\alpha^*_{\boldsymbol{\omega}_1}/2)}{\Gamma(\alpha^*_{\boldsymbol{\omega}_2}/2)}\frac{\Gamma(\alpha_{\boldsymbol{\omega}_2}/2)}{\Gamma(\alpha_{\boldsymbol{\omega}_1}/2)}\right)$$
$$\times \left(\frac{(\gamma_{\boldsymbol{\omega}_1})^{\alpha_{\omega_1}/2}}{(\gamma_{\boldsymbol{\omega}_2})^{\alpha_{\omega_2}/2}}\frac{(\gamma^*_{\boldsymbol{\omega}_2})^{\alpha^*_{\omega_2}/2}}{(\gamma^*_{\boldsymbol{\omega}_1})^{\alpha^*_{\omega_1}/2}}\right), \tag{B.13}$$

where $K_{\boldsymbol{\omega}_1}$ and $K_{\boldsymbol{\omega}_2}$ are the number of parameters in models $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ respectively, and the other quantities are defined in section B.1.

B.4. Bayesian confidence intervals (credible sets)

Simultaneously for all vectors $\boldsymbol{a}$, a $100(1-\delta)\%$ confidence interval is given by

$$\boldsymbol{a}'\boldsymbol{\beta}^*_{\boldsymbol{\omega}} \pm \left\{K_{\boldsymbol{\omega}}\left(\frac{\gamma^*_{\boldsymbol{\omega}}}{\alpha^*_{\boldsymbol{\omega}}}\right)\boldsymbol{a}'\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}}\boldsymbol{a}F_{K_{\omega},\alpha^*_{\omega}}(\delta)\right\}^{1/2}, \tag{B.14}$$

where $F_{K_{\omega},\alpha^*_{\omega}}(\delta)$ represents the $100(1-\delta)$th percentile of Fisher's F distribution with degrees of freedom $K_{\boldsymbol{\omega}}$ and $\alpha^*_{\boldsymbol{\omega}}$, and the other quantities are defined as before.

Chapitre 3

FAMILLE DE DISTRIBUTIONS ET ESTIMATION POUR L'ARTICLE 2

Dans le présent chapitre, la famille de distributions statistiques que nous proposons est mise de l'avant. Cette famille nous permet d'avoir un contexte de modélisation non paramétrique assez général et ainsi d'étudier une diversité de distributions statistiques desquelles peuvent provenir les observations de débits. Nous présentons aussi les détails techniques reliés à l'estimation des paramètres définissant le type de modèles qui nous intéresse. Plus spécifiquement, nous déterminons les estimateurs du maximum de vraisemblance et du maximum de la distribution *a posteriori* pour les paramètres. Ces estimateurs sont utilisés afin d'estimer les fonctions à modéliser, mais ils interviennent aussi dans les deux approximations considérées de la distribution marginale des observations (voir article 2). En effet, la première approximation est basée sur le critère de Schwarz (Schwarz, 1978) qui fait appel aux estimateurs du maximum de vraisemblance. La seconde approximation, basée sur l'approximation de Laplace d'une intégrale, est employée pour le modèle bayésien complet et dépend du maximum de la distribution *a posteriori*.

3.1. La famille de distributions

Le cadre présenté dans ce qui suit est général et il nous permet de traiter un ensemble de distributions d'une façon uniforme. Il serait possible d'aborder certaines distributions en tenant compte de leurs particularités spécifiques mais

ceci n'est pas fait ci-après. Toutes les distributions que nous étudions dans les trois articles appartiennent à la famille de distributions, $\mathcal{F}$, pour laquelle la densité de probabilité s'écrit

$$f(y|\zeta,\phi) = \exp\left\{\phi[\zeta z - \psi(\zeta)] + \kappa(y,\phi)\right\}, \tag{3.1.1}$$

où ζ et ϕ sont respectivement le paramètre canonique de position et le paramètre de précision, $z = z(y)$ est une fonction bijective de y, alors que les fonctions ψ et κ définissent les distributions spécifiques. Les densités de cette famille possèdent la forme de celles appartenant à la famille exponentielle, mais elles nous permettent aussi de considérer des transformations de variables provenant de la famille exponentielle.

La famille $\mathcal{F}$ possède de bonnes propriétés de régularité (voir par exemple Casella et Berger, 2001 ; McCullagh et Nelder, 1989), ce qui permet d'obtenir des expressions simples entre les moments centrés et les dérivées de la fonction $\psi(\zeta)$. Plus spécifiquement, pour nos besoins, nous notons que la moyenne et la variance de la variable transformée sont données par

$$\mathbb{E}(z) = \mu = \nabla_\zeta\left[\psi(\zeta)\right], \tag{3.1.2}$$

$$\mathbb{V}(z) = \phi^{-1}v(\mu) = \phi^{-1}\nabla_\zeta\nabla_\zeta\left[\psi(\zeta)\right] = \phi^{-1}\nabla_\zeta\left[\mu(\zeta)\right], \tag{3.1.3}$$

où ∇_ζ est la dérivée partielle par rapport à ζ (voir l'Annexe A du présent chapitre pour les détails concernant les opérateurs différentiels utilisés dans la présente section) et $v(\mu)$ représente la fonction variance, c'est-à-dire la relation entre la variance et la moyenne. Ces deux propriétés nous indiquent le lien intrinsèque qui existe entre le paramètre canonique ζ et le paramètre μ.

Les densités de probabilité étudiées dans les applications sont données au tableau 3.1. Nous considérons les distributions suivantes : normale ($\mathcal{N}$), gamma ($\mathcal{G}$), inverse gaussienne (IG), lognormale ($\mathcal{LN}$) et réciproque inverse gaussienne (RIG). Le tableau 3.1 présente explicitement les différentes fonctions pour spécifier les densités, les relations entre le paramètre canonique et le paramètre μ, ainsi que les fonctions variance, v (voir équation (3.1.3)). Dans le tableau, nous faisons appel à la décomposition suivante : $\kappa(y,\phi) = \kappa_1(y,\phi) + \kappa_2(\phi) + \kappa_3(y) + \kappa_4$, où

κ_4 est un terme constant. L'espérance et la variance de la variable non transformée sont ensuite explicitées. Finalement, les fonctions lien canonique, $g(\mu)$, des différentes distributions sont données (voir ci-après).

TABLE 3.1. Information pour les densités de probabilité étudiées appartenant à la famille $\mathcal{F}$, soient les distributions normale ($\mathcal{N}$), gamma ($\mathcal{G}$), inverse gaussienne (IG), lognormale ($\mathcal{LN}$) et réciproque inverse gaussienne (RIG).

	$\mathcal{N}(\mu,\phi)$	$\mathcal{G}(\mu,\phi)$	IG (μ,ϕ)	$\mathcal{LN}(\mu,\phi)$	RIG (μ,ϕ)
$z(y)$	y	y	y	$\log(y)$	y^{-1}
$\psi(\zeta)$	$\frac{\zeta^2}{2}$	$-\log(-\zeta)$	$-(-2\zeta)^{1/2}$	$\frac{\zeta^2}{2}$	$-\log(-\zeta)$
$\zeta(\mu)$	μ	$-\mu^{-1}$	$-\frac{1}{2}\mu^{-2}$	μ	$-\frac{1}{2}\mu^{-2}$
$\psi[\zeta(\mu)]$	$\frac{\mu^2}{2}$	$\log(\mu)$	$-\mu^{-1}$	$\frac{\mu^2}{2}$	$-\mu^{-1}$
$v(\mu)$	1	μ^2	μ^3	1	μ^3
$\kappa_1(y,\phi)$	$-\phi\frac{1}{2}y^2$	$\phi\log(y)$	$-\phi\frac{1}{2}y^{-1}$	$-\phi\frac{1}{2}[\log(y)]^2$	$-\phi\frac{1}{2}y$
$\kappa_2(\phi)$	$\frac{1}{2}\log(\phi)$	$\phi\log(\phi)-\log[\Gamma(\phi)]$	$\frac{1}{2}\log(\phi)$	$\frac{1}{2}\log(\phi)$	$\frac{1}{2}\log(\phi)$
$\kappa_3(y)$	0	$-\log(y)$	$-\frac{1}{2}\log(y^3)$	$-\log(y)$	$-\frac{1}{2}\log(y)$
κ_4	$-\frac{1}{2}\log(2\pi)$	0	$-\frac{1}{2}\log(2\pi)$	$-\frac{1}{2}\log(2\pi)$	$-\frac{1}{2}\log(2\pi)$
$\mu_y=\mathbb{E}(y)$	μ	μ	μ	$\exp(\mu+\phi^{-1}/2)$	$\mu^{-1}+\phi^{-1}$
$\mathbb{V}(y)$	ϕ^{-1}	$\phi^{-1}\mu^2$	$\phi^{-1}\mu^3$	$\mu_y^2(\exp(\phi^{-1})-1)$	$\phi^{-1}\mu_y+\phi^{-2}$
$g(\mu)$	μ	μ^{-1}	μ^{-2}	μ	μ^{-2}

Considérons maintenant un échantillon $\boldsymbol{y}=(y_1,\ldots,y_j,\ldots,y_n)'$ pour lequel les observations sont indépendamment distribuées selon la famille $\mathcal{F}(\zeta_j,\phi)$, c'est-à-dire $y_j\sim\mathcal{F}(\zeta_j,\phi)$. La densité conjointe est alors donnée par

$$f(\boldsymbol{y}|\boldsymbol{\zeta},\phi)=\prod_{j=1}^{n}\exp\{\phi[\zeta_j z_j-\psi(\zeta_j)]+\kappa(y_j,\phi)\}, \tag{3.1.4}$$

$$=\exp\{\phi[\boldsymbol{\zeta}'\boldsymbol{z}-\boldsymbol{\psi}(\boldsymbol{\zeta})'\mathbf{1}]+\boldsymbol{\kappa}(\boldsymbol{y},\phi)'\mathbf{1}\}, \tag{3.1.5}$$

$$=\exp\{\mathcal{L}\}, \tag{3.1.6}$$

où

$$\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_j, \ldots, \zeta_n)', \tag{3.1.7}$$

$$\boldsymbol{z} = (z_1, \ldots, z_j, \ldots, z_n)', \tag{3.1.8}$$

$$\boldsymbol{\psi}(\boldsymbol{\zeta}) = (\psi(\zeta_1), \ldots, \psi(\zeta_j), \ldots, \psi(\zeta_n))', \tag{3.1.9}$$

$$\boldsymbol{\kappa}(\boldsymbol{y}, \phi) = (\kappa(y_1, \phi), \ldots, \kappa(y_j, \phi), \ldots, \kappa(y_n, \phi))', \tag{3.1.10}$$

$\mathbf{1}$ est un vecteur unitaire de taille $n \times 1$ et $\mathcal{L}$ représente le logarithme de la densité conjointe, ou de la vraisemblance. La représentation vectorielle, donnée à l'équation (3.1.5), est celle que nous utilisons lors de l'obtention des estimateurs.

D'une façon analogue aux modèles linéaires généralisés (McCullagh et Nelder, 1989), nous définissons les modèles statistiques étudiés par les trois composantes suivantes :

(1) une composante aléatoire : une distribution statistique provenant de $\mathcal{F}$, $y_j \sim \mathcal{F}(\zeta_j, \phi)$;

(2) une composante systématique : une fonction linéaire basée sur de l'information auxiliaire, $u_j = \boldsymbol{x}_j' \boldsymbol{\beta}$, où $\boldsymbol{x}_j$ est un vecteur $K_{\boldsymbol{\beta}} \times 1$ d'information auxiliaire et $\boldsymbol{\beta}$ est un vecteur de paramètres $K_{\boldsymbol{\beta}} \times 1$;

(3) une fonction lien, g, qui relie la moyenne de la composante aléatoire et la composante systématique : $g(\mathbb{E}(z_j)) = u_j = \boldsymbol{x}_j' \boldsymbol{\beta}$.

Dans notre contexte de modélisation, nous ne disposons pas d'information auxiliaire mais notre modèle non paramétrique s'écrit sous la même forme qu'un modèle de régression. Par conséquent, l'information auxiliaire doit ici être comprise comme une base de modélisation. Plusieurs fonctions lien peuvent être explorées dans ce contexte de modélisation mais nous nous limitons à trois fonctions : identité, inverse et logarithmique. Certaines de ces fonctions correspondent aux liens canoniques des distributions étudiées (voir tableau 3.1) ; l'utilisation du lien canonique implique que le paramètre canonique de position d'une distribution est modélisé par la composante systématique. Nous notons que lors de la dérivation

des estimateurs, nous travaillons avec l'inverse de la fonction lien. Plus précisément, nous utilisons : $\mathbb{E}(z_j) = h\left(u_j\right) = g^{-1}\left(u_j\right)$, où g^{-1} est l'inverse de la fonction lien.

3.2. Estimation par maximum de vraisemblance

Dans tout ce qui suit, nous considérons la matrice d'incidence comme fixe, ce qui correspond à une série de noeuds intérieurs fixe dans le contexte de la modélisation avec les splines de régression. Puisque nous travaillons vectoriellement, il est utile de noter que le vecteur des composantes systématiques s'écrit $\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta}$, et que le vecteur des moyennes peut être exprimé relativement à l'inverse de la fonction lien, soit $\boldsymbol{\mu} = \boldsymbol{h}\left(\boldsymbol{u}\right)$. La matrice d'incidence, $\boldsymbol{X}$, est de dimension $n \times K_{\boldsymbol{\beta}}$ et la ligne j est donnée par $\boldsymbol{x}_j'$. De plus, la notation vectorielle pour l'inverse de la fonction lien doit être comprise comme l'application aux éléments individuels de cet inverse de la fonction lien.

Nous cherchons la solution de $\nabla_{\boldsymbol{\beta}}\mathcal{L} = \mathbf{0}$, où $\mathcal{L}$ est le logarithme de la vraisemblance défini aux équations (3.1.5) et (3.1.6). Nous avons

$$\nabla_{\boldsymbol{\beta}}\mathcal{L} = \phi\left\{\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\zeta}'\boldsymbol{z}\right] + \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\psi}(\boldsymbol{\zeta})'\mathbf{1}\right]\right\}, \tag{3.2.1}$$

$$= \phi\left\{\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\zeta}'\right]\boldsymbol{z} + \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\psi}(\boldsymbol{\zeta})'\right]\mathbf{1}\right\}. \tag{3.2.2}$$

En appliquant la règle des dérivées en chaîne aux deux expressions à différentier, nous obtenons

$$\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\zeta}'\right] = \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{u}'\right]\nabla_{\boldsymbol{u}}\left[\boldsymbol{\mu}'\right]\nabla_{\boldsymbol{\mu}}\left[\boldsymbol{\zeta}'\right], \tag{3.2.3}$$

$$\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\psi}(\boldsymbol{\zeta})'\right] = \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{u}'\right]\nabla_{\boldsymbol{u}}\left[\boldsymbol{\mu}'\right]\nabla_{\boldsymbol{\mu}}\left[\boldsymbol{\zeta}'\right]\nabla_{\boldsymbol{\zeta}}\left[\boldsymbol{\psi}(\boldsymbol{\zeta})'\right]. \tag{3.2.4}$$

En utilisant les deux propriétés données aux équations (3.1.2) et (3.1.3), ainsi que les propriétés de l'Annexe A du présent chapitre, nous calculons

$$\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{u}'\right] = \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}'\boldsymbol{X}'\right] = \boldsymbol{X}', \tag{3.2.5}$$

$$\nabla_{\boldsymbol{u}}\left[\boldsymbol{\mu}'\right] = \nabla_{\boldsymbol{u}}\left[\boldsymbol{h}\left(\boldsymbol{u}\right)'\right] = \operatorname{Diag}\left\{\dot{h}_j\right\}, \tag{3.2.6}$$

$$\nabla_{\boldsymbol{\mu}}\left[\boldsymbol{\zeta}'\right] = \operatorname{Diag}\left\{v_j^{-1}\right\}, \tag{3.2.7}$$

$$\nabla_{\boldsymbol{\zeta}}\left[\boldsymbol{\psi}(\boldsymbol{\zeta})'\right] = \operatorname{Diag}\left\{\mu_j\right\}, \tag{3.2.8}$$

où $\dot{h}_j = \nabla_{u_j}[h(u_j)]$ et $v_j = v(\mu_j)$. Ainsi, nous pouvons maintenant écrire

$$\nabla_{\boldsymbol{\beta}}\mathcal{L} = \phi \boldsymbol{X}' \text{Diag}\left\{\dot{h}_j\right\} \text{Diag}\left\{v_j^{-1}\right\} (\boldsymbol{z} + \text{Diag}\{\mu_j\}\,\mathbf{1}), \tag{3.2.9}$$

$$= \phi \boldsymbol{X}' \boldsymbol{W} (\boldsymbol{z} - \boldsymbol{\mu}), \tag{3.2.10}$$

où $\boldsymbol{W} = \text{Diag}\left\{v_j^{-1}\dot{h}_j\right\}$.

En posant $\nabla_{\boldsymbol{\beta}}\mathcal{L} = \mathbf{0}$, nous trouvons que l'estimateur du maximum de vraisemblance, $\widehat{\boldsymbol{\beta}}$, est la solution du système d'équations suivant

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{\mu}, \tag{3.2.11}$$

Il est intéressant de noter que le système d'équations ne dépend pas du paramètre de précision ϕ. Bien que le système d'équations pour l'estimateur du maximum de vraisemblance soit bien défini, il est clair qu'il n'est pas linéaire en $\boldsymbol{\beta}$, étant donné la dépendance de la matrice $\boldsymbol{W}$ en $\boldsymbol{\beta}$. Les deux approches habituellement utilisées afin de résoudre pour $\widehat{\boldsymbol{\beta}}$ sont la méthode de Newton-Raphson et la méthode par score de Fisher. Afin de mettre en oeuvre une de ces deux méthodes, nous procédons maintenant au calcul de la matrice hessienne.

En notant $\nabla'_{\boldsymbol{\beta}}\mathcal{L} = \{\nabla_{\boldsymbol{\beta}}\mathcal{L}\}'$, la matrice hessienne est donnée par

$$\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\beta}}\mathcal{L} = \phi\nabla_{\boldsymbol{\beta}}\left[(\boldsymbol{z} - \boldsymbol{\mu})'\boldsymbol{W}\boldsymbol{X}\right], \tag{3.2.12}$$

$$= \phi\nabla_{\boldsymbol{\beta}}\left[(\boldsymbol{z} - \boldsymbol{\mu})'\boldsymbol{W}\right]\boldsymbol{X}. \tag{3.2.13}$$

Afin d'appliquer la règle du produit, il est utile de noter que le terme à différentier peut être écrit sous les trois formes suivantes

$$(\boldsymbol{z} - \boldsymbol{\mu})'\boldsymbol{W} = \begin{cases} \boldsymbol{a}_1'\boldsymbol{A}_1 \\ \boldsymbol{a}_2'\boldsymbol{A}_2 \\ \boldsymbol{a}_3'\boldsymbol{A}_3, \end{cases} \tag{3.2.14}$$

où

$$\boldsymbol{a}_1 = (\boldsymbol{z} - \boldsymbol{\mu}), \tag{3.2.15}$$

$$\boldsymbol{a}_2 = \left(v_1^{-1}, \ldots, v_j^{-1}, \ldots, v_n^{-1}\right)', \tag{3.2.16}$$

$$\boldsymbol{a}_3 = \left(\dot{h}_1, \ldots, \dot{h}_j, \ldots, \dot{h}_n\right)', \tag{3.2.17}$$

$$\boldsymbol{A}_1 = \text{Diag}\left\{v_j^{-1}\dot{h}_j\right\}, \tag{3.2.18}$$

$$\boldsymbol{A}_2 = \text{Diag}\left\{(z_j - \mu_j)\,\dot{h}_j\right\}, \tag{3.2.19}$$

$$\boldsymbol{A}_3 = \text{Diag}\left\{(z_j - \mu_j)\,v_j^{-1}\right\}. \tag{3.2.20}$$

En appliquant la règle du produit, nous avons alors

$$\nabla_{\boldsymbol{\beta}}\left[(\boldsymbol{z} - \boldsymbol{\mu})'\,\boldsymbol{W}\right] = \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{a}_1'\right]\boldsymbol{A}_1 + \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{a}_2'\right]\boldsymbol{A}_2 + \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{a}_3'\right]\boldsymbol{A}_3. \tag{3.2.21}$$

Le calcul des différents termes donne

$$\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{a}_1'\right] = -\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\mu}'\right] = -\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{u}'\right]\nabla_{\boldsymbol{u}}\left[\boldsymbol{h}\,(\boldsymbol{u})'\right] = -\boldsymbol{X}'\text{Diag}\left\{\dot{h}_j\right\}, \tag{3.2.22}$$

$$\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{a}_2'\right] = \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{u}'\right]\nabla_{\boldsymbol{u}}\left[\boldsymbol{h}\,(\boldsymbol{u})'\right]\nabla_{\boldsymbol{\mu}}\left[\boldsymbol{a}_2'\right] = -\boldsymbol{X}'\text{Diag}\left\{\dot{h}_j\dot{v}_j v_j^{-2}\right\}, \tag{3.2.23}$$

$$\nabla_{\boldsymbol{\beta}}\left[\boldsymbol{a}_3'\right] = \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{u}'\right]\nabla_{\boldsymbol{u}}\left[\boldsymbol{a}_3'\right] = \boldsymbol{X}'\text{Diag}\left\{\ddot{h}_j\right\}, \tag{3.2.24}$$

où $\dot{v}_j = \nabla_{\mu_j}\left[v(\mu_j)\right]$ et $\ddot{h}_j = \nabla_{u_j}\nabla'_{u_j}\left[h(u_j)\right]$. Avec ces résultats, la matrice hessienne peut maintenant être écrite de la façon suivante

$$\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\beta}}\mathcal{L} = \phi\boldsymbol{X}'\text{Diag}\left\{v_j^{-1}\left[(z_j - \mu_j)(\ddot{h}_j - v_j^{-1}\dot{v}_j\dot{h}_j^2) - \dot{h}_j^2\right]\right\}\boldsymbol{X}, \tag{3.2.25}$$

$$= \phi\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}, \tag{3.2.26}$$

où $\boldsymbol{V} = \text{Diag}\left\{v_j^{-1}\left[(z_j - \mu_j)(\ddot{h}_j - v_j^{-1}\dot{v}_j\dot{h}_j^2) - \dot{h}_j^2\right]\right\}$.

Nous pouvons directement calculer la matrice d'information de Fisher, $\mathcal{I}$, en notant que la matrice hessienne peut s'écrire sous la forme suivante

$$\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\beta}}\mathcal{L} = \phi\sum_{j=1}^{n}\left\{v_j^{-1}\left[(z_j - \mu_j)(\ddot{h}_j - v_j^{-1}\dot{v}_j\dot{h}_j^2) - \dot{h}_j^2\right]\right\}\boldsymbol{x}_j\boldsymbol{x}_j'. \tag{3.2.27}$$

La matrice d'information de Fisher est alors donnée par

$$\mathcal{I} = -\mathbb{E}\left\{\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\beta}}\mathcal{L}\right\}, \tag{3.2.28}$$

$$= -\phi\sum_{j=1}^{n}\left\{v_j^{-1}\left[(\ddot{h}_j - v_j^{-1}\dot{v}_j\dot{h}_j^2)\mathbb{E}(z_j-\mu_j) - \dot{h}_j^2\right]\right\}\boldsymbol{x}_j\boldsymbol{x}'_j, \tag{3.2.29}$$

$$= \phi\sum_{j=1}^{n}\left\{v_j^{-1}\dot{h}_j^2\right\}\boldsymbol{x}_j\boldsymbol{x}'_j, \tag{3.2.30}$$

$$= \phi\boldsymbol{X}'\boldsymbol{U}\boldsymbol{X}, \tag{3.2.31}$$

où $\boldsymbol{U} = \text{Diag}\left\{v_j^{-1}\dot{h}_j^2\right\}$.

Nous pouvons maintenant expliciter l'algorithme de Newton-Raphson et la méthode par score de Fisher. Dans le présent contexte, l'algorithme de Newton-Raphson est donné par :

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \left\{\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\beta}}\mathcal{L}\right\}^{-1}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}\left\{\nabla_{\boldsymbol{\beta}}\mathcal{L}\right\}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \tag{3.2.32}$$

$$= \boldsymbol{\beta}^{(m)} - \left\{\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\right\}^{-1}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}\left\{\boldsymbol{X}'\boldsymbol{W}\left(\boldsymbol{z}-\boldsymbol{\mu}\right)\right\}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \tag{3.2.33}$$

où $\boldsymbol{\beta}^{(m)}$ représente la valeur calculée à l'itération m.

La méthode par score de Fisher est également beaucoup utilisée dans les logiciels statistiques. Cette méthode est essentiellement une variante de l'algorithme de Newton-Raphson où la matrice hessienne est remplacée par la matrice d'information de Fisher multipliée par -1. Cet algorithme prend donc la forme suivante

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left\{\mathcal{I}\right\}^{-1}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}\left\{\nabla_{\boldsymbol{\beta}}\mathcal{L}\right\}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \tag{3.2.34}$$

$$= \boldsymbol{\beta}^{(m)} + \left\{\boldsymbol{X}'\boldsymbol{U}\boldsymbol{X}\right\}^{-1}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}\left\{\boldsymbol{X}'\boldsymbol{W}\left(\boldsymbol{z}-\boldsymbol{\mu}\right)\right\}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}. \tag{3.2.35}$$

Il est intéressant de noter que les deux algorithmes sont identiques lorsque le lien canonique d'une distribution est utilisé, c'est-à-dire lorsque $(\ddot{h}_j - v_j^{-1}\dot{v}_j\dot{h}_j^2) = 0$.

Une fois que l'on a estimé $\boldsymbol{\beta}$, nous pouvons calculer l'estimateur du maximum de vraisemblance du paramètre de précision ϕ. En notant le logarithme de la vraisemblance maximisée pour $\boldsymbol{\beta}$ par $\mathcal{L}(\widehat{\boldsymbol{\beta}})$, nous cherchons la solution de

$\nabla_\phi \left[\mathcal{L}(\widehat{\boldsymbol{\beta}})\right] = 0$. Nous avons

$$\nabla_\phi \left[\mathcal{L}(\widehat{\boldsymbol{\beta}})\right] = \nabla_\phi \left\{\phi \left[\hat{\boldsymbol{\zeta}}' \boldsymbol{z} - \boldsymbol{\psi}(\hat{\boldsymbol{\zeta}})' \mathbf{1}\right]\right\} + \nabla_\phi \boldsymbol{\kappa}(\boldsymbol{y}, \phi)' \mathbf{1}, \tag{3.2.36}$$

$$= \left[\hat{\boldsymbol{\zeta}}' \boldsymbol{z} - \boldsymbol{\psi}(\hat{\boldsymbol{\zeta}})' \mathbf{1}\right] + \nabla_\phi \left[\boldsymbol{\kappa}(\boldsymbol{y}, \phi)'\right] \mathbf{1}. \tag{3.2.37}$$

En posant $\nabla_\phi \mathcal{L}(\widehat{\boldsymbol{\beta}}) = 0$, l'estimateur du maximum de vraisemblance est la solution de l'équation suivante

$$\left[\nabla_\phi \boldsymbol{\kappa}_1(\boldsymbol{y}, \phi)' + \nabla_\phi \kappa_2(\phi) \mathbf{1}'\right] \mathbf{1} = -\left[\hat{\boldsymbol{\zeta}}' \boldsymbol{z} - \boldsymbol{\psi}(\hat{\boldsymbol{\zeta}})' \mathbf{1}\right], \tag{3.2.38}$$

où $\boldsymbol{\kappa}_1(\boldsymbol{y}, \phi) = (\kappa_1(y_1, \phi), \ldots, \kappa_1(y_j, \phi), \ldots, \kappa_1(y_n, \phi))'$, et les fonctions κ_1 et κ_2 proviennent de la décomposition de κ donnée au tableau 3.1.

Avec les distributions du tableau 3.1, nous avons deux cas particuliers et nous donnons maintenant les solutions explicites de chacun des cas. Pour les distributions $\mathcal{N}$, IG, $\mathcal{LN}$ et RIG, nous avons

$$\kappa_1(y_j, \phi) = -\phi s(y_j) \qquad \text{et} \qquad \kappa_2(\phi) = \frac{1}{2} \log(\phi), \tag{3.2.39}$$

où $s(y_j)$ est une fonction spécifique pour chacune des distributions (voir tableau 3.1). Par conséquent,

$$\nabla_\phi \boldsymbol{\kappa}_1(\boldsymbol{y}, \phi)' = -\boldsymbol{s}(\boldsymbol{y})' \qquad \text{et} \qquad \nabla_\phi \kappa_2(\phi) = (2\phi)^{-1}. \tag{3.2.40}$$

En remplaçant ces quantités dans l'équation (3.2.38), nous obtenons

$$\hat{\phi} = \frac{n}{2\left\{\boldsymbol{s}(\boldsymbol{y})' \mathbf{1} - \left[\hat{\boldsymbol{\zeta}}' \boldsymbol{z} - \boldsymbol{\psi}(\hat{\boldsymbol{\zeta}})' \mathbf{1}\right]\right\}}, \tag{3.2.41}$$

où nous utilisons le fait que $n = \mathbf{1}'\mathbf{1}$. Pour les 4 distributions données ci-dessus, la quantité au dénominateur correspond à la déviance, où la déviance est définie comme deux fois la différence entre le maximum du logarithme de la vraisemblance et la valeur de celui-ci sous un modèle à l'étude, pour un paramètre de précision unitaire (voir chapitre 2 de McCullagh et Nelder, 1989 ; et l'annexe A du chapitre 4).

Le second cas correspond à la distribution gamma ($\mathcal{G}$). Dans ce cas, nous avons

$$\kappa_1(y_j, \phi) = -\phi s(y_j) \qquad \text{et} \qquad \kappa_2(\phi) = \phi \log(\phi) - \log\left[\Gamma(\phi)\right], \tag{3.2.42}$$

où $s(y_j) = -\log(y_j)$. Conséquemment, nous obtenons

$$\nabla_\phi \boldsymbol{\kappa}_1(\boldsymbol{y}, \phi)' = -\boldsymbol{s}(\boldsymbol{y})' \qquad \text{et} \qquad \nabla_\phi \kappa_2(\phi) = \log(\phi) - \frac{\dot{\Gamma}(\phi)}{\Gamma(\phi)} + 1, \tag{3.2.43}$$

où $\dot{\Gamma}(\phi) = \nabla_\phi \Gamma(\phi)$. Pour obtenir $\hat{\phi}$, nous devons alors résoudre l'équation suivante

$$\log(\phi) - \frac{\dot{\Gamma}(\phi)}{\Gamma(\phi)} = \frac{\boldsymbol{s}(\boldsymbol{y})'\mathbf{1} - \left[\hat{\boldsymbol{\zeta}}'\boldsymbol{z} - \boldsymbol{\psi}(\hat{\boldsymbol{\zeta}})'\mathbf{1}\right] - n}{n}, \tag{3.2.44}$$

où nous utilisons $n = \mathbf{1}'\mathbf{1}$. Ici, deux fois la quantité au numérateur correspond à la déviance de la distribution gamma. L'approximation de la marginale basée sur le critère de Schwarz utilise les estimateurs $\widehat{\boldsymbol{\beta}}$ et $\hat{\phi}$.

3.3. Estimation du maximum de la distribution *a posteriori*

Bien qu'il soit possible de définir des distributions *a priori* conjugées pour les paramètres canoniques de la famille de distributions $\mathcal{F}$ (voir par exemple Albert, 1988 ; West et Harrison, 1998), nous n'adoptons pas cette stratégie dans ce qui suit. Notre but étant de travailler avec le vecteur de paramètres $\boldsymbol{\beta}$, il ne nous semble pas utile d'ajouter un niveau de modélisation probabiliste. Nous procédons donc d'une façon analogue à ce que nous avons fait pour le modèle linéaire de l'article 1 en ce qui concerne les distributions *a priori*.

À la section précédente, nous avons vu que la forme de la fonction $\kappa(y, \phi)$ pour la distribution $(\mathcal{G})$ diffère de celle des autres distributions étudiées. Nous avons donc cherché à obtenir une approximation de la distribution gamma afin d'avoir une modélisation uniforme pour ϕ dans le modèle bayésien. Il s'avère qu'une approximation par point de selle de la distribution $\mathcal{G}$ peut être employée à cette fin (voir par exemple Jorgensen, 1997). L'approximation par point de selle de la distribution gamma est équivalente à remplacer la fonction $\Gamma(\phi)$ par l'approximation de Stirling. En considérant le résultat suivant (Abramowitz and Stegun, 1964)

$$\Gamma(\phi) = (2\pi)^{1/2}\phi^{\phi-1/2}\exp(-\phi)\left\{1 + O(\phi^{-1})\right\}, \tag{3.3.1}$$

l'approximation de Stirling est obtenue lorsque le terme $O(\phi^{-1})$ est considéré négligeable, c'est-à-dire lorsque le paramètre de précision, ϕ, est grand. En réécrivant la fonction $\kappa(y, \phi)$, telle que donnée au tableau 3.1, et en utilisant ensuite l'approximation de Stirling, nous calculons

$$\kappa(y, \phi) = \phi \log(y) + \phi \log(\phi) - \log\{\Gamma(\phi)\} - \log(y), \tag{3.3.2}$$

$$\approx \phi\,[\log(y) + 1] + \frac{1}{2}\log(\phi) - \log(y) - \frac{1}{2}\log(2\pi). \tag{3.3.3}$$

Avec cette dernière forme, les fonctions $\kappa(y, \phi)$ des distributions données au tableau 3.1 peuvent toutes être écrites sous la forme suivante

$$\kappa(y, \phi) \cong -\phi s(y) + \frac{1}{2}\log(\phi) - \frac{1}{2}\log(2\pi) + t(y), \tag{3.3.4}$$

où le symbole $\cong$ est utilisé pour indiquer que la fonction est exacte pour toutes les distributions sauf pour la distribution gamma. Le tableau 3.2 donne les fonctions $s(y)$ et $t(y)$ pour chacune des distributions étudiées. Cette décomposition est importante pour deux raisons. D'une part elle est utilisée afin de traiter le paramètre de précision d'une façon uniforme dans le modèle bayésien. D'autre part elle est au centre de la modélisation simultanée de la courbe moyenne et de la courbe de dispersion du troisième article (voir chapitres 5 et 6).

TABLE 3.2. La décomposition de $\kappa(y, \phi)$ donnée à l'équation (3.3.4) pour les différentes distributions étudiées.

	$\mathcal{N}(\mu, \phi)$	$\mathcal{G}(\mu, \phi)$	IG (μ, ϕ)	$\mathcal{LN}(\mu, \phi)$	RIG (μ, ϕ)
$s(y)$	$\frac{1}{2}y^2$	$-\log(y) - 1$	$\frac{1}{2}y^{-1}$	$\frac{1}{2}[\log(y)]^2$	$\frac{1}{2}y$
$t(y)$	0	$-\log(y)$	$-\frac{1}{2}\log(y^3)$	$-\log(y)$	$-\frac{1}{2}\log(y)$

Avant de spécifier les distributions *a priori* pour $\boldsymbol{\beta}$ et ϕ, nous notons qu'en utilisant la décomposition de $\kappa(y, \phi)$ donnée à l'équation (3.3.4) le logarithme de la vraisemblance s'écrit maintenant

$$\mathcal{L} = \phi\,[\boldsymbol{\zeta}'\boldsymbol{z} - \boldsymbol{\psi}(\boldsymbol{\zeta})'\mathbf{1}] + \boldsymbol{\kappa}(\boldsymbol{y}, \phi)'\mathbf{1}, \tag{3.3.5}$$

$$\cong \phi\,[\boldsymbol{\zeta}'\boldsymbol{z} - \boldsymbol{\psi}(\boldsymbol{\zeta})'\mathbf{1}] - \phi\boldsymbol{s}(\boldsymbol{y})'\mathbf{1} - \frac{1}{2}\log(\phi)\mathbf{1}'\mathbf{1} + \boldsymbol{t}(\boldsymbol{y})'\mathbf{1} - \frac{1}{2}\log(2\pi)\mathbf{1}'\mathbf{1}, \tag{3.3.6}$$

$$= -\phi\,\{\boldsymbol{s}(\boldsymbol{y})'\mathbf{1} - [\boldsymbol{\zeta}'\boldsymbol{z} - \boldsymbol{\psi}(\boldsymbol{\zeta})'\mathbf{1}]\} - \frac{n}{2}\log(\phi) + \boldsymbol{t}(\boldsymbol{y})'\mathbf{1} - \frac{n}{2}\log(2\pi), \tag{3.3.7}$$

où le symbole $\cong$ indique que l'expression est exacte sauf pour la distribution gamma.

Avec le modèle probabiliste donné ci-haut, nous considérons les distributions *a priori* suivantes

$$\boldsymbol{\beta}|\phi \sim \mathcal{N}_{K_{\boldsymbol{\beta}}}\left(\boldsymbol{\beta}^0, \phi^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \tag{3.3.8}$$

$$\phi \sim \mathcal{G}_*\left(\alpha, \gamma\right), \tag{3.3.9}$$

où la notation $\mathcal{G}_*$ désigne la distribution gamma avec un paramètre de forme α et un paramètre d'échelle γ. La distribution *a priori* pour les coefficients de la composante systématique est définie conditionnellement au paramètre de précision, d'une façon analogue au traitement bayésien du modèle linéaire étudié dans l'article 1. La distribution gamma pour le paramètre de précision est choisie à cause de la forme du logarithme de la vraisemblance donné à l'équation (3.3.7). Les hyperparamètres sont considérés fixes dans ce qui suit, mais en pratique nous les spécifions à partir d'information historique (voir chapitre 4).

Avec ces distributions *a priori*, la distribution conjointe des observations et des paramètres est donnée par

$$D(\boldsymbol{y}, \boldsymbol{\beta}, \phi) = f(\boldsymbol{y}|\boldsymbol{\zeta}, \phi)\pi(\boldsymbol{\beta}|\boldsymbol{\beta}^0, \phi)\pi(\phi|\alpha, \gamma), \tag{3.3.10}$$

$$= \exp\left\{\mathcal{L} + \varpi_{\boldsymbol{\beta}} + \varpi_{\phi}\right\}, \tag{3.3.11}$$

où $\mathcal{L}$ correspond au logarithme de la vraisemblance, $\varpi_{\boldsymbol{\beta}}$ et ϖ_{ϕ} correspondent respectivement aux logarithmes de la distribution *a priori* $\pi(\boldsymbol{\beta}|\boldsymbol{\beta}^0, \phi)$ et de la distribution *a priori* $\pi(\phi|\alpha, \gamma)$. Explicitement, nous avons

$$\varpi_{\boldsymbol{\beta}} = -\frac{\phi}{2}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^0\right)' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^0\right) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right| - \frac{K_{\boldsymbol{\beta}}}{2}\log\left(2\pi\right), \tag{3.3.12}$$

$$\varpi_{\phi} = (\alpha - 1)\log(\phi) - \gamma\phi + \alpha\log(\gamma) - \log\Gamma(\alpha). \tag{3.3.13}$$

Il nous est possible d'obtenir le maximum de la distribution *a posteriori* (MAP) de $\boldsymbol{\beta}$ à partir de la distribution conjointe $D(\boldsymbol{y}, \boldsymbol{\beta}, \phi)$ étant donné que maximiser la distribution conjointe en $\boldsymbol{\beta}$ donne le MAP. Le MAP de $\boldsymbol{\beta}$ est donné par

$$\boldsymbol{\beta}^* = \text{argmax}_{\boldsymbol{\beta}}\left\{\mathcal{L} + \varpi_{\boldsymbol{\beta}}\right\}, \tag{3.3.14}$$

et nous cherchons donc la solution de $\nabla_{\boldsymbol{\beta}}\left[\mathcal{L}+\varpi_{\boldsymbol{\beta}}\right]=\mathbf{0}$

Nous avons calculé $\nabla_{\boldsymbol{\beta}}\mathcal{L}$ à la section précédente, alors afin de calculer le système d'équations correspondant au MAP, il nous suffit maintenant de calculer $\nabla_{\boldsymbol{\beta}}\varpi_{\boldsymbol{\beta}}$. Nous calculons

$$\nabla_{\boldsymbol{\beta}}\varpi_{\boldsymbol{\beta}}=-\phi\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{0}\right). \tag{3.3.15}$$

Le MAP, $\boldsymbol{\beta}^{*}$, est par conséquent la solution du système d'équations suivant

$$\nabla_{\boldsymbol{\beta}}\left[\mathcal{L}+\varpi_{\boldsymbol{\beta}}\right]=\mathbf{0}, \tag{3.3.16}$$

$$\boldsymbol{X}'\boldsymbol{W}\left(\boldsymbol{z}-\boldsymbol{\mu}\right)-\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{0}\right)=\mathbf{0}. \tag{3.3.17}$$

Nous notons que ce système d'équations ne dépend pas du paramètre ϕ à cause de la structure de covariance de $\boldsymbol{\beta}$ spécifiée à l'équation (3.3.8). Puisque nous utilisons l'algorithme de Newton-Raphson afin de déterminer le MAP, nous devons aussi calculer $\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\beta}}\left[\mathcal{L}+\varpi_{\boldsymbol{\beta}}\right]$. Le premier de ces termes fut calculé à la section précédente, alors nous devons seulement calculer le deuxième terme. Nous avons

$$\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\beta}}\varpi_{\boldsymbol{\beta}}=-\phi\nabla_{\boldsymbol{\beta}}\left[\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{0}\right)'\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right], \tag{3.3.18}$$

$$=-\phi\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}. \tag{3.3.19}$$

Pour calculer le MAP, $\boldsymbol{\beta}^{*}$, l'algorithme de Newton-Raphson est alors donné par

$$\boldsymbol{\beta}^{(m+1)}=\boldsymbol{\beta}^{(m)}-\left\{\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\beta}}\left[\mathcal{L}+\varpi_{\boldsymbol{\beta}}\right]\right\}^{-1}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}\left\{\nabla_{\boldsymbol{\beta}}\left[\mathcal{L}+\varpi_{\boldsymbol{\beta}}\right]\right\}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \tag{3.3.20}$$

$$=\boldsymbol{\beta}^{(m)}-\left\{\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}-\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right\}^{-1}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}} \tag{3.3.21}$$

$$\times\left\{\boldsymbol{X}'\boldsymbol{W}\left(\boldsymbol{z}-\boldsymbol{\mu}\right)-\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{0}\right)\right\}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}, \tag{3.3.22}$$

où $\boldsymbol{\beta}^{(m)}$ représente la valeur calculée à l'itération m.

Une fois la solution $\boldsymbol{\beta}^{*}$ obtenue, nous utilisons l'approximation de Laplace pour l'intégrale par rapport à $\boldsymbol{\beta}$. Cette aproximation est équivalente à supposer que la distribution *a posteriori* du vecteur $\boldsymbol{\beta}$ est une loi normale multivariée. Ainsi, avec la distribution *a priori* donnée à l'équation (3.3.8), l'approximation de Laplace donne un résultat similaire à celui du modèle linéaire traité dans l'article 1 (voir chapitre 2). Finalement, avec la forme du logarithme de la vraisemblance

de l'équation (3.3.7) et la distribution *a priori* pour le paramètre de précision de l'équation (3.3.9), il est possible d'intégrer le paramètre de précision pour obtenir la marginale complète.

Annexe A. Opérateurs différentiels

Pour ce qui suit, nous utilisons les vecteurs suivants :

$$\boldsymbol{x} = (x_1, \ldots, x_p)', \tag{A.1}$$

$$\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) = (y_1(\boldsymbol{x}), \ldots, y_q(\boldsymbol{x}))', \tag{A.2}$$

$$\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y}) = (z_1(\boldsymbol{y}), \ldots, z_r(\boldsymbol{y}))', \tag{A.3}$$

et la matrice $\boldsymbol{A}$ est une matrice indépendante des vecteurs précédents. Nous employons aussi un scalaire, a, et une matrice $\boldsymbol{B}$, qui dépend de a.

La dérivée par rapport au vecteur $\boldsymbol{x}$ est définie par

$$\nabla_{\boldsymbol{x}} = \frac{\partial}{\partial \boldsymbol{x}} = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_p} \right)'. \tag{A.4}$$

La différentiation de $\boldsymbol{y}'$ par rapport à $\boldsymbol{x}$ est donnée par

$$\nabla_{\boldsymbol{x}} (\boldsymbol{y}') = (\nabla_{\boldsymbol{x}} y_1(\boldsymbol{x}), \ldots, \nabla_{\boldsymbol{x}} y_q(\boldsymbol{x})), \tag{A.5}$$

une matrice $p \times q$. Pour le cas particulier $\boldsymbol{y} = \boldsymbol{x}$, nous avons alors

$$\nabla_{\boldsymbol{x}} (\boldsymbol{x}') = \mathbb{I}_p, \tag{A.6}$$

où $\mathbb{I}_p$ est la matrice identité de dimension $p \times p$. La différentiation du produit $\boldsymbol{y}'\boldsymbol{A}$ par rapport à $\boldsymbol{x}$ est

$$\nabla_{\boldsymbol{x}} (\boldsymbol{y}'\boldsymbol{A}) = \nabla_{\boldsymbol{x}} (\boldsymbol{y}') \boldsymbol{A}. \tag{A.7}$$

La différentiation en chaîne est donnée par

$$\nabla_{\boldsymbol{x}} (\boldsymbol{z}') = \nabla_{\boldsymbol{x}} (\boldsymbol{y}') \nabla_{\boldsymbol{y}} (\boldsymbol{z}'). \tag{A.8}$$

La dérivée d'une forme quadratique est donnée par

$$\nabla_{\boldsymbol{x}} (\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}'\boldsymbol{x}, \tag{A.9}$$

et dans le cas particulier où $\boldsymbol{A}$ est une matrice symétrique, nous obtenons

$$\nabla_{\boldsymbol{x}} (\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}. \tag{A.10}$$

En ce qui a trait à la différentiation d'une matrice par un scalaire, nous avons pour l'inverse

$$\nabla_a \left(\boldsymbol{B}^{-1}\right) = -\boldsymbol{B}^{-1} \nabla_a \left(\boldsymbol{B}\right) \boldsymbol{B}^{-1}. \tag{A.11}$$

Enfin, pour le logarithme du déterminant d'une matrice, nous avons

$$\nabla_a \left(\log |\boldsymbol{B}|\right) = tr \left\{\boldsymbol{B}^{-1} \nabla_a \left(\boldsymbol{B}\right)\right\}. \tag{A.12}$$

Chapitre 4

MODELLING THE AVERAGE BEHAVIOUR OF A SAMPLE OF CURVES WITH BAYESIAN REGRESSION SPLINES

Ce chapitre présente le deuxième article rédigé dans le cadre de cette thèse. L'article fut soumis à la Revue canadienne de statistique au mois de juin 2009.

Abstract

This paper is concerned with modelling longitudinal data nonparametrically in a Bayesian context. We develop a methodology, based on free-knot regression splines, in which the observations arise from a class of distributions that includes the continuous distributions of the exponential family. Approximations of the marginal distribution of the data are considered since this distribution plays a key role in the MCMC algorithm used to explore the space of knot configurations. A strategy is proposed to discriminate which statistical distribution is the most adequate for a given data set. The methodology is applied to a hydrological sample of curves.

4.1. Introduction

In many scientific disciplines, researchers are faced with the task of modelling longitudinal data which consist of repeated measurements of a response variable on an experimental unit. This situation often occurs in life sciences and environmental sciences where measurements are made through time. For long sequences

of longitudinal data, such as atmospheric variable measurements which can span several years, it is possible to use time series methods that take into account the seasonality of the series or functional data analysis approaches which consider each yearly series as a function. This latter methodology, set in a Bayesian framework, is the one we adopt in this paper by using a nonparametric method to capture the shape of yearly series. Our research motivation stems from a hydrological problem for which we seek to capture the average behaviour of weekly water flows and although the statistical model is constructed for this purpose, the methodology can easily be adapted to model a single curve. The functional representation of a sequence of water flows is called a hydrograph, and Figure 4.1 shows a sample of yearly hydrographs that have been landmark registered in order to make them similar relative to the time at which salient features happen (see Ramsay and Silverman (2005) on landmark registration and section 4.3.1 below). This data set represents a typical example of a sample of curves for which we want to model the average nonparametrically and our approach will be studied on this sample in the application section.

Nonparametric methods to model data have evolved significantly during the last few decades with the advent of several powerful automated techniques. Smoothing splines (Hastie and Tibshirani, 1990 ; Wahba, 1990 ; Green and Silverman, 1994), automated free-knot regression splines (Smith and Khon, 1996 ; Denison *et al.*, 1998), penalized regression splines (Eilers and Marx, 1996 ; Ruppert *et al.*, 2003) and wavelets (Donoho and Johnstone, 1994 ; Ogden, 1997), all constitute examples of these automatic methods. Some of these techniques have only been studied when the observations are considered to be normally distributed, while others have been extended to distributions belonging to the exponential family. For example, smoothing splines, which consist of placing knots at each data point and penalizing for the curvature of a fit through a smoothing parameter, have been extended to exponential families by the penalized log likelihood approach (chapter 6 of Hastie and Tibshirani (1990), and chapter 5 of Green and Silverman (1994)). The MCMC driven methods for automated free-knot regression splines mentioned above concentrated on the normal distribution in a Bayesian context.

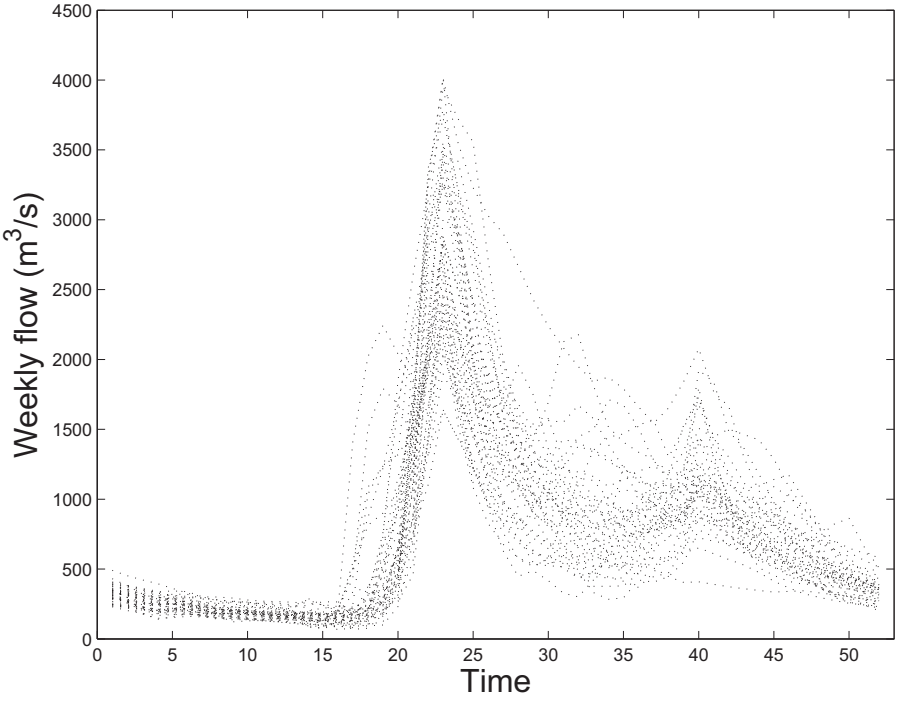


FIGURE 4.1. A sample of 42 landmark registered yearly hydrographs with weekly measurements from a watershed situated in northern Québec.

The methodology of Denison *et al.* (1998) was put in a more formal setting and improved in DiMatteo *et al.* (2001), who also extended the approach to nonparametric Poisson regression.

In this paper, we use a Bayesian probabilistic model based on free-knot regression splines to capture the average behaviour of a sample of curves when no auxiliary information is available. Free-knot regression splines are used since they represent one of the most efficient nonparametric modelling tools with regards to parsimony, a desired property in our modelling context. The difficulty of finding adequate knot configurations, which define the basis that models the data, is addressed through a reversible jump MCMC algorithm that explores the posterior distribution of the knot configurations in a similar fashion to the approaches of Denison *et al.* (1998) and DiMatteo *et al.* (2001). We develop a fairly general methodology in which the observations are thought to arise from a class of distributions that includes the continuous distributions of the exponential family ; more specifically, we consider continuous distributions for which the mean and

the variance possess a relation of the following type : variance $\propto \{\text{mean}\}^p$, where $p = 0, 1, 2, 3$. Since the marginal distribution of the data plays a key role in the evolution of the reversible jump MCMC and since it is not explicit in general, some approximations are studied. In doing so, we extend some of the results of Raftery (1996) in which the author considered marginal distribution approximations for generalized linear model selection purposes. Finally, we propose a method to determine the statistical distribution, from the aforementioned class, which is best suited to model a given data set. The Bayesian probabilistic model and its ramifications are exposed in the next section, while the third section presents an application of the proposed methodology to hydrographs.

4.2. Statistical model

Our methodology is closely related to that of generalized linear models (McCullagh and Nelder, 1989 ; McCulloch and Searle, 2001) and we therefore adopt the terminology associated with this class of models. The next two subsections cast our statistical problem in this framework by describing the observational random component, the systematic component and the link function. The third subsection gives the prior distributions of the parameters, while the fourth discusses sampling from the posterior distribution of knot configurations and model selection. Finally, estimation issues and the construction of approximate credible sets are addressed in the last subsection. For alternative treatments of generalized linear models in a Bayesian setting, the interested reader can refer to Albert (1988), Gelman *et al.* (1995), and West and Harrison (1998).

4.2.1. Random component : statistical distribution of the observations

We assume that we have a sample of N curves at our disposal and that each of these share a common period of time over which the measurements are taken. For a given curve i $(i = 1, \ldots, N)$, we have the following data

$$(x_1, y_{i1}), \ldots, (x_j, y_{ij}), \ldots, (x_n, y_{in}),$$

where x_j represents the time associated with the response variable y_{ij}. The x_j's are taken to be fixed and we consider the y_{ij}'s as random variables to be modelled.

We make the hypothesis that the observations of a given curve i are conditionally independent and that the data, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in})'$, has a joint distribution which belongs to the following family of distributions

$$f(\boldsymbol{y}_i|\boldsymbol{\zeta}_i, \boldsymbol{\phi}_i) = \prod_{j=1}^{n} \exp\left\{\phi_{ij}\left[\zeta_{ij}z_{ij} - \psi(\zeta_{ij})\right] + \kappa(y_{ij}, \phi_{ij})\right\}, \tag{4.2.1}$$

where z_{ij} is the value of a one-to-one function, $z(\cdot)$, evaluated at y_{ij}, $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{ij}, \ldots, \zeta_{in})'$ represents the vector of canonical parameters, the vector of precision parameters is $\boldsymbol{\phi}_i = (\phi_{i1}, \ldots, \phi_{ij}, \ldots, \phi_{in})'$; the functions ψ, κ, and z define the different statistical distributions. When $z(\cdot)$ is the identity function, this family of distributions corresponds to the exponential family and although we are not interested in studying a wide variety of data transformation functions, this framework is useful for what follows. Considering the individual curves to be independent, the joint distribution of all the curves, $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_i', \ldots, \boldsymbol{y}_N')'$, is given by

$$f(\boldsymbol{y}|\boldsymbol{\zeta}, \boldsymbol{\phi}) = \prod_{i=1}^{N} f(\boldsymbol{y}_i|\boldsymbol{\zeta}_i, \boldsymbol{\phi}_i) = \prod_{i=1}^{N}\prod_{j=1}^{n} \exp\left\{\phi_{ij}\eta_{ij} + \kappa_{ij}\right\}, \tag{4.2.2}$$

where we define $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1', \ldots, \boldsymbol{\zeta}_i', \ldots, \boldsymbol{\zeta}_N')'$, $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \ldots, \boldsymbol{\phi}_i', \ldots, \boldsymbol{\phi}_N')'$, and to simplify the notation, $\eta_{ij} = \eta(y_{ij}, \zeta_{ij}) = [\zeta_{ij}z_{ij} - \psi(\zeta_{ij})]$ and $\kappa_{ij} = \kappa(y_{ij}, \phi_{ij})$.

As mentioned above, we want to model the average process of a sample of curves and we thus take the canonical parameter at each measurement time to be the same for all the experimental units; furthermore, we consider a global precision parameter. More specifically, we set : 1. $\zeta_{ij} = \zeta_j$, $\forall i$, and 2. $\phi_{ij} = \phi$, $\forall i, j$. These two working assumptions lead to the following joint distribution

$$f(\boldsymbol{y}|\boldsymbol{\zeta}, \phi) = \prod_{j=1}^{n} \exp\left\{N\phi\bar{\eta}_{\cdot j} + \sum_{i=1}^{N}\kappa_{ij}\right\} = \exp\left\{N\phi\sum_{j=1}^{n}\bar{\eta}_{\cdot j} + \kappa_{\cdot\cdot}\right\}, \tag{4.2.3}$$

where we now have $\bar{\eta}_{\cdot j} = N^{-1}\sum_{i=1}^{N}\eta_{ij} = [\zeta_j\bar{z}_{\cdot j} - \psi(\zeta_j)]$, $\bar{z}_{\cdot j} = N^{-1}\sum_{i=1}^{N} z_{ij}$, $\kappa_{ij} = \kappa(y_{ij}, \phi)$, and $\kappa_{\cdot\cdot} = \sum_{i,j}\kappa_{ij}$; for what follows, we write $\bar{\boldsymbol{z}} = (\bar{z}_{\cdot 1}, \ldots, \bar{z}_{\cdot j}, \ldots, \bar{z}_{\cdot n})'$.

The data we wish to model are continuous and we therefore consider the following distributions : normal ($\mathcal{N}$), gamma ($\mathcal{G}$), inverse gaussian (IG), lognormal

($\mathcal{LN}$) and reciprocal inverse gaussian (RIG). Even if our data is strictly positive and possibly skewed, we still consider the normal distribution because of its wide use in applications. As indicated in Appendix A, each distribution possesses a specific relation between the mean and the variance, namely a relation of the following type : variance $\propto \{\text{mean}\}^p$, where $p = 0, 1, 2, 3$. One of our goals is to propose a framework in which it is possible to discriminate between these different variance structures. Appendix A also supplies the function η relative to the mean μ for each distribution, as well as a decomposition of the individual functions κ, which is necessary to specify a prior distribution for the precision parameter ϕ.

4.2.2. Systematic component and link function

In the usual generalized linear model framework, one specifies a systematic component, *i.e.* a linear combination of auxiliary variables, which is related to the mean of the random variable through a link function. As mentioned previously, there is no auxiliary information in our present modelling context and we therefore rely on a nonparametric method to model the mean process. We choose to work with polynomial regression spline functions (Smith, 1979 ; Friedman and Silverman, 1989 ; Smith and Kohn, 1996 ; Denison *et al.*, 1998 ; DiMatteo *et al.*, 2001) as a basis. More specifically, we use M-spline functions which are closely related to B-spline functions and only differ in their normalizing constants. The reason for choosing M-spline functions is that by integrating these functions, we get the so called I-spline functions (Ramsay, 1988) which are useful for our operational purposes since there is also a practical interest in the cumulative function of the data we wish to model. With this choice, the systematic component can be written as

$$u_{j,\boldsymbol{\omega}} = \sum_{k=1}^{K_{\boldsymbol{\omega}}} \beta_{k,\boldsymbol{\omega}} b_{k,\boldsymbol{\omega}}(x_j) = \boldsymbol{b}_{\boldsymbol{\omega}}(x_j)' \boldsymbol{\beta}_{\boldsymbol{\omega}}, \tag{4.2.4}$$

where $\beta_{k,\boldsymbol{\omega}}$ represents the coefficient of the M-spline basis element $b_{k,\boldsymbol{\omega}}(\cdot)$, and $K_{\boldsymbol{\omega}}$ is the number of elements in the basis. In vector notation, we have $\boldsymbol{b}_{\boldsymbol{\omega}}(x_j) = (b_{1,\boldsymbol{\omega}}(x_j), \ldots, b_{K_{\boldsymbol{\omega}},\boldsymbol{\omega}}(x_j))'$, a $K_{\boldsymbol{\omega}} \times 1$ vector of the basis elements evaluated at x_j, and $\boldsymbol{\beta}_{\boldsymbol{\omega}} = (\beta_{1,\boldsymbol{\omega}}, \ldots, \beta_{K_{\boldsymbol{\omega}},\boldsymbol{\omega}})'$, a $K_{\boldsymbol{\omega}} \times 1$ vector of parameters. Once the order l of the spline polynomial functions is fixed, the basis elements $b_{k,\boldsymbol{\omega}}(\cdot)$ are determined

by the number of interior knots, m, and the ordered location of these knots, $\boldsymbol{r}^{(m)} = (r_1, ..., r_m)$; we summarize this information in the model parameter $\boldsymbol{\omega} = (m, \boldsymbol{r}^{(m)})$, and the number of elements in the basis is then given by $K_{\boldsymbol{\omega}} = l + m$. We treat $\boldsymbol{\omega}$ as a random parameter and its parameter space is explored through a reversible jump MCMC algorithm described in section 4.2.4 and Appendix B. The vector of systematic components associated with the vector of observed statistics, $\bar{\boldsymbol{z}}$, can be written as

$$\boldsymbol{u}_{\boldsymbol{\omega}} = \boldsymbol{B}_{\boldsymbol{\omega}}\boldsymbol{\beta}_{\boldsymbol{\omega}}, \tag{4.2.5}$$

where $\boldsymbol{u}_{\boldsymbol{\omega}} = (u_{1,\boldsymbol{\omega}}, \ldots, u_{n,\boldsymbol{\omega}})'$ is a $n \times 1$ vector and $\boldsymbol{B}_{\boldsymbol{\omega}} = (\boldsymbol{b}_{\boldsymbol{\omega}}(x_1), \ldots, \boldsymbol{b}_{\boldsymbol{\omega}}(x_n))'$, a $n \times K_{\boldsymbol{\omega}}$ matrix.

For the link function, one can use the canonical link of a given distribution since it possesses good statistical properties such as sufficient statistics which are linear in the data, etc. Here we decide to study several link functions and we therefore also consider this aspect of the problem as one of model selection. An important reason for doing so is that the choice of the link function directly influences the marginal covariance structure of the observations (chapter 8 of McCulloch and Searle, 2001). The following link functions are considered : the identity link (IDL), the logarithmic link (LOL) and the inverse link (INL) ; more explicitely, we have

$$\text{IDL} \quad : \quad \mathbb{E}\left(\bar{\boldsymbol{z}}\right) = \boldsymbol{u}_{\boldsymbol{\omega}}, \tag{4.2.6}$$

$$\text{LOL} \quad : \quad \log\{\mathbb{E}\left(\bar{\boldsymbol{z}}\right)\} = \boldsymbol{u}_{\boldsymbol{\omega}}, \tag{4.2.7}$$

$$\text{INL} \quad : \quad \{\mathbb{E}\left(\bar{\boldsymbol{z}}\right)\}^{-1} = \boldsymbol{u}_{\boldsymbol{\omega}}, \tag{4.2.8}$$

where the logarithm and inverse functions are applied componentwise.

4.2.3. Prior distributions

We consider the coefficients of the spline functions, $\boldsymbol{\beta}_{\boldsymbol{\omega}}$, to arise from the following multivariate normal distribution

$$\boldsymbol{\beta}_{\boldsymbol{\omega}} | \phi_{\boldsymbol{\omega}} \sim \mathcal{N}_{K_{\boldsymbol{\omega}}}\left(\boldsymbol{\beta}^0_{\boldsymbol{\omega}}, \phi^{-1}_{\boldsymbol{\omega}}\boldsymbol{\Sigma}_{\boldsymbol{\omega}}\right), \tag{4.2.9}$$

where $\boldsymbol{\beta}_{\boldsymbol{\omega}}^{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ are taken to be fixed and are specified in the application section. As shown in Appendix A, a convenient prior distribution for the precision parameter is a gamma distribution

$$\phi_{\boldsymbol{\omega}} \sim \mathcal{G}_{*}\left(\alpha_{\boldsymbol{\omega}}, \gamma_{\boldsymbol{\omega}}\right), \tag{4.2.10}$$

where $\mathcal{G}_{*}$ refers to a gamma distribution with shape parameter $\alpha_{\boldsymbol{\omega}}$ and inverse scale parameter $\gamma_{\boldsymbol{\omega}}$. Since we consider the number of knots and their positions to be random quantities, we also need to specify a prior distribution for the model parameter $\boldsymbol{\omega}$. Following Denison *et al.* (1998), we consider the prior distribution for $\boldsymbol{\omega}$ to be given by

$$\pi(\boldsymbol{\omega}) = \pi(m, \boldsymbol{r}^{(m)}) = \pi_2(\boldsymbol{r}^{(m)}|m)\pi_1(m), \tag{4.2.11}$$

where $m \sim \mathcal{P}(\lambda) I_{\{0,1,\ldots,M\}}(m)$, a truncated Poisson distribution ; each component r_i of $\boldsymbol{r}^{(m)}$ is taken to be the order statistic i from a uniform discrete distribution on a support set containing M possible knot locations (see section 4.3.2.2 for more details). We therefore have

$$\pi_2(\boldsymbol{r}^{(m)}|m) = m! M^{-m}. \tag{4.2.12}$$

For a different prior specification for $\boldsymbol{\omega}$, see DiMatteo *et al.* (2001).

4.2.4. Knot configuration exploration and model selection

The difficulty in practice with regression splines is the determination of 'optimal' knot configurations, but as was initially shown by Denison *et al.* (1998), this problem can be alleviated by using the MCMC reversible jump algorithm developed by Green (1995). As mentioned in the introduction, some modifications were brought to the methodology in the paper of DiMatteo *et al.* (2001). The algorithm used here employs some of the aspects of Denison *et al.* (1998) and some from DiMatteo *et al.* (2001). The possible knot configurations are based on the former as indicated by the prior distribution for $\boldsymbol{\omega}$, while the transition probabilities are similar in spirit to the latter. As discussed in Appendix B, the transition probabilities depend on the partial marginal distribution $\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega})$; the integration of the coefficients of the spline functions and the precision parameter

is important for two reasons : the speed of convergence of the chain and the fact that it simplifies the algorithm with regards to issues related to switching dimensions. With the distributional assumptions outlined in section 4.2.1, the marginal distribution $\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega})$ does not possess an analytic form except in the normal case ; we therefore need to use an approximation and in what follows, we consider two such approximations.

4.2.4.1. *Marginal distribution approximations*

The first approximation we consider for the marginal distribution, which was used by DiMatteo *et al.* (2001) in the context of Poisson nonparametric regression, relies on the information criterion proposed by Schwarz (1978). It is given by

$$\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}) \approx \mathrm{m}_1(\boldsymbol{y}|\boldsymbol{\omega}) = \exp\left(S_{\boldsymbol{\omega}}\right), \tag{4.2.13}$$

where $S_{\boldsymbol{\omega}}$ is Schwarz' information criterion. With the present notation, we have

$$S_{\boldsymbol{\omega}} = N\widehat{\phi}_{\boldsymbol{\omega}} \sum_{j=1}^{n} \widehat{\bar{\eta}}_{\cdot j} + \widehat{\kappa}_{\cdot\cdot} - \frac{(K_{\boldsymbol{\omega}}+1)}{2}\log(nN), \tag{4.2.14}$$

where $\widehat{\bar{\eta}}_{\cdot j}$ and $\widehat{\kappa}_{\cdot\cdot}$ are evaluated for $\widehat{\phi}_{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}$, the maximum likelihood estimates. The criterion measures goodness of fit through the first two terms, the maximized log likelihood, and penalizes for model complexity via the last term, which is proportional to the number of evaluated parameters.

Kass and Wasserman (1995) showed that this quantity can be used effectively to calculate accurate ratios of marginal distributions, or Bayes factors (Kass and Raftery, 1995), for nested models under the assumptions that the 'regression' parameters follow a unit information prior (see section 4.3.2.2 below) and that the precision parameter is known. In our context, the precision parameter is unknown but we nonetheless use this expression directly to compare its performance to that of the second approximation since it is easy and convenient to use in practice, relying only on the maximized log likelihood.

The second approximation is based on the basic form of Laplace's approximation to an integral. Tierney and Kadane (1986), Tierney *et al.* (1989), Shun and McCullagh (1995), all discuss this basic form and higher-order approximations.

We use it to approximate the integral over the parameter space of the coefficients associated with the spline functions. After performing this operation, the precision parameter is integrated (Appendix C provides details). With the prior distributions specified in the previous section, the second approximation to the marginal distribution is given by

$$\begin{aligned} \mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}) \approx \mathrm{m}_2(\boldsymbol{y}|\boldsymbol{\omega}) = & \left(\frac{|\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}}|}{|\boldsymbol{\Sigma}_{\boldsymbol{\omega}}|}\right)^{1/2} \\ & \times \left(\frac{\Gamma(\alpha^*_{\boldsymbol{\omega}})}{(2\pi)^{Nn/2}\Gamma(\alpha_{\boldsymbol{\omega}})}\right)\left(\frac{(\gamma_{\boldsymbol{\omega}})^{\alpha_{\boldsymbol{\omega}}}}{(\gamma^*_{\boldsymbol{\omega}})^{\alpha^*_{\boldsymbol{\omega}}}}\right)\exp(t_{\cdot\cdot}), \end{aligned} \tag{4.2.15}$$

where

$$\alpha^*_{\boldsymbol{\omega}} = Nn/2 + \alpha_{\boldsymbol{\omega}}, \tag{4.2.16}$$

$$\gamma^*_{\boldsymbol{\omega}} = \frac{N}{2}\sum_{j=1}^{n}\bar{d}^*_{\cdot j} + \frac{1}{2}(\boldsymbol{\beta}^*_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}})'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\omega}}(\boldsymbol{\beta}^*_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}}) + \gamma_{\boldsymbol{\omega}}, \tag{4.2.17}$$

$$\bar{d}^*_{\cdot j} = 2[\bar{s}_{\cdot j} - \bar{\eta}^*_{\cdot j}], \tag{4.2.18}$$

$$\boldsymbol{\beta}^*_{\boldsymbol{\omega}} = \mathrm{argmax}_{\boldsymbol{\beta}_{\boldsymbol{\omega}}}\left\{N\sum_{j=1}^{n}\bar{\eta}_{\cdot j} - \frac{1}{2}(\boldsymbol{\beta}_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}})'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\omega}}(\boldsymbol{\beta}_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}})\right\}, \tag{4.2.19}$$

$$\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}} = \left(\boldsymbol{\Lambda}^*_{\boldsymbol{\omega}} + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\omega}}\right)^{-1}, \tag{4.2.20}$$

$$\boldsymbol{\Lambda}^*_{\boldsymbol{\omega}} = -N\left\{\frac{\partial^2}{\partial\boldsymbol{\beta}_{\boldsymbol{\omega}}\partial\boldsymbol{\beta}'_{\boldsymbol{\omega}}}\sum_{j=1}^{n}\bar{\eta}_{\cdot j}\right\}_{\boldsymbol{\beta}^*_{\boldsymbol{\omega}}} = N\boldsymbol{B}'_{\boldsymbol{\omega}}\boldsymbol{\Delta}^*_{\boldsymbol{\omega}}\boldsymbol{B}_{\boldsymbol{\omega}}, \tag{4.2.21}$$

and $\boldsymbol{\Delta}^*_{\boldsymbol{\omega}}$ is a diagonal matrix, evaluated at $\boldsymbol{\beta}^*_{\boldsymbol{\omega}}$. This matrix, which possesses an analytic form, depends on the data, the statistical distribution and the link function. The quantities $t_{\cdot\cdot}$, $\bar{s}_{\cdot j}$ and $\bar{d}_{\cdot j}$ depend on the different statistical distributions, and it is interesting to note that this last quantity is an average of individual data contributions to the deviance (see Appendix A for the details).

This second approximation is reminiscent of the marginal distribution encountered in the Bayesian treatment of the linear model (see Lempers (1971), and Pericchi (1984) for similar expressions). For the usual normal linear model, we have $t_{\cdot\cdot} = 0$ and $\boldsymbol{\Delta}^*_{\boldsymbol{\omega}} = \mathbb{I}_n$. Therefore, the main differences between this more general expression and the one for the linear model lie in the exponential term and the presence of the diagonal matrix $\boldsymbol{\Delta}^*_{\boldsymbol{\omega}}$ in the determinant. For generalized linear

models with known precision parameter, Raftery (1996) studied approximations of the marginal distribution based on the basic form of Laplace's approximation. Equation (4.2.15) generalizes some of his results since the precision parameter here is not assumed known and furthermore, we perform a 'complete' maximisation of equation (4.2.19), while Raftery (1996) only considered a single step of the Newton-Raphson algorithm. This latter choice was made in order to directly use standard outputs of generalized linear models software.

Although expression (4.2.15) does not seem similar to the one obtained from the Schwarz criterion, some similarities are drawn in Appendix C. In the application section, we compare the behaviours of $\mathrm{m}_1(\boldsymbol{y}|\boldsymbol{\omega})$, based on $S_{\boldsymbol{\omega}}$, and $\mathrm{m}_2(\boldsymbol{y}|\boldsymbol{\omega})$, based on Laplace's approximation.

4.2.4.2. *Knot configuration selection*

The MCMC reversible jump algorithm, detailed in Appendix B, explores the space of the model parameter, $\boldsymbol{\omega}$, and reaches a stationary distribution once convergence is attained. The mode $\boldsymbol{\omega}^{\dagger}$ of the converged chain is used as an 'optimal' knot configuration and this constitutes our selected knot configuration. Although it would be possible to average the output of the MCMC algorithm in a fashion similar to that of DiMatteo *et al.* (2001), we choose to select a knot configuration for operational purposes, *i.e.* to have a model in the parameter space of the spline coefficients. In practice, the mode of a given chain is determined by first finding the mode of the posterior distribution of m, the number of knots, which we write as $m^{\dagger}$; once this mode is established, the mode of the knot configurations containing $m^{\dagger}$ knots is determined, and these two results give $\boldsymbol{\omega}^{\dagger}$.

4.2.4.3. *Selection of an adequate statistical distribution and link function*

In order to compare the set of distributions and link functions given in sections 4.2.1 and 4.2.2, we propose to use ratios of marginal distributions evaluated at the modes $\boldsymbol{\omega}^{\dagger}$ of the converged chains. This is similar in spirit to Bayes factors (Jeffreys, 1961 ; Kass and Raftery, 1995) which compare two models through the ratio of their marginal distributions, and which can be interpreted as the weight of evidence in favour of a model relative to another. Strictly speaking, in our

modelling context, we would need to integrate over the space of $\boldsymbol{\omega}$ to get the full marginal distribution and then calculate the Bayes factors. We nonetheless think that the modal evaluation of the marginal distribution supplies reasonable information concerning the adequacy of a given model.

To indicate that the ratios are evaluated at the modal quantities of $\boldsymbol{\omega}$, we write

$$BF^{\dagger}(A,B) = \frac{\mathrm{m}^{(A)}(\boldsymbol{y}|\boldsymbol{\omega}^{\dagger})}{\mathrm{m}^{(B)}(\boldsymbol{y}|\boldsymbol{\omega}^{\dagger})}, \tag{4.2.22}$$

where $\mathrm{m}^{(A)}(\boldsymbol{y}|\boldsymbol{\omega}^{\dagger})$ and $\mathrm{m}^{(B)}(\boldsymbol{y}|\boldsymbol{\omega}^{\dagger})$ are the marginal distributions of the two models A and B, evaluated at their respective modes ; a statistical distribution and a link function determine each of these models. In the application section, we study these ratios with the two marginal distribution approximations given above.

4.2.5. Function estimation and approximate credible sets

Our Bayesian model yields approximate posterior distributions for the parameters of the spline functions, $\boldsymbol{\beta}_{\boldsymbol{\omega}}$, and for the precision parameter, $\phi_{\boldsymbol{\omega}}$. More specifically, the Laplace approximation is equivalent to assume that $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ is normally distributed as follows

$$\boldsymbol{\beta}_{\boldsymbol{\omega}}|\phi_{\boldsymbol{\omega}}, \boldsymbol{y} \sim \mathcal{N}_{K_{\omega}}\left(\boldsymbol{\beta}^{*}_{\boldsymbol{\omega}}, \phi^{-1}_{\boldsymbol{\omega}}\boldsymbol{\Sigma}^{*}_{\boldsymbol{\omega}}\right), \tag{4.2.23}$$

where $\boldsymbol{\beta}^{*}_{\boldsymbol{\omega}}$ and $\boldsymbol{\Sigma}^{*}_{\boldsymbol{\omega}}$ are respectively given in equations (4.2.19) and (4.2.20). Furthermore, with the prior distribution for the precision parameter given in equation (4.2.10), we have the following posterior distribution

$$\phi_{\boldsymbol{\omega}}|\boldsymbol{y} \sim \mathcal{G}_{*}\left(\alpha^{*}_{\boldsymbol{\omega}}, \gamma^{*}_{\boldsymbol{\omega}}\right), \tag{4.2.24}$$

where $\alpha^{*}_{\boldsymbol{\omega}}$ and $\gamma^{*}_{\boldsymbol{\omega}}$ are given in equations (4.2.16) and (4.2.17) respectively. Given this last posterior distribution, it is possible to integrate the precision parameter to get the posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ independent of $\phi_{\boldsymbol{\omega}}$:

$$\boldsymbol{\beta}_{\boldsymbol{\omega}}|\boldsymbol{y} \sim \mathcal{T}_{K_{\omega}}\left(2\alpha^{*}_{\boldsymbol{\omega}}, \boldsymbol{\beta}^{*}_{\boldsymbol{\omega}}, \left(\frac{\gamma^{*}_{\boldsymbol{\omega}}}{\alpha^{*}_{\boldsymbol{\omega}}}\right)\boldsymbol{\Sigma}^{*}_{\boldsymbol{\omega}}\right). \tag{4.2.25}$$

We note that these posterior distributions parallel the ones obtained for the linear model (see for example Zellner, 1971 ; Robert, 1994).

4.2.5.1. *Function estimation*

In our modelling context, we model the vector $\bar{\boldsymbol{z}}$ and as shown in equation (4.2.4), the jth systematic component is given by $u_{j,\boldsymbol{\omega}} = \boldsymbol{b}_{\boldsymbol{\omega}}(x_j)'\boldsymbol{\beta}_{\boldsymbol{\omega}}$. Using $h(\cdot)$ to denote the inverse of the link function, we have

$$\mathbb{E}(\bar{z}_{.j}|\boldsymbol{\beta}_{\boldsymbol{\omega}}) = h_j, \tag{4.2.26}$$

where $h_j = h\left(u_{j,\boldsymbol{\omega}}\right) = h\left(\boldsymbol{b}_{\boldsymbol{\omega}}(x_j)'\boldsymbol{\beta}_{\boldsymbol{\omega}}\right)$. Except in the case of the identity link, the model for $\bar{\boldsymbol{z}}$ is not a linear function of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ and we use a Taylor series expansion of order 1 to approximate the inverse of the link function. This yields

$$h_j \approx h_j^* + \dot{\boldsymbol{h}}_j^{'}\left(\boldsymbol{\beta}_{\boldsymbol{\omega}} - \boldsymbol{\beta}_{\boldsymbol{\omega}}^*\right), \tag{4.2.27}$$

where $h_j^* = h\left(\boldsymbol{b}_{\boldsymbol{\omega}}(x_j)'\boldsymbol{\beta}_{\boldsymbol{\omega}}^*\right)$ and $\dot{\boldsymbol{h}}_j = \left\{\frac{\partial h_j}{\partial \boldsymbol{\beta}_{\boldsymbol{\omega}}}\right\}_{\boldsymbol{\beta}_{\boldsymbol{\omega}}^*}$. We then get $\mathbb{E}(\bar{z}_{.j}) \approx h_j^*$ by taking the expectation relative to the posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$.

For the untransformed data, *i.e.* the $\mathcal{N}$ (normal), $\mathcal{G}$ (gamma) and IG (inverse gaussian) distributions, we have $\bar{\boldsymbol{z}} = \bar{\boldsymbol{y}}$ and therefore $\mathbb{E}(\bar{y}_{.j}) = \mathbb{E}(\bar{z}_{.j}) \approx h_j^*$. For the $\mathcal{LN}$ (lognormal) and RIG (reciprocal inverse gaussian) distributions, the data are transformed by applying the logarithm and inverse functions respectively ; in these two cases, $\bar{\boldsymbol{z}}$ thus represents the average of the logarithm and inverse of the data. To get a representation for the untransformed data, we use the expressions for the mean given in Appendix A and a Taylor series expansion as above ; in this case, it is important to note that the mean on the untransformed scale depends on the precision parameter $\phi_{\boldsymbol{\omega}}$. Writing the function for the mean as $g(\cdot)$, we then have

$$\mathbb{E}(\bar{y}_{.j}|\boldsymbol{\beta}_{\boldsymbol{\omega}}, \phi_{\boldsymbol{\omega}}) = g_j, \tag{4.2.28}$$

where $g_j = g\left\{\mathbb{E}(\bar{z}_{.j}|\boldsymbol{\beta}_{\boldsymbol{\omega}}), \phi_{\boldsymbol{\omega}}\right\} = g\left\{h_j, \phi_{\boldsymbol{\omega}}\right\}$. The Taylor series expansion now gives

$$g_j \approx g_j^* + \dot{\boldsymbol{g}}_j'\left\{\left(\boldsymbol{\beta}_{\boldsymbol{\omega}} - \boldsymbol{\beta}_{\boldsymbol{\omega}}^*\right)', \left(\phi_{\boldsymbol{\omega}} - \phi_{\boldsymbol{\omega}}^*\right)\right\}', \tag{4.2.29}$$

where $g_j^* = g\left\{h_j^*, \phi_{\boldsymbol{\omega}}^*\right\}$, $\dot{\boldsymbol{g}}_j = \left\{\left(\frac{\partial g_j}{\partial h_j}\right)\left(\frac{\partial h_j}{\partial \boldsymbol{\beta}_{\boldsymbol{\omega}}}\right)', \frac{\partial g_j}{\partial \phi_{\boldsymbol{\omega}}}\right\}'_{\boldsymbol{\beta}_{\boldsymbol{\omega}}^*, \phi_{\boldsymbol{\omega}}^*}$, and $\phi_{\boldsymbol{\omega}}^* = \alpha_{\boldsymbol{\omega}}^*/\gamma_{\boldsymbol{\omega}}^*$, the expectation of $\phi_{\boldsymbol{\omega}}$ under the posterior distribution given in (4.2.24). Taking the expectation relative to the posterior distributions given in equations (4.2.23) and (4.2.24) yields : $\mathbb{E}(\bar{y}_{.j}) \approx g_j^*$.

4.2.5.2. *Approximate credible sets*

We are also interested in constructing credible sets for the sample of curves under study. Simultaneous credible sets can be obtained fairly directly for the case of the normal distribution and the identity link, but for other distributions and link functions one needs to construct approximate credible sets.

The approximate credible sets can be obtained from the Taylor series expansions of equations (4.2.27) and (4.2.29), and the posterior distributions of (4.2.23), (4.2.24) and (4.2.25). For $\mathbb{E}(\bar{z}_{.j})$, simultaneous $100(1-\delta)\%$ credible sets are constructed from equation (4.2.27) and the posterior distribution given in equation (4.2.25). They are given by

$$h_j^* \pm \left\{ K_{\boldsymbol{\omega}} \left(\frac{\gamma_{\boldsymbol{\omega}}^*}{\alpha_{\boldsymbol{\omega}}^*} \right) \dot{\boldsymbol{h}}_j' \boldsymbol{\Sigma}_{\boldsymbol{\omega}}^* \dot{\boldsymbol{h}}_j F_{K_{\boldsymbol{\omega}}, 2\alpha_{\boldsymbol{\omega}}^*}(\delta) \right\}^{1/2}, \tag{4.2.30}$$

where $F_{K_{\boldsymbol{\omega}}, 2\alpha_{\boldsymbol{\omega}}^*}(\delta)$ represents the $100(1-\delta)$th percentile of Fisher's F distribution with degrees of freedom $K_{\boldsymbol{\omega}}$ and $2\alpha_{\boldsymbol{\omega}}^*$. These approximate simultaneous credible sets can be used for $\mathbb{E}(\bar{y}_{.j})$ in the case of the $\mathcal{N}$, $\mathcal{G}$ and IG distributions.

Concerning the $\mathcal{LN}$ and RIG distributions, a further assumption needs to be made to obtain simultaneous approximate credible sets for $\mathbb{E}(\bar{y}_{.j})$. Since $\mathbb{E}(\bar{y}_{.j})$ is a function of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ and $\phi_{\boldsymbol{\omega}}$, we assume that

$$\boldsymbol{\theta}_{\boldsymbol{\omega}} = (\boldsymbol{\beta}_{\boldsymbol{\omega}}', \phi_{\boldsymbol{\omega}})' \sim \mathcal{N}_{K_{\boldsymbol{\omega}}+1} \left(\boldsymbol{\theta}_{\boldsymbol{\omega}}^*, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{\omega}}}^* \right), \tag{4.2.31}$$

where $\boldsymbol{\theta}_{\boldsymbol{\omega}}^* = ((\boldsymbol{\beta}_{\boldsymbol{\omega}}^*)', \phi_{\boldsymbol{\omega}}^*)'$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{\omega}}}^* = \text{Diag}\left\{ \boldsymbol{\Sigma}_{\boldsymbol{\omega}}^*, \frac{\alpha_{\boldsymbol{\omega}}^*}{(\gamma_{\boldsymbol{\omega}}^*)^2} \right\}$ with all the other quantities defined as above. This approximation is reasonable when $\alpha_{\boldsymbol{\omega}}^*$ (defined in equation (4.2.16)), of equations (4.2.24) and (4.2.25), is large, *i.e.* when the scale parameter of the gamma distribution and the degrees of freedom of the multivariate Student's t distribution are large. Finally, using equation (4.2.29), one gets the following simultaneous $100(1-\delta)\%$ credible sets

$$g_j^* \pm \left\{ \dot{\boldsymbol{g}}_j' \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{\omega}}}^* \dot{\boldsymbol{g}}_j \chi^2_{K_{\boldsymbol{\omega}}+1}(\delta) \right\}^{1/2}, \tag{4.2.32}$$

where $\chi^2_{K_{\boldsymbol{\omega}}+1}(\delta)$ represents the $100(1-\delta)$th percentile of a χ^2 distribution with degrees of freedom $K_{\boldsymbol{\omega}}+1$.

4.3. Application

In hydrology, a time series of considerable interest is the one made up of water flows. This data needs to be modelled effectively and accurately in order to make important decisions concerning water management, but also to prevent environmental and human casualties which could happen as a consequence of extreme events. Since the typical shape of a watershed's hydrographs is often of interest for various operational purposes, we now apply the statistical model developed in the previous section to capture the average behaviour of the sample of yearly hydrographs shown in Figure 4.1.

As pointed out in section 4.2.1, our modelling framework enables us to consider a variety of continuous distributions which possess different structures of dependence between the variance and the mean. Several distributions that we consider have been used in the past to model water flow data. The $\mathcal{G}$ and $\mathcal{LN}$ distributions are discussed in Gumbel (1958) and Markovic (1965), while the IG distribution was used for this purpose by Folks and Chhikara (1978) ; for a more recent general discussion, the reader can refer to Aksoy (2000). These distributions were studied in this context because the variable of interest is often skewed to the right. As mentioned in section 4.2.1, we also analyze the $\mathcal{N}$ distribution since it is often used in smoothing problems ; finally the RIG distribution is considered because its support is stricly positive and it can be seen as a middle ground between the $\mathcal{N}$ distribution, no dependence between the mean and the variance, and the $\mathcal{G}$ and $\mathcal{LN}$ distributions that possess a quadratic dependence (see Appendix A for details).

4.3.1. Hydrological data

A sample of 42 yearly hydrographs with weekly measurements from a watershed in northern Québec is shown in Figure 4.1 ; this is the sample for which we want to model the average nonparametrically. In total, we have a sample of 53 yearly hydrographs covering the period which extends from 1950 to 2002. The first 11 annual hydrographs are known to be of lesser quality since the data have been reconstructed from auxiliary information, and we use these to specify our prior

distributions (see section 4.3.2.2) ; our effective sample of curves thus contains the 42 yearly hydrographs shown in Figure 4.1. The first data point of each curve corresponds to the measurement made on the first week of January, while the last one to the measurement recorded on the last week of December. The first weeks of each hydrograph show a steady decline in water flow during the winter period, and the variability across hydrographs is weak. With the advent of warmer temperatures, spring floods begin and a strong increase in water flow is observed when the accumulated snow starts melting. A spring flood is characterized by a global maximum and can also possess secondary peaks from heavy rainfalls during this time of the year. In Figure 4.1, the global maxima of the spring flood for the different hydrographs happen simultaneously since the hydrographs have been registered. Although the spring floods all share a common global structure, it is clear that the variability of water flows is high for this period. After the spring flood is over, water flow is mainly governed by rainfall during summer and autumn. The autumn maxima happen simultaneously on the figure since these features were also used to perform the registration. The variability across hydrographs is also considerable during the summer and autumn weeks, due to the random nature of important rainfall events. Finally, at the end of the domain, water flow starts decreasing again at the onset of winter.

4.3.2. Model specifications

4.3.2.1. *Spline basis*

The order l of the spline functions which form the modelling basis needs to be specified in order to apply our method, and although we could treat this quantity as a parameter to be estimated in the procedure, we will consider it to be fixed as is usually done in practice. We choose to work with M-spline functions of order $l = 3$ which means that the basis elements are quadratic by parts.

4.3.2.2. *Prior distributions*

We first note that if one uses the marginal distribution approximation based on the Schwarz criterion, it is not necessary to specify prior distributions in order to

explore knot configurations through the MCMC algorithm. DiMatteo *et al.* (2001) nonetheless justify this approach by specifying only the covariance structure of the coefficients of the spline functions, which they take to be a unit information prior (see below for details).

In our present modelling context, it is possible to specify the parameters of the prior distributions since we have data of lesser quality which contains information about the curves we wish to model. In the same fashion as before, we write $\boldsymbol{y}^0$ to represent the vector made up of the $N_0 = 11$ yearly hydrographs of lesser quality. The prior location parameter $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$ is obtained as the maximizer of $\sum_{j=1}^{n} \left[\zeta_j \bar{z}^0_{\cdot j} - \psi(\zeta_j)\right]$, which has the same form as the first term of the joint distribution given in (4.2.3) but where it is now applied to the prior sample. Concerning $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$, we consider the two following structures : (a) $\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = (cN_0 \boldsymbol{B}'_{\boldsymbol{\omega}} \boldsymbol{B}_{\boldsymbol{\omega}})^{-1}$ and (b) $\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = (cN_0)^{-1} \mathbb{I}_{K_{\boldsymbol{\omega}}}$. The multiplicative factor cN_0 is chosen to indicate that the number of hydrographs in the prior sample is multiplied by a factor c varying between 0 and 1 ; when $c = 0$, the prior information is null since the covariance matrix is infinite, while for $c = 1$, each curve in the prior sample contributes as much as a curve in the effective sample. In the context of linear models, the structure given in (a) is a form that was first suggested by Zellner (1986) under the name of g-priors. It was later used in a theoretical framework by Kass and Wasserman (1995) who called it the unit information prior when $cN_0 = 1/n$, since the prior information regarding the regression parameters then contributes the weight of one observation, and Pauler (1998) ; in a more applied setting, Smith and Khon (1996) used this form of covariance structure. The form given in (b) implies that *a priori* we consider the coefficients of the spline functions to be independent and reflects a form of ignorance.

The prior parameters for the distribution of the precision parameter are determined in the same manner as above. The shape parameter is given by $\alpha_{\boldsymbol{\omega}} = c(nN_0/2)$, while the scale parameter is

$$\gamma_{\boldsymbol{\omega}} = c \left\{ \frac{N_0}{2} \sum_{j=1}^{n} \bar{d}^0_{\cdot j} \right\}, \tag{4.3.1}$$

where $\bar{d}^0_{\cdot j} = 2[\bar{s}^0_{\cdot j} - \bar{\eta}^0_{\cdot j}]$, as in equation (4.2.18), and $\bar{\eta}^0_{\cdot j}$ is evaluated at $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$. It should be noted that for each knot configuration $\boldsymbol{\omega}$, explored in the MCMC algorithm, the prior distributions are recomputed accordingly. For the numerical implementation, we take $c = 1/N_0$ and the average of the N_0 curves thus contributes the weight of one curve in the global model.

The prior specification of M, the number of possible knot positions, is taken to be twice the number of data points and the support consists of equally spaced potential knot positions which cover the domain of the data. Furthermore, similar to the approach of Denison *et al.* (1998), we put a constraint on the proximity of interior knots ; more specifically, we constrain the knots in order that there is at least one data point between adjacent knots. Since we know our data forms a continuous function, there is no need for interior knot multiplicity which would be useful when the data to be modelled features abrupt jumps. Some experimentation has been conducted as regards to the number of equally spaced possible knot positions and it was found that the quality of modelling was not seriously affected by the fineness of the grid.

4.3.3. Results

The MCMC reversible jump algorithm is implemented in the following way. The different algorithms are run for 5000 iterations and we consider the first 2500 to constitute the burn-in period ; the last 2500 iterations are used to determine the mode of the knot configurations. We study 3 types of algorithms for a given distribution and a given link function. The first one uses the transition probabilities based on $\mathrm{m}_1(\boldsymbol{y}|\boldsymbol{\omega})$, the Schwarz criterion approximation ; in this instance, we write the mode of the knot configuration as $\boldsymbol{\omega}^\dagger_1$. The other 2 algorithms rely on $\mathrm{m}_2(\boldsymbol{y}|\boldsymbol{\omega})$ to calculate the transition probabilities. In this case, we study two different prior covariance structures for the coefficients of the spline functions and we write the modes as $\boldsymbol{\omega}^\dagger_{2a}$ and $\boldsymbol{\omega}^\dagger_{2b}$ corresponding to the covariance structures specified above.

4.3.3.1. *Model selection*

Table 4.1 gives the logarithm of the ratio of the partial marginal distributions evaluated at the corresponding modes of the converged chains. For a given column, the models are compared to a reference model which is taken to be the normal distribution with the identity link function ($\mathcal{N}$, IDL), *i.e.* the standard linear model. Positive values indicate better performance than the reference model, while negative values indicate the opposite. Table 4.1 also presents, in parentheses, the number of knots of the 'best' or modal model for a given distribution and a given link function.

According to these results, the 'best' model for all the selection criteria is the lognormal distribution ($\mathcal{LN}$) with the identity link (IDL) ; although the other links for this distribution also give high values, we still only consider the highest values since we want the procedure to be automatic for operational purposes. If we only look at the best link for each distribution, we find that the distributions are ordered in the following way : $\mathcal{LN}$, $\mathcal{G}$, IG, RIG, and $\mathcal{N}$. For the studied sample of curves, it thus appears that a distribution with a variance proportional to the square of the mean represents the best distribution, *i.e.* the $\mathcal{LN}$ and the $\mathcal{G}$ distributions.

For a given distribution, we find that in the normal case, the best link is the identity for both $\log(BF^{\dagger}_{2a})$ and $\log(BF^{\dagger}_{2b})$, while the logarithmic link is slightly better for $\log(BF^{\dagger}_{1})$. For the gamma distribution, the highest values are obtained for the logarithmic link and this is true for all the selection criteria ; for the inverse gaussian, the inverse link constitutes the best link according to all the criteria. Finally, for the reciprocal inverse gaussian, we find that the best link function is always the logarithmic link.

It is instructive to look at the different selection criteria separately. In the case of the Schwarz approximation, the values in Table 4.1 for a fixed distribution are all very similar, except in the case of the RIG distribution. This reflects the fact that, in general, the spline models are flexible enough to model the average of the sample whatever link function is chosen. Furthermore, since this criterion does not take into account the marginal covariance of the observations, it is not

TABLE 4.1. Comparison of the different models according to the logarithm of expression (4.2.22), where the reference model for all the calculations in a column is $B = (\mathcal{N},\ \text{IDL})$. The modal number of interior knots is also given in parentheses.

		$\log(BF_1^\dagger)$ $(m_1^\dagger)$	$\log(BF_{2a}^\dagger)$ $(m_{2a}^\dagger)$	$\log(BF_{2b}^\dagger)$ $(m_{2b}^\dagger)$
$\mathcal{N}$	IDL	0 (7)	0 (9)	0 (10)
	LOL	5 (5)	-45 (5)	-39 (5)
	INL	1 (5)	-85 (3)	-80 (3)
$\mathcal{G}$	IDL	1171 (7)	-5799 (4)	-7459 (21)
	LOL	1174 (6)	1191 (7)	1188 (12)
	INL	1170 (6)	1147 (4)	1153 (4)
IG	IDL	977 (7)	-12743 (5)	-14250 (14)
	LOL	980 (6)	564 (5)	-354 (20)
	INL	984 (5)	960 (5)	966 (5)
$\mathcal{LN}$	IDL	1257 (5)	1250 (7)	1251 (9)
	LOL	1255 (5)	1241 (5)	1247 (5)
	INL	1255 (5)	1229 (5)	1235 (5)
RIG	IDL	882 (7)	867 (5)	874 (5)
	LOL	955 (5)	928 (5)	930 (5)
	INL	-1557 (7)	211 (8)	-591 (22)

able to discriminate the adequacy of the covariance structure which depends on the link function. Important differences are nonetheless present between several distributions, and this approximation thus seems able to determine the adequacy of the statistical distributions.

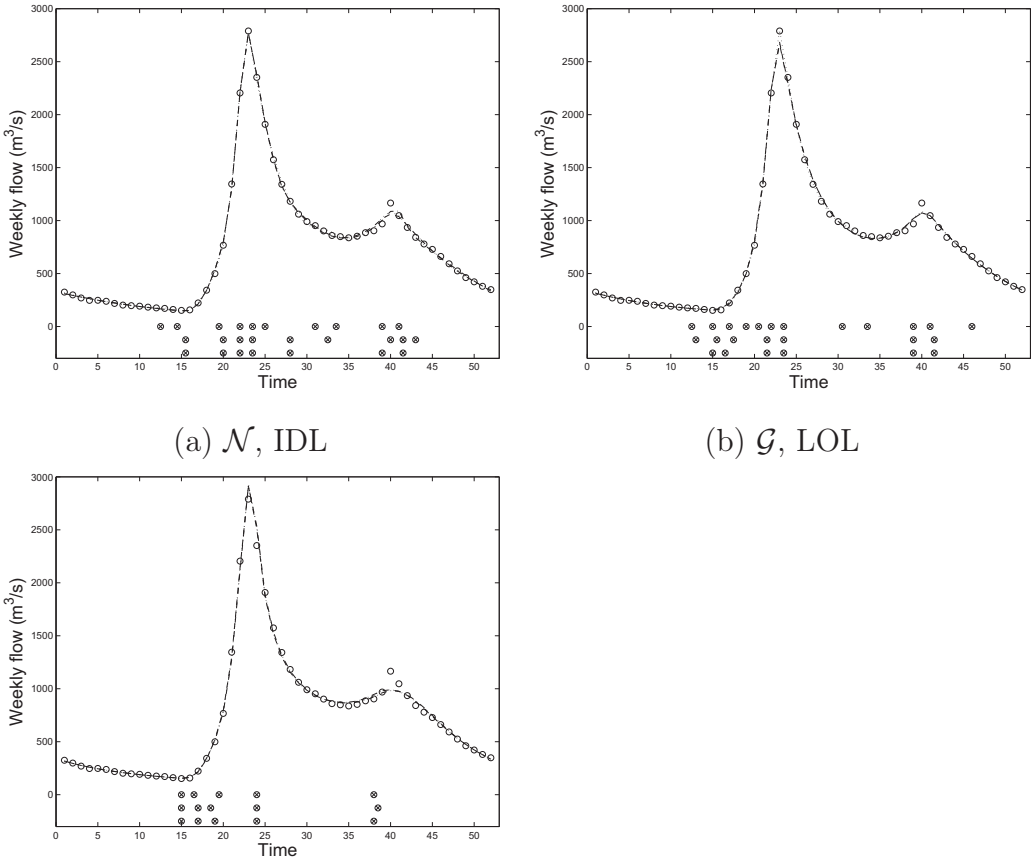
For the other two columns, it should be noticed that certain entries have large negative values ; there are two explanations for these results. The first is that the chain converged to a model configuration which was unable to model the average curve properly, for example the model ($\mathcal{G}$, IDL) in the second column ; when the chain converges to a knot configuration that does not model the average adequately, it indicates that the chain was not able to move to higher dimensions because

of low transition probabilities associated with high dimension models. The second explanation is that the number of knots is excessively large, for example ($\mathcal{G}$, IDL) in the third column. In this latter case, although the average curve is well modelled, a large penalty for dimensionality is introduced. It is interesting to note that in this case the chain manages to move to higher dimensions, which comes from the fact that the prior covariance structure of the spline coefficients is taken to be the identity matrix. In a similar fashion to ridge regression, multicollinearity effects are attenuated with this covariance structure and higher dimensional models can be obtained.

4.3.3.2. *Function estimation and approximate credible sets*

Figure 4.2 shows the estimated functions associated with the best link functions, according to the criteria given in Table 4.1, for the $\mathcal{N}$, $\mathcal{G}$, and IG distributions, *i.e.* the distributions for which the data is not transformed. The only exception to this rule is in the case of the normal distribution according to the selection criterion based on the Schwarz approximation. In this instance, we also consider the identity link function since it is the best link for the other selection criteria and for comparative purposes. The modal knot configurations of each of the Markov chains are also displayed at the bottom of Figure 4.2. For the cases shown, all the algorithms model the average very well by capturing the spring flood and the global behaviour of the data. The ($\mathcal{N}$, IDL) model seems to give the best fit since it captures the spring and autumn flood peaks. The ($\mathcal{G}$, LOL) model falls a bit short concerning these two maxima but nonetheless performs well globally, while the combination (IG, INL) fully replicates the spring maximum but has difficulty with the autumn peak because it appears to over smooth the summer-autumn period.

Figure 4.3 gives the estimated functions corresponding to the best link functions for the $\mathcal{LN}$ and RIG distributions according to the criteria given in Table 4.1. The results for both the transformed and the untransformed scales are presented since the systematic component, or the spline functions, model the former, while the latter is the scale of interest in practice. Furthermore, it is interesting



(a) $\mathcal{N}$, IDL

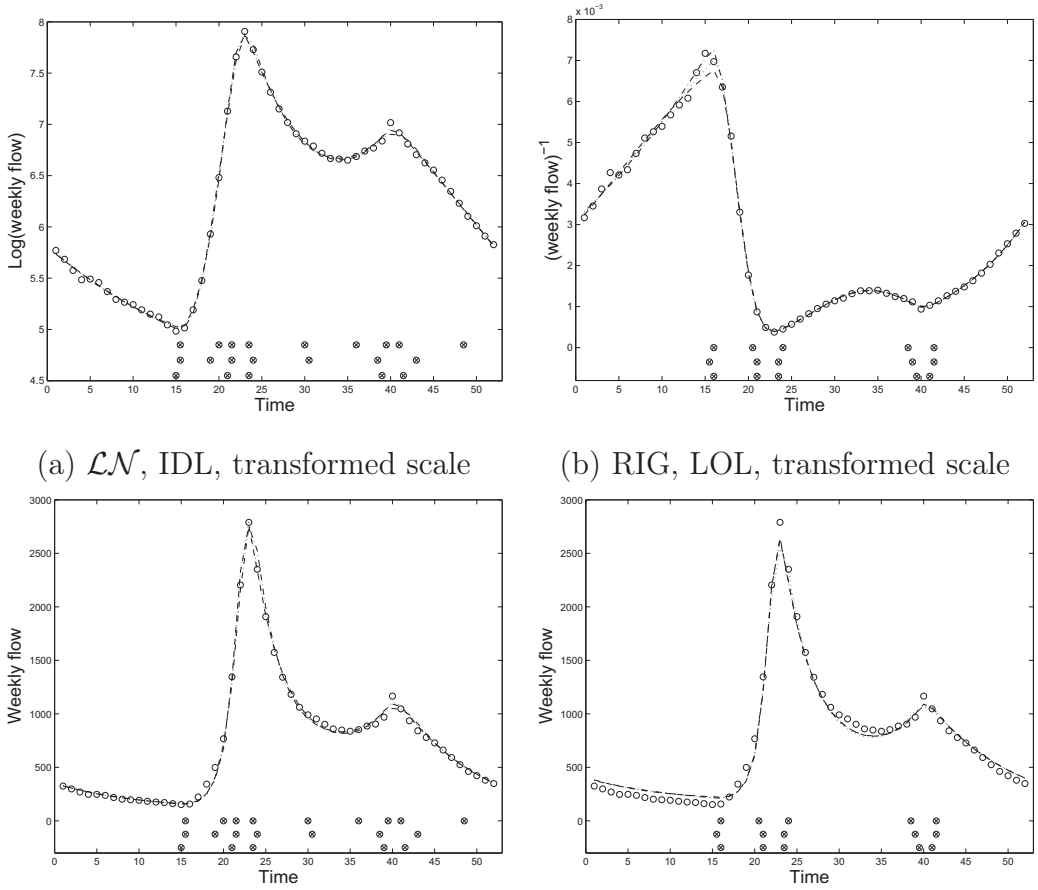
(b) $\mathcal{G}$, LOL

(c) IG, INL

FIGURE 4.2. Function estimation for different distributions and link functions. The open circles represent the weekly average of the sample of curves, while the crossed circles at the bottom represent the modal knot configurations. From bottom to top, these are $\boldsymbol{\omega}_1^\dagger$, $\boldsymbol{\omega}_{2a}^\dagger$, and $\boldsymbol{\omega}_{2b}^\dagger$; the corresponding models are dashed, dot-dashed, and dotted respectively.

to see how the spline functions perform on the transformed scale for which the sample average can exhibit different features from those obtained on the untransformed scale. For the lognormal distribution, the results on both the transformed and the untransformed scales are very adequate. Concerning the reciprocal inverse gaussian distribution, the models seem to perform fairly well on the transformed scale, but appear to be the least adequate models on the untransformed scale.

For a given distribution and a given link function, in both Figures 4.2 and 4.3, it is important to note the similarity of the fits for the different modal knot



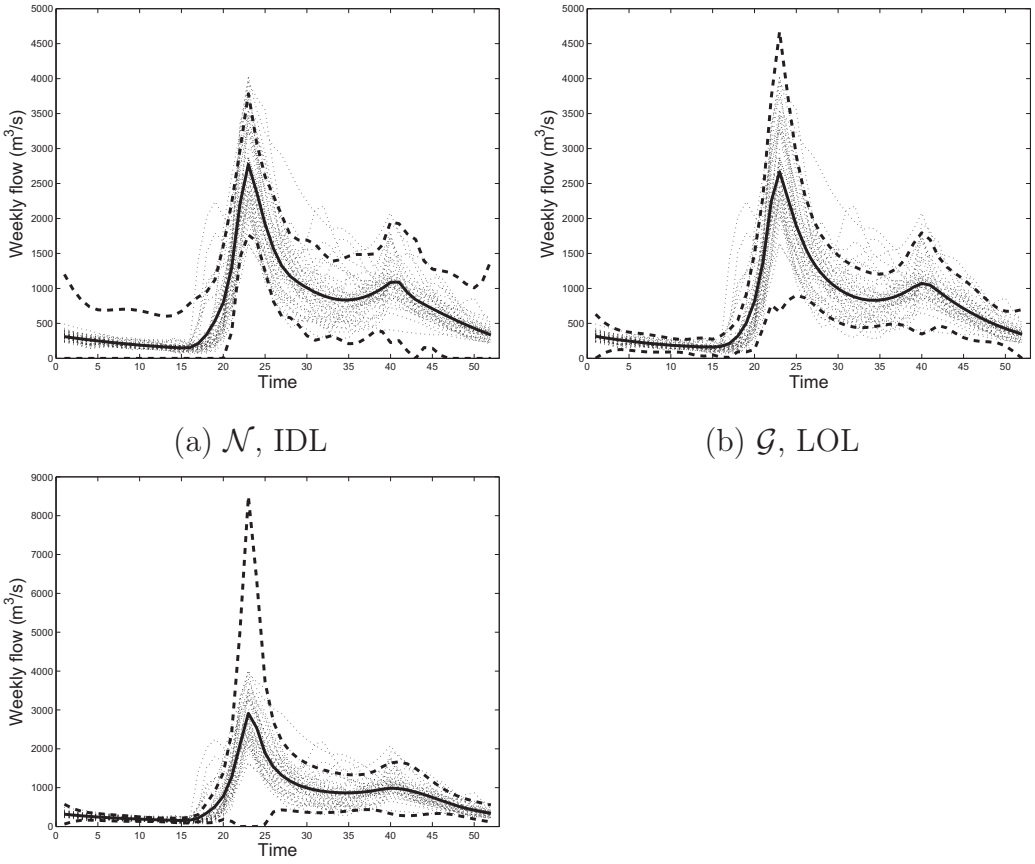
(a) $\mathcal{LN}$, IDL, transformed scale (b) RIG, LOL, transformed scale

(c) $\mathcal{LN}$, IDL, untransformed scale (d) RIG, LOL, untransformed scale

FIGURE 4.3. Function estimation for different distributions and link functions. The open circles represent the weekly average of the sample of curves, while the crossed circles at the bottom represent the modal knot configurations. From bottom to top, these are $\boldsymbol{\omega}_1^\dagger$, $\boldsymbol{\omega}_{2a}^\dagger$, and $\boldsymbol{\omega}_{2b}^\dagger$; the corresponding models are dashed, dot-dashed, and dotted respectively.

configurations and also the fact that all of these basically superpose each other. Since the modal quantities exhibit some variability, this confirms the fact that many spline solutions can give an adequate fit to a data set.

Figures 4.4 and 4.5 show approximate credible sets, constructed from the methodology exposed in section 4.2.5.2, for the combinations of distribution and link function shown in Figure 4.2. Figure 4.4 displays the credible sets for $\boldsymbol{\omega}_{2a}^\dagger$, while



(a) $\mathcal{N}$, IDL

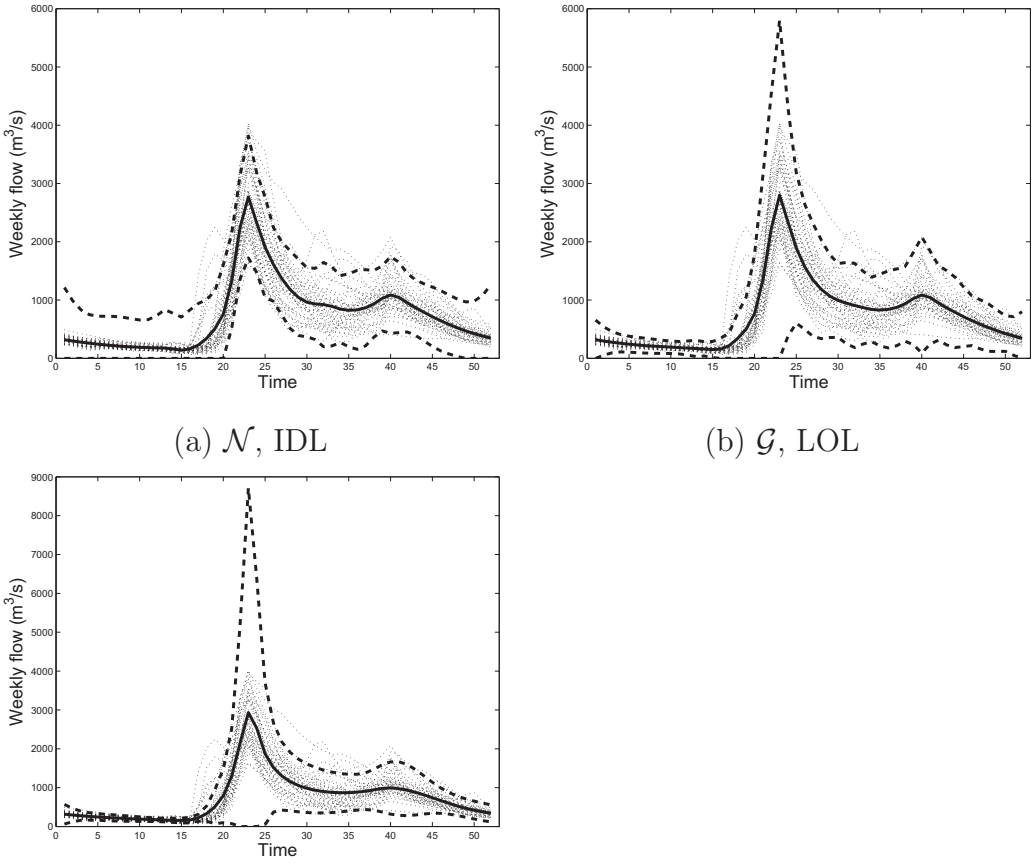
(b) $\mathcal{G}$, LOL

(c) IG, INL

FIGURE 4.4. Approximate credible sets (bold dashed lines), obtained from the method described in section 4.2.5, for the models corresponding to $\boldsymbol{\omega}_{2a}^{\dagger}$ of Figure 4.2. The sample of hydrographs are shown as dotted lines, while the full bold line represents the function estimate.

Figure 4.5 presents those obtained from $\boldsymbol{\omega}_{2b}^{\dagger}$. The credible sets have been truncated at zero since water flow measurements are always nonnegative. Although it is not of major importance in this study since the main interest of the model does not lie in these regions, we notice the boundary effects of the spline basis; this could be avoided by putting different constraints on the boundaries of the domain by using natural splines for example (Ruppert *et al.*, 2003).

It is insightful to look at the individual cases of Figure 4.4 separately to understand the differences in the credible sets. The model ($\mathcal{N}$, IDL) clearly exhibits



(a) $\mathcal{N}$, IDL

(b) $\mathcal{G}$, LOL

(c) IG, INL

FIGURE 4.5. Approximate credible sets (bold dashed lines), obtained from the method described in section 4.2.5, for the models corresponding to $\boldsymbol{\omega}_{2b}^{\dagger}$ of Figure 4.2. The sample of hydrographs are shown as dotted lines, while the full bold line represents the function estimate.

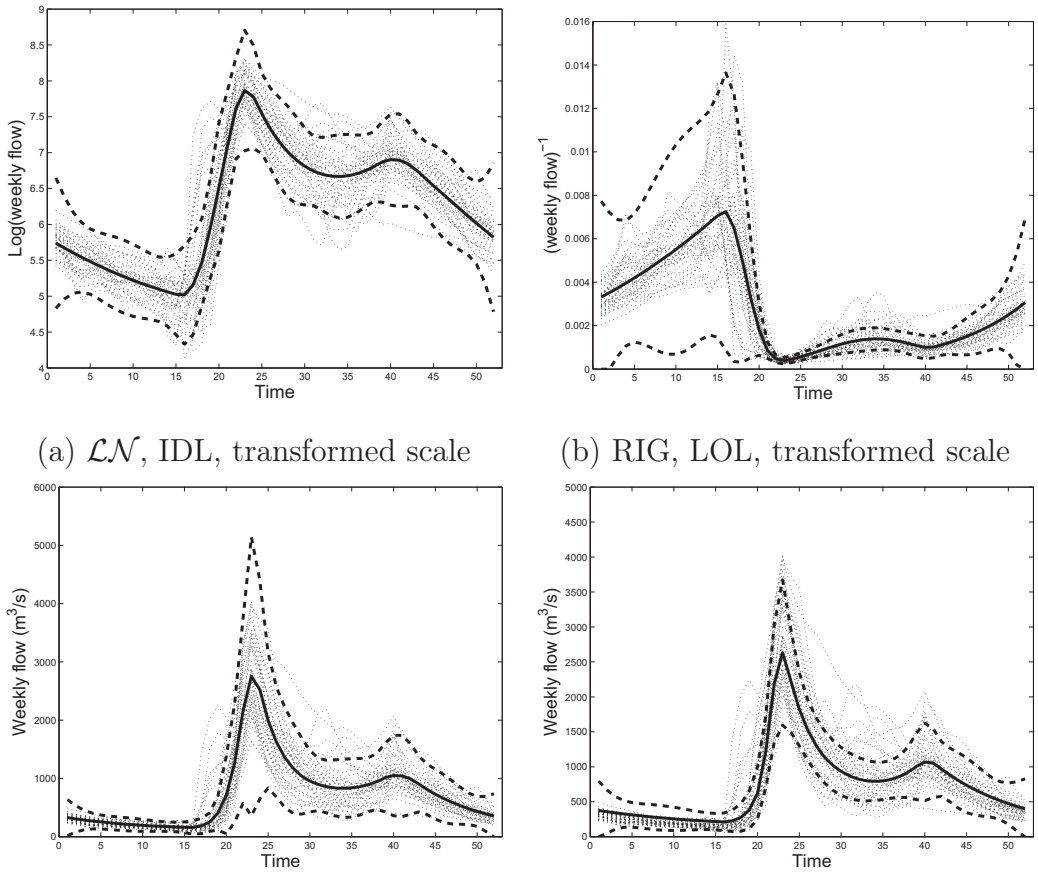
the independence of the mean and variance associated with the normal distribution; for a global precision parameter that represents the inverse of a constant variance, this leads to credible sets which are too wide in certain parts of the domain, such as the winter period, and that probably underestimate the spread of the data in highly variable regions such as the spring flood. On the other extreme, the dependence between the variance and the mean of the inverse gaussian is cubic (see Appendix A), and this is reflected markedly in the flood region where

the credible sets become quite large. The model ($\mathcal{G}$, LOL) seems to capture satisfactorily the observed variability of the hydrographs compared with the two preceding models.

Figure 4.5 exhibits the same type of behaviours as the ones indicated above, although some differences can be put forward. The contribution of the design matrix to the covariance structure, through the matrix $\mathbf{\Lambda}^*_{\boldsymbol{\omega}}$ (see sections 4.2.4.1 and 4.2.5), can increase the variance substantially when more knots are present in a certain region of the domain. For example, if we compare the autumn regions of the ($\mathcal{N}$, IDL) models shown in panel (a) of Figures 4.4 and 4.5, we notice that the width of the credible sets of the former is greater than those of the latter. This happens because of an extra knot in this region which can be seen in Figure 4.2. Another instance of this can be seen for the models ($\mathcal{G}$, LOL) in the region of the spring flood. The credible sets shown in Figure 4.5 (b) are wider than those of Figure 4.4 (b), which again happens because of extra knots (see panel (b) of Figure 4.2).

Figures 4.6 and 4.7 present approximate credible sets for the combinations of distribution and link function shown in Figure 4.3. Figure 4.6 displays the credible sets for $\boldsymbol{\omega}^{\dagger}_{2a}$, while Figure 4.7 gives those obtained from $\boldsymbol{\omega}^{\dagger}_{2b}$. The credible sets for both the transformed and untransformed scales are shown. We see that the credible sets for the lognormal distribution, on the transformed scale, possess roughly a constant width throughout the domain and they seem to adequately cover the data on the logarithmic scale. For the reciprocal inverse gaussian distribution, the variance is proportional to the cube of the mean on the reciprocal scale and the credible sets appear to be too wide on the left of the domain, while being too narrow in the middle region.

On the untransformed scales, we obtain the mean and variance relations given in Appendix A for the $\mathcal{LN}$ and RIG distributions, *i.e.* a quadratic and linear relation respectively. The ($\mathcal{LN}$, IDL) models give credible sets that are quite similar to those of the ($\mathcal{G}$, LOL) models, which reflects their common quadratic mean-variance relation. As discussed above, the influence of the covariance structure is seen in panel (c) of Figures 4.6 and 4.7, where the effect on the credible sets is



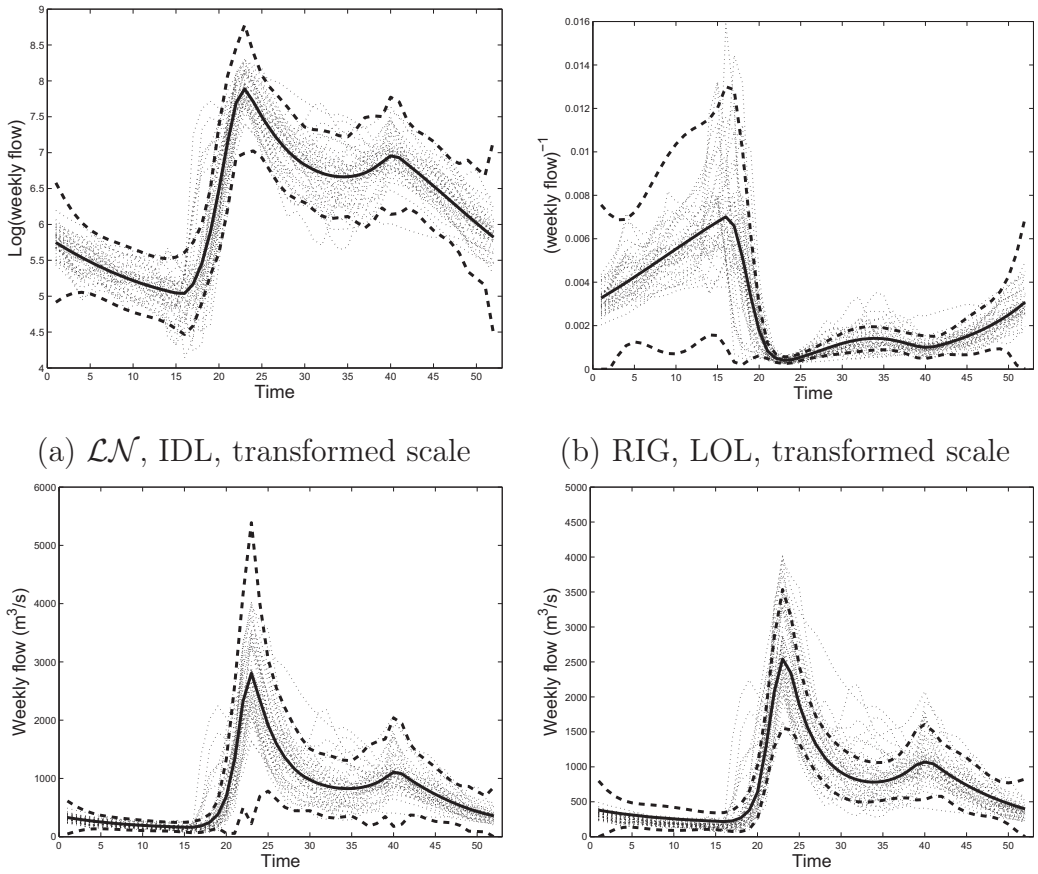
(a) $\mathcal{LN}$, IDL, transformed scale (b) RIG, LOL, transformed scale

(c) $\mathcal{LN}$, IDL, untransformed scale (d) RIG, LOL, untransformed scale

FIGURE 4.6. Approximate credible sets (bold dashed lines), obtained from the method described in section 4.2.5, for the models corresponding to $\boldsymbol{\omega}_{2a}^{\dagger}$ of Figure 4.3. The sample of hydrographs are shown as dotted lines, while the full bold line represents the function estimate.

mostly present in the autumn period. The credible sets of the (RIG, LOL) models are tighter than those of the ($\mathcal{N}$, IDL) models but they do not capture the full variability of the sample of hydrographs, as can be seen in panel (d) of Figures 4.6 and 4.7.

Finally to sum up the results of the application, Table 4.1 indicates that the $\mathcal{LN}$ and $\mathcal{G}$ distributions appear to be the most adequate to model the sample of curves. This is confirmed by the approximate credible sets shown in Figures 4.4, 4.5, 4.6, and 4.7. These results are in agreement with many hydrological studies which have used these two distributions to adequately model water flow data.



(a) $\mathcal{LN}$, IDL, transformed scale (b) RIG, LOL, transformed scale

(c) $\mathcal{LN}$, IDL, untransformed scale (d) RIG, LOL, untransformed scale

FIGURE 4.7. Approximate credible sets (bold dashed lines), obtained from the method described in section 4.2.5, for the models corresponding to $\boldsymbol{\omega}_{2b}^{\dagger}$ of Figure 4.3. The sample of hydrographs are shown as dotted lines, while the full bold line represents the function estimate.

4.4. CONCLUSION

We have presented a formal Bayesian methodology, based on free-knot regression splines, to model the average behaviour of a sample of univariate functional data, thought to arise from a class of continuous distributions that includes those of the exponential family. Inspired by Denison *et al.* (1998) and DiMatteo *et al.* (2001), we use a reversible jump MCMC algorithm to explore the posterior distribution of knot configurations. To simplify the transition probabilities of the

Metropolis-Hastings step and also speed up the convergence of the chain, the algorithm relies on the partial marginal distribution of the observations. Since this distribution is not explicit in the general framework, we considered two approximations, one based on the Schwarz criterion, and a second obtained by using the basic form of Laplace's approximation of an integral. We also put forward a way of determining the most adequate statistical distribution for a data set and gave expressions to construct approximate credible sets for a sample of curves.

It should be clear that the statistical model can directly be applied to a single set of functional data instead of a sample of curves. A direct extension of the methodology is to generalize it to include several additive 'predictors', instead of only one, as suggested in Denison *et al.* (1998) in the case of the normal distribution. Since the exploration of the knot configurations, which define the modelling basis and therefore the design matrix, is similar to a variable selection procedure, the methodology can also be applied for variable selection purposes in generalized linear models.

Finally, another possible extension is to model simultaneously the mean and dispersion processes with Bayesian free-knot regression splines in the same spirit as suggested by McCullagh and Nelder (1989) for generalized linear models.

Appendix A. Statistical distributions

Table 4.2 gives explicit information about distributional quantities referred to in the main text. The first two lines present the expectation and the variance of each distribution ; the second line thus shows how the variance depends on the mean and establishes the value of the exponent p in the relation : variance $\propto \{\text{mean}\}^p$. We also see that for the $\mathcal{LN}$ and RIG distributions, both the expectation and the variance depend on the precision parameter ϕ. The third line relates η of the main text and the parameter μ which is modelled by a function of the regression splines, while the fourth line gives the canonical link (CL) of each distribution. Concerning the last two lines, they are associated with the following decomposition of the function κ that is useful to determine a prior distribution for the precision parameter ϕ. For all the distributions except the gamma distribution,

TABLE 4.2. Information concerning the different statistical distributions : normal ($\mathcal{N}$), gamma ($\mathcal{G}$), inverse gaussian (IG), lognormal ($\mathcal{LN}$), and reciprocal inverse gaussian (RIG) distributions.

	$\mathcal{N}(\mu,\phi)$	$\mathcal{G}(\mu,\phi)$	IG(μ,ϕ)	$\mathcal{LN}(\mu,\phi)$	RIG(μ,ϕ)
$\mathbb{E}(x)$	μ	μ	μ	$\mu_x = \exp\left(\mu + \phi^{-1}/2\right)$	$\mu_x = \mu^{-1} + \phi^{-1}$
$\mathbb{V}(x)$	ϕ^{-1}	$\phi^{-1}\mu^2$	$\phi^{-1}\mu^3$	$\mu_x^2\left(\exp(\phi^{-1}) - 1\right)$	$\phi^{-1}\mu_x + \phi^{-2}$
$\eta(x,\mu)$	$\mu x - \frac{\mu^2}{2}$	$-\frac{x}{\mu} - \log(\mu)$	$-\frac{x}{2\mu^2} + \frac{1}{\mu}$	$\mu\log(x) - \frac{\mu^2}{2}$	$-\frac{1}{2\mu^2 x} + \frac{1}{\mu}$
CL	IDL $(\mu = \zeta)$	INL $(\mu^{-1} = \zeta)$	ISL $(\mu^{-2} = \zeta)$	IDL $(\mu = \zeta)$	ISL $(\mu^{-2} = \zeta)$
$s(x)$	$\frac{1}{2}x^2$	$-\log(x) - 1$	$\frac{1}{2}x^{-1}$	$\frac{1}{2}[\log(x)]^2$	$\frac{1}{2}x$
$t(x)$	0	$-\log(x)$	$-\frac{1}{2}\log(x^3)$	$-\log(x)$	$-\frac{1}{2}\log(x)$

the quantity $\kappa(x,\phi)$ can be written as

$$\kappa(x,\phi) = -\phi s(x) + \frac{1}{2}\log(\phi) - \frac{1}{2}\log(2\pi) + t(x). \tag{A.1}$$

For the gamma distribution, we have

$$\kappa(x,\phi) = \phi\log(x) + \phi\log(\phi) - \log\left[\Gamma(\phi)\right] - \log(x), \tag{A.2}$$

which displays a more complex dependence of κ relative to the precision parameter ϕ. We can write (Abramowitz and Stegun, 1964)

$$\Gamma(\phi) = (2\pi)^{1/2}\phi^{\phi-1/2}\exp(-\phi)\left\{1 + O(\phi^{-1})\right\}, \tag{A.3}$$

and when ϕ is large, we obtain : $\Gamma(\phi) \approx (2\pi)^{1/2}\phi^{\phi-1/2}\exp(-\phi)$, which is Stirling's approximation. Using this latter approximation, we get the same form as the one in equation (A.1) with $s(x)$ and $t(x)$ given in Table 4.2; we thus see that the common form of $\kappa(x,\phi)$ suggests a gamma prior for the precision parameter. As noted by Jorgensen (1997, p. 104), using Stirling's approximation is equivalent to a saddlepoint approximation of the gamma distribution.

With the given form of $\kappa(x,\phi)$, the distribution of one observation is given by

$$f(x|\mu,\phi) \cong \exp\left\{-\phi\left[s(x) - \eta(x,\mu)\right] + \frac{1}{2}\log(\phi) - \frac{1}{2}\log(2\pi) + t(x)\right\}, \tag{A.4}$$

where we use the symbol $\cong$ to indicate that the distribution is exact except in the case of the gamma distribution. The quantity in square brackets can be related to the deviance which is the standard measure of goodness of fit for

generalized linear models. For a sample of observations, the deviance is defined as twice the difference between the maximum log likelihood achievable and that achieved by a model under consideration for a fixed unit dispersion (or precision) parameter (see chapter 2 of McCullagh and Nelder, 1989). With the present notation, the contribution of one observation to the deviance is given by $\hat{d} = d(x, \hat{\mu}) = 2\left[\eta(x, z) - \eta(x, \hat{\mu})\right]$, where $z = z(x)$ is the transformation function of the data and $\hat{\mu}$ represents the maximum likelihood estimate of μ for the model under consideration. For the distributions considered, we find that

$$[s(x) - \eta(x, \hat{\mu})] = \frac{1}{2}\hat{d}, \tag{A.5}$$

which implies that the quantity on the left is always larger or equal to zero from the definition of the deviance.

In the context of modelling the sample of curves, we use $s_{ij} = s(y_{ij})$, $t_{ij} = t(y_{ij})$, and $d_{ij} = d(y_{ij})$, where this last quantity represents the contribution of the observation y_{ij} to the total deviance. With this notation at hand, we have $\kappa_{ij} = -\phi s_{ij} + \frac{1}{2}\log(\phi) - \frac{1}{2}\log(2\pi) + t_{ij}$, and $d_{ij} = 2[s_{ij} - \eta_{ij}]$; therefore the joint distribution of the observations can be written under the two following forms

$$f(\boldsymbol{y}|\boldsymbol{\zeta}, \phi) = \exp\left\{N\phi \sum_{j=1}^{n} \bar{\eta}_{\cdot j} + \kappa_{\cdot\cdot}\right\} \tag{A.6}$$

$$\cong \exp\left\{-\phi\left[\frac{N}{2}\sum_{j=1}^{n} \bar{d}_{\cdot j}\right] + t_{\cdot\cdot} + \frac{Nn}{2}\left[\log(\phi) - \log(2\pi)\right]\right\}, \tag{A.7}$$

where $\cong$ is again used to indicate that the expression is exact for all the distributions except for the gamma distribution.

Appendix B. MCMC reversible jump algorithm

The MCMC reversible jump algorithm relies on a Metropolis-Hastings step at each iteration. The acceptance probabilites of this step, as originally proposed by Green (1995), are given by

$$\rho_t = \min\left(1, \text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio}\right), \tag{B.1}$$

for a particular move type t. In our application, the possible move types are $t = a$ for the addition of a knot, $t = s$ for the suppression of a knot and $t = d$ for

the displacement of a knot. In our approach, the likelihoods are replaced by the marginal distributions $\mathrm{m}_1(\boldsymbol{y}|\boldsymbol{\omega})$ or $\mathrm{m}_2(\boldsymbol{y}|\boldsymbol{\omega})$, which constitute two approximations to the integration of the parameters $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ and $\phi_{\boldsymbol{\omega}}$. In this respect, our methodology is similar to that of DiMatteo *et al.* (2001), who worked with the first of these approximations; Denison *et al.* (1998) directly considered the likelihoods since they did not use a fully Bayesian model in the sense that no prior distribution was assigned to $\boldsymbol{\beta}_{\boldsymbol{\omega}}$.

The prior ratios are based on the prior distributions given in section 4.2.3; for example when $t = s$ and the model currently contains m knots, the prior ratio is given by

$$\text{prior ratio} = \frac{\pi_2(\boldsymbol{r}^{(m-1)}|m-1)\pi_1(m-1)}{\pi_2(\boldsymbol{r}^{(m)}|m)\pi_1(m)}. \tag{B.2}$$

Concerning the proposal ratios, care needs to be taken in order to have a balanced chain. Writing the probabilities of choosing $t = \{a, s, d\}$, when the model contains m knots, as a_m, s_m and d_m respectively, we have the necessary constraint : $a_m + s_m + d_m = 1$. Following Denison *et al.* (1998), we take

$$a_m = c \times \min\left(1, \frac{\pi_1(m+1)}{\pi_1(m)}\right), \tag{B.3}$$

$$s_m = c \times \min\left(1, \frac{\pi_1(m-1)}{\pi_1(m)}\right), \tag{B.4}$$

which ensures that

$$\frac{a_m}{s_{m+1}} = \frac{\pi_1(m+1)}{\pi_1(m)}. \tag{B.5}$$

The constant c determines the rate at which the dimension of the model changes and it needs to be between 0 and 0.5 for the sum of the move type probabilities to equal one. For the application section we use $c = 0.4$ like Denison *et al.* (1998) and DiMatteo *et al.* (2001). For example, the proposal ratio for $t = s$, when the model possesses m knots, is taken to be

$$\text{proposal ratio} = \frac{a_{m-1}/(M - I_{m-1})}{s_m/m}, \tag{B.6}$$

where I_{m-1} represents the number of impossible positions when there are $m-1$ knots; the unavailable positions are fixed by constraints on the proximity of

the knots. In this framework, for a model containing m knots, the acceptance probabilities are given by

$$\rho_a = \min\left(1, \frac{\mathrm{m}(\boldsymbol{y}|m+1, \boldsymbol{r}_*^{(m+1)})}{\mathrm{m}(\boldsymbol{y}|m, \boldsymbol{r}^{(m)})} \times \frac{M - I_m}{M}\right), \tag{B.7}$$

$$\rho_s = \min\left(1, \frac{\mathrm{m}(\boldsymbol{y}|m-1, \boldsymbol{r}_*^{(m-1)})}{\mathrm{m}(\boldsymbol{y}|m, \boldsymbol{r}^{(m)})} \times \frac{M}{M - I_{m-1}}\right), \tag{B.8}$$

$$\rho_d = \min\left(1, \frac{\mathrm{m}(\boldsymbol{y}|m, \boldsymbol{r}_*^{(m)})}{\mathrm{m}(\boldsymbol{y}|m, \boldsymbol{r}^{(m)})}\right), \tag{B.9}$$

where $\boldsymbol{r}_*^{(\cdot)}$ indicates a proposed knot configuration and $\boldsymbol{r}^{(\cdot)}$, the current knot configuration.

Appendix C. Derivation of $\mathrm{m}_2(\boldsymbol{y}|\boldsymbol{\omega})$

The joint distribution, given a knot configuration $\boldsymbol{\omega}$ and taking $\boldsymbol{\Sigma}$ to be fixed, is

$$D(\boldsymbol{y}, \boldsymbol{\beta}, \phi|\boldsymbol{\omega}) = f(\boldsymbol{y}|\boldsymbol{\zeta}, \phi, \boldsymbol{\omega})\pi(\boldsymbol{\beta}|\boldsymbol{\beta}^0, \phi, \boldsymbol{\omega})\pi(\phi|\alpha, \gamma, \boldsymbol{\omega}), \tag{C.1}$$

where all subscripts $\boldsymbol{\omega}$ have been removed to simplify the notation and $\boldsymbol{\zeta}$ is a function of $\boldsymbol{\beta}$ which depends on the distribution and the link function. We want to approximate the integral $\int_{\boldsymbol{\beta},\phi} D(\boldsymbol{y}, \boldsymbol{\beta}, \phi|\boldsymbol{\omega})\mathrm{d}\boldsymbol{\beta}\mathrm{d}\phi$. We first approximate the integral over the vector of parameters $\boldsymbol{\beta}$ by using the basic form of Laplace's approximation. Using equation (A.7) and the prior distribution for $\boldsymbol{\beta}$, we can write

$$\begin{aligned} D(\boldsymbol{y}, \boldsymbol{\beta}, \phi|\boldsymbol{\omega}) \cong & \pi(\phi|\alpha, \gamma, \boldsymbol{\omega})\phi^{(Nn+K)/2} \\ & \exp\left\{-\phi\left[\frac{N}{2}\sum_{j=1}^{n}\bar{d}_{\cdot j} + q(\boldsymbol{\beta})\right] + t_{\cdot\cdot} + c_1\right\}, \end{aligned} \tag{C.2}$$

where $\cong$ is used to indicate that the distribution is exact except in the case of the gamma distribution, $q(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)$ and $c_1 = -\frac{1}{2}\log\{|\boldsymbol{\Sigma}|\} - \frac{Nn+K}{2}\log(2\pi)$. Applying Laplace's approximation, one gets

$$\begin{aligned} D(\boldsymbol{y}, \phi|\boldsymbol{\omega}) \approx & \pi(\phi|\alpha, \gamma, \boldsymbol{\omega})\phi^{Nn/2} \\ & \exp\left\{-\phi\left[\frac{N}{2}\sum_{j=1}^{n}\bar{d}^*_{\cdot j} + q(\boldsymbol{\beta}^*)\right] + t_{\cdot\cdot} + c_2\right\}, \end{aligned} \tag{C.3}$$

where $\bar{d}^*_{\cdot j}$ and $q(\boldsymbol{\beta}^*)$ are evaluated at $\boldsymbol{\beta}^*$ (given in equation (4.2.19)),

$$c_2 = c_1 - \frac{1}{2}\log\left\{|\boldsymbol{\Lambda}^* + \boldsymbol{\Sigma}^{-1}|\right\} + \frac{K}{2}\log(2\pi) \tag{C.4}$$

$$= \frac{1}{2}\log\{|\boldsymbol{\Sigma}^*|\} - \frac{1}{2}\log\{|\boldsymbol{\Sigma}|\} - \frac{Nn}{2}\log(2\pi), \tag{C.5}$$

$\boldsymbol{\Lambda}^*$ and $\boldsymbol{\Sigma}^*$ are respectively given in equations (4.2.21) and (4.2.20). Forgetting the prior density of ϕ, writing $\phi^{Nn/2}$ in the exponential, and using equation (A.6), we then see that the expression which is exponentiated possesses a form similar to the Schwarz criterion :

$$N\phi\sum_{j=1}^{n}\bar{\eta}^*_{\cdot j} + \kappa_{\cdot\cdot} - \frac{\phi}{2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0) + c_2. \tag{C.6}$$

The first term is the same as that of the Schwarz criterion except for where it is evaluated in the parameter space, while the second term is identical. The third term measures the discrepancy between the prior location of $\boldsymbol{\beta}$ and the mode of the posterior distribution, and the last term takes into account dimensionality through the ratios of determinants. We thus see that the Schwarz criterion penalizes only for dimensionality, while the last two terms above penalize for both dimensionality and prior information, *i.e.* prior location, prior covariance and prior structure of the model. Since we work with a prior distribution for ϕ, we can now integrate ϕ instead of using an estimate. To get the expression given in (4.2.15), we use equation (C.3) and the prior distribution of ϕ given in section 4.2.3.

Chapitre 5

ESTIMATION POUR L'ARTICLE 3

Dans l'article 3 (chapitre 6), nous présentons un modèle basé sur la même famille de distributions que celle explorée dans l'article 2 (chapitre 4), mais le modèle proposé est plus général puisque nous modélisons simultanément la tendance centrale et la dispersion. Le présent chapitre indique comment la famille étudiée peut être utilisée à cette fin. De plus, d'une façon analogue au chapitre 3, nous obtenons les estimateurs du maximum de vraisemblance et du maximum de la distribution *a posteriori* qui sont employés dans l'article 3.

5.1. Modélisation de la tendance centrale et de la dispersion

Nous étudions les mêmes distributions statistiques appartenant à la famille $\mathcal{F}$ (voir section 3.1) dans cet article ; par contre, nous considérons maintenant que le paramètre de dispersion (ou de précision) peut varier en fonction des observations. Plus précisément, nous étudions des échantillons, $\boldsymbol{y} = (y_1, \ldots, y_j, \ldots, y_n)'$, dont les observations sont indépendamment distribuées selon la famille $\mathcal{F}$, c'est-à-dire $y_j \sim \mathcal{F}(\zeta_j, \varphi_j)$, où le deuxième paramètre, φ_j, est maintenant le paramètre de dispersion. Ici, nous travaillons avec le paramètre de dispersion au lieu du paramètre de précision afin de rendre l'exposé plus clair.

La densité conjointe d'un tel échantillon s'exprime comme

$$f_1(\boldsymbol{y}|\boldsymbol{\zeta},\boldsymbol{\varphi}) = \prod_{j=1}^{n} \exp\left\{\frac{[\zeta_j z_j - \psi(\zeta_j)]}{\varphi_j} + \kappa(y_j,\varphi_j)\right\}, \tag{5.1.1}$$

$$= \exp\left\{\left[\boldsymbol{\zeta}'\boldsymbol{\Phi}^{-1}\boldsymbol{z} - \boldsymbol{\psi}(\boldsymbol{\zeta})'\boldsymbol{\Phi}^{-1}\mathbf{1}\right] + \boldsymbol{\kappa}(\boldsymbol{y},\boldsymbol{\varphi})'\mathbf{1}\right\}, \tag{5.1.2}$$

$$= \exp\left\{\mathcal{L}_1\right\}, \tag{5.1.3}$$

où $\boldsymbol{\zeta}$, $\boldsymbol{z}$ et $\boldsymbol{\psi}(\boldsymbol{\zeta})$ sont définis à la section 3.1 ; $\boldsymbol{\Phi} = \text{Diag}\,(\varphi_j)$ est une matrice diagonale $n \times n$ et $\boldsymbol{\kappa}(\boldsymbol{y},\boldsymbol{\varphi}) = (\kappa(y_1,\varphi_1),\ldots,\kappa(y_j,\varphi_j),\ldots,\kappa(y_n,\varphi_n))'$. L'indice pour la densité conjointe et le logarithme de la vraisemblance est utilisé puisqu'il est pratique d'employer deux représentations de ces deux quantités selon que nous considérons le modèle pour la tendance centrale ou celui pour la dispersion. Dans le présent contexte de modélisation, nous utilisons la décomposition donnée à l'équation (3.3.4), mais où le paramètre de précision est remplacé par le paramètre de dispersion. Nous avons alors

$$\kappa(y_j,\varphi_j) \cong -\frac{s(y_j)}{\varphi_j} - \frac{1}{2}\log(\varphi_j) + t(y_j) - \frac{1}{2}\log(2\pi), \tag{5.1.4}$$

où le symbole $\cong$ indique que cette décomposition est exacte pour toutes les distributions étudiées sauf la distribution gamma. Pour la distribution gamma, cette décomposition correspond à une approximation par point de selle comme discuté à la section 3.3. En posant $\lambda(\varphi_j) = \log(\varphi_j)$ et $\boldsymbol{c} = -\frac{1}{2}\log(2\pi)\mathbf{1}$, nous avons

$$\boldsymbol{\kappa}(\boldsymbol{y},\boldsymbol{\varphi})'\mathbf{1} \cong -\boldsymbol{s}(\boldsymbol{y})'\boldsymbol{\Phi}^{-1}\mathbf{1} - \frac{1}{2}\boldsymbol{\lambda}(\boldsymbol{\varphi})'\mathbf{1} + \boldsymbol{t}(\boldsymbol{y})'\mathbf{1} + \boldsymbol{c}'\mathbf{1}. \tag{5.1.5}$$

En utilisant (5.1.5) et en regroupant les termes qui contiennent $\boldsymbol{\Phi}^{-1}$ à l'équation (5.1.2), nous obtenons

$$\boldsymbol{\zeta}'\boldsymbol{\Phi}^{-1}\boldsymbol{z} - \boldsymbol{\psi}(\boldsymbol{\zeta})'\boldsymbol{\Phi}^{-1}\mathbf{1} - \boldsymbol{s}(\boldsymbol{y})'\boldsymbol{\Phi}^{-1}\mathbf{1} = \frac{1}{2}\boldsymbol{\xi}'\boldsymbol{d}, \tag{5.1.6}$$

où

$$\boldsymbol{\xi} = (\xi_1,\ldots,\xi_j,\ldots,\xi_n)' = \left(-\varphi_1^{-1},\ldots,-\varphi_j^{-1},\ldots,-\varphi_n^{-1}\right)', \tag{5.1.7}$$

$$\boldsymbol{d} = 2\left[\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{\eta}\left(\boldsymbol{y},\boldsymbol{\zeta}\right)\right], \tag{5.1.8}$$

$$\boldsymbol{\eta}\left(\boldsymbol{y},\boldsymbol{\zeta}\right) = \left[\zeta_1 z_1 - \psi(\zeta_1),\ldots,\zeta_j z_j - \psi(\zeta_j),\ldots,\zeta_n z_n - \psi(\zeta_n)\right]'. \tag{5.1.9}$$

Avec cette représentation, nous pouvons maintenant écrire la densité conjointe sous la forme suivante

$$f_2(\boldsymbol{y}|\boldsymbol{\zeta},\boldsymbol{\varphi}) \cong \exp\left\{\frac{1}{2}\left[\boldsymbol{\xi}'\boldsymbol{d} - \boldsymbol{\lambda}(\boldsymbol{\xi})'\mathbf{1}\right] + \boldsymbol{t}(\boldsymbol{y})'\mathbf{1} + \boldsymbol{c}'\mathbf{1}\right\}, \tag{5.1.10}$$

$$= \exp\left\{\left[\boldsymbol{\xi}'\boldsymbol{D}^{-1}\boldsymbol{d} - \boldsymbol{\lambda}(\boldsymbol{\xi})'\boldsymbol{D}^{-1}\mathbf{1}\right] + \boldsymbol{t}(\boldsymbol{y})'\mathbf{1} + \boldsymbol{c}'\mathbf{1}\right\}, \tag{5.1.11}$$

$$= \exp\left\{\mathcal{L}_2\right\}, \tag{5.1.12}$$

où $\boldsymbol{\lambda}(\boldsymbol{\xi}) = (\lambda(\xi_1), \ldots, \lambda(\xi_j), \ldots, \lambda(\xi_n))'$, $\lambda(\xi_j) = -\log(-\xi_j)$ et $\boldsymbol{D} = \text{Diag}\,(2)$; cette dernière matrice diagonale est utilisée afin d'obtenir deux représentations tout à fait analogues de la densité conjointe. Comme précédemment, l'indice pour la densité conjointe et le logarithme de la vraisemblance désigne la seconde représentation de ces quantités.

Nous voyons donc que les deux représentations données aux équations (5.1.2) et (5.1.11) possèdent la même forme en ce qui a trait aux deux quantités qui se retrouvent entre les crochets. La première représentation, équation (5.1.2), est utile lorsque nous considérons le modèle pour la tendance centrale, alors que la seconde, équation (5.1.11), l'est quand le modèle pour la dispersion est étudié. Il est important de noter que le modèle au niveau de la tendance centrale dépend de la distribution statistique au niveau des observations, alors que le modèle pour la dispersion possède toujours la forme d'un modèle gamma pour le vecteur $\boldsymbol{d}$, où $\boldsymbol{\varphi}$ représente son espérance (voir tableau 3.1).

Puisque nous cherchons à modéliser simultanément la tendance centrale et la dispersion, nous travaillons maintenant avec deux composantes systématiques. Nous notons le vecteur des composantes systématiques associées à la tendance centrale et le vecteur de celles associées à la dispersion par

$$\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta} \qquad \text{et} \qquad \boldsymbol{w} = \mathbf{X}\boldsymbol{\gamma}, \tag{5.1.13}$$

où $\boldsymbol{X}$ et $\boldsymbol{\beta}$ dénotent respectivement la matrice d'incidence $n \times K_{\boldsymbol{\beta}}$ et le vecteur des paramètres $K_{\boldsymbol{\beta}} \times 1$ pour la tendance centrale, alors que $\mathbf{X}$ et $\boldsymbol{\gamma}$ représentent la matrice d'incidence $n \times K_{\boldsymbol{\gamma}}$ et le vecteur des paramètres $K_{\boldsymbol{\gamma}} \times 1$ pour la dispersion. Ici, et dans ce qui suit, nous utilisons des caractères en italique pour les quantités

associées à la tendance centrale et les caractères usuels pour les quantités similaires associées à la dispersion. Les vecteurs $n \times 1$ des moyennes et des dispersions peuvent être écrits relativement aux inverses des fonctions liens. Ainsi,

$$\boldsymbol{\mu} = \boldsymbol{h}(\boldsymbol{u}) \qquad \text{et} \qquad \boldsymbol{\varphi} = \mathbf{h}(\boldsymbol{w}), \tag{5.1.14}$$

où $\boldsymbol{h}(\boldsymbol{u}) = (h(u_1), \ldots, h(u_j), \ldots, h(u_n))'$ et $\mathbf{h}(\boldsymbol{w}) = (\mathrm{h}(w_1), \ldots, \mathrm{h}(w_j), \ldots, \mathrm{h}(w_n))'$; l'inverse de la fonction lien pour la tendance centrale est donnée par h, alors que celle pour la dispersion, par h.

5.2. Estimation par maximum de vraisemblance

Afin de calculer les estimateurs du maximum de vraisemblance, nous calculons les premières et deuxièmes dérivées partielles de la vraisemblance puisque nous utilisons l'algorithme de Newton-Raphson pour déterminer ces estimateurs. Afin d'exposer les calculs clairement, il est utile d'employer les deux représentations du logarithme de la vraisemblance, $\mathcal{L}_1$ (équation (5.1.3)) et $\mathcal{L}_2$ (équation (5.1.12)), selon que nous calculons les dérivées par rapport à $\boldsymbol{\beta}$ ou par rapport à $\boldsymbol{\gamma}$.

Les deux premières dérivées du logarithme de la vraisemblance par rapport à $\boldsymbol{\beta}$ sont très similaires à celles calculées à la section 3.2, sauf pour la présence de la matrice diagonale $\boldsymbol{\Phi}^{-1}$. En utilisant les résultats donnés aux équations (3.2.3)-(3.2.8), nous obtenons

$$\nabla_{\boldsymbol{\beta}}\mathcal{L}_1 = \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\zeta}'\right]\boldsymbol{\Phi}^{-1}\boldsymbol{z} + \nabla_{\boldsymbol{\beta}}\left[\boldsymbol{\psi}(\boldsymbol{\zeta})'\right]\boldsymbol{\Phi}^{-1}\mathbf{1}, \tag{5.2.1}$$

$$= \boldsymbol{X}'\boldsymbol{W}\boldsymbol{\Phi}^{-1}\boldsymbol{z} + \boldsymbol{X}'\boldsymbol{W}\mathrm{Diag}\left\{\mu_j\right\}\boldsymbol{\Phi}^{-1}\mathbf{1}, \tag{5.2.2}$$

$$= \boldsymbol{X}'\boldsymbol{W}\boldsymbol{\Phi}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}), \tag{5.2.3}$$

puisque $\mathrm{Diag}\left\{\mu_j\right\}\boldsymbol{\Phi}^{-1} = \boldsymbol{\Phi}^{-1}\mathrm{Diag}\left\{\mu_j\right\}$. Comme à la section 3.2, nous avons la matrice diagonale $\boldsymbol{W} = \mathrm{Diag}\left\{v_j^{-1}\dot{h}_j\right\}$, où $v_j = v(\mu_j)$ et $\dot{h}_j = \nabla_{u_j}\left[h(u_j)\right]$. Pour le calcul de la deuxième dérivée, nous notons que

$$\nabla'_{\boldsymbol{\beta}}\mathcal{L}_1 = (\boldsymbol{z} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}\boldsymbol{W}\boldsymbol{X}, \tag{5.2.4}$$

$$= (\boldsymbol{z} - \boldsymbol{\mu})'\boldsymbol{W}\boldsymbol{\Phi}^{-1}\boldsymbol{X}, \tag{5.2.5}$$

puisque les matrices $\boldsymbol{W}$ et $\boldsymbol{\Phi}^{-1}$ sont des matrices diagonales. Nous pouvons alors appliquer la règle du produit avec les différentes formes données à l'équation (3.2.21). Ainsi, avec les résultats des équations (3.2.22)-(3.2.24), nous obtenons

$$\nabla_{\boldsymbol{\beta}} \nabla'_{\boldsymbol{\beta}} \mathcal{L}_1 = \boldsymbol{X}' \boldsymbol{V} \boldsymbol{\Phi}^{-1} \boldsymbol{X}, \tag{5.2.6}$$

où $\boldsymbol{V} = \text{Diag}\left\{ v_j^{-1} \left[(z_j - \mu_j)(\ddot{h}_j - v_j^{-1} \dot{v}_j \dot{h}_j^2) - \dot{h}_j^2 \right] \right\}$, $\ddot{h}_j = \nabla_{u_j} \nabla'_{u_j} [h(u_j)]$ et $\dot{v}_j = \nabla_{\mu_j} [v(\mu_j)]$. En ce qui concerne les deux premières dérivées de la vraisemblance par rapport à $\boldsymbol{\gamma}$, nous pouvons maintenant utiliser la deuxième représentation du logarithme de la vraisemblance afin d'obtenir des résultats totalement similaires à ceux qui viennent d'être obtenus. Nous avons alors

$$\nabla_{\boldsymbol{\gamma}} \mathcal{L}_2 = \mathbf{X}' \mathbf{W} \boldsymbol{D}^{-1} (\boldsymbol{d} - \boldsymbol{\varphi}), \tag{5.2.7}$$

$$\nabla_{\boldsymbol{\gamma}} \nabla'_{\boldsymbol{\gamma}} \mathcal{L}_2 = \mathbf{X}' \mathbf{V} \boldsymbol{D}^{-1} \mathbf{X}, \tag{5.2.8}$$

où

$$\mathbf{W} = \text{Diag}\left\{ \mathrm{v}_j^{-1} \dot{\mathrm{h}}_j \right\}, \tag{5.2.9}$$

$$\mathbf{V} = \text{Diag}\left\{ \mathrm{v}_j^{-1} \left[(d_j - \varphi_j)(\ddot{\mathrm{h}}_j - \mathrm{v}_j^{-1} \dot{\mathrm{v}}_j \dot{\mathrm{h}}_j^2) - \dot{\mathrm{h}}_j^2 \right] \right\}, \tag{5.2.10}$$

$$\mathrm{v}_j = \mathrm{v}(\varphi_j), \tag{5.2.11}$$

$$\dot{\mathrm{h}}_j = \nabla_{w_j} [\mathrm{h}(w_j)], \tag{5.2.12}$$

$$\ddot{\mathrm{h}}_j = \nabla_{w_j} \nabla'_{w_j} [\mathrm{h}(w_j)], \tag{5.2.13}$$

$$\dot{\mathrm{v}}_j = \nabla_{\varphi_j} [\mathrm{v}(\varphi_j)]. \tag{5.2.14}$$

Afin de complètement spécifier la matrice hessienne du logarithme de la vraisemblance, il ne reste qu'à calculer les dérivées croisées. Nous avons

$$\nabla_{\boldsymbol{\gamma}} \nabla'_{\boldsymbol{\beta}} \mathcal{L}_1 = \nabla_{\boldsymbol{\gamma}} \left\{ (\boldsymbol{z} - \boldsymbol{\mu})' \boldsymbol{\Phi}^{-1} \boldsymbol{W} \boldsymbol{X} \right\}, \tag{5.2.15}$$

$$= \nabla_{\boldsymbol{\gamma}} \{ \boldsymbol{a}' \} \boldsymbol{A} \boldsymbol{X}, \tag{5.2.16}$$

en utilisant comme décomposition le vecteur $\boldsymbol{a} = \left(\varphi_1^{-1}, \ldots, \varphi_j^{-1}, \ldots, \varphi_n^{-1} \right)'$ et la matrice $\boldsymbol{A} = \text{Diag}\left\{ v_j^{-1} \dot{h}_j (z_j - \mu_j) \right\}$. Selon la règle des dérivées en chaîne, nous avons

$$\nabla_{\boldsymbol{\gamma}} \{ \boldsymbol{a}' \} = \nabla_{\boldsymbol{\gamma}} [\boldsymbol{w}'] \nabla_{\boldsymbol{w}} [\boldsymbol{\varphi}'] \nabla_{\boldsymbol{\varphi}} [\boldsymbol{a}']. \tag{5.2.17}$$

De la même façon qu'aux équations (3.2.5) et (3.2.6), nous calculons

$$\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{w}'\right] = \mathbf{X}', \tag{5.2.18}$$

$$\nabla_{\boldsymbol{w}}\left[\boldsymbol{\varphi}'\right] = \nabla_{\boldsymbol{w}}\left[\mathbf{h}\left(\boldsymbol{w}\right)'\right] = \text{Diag}\left\{\dot{\text{h}}_j\right\}. \tag{5.2.19}$$

Enfin, la dernière dérivée est donnée par

$$\nabla_{\boldsymbol{\varphi}}\left[\boldsymbol{a}'\right] = -\text{Diag}\left\{\varphi_j^{-2}\right\} = -\text{Diag}\left\{\text{v}_j^{-1}\right\}, \tag{5.2.20}$$

puisque pour la modélisation de $\boldsymbol{\varphi}$, la fonction variance est quadratique, $\text{v} = \varphi^2$, car c'est un modèle gamma. En regroupant tous ces termes, nous obtenons

$$\nabla_{\boldsymbol{\gamma}}\nabla'_{\boldsymbol{\beta}}\mathcal{L}_1 = -\mathbf{X}'\boldsymbol{U}\boldsymbol{X}, \tag{5.2.21}$$

où $\boldsymbol{U} = \text{Diag}\left\{v_j^{-1}\dot{h}_j\text{v}_j^{-1}\dot{\text{h}}_j(z_j - \mu_j)\right\}$. D'une façon analogue, nous trouvons

$$\nabla_{\boldsymbol{\beta}}\nabla'_{\boldsymbol{\gamma}}\mathcal{L}_2 = -\boldsymbol{X}'\boldsymbol{U}\mathbf{X}. \tag{5.2.22}$$

Ainsi, comme à la section 3.2, l'algorithme de Newton-Raphson peut maintenant être appliqué afin de simultanément calculer les estimateurs du maximum de vraisemblance : $\widehat{\boldsymbol{\beta}}$ et $\widehat{\boldsymbol{\gamma}}$.

5.3. Estimation du maximum de la distribution *a posteriori*

Les distributions *a priori* considérées au niveau des paramètres des deux composantes systématiques sont les suivantes

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \mathcal{N}_{K_{\boldsymbol{\beta}}}\left(\boldsymbol{\beta}^0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \tag{5.3.1}$$

$$\boldsymbol{\gamma} \sim \mathcal{N}_{K_{\boldsymbol{\gamma}}}\left(\boldsymbol{\gamma}^0, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right). \tag{5.3.2}$$

La distribution *a priori* de $\boldsymbol{\beta}$ est conditionnelle au vecteur $\boldsymbol{\gamma}$ puisque nous choisissons de travailler avec une matrice de covariance qui dépend du paramètre de dispersion. Plus précisément, la forme de la matrice de covariance est donnée par

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(n_{\boldsymbol{\beta}}\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}, \tag{5.3.3}$$

où $n_{\boldsymbol{\beta}}$ est un facteur multiplicatif et $\boldsymbol{\Phi}$ est la matrice diagonale définie précédemment, pour laquelle les éléments sont modélisés par la composante systématique

qui dépend de $\boldsymbol{\gamma}$. Pour sa part, la matrice de covariance pour le vecteur $\boldsymbol{\gamma}$ est donnée par

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \left(n_{\boldsymbol{\gamma}} \mathbf{X}' \boldsymbol{D}^{-1} \mathbf{X}\right)^{-1}, \tag{5.3.4}$$

où $n_{\boldsymbol{\gamma}}$ est un facteur multiplicatif et $\boldsymbol{D} = \text{Diag}\,(2)$. Nous notons que les matrices de covariance données aux équations (5.3.3) et (5.3.4) possèdent des formes similaires à celles employées dans les articles 1 (chapitre 2) et 2 (chapitre 4). En ce qui concerne la matrice de covariance pour $\boldsymbol{\beta}$, nous introduisons la matrice $\boldsymbol{\Phi}$, qui dépend des paramètres de dispersion, afin d'avoir une distribution *a priori* conditionnelle aux paramètres de dispersion ; ceci est tout à fait analogue aux distributions *a priori* utilisées pour les paramètres associés à la tendance centrale dans les articles 1 et 2.

Les logarithmes des deux distributions *a priori* s'écrivent

$$\varpi_{\boldsymbol{\beta}} = -\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^0\right)' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^0\right) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right| - \frac{K_{\boldsymbol{\beta}}}{2}\log\left(2\pi\right), \tag{5.3.5}$$

$$\varpi_{\boldsymbol{\gamma}} = -\frac{1}{2}\left(\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right)' \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \left(\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right| - \frac{K_{\boldsymbol{\gamma}}}{2}\log\left(2\pi\right). \tag{5.3.6}$$

Nous procédons maintenant au calcul des premières et deuxièmes dérivées de ces deux quantités.

En ce qui concerne les deux premières dérivées de $\varpi_{\boldsymbol{\beta}}$ par rapport à $\boldsymbol{\beta}$, nous obtenons les mêmes résultats qu'à la section 3.3. Nous calculons

$$\nabla_{\boldsymbol{\beta}} \varpi_{\boldsymbol{\beta}} = -\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^0\right) = -n_{\boldsymbol{\beta}} \boldsymbol{X}' \boldsymbol{\Phi}^{-1} \boldsymbol{X}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^0\right), \tag{5.3.7}$$

$$\nabla_{\boldsymbol{\beta}} \nabla'_{\boldsymbol{\beta}} \varpi_{\boldsymbol{\beta}} = -\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} = -n_{\boldsymbol{\beta}} \boldsymbol{X}' \boldsymbol{\Phi}^{-1} \boldsymbol{X}. \tag{5.3.8}$$

De la même façon, les deux premières dérivées de $\varpi_{\boldsymbol{\gamma}}$ par rapport à $\boldsymbol{\gamma}$ sont données par

$$\nabla_{\boldsymbol{\gamma}} \varpi_{\boldsymbol{\gamma}} = -\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\left(\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right) = -n_{\boldsymbol{\gamma}} \mathbf{X}' \boldsymbol{D}^{-1} \mathbf{X}\left(\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right), \tag{5.3.9}$$

$$\nabla_{\boldsymbol{\gamma}} \nabla'_{\boldsymbol{\gamma}} \varpi_{\boldsymbol{\gamma}} = -\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} = -n_{\boldsymbol{\gamma}} \mathbf{X}' \boldsymbol{D}^{-1} \mathbf{X}. \tag{5.3.10}$$

Puisque la matrice de covariance *a priori* de $\boldsymbol{\beta}$ dépend des paramètres de dispersion et donc, du vecteur de paramètres $\boldsymbol{\gamma}$, nous calculons maintenant la première dérivée de $\varpi_{\boldsymbol{\beta}}$ par rapport à $\boldsymbol{\gamma}$. En spécifiant la matrice de covariance, donnée à l'équation (5.3.3), dans l'expression de $\varpi_{\boldsymbol{\beta}}$ (équation (5.3.5)), nous voyons

alors que les deux termes qui dépendent directement de $\boldsymbol{\gamma}$ sont

$$T_1 = -\frac{n_{\boldsymbol{\beta}}}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^0\right)'\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^0\right), \tag{5.3.11}$$

$$T_2 = \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}|. \tag{5.3.12}$$

Le premier terme peut être réécrit de la façon suivante

$$T_1 = -\frac{n_{\boldsymbol{\beta}}}{2}\left(\boldsymbol{u}-\boldsymbol{u}^0\right)'\boldsymbol{\Phi}^{-1}\left(\boldsymbol{u}-\boldsymbol{u}^0\right), \tag{5.3.13}$$

$$= -\frac{n_{\boldsymbol{\beta}}}{2}\boldsymbol{a}'\boldsymbol{A}\left(\boldsymbol{u}-\boldsymbol{u}^0\right), \tag{5.3.14}$$

où $\boldsymbol{u}=\boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{u}^0=\boldsymbol{X}\boldsymbol{\beta}^0$, $\boldsymbol{a}=\left(\varphi_1^{-1},\ldots,\varphi_j^{-1},\ldots,\varphi_n^{-1}\right)'$ et $\boldsymbol{A}=\mathrm{Diag}\left\{\left(u_j-u_j^0\right)\right\}$. Ainsi, la première dérivée par rapport à $\boldsymbol{\gamma}$ de T_1 est donnée par

$$\nabla_{\boldsymbol{\gamma}}T_1 = -\frac{n_{\boldsymbol{\beta}}}{2}\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}'\right]\boldsymbol{A}\left(\boldsymbol{u}-\boldsymbol{u}^0\right). \tag{5.3.15}$$

En utilisant les résultats obtenus aux équations (5.2.17)-(5.2.20), nous calculons

$$\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}'\right] = \mathbf{X}'\mathrm{Diag}\left\{\mathrm{v}_j^{-1}\dot{\mathrm{h}}_j\right\}. \tag{5.3.16}$$

Ainsi, nous obtenons

$$\nabla_{\boldsymbol{\gamma}}T_1 = \frac{n_{\boldsymbol{\beta}}}{2}\mathbf{X}'\mathrm{Diag}\left\{\mathrm{v}_j^{-1}\dot{\mathrm{h}}_j\left(u_j-u_j^0\right)^2\right\}\mathbf{1}. \tag{5.3.17}$$

La dérivée du second terme est donnée par

$$\nabla_{\boldsymbol{\gamma}}T_2 = \frac{1}{2}\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{w}'\right]\nabla_{\boldsymbol{w}}\left[\boldsymbol{\varphi}'\right]\nabla_{\boldsymbol{\varphi}}\left[\log|\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}|\right]. \tag{5.3.18}$$

Les deux premiers termes ont été calculés à la section précédente et nous notons que pour un élément du dernier terme, nous avons

$$\nabla_{\varphi_j}\left[\log|\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}|\right] = tr\left\{\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\nabla_{\varphi_j}\left[\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right]\right\}. \tag{5.3.19}$$

De plus,

$$\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X} = \sum_{j=1}^{n}\varphi_j^{-1}\boldsymbol{x}_j\boldsymbol{x}_j', \tag{5.3.20}$$

alors nous calculons

$$\nabla_{\varphi_j}\left[\sum_{j=1}^{n}\varphi_j^{-1}\boldsymbol{x}_j\boldsymbol{x}_j'\right] = \sum_{j=1}^{n}\nabla_{\varphi_j}\left[\varphi_j^{-1}\right]\boldsymbol{x}_j\boldsymbol{x}_j', \tag{5.3.21}$$

$$= -\varphi_j^{-2}\boldsymbol{x}_j\boldsymbol{x}_j'. \tag{5.3.22}$$

Ainsi,

$$\nabla_{\varphi_j}\left[\log|\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}|\right] = -tr\left\{\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\varphi_j^{-2}\boldsymbol{x}_j\boldsymbol{x}_j'\right\}, \tag{5.3.23}$$

$$= -\varphi_j^{-2}\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j. \tag{5.3.24}$$

En regroupant tous les termes, nous obtenons

$$\nabla_{\boldsymbol{\gamma}}T_2 = -\frac{1}{2}\mathbf{X}'\mathrm{Diag}\left\{\varphi_j^{-2}\dot{\mathrm{h}}_j\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right\}\mathbf{1}. \tag{5.3.25}$$

La première dérivée de $\varpi_{\boldsymbol{\beta}}$ par rapport à $\boldsymbol{\gamma}$ est donc donnée par

$$\nabla_{\boldsymbol{\gamma}}\varpi_{\boldsymbol{\beta}} = \nabla_{\boldsymbol{\gamma}}\left[T_1 + T_2\right], \tag{5.3.26}$$

$$= \frac{1}{2}\mathbf{X}'\mathrm{Diag}\left\{\mathrm{v}_j^{-1}\dot{\mathrm{h}}_j\left[n_{\boldsymbol{\beta}}\left(u_j - u_j^0\right)^2 - \boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right]\right\}\mathbf{1}. \tag{5.3.27}$$

Pour calculer la deuxième dérivée de $\varpi_{\boldsymbol{\beta}}$ par rapport à $\boldsymbol{\gamma}$, nous procédons encore avec les deux termes séparément. En ce qui concerne la deuxième dérivée du premier terme, nous pouvons écrire

$$\nabla_{\boldsymbol{\gamma}}'T_1 = \frac{n_{\boldsymbol{\beta}}}{2}\mathbf{1}'\mathrm{Diag}\left\{\mathrm{v}_j^{-1}\dot{\mathrm{h}}_j\left(u_j - u_j^0\right)^2\right\}\mathbf{X}, \tag{5.3.28}$$

$$= \frac{n_{\boldsymbol{\beta}}}{2}\boldsymbol{a}_1'\boldsymbol{A}_1\mathbf{X}, \tag{5.3.29}$$

$$= \frac{n_{\boldsymbol{\beta}}}{2}\boldsymbol{a}_2'\boldsymbol{A}_2\mathbf{X}, \tag{5.3.30}$$

où

$$\boldsymbol{a}_1 = \left(\mathrm{v}_1^{-1},\ldots,\mathrm{v}_j^{-1},\ldots,\mathrm{v}_n^{-1}\right)', \tag{5.3.31}$$

$$\boldsymbol{a}_2 = \left(\dot{\mathrm{h}}_1,\ldots,\dot{\mathrm{h}}_j,\ldots,\dot{\mathrm{h}}_n\right)', \tag{5.3.32}$$

$$\boldsymbol{A}_1 = \mathrm{Diag}\left\{\dot{\mathrm{h}}_j\left(u_j - u_j^0\right)^2\right\}, \tag{5.3.33}$$

$$\boldsymbol{A}_2 = \mathrm{Diag}\left\{\mathrm{v}_j^{-1}\left(u_j - u_j^0\right)^2\right\}. \tag{5.3.34}$$

Les deux formes, données aux équations (5.3.29) et (5.3.30), peuvent être utilisées pour appliquer la règle du produit. Nous avons alors

$$\nabla_{\boldsymbol{\gamma}}\nabla_{\boldsymbol{\gamma}}'T_1 = \frac{n_{\boldsymbol{\beta}}}{2}\left\{\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_1'\right]\boldsymbol{A}_1\mathbf{X} + \nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_2'\right]\boldsymbol{A}_2\mathbf{X}\right\}. \tag{5.3.35}$$

Nous calculons

$$\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_1'\right] = -\mathbf{X}'\mathrm{Diag}\left\{\dot{\mathrm{h}}_j\dot{\mathrm{v}}_j\mathrm{v}_j^{-2}\right\}, \tag{5.3.36}$$

$$\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_2'\right] = \mathbf{X}'\mathrm{Diag}\left\{\ddot{\mathrm{h}}_j\right\}. \tag{5.3.37}$$

Ainsi,

$$\nabla_{\boldsymbol{\gamma}}\nabla_{\boldsymbol{\gamma}}'T_1 = \frac{n_{\boldsymbol{\beta}}}{2}\mathbf{X}'\mathrm{Diag}\left\{\mathrm{v}_j^{-1}\left(u_j - u_j^0\right)^2\left[\ddot{\mathrm{h}}_j - \mathrm{v}_j^{-1}\dot{\mathrm{v}}_j\dot{\mathrm{h}}_j^2\right]\right\}\mathbf{X}. \tag{5.3.38}$$

Pour la deuxième dérivée du second terme, nous notons que nous pouvons écrire

$$\nabla_{\boldsymbol{\gamma}}'T_2 = -\frac{1}{2}\mathbf{1}'\mathrm{Diag}\left\{\varphi_j^{-2}\dot{\mathrm{h}}_j\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right\}\mathbf{X}, \tag{5.3.39}$$

$$= -\frac{1}{2}\boldsymbol{a}_1'\boldsymbol{A}_1\mathbf{X}, \tag{5.3.40}$$

$$= -\frac{1}{2}\boldsymbol{a}_2'\boldsymbol{A}_2\mathbf{X}, \tag{5.3.41}$$

$$= -\frac{1}{2}\boldsymbol{a}_3'\boldsymbol{A}_3\mathbf{X}, \tag{5.3.42}$$

où

$$\boldsymbol{a}_1 = \left(\mathrm{v}_1^{-1}, \ldots, \mathrm{v}_j^{-1}, \ldots, \mathrm{v}_n^{-1}\right)', \tag{5.3.43}$$

$$\boldsymbol{a}_2 = \left(\dot{\mathrm{h}}_1, \ldots, \dot{\mathrm{h}}_j, \ldots, \dot{\mathrm{h}}_n\right)', \tag{5.3.44}$$

$$\boldsymbol{a}_3 = \left(\ldots, \boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j, \ldots\right)',$$

$$\boldsymbol{A}_1 = \mathrm{Diag}\left\{\dot{\mathrm{h}}_j\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right\}, \tag{5.3.45}$$

$$\boldsymbol{A}_2 = \mathrm{Diag}\left\{\mathrm{v}_j^{-1}\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right\}, \tag{5.3.46}$$

$$\boldsymbol{A}_3 = \mathrm{Diag}\left\{\mathrm{v}_j^{-1}\dot{\mathrm{h}}_j\right\}. \tag{5.3.47}$$

Pour utiliser la règle du produit, les trois formes données aux équations (5.3.40), (5.3.41) et (5.3.42) peuvent être employées. Ainsi, nous avons

$$\nabla_{\boldsymbol{\gamma}}\nabla_{\boldsymbol{\gamma}}'T_2 = -\frac{1}{2}\left\{\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_1'\right]\boldsymbol{A}_1\mathbf{X} + \nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_2'\right]\boldsymbol{A}_2\mathbf{X} + \nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_3'\right]\boldsymbol{A}_3\mathbf{X}\right\}. \tag{5.3.48}$$

Les deux dérivées $\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_1'\right]$ et $\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_2'\right]$ sont données aux équations (5.3.36) et (5.3.37). Il ne reste donc qu'à calculer la dérivée $\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{a}_3'\right]$. Pour un élément de

$\boldsymbol{a}_3'$, nous avons

$$\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right] = \nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{\varphi}'\right]\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right], \tag{5.3.49}$$

$$= \mathbf{X}'\mathrm{Diag}\left\{\dot{\mathrm{h}}_j\right\}\nabla_{\boldsymbol{\varphi}}\left[\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right], \tag{5.3.50}$$

en utilisant les résultats trouvés précédemment. Pour un élément de ce dernier terme, nous pouvons écrire

$$\nabla_{\varphi_i}\left[\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right] = \nabla_{\varphi_i}\left[tr\left\{\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\boldsymbol{x}_j'\right\}\right], \tag{5.3.51}$$

$$= tr\left\{\nabla_{\varphi_i}\left[\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\right]\boldsymbol{x}_j\boldsymbol{x}_j'\right\}. \tag{5.3.52}$$

Nous avons

$$\nabla_{\varphi_i}\left[\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\right] = -\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\nabla_{\varphi_i}\left[\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right]\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}, \tag{5.3.53}$$

$$= -\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\nabla_{\varphi_i}\left[\sum_{i=1}^{n}\varphi_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right]\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}, \tag{5.3.54}$$

$$= \varphi_i^{-2}\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}. \tag{5.3.55}$$

Donc nous pouvons écrire

$$\nabla_{\varphi_i}\left[\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right] = \varphi_i^{-2}\left[\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\right]^2. \tag{5.3.56}$$

En posant $\boldsymbol{c}_j = \nabla_{\boldsymbol{\varphi}}\left[\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right]$, pour lequel l'élément i est donné par (5.3.56), nous pouvons maintenant écrire

$$\nabla_{\boldsymbol{\gamma}}\left[\boldsymbol{x}_j'\left(\boldsymbol{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_j\right] = \mathbf{X}'\mathrm{Diag}\left\{\dot{\mathrm{h}}_j\right\}\boldsymbol{c}_j. \tag{5.3.57}$$

Ainsi, il est possible d'utiliser ce dernier résultat et les résultats précédents afin de complètement déterminer $\nabla_{\boldsymbol{\gamma}}\nabla_{\boldsymbol{\gamma}}'T_2$. Finalement, nous pouvons calculer

$$\nabla_{\boldsymbol{\gamma}}\nabla_{\boldsymbol{\gamma}}'\varpi_{\boldsymbol{\beta}} = \nabla_{\boldsymbol{\gamma}}\nabla_{\boldsymbol{\gamma}}'\left[T_1 + T_2\right]. \tag{5.3.58}$$

Dans un dernier temps, il reste à calculer les termes des dérivées croisées pour $\varpi_{\boldsymbol{\beta}}$. En utilisant (5.3.27), nous trouvons

$$\nabla_{\boldsymbol{\beta}}\nabla_{\boldsymbol{\gamma}}'\varpi_{\boldsymbol{\beta}} = n_{\boldsymbol{\beta}}\boldsymbol{X}'\mathrm{Diag}\left\{\mathrm{v}_j^{-1}\dot{\mathrm{h}}_j\left(u_j - u_j^0\right)\right\}\mathbf{X}, \tag{5.3.59}$$

et d'une façon similaire

$$\nabla_{\boldsymbol{\gamma}} \nabla'_{\boldsymbol{\beta}} \varpi_{\boldsymbol{\beta}} = n_{\boldsymbol{\beta}} \mathbf{X}' \mathrm{Diag} \left\{ \mathrm{v}_j^{-1} \dot{\mathrm{h}}_j \left(u_j - u_j^0 \right) \right\} \boldsymbol{X}. \tag{5.3.60}$$

Finalement, toutes ces dérivées partielles, ainsi que celles de la vraisemblance de la section précédente, peuvent être employées afin de faire appel à l'algorithme de Newton-Raphson pour calculer simultanément les estimateurs du maximum de la distribution *a posteriori* $\boldsymbol{\beta}^*$ et $\boldsymbol{\gamma}^*$. Ce sont ces estimateurs qui sont utilisés pour effectuer l'approximation de Laplace.

Chapitre 6

SIMULTANEOUS MODELLING OF THE MEAN AND DISPERSION FUNCTIONS OF A SAMPLE OF CURVES WITH BAYESIAN REGRESSION SPLINES

Ce chapitre présente le troisième article rédigé dans le cadre de cette thèse. L'article sera soumis à la revue Journal of the American Statistical Association au mois de septembre 2009.

Abstract

In this paper, we propose a Bayesian model, based on free-knot regression splines, to simultaneously model the mean and dispersion functions of a sample of longitudinal curves, or of a single curve, for which the observations arise from a fairly broad class of continuous distributions that includes those of the exponential family. Modelling both the mean and dispersion functions enhances the flexibility of the model and can prevent inferential difficulties resulting from a misspecification of the statistical distribution of the observations. The class of distributions makes it possible to study several statistical distributions under the same modelling framework and a strategy is put forward to discriminate the most adequate distribution for a given data set. The methodology is applied to a sample of hydrological curves.

6.1. Introduction

The present paper is concerned with simultaneously modelling the mean and dispersion functions of a sample of longitudinal curves using nonparametric regression when no auxiliary information is available ; the framework is also applicable to situations where several individuals' response variable depends nonlinearly on a single predictor and for which there is an interest in the mean and dispersion curves of the group of individuals. Although the methodology is developed for a sample of curves, or individual units, it can also be applied to a single curve. We consider a fairly broad class of continuous distributions from which the data might have arisen and treat the whole problem in a Bayesian setting. Free-knot regression splines are used as the functional modelling tool since they constitute an efficient nonparametric technique and lead to parsimonious models. We also develop a strategy to find the most adequate distribution to model the data.

The problem we address in the present paper is closely related to problems which have been around for a number of years in regression models. This problem is to model the mean behaviour of a process but also to adequately capture the variability of the process about this mean behaviour. In order to put our methodology in perspective, we briefly summarize developments concerning distributional hypotheses and fitting procedures in regression models.

The transformation method, put forward by Box and Cox (1964), consists in transforming the response variable in order for the residuals of a linear model to be roughly homoscedastic. Given an adequate transformation, the residuals of the model applied to the transformed response can then be considered normally distributed with a constant variance and the wealth of results related to the normal linear model can be applied. Generalized Linear Models (GLMs), introduced by Nelder and Wedderburn (1972), treat observations which arise from a distribution belonging to the exponential family in a regression setting. For these observations, this leads to a variance that is a function of the mean, where the function is determined by the distribution considered ; this structural relation between the variance and the mean is therefore fixed once a distribution is chosen.

Carroll and Ruppert (1982, 1988) relax this assumption by considering a broader class of variance functions which include those of the distributions of the exponential family. In order to do so, they elaborate the so-called pseudo-likelihood approach which is essentially a normal distribution where the variance is a function of the mean. They show how this methodology, on an estimation level, is equivalent to the GLM fitting procedure for distributions belonging to the exponential family since both algorithms basically rely on iterative weighted least-squares. Nelder and Pregibon (1987) broaden the scope of GLMs by extending the range of available variance functions associated with the exponential family and by also proposing a method to simultaneously model the mean and dispersion as functions of predictor variables. In the case of the former breakthrough, the estimation problem is solved by generalizing the quasi-likelihood approach developed by Wedderburn (1974), which can be viewed as an estimation technique based on the first two moments of a distribution ; for the latter breakthrough, a new method of estimation based on the extended quasi-likelihood is put forward (see also chapters 9 and 10 of McCullagh and Nelder, 1989). As noted in McCullagh and Nelder (1989), these estimation procedures can be understood in the general framework of estimating equations. Independently, Efron (1986) introduces the double exponential family, not to be confused with the double exponential, or Laplace, distribution. In this paper, he shows how it is possible to simultaneously model the mean and the dispersion in a 'distribution' based framework. Although the author's main motivation seems to be to model over-dispersion for discrete data belonging to the exponential family, he nonetheless indicates how his theory can be viewed as a 'distribution' based justification for the extended quasi-likelihood approach.

The present paper is closely related to the original formulation of GLMs with respect to the family of distributions from which the observations are thought to arise, but the framework developed also considers possible transformations of the data. More specifically, our method is based on a family of continuous distributions which can account for a transformation of the response variable and it does not directly rely on the extended quasi-likelihood methodology to

simultaneously model the mean and dispersion functions, although some parallels can and will be made in what follows.

The paper has several goals. First, it aims to develop a nonparametric modelling context where the distributional hypotheses concerning the observations are not restrictive in the sense that several statistical distributions for continuous data can be considered. This framework makes it possible to study different structural relationships between the variance and the mean which result from the choice of a given statistical distribution. Second, it relaxes the assumption that the dispersion parameter is constant for a given statistical distribution. By modelling both the mean and the dispersion nonparametrically, the flexibility of the model is enhanced and a misspecification of the structural relationship for the variance can be 'rectified' by the dispersion model to obtain adequate confidence bounds for the data. Third, in the Bayesian framework, we are able to propose a methodology to determine the most appropriate models for a given data set.

6.2. Statistical model

For what follows, we suppose that we have a sample of N curves at our disposal and each of these share a common period of time over which the measurements are taken. For a given curve i $(i = 1, \ldots, N)$, we have the data

$$(x_1, y_{i1}), \ldots, (x_j, y_{ij}), \ldots, (x_n, y_{in}),$$

where x_j represents the time associated with the response variable y_{ij}. The x_j's are taken to be fixed and the y_{ij}'s are considered random variables to be modelled. We assume that these random variables are continuous and in the application section, we study the following distributions : normal ($\mathcal{N}$), gamma ($\mathcal{G}$), inverse Gaussian (IG), lognormal ($\mathcal{LN}$) and reciprocal inverse Gaussian (RIG). As indicated in Appendix A, each distribution possesses a specific relation between the mean and the variance, namely a relation of the following type : variance $\propto \{\text{mean}\}^p$, where $p = 0, 1, 2, 3$. Although the application focuses on these five distributions, we present a unified framework to simultaneously model nonparametrically the mean and dispersion behaviours of a curve (or a sample of curves). As will be

seen shortly, it would be possible to consider other statistical distributions which would lead to variance functions of a different kind than the type given above.

The following subsection describes the family of distributions which enables us to consider the aforementioned statistical distributions in a general framework and also indicates how we simultaneously model the mean and dispersion functions. We present the nonparametric systematic components used to model the mean and dispersion functions in the second subsection, as well as the link functions studied. The last three subsections respectively discuss the prior distributions of the parameters, the exploration of the posterior distribution of the knot configurations which define the spline function bases, and the construction of approximate credible sets ; issues of model selection are also addressed in the fourth subsection.

6.2.1. Random component : statistical distribution of the observations

We make the hypothesis that the observations of a given curve i are conditionally independent and that the data, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in})'$, has a joint distribution which belongs to the following family of distributions

$$f_1(\boldsymbol{y}_i|\boldsymbol{\zeta}_i, \boldsymbol{\phi}_i) = \prod_{j=1}^{n} \exp\left\{ \frac{[\zeta_{ij} z_{ij} - \psi(\zeta_{ij})]}{\phi_{ij}} + \kappa(y_{ij}, \phi_{ij}) \right\}, \tag{6.2.1}$$

where z_{ij} is the value of a one-to-one function, $z(\cdot)$, evaluated at y_{ij}, $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{ij}, \ldots, \zeta_{in})'$ represents the vector of canonical parameters, the vector of dispersion parameters is $\boldsymbol{\phi}_i = (\phi_{i1}, \ldots, \phi_{ij}, \ldots, \phi_{in})'$; the functions ψ, κ, and z define the different statistical distributions. As will become clear in the next paragraphs, f_1 is used to indicate the first representation of the joint distribution of the observations. To simplify the notation, we define

$$\eta_{ij} = \eta(y_{ij}, \zeta_{ij}) = \zeta_{ij} z_{ij} - \psi(\zeta_{ij}), \tag{6.2.2}$$

where the canonical parameter is a function of the expected value $\mu_{ij} = \mathbb{E}(z_{ij})$, *i.e.* $\zeta_{ij} = \zeta(\mu_{ij})$, and this function is determined by the statistical distribution.

We also define $\kappa_{ij} = \kappa(y_{ij}, \phi_{ij})$ and assume that this function can be written as

$$\kappa_{ij} \cong -\frac{s_{ij}}{\phi_{ij}} - \frac{1}{2}\log(\phi_{ij}) + t_{ij} - \frac{1}{2}\log(2\pi), \tag{6.2.3}$$

where the symbol $\cong$ is used here, and throughout the paper, to indicate that the expression is exact except for the gamma distribution (see Appendix A for details) ; s_{ij} and t_{ij} are used for the one-to-one functions, $s(\cdot)$ and $t(\cdot)$, evaluated at y_{ij}. The explicit functions for the studied distributions are given in Appendix A.

Using (6.2.2) and (6.2.3), the decomposition of κ_{ij}, expression (6.2.1) becomes

$$f_1(\boldsymbol{y}_i|\boldsymbol{\zeta}_i, \boldsymbol{\phi}_i) \cong \prod_{j=1}^{n} \exp\left\{\frac{\eta_{ij} - s_{ij}}{\phi_{ij}} - \frac{1}{2}\log(\phi_{ij}) + t_{ij} - \frac{1}{2}\log(2\pi)\right\}, \tag{6.2.4}$$

$$\cong \prod_{j=1}^{n} \exp\left\{\frac{1}{2}\left[-\frac{d_{ij}}{\phi_{ij}} - \log(\phi_{ij})\right] + t_{ij} - \frac{1}{2}\log(2\pi)\right\}, \tag{6.2.5}$$

where $d_{ij} = 2\left[s_{ij} - \eta_{ij}\right]$. For the continuous distributions considered here, when $z_{ij} = y_{ij}$, the expression in the exponent of equation (6.2.5) corresponds to the contribution of one observation to the extended quasi-likelihood (see chapter 10 of McCullagh and Nelder, 1989). The distribution of one observation, given in the same equation, belongs to the approximate form of the double exponential family which Efron (1986) used in practice. To get a representation of equation (6.2.5) with the same form as expression (6.2.1), we set $\xi_{ij} = \xi\left(\phi_{ij}\right) = -\phi_{ij}^{-1}$, which leads to

$$f_2(\boldsymbol{y}_i|\boldsymbol{\zeta}_i, \boldsymbol{\xi}_i) \cong \prod_{j=1}^{n} \exp\left\{\frac{\left[\xi_{ij} d_{ij} - \lambda(\xi_{ij})\right]}{2} + t_{ij} - \frac{1}{2}\log(2\pi)\right\}, \tag{6.2.6}$$

with $\lambda(\xi_{ij}) = \log(-\xi_{ij}^{-1})$ and where f_2 is used to indicate the second representation of the joint distribution of the observations. We thus see that both representations given by equations (6.2.1) and (6.2.6) possess similar forms with regards to the expressions in square brackets, *i.e.* the 'kernel' of an exponential family.

In a GLM framework, equation (6.2.1) constitutes a model for $\boldsymbol{\mu}_i = \mathbb{E}\left(\boldsymbol{z}_i\right)$, the vector of expected values of the statistics $\boldsymbol{z}_i$, when the vector of dispersion parameters, $\boldsymbol{\phi}_i$, is fixed. In a similar fashion, if the mean vector $\boldsymbol{\mu}_i$ is fixed, equation (6.2.6) corresponds to a model for $\boldsymbol{\phi}_i = \mathbb{E}\left(\boldsymbol{d}_i\right)$, the vector of expected values of the statistics $\boldsymbol{d}_i$. The components of this vector, the d_{ij}'s, represent the discrepancy

between the statistic z_{ij} (a function of y_{ij}) and μ_{ij} on a distance scale determined by a given statistical distribution (see Table 6.2 in Appendix A). As discussed in Appendix A, the dispersion statistics can be related to the unit deviance function of Jorgensen (1997), the Kullback-Leibler information (Kullback, 1968), and also the standard measure of goodness of fit in GLMs, the deviance of McCullagh and Nelder (1989).

Equation (6.2.1) thus has the structure of a GLM for the mean, while equations (6.2.5) and (6.2.6) possess the structure of a GLM for the dispersion with a gamma random component and a constant dispersion parameter of 2. This duality of the joint distribution indicates a certain symmetry between modelling the mean when the dispersion parameters are known and modelling the dispersion when the mean parameters are known. In the case of the first representation, if $\boldsymbol{\phi}_i$ is known and a model is considered for $\boldsymbol{\zeta}_i$, the maximum likelihood solution for the model is given by maximizing $\sum_{j=1}^{n} \phi_{ij}^{-1} [\zeta_{ij} z_{ij} - \psi(\zeta_{ij})]$, where each ϕ_{ij}^{-1} acts as a weight associated to z_{ij}. For the second representation, assuming $\boldsymbol{\zeta}_i$ is known and a model is considered for $\boldsymbol{\xi}_i$, then the solution is given by maximizing $\sum_{j=1}^{n} [\xi_{ij} d_{ij} - \lambda(\xi_{ij})]$. Although the analysis which follows is aimed at modelling the mean and dispersion functions of a sample of curves, it should be clear from the previous discussion that the methodology could very well be applied to a single curve where the objective would be to model the curve and the dispersion between the mean function model and the curve.

In order to model the mean and dispersion functions of a sample of curves, we take the canonical parameters of the experimental units to be the same at each measurement time. We thus set : $\zeta_{ij} = \zeta_j$ and $\xi_{ij} = \xi_j$ $(\forall i)$, which is equivalent to : $\mu_{ij} = \mu_j$ and $\phi_{ij} = \phi_j$ $(\forall i)$, in terms of the expected values of z_{ij} and d_{ij}. We consider the individual curves to be independent and the joint distribution of all the curves, $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_i', \ldots, \boldsymbol{y}_N')'$, is then given by $f_1(\boldsymbol{y}|\boldsymbol{\zeta}, \boldsymbol{\phi}) = \prod_{i=1}^{N} f_1(\boldsymbol{y}_i|\boldsymbol{\zeta}, \boldsymbol{\phi})$ under the first representation, or $f_2(\boldsymbol{y}|\boldsymbol{\zeta}, \boldsymbol{\xi}) = \prod_{i=1}^{N} f_2(\boldsymbol{y}_i|\boldsymbol{\zeta}, \boldsymbol{\xi})$ under the second representation.

After some algebra, the first representation of the joint distribution leads to

$$f_1(\boldsymbol{y}|\boldsymbol{\zeta},\boldsymbol{\phi}) = \exp\left\{ N\sum_{j=1}^{n} \frac{[\zeta_j \bar{z}_{\cdot j} - \psi(\zeta_j)]}{\phi_j} + \sum_{i,j} \kappa_{ij} \right\}, \tag{6.2.7}$$

where $\bar{z}_{\cdot j} = N^{-1}\sum_{i=1}^{N} z_{ij}$ and we define the vector of these statistics by $\bar{\boldsymbol{z}} = (\bar{z}_{\cdot 1}, \ldots, \bar{z}_{\cdot j}, \ldots, \bar{z}_{\cdot n})'$. The second representation of the joint distribution gives

$$f_2(\boldsymbol{y}|\boldsymbol{\zeta},\boldsymbol{\xi}) \cong \exp\left\{ \frac{N}{2}\sum_{j=1}^{n} \left[\xi_j \bar{d}_{\cdot j} - \lambda(\xi_j)\right] + Nn\left[\bar{t}_{\cdot\cdot} - \frac{1}{2}\log(2\pi)\right] \right\}, \tag{6.2.8}$$

where $\bar{d}_{\cdot j} = 2\left[\bar{s}_{\cdot j} - \bar{\eta}_{\cdot j}\right]$, $\bar{s}_{\cdot j} = N^{-1}\sum_{i=1}^{N} s_{ij}$, $\bar{\eta}_{\cdot j} = N^{-1}\sum_{i=1}^{N} \eta_{ij} = [\zeta_j \bar{z}_{\cdot j} - \psi(\zeta_j)]$, and $\bar{t}_{\cdot\cdot} = (Nn)^{-1}\sum_{i,j} t_{ij}$; for what follows, we write $\bar{\boldsymbol{d}} = (\bar{d}_{\cdot 1}, \ldots, \bar{d}_{\cdot j}, \ldots, \bar{d}_{\cdot n})'$.

6.2.2. Systematic components and link functions

The systematic components associated with the vectors $\bar{\boldsymbol{z}}$ and $\bar{\boldsymbol{d}}$ need to be adaptable in order to capture the shape of the corresponding functions. We use free-knot regression M-spline functions (Ramsay, 1988) for each systematic component because of their flexibility and the fact that they lead to parsimonious models when free-knots are used. The shape of the basis elements depend on the order of the spline functions, the number of interior knots and the location of these knots ; we consider these last two characteristics as random variables. For the systematic component of $\bar{\boldsymbol{z}}$, we write the parameter describing the interior knots as $\boldsymbol{\omega} = (m_1, \boldsymbol{r}_1^{(m_1)})$, where m_1 is the number of interior knots and $\boldsymbol{r}_1^{(m_1)} = (r_{1,1}, ..., r_{m_1,1})$ is the vector of ordered locations of the knots. In a similar fashion, the knot parameter of the systematic component associated with $\bar{\boldsymbol{d}}$ is written as $\boldsymbol{\nu} = (m_2, \boldsymbol{r}_2^{(m_2)})$, where as previously m_2 is the number of interior knots and $\boldsymbol{r}_2^{(m_2)} = (r_{1,2}, ..., r_{m_2,2})$ is the vector of ordered locations of the knots. We write the order of the spline functions used in each case by l_1 and l_2, and we assume that these two quantities are fixed.

The vector of systematic components corresponding to $\bar{\boldsymbol{z}}$ is given by

$$\boldsymbol{u}_{\boldsymbol{\omega}} = \boldsymbol{B}_{\boldsymbol{\omega}}\boldsymbol{\beta}_{\boldsymbol{\omega}}, \tag{6.2.9}$$

where $\boldsymbol{u}_{\boldsymbol{\omega}} = (u_{1,\boldsymbol{\omega}}, \ldots, u_{n,\boldsymbol{\omega}})'$ is a $n \times 1$ vector ; $\boldsymbol{B}_{\boldsymbol{\omega}} = (\boldsymbol{b}_{\boldsymbol{\omega}}(x_1), \ldots, \boldsymbol{b}_{\boldsymbol{\omega}}(x_n))'$, a $n \times K_{\boldsymbol{\omega}}$ matrix, with $\boldsymbol{b}_{\boldsymbol{\omega}}(x_j) = (b_{1,\boldsymbol{\omega}}(x_j), \ldots, b_{K_{\boldsymbol{\omega}},\boldsymbol{\omega}}(x_j))'$, a $K_{\boldsymbol{\omega}} \times 1$ vector of

the basis elements evaluated at x_j ; $\boldsymbol{\beta_\omega} = (\beta_{1,\boldsymbol{\omega}}, \ldots, \beta_{K_{\boldsymbol{\omega}},\boldsymbol{\omega}})'$, a $K_{\boldsymbol{\omega}} \times 1$ vector of parameters. The number of basis elements is given by the sum of the order of the spline functions and the number of interior knots : $K_{\boldsymbol{\omega}} = l_1 + m_1$. For $\bar{\boldsymbol{d}}$, we write the vector of systematic components as

$$\boldsymbol{w_\nu} = \boldsymbol{B_\nu}\boldsymbol{\gamma_\nu}, \tag{6.2.10}$$

where $\boldsymbol{w_\nu} = (w_{1,\boldsymbol{\nu}}, \ldots, w_{n,\boldsymbol{\nu}})'$ is a $n \times 1$ vector ; $\boldsymbol{B_\nu} = (\boldsymbol{b_\nu}(x_1), \ldots, \boldsymbol{b_\nu}(x_n))'$, a $n \times K_{\boldsymbol{\nu}}$ matrix, with $\boldsymbol{b_\nu}(x_j) = (b_{1,\boldsymbol{\nu}}(x_j), \ldots, b_{K_{\boldsymbol{\nu}},\boldsymbol{\nu}}(x_j))'$, a $K_{\boldsymbol{\nu}} \times 1$ vector of the basis elements evaluated at x_j ; $\boldsymbol{\gamma_\nu} = (\gamma_{1,\boldsymbol{\nu}}, \ldots, \gamma_{K_{\boldsymbol{\nu}},\boldsymbol{\nu}})'$, a $K_{\boldsymbol{\nu}} \times 1$ vector of parameters. Here, the number of basis elements is given by : $K_{\boldsymbol{\nu}} = l_2 + m_2$.

The link functions relate $\boldsymbol{\mu} = \mathbb{E}(\bar{\boldsymbol{z}})$ and $\boldsymbol{u_\omega}$ on the one hand, and on the other hand, $\boldsymbol{\phi} = \mathbb{E}(\bar{\boldsymbol{d}})$ and $\boldsymbol{w_\nu}$. We study three link functions in what follows : the identity link (IDL), the logarithmic link (LOL), and the inverse link (INL). For the mean and dispersion functions, we thus have

$$\text{IDL} \quad : \quad \boldsymbol{\mu} = \boldsymbol{u_\omega} \quad \text{and} \quad \boldsymbol{\phi} = \boldsymbol{w_\nu}, \tag{6.2.11}$$

$$\text{LOL} \quad : \quad \log(\boldsymbol{\mu}) = \boldsymbol{u_\omega} \quad \text{and} \quad \log(\boldsymbol{\phi}) = \boldsymbol{w_\nu}, \tag{6.2.12}$$

$$\text{INL} \quad : \quad \boldsymbol{\mu}^{-1} = \boldsymbol{u_\omega} \quad \text{and} \quad \boldsymbol{\phi}^{-1} = \boldsymbol{w_\nu}, \tag{6.2.13}$$

where the functions for LOL and INL are applied componentwise.

6.2.3. Prior distributions

We consider the coefficients of the spline functions, $\boldsymbol{\beta_\omega}$, associated with modelling the mean function, to arise from the following multivariate normal distribution

$$\boldsymbol{\beta_\omega}|\boldsymbol{\gamma_\nu} \sim \mathcal{N}_{K_\omega}\left(\boldsymbol{\beta}^0_{\boldsymbol{\omega}}, \boldsymbol{\Sigma_\omega}\right), \tag{6.2.14}$$

where $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$ and $\boldsymbol{\Sigma_\omega}$ are fixed. As discussed below, we consider the covariance matrix to depend on the dispersion vector $\boldsymbol{\phi}$ and since this vector is modelled by the vector of parameters $\boldsymbol{\gamma_\nu}$, we make this explicit by writing the conditional distribution relative to this vector.

The vector of coefficients for the model of the dispersion function, $\boldsymbol{\gamma_\nu}$, is taken to be distributed as

$$\boldsymbol{\gamma_\nu} \sim \mathcal{N}_{K_\nu}\left(\boldsymbol{\gamma}_{\boldsymbol{\nu}}^0, \boldsymbol{\Sigma_\nu}\right), \tag{6.2.15}$$

with $\boldsymbol{\gamma}_{\boldsymbol{\nu}}^0$ and $\boldsymbol{\Sigma_\nu}$ fixed (see below).

The prior distributions for the parameters which define the elements of the two bases of the spline functions, $\boldsymbol{\omega} = (m_1, \boldsymbol{r}_1^{(m_1)})$ and $\boldsymbol{\nu} = (m_2, \boldsymbol{r}_2^{(m_2)})$, are defined as independent and they are given by

$$\pi(\boldsymbol{\omega}) = \pi(m_1, \boldsymbol{r}_1^{(m_1)}) = \pi(\boldsymbol{r}_1^{(m_1)}|m_1)\pi(m_1), \tag{6.2.16}$$

$$\pi(\boldsymbol{\nu}) = \pi(m_2, \boldsymbol{r}_2^{(m_2)}) = \pi(\boldsymbol{r}_2^{(m_2)}|m_2)\pi(m_2), \tag{6.2.17}$$

where, for $i = \{1, 2\}$, we have $\pi(m_i) \equiv \text{Poisson}(\lambda_i) I_{\{0,1,\ldots,M_i\}}(m_i)$, a truncated Poisson distribution at M_i, and $\pi(\boldsymbol{r}_i^{(m_i)}|m_i) = m_i! M_i^{-m_i}$. The components of $\boldsymbol{r}_i^{(m_i)}$ are thus taken to be the ordered statistics from a uniform discrete distribution on a support set containing M_i possible knot locations. For a given parameter $(m_i, \boldsymbol{r}_i^{(m_i)})$, the prior distributions correspond to the ones used by Denison *et al.* (1998) in the context of modelling the mean with free-knot regression splines for normally distributed data.

The fact that the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$ are considered random and that their support is explored through a reversible jump MCMC algorithm (see next section) makes it difficult to specify the prior location parameters $\boldsymbol{\beta}_{\boldsymbol{\omega}}^0$ and $\boldsymbol{\gamma}_{\boldsymbol{\nu}}^0$ from prior knowledge. Since we have historic data of lesser quality at our disposal, it is used to determine $\boldsymbol{\beta}_{\boldsymbol{\omega}}^0$ and $\boldsymbol{\gamma}_{\boldsymbol{\nu}}^0$ for each combination of $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$ explored by the algorithm. The exact manner in which this is done is given in the application section (section 6.3.2.2). Concerning the prior covariance matrices $\boldsymbol{\Sigma_\omega}$ and $\boldsymbol{\Sigma_\nu}$, we consider their structure to be of the type proposed by Zellner (1986). This kind of prior distribution, named g-prior by the author, was introduced in the treatment of the Bayesian linear model. As first proposed, the covariance matrix depends directly on the design matrix, the dispersion parameter and a multiplicative factor, which Zellner wrote as g and which we write as n_0. More specifically, for a design matrix $\boldsymbol{X}$, the covariance matrix is given by : $\boldsymbol{\Sigma} = \phi\left(n_0 \boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, where ϕ is a constant dispersion parameter. In the case of the normal linear model with

a constant variance, one notices that, except for the presence of n_0, the matrix represents the covariance matrix of the maximum likelihood estimate of the regression coefficients; it is thus fairly direct to interpret n_0 as a weighting factor for the influence of the prior information concerning the regression coefficients (see Zellner, 1986).

In our modelling context, the dispersion parameter of the data is not constant and we therefore need to extend the methodology. In order to do so, we consider a modification of the covariance structure which reduces to the original prescription when the dispersion is constant. For the coefficients of the spline functions which model the mean, we set

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \left(n_{\boldsymbol{\omega}} \boldsymbol{B}'_{\boldsymbol{\omega}} \boldsymbol{\Phi}^{-1} \boldsymbol{B}_{\boldsymbol{\omega}}\right)^{-1}, \tag{6.2.18}$$

where $n_{\boldsymbol{\omega}}$ is a multiplicative factor, $\boldsymbol{B}_{\boldsymbol{\omega}}$ is the design matrix associated with the knot configuration $\boldsymbol{\omega}$ and $\boldsymbol{\Phi} = \text{Diag}\{\phi_j\}$, a diagonal matrix with the dispersion parameters on the diagonal. Since the dispersion parameters are modelled by the systematic components given in equation (6.2.10), which depend on the vector of parameters $\boldsymbol{\gamma}_{\boldsymbol{\nu}}$, this explains the conditional prior distribution of equation (6.2.14).

Concerning the parameters of the coefficients that model the dispersion, we notice that the expression given in equation (6.2.8) possesses a constant dispersion parameter of 2. We therefore define the covariance matrix of these coefficients as

$$\boldsymbol{\Sigma}_{\boldsymbol{\nu}} = 2\left(n_{\boldsymbol{\nu}} \boldsymbol{B}'_{\boldsymbol{\nu}} \boldsymbol{B}_{\boldsymbol{\nu}}\right)^{-1}, \tag{6.2.19}$$

where $n_{\boldsymbol{\nu}}$ is a multiplicative factor and $\boldsymbol{B}_{\boldsymbol{\nu}}$ is the design matrix associated with the knot configuration $\boldsymbol{\nu}$.

6.2.4. Posterior distributions of the knot configurations and model selection

In the Bayesian model, we consider the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$ as random variables. This leads to posterior distributions for these two parameters. Although these distributions are not explicit because of the complexity of their support, it is nonetheless possible to sample from them using a MCMC Reversible Jump

(MCMCRJ) algorithm (Green, 1995). The MCMCRJ algorithm was first used in the context of free-knot regression splines by Denison *et al.* (1998) and later modified by DiMatteo *et al.* (2001). An important contribution of the latter authors was to integrate out the regression parameters in order to substantially simplify the algorithm by alleviating the switching dimension difficulties associated with the MCMCRJ methodology. Furthermore, the algorithm, which uses the marginal distribution of the observations instead of the likelihood in the expression of the acceptance probabilities (see Appendix B), explores the posterior distribution of knot configurations faster since the parameter spaces to be sampled from are considerably reduced.

In the present context, we are thus interested in the partial marginal distribution $\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$ for which the two vectors $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ and $\boldsymbol{\gamma}_{\boldsymbol{\nu}}$ have been integrated out. This marginal distribution does not possess a closed form with the distributional assumptions outlined above and we are thus led to consider some approximations.

6.2.4.1. *Marginal distribution approximations*

The first approximation is based on the Schwarz information criterion (Schwarz, 1978) and it is given by

$$\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu}) = (Nn)^{-\frac{(K_{\boldsymbol{\omega}}+K_{\boldsymbol{\nu}})}{2}}\exp\Bigg\{\frac{N}{2}\sum_{j=1}^{n}\left[\widehat{\xi}_j\widehat{\bar{d}}_{\cdot j} - \lambda(\widehat{\xi}_j)\right] + Nn\left[\bar{t}_{\cdot\cdot} - \frac{1}{2}\log(2\pi)\right]\Bigg\}, \quad (6.2.20)$$

where $\widehat{\xi}_j$ and $\widehat{\bar{d}}_{\cdot j}$ are evaluated at the maximum likelihood estimates $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\nu}}$ (see Appendix C). Although the Schwarz information criterion was originally derived in a Bayesian setting, it relies on an asymptotic argument and it therefore does not take into account the specific form of the prior distributions. It should be noted though that Kass and Wasserman (1995), Pauler (1998), and DiMatteo *et al.* (2001) justify the use of this approximation to the marginal distribution through a normal prior distribution for the regression parameters with a specific covariance structure.

Although we have defined a fully Bayesian model in this section, we want to study the performance of this approximation for pragmatic reasons. This approximation is fairly easy to calculate since it only relies on the maximum likelihood estimates which can be obtained quite directly from various statistical packages; furthermore, it does not require any specification for the prior distributions of the regression parameters $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ and $\boldsymbol{\gamma}_{\boldsymbol{\nu}}$, which can be useful when no prior information is available.

The second approximation considered is based on the basic Laplace approximation to an integral (see Appendix C for details) and it is given by

$$\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu}) = \left(\frac{|\boldsymbol{\Sigma}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}|}{|\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}}||\boldsymbol{\Sigma}_{\boldsymbol{\nu}}|}\right)^{1/2} \exp\left\{\frac{N}{2}\sum_{j=1}^{n}\left[\xi^*_j\bar{d}^*_{\cdot j} - \lambda(\xi^*_j)\right] + Nn\left[\bar{t}_{\cdot\cdot} - \frac{1}{2}\log(2\pi)\right] - \frac{1}{2}q\left(\boldsymbol{\theta}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}\right)\right\}, \quad (6.2.21)$$

where we now have defined $\boldsymbol{\theta}_{\boldsymbol{\omega},\boldsymbol{\nu}} = (\boldsymbol{\beta}'_{\boldsymbol{\omega}}, \boldsymbol{\gamma}'_{\boldsymbol{\nu}})'$ and all starred quantities are evaluated at

$$\boldsymbol{\theta}^*_{\boldsymbol{\omega},\boldsymbol{\nu}} = \mathrm{argmax}_{\boldsymbol{\theta}_{\boldsymbol{\omega},\boldsymbol{\nu}}}\left\{\frac{N}{2}\sum_{j=1}^{n}\left[\xi_j\bar{d}_{\cdot j} - \lambda(\xi_j)\right] - \frac{1}{2}q\left(\boldsymbol{\theta}_{\boldsymbol{\omega},\boldsymbol{\nu}}\right) - \frac{1}{2}\log\left(|\boldsymbol{\Sigma}_{\boldsymbol{\omega}}|\right)\right\}, \quad (6.2.22)$$

with $q\left(\boldsymbol{\theta}_{\boldsymbol{\omega},\boldsymbol{\nu}}\right) = (\boldsymbol{\beta}_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}})'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\omega}}(\boldsymbol{\beta}_{\boldsymbol{\omega}} - \boldsymbol{\beta}^0_{\boldsymbol{\omega}}) + (\boldsymbol{\gamma}_{\boldsymbol{\nu}} - \boldsymbol{\gamma}^0_{\boldsymbol{\nu}})'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\nu}}(\boldsymbol{\gamma}_{\boldsymbol{\nu}} - \boldsymbol{\gamma}^0_{\boldsymbol{\nu}})$. The other quantities are given by

$$\boldsymbol{\Sigma}^*_{\boldsymbol{\omega},\boldsymbol{\nu}} = \left\{\boldsymbol{\Lambda}^*_{\boldsymbol{\omega},\boldsymbol{\nu}} + \mathrm{Diag}\left(\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}}\right)^{-1} + \boldsymbol{\Upsilon}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}\right\}^{-1}, \quad (6.2.23)$$

$$\begin{aligned}\boldsymbol{\Lambda}^*_{\boldsymbol{\omega},\boldsymbol{\nu}} &= -N\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}_{\boldsymbol{\omega},\boldsymbol{\nu}}\partial\boldsymbol{\theta}'_{\boldsymbol{\omega},\boldsymbol{\nu}}}\sum_{j=1}^{n}\frac{1}{2}\left[\xi_j\bar{d}_{\cdot j} - \lambda(\xi_j)\right]\right\}_{\boldsymbol{\theta}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}} \\ &= N\boldsymbol{B}'_{\boldsymbol{\omega},\boldsymbol{\nu}}\boldsymbol{\Delta}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}\boldsymbol{B}_{\boldsymbol{\omega},\boldsymbol{\nu}}, \quad (6.2.24)\end{aligned}$$

$$\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}} = \left\{n_{\boldsymbol{\omega}}\boldsymbol{B}'_{\boldsymbol{\omega}}\mathrm{Diag}\left(\phi^*_j\right)^{-1}\boldsymbol{B}_{\boldsymbol{\omega}}\right\}^{-1}, \quad (6.2.25)$$

$$\boldsymbol{B}_{\boldsymbol{\omega},\boldsymbol{\nu}} = \mathrm{Diag}\left(\boldsymbol{B}_{\boldsymbol{\omega}}, \boldsymbol{B}_{\boldsymbol{\nu}}\right). \quad (6.2.26)$$

If we study the structure of the matrix $\boldsymbol{\Sigma}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}$, we see that $\boldsymbol{\Lambda}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}$ is minus the hessian matrix of the log-likelihood evaluated at $\boldsymbol{\theta}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}$. The prior covariance matrices

appear in the block diagonal matrix Diag $(\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^{-1}$, where $\boldsymbol{\Sigma}^*_{\boldsymbol{\omega}}$ which depends on the dispersion parameter is evaluated at $\boldsymbol{\gamma}^*_{\boldsymbol{\nu}}$, and finally $\boldsymbol{\Upsilon}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}$ is a matrix of second derivatives, also evaluated at $\boldsymbol{\gamma}^*_{\boldsymbol{\nu}}$, which takes into account the dependency of $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ relative to the dispersion vector $\boldsymbol{\phi}$ (see section 6.2.3).

It is interesting to note the similarities and differences between the two approximations $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$ and $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$. The two terms in the first square bracket of the exponent in equations (6.2.20) and (6.2.21) are identical except for the value of the parameter spaces where they are evaluated, *i.e.* the maximum likelihood estimates and the maximum *a posteriori* estimates; the two terms in the second square bracket are identical, while the last term in the exponent of $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$, $q\left(\boldsymbol{\theta}^*_{\boldsymbol{\omega},\boldsymbol{\nu}}\right)$, captures the discrepancy between the prior location parameters and the posterior maxima. Concerning the first factor in equations (6.2.20) and (6.2.21), it represents a term that takes into account the dimensionality of the model, directly via the number of parameters in the case of $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$, and through the determinants in the case of $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$.

6.2.4.2. *Knot configuration selection*

The MCMCRJ algorithm, detailed in Appendix B, explores the space of the model parameters, $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$, and reaches a stationary distribution once convergence is attained. The modes, $\boldsymbol{\omega}^\dagger$ and $\boldsymbol{\nu}^\dagger$, of the converged chain are used as 'optimal' knot configurations and they constitute our selected knot configurations. We choose to select specific knot configurations for operational purposes, *i.e.* to have a model in the parameter space of the spline coefficients. It should be noted though that it would be possible to average the output of the MCMCRJ algorithm in a fashion similar to that of DiMatteo *et al.* (2001).

6.2.4.3. *Selection of adequate statistical distribution and link functions*

We use ratios of marginal distributions evaluated at the modes $\boldsymbol{\omega}^\dagger$ and $\boldsymbol{\nu}^\dagger$ of the converged chains to compare the set of distributions and link functions given in sections 6.2.1 and 6.2.2. This approach is similar to Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995) which compare two models through the ratio of their marginal distributions, and which can be interpreted as the weight of evidence in

favour of a model relative to another. Strictly speaking, in our modelling context, we would need to integrate over the spaces of $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$ to get the full marginal distribution and then calculate the Bayes factors. We nonetheless think that the modal evaluation of the marginal distribution supplies reasonable information concerning the adequacy of a given model.

In order to indicate that the ratios are evaluated at the modal quantities of $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$, we write

$$BF^{\dagger}(A,B) = \frac{\mathrm{m}^{(A)}(\boldsymbol{y}|\boldsymbol{\omega}^{\dagger},\boldsymbol{\nu}^{\dagger})}{\mathrm{m}^{(B)}(\boldsymbol{y}|\boldsymbol{\omega}^{\dagger},\boldsymbol{\nu}^{\dagger})}, \tag{6.2.27}$$

where $\mathrm{m}^{(A)}(\boldsymbol{y}|\boldsymbol{\omega}^{\dagger},\boldsymbol{\nu}^{\dagger})$ and $\mathrm{m}^{(B)}(\boldsymbol{y}|\boldsymbol{\omega}^{\dagger},\boldsymbol{\nu}^{\dagger})$ are the marginal distributions of the two models A and B, evaluated at their respective modes; a statistical distribution and two link functions determine each of these models. In the application section, we study these ratios with the two marginal distribution approximations given in section 6.2.4.1.

6.2.5. Function estimation and approximate credible sets

The Bayesian model which uses $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$ yields approximate posterior distributions since using Laplace's approximation is equivalent to assume that the posterior distribution of the spline function coefficients are normally distributed (see Robert, 2001, and Appendix C). The approximate posterior distribution of $\boldsymbol{\theta}_{\boldsymbol{\omega},\boldsymbol{\nu}} = (\boldsymbol{\beta}'_{\boldsymbol{\omega}}, \boldsymbol{\gamma}'_{\boldsymbol{\nu}})'$ is

$$\boldsymbol{\theta}_{\boldsymbol{\omega},\boldsymbol{\nu}}|\boldsymbol{y} \sim \mathcal{N}_{K_{\boldsymbol{\omega}}+K_{\boldsymbol{\nu}}}\left(\boldsymbol{\theta}^{*}_{\boldsymbol{\omega},\boldsymbol{\nu}}, \boldsymbol{\Sigma}^{*}_{\boldsymbol{\omega},\boldsymbol{\nu}}\right), \tag{6.2.28}$$

where $\boldsymbol{\theta}^{*}_{\boldsymbol{\omega},\boldsymbol{\nu}}$ and $\boldsymbol{\Sigma}^{*}_{\boldsymbol{\omega},\boldsymbol{\nu}}$ are defined in equations (6.2.22) and (6.2.23) respectively.

For what follows, we drop the knot configuration parameters $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$, and we use $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)' = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$; we now write $\dim(\boldsymbol{\theta}_i) = K_i$, for $i = 1, 2$, and $K = K_1 + K_2$. The full matrix $\boldsymbol{\Sigma}^{*}$ is partitioned in the usual manner with matrix elements $\boldsymbol{\Sigma}^{*}_{11}$, $\boldsymbol{\Sigma}^{*}_{12}$, $\boldsymbol{\Sigma}^{*}_{21}$, $\boldsymbol{\Sigma}^{*}_{22}$, which are respectively of dimension $(K_1 \times K_1)$, $(K_1 \times K_2)$, $(K_2 \times K_1)$, and $(K_2 \times K_2)$. Furthermore, the basis elements associated with $\boldsymbol{\theta}_i$ are written $\boldsymbol{b}_i(\cdot)$, the inverse of the link function for $\boldsymbol{\theta}_1$ is given by $g(\cdot)$ and that of the link function for $\boldsymbol{\theta}_2$ is given by $h(\cdot)$.

With this notation, the different approximate posterior distributions are given by

$$\boldsymbol{\theta}|\boldsymbol{y} \sim \mathcal{N}_K\left(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}^*\right), \tag{6.2.29}$$

$$\boldsymbol{\theta}_1|\boldsymbol{y} \sim \mathcal{N}_{K_1}\left(\boldsymbol{\theta}_1^*, \boldsymbol{\Sigma}_{11}^*\right), \tag{6.2.30}$$

$$\boldsymbol{\theta}_2|\boldsymbol{y} \sim \mathcal{N}_{K_2}\left(\boldsymbol{\theta}_2^*, \boldsymbol{\Sigma}_{22}^*\right). \tag{6.2.31}$$

6.2.5.1. *Function estimation*

We model the mean function, $\bar{\boldsymbol{z}}$, with the systematic component $\boldsymbol{u}$, and the dispersion function, $\bar{\boldsymbol{d}}$, with the systematic component $\boldsymbol{w}$. Given the previous notation, we can write

$$\mathbb{E}\left(\bar{z}_{\cdot j}|\boldsymbol{\theta}_1\right) = g\left(u_j\right), \tag{6.2.32}$$

$$\mathbb{E}\left(\bar{d}_{\cdot j}|\boldsymbol{\theta}_2\right) = h\left(w_j\right), \tag{6.2.33}$$

where $u_j = \boldsymbol{b}_1(x_j)'\boldsymbol{\theta}_1$ and $w_j = \boldsymbol{b}_2(x_j)'\boldsymbol{\theta}_2$. If the link functions are not identity links, the models for $\bar{z}_{\cdot j}$ and $\bar{d}_{\cdot j}$ are not linear functions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. To approximate the inverse of a link function, we rely on a Taylor series expansion of order 1. We thus have

$$g_j \approx g_j^* + \dot{\boldsymbol{g}}_j'\left(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1^*\right), \tag{6.2.34}$$

$$h_j \approx h_j^* + \dot{\boldsymbol{h}}_j'\left(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_2^*\right), \tag{6.2.35}$$

where $g_j = g\left(u_j\right)$, $g_j^* = g\left(\boldsymbol{b}_1(x_j)'\boldsymbol{\theta}_1^*\right)$, $\dot{\boldsymbol{g}}_j = \left\{\frac{\partial g_j}{\partial \boldsymbol{\theta}_1}\right\}_{\boldsymbol{\theta}_1^*}$, $h_j = h\left(w_j\right)$, $h_j^* = h\left(\boldsymbol{b}_2(x_j)'\boldsymbol{\theta}_2^*\right)$, $\dot{\boldsymbol{h}}_j = \left\{\frac{\partial h_j}{\partial \boldsymbol{\theta}_2}\right\}_{\boldsymbol{\theta}_2^*}$. Taking the expectations of the expressions given in (6.2.34) and (6.2.35) relative to the posterior distributions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively (given in equations (6.2.30) and (6.2.31)), leads to

$$\mathbb{E}\left(\bar{z}_{\cdot j}\right) \approx g_j^* \qquad \text{and} \qquad \mathbb{E}\left(\bar{d}_{\cdot j}\right) \approx h_j^* \,. \tag{6.2.36}$$

The first expression can be used to model $\bar{\boldsymbol{y}}$ for the $\mathcal{N}$ (normal), $\mathcal{G}$ (gamma), and IG (inverse Gaussian) distributions since we then have $\bar{\boldsymbol{z}} = \bar{\boldsymbol{y}}$. For the $\mathcal{LN}$ (lognormal) and RIG (reciprocal inverse Gaussian) distributions, it corresponds to a model for the mean function on the transformed scale. The second expression

can be used directly for all the distributions since it represents a model for the jth component of the dispersion vector $\boldsymbol{\phi}$.

If we seek a model on the untransformed scale for the $\mathcal{LN}$ and RIG distributions, we need to rely on the equations for the mean given in Table 6.2 of Appendix A. These expressions depend on both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ since they are functions of the mean on the transformed scale, $\boldsymbol{\mu}$, and the dispersion parameters, $\boldsymbol{\phi}$. Writing the function for the mean on the untransformed scale as $e(\cdot,\cdot)$, we have

$$\mathbb{E}\left(\bar{y}_{\cdot j}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right) = e\left(g_j,h_j\right), \tag{6.2.37}$$

where $g_j = \mathbb{E}\left(\bar{z}_{\cdot j}|\boldsymbol{\theta}_1\right)$ and $h_j = \mathbb{E}\left(\bar{d}_{\cdot j}|\boldsymbol{\theta}_2\right)$, as above. Using a Taylor series expansion of order 1 again, we obtain

$$e_j \approx e_j^* + \dot{\boldsymbol{e}}_j'\left\{\left(\boldsymbol{\theta}_1-\boldsymbol{\theta}_1^*\right)',\left(\boldsymbol{\theta}_2-\boldsymbol{\theta}_2^*\right)'\right\}', \tag{6.2.38}$$

where $e_j = e\left(g_j,h_j\right)$, $e_j^* = e\left(g_j^*,h_j^*\right)$, and $\dot{\boldsymbol{e}}_j = \left\{\left(\frac{\partial e_j}{\partial g_j}\right)\left(\frac{\partial g_j}{\partial \boldsymbol{\theta}_1}\right)', \left(\frac{\partial e_j}{\partial h_j}\right)\left(\frac{\partial h_j}{\partial \boldsymbol{\theta}_2}\right)'\right\}'_{\boldsymbol{\theta}^*}$. Finally, we get : $\mathbb{E}\left(\bar{y}_{\cdot j}\right) \approx e_j^*$, by taking the expectation relative to the posterior distribution of $\boldsymbol{\theta}$ given in equation (6.2.29).

6.2.5.2. *Approximate credible sets*

Following the discussion of the previous section, it is possible to construct approximate credible sets for the sample of curves and also the dispersion function. The former is of obvious interest and the latter can be used to study if the hypothesis of a constant dispersion parameter is a reasonable assumption. We can get expressions for approximate credible sets from the Taylor expansions given in equations (6.2.34), (6.2.35), and (6.2.38).

Simultaneous, for all j's, approximate $100(1-\delta)\%$ credible sets for $\mathbb{E}\left(\bar{z}_{\cdot j}\right)$ are obtained from equation (6.2.34) and the posterior distribution of $\boldsymbol{\theta}_1$. They are given by

$$g_j^* \pm \left\{\dot{\boldsymbol{g}}_j'\boldsymbol{\Sigma}_{11}^*\dot{\boldsymbol{g}}_j\chi^2_{K_1}(\delta)\right\}^{1/2}, \tag{6.2.39}$$

where $\chi^2_{K_1}(\delta)$ represents the $100(1-\delta)$th percentile of a χ^2 distribution with degrees of freedom K_1. These credible sets can be used for the mean of the $\mathcal{N}$, $\mathcal{G}$, and IG distributions on the untransformed scales, and the mean of the $\mathcal{LN}$ and RIG distributions on the transformed scales.

In a similar fashion for $\mathbb{E}\left(\bar{d}_{\cdot j}\right)$, we have

$$h_j^* \pm \left\{\dot{\boldsymbol{h}}_j' \boldsymbol{\Sigma}_{22}^* \dot{\boldsymbol{h}}_j \chi_{K_2}^2(\delta)\right\}^{1/2}. \tag{6.2.40}$$

Finally, for the $\mathcal{LN}$ and RIG distributions, the credible sets of the mean on the untransformed scale are given by

$$e_j^* \pm \left\{\dot{\boldsymbol{e}}_j' \boldsymbol{\Sigma}^* \dot{\boldsymbol{e}}_j \chi_K^2(\delta)\right\}^{1/2}. \tag{6.2.41}$$

6.3. Application

The time series of water flows is of particular interest in hydrology. This data needs to be modelled effectively and accurately to make decisions concerning water management, and to prevent environmental and human casualties which could happen as a consequence of extreme events. A hydrograph is defined as the functional representation of a sequence of water flows, and Figure 6.1 shows a sample of yearly hydrographs with weekly measurements from a watershed in northern Québec. The yearly hydrographs have been landmark registered (Ramsay and Silverman, 2005) in order to make them similar relative to the time at which important features happen. Since the typical shape of a watershed's hydrographs and the variability about this typical hydrograph are often of interest for various operational purposes, we use the statistical model developed in the previous section to capture the average and dispersion behaviours of the sample of yearly hydrographs shown in Figure 6.1.

As noted in section 6.2, we are able to study continuous distributions with different structural relations between the variance and the mean in our general modelling framework. Several distributions considered here have been used in the past to model water flow data. For example, the gamma ($\mathcal{G}$) distribution is discussed in this context by Gumbel (1958), and Bobée and Ashkar (1991); the lognormal ($\mathcal{LN}$) distribution in Gumbel (1958), and Rao and Hamed (2000); and the inverse Gaussian (IG) distribution was used for this purpose by Folks and Chhikara (1978). It needs to be noted that the inverse Gaussian (IG) and reciprocal inverse Gaussian (RIG) distributions are particular cases of Halphen's laws which were developed to model water flow data (see Perreault *et al.* (1999a)

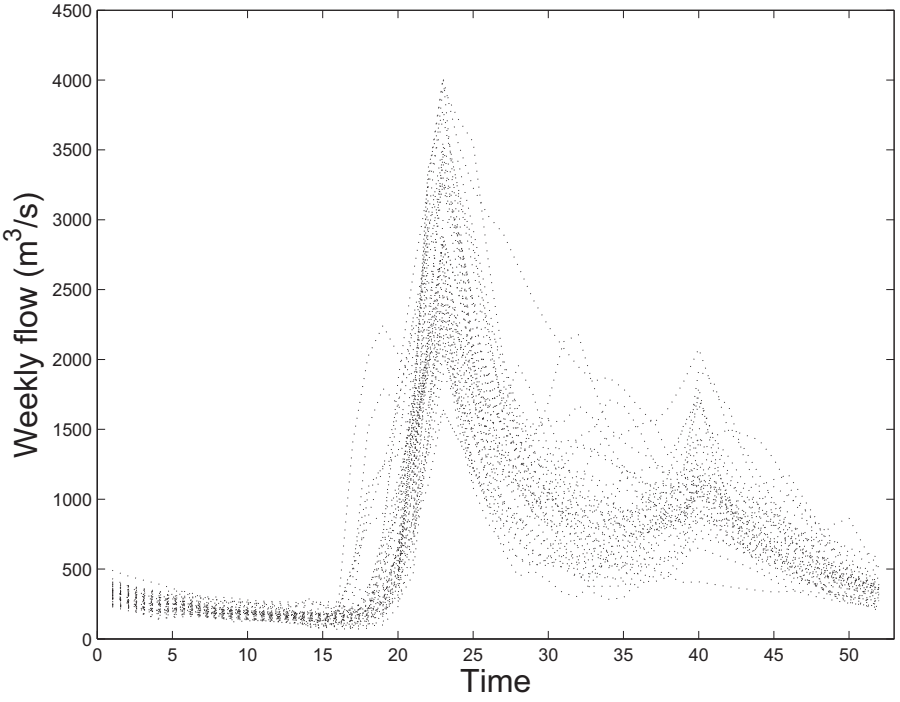


FIGURE 6.1. A sample of 42 yearly landmark registered hydrographs with weekly measurements from a watershed situated in northern Québec.

for a thorough discussion of Halphen's laws and Perreault *et al.* (1999b) for an application in hydrology). As mentioned in section 6.2, we also study the normal ($\mathcal{N}$) distribution since it is often used in smoothing problems.

6.3.1. Hydrological data

We want to model the mean and dispersion functions of the sample of 42 yearly hydrographs shown in Figure 6.1. The full sample at our disposal contains 53 yearly hydrographs covering the period which extends from 1950 to 2002. It is known that the first 11 annual hydrographs are of lesser quality since they have been reconstructed from regional information, and we use these to specify the prior distributions (see below). The effective sample of curves is thus made up of the 42 yearly hydrographs shown in Figure 6.1. The first data point of each curve represents the measurement recorded on the first week of January, while the last one corresponds to the measurement made on the last week of December. The first weeks of the hydrographs exhibit a steady decline during the winter period,

and the hydrographs show little variability. With the advent of warmer temperatures, spring floods start and a strong increase in water flow is observed when the accumulated snow begins to melt. A global maximum characterizes a spring flood but secondary peaks resulting from heavy rainfalls during this time of the year can also be present. In Figure 6.1, the global maxima of the spring floods happen simultaneously since the hydrographs have been landmark registered relative to this event. Even if the spring floods all share a common global structure, the variability of water flows across hydrographs is high for this period. After the spring flood ends, water flow is mostly due to rainfall during summer and autumn. Similarly to the spring maxima, the autumn maxima happen simultaneously on the figure since these features were also used to perform the registration. During the summer and autumn weeks, the variability across hydrographs is also important because of the random nature of important rainfall events. At the end of the year, water flow starts decreasing again with the beginning of winter.

6.3.2. Model specifications

6.3.2.1. *Spline bases*

In section 6.2.2, it was pointed out that the modelling bases made up of spline functions are fully defined by the order of these functions, the number of interior knots and the position of these knots. The last two quantities are treated with the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$ for the basis which models the mean and the one that models the dispersion respectively. To apply our method, we now need to fix the order of the spline functions. We choose to work with M-spline functions of order 3 ($l_1 = l_2 = 3$) which means that the basis elements are quadratic by parts.

6.3.2.2. *Prior distributions*

As discussed in section 6.2.4.1, if one uses the marginal distribution approximation based on the Schwarz criterion, it is not necessary to specify prior distributions in order to explore knot configurations through the MCMCRJ algorithm.

Since we have data of lesser quality which contains information about the curves we wish to model (see section 6.3.1), we can specify the parameters of the

prior distributions. With the same notation as in section 6.2.1, the vector made up of the $N_0 = 11$ yearly hydrographs of lesser quality is written as $\boldsymbol{y}^0$. The prior location parameters, $\boldsymbol{\beta}^0_{\boldsymbol{\omega}}$ and $\boldsymbol{\gamma}^0_{\boldsymbol{\nu}}$, are obtained by maximizing $\sum_{j=1}^{n} \left[\xi_j \bar{d}^0_{\cdot j} - \lambda(\xi_j)\right]$, where $\bar{d}^0_{\cdot j} = 2\left[\bar{s}^0_{\cdot j} - \bar{\eta}^0_{\cdot j}\right]$, $\bar{s}^0_{\cdot j} = N^{-1}\sum_{i=1}^{N} s^0_{ij}$, $\bar{\eta}^0_{\cdot j} = N^{-1}\sum_{i=1}^{N} \eta^0_{ij} = \left[\zeta_j \bar{z}^0_{\cdot j} - \psi(\zeta_j)\right]$.

For the covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\nu}}$, we only need to specify the multiplying factors $n_{\boldsymbol{\omega}}$ and $n_{\boldsymbol{\nu}}$ (see section 6.2.3). These two factors can roughly be interpreted as the weight the prior distributions have in the Bayesian estimation of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ and $\boldsymbol{\gamma}_{\boldsymbol{\nu}}$. We thus consider them to be of the following form : $n_{\boldsymbol{\omega}} = c_{\boldsymbol{\omega}} N_0$ and $n_{\boldsymbol{\nu}} = c_{\boldsymbol{\nu}} N_0$, where $c_{\boldsymbol{\omega}}$ and $c_{\boldsymbol{\nu}}$ vary between 0 and 1. The two limiting cases, 0 and 1, respectively represent non-informative priors and priors for which the N_0 curves each contribute as much as a curve in the effective sample. For the numerical implementation, we take $c_{\boldsymbol{\omega}} = 1/N_0$ and $c_{\boldsymbol{\nu}} = 1/(N_0 n)$, therefore the average of the N_0 curves contributes the weight of one curve concerning the parameters to model the mean, while it contributes the weight of one observation for the parameters that model the dispersion. The reason for this asymmetry is that the curves are registered according to their average behaviour, and the dispersion of the prior sample might not necessarily resemble that of the effective sample.

The number of possible knot positions for the mean and dispersion models, M_i $(i = 1, 2)$, and the potential knot positions need to be specified in the prior distributions of equations (6.2.16) and (6.2.17). Here, we take M_i as twice the number of data points and the potential knot positions are taken to be equally spaced throughout the domain of the data. A proximity constraint for the interior knots is also applied in the same spirit as in Denison *et al.* (1998). The knots are constrained in order to have at least one data point between them.

6.3.3. Results

The MCMCRJ algorithms are implemented with the two marginal distribution approximations, $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu})$ and $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu})$. For a given distribution, two link functions, and a marginal distribution approximation, the MCMCRJ algorithm is run for 10000 iterations and we consider the first 5000 iterations to be the

burn-in period, while the last 5000 iterations are used to determine the modal knot configurations, $\boldsymbol{\omega}^{\dagger}$ and $\boldsymbol{\nu}^{\dagger}$.

TABLE 6.1. Comparison of different models according to the logarithm of expression (6.2.27), where the reference model for all the calculations in a column is $B = \{\mathcal{N}, (\text{ IDL, LOL})\}$. In parentheses, the modal number of interior knots for the mean and dispersion models are given.

	$\log(BF_a^{\dagger})$ $(m_1^{\dagger}, m_2^{\dagger})$	$\log(BF_b^{\dagger})$ $(m_1^{\dagger}, m_2^{\dagger})$
$\mathcal{N}$ (IDL, LOL)	0 (8, 5)	0 (9, 9)
$\mathcal{G}$ (LOL, INL)	192 (5, 5)	208 (12, 12)
IG (INL, INL)	191 (5, 6)	137 (8, 10)
$\mathcal{LN}$ (IDL, INL)	205 (6, 4)	215 (12, 12)
RIG (IDL, LOL)	195 (7, 4)	112 (5, 5)

Table 6.1 presents the logarithm of the ratio of the partial marginal distributions evaluated at the modes of the corresponding chains, and also the modal number of interior knots. For a given column, the models are compared to the reference model which is taken to be $B = \{\mathcal{N}, (\text{ IDL, LOL})\}$, in equation (6.2.27). Since all the values are positive, the magnitude of a value indicates the improvement in quality of a model compared to the reference model. It is interesting to note that, in general, the use of the first marginal distribution approximation, the one based on the Schwarz criterion, leads to models which are more parsimonious than when the Laplace approximation is used. Although the three link functions given in section 6.2.2 were implemented for the two systematic components, Table 6.1 only gives the results of the combinations of link functions which gave the best results for a given statistical distribution. The quality of the results were judged by the second approximation, which involves the covariance matrices, since the first approximation can not discriminate for link functions because of the flexibility of the spline functions (technical report by the same authors).

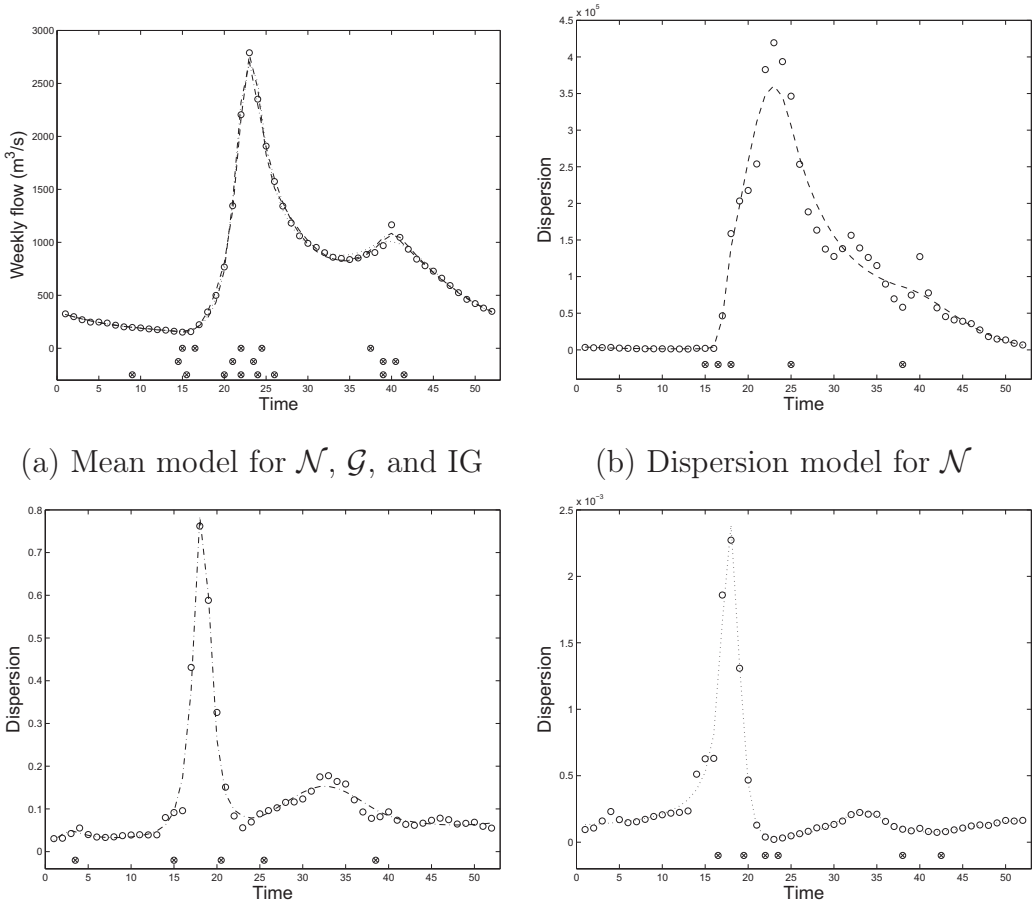
According to both columns, the best distribution is the $\mathcal{LN}$ distribution which possesses a variance that depends quadratically on the mean (see Appendix A).

The quantity $BF_a^\dagger$ does not seem to discriminate substantially between the $\mathcal{G}$, IG and RIG distributions, although all of them seem to perform better than the $\mathcal{N}$ distribution. The quantity $BF_b^\dagger$ gives results which are more nuanced, placing the $\mathcal{LN}$ and $\mathcal{G}$ distributions at the top, the IG and RIG distributions in the middle and the $\mathcal{N}$ distribution at the bottom. This quantity thus places the quadratic dependency of the variance relative to the mean as the most adequate relation, and in general, it seems to indicate that the variance needs to depend on the mean to properly model the data.

6.3.3.1. *Function Estimation*

Figure 6.2 presents the simultaneous models for the mean and dispersion functions under the hypothesis that the data come from the normal ($\mathcal{N}$), gamma ($\mathcal{G}$), and inverse Gaussian (IG) distributions; here, the MCMCRJ algorithm was applied using the first marginal distribution approximation, $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$. Panel (a) shows the mean models for the three distributions, while panels (b), (c), and (d) give the dispersion models for the different distributions. The observed weekly mean and dispersion functions are shown as circles, while the respective modal knot configurations, $\boldsymbol{\omega}^\dagger$ and $\boldsymbol{\nu}^\dagger$, are indicated by crossed circles. In panel (a), we see that the mean models for the different distributions are all very similar. The dispersion models for the $\mathcal{G}$ and IG distributions, shown in panels (c) and (d), appear to reproduce well the characteristics of the observed weekly dispersion functions; the dispersion model for the $\mathcal{N}$ distribution, given in panel (b), captures the global shape of the observed dispersions but does not reproduce all the observed features.

For the same distributions studied in Figure 6.2, Figure 6.3 presents the simultaneous models when the second marginal distribution approximation, $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$, is used in the MCMCRJ algorithm. As in Figure 6.2, the different mean models, given in panel (a), seem to reproduce well the observed weekly averages. The dispersion model for the $\mathcal{N}$ distribution, shown in panel (b), appears to better capture the main features of the observed weekly dispersions than the same model of Figure 6.2; the opposite can be said for the dispersion model of the IG



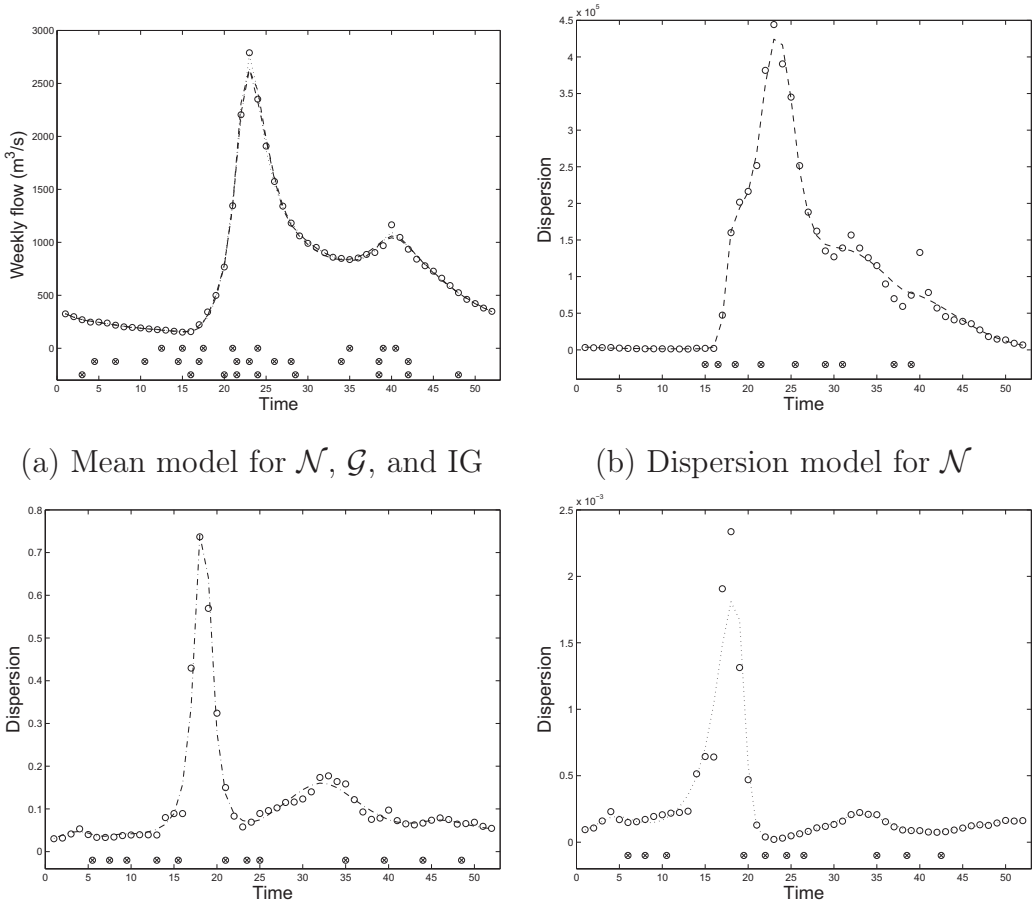
(a) Mean model for $\mathcal{N}$, $\mathcal{G}$, and IG

(b) Dispersion model for $\mathcal{N}$

(c) Dispersion model for $\mathcal{G}$

(d) Dispersion model for IG

FIGURE 6.2. Models obtained with the use of $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu})$ in the MCMCRJ algorithm. Panel (a) : mean function models for the $\mathcal{N}$ (dashed), $\mathcal{G}$ (dot-dashed), and IG (dotted) distributions. The open circles represent the weekly average of the sample of curves, while the crossed circles at the bottom represent the modal knot configurations ($\boldsymbol{\omega}^\dagger$) corresponding to, from bottom to top, the $\mathcal{N}$, $\mathcal{G}$, and IG distributions respectively. Panels (b), (c), and (d) give the weekly dispersion of the sample of curves (open circles) and dispersion models corresponding to the different statistical distributions with the same line types as in (a). The crossed circles at the bottom represent the modal knot configurations ($\boldsymbol{\nu}^\dagger$).



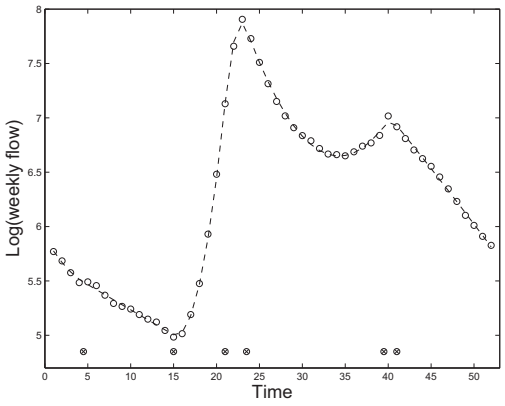
(a) Mean model for $\mathcal{N}$, $\mathcal{G}$, and IG

(b) Dispersion model for $\mathcal{N}$

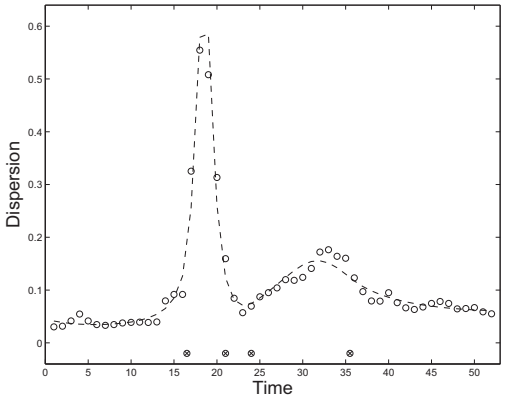
(c) Dispersion model for $\mathcal{G}$

(d) Dispersion model for IG

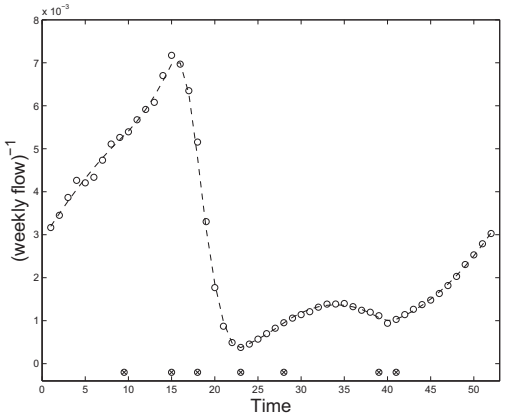
FIGURE 6.3. Models obtained with the use of $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu})$ in the MCMCRJ algorithm. Panel (a) : mean function models for the $\mathcal{N}$ (dashed), $\mathcal{G}$ (dot-dashed), and IG (dotted) distributions. The open circles represent the weekly average of the sample of curves, while the crossed circles at the bottom represent the modal knot configurations ($\boldsymbol{\omega}^\dagger$) corresponding to, from bottom to top, the $\mathcal{N}$, $\mathcal{G}$, and IG distributions respectively. Panels (b), (c), and (d) give the weekly dispersion of the sample of curves (open circles) and dispersion models corresponding to the different statistical distributions with the same line types as in (a). The crossed circles at the bottom represent the modal knot configurations ($\boldsymbol{\nu}^\dagger$).



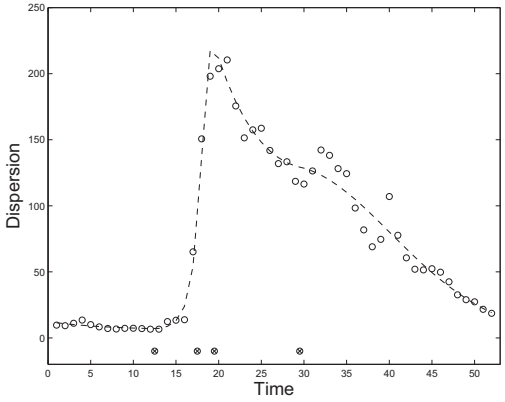
(a) Mean model for $\mathcal{LN}$ (transformed scale)



(b) Dispersion model for $\mathcal{LN}$



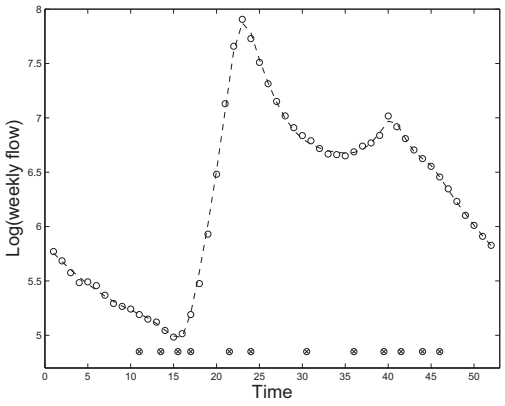
(c) Mean model for RIG (transformed scale)



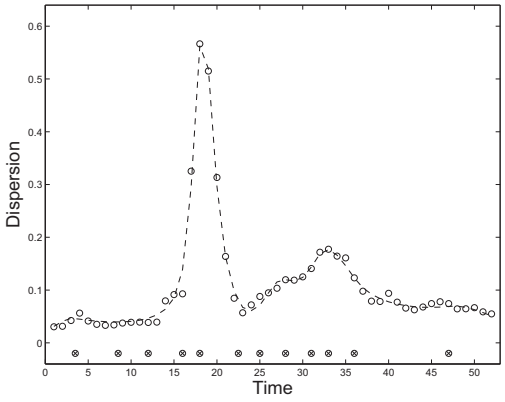
(d) Dispersion model for RIG

FIGURE 6.4. Models obtained with the use of $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$ in the MCMCRJ algorithm. Panels (a) and (c) : mean function models for the $\mathcal{LN}$ and RIG distributions; the open circles represent the weekly average of the sample of curves, while the crossed circles at the bottom represent the modal knot configurations ($\boldsymbol{\omega}^\dagger$). Panels (b) and (d) : dispersion function models for the $\mathcal{LN}$ and RIG distributions; the open circles give the weekly dispersion of the sample of curves and the crossed circles at the bottom represent the modal knot configurations ($\boldsymbol{\nu}^\dagger$).

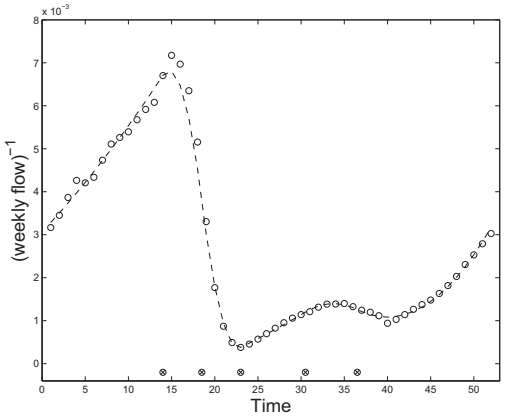
distribution shown in panel (d). The model for the $\mathcal{G}$ distribution, given in panel (c), appears to adequately reproduce the characteristics of the observed values, as was the case for panel (c) of Figure 6.2.



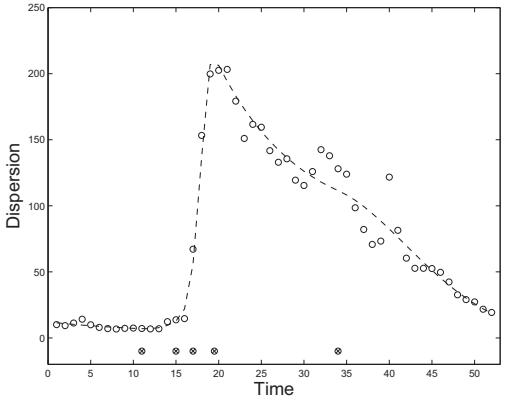
(a) Mean model for $\mathcal{LN}$ (transformed scale)



(b) Dispersion model for $\mathcal{LN}$



(c) Mean model for RIG (transformed scale)



(d) Dispersion model for RIG

FIGURE 6.5. Models obtained with the use of $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$ in the MCMCRJ algorithm. Panels (a) and (c) : mean function models for the $\mathcal{LN}$ and RIG distributions; the open circles represent the weekly average of the sample of curves, while the crossed circles at the bottom represent the modal knot configurations ($\boldsymbol{\omega}^\dagger$). Panels (b) and (d) : dispersion function models for the $\mathcal{LN}$ and RIG distributions; the open circles give the weekly dispersion of the sample of curves and the crossed circles at the bottom represent the modal knot configurations ($\boldsymbol{\nu}^\dagger$).

Figures 6.4 and 6.5 show the simultaneous models for the mean (on the transformed scale) and dispersion functions under the hypothesis that the data come from the lognormal ($\mathcal{LN}$) and reciprocal inverse Gaussian (RIG) distributions. Panels (a) and (c) present the model for the average on the transformed scales,

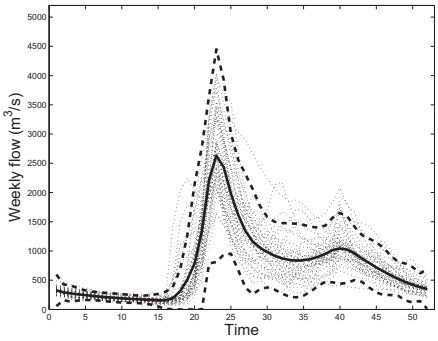
i.e. the logarithmic and reciprocal scales for the $\mathcal{LN}$ and RIG respectively, while panels (b) and (d) give the corresponding dispersion models. In Figure 6.4, the models were obtained with the use of $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$ in the MCMCRJ algorithm, and those of Figure 6.5 were obtained with the algorithm based on $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$. It is interesting to note the similarity between panel (b) of both Figures (the $\mathcal{LN}$ dispersion function) with panel (c) of Figures 6.2 and 6.3 (the $\mathcal{G}$ dispersion function).

Globally, the models shown in Figures 6.4 and 6.5 perform well and they capture the features of the different functions adequately. An exception is the RIG dispersion function, where the main shape of the curve is reproduced but specific features are not. This is similar to the case of the $\mathcal{N}$ dispersion function shown in panel (b) of Figures 6.2 and 6.3.

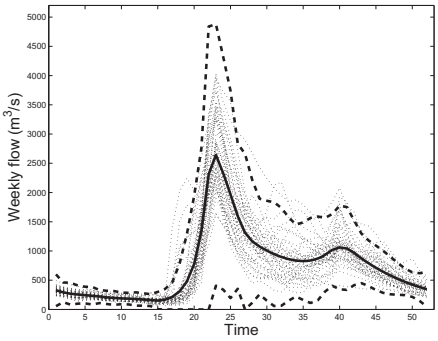
6.3.3.2. *Approximate credible sets*

Figure 6.6 gives the 95% approximate credible sets for the sample of curves according to the expressions given in equations (6.2.39) for the $\mathcal{N}$, $\mathcal{G}$, and IG distributions, and in (6.2.41) for the $\mathcal{LN}$ and RIG distributions. We see that most approximate credible sets give fairly good coverage but that the ones for the $\mathcal{LN}$ and $\mathcal{G}$ distributions seem to be the most adequate to reproduce the variability of the sample, which is in agreement with Table 6.1. For example, the former is able to capture the variability of the curves before and after the spring flood, and furthermore, the variability about the autumn maximum. Although the quadratic dependency of the variance on the mean seems to be the best relation to model the data, the approximate credible sets for the other distributions are still reasonable, which would not be the case if the mean and dispersion would not be modelled simultaneously.

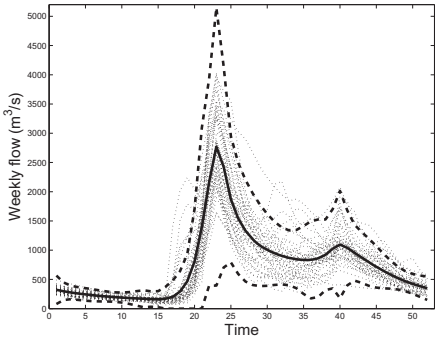
Finally, Figure 6.7 gives the 95% approximate credible sets for the dispersion functions. These can be used to determine whether the dispersion parameters can be considered equal. For the $\mathcal{N}$, IG and RIG distributions, it is clear that the dispersion parameters are not equal across the time domain. Regarding the $\mathcal{LN}$ and $\mathcal{G}$ distributions, besides the features in the intervals (15, 20) and (30, 35),



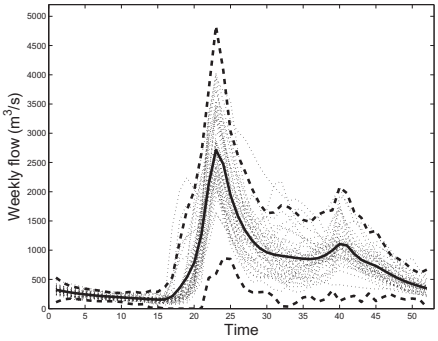
(a) $\mathcal{N}$ distribution



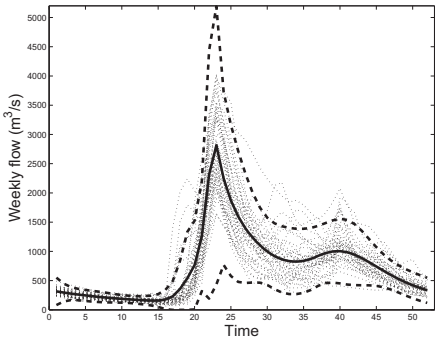
(b) $\mathcal{G}$ distribution



(c) IG distribution



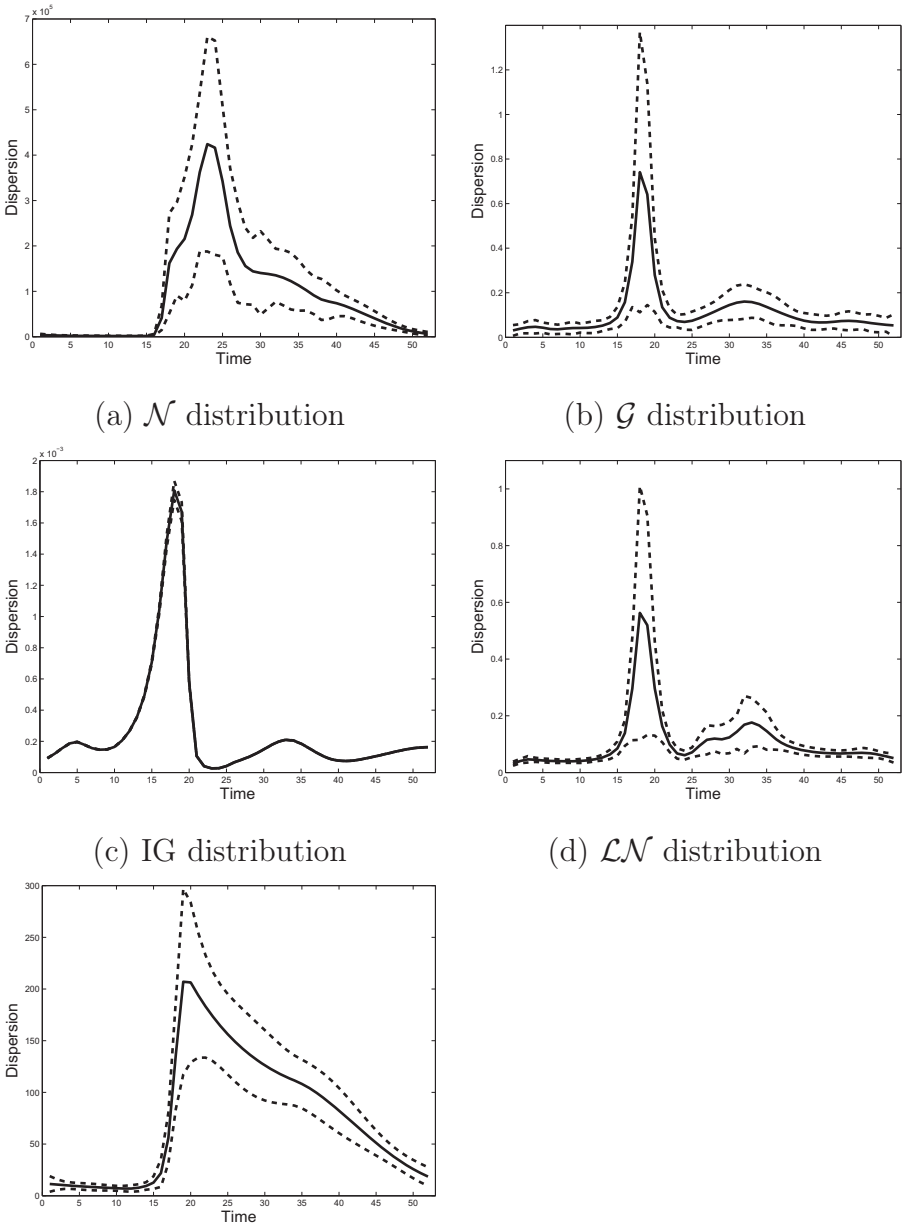
(d) $\mathcal{LN}$ distribution



(e) RIG distribution

FIGURE 6.6. 95% approximate credible sets for the sample of curves under the different distributions. The credible sets are given by the dashed lines, the models for the mean function are represented by full lines, and the dotted curves show the sample of functions given in Figure 6.1.

it appears that the dispersion is pretty much constant throughout the domain. Nonetheless, modelling these features results in more realistic credible sets as shown in Figure 6.6.



(a) $\mathcal{N}$ distribution

(b) $\mathcal{G}$ distribution

(c) IG distribution

(d) $\mathcal{LN}$ distribution

(e) RIG distribution

FIGURE 6.7. 95% approximate credible sets for the dispersion function under the different distributions. The credible sets are given by the dashed lines, while the full line gives the model for the dispersion function.

6.4. CONCLUSION

We have presented a Bayesian approach based on free-knot regression splines to simultaneously model the mean and dispersion functions of a sample of curves

in a fairly general distributional setting. We also addressed issues of model selection and the construction of credible sets for the mean and dispersion functions. The method was applied to a sample of hydrological functions and the results indicate that the methodology works well in practice.

There are several possible extensions to the present work. Here, both the mean and dispersion functions were modelled with Bayesian regression splines but it would be possible to apply the same methodology when the mean and/or the dispersion are modelled with multiple regression models. The approach based on the MCMCRJ methodology can also be applied for variable selection since it is constructed to move on finite spaces. For example, this implies that the mean of a generalized linear model could be modelled by a possible set of explanatory variables while the dispersion could be modelled simultaneously on an arbitrary domain such as case numbers. In this context, the model could choose the most adequate explanatory variables while at the same time determine if the dispersion can be considered constant for all observations.

Another extension is to model the mean and dispersion functions of a data set using generalized additive models (Hastie and Tibshirani, 1990). In the case of normally distributed data, Denison *et al.* (1998) showed how Bayesian regression splines can be used to model the mean by using generalized additive models. It is thus possible to extend our methodology in a similar direction to simultaneously model the mean and dispersion functions.

Appendix A. Statistical distributions

Table 6.2 gives explicit information about distributional quantities referred to in the main text. The first two lines present the expectation and the variance of a random variable from each distribution; the second line thus shows how the variance depends on the mean and establishes the value of the exponent p in the relation : variance $\propto \{\text{mean}\}^p$. We also see that for the $\mathcal{LN}$ and RIG distributions, both the expectation and the variance depend on the dispersion parameter ϕ. The third line gives η of equation (6.2.2) relative to the parameter μ which is modelled by a function of the regression splines defined by the nodal

parameter $\boldsymbol{\omega}$; the function $z(y)$ for the different statistical distributions can be read directly from this row. The fourth and fifth lines present the functions of the random variable for the decomposition of $\kappa(y,\phi)$ given in equation (6.2.3). The dispersion functions discussed in section 6.2.1 appear in the sixth line where they are now written as functions of y and μ (see equation (A.8) below). The last line gives the expectation of the dispersion functions for the different statistical distributions.

TABLE 6.2. Information concerning the normal ($\mathcal{N}$), gamma ($\mathcal{G}$), inverse gaussian (IG), lognormal ($\mathcal{LN}$), and reciprocal inverse gaussian (RIG) distributions.

	$\mathcal{N}(\mu,\phi)$	$\mathcal{G}(\mu,\phi)$	IG(μ,ϕ)	$\mathcal{LN}(\mu,\phi)$	RIG(μ,ϕ)
$\mathbb{E}(y)$	μ	μ	μ	$\mu_y=\exp(\mu+\phi/2)$	$\mu_y=\mu^{-1}+\phi$
$\mathbb{V}(y)$	ϕ	$\phi\mu^2$	$\phi\mu^3$	$\mu_y^2(\exp(\phi)-1)$	$\phi\mu_y+\phi^2$
$\eta[y,\zeta(\mu)]$	$\mu y-\frac{\mu^2}{2}$	$-\frac{y}{\mu}-\log(\mu)$	$-\frac{y}{2\mu^2}+\frac{1}{\mu}$	$\mu\log(y)-\frac{\mu^2}{2}$	$-\frac{1}{2\mu^2 y}+\frac{1}{\mu}$
$s(y)$	$\frac{1}{2}y^2$	$-\log(y)-1$	$\frac{1}{2}y^{-1}$	$\frac{1}{2}[\log(y)]^2$	$\frac{1}{2}y$
$t(y)$	0	$-\log(y)$	$-\frac{1}{2}\log(y^3)$	$-\log(y)$	$-\frac{1}{2}\log(y)$
$d(y,\mu)$	$(y-\mu)^2$	$2\left[\frac{(y-\mu)}{\mu}-\log\left(\frac{y}{\mu}\right)\right]$	$\frac{(y-\mu)^2}{\mu^2 y}$	$[\log(y)-\mu]^2$	$\frac{y\left(y^{-1}-\mu\right)^2}{\mu^2}$
$\mathbb{E}(d(y,\mu))$	ϕ	ϕ (for small ϕ)	ϕ	ϕ	ϕ

For the gamma distribution ($\mathcal{G}$), the decomposition of $\kappa(y,\phi)$ is obtained by using Stirling's approximation for $\Gamma(\phi^{-1})$, which should be accurate for small ϕ; this corresponds to a saddlepoint approximation of the gamma distribution (see Jorgensen, 1997). More specifically, setting $\varphi=\phi^{-1}$, Stirling's formula is given by (Abramowitz and Stegun, 1964)

$$\Gamma(\varphi)=(2\pi)^{1/2}\varphi^{\varphi-1/2}\exp(-\varphi)\left\{1+O(\varphi^{-1})\right\}, \tag{A.1}$$

and when φ is large, we obtain : $\Gamma(\varphi)\approx(2\pi)^{1/2}\varphi^{\varphi-1/2}\exp(-\varphi)$, which is Stirling's approximation. Replacing φ by ϕ^{-1} and using the approximation in the gamma density leads to the functions $s(y)$ and $t(y)$ given in Table 6.2.

The dispersion functions for the $\mathcal{N}$, $\mathcal{G}$, and IG distributions, given in Table 6.2, are the corresponding unit deviance functions of Jorgensen (1997). Furthermore, it

is possible to understand these functions from an information theory perspective ; in the same fashion as Efron (1986), we proceed to show this for the family of distributions considered here. Starting with the distribution of one observation with $\phi = 1$ (see equation (6.2.1)), we have

$$f(y|\zeta) = \exp\left\{\eta(y,\zeta) + \kappa(y,1)\right\}, \tag{A.2}$$

where, as in the main text, $\eta(y,\zeta) = \zeta z - \psi(\zeta)$, ζ is a function of the mean μ, z is a function of y, and $\kappa(y,1)$ is a function which now only depends on y. For a given distribution, the theoretical Kullback-Leibler information (Kullback, 1968) to discriminate between two models with $\zeta_1 = \zeta(\mu_1)$ and $\zeta_2 = \zeta(\mu_2)$ is given by

$$\Im = \mathbb{E}\left\{\log\left[\frac{f(y|\zeta_1)}{f(y|\zeta_2)}\right]\right\}, \tag{A.3}$$

$$= \mathbb{E}\left\{\eta(y,\zeta_1) - \eta(y,\zeta_2)\right\}, \tag{A.4}$$

$$= \mu_1\left\{\zeta_1 - \zeta_2\right\} - \left\{\psi(\zeta_1) - \psi(\zeta_2)\right\}, \tag{A.5}$$

where the expectation is taken relative to $f(y|\zeta_1)$. The Kullback-Leibler information statistic (Kullback, 1968) is obtained by setting $\mu_1 = z$ and $\mu_2 = \mu$, which gives

$$\Im_o = z\left\{\zeta(z) - \zeta(\mu)\right\} - \left\{\psi[\zeta(z)] - \psi[\zeta(\mu)]\right\}, \tag{A.6}$$

$$= \eta\left[y,\zeta(z)\right] - \eta\left[y,\zeta(\mu)\right], \tag{A.7}$$

The expression in (A.6) is the same as the expression of Efron (1986) when $z(y) = y$. For the family of distributions studied, we have $s(y) = \eta\left[y,\zeta(z)\right]$; by using the definition of $d(y,\mu)$ given after equation (6.2.5), we can therefore write

$$d(y,\mu) = 2\left\{s(y) - \eta\left[y,\zeta(\mu)\right]\right\} = 2\left\{\eta\left[y,\zeta(z)\right] - \eta\left[y,\zeta(\mu)\right]\right\} = 2\Im_o. \tag{A.8}$$

Thus, the model for the dispersion parameters, given in equations (6.2.5) and (6.2.6), can be understood as a model for twice the Kullback-Leibler information statistics. Since these statistics can be considered as measures of distance (Efron, 1986 ; Robert, 2001) under a given statistical distribution, a dispersion parameter therefore measures the discrepancy between $z(y)$ and μ on the Kullback-Leibler information scale.

For a sample of observations, McCullagh and Nelder (1989) define the deviance as twice the difference between the maximum log-likelihood achievable and that achieved by a model under consideration for a fixed unit dispersion parameter. The contribution of one observation to the deviance, with the present notation, is given by $\hat{d} = d(y, \hat{\mu}) = 2\left\{\eta\left[y, \zeta(z)\right] - \eta\left[y, \zeta(\hat{\mu})\right]\right\}$, where $\hat{\mu}$ represents the maximum likelihood estimate of μ for the model under consideration. When the mean parameter, μ, is unknown and replaced by an estimate, the dispersion parameter thus models the contribution of an observation to the deviance.

Appendix B. MCMC reversible jump algorithm

The MCMC reversible jump algorithm relies on a Metropolis-Hastings step at each iteration. Based on the algorithm suggested by Green (1995), we consider the acceptance probabilites of this step to be given by

$$\rho_{t_1,t_2} = \min\left(1, \text{marginal ratio} \times \text{prior ratio} \times \text{proposal ratio}\right), \tag{B.1}$$

for a set of particular move types t_1 and t_2, where t_1 is the move for the parameter $\boldsymbol{\omega} = (m_1, \boldsymbol{r}_1^{(m_1)})$ and t_2 is the move for $\boldsymbol{\nu} = (m_2, \boldsymbol{r}_2^{(m_2)})$. In our application (for $i = 1, 2$), the possible move types are $t_i = a$ for the addition of a knot, $t_i = s$ for the suppression of a knot and $t_i = d$ for the displacement of a knot. In our approach, the marginal distributions are replaced by the approximations $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu})$ or $\mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu})$, which constitute two approximations to the integration of the parameters $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ and $\boldsymbol{\gamma}_{\boldsymbol{\nu}}$.

The prior ratios are based on the prior distributions, $\pi(\boldsymbol{\omega})$ and $\pi(\boldsymbol{\nu})$, given in section 6.2.3; for example when $t_1 = s$ and $t_2 = s$, and the model currently contains m_1 and m_2 knots, the prior ratio is given by

$$\text{prior ratio} = \frac{\pi(\boldsymbol{r}_1^{(m_1-1)}|m_1-1)\pi(m_1-1)}{\pi(\boldsymbol{r}_1^{(m_1)}|m_1)\pi(m_1)} \frac{\pi(\boldsymbol{r}_2^{(m_2-1)}|m_2-1)\pi(m_2-1)}{\pi(\boldsymbol{r}_2^{(m_2)}|m_2)\pi(m_2)}. \tag{B.2}$$

Concerning the proposal ratios, care needs to be taken in order to have a balanced chain. For $i = 1, 2$, writing the probabilities of choosing $t_i = a, s, d$, when the

model contains m_i knots, as a_{m_i}, s_{m_i} and d_{m_i} respectively, we have the necessary constraint : $a_{m_i} + s_{m_i} + d_{m_i} = 1$. Following Denison *et al.* (1998), we take

$$a_{m_i} = c \times \min\left(1, \frac{\pi(m_i+1)}{\pi(m_i)}\right), \tag{B.3}$$

$$s_{m_i} = c \times \min\left(1, \frac{\pi(m_i-1)}{\pi(m_i)}\right), \tag{B.4}$$

which ensures that

$$\frac{a_{m_i}}{s_{m_i+1}} = \frac{\pi(m_i+1)}{\pi(m_i)}. \tag{B.5}$$

The constant c determines the rate at which the dimension of a nodal model changes and it needs to be between 0 and 0.5 for the sum of the move type probabilities to equal one. For the application section we use $c = 0.4$ as in Denison *et al.* (1998) and DiMatteo *et al.* (2001). For example, the proposal ratio for $t_1 = s$ and $t_2 = s$, when the model possesses m_1 and m_2 knots, is taken to be

$$\text{proposal ratio} = \frac{a_{m_1-1}/(M_1 - I_{m_1-1})}{s_{m_1}/m_1} \frac{a_{m_2-1}/(M_2 - I_{m_2-1})}{s_{m_2}/m_2}, \tag{B.6}$$

where I_{m_i-1} represents the number of impossible positions when there are $m_i - 1$ knots ; the unavailable positions are fixed by constraints on the proximity of the knots.

In general, we can write the acceptance probabilities under the form

$$\rho_{t_1,t_2} = \min\left(1, \varrho_{t_1,t_2}\right). \tag{B.7}$$

The following table gives the expressions for the various ϱ_{t_1,t_2}. The current model of the chain is considered to have the knot parameters : $\boldsymbol{\omega}_c = (m_1, \boldsymbol{r}_1^{(m_1)})$ and $\boldsymbol{\nu}_c = (m_2, \boldsymbol{r}_2^{(m_2)})$. The proposed knot configurations are as follows : $\boldsymbol{\omega}_a^* = (m_1 + 1, \boldsymbol{r}_{1*}^{(m_1+1)})$, $\boldsymbol{\omega}_s^* = (m_1 - 1, \boldsymbol{r}_{1*}^{(m_1-1)})$, $\boldsymbol{\omega}_d^* = (m_1, \boldsymbol{r}_{1*}^{(m_1)})$, $\boldsymbol{\nu}_a^* = (m_2 + 1, \boldsymbol{r}_{2*}^{(m_2+1)})$, $\boldsymbol{\nu}_s^* = (m_2 - 1, \boldsymbol{r}_{2*}^{(m_2-1)})$, and $\boldsymbol{\nu}_d^* = (m_2, \boldsymbol{r}_{2*}^{(m_2)})$; where $\boldsymbol{r}_{i*}^{(\cdot)}$ indicates a proposed knot configuration.

Appendix C. Marginal distribution approximations

$\mathbf{m}_a(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu})$

The information criterion, proposed by Schwarz (1978), consists of measuring goodness of fit through the maximized log-likelihood and penalizing for model

TABLE 6.3. Explicit expressions of ϱ_{t_1,t_2} for all move types t_1 and t_2.

	$t_2 = a$	$t_2 = s$	$t_2 = d$
$t_1 = a$	$\frac{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_a^*,\boldsymbol{\nu}_a^*)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)} \frac{M_1-I_{m_1}}{M_1} \frac{M_2-I_{m_2}}{M_2}$	$\frac{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_a^*,\boldsymbol{\nu}_s^*)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)} \frac{M_1-I_{m_1}}{M_1} \frac{M_2}{M_2-I_{m_2}}$	$\frac{\mathrm{m}\left(\boldsymbol{y}\|\boldsymbol{\omega}_a^*,\boldsymbol{\nu}_d^*\right)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)} \frac{M_1-I_{m_1}}{M_1}$
$t_1 = s$	$\frac{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_s^*,\boldsymbol{\nu}_a^*)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)} \frac{M_1}{M_1-I_{m_1}} \frac{M_2-I_{m_2}}{M_2}$	$\frac{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_s^*,\boldsymbol{\nu}_s^*)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)} \frac{M_1}{M_1-I_{m_1}} \frac{M_2}{M_2-I_{m_2}}$	$\frac{\mathrm{m}\left(\boldsymbol{y}\|\boldsymbol{\omega}_s^*,\boldsymbol{\nu}_d^*\right)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)} \frac{M_1}{M_1-I_{m_1}}$
$t_1 = d$	$\frac{\mathrm{m}\left(\boldsymbol{y}\|\boldsymbol{\omega}_d^*,\boldsymbol{\nu}_a^*\right)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)} \frac{M_2-I_{m_2}}{M_2}$	$\frac{\mathrm{m}\left(\boldsymbol{y}\|\boldsymbol{\omega}_d^*,\boldsymbol{\nu}_s^*\right)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)} \frac{M_2}{M_2-I_{m_2}}$	$\frac{\mathrm{m}\left(\boldsymbol{y}\|\boldsymbol{\omega}_d^*,\boldsymbol{\nu}_d^*\right)}{\mathrm{m}(\boldsymbol{y}\|\boldsymbol{\omega}_c,\boldsymbol{\nu}_c)}$

complexity via a term which depends on the number of parameters in the model and the number of observations used to fit the model. In the context of this paper (see equations (6.2.7) and (6.2.8)), it is given by

$$S_{\boldsymbol{\omega},\boldsymbol{\nu}} = \frac{N}{2} \sum_{j=1}^{n} \left[\widehat{\xi}_j \widehat{\bar{d}}_{\cdot j} - \lambda(\widehat{\xi}_j) \right] + Nn \left[\bar{t}_{\cdot\cdot} - \frac{1}{2} \log(2\pi) \right] - \frac{(K_{\boldsymbol{\omega}} + K_{\boldsymbol{\nu}})}{2} \log(Nn), \quad \text{(C.1)}$$

where $\widehat{\xi}_j$ and $\widehat{\bar{d}}_{\cdot j}$ are evaluated at the maximum likelihood estimates, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\nu}}$, which are obtained by maximizing $\sum_{j=1}^{n} \left[\xi_j \bar{d}_{\cdot j} - \lambda(\xi_j) \right]$; $K_{\boldsymbol{\omega}} = \dim(\boldsymbol{\beta}_{\boldsymbol{\omega}})$ and $K_{\boldsymbol{\nu}} = \dim(\boldsymbol{\gamma}_{\boldsymbol{\nu}})$. The first approximation is obtained by exponentiating the Schwarz criterion : $\mathrm{m}_a(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu}) = \exp(S_{\boldsymbol{\omega},\boldsymbol{\nu}})$. Kass and Wasserman (1995) showed that this quantity can be used effectively to calculate accurate ratios of marginal distributions for nested models.

$\mathbf{m}_b(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu})$

Given two knot configurations, $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$, and the prior distributions given in the main text (section 6.2.3), the joint distribution is

$$D(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{\omega},\boldsymbol{\nu}) = f(\boldsymbol{y}|\boldsymbol{\zeta},\boldsymbol{\xi},\boldsymbol{\omega},\boldsymbol{\nu})\pi(\boldsymbol{\beta}|\boldsymbol{\beta}^0,\boldsymbol{\gamma},\boldsymbol{\omega},\boldsymbol{\nu})\pi(\boldsymbol{\gamma}|\boldsymbol{\gamma}^0,\boldsymbol{\omega},\boldsymbol{\nu}), \quad \text{(C.2)}$$

where all subscripts $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$ have been removed to simplify the notation; $\boldsymbol{\zeta}$ is a function of $\boldsymbol{\beta}$ which depends on the distribution and a link function, and $\boldsymbol{\xi}$ is a function of $\boldsymbol{\gamma}$, also through a link function. We want to approximate the integral $\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega},\boldsymbol{\nu}) = \int_{\boldsymbol{\beta},\boldsymbol{\gamma}} D(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{\omega},\boldsymbol{\nu}) \mathrm{d}\boldsymbol{\beta}\mathrm{d}\boldsymbol{\gamma}$. We use the basic form of Laplace's approximation (Tierney and Kadane, 1986; Shun and McCullagh, 1995) to do this, but in order to do so, we need to maximize the joint distribution relative

to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Since the two maximisations are intertwined, *i.e.* the maximisation of $\boldsymbol{\beta}$ depends on $\boldsymbol{\gamma}$ and vice versa, we propose to maximize the joint distribution simultaneously for both parameters. We thus consider the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ and we can rewrite the marginal distribution as

$$\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu}) = \int_{\boldsymbol{\theta}} D(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{\omega}, \boldsymbol{\nu}) \mathrm{d}\boldsymbol{\theta}. \tag{C.3}$$

Using equation (6.2.8) and the prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the joint distribution can be written as

$$D(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{\omega}, \boldsymbol{\nu}) \cong \exp\Bigg\{ \frac{N}{2} \sum_{j=1}^{n} \left[\xi_j \bar{d}_{\cdot j} - \lambda(\xi_j)\right] - \frac{1}{2} q(\boldsymbol{\theta}) - \frac{1}{2} \log\left(|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|\right) + c_1 \Bigg\}, \tag{C.4}$$

where $q(\boldsymbol{\theta})$ is defined in the main text after equation (6.2.22), $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is the prior covariance of $\boldsymbol{\beta}$ which depends on $\boldsymbol{\gamma}$, $c_1 = t_{\cdot\cdot} - \frac{Nn + K_{\boldsymbol{\beta}} + K_{\boldsymbol{\gamma}}}{2} \log(2\pi) - \frac{1}{2} \log\left(|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|\right)$, $K_{\boldsymbol{\beta}} = \dim(\boldsymbol{\beta})$, $K_{\boldsymbol{\gamma}} = \dim(\boldsymbol{\gamma})$, and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ is the prior covariance matrix of $\boldsymbol{\gamma}$. Laplace's approximation then yields

$$\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu}) \approx \mathrm{m}_b(\boldsymbol{y}|\boldsymbol{\omega}, \boldsymbol{\nu}) = \exp\Bigg\{ \frac{N}{2} \sum_{j=1}^{n} \left[\xi_j^* \bar{d}_{\cdot j}^* - \lambda(\xi_j^*)\right] - \frac{1}{2} q(\boldsymbol{\theta}^*) - \frac{1}{2} \log\left(|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^*|\right) + c_2 \Bigg\}, \tag{C.5}$$

where all starred quantities are evaluated at $\boldsymbol{\theta}^*$ given in equation (6.2.22) and

$$c_2 = c_1 + \frac{1}{2} \log\left(|\boldsymbol{\Sigma}_{\boldsymbol{\beta},\boldsymbol{\gamma}}^*|\right) + \frac{K_{\boldsymbol{\beta}} + K_{\boldsymbol{\gamma}}}{2} \log(2\pi), \tag{C.6}$$

$$= Nn \left[\bar{t}_{\cdot\cdot} - \frac{1}{2} \log(2\pi)\right] + \frac{1}{2} \log\left(|\boldsymbol{\Sigma}_{\boldsymbol{\beta},\boldsymbol{\gamma}}^*|\right) - \frac{1}{2} \log\left(|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|\right), \tag{C.7}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\beta},\boldsymbol{\gamma}}^*$ given in equation (6.2.23). Rearranging the terms, one gets equation (6.2.21).

Chapitre 7

CONCLUSION

7.1. Contributions de cette thèse

Dans cette thèse, nous avons présenté trois articles sur la modélisation bayésienne avec des splines de données fonctionnelles provenant du domaine de l'hydrologie.

Le premier article montre la pertinence de l'analyse bayésienne de données fonctionnelles en hydrologie. En effet, avec le contexte de modélisation proposé, nous sommes en mesure de considérer les hydrogrammes annuels comme un échantillon de courbes pour lequel nous cherchons à obtenir un profil moyen, ainsi qu'à modéliser ce dernier. La synchronisation des hydrogrammes annuels nous permet d'avoir un échantillon de courbes plus homogène au niveau temporel, alors que la modélisation bayésienne avec des splines de régression nous permet de reproduire le profil moyen d'un échantillon de courbes synchronisées. Cette approche en deux étapes nous permet d'obtenir une représentation parcimonieuse de l'échantillon de courbes hydrologiques. Finalement, il doit être noté que la méthodologie présentée dans le premier article est en voie d'être implantée à Hydro-Québec.

Le deuxième article propose un cadre de modélisation plus général que celui présenté dans le premier article. Le modèle probabiliste est élargi en mettant de l'avant une famille de distributions statistiques qui inclut celles de la famille exponentielle. Ainsi, nous pouvons étudier une diversité de distributions statistiques desquelles peuvent provenir les données fonctionnelles à modéliser. De plus, nous développons une méthodologie, basée sur le MCMC avec sauts réversibles, pour

traiter les noeuds intérieurs comme des quantités aléatoires. Ces noeuds intérieurs déterminent la base de modélisation et en les traitant comme des quantités aléatoires, notre modèle est beaucoup plus flexible. Nous proposons aussi une stratégie pour déterminer la distribution statistique la plus adéquate pour un échantillon de données fonctionnelles.

Le troisième article présente une méthodologie pour simultanément modéliser, avec des splines de régression, la courbe moyenne et la courbe de dispersion d'un échantillon de données fonctionnelles, et ce sous l'hypothèse que les observations proviennent de la famille de distributions étudiée dans l'article 2. Le modèle proposé possède par conséquent une grande flexibilité car il permet de prendre en compte la variabilité de données provenant de diverses distributions statistiques, tout en modélisant la tendance centrale de ces données. Les noeuds intérieurs des deux bases de modélisation, celle pour la tendance centrale et celle pour la dispersion, sont considérés comme des quantités aléatoires et leurs distributions *a posteriori* sont ici aussi explorées avec le MCMC avec sauts réversibles.

Il est important de noter que les méthodologies développées dans les trois articles ne se limitent pas seulement à la modélisation d'échantillons de courbes avec des splines de régression. En effet, les méthodes peuvent aussi bien être appliquées pour la modélisation d'une seule courbe avec des splines de régression. Les méthodologies proposées dans les articles 2 et 3 peuvent aussi être mises en oeuvre avec des méthodes non paramétriques telles les séries de Fourier et les ondelettes, où au lieu de traiter les noeuds intérieurs comme des quantités aléatoires, les éléments des bases de fonctions pourraient être indexés et les indices seraient traités comme des quantités aléatoires. Ceci est possible car nous utilisons un support fini pour les noeuds intérieurs. Finalement, pour la même raison, il est naturellement possible d'utiliser les méthodes des articles 2 et 3 pour effectuer la sélection de variables afin de modéliser la tendance centrale, et aussi la dispersion, d'observations provenant de la famille de distributions mise de l'avant.

7.2. Avenues de recherche

Dans un premier temps, une attention limitée a été donnée à la synchronisation des hydrogrammes dans cette thèse. Nous croyons qu'il existe probablement des variables physiques qui peuvent nous informer par rapport aux moments des crues printanières par exemple. Plus précisément, ces variables physiques pourraient être utilisées dans un cadre prédictif afin de déterminer si une crue printanière d'une certaine année sera hâtive ou tardive. Ainsi, il serait possible d'élaborer une méthode statistique afin de modéliser les fonctions de synchronisation à partir de variables auxiliaires.

Dans la thèse, nous nous sommes concentrés sur la modélisation des propriétés globales d'un échantillon de courbes, mais il pourrait aussi être d'intérêt de modéliser les courbes individuelles. Afin d'accomplir cette tâche, un modèle bayésien hiérarchique pourrait être construit. En ajoutant un niveau pour la modélisation des courbes individuelles, il serait possible de simultanément obtenir des modèles pour les courbes individuelles et une courbe moyenne. Nous travaillons présentement sur ce problème.

Finalement, les méthodes présentées peuvent être adaptées au cadre des modèles additifs généralisés. Ainsi, pour chacune des variables auxiliaires, une base formée de splines de régression pourrait être utilisée comme outil non paramétrique.

BIBLIOGRAPHIE

Abramowitz, M., and Stegun, I. (1964). *Handbook of Mathematical Functions*. Dover, New York.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*. (eds Petrov, B.N., et Csaki, F.). Budapest : Akademiai Kaido.

Aksoy (2000). Use of Gamma Distribution in Hydrological Analysis. *Turkish Journal of Engineering and Environmental Sciences* **24**, 419-428.

Albert, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association* **83**, 1037-1044.

Angers, J.-F., Merleau, J., et Perreault, L. (2005). Landmark Registration of Hydrographs and Bayesian Estimation of a Mean Hydrograph. In *Proceedings of the International Sri Lankan Statistical Conference : Visions of Futuristic Methodologies*. (eds. de Silva, B. M. and Mukhopadhyay, N.). Kandy, Sri Lanka, p. 47-60.

Bernardo, J. M., et Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

Biller, C. (2004). Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models. *Journal of Computational and Graphical Statistics* **9**, 122-140.

Bobée, B., et Ashkar, F. (1991). *The Gamma Family and Derived Distributions*. Colorado : Water Ressources Publications.

Box, G. E. P., et Cox, D. R. (1964). An analysis of transformations. *Journal of Royal Statistical Society* B **26**, 211-246.

Box, G.E.P., Jenkins, G.M., et Reinsel, G.C. (1994). *Time Series Analysis : Forecasting and Control*, 3rd Ed. Prentice-Hall, New Jersey.

Brumback, C. L., et Lindstrom, J. M. (2004). Self-Modeling with Flexible, Random Time Transformations. *Biometrics* **60**, 461-470.

Carroll, R. J., et Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Annals of Statistics* **10**, 429-441.

Carroll, R. J., et Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

Casella, G., et Berger, R. (2001). *Statistical Inference*. Wadsworth, Belmont, CA.

Chow, V. T. (1964). Runoff. *Handbook of Applied Hydrology*. McGraw-Hill, New York.

Chow, V. T., Maidment, D. R., et Mays, L. W. (1988). *Applied Hydrology*. McGraw-Hill, New York.

Coles, S. (2001). *An Introduction to Statistical Modelling of Extreme Values*. Springer-Verlag, London.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer, New York.

Denison, D. G. T., Mallick B. K., et Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of Royal Statistical Society* B **60**, 333-350.

DiMatteo, I., Genovese, C. R., et Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055-1072.

Donoho, D. L., et Johnston, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Doodge, J. C. I. (1959). A general theory of the unit hydrograph. *Journal of Geophysical Resources* **64**, 241-256.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* **81**, 709-721.

Eilers, P. H. C., et Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science* **11**, 89-121.

Favre, A.-C., Meylan, P., et Musy, A. (2008). *Hydrologie fréquentielle : Une science prédictive*. Presses polytechniques et universitaires romandes, Lausanne.

Folks, J. L., et Chhikara, R. S. (1978). The Inverse Gaussian Distribution and its Statistical Application-A Review. *Journal of Royal Statistical Society* B **40**, 263-289.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., et Villeneuve, J.-P. (2001a). Distributed watershed compatible with remote sensing and GIS data. I : Description of the model. *Journal of Hydrologic Engineering* **6**, 91-99.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., et Villeneuve, J.-P. (2001b). Distributed watershed compatible with remote sensing and GIS data. II : Application to Chaudiere watershed. *Journal of Hydrologic Engineering* **6**, 100-108.

Fortin, V. (2000). Le modèle météo-apport HSAMI : historique, théorie et application. *Document interne Hydro-Québec no. IREQ-99-255*.

Friedman, J. H., et Silverman, B. W. (1989). Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics* **31**, 3-39.

Gelman, A., Carlin, J. B., Stern, H. S., et Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, Boca Raton, FL.

Genest, C., et Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* **12**, 347-368

George, E. I., et McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association* **88**, 881-889.

George, E. I., et McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-374.

Gervini, D., et Gasser, T. (2004). Self-Modelling Warping Functions. *Journal of Royal Statistical Society* B **66**, 959-971.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Green, P. J., et Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

Gumbel, E. (1958). *Statistics of Extremes*. Columbia University Press, New York.

Hastie, T., et Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

He, X., et Shi, P. (1998). Monotone B-Spline Smoothing. *Journal of the American Statistical Association* **93**, 643-650.

IAWCD (1982). Guidelines for determining flood flow frequency. Technical Report Bulletin 17B (revised and corrected). InterAgency Committee on Water Data (IAWCD) - Hydrology Subcommittee. Washington, DC.

Javelle, P., Ouarda, T. B. M. J., Lang, M., Bobée, B., Galéa, J., et Grésillon, J.-M. (2002). Development of regional flow-duration-frequency curves based on the index-flood method. *Journal of Hydrology* **258**, 249-259.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Clarendon, Oxford.

Johnson, R. A., et Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*, 3rd ed. Prentice Hall, New Jersey.

Jorgensen, B. (1997). *The Theory of Dispersion Models*. Chapman and Hall, London.

Kass, R. E., et Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

Kass, R. E., et Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928-934.

Kneip, A., et Engel, J. (1995). Model Estimation in nonlinear regression under shape invariance. *The Annals of Statistics* **23**, 551-570.

Kneip, A., et Gasser, T. (1988). Convergence and consistency results for self-modeling nonlinear regression. *Annals of Statistics* **16**, 82-112.

Kneip, A., et Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics* **20**, 1266-1305.

Kneip, A., Li, X., MacGibbon, K. B., et Ramsay, J. O. (2000). Curve registration by local regresion. *Canadian Journal of Statistics* **28**, 19-29.

Kullback, S. (1968). *Information Theory and Statistics*, 2nd ed. Dover, New York.

Lee, P. (1989), *Bayesian Statistics : An introduction*. Oxford University Press, London.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*. Rotterdam University Press.

Lindstrom, M. J. (2002). Bayesian Estimation of Free-knot Splines using Reversible Jumps. *Computational Statistics and Data Analysis* **41**, 255-269.

Liu, X., et Muller, H. (2004). Functional Averaging and Synchronization for Time-Warped Random Curves. *Journal of the American Statistical Association* **99**, 687-699.

Markovic, R. D. (1965). *Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff*, Hydrology Paper, No.8, Colorado State University, Fort Collins, Colorado.

McCullagh, P., et Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

McCulloch, C. E., et Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

Merleau, J., Evin, G., Perreault, L., Favre, A.-C., Tremblay, D., et Angers, J.-F. (2005). Analyse descriptive des prévisions hydrologiques d'ensemble et modélisation par un mélange de deux lois gamma. Création d'hydrogrammes de base et d'hydrogrammes prévisionnels à pas de temps journalier. *Rapport technique R819*, INRS-Eau, Terre et Environnement, Québec.

Nelder, J. A., et Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221-232.

Nelder, J. A., et Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of Royal Statistical Society* A **135**, 370-384.

Nezhikhovsky, R. A. (1971). *Channel network of the basin and runoff formation.* Hydrometeorological, Leningrad, Russia.

Nott, David J. (2006). Semiparametric estimation of mean and variance functions for non-Gaussian data. *Computational Statistics* **21**, 603-620.

Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis.* Birkhauser, Boston.

Ouhib, L. (2005). Modélisation des apports naturels de réservoirs. *Mémoire de maîtrise*, Université de Montréal, Montréal.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models, *Biometrika* **85**, 13-27.

Pericchi, L. R. (1984). An alternative to the standard Bayesian procedure for discrimination between normal linear models, *Biometrika* **71**, 575-586.

Perreault, L., Bobée, B., et Rasmussen, P. F. (1999a). Halphen Distribution System. I : Mathematical and Statistical Properties. *Journal of Hydrologic Engineering*, July 1999, 189-199.

Perreault, L., Bobée, B., et Rasmussen, P. F. (1999b). Halphen Distribution System. II : Parameter and Quantile Estimation. *Journal of Hydrologic Engineering*, July 1999, 200-208.

Perreault, L., et Latraverse, M. (2001). Modélisation des apports naturels pour la prise en compte de leur aléa dans la méthode SDDP de planification de la production. *Rapport technique*, Institut de recherche d'Hydro-Québec, Varennes.

Pilgrim, D. H., et Cordery, I. (1993). Flood runoff. *Handbook of Hydrology.* McGraw-Hill, New York.

Raiffa, H., et Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston : Graduate School of Business, Harvard University.

Raftery, A. E. (1996). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models. *Biometrika* **83**, 251-266.

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science* **3**, 425-441.

Ramsay, J. O., et Li, X. (1998). Curve Registration. *Journal of Royal Statistical Society* B **60**, 351-363.

Ramsay, J. O., et Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York.

Rao, A. R., et Hamed, K. H. (2000). *Flood Frequency Analysis*. CRC Press, Boca Raton, FL.

Rasmussen, P. F., Salas, J. D., Fagherazzi, L., Rassam, J. C., et Bobée, B. (1996). Estimation and validation of contemporaneous PARMA models for streamflow simulation. *Water Resources Research* **32**, 3151-3160.

Robert, C. P. (1994). *The Bayesian Choice*. Springer, New York.

Robert, C. P. (2001). *The Bayesian Choice*, 2nd ed. Springer, New York.

Ronn, B. (2001). Nonparametric Maximum Likelihood Estimation for Shifted Curve. *Journal of Royal Statistical Society* B **63**, 243-259.

Ruppert, D., Wand, M. P., et Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Salas, J. D., Delleur, J. W., Yevjevich, V., et Lane, W. L. (1980). *Applied Modelling of Hydrologic Time Series*. Water Resources Publication, Fort Collins, Colorado.

Salas, J. D., Boes, D. C., et Smith, R. A. (1982). Estimation of ARMA models with seasonal parameters. *Water Resources Research* **18**, 1006-1010.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* **6**, 461-464.

Sherman, L. K. (1932). Streamflow from rainfall by the unit unit-graph method. *Eng. News-Rec.* **108**, 501-505.

Shun, Z., et McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of Royal Statistical Society* B **57**, 749-760.

Smith, P. (1979). Splines as a Useful and Convenient Statistical Tool. *The American Statistician* **33**, 57-62.

Smith, M., et Khon, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317-343.

Snyder, F. F. (1938). Synthetic unit-graphs. *Trans. Am. Geophys. Union* **19**, 447-454.

Sokolov, A. A., Rantz, S. E., et Roche, M. (1976). Methods of developing design-flood hydrographs. *Flood computation methods compiled from world experience*, UNESCO, Paris.

Stone, C. J., Hansen, M., Kooperberg, C., et Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics* **25**, 1371-1425.

Telesca, D., et Inoue, L. Y. T. (2008). Bayesian Hierarchical Curve Registration. *Journal of the American Statistical Association* **103**, 328-339.

Tierney, L., et Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82-86.

Tierney, L., Kass, R. E., et Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions. *Journal of the American Statistical Association* **84**, 710-716.

Titterington, D. M., Smith, A. F. M., et Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

U.S. Soil Conservation Service (US-SCS) (1985). Hydrology. *National Engineering Handbook*, U.S. Dept of Agriculture, Washington, D.C.

Vecchia, A. V., Obeysekera, J. T. B., Salas, J. D., et Boes, D. C. (1983). Aggregation and estimation of low-order periodic ARMA models, *Water Resources Research* **19**, 1297-1306.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia.

Wang, X., et George, E. I. (2007). Adaptive Bayesian Criteria in Variable Selection for Generalized Linear Models. *Statistica Sinica* **17**, 667-690.

Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439-447.

West, M., and Harrison, J. (1998). *Bayesian Forecasting and Dynamic Models*, 2nd ed. Springer-Verlag, New York.

Yue, S., et Hashino, M. (2000). Unit hydrographs to model quick and slow runoff components of streamflow. *Journal of Hydrology* **227**, 195-206.

Yue, S., Ouarda, T. B. M. J., Bobée, B., Legendre, P., et Bruneau, P. (1999). The Gumbel mixed model for flood frequency analysis. *Journal of Hydrology* **226**, 88-100.

Yue, S., Ouarda, T. B. M. J., Bobée, B., Legendre, P., et Bruneau, P. (2002). Approach for Describing Statistical Properties of Flood Hydrograph. *Journal of Hydrologic Engineering* **7**, 147-153.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*, P. Goel and A. Zellner (Eds.), 233-243. Elsevier, North-Holland, Amsterdam.

Annexe A

AUTORISATION POUR L'ARTICLE 1

We are pleased to grant permission for the use of the material requested for inclusion in your thesis.

Michael Connolly
Journals Publications Specialist
AGU
2000 Florida Avenue, NW
Washington, DC 20009
USA
202-777-7365