Université du Québec
INRS-ETE

La calibration de modèle hydrologique par analyse Bayésienne

Théorie et application aux modèles "abc" et GR2M

par
DAVID HUARD

Thèse présentée
pour l'obtention
du grade de Philosophiæ doctor (Ph.D.)
en sciences de l'eau

Jury d'évaluation

Président du jury et examinateur externe	M. Jim Freer, professeur Centre for Sustainable Water Management Lancaster Environment Centre
Examinateur externe	François Brissette, professeur École de Technologie Supérieure
Examinateur interne	M. Taha Ouarda, professeur INRS-ETE Université du Québec
Directeur de recherche	M. Alain Mailhot, professeur INRS-ETE Université du Québec

Thèse présentée le 7 septembre 2007

Résumé

Les méthodes de calibration standard font l'hypothèse que les données d'entrée sont exactes. Lorsque cette condition n'est pas satisfaite, les paramètres estimés lors de la calibration du modèle risquent d'être biaisés. Or, en hydrologie et dans la plupart des sciences environnementales, il est difficile d'obtenir des valeurs expérimentales qui ne soient pas affectées par des erreurs de toutes sortes. Bien qu'il soit difficile d'estimer l'impact de telles erreurs sur les prédictions des modèles, les hydrologues s'entendent pour dire que le développement de méthodes de calibration tenant compte des diverses sources d'incertitudes constitue une étape essentielle au développement des modèles hydrologiques. Cette thèse présente une telle méthode, basée sur l'analyse Bayésienne. Cette méthode greffe au modèle hydrologique divers modèles d'erreurs (entrée, sortie, structural), ce qui permet lors de la calibration d'interpréter les données afin d'estimer de manière cohérente la probabilité d'un jeu de paramètres. Bien que conceptuellement simple, cette méthode nécessite la résolution d'intégrales de très grandes dimensions, ce qui freine son application aux modèles hydrologiques journaliers utilisés couramment. Afin d'en étudier les propriétés, elle est donc appliquée à deux modèles hydrologiques mensuels simples : "abc" et GR2M. Le premier étant linéaire, il est possible, sous certaines conditions, de dériver une solution analytique permettant le calcul de la distribution *a posteriori* (*posterior*) de manière directe. Les résultats issus de ces simulations indiquent que la méthode permet effectivement de tenir compte des incertitudes sur les données d'entrée lors

de la calibration. Pour les modèles non linéaires tels que GR2M, l'intégration doit se faire de manière numérique. La solution choisie utilise un algorithme de type *chaînes de Markov Monte Carlo* (MCMC), permettant de générer un ensemble de paramètres dont la distribution converge vers le *posterior* lorsque le nombre d'itérations Monte Carlo augmente. Bien que cette solution soit fonctionnelle, son application aux modèles journaliers (ou horaires) requiera des ressources numériques importantes.

L'analyse des résultats obtenus avec "abc" et GR2M suggèrent que des modèles d'erreurs vagues (peu informatifs) ne permettent pas d'améliorer significativement l'efficacité des modèles et la qualité des prédictions. Par contre, en spécifiant des modèles d'erreurs qui décrivent fidèlement la probabilité d'occurrence des erreurs, on s'assure que l'incertitude sur les prédictions reflète l'incertitude sous-jacentes aux données et au modèle. Il apparaît donc que l'intérêt d'utiliser une telle méthode de calibration est étroitement liée à la capacité du modélisateur de formuler des modèles d'erreurs réalistes et informatifs, ce qui, en hydrologie est loin d'être trivial. Il n'en reste pas moins que la calibration bayésienne de ces modèles simples permet de tirer des conclusions générales et d'identifier certains enjeux importants. En effet, l'application au modèle "abc" souligne l'importance de la sélection du *prior* sur les vraies données d'entrée. L'application au modèle GR2M, quant à elle, aborde l'effet de l'incertitude des conditions initiales, l'interprétion des résultats de calibration dans le contexte de multiples sources d'incertitude, l'effet du choix des modèles erreurs et la validation des hypothèses sous-jacentes à ces modèles d'erreurs. De plus, l'application montre que l'intégration explicite des erreurs structurales permet d'inférer séparement les différentes erreurs, et par exemple, de discerner les erreurs structurales des erreurs de mesures.

Abstract

Standard calibration methods assume that the input data is known exactly. In hydrology, as well as in most environmental sciences, it is extremely difficult to gather experimental data uncorrupted by all kinds of errors. When this exactness condition is not met, hydrologists face the possibility that the parameters estimated during calibration are biased by the input errors. Although it is difficult to generalize the effect of such biases on the predictions of calibrated models, the concensus in the hydrological community is that a crucial step in the improvement of hydrological models is the development of calibration methods taking into account multiple sources of uncertainties. This thesis proposes such a method, based on Bayesian analysis. This method grafts various error models (input, output, structural) to the hydrological model, which interpret the data for the model and allow the estimation of coherent probabilities for parameter sets. Although conceptually simple, the use of the method relies on the resolution of high-dimensional integrals, hindering its application to the daily hydrological models commonly used. In order to study the method's properties and better understand the effect of errors on calibration, we apply it to two simple monthly models: "abc" and GR2M. The first one being linear, it is possible, under certain conditions, to derive an analytical solution and compute directly the *posterior* probability. Results obtained from these simulations suggest that Bayesian calibration reduces slightly the bias on the parameters. For nonlinear models such as GR2M, the integration cannot be done analytically. The

solution adopted was to perform Markov chain sampling using the Metropolis algorithm (MCMC) to approximate the posterior distribution. Although this solution is functional, its application to daily (or hourly) models requires significant numerical resources due, again, to the high dimensionality of the problem.

Results obtained for the two models suggest that vague errors models (uniformative) do not improve significantly the efficiency of the calibrated models. On the other hand, specifying realistic error models ensures that the prediction uncertainty is coherent with data and model uncertainty. Hence, the interest for using such a calibration method is closely dependent on the modeller's ability to define realistic and informative error models, which is not a trivial task in hydrology. Nevertheless, the calibration of simple models by a Bayesian uncertainty assessment method allows certain general conclusions to be drawn and the identification of a number of interesting issues. Indeed, the "abc" application underlines the importance of the prior selection for the true input on parameter inference. The GR2M application discusses the effect of initial state uncertainty, the interpretation of calibration results in the context of multiple sources of error, the effect of error models on the parameters and the validation of their underlying assumptions. Also, the GR2M study shows that by explicitly integrating structural errors, the different types of errors can be inferred, and structural errors can be explictly separated from data errors.

Avant propos / Foreword

Cette thèse présente les travaux de recherche menés au cours de mes études doctorales. Les résultats principaux se trouvent sous la forme d'articles insérés en deuxième partie. La première partie consiste en une synthèse permettant de donner au lecteur une vue d'ensemble des travaux effectués, ainsi qu'une occasion de tirer des conclusions générales sur leur portée. Pour les lecteurs ne maitrisant pas l'anglais, un résumé en français de la synthèse est inclu au premier chapitre de la première partie (page 1).

This document covers the work done during my PhD studies. Its format is that of a thesis by article, in which a relatively breve discussion about the context, methodology, general results and conclusion (part I) precedes published or submitted articles (part II). The reader will also find in the first chapter a French synthesis of this document. Although this thesis may be read from first page to last, I would suggest reading the articles just before chapter 4, as some of the topics rely heavily on the theory and results presented in the papers.

Papers and authors contribution

David Huard and Alain Mailhot. A bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc". *Water Resources Research*, 42:W07416, 2006. doi: 10.1029/2005WR004661.

David Huard and Alain Mailhot. Calibration of hydrological model GR2M using Bayesian uncertainty analysis. *Submitted to Water Ressources Research*, 2007.

David Huard, Guillaume Évin and Anne-Catherine Favre. Bayesian Copula selection *Computation statistics & data analysis*, 51, pp, 809–822, 2006. doi: 10.1016/j.csda.2005.08.010

In the first article, A. Mailhot is responsible for the theoretical computations related to the matrix form of the ABC model and the integration of input errors. D. Huard extended the framework suggested by Kavetski by adding a structural error term and realizing the importance of the prior for the true input. Numerical implementation, simulations and redaction of the article were done by D. Huard.

The second article is entirely the result of D. Huard work. A. Mailhot reviewed the final version of the written paper.

The original idea and the theoretical development leading to the third paper are constributions from D. Huard. A-C. Favre and G. Évin wrote respectively the sections on copula theory and the literature review on copula selection methods. Coding and simulations were done collaboratively by G. Évin and D. Huard.

Contribution to science

The thesis discusses issues related to the calibration of hydrological model in presence of multiple sources of uncertainties. More precisely, the work explores the subtleties that affect calibration and influence the calibration results and their interpretation. The methods used in this thesis, Bayesian analysis and various integration methods, are commonplace, and the originality of the thesis lies in their application to specific issues such as: the influence of the prior for the true input variable, the influence of the initial state uncertainty, the selection of priors for model comparison, the effect of error model assumptions on posterior parameter distributions and the validation of models calibrated in presence of input errors. The thesis also shows how structural errors can be identified and separated from input and output errors based on assumptions about their respective distribution. Although the conclusions are based on simulations run with monthly models, the general ideas and constatations are likely to hold for higher resolution models, once appropriate error models are defined for those cases.

Remerciements

J'ai peur de décevoir le lecteur qui s'attendrait au récit de mes épreuves car, pour ma part, l'expérience doctorale se compare à un fleuve tranquille, parsemé de quelques rapides en fin de parcours, certes, mais sans chavirement, et dans l'ensemble, fort agréable (ahh, c'est beau, le recul).

Il n'en reste pas moins qu'éviter les ennuis est un art bien plus grand que de les résoudre, et c'est en ce sens que vont mes remerciements. À Alain, donc, pour la direction de ce doctorat, les discussions, les voyages, l'appui inconditionel, la confiance et les encouragements ; j'espère un jour pouvoir te rendre un service comparable à celui que tu m'as rendu. À mes collègues de l'INRS, Guillaume, Romain, Sophie, Dany, Jean-Philippe, Renaud, Julia pour les bons moments passés ensemble, l'amitié et la camaraderie qui sont les éléments nutritifs essentiels d'un milieu de travail. À ceux de McGill (LP, JF, Yan, Bruno) pour leur accueil chaleureux, la pétanque et les barbecue. À Anne-Catherine pour le support, les encouragements et les multiples lettres de références. À Marie-Claude et Bernard pour les vacances reposantes. À Pierre et Marcelle pour tout. Et finalement, à Laurence, Jeanne, Madeleine et Marie qui sont la source de mon inspiration et mon énergie, le moteur et le catalyste de ce travail.

Contents

List of Tables

List of Figures

Résumé

1.1 Introduction

Un modèle hydrologique est un algorithme permettant de prédire l'écoulement de l'eau d'un bassin versant vers un exutoire à partir de mesures expérimentales. Ces mesures peuvent être, par exemple, de type météorologiques (précipitation, température, ensoleillement), géographiques (topographie), pédologique (perméabilité, niveau de la nappe) ou humaines (présence de puits, gestion de barrage). Il existe une énorme variété de modèles, certains possédant seulement quelques paramètres et une structure simple, et d'autres plus complexes, décrivant l'écoulement souterrain, les interactions de surface, l'utilisation du sol, etc, et ce à des résolutions spatiales et temporelles élevées. Pour pratiquement tous ces modèles, une étape préalable à leur utilisation s'impose : la calibration. Calibrer un modèle consiste à estimer les paramètres permettant à celui-ci de satisfaire optimalement à une ou plusieurs conditions. Généralement, les paramètres sont choisis de manière à maximiser une fonction décrivant la concordance entre les prédictions du modèle et des mesures expérimentales. Or, pour être utilisé, le modèle doit être conditionné par des données d'entrée (par exemple la précipitation et l'évapotranspiration), qui sont mesurées sur le terrain et souvent entachées d'incertitudes de toutes sortes.

Une question préoccupant énormément les hydrologues est de savoir quelle sera l'influence de ces erreurs sur la calibration du modèle. Il se trouve que plusieurs études ont montrées que les erreurs sur les données d'entrée des modèles corrompent la calibration, et donc peuvent fausser les prédictions subséquentes de ces modèles. La réaction des hydrologues a été de chercher des méthodes de calibration permettant de tenir compte de ces erreurs sur les données, ainsi que des erreurs commises par le modèle lui-même. Cette prise en compte du modèle d'erreur structural permet de définir une approche qui intègre l'ensemble des sources possibles d'incertitude associés à la mise en place d'un modèle hydrologique. Ce travail de doctorat s'inscrit dans cette mouvance et étudie une méthode de calibration permettant de tenir compte de multiples sources d'incertitude.

La méthode proposée s'appuie sur un article publié récemment par *Kavetski et al.* (2003) qui propose de considérer les erreurs sur les données d'entrée et de sortie par le biais de modèles d'erreurs. Ces modèles d'erreurs, qui permettent d'interpréter les données en spécifiant les caractéristiques des erreurs (biais, corrélation, hétéroscédascité) sont en quelque sorte *greffés* au modèle hydrologique via l'analyse bayésienne. Dans cette thèse, nous généralisons cette idée par l'ajout d'un modèle d'erreur structural permettant de traiter les erreurs imputables au modèle lui-même. La thèse présente deux applications de la méthode à deux modèles hydrologiques simples. Bien que les modèles hydrologiques soient d'un intérêt pratique limité, ils permettent d'étudier certaines questions d'ordre théorique : l'influence des modèles d'erreurs, leur sélection, la validation des modèles dans un contexte d'erreur sur les entrées, l'impact des distributions *a priori* (*priors*) et la mesure de l'efficacité de la calibration.

Dans les pages qui suivent, nous discuterons d'abord plus en détail de la problématique et des méthodes de calibration ayant été publiées au cours des dernières

années. Suit une description de la méthode proposée, puis les résultats principaux issues des expériences menées avec les modèles "abc" et GR2M. Finalement, des conclusions générales sur les défis soulevés par la calibration de modèles plus complexes sont émises, en parallèle avec des pistes de solutions à envisager.

1.2 Revue de littérature

Selon *Gupta et al.* (2003), l'amélioration de la calibration des modèles hydrologiques passe par la résolution de deux problèmes majeurs : arriver à tenir compte des incertitudes sur les données et le modèle, et élargir le spectre des indicateurs de performance des modèles. Cet intérêt des hydrologues pour les incertitudes vient de la combinaison de deux facteurs : 1. les données d'entrée des modèles hydrologiques (précipitation, évapotranspiration) sont parfois entachées d'incertitudes importantes, 2. ces erreurs sur les données biaisent les paramètres estimés par les méthodes de calibration traditionnelles.

1.2.1 Les moindres carrés

La méthode la mieux connue et probablement la plus utilisée pour calibrer un modèle est la méthode imaginée par Gauss il y a plus de deux cent ans, les moindres carrés (SLS). Cette méthode consiste à estimer les paramètres $\boldsymbol{\theta}$ minimisant la différence élevée au carrée entre les prédictions d'un modèle $f(\boldsymbol{x}, \boldsymbol{\theta})$ et des données expérimentales $\tilde{\boldsymbol{y}}$:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \sum_i [\tilde{y}_i - \mathcal{M}(\tilde{x}_i, \boldsymbol{\theta})]^2 \right\}.$$

Cette méthode fonctionne admirablement bien lorsqu'une condition essentielle est respectée, soit que les données d'entrée du modèle, $\tilde{\boldsymbol{x}}$, sont exactes. Le succès obtenu

par cette méthode est dû en grande partie au fait qu'elle se base sur la représentation du modèle d'erreur par une distribution normale. En effet, la minimisation des différences carrées est équivalente à la maximisation de la probabilité d'occurrence des données lorsque la probabilité est définie par une loi normale centrée :

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \ \{\mathcal{N}(\tilde{\boldsymbol{y}} \mid f(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}), \sigma)\},$$

où σ dénote la déviation standard des erreurs. Or, l'une des propriétés de la loi normale est qu'elle est la distribution maximisant l'entropie lorsque seule la moyenne et la variance d'une variable stochastique sont connues. Autrement dit, c'est la distribution la plus générale qui soit pour décrire une erreur aléatoire typique. En conséquence, en l'absence d'information sur la distribution des résidus, il est tout à fait correct d'utiliser la méthode des moindres carrés, et ce, même si, *a posteriori*, il se trouve que la distribution des résidus ne suit pas une loi normale.

Malgré les remarquables propriétés de cette méthode, certaines difficultés surviennent lorsque le nombre de paramètres à estimer augmente. En effet, à mesure que l'on ajoute des paramètres, la dimension de l'espace à explorer pour trouver le minimum augmente, ainsi que les risques de tomber sur un minimum local, plutôt que le minimum global. Afin de pallier ces difficultés, des méthodes telles que SCE-UA (*Duan et al.*, 1992) ou les algorithmes génétiques sont utilisées (voir *Duan* (2003) pour une revue des méthodes d'optimisation globales.)

En hydrologie, l'utilisation de la méthode des moindres carrés pose également un autre type de difficulté. En effet, les sorties d'un modèle hydrologique incluent généralement des régimes au comportement fort différents : les crues et les étiages. En appliquant directement la méthode des moindres carrés, il arrive souvent que les paramètres choisis reproduisent mal les périodes d'étiage pour la simple raison que

les sorties sont faibles et donc les erreurs commises par une mauvaise paramétrisation sont également comparativement faibles. Une solution peut être alors d'utiliser une loi normale hétéroscédatique (*Sorooshian and Dracup*, 1980) ou bien de minimiser d'autres fonctions qu'une différence élevée au carré (*Gupta et al.*, 1998).

La calibration de modèles hydrologiques constitue donc un problème complexe : le comportement à reproduire possède une grande variabilité, le nombre de paramètres est souvent élevé, et finalement, les données servant à calibrer le modèle sont incertaines, ce qui invalide l'hypothèse de base de la méthode des moindres carrés.

1.2.2 Les sources d'incertitudes

Les données de base servant à forcer et à calibrer les modèles hydrologiques sont, de façon générale, les précipitations, l'évapotranspiration et les débits. D'autres variables peuvent être considérées selon le modèle, telles que le niveau de la nappe phréatique, le pompage (lorsqu'il est connu), l'utilisation du sol, les cycles végétatifs, etc. Dans les climats humides, la variable qui domine est la précipitation. Les précipitations sont traditionnellement mesurées par une station météorologique, typiquement un entonnoir collectant l'eau de pluie. Ces stations sont parfois automatisées, parfois opérées par du personnel. Dans tous les cas, les mesures tendent à sous-estimer la précipitation réelle en raison de nombreux facteurs, le principal étant le vent. En effet, à mesure que la vitesse du vent augmente, il se crée une zone de haute pression autour de l'appareil, qui dévie les gouttes de pluie de l'entonnoir. Les pertes peuvent aller jusqu'à 30% de la précipitation totale, et jusqu'à 100% pour la neige (*Yang et al.*, 1999). Une autre source d'erreur provient de l'évaporation des gouttelettes d'eau qui adhèrent à l'entonnoir. Enfin, les stations automatisées fonctionnant sous le principe de l'auge à bascule doivent être calibrées (ajustement

de la fréquence de l'oscillation à l'intensité de la pluie), et cette calibration peut constituer une source significative d'incertitude. À ces différentes sources d'erreurs vient s'ajouter des considérations plus pragmatiques : le déplacement de la station, le changement du type de station, le remplacement du personnel, la modification de l'environnement (la végétation qui grandit, la construction de bâtiments), le dysfonctionnement de l'appareil, la négligence de l'opérateur, les pertes de données, etc.

Enfin, même si la mesure de la précipitation était parfaite, un problème de taille subsiste, nommément la représentativité de la mesure. En effet, une station ne mesure la précipitation que sur une surface équivalente à celle d'une assiette. De l'autre côté, le modèle hydrologique nécessite une mesure de la précipitation à l'échelle du bassin versant ou dans les modèles distribuées, à l'échelle des points de grilles, espacés dans le meilleur des cas de quelques kilomètres. Or, les précipitations peuvent exhiber une variabilité spatiale importante. La mesure expérimentale (en un point précis) et la variable du modèle (moyenne spatiale) sont donc fondamentalement différentes. Pour estimer la précipitation globale, on doit ainsi extrapoler les valeurs ponctuelles à l'aide de méthodes de krigeage. La validité de l'extrapolation dépendra de la superficie sur laquelle elle est appliquée, du nombre de stations incorporées et du type d'averse. Les orages courts et intenses sont généralement très localisés alors que les averses qui s'étalent sur plusieurs heures couvrent des territoires plus vastes. Si une averse passe directement au-dessus de la station et possède une superficie inférieure à celle du territoire d'extrapolation, la précipitation sera surévaluée. À l'inverse, si l'averse évite la station, la précipitation sera sous-estimée. Ce problème de représentativité de la mesure est discuté dans *Habib et al.* (2001).

La situation n'est pas tellement plus reluisante pour l'évapotranspiration et les

débits. L'évapotranspiration dépend principalement de la température, du rayonnement solaire et du type de végétation. Le grand nombre de modèle d'évapotranspiration en circulation témoigne de la difficulté d'estimer celle-ci adéquatement (*Oudin et al.*, 2005b). Les débits, quant à eux, sont mesurés par des relations hauteur-débit (*stage-discharge*), ou plus récemment par des sonars Doppler. Les incertitudes sur les décharges proviennent donc directement de l'incertitude sur la relation entre le niveau d'eau et le débit, affectée, entre autre, par la morphologie changeante du lit des rivières et les courtes périodes de mesures disponibles pour en effectuer la calibration.

Ces sources d'incertitudes qui affectent les données de calibration des modèles hydrologiques se mêlent aux incertitudes issues du modèle lui-même. Ces incertitudes structurales peuvent être divisées en deux grandes catégories: l'incertitude associée aux processus stochastiques, et celle provenant du fait que le modèle est une représentation simplifiée des processus naturels. Bien que cette division soit légèrement artificielle (un processus stochastique n'est au fond qu'un processus déterministe extrêmement sensible aux conditions environnantes), elle en demeure pertinente puisqu'elle sous-entend qu'il est possible d'améliorer la réprésentation des processus hydrologiques de manière à ce que l'erreur structurale ne soit finalement dû qu'aux processus stochastiques.

1.2.3 Les impacts des incertitudes

Les incertitudes associées aux données sont par nature aléatoire. On pourrait donc espérer que ces erreurs, en moyenne, s'annulent et que le comportement du modèle n'est pas affecté outre mesure par ces erreurs. Ce n'est toutefois pas la cas. En effet, il semble y avoir concensus parmi la communauté hydrologique pour dire que les

erreurs aléatoires sur les précipitations peuvent affecter les modèles hydrologiques et nuire à la qualité des prédictions (*Oudin et al.*, 2005a; *Kavetski et al.*, 2003; *Andréassian et al.*, 2001; *Troutman*, 1982) ; c'est-à-dire que la performance des modèles, mesurée notamment par le critère de Nash-Sutcliffe, diminue à mesure que l'amplitude des erreurs augmente. Il est à noter que l'influence des erreurs d'entrée varie considérablement selon le modèle, le climat et plus généralement, par le type de bassin considéré.

1.2.4 Méthodes de calibration

L'intérêt pour des méthodes de calibration tenant compte des incertitudes ne date pas d'hier et il existe bon nombre de méthode de calibration cherchant à réduire l'effet des erreurs sur la calibration. Les prochaines pages présentent brièvement quelques unes des méthodes proposées récemment permettant de tenir compte des multiples sources d'incertitude pouvant affecter la modélisation hydrologique.

BATEA

BATEA, *Bayesian Total Error Analysis* (*Kavetski et al.*, 2003) est une méthode bayésienne fondée sur les moindres carrés totaux (Total Least Squares). L'idée de base est de ne plus assumer que les entrées sont exactes, mais plutôt d'inférer leur valeurs, au même titre que les paramètres du modèle. Ainsi, en faisant la distinction entre les vraies entrées $\boldsymbol{x}$ et les entrées mesurées $\tilde{\boldsymbol{x}}$, nous pouvons écrire :

$$p(\boldsymbol{\theta}, \boldsymbol{x} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{x})\, p(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, p(\boldsymbol{\theta}, \boldsymbol{x}), \tag{1.1}$$

où les termes de droites représentent respectivement le modèle d'erreur de sortie (la vraisemblance), le modèle d'erreur d'entrée et les *prior* des paramètres et des vraies valeurs d'entrée. Dans l'article cité, les auteurs appliquent cette méthode au modèle "abc" (sur une base horaire) et à TopModel. En utilisant des données synthétiques auxquelles ils imposent un bruit aléatoire, ils parviennent à montrer que la méthode de calibration standard (SLS) ne converge pas vers les vraies valeurs de paramètres. Par contre, en utilisant BATEA, ils parviennent à estimer adéquatement la distribution des paramètres.

Afin de bien comprendre les problématiques reliées à la calibration, il faut réaliser que le fait que SLS ne parvienne pas à estimer les vrais paramètres du modèle n'est pas problématique en soi. En effet, les paramètres des modèles sont très souvent de nature conceptuels ; ils n'ont pas d'équivalent physique mesurable. Le fait que l'on estime ou non leur vraie valeur n'a pas réellement d'importance, tant que les prédictions réalisées par les paramètres demeurent réalistes. Nous reviendrons sur cet aspect de la validation plus loin.

Fait à noter, le modèle d'erreur d'entrée utilisé par les auteurs de BATEA découpe la série de précipitations en averses : ils supposent que, pour une averse donnée, toutes les précipitations sont affectées par la même erreur (un facteur multiplicatif). Autrement dit, si une averse dure 15 heures, ils supposent qu'en multipliant les 15 mesures de précipitations par un seul facteur, on obtiendra les 15 vraies valeurs de précipitation.

Le schéma d'incertitude utilisé dans cette thèse ajoute à celui de *Kavetski et al.* (2003) deux sources d'incertitude: l'incertitude structurale et celle dûe à l'ignorance des conditions initiales. De plus, les hypothèses sur les erreurs d'entrée sont relaxées, permettant de considérer n'importe quel type d'autocorrélation entre erreurs.

SODA

Dans un article de *Vrugt et al.* (2005), la méthode de calibration SODA est proposée, basée sur les ensembles de filtres de Kalman. L'idée maîtresse est que l'évolution d'un modèle déterministe dépend étroitement des conditions initiales qui sont souvent difficiles à estimer. Autrement dit, pour décrire de manière réaliste les prédictions d'un modèle, il faut considérer un ensemble de conditions initiales (des états), et faire évoluer chacun de ces états afin d'obtenir un portrait complet de la situation. Évidemment, pour des modèles complexes comme les modèles atmosphériques ou océaniques, il est hors de question de faire tourner le modèle les milliers de fois nécessaires pour obtenir ce type de portrait.

Au lieu de décrire l'évolution individuelle des états, les équations de Kalman permettent de calculer l'évolution moyenne du modèle, ainsi que la matrice de covariance décrivant l'étalement des états autour de la moyenne. L'avantage de cette méthode est qu'elle permet de tenir compte en une seule itération de l'incertitude sur les conditions initiales. Le désavantage majeur est que ce faisant, elle linéarise le comportement du modèle. Pour des modèles hautement nonlinéaires tels les modèles hydrologiques, cela peut conduire à des résultats erronés. La solution à ce problème est de considérer un ensemble de solutions moyennes (EnKF, *Ensemble Kalman Filters*, (*Evensen*, 1994)). De plus, en joignant une boucle Monte Carlo, il devient possible d'intégrer les incertitudes structurales et les incertitudes sur les données d'entrée à l'analyse, plutôt que de se limiter à l'incertitude sur les conditions initiales. Toutefois, la méthode EnKF fait l'hypothèse que les paramètres sont fixées, et donc ne permet pas directement d'en estimer la valeur à partir d'une série de données expérimentales.

La méthode SODA vient combler cette lacune et proposant de combiner EnKF

avec une boucle d'estimation des paramètres. L'idée est donc d'utiliser EnKF pour calculer une série temporelle de sorties pour un ensemble de paramètres, calculer la probabilité *a posteriori* de la série générée par chaque paramètre, et se servir de cette probabilité pour explorer l'espace des paramètres. Cette exploration utilise la méthode SCEM-UA (*Vrugt et al.*, 2003) afin de minimiser le nombre de calculs et converger rapidement vers la distribution *a posteriori* des paramètres. L'avantage principal de cette combinaison EnKF+optimisation est qu'il devient possible d'estimer récursivement les paramètres. Autrement dit, si une nouvelle donnée devient disponible, il est possible d'utiliser cette méthode pour mettre à jour la distribution de paramètres, sans avoir à traiter à nouveau la série entière. Les auteurs appliquent SODA au modèle hydrologique HYMOD et rapportent des résultats jugés satisfaisants (basés sur des données synthétiques).

Filtre de particules

Les filtres de particules (*Doucet et al.*, 2000) permettent de calculer la distribution des paramètres d'un modèle de manière relativement simple (du moins en théorie.) L'avantage principal de cette méthode est qu'elle se prête naturellement à l'estimation récursive des paramètres. L'idée maîtresse est de générer des particules, chacune représentant une réalisation aléatoire des paramètres du modèle, et d'assigner à chacune d'elles un poids proportionnel à la probabilité des paramètres qu'elle représente. Lorsqu'un nouveau jeu de donnée est disponible, il est aisé d'en calculer la vraisemblance et mettre à jour le poids de chaque particule. Les particules associées à des paramètres probables finissent par obtenir un poids élevé, alors que le poids des particules improbables tend vers zéro. L'espérance d'une fonction peut être calculée en pesant la contribution de chaque particule par son poids.

Si la méthode est simple en théorie, son application comporte quelques difficultés techniques. La principale vient du fait qu'après quelques itérations, la plupart des particules obtiennent des poids proches de zéro. Conséquemment, le nombre de particules qui contribuent de manière significative à l'espérance d'une fonction diminue rapidement, ce qui dégrade la qualité des résultats. Il est possible de pallier ce problème en ré-échantillonnant des particules à chaque itération, mais cela entraîne un appauvrissement de la diversité des particules, ce qui exige d'autres mesures correctives.

Moradkhani et al. (2005) applique le filtre de particules afin de calibrer le modèle hydrologique HYMOD. Bien que les auteurs rapportent des résultats satisfaisants, il est à noter que, dans leurs simulations, les incertitudes sur les données d'entrée et les erreurs structurales ne sont pas considérées. Néanmoins, le filtre de particule se prête admirablement bien à l'estimation récursive des paramètres et cela constitue probablement son principal atout.

1.3 Discussion

Le point de départ de toute méthode de calibration est le schéma d'incertitude qui décrit la relation entre les données, les erreurs et le modèle. Dans les lignes qui suivent, nous définissons un tel schéma englobant les erreurs sur les données d'entrée, les données de sortie et sur le modèle lui-même. Par la suite, le schéma est converti, grâce à l'analyse bayésienne, en une équation décrivant la distribution *a posteriori* des paramètres. Cette équation est résolue analytiquement pour le modèle "abc", puis par simulations Monte Carlo pour le modèle GR2M. Ces deux méthodes de résolution font l'objet des articles constituant la seconde partie de cette thèse. Les

résultats issus de ces simulations sont analysés et discutés dans les sections 1.3.2 et 1.3.3.

1.3.1 Schéma conceptuel

Avant d'attaquer la question de la calibration en présence d'incertitudes, un schéma décrivant la relation entre les différentes sources d'incertitudes, les données et le modèle doit d'abord être proposé. Ce schéma est décrit dans *Huard and Mailhot* (2006) et reproduit à la figure 4.1. L'hypothèse de base est qu'il existe des variables, les *vraies* données d'entrée et les *vraies* données de sortie, liées entre elles par un *vrai* processus physique. Les mesures de laboratoires ou de terrain, quant à elles, sont une approximation des vraies valeurs, et le modèle, une simplification du vrai processus. Pour arriver à calibrer le modèle, on doit inférer les vraies valeurs à partir des données, et comparer les sorties simulées aux sorties mesurées. L'inférence des vraies valeurs se base sur des modèles d'erreurs, décrivant la distribution de probabilité pour chaque source d'erreur. On retrouve donc un modèle d'erreurs d'entrée, de sortie et un modèle d'erreurs structurales, qui doivent être défini *a priori*.

À partir du schéma conceptuel, il est relativement aisé de passer aux équations en utilisant le théorème de Bayes (une dérivation complète est donnée dans l'annexe A de *Huard and Mailhot* (2006)) :

$$\begin{aligned} \mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \boldsymbol{\phi_0}) &= \iint \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \boldsymbol{\phi_0}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \\ &= \iint \mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) \, \mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x}) \, \mathrm{p}_{str}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi_0}) \, \pi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \cdot \frac{\pi(\boldsymbol{\theta})}{\mathrm{p}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \boldsymbol{\phi_0})}, \end{aligned} \tag{1.2}$$

où p_{out}, p_{in}, p_{str} sont respectivement les modèles d'erreurs de sortie, d'entrée et

structurales. La distribution *a posteriori* $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \boldsymbol{\phi_0})$ décrit la probabilité des paramètres d'un modèle, connaissant les données d'entrée et de sorties mesurées, ainsi que les conditions initiales $\boldsymbol{\phi_0}$. La formule (1.2) contient une intégration multiple sur $\boldsymbol{x}$ et sur $\boldsymbol{y}$, les variables de nuisance décrivant les vraies valeurs d'entrée et de sortie. C'est en traitant les vraies valeurs d'entrée comme des variables de nuisance et en intégrant sur celles-ci qu'il est possible de tenir compte de l'effet des erreurs d'entrée lors de la calibration et de la validation. Toutefois, cette intégrale est typiquement insoluble analytiquement, sauf dans quelques cas où le modèle est linéaire et les modèles d'erreurs ont une forme gaussienne. La première application de la méthode considère un tel cas, où il est possible d'intégrer sur les variables de nuisance de manière relativement directe. La deuxième application, traite un cas plus général de modèle nonlinéaire, où l'intégrale doit se faire de manière numérique en utilisant les méthodes de chaînes de Markov Monte Carlo (MCMC).

1.3.2 Solution analytique : application au modèle "abc"

L'équation 1.2 possède une solution analytique si certaines conditions précises sont satisfaites : le modèle est une fonction linéaire des données d'entrée, les modèles d'erreurs sont de types gaussiens, tout comme la distribution *a priori* des données d'entrée. La dérivation de la solution pour la distribution *a posteriori* est donnée dans ce cas par l'équation (18) de *Huard and Mailhot* (2006). Ceci dit, la grande majorité des modèles hydrologiques sont fortement nonlinéaires. La solution analytique a donc un intérêt pratique limitée, et nous l'étudions ici afin de de vérifier et d'étudier un certain nombre de questions théoriques. Le modèle choisi est "abc" (*Fiering*, 1967), un modèle mensuel à trois paramètres, fréquemment utilisé parce que très simple et pédagogique.

En utilisant des données synthétiques, il nous est possible de comparer le résultat de la calibration avec les vraies valeurs de paramètres ayant servi à générer les données. Les résultats suggèrent que le fait d'intégrer l'incertitude sur les données d'entrée permet d'améliorer légèrement l'estimation des paramètres. Cette amélioration est cependant beaucoup moins spectaculaire que celle obtenue par *Kavetski et al.* (2003). Deux facteurs contribuent à expliquer cette différence. Le premier est que nos séries de données ne dépassent pas 100 mois[1]. Pour estimer trois paramètres en présence d'incertitudes importantes, cela reste assez peu. Deuxièmement, notre modèle d'erreurs d'entrée est beaucoup plus permissif, allouant à chaque mesure de précipitation une erreur propre, alors que Kavetski attribue une seule erreur par averse.

Afin de mieux comprendre l'effet des modèles d'erreur, nous avons effectué la calibration d'une ligne droite en présence d'incertitudes sur les données d'entrée. Cela nous a permis de constater un fait important : en intégrant l'incertitude sur les données d'entrée, la distribution du paramètre (ici la pente de la droite) converge vers la vraie valeur pour un nombre suffisant de données, si et seulement si le *prior* pour les vraies entrées $\pi(\boldsymbol{x})$ est adéquat. L'influence de ce *prior* varie évidemment selon le modèle d'erreur d'entrée qui est spécifié. Si le modèle d'erreur suppose que les données d'entrée sont précises alors que ce n'est pas le cas, le *prior* joue un rôle insignifiant et on observera un biais sur la pente similaire à celui obtenu avec SLS. Si par contre le modèle d'erreur est très vague, le *prior* sur les vraies données d'entrée prend une influence considérable et peut jouer un rôle prépondérant sur la forme de la distribution de paramètres. Plus spécifiquement, l'effet sera de moyenner l'hydrographe : écrêter les crues et augmenter le niveau des étiages.

[1]Le temps de calcul est *grosso modo* proportionnel à n^2, où n est la longueur de la série. Pour des séries de plus de 100 mois, les calculs deviennent prohibitivement longs.

La conclusion de ces analyses est donc que le choix des modèles d'erreurs et du *prior* sur les vraies valeurs est extrêmement important afin de garantir la fiabilité de la calibration. Pécher par excès de prudence en spécifiant des modèles d'erreurs exagérément vague peut conduire à des résultats pires que ceux obtenus par une calibration standard. Il importe donc de bien savoir définir ces modèles d'erreurs, ce qui est loin d'être simple lorsqu'il s'agit de vraies données et non de séries synthétiques.

1.3.3 Solution générale : Application à GR2M

Étant donné que la grande majorité des modèles hydrologiques sont fortement nonlinéaires, il importe de savoir résoudre l'équation (1.2) de manière générale. La solution la plus simple est dans ce cas l'utilisation de l'algorithme Metropolis Monte Carlo. Cela nous permet d'utiliser n'importe quel modèle hydrologique et nous laisse une totale liberté en ce qui a trait au choix des modèles d'erreurs et des *priors*. Pour ce qui est du modèle, nous avons opté pour GR2M (*Mouelhi et al.*, 2006), un modèle parcimonieux comptant seulement deux paramètres et deux variables d'états.

Pour cette application (*Huard and Mailhot*, 2007), l'objectif est d'effectuer une véritable calibration utilisant de vraies données plutôt que des données synthétiques. Étant donné que GR2M ne comporte pas de module de neige, nous avons porté notre choix sur des bassins chauds et humides des États-Unis. Les données font partie de la base de données MOPEX (Model Parameter Experiment) (*Schaake et al.*, 2006b), une collection de séries hydrologiques destinées expressément à l'étude de méthodes de calibration. Le défi principal consiste à choisir des modèles d'erreurs réalistes et le plus informatifs possible. Malheureusement, la littérature à ce sujet est relativement mince, c'est-à-dire que de nombreuses études s'attardent à la qualité

des estimés de précipitations, mais traduire cette information en un modèle d'erreur n'est pas trivial. De plus, il n'est pas dit qu'un modèle d'erreur validé en un site soit applicable en un autre. Enfin, les données du MOPEX sont le résultat d'une ré-analyse poussée, où les précipitations ponctuelles sont extrapolées sur les points d'une grille puis intégrées sur la superficie couverte par le bassin versant de chaque station. Il est donc difficile de définir un modèle qui décrive les erreurs de mesures après un tel traitement. Ironiquement, les efforts déployés par MOPEX pour fournir des données de qualité compliquent l'élaboration des modèles d'erreurs.

En calibrant GR2M par analyse Bayésienne, certaines constats ont pu être dégagés. Tout d'abord, il importe de faire la différence entre les simulations d'un modèle et les observations. Une observation se sortie est une vraie valeur de sortie corrompue par une erreur de mesure (ou de représentativité), alors qu'une simulation est une vraie valeur de sortie corrompue par une erreur structurale. En ce sens, il est périeux de comparer ces deux quantités, puisque qu'elles sont fondamentalement diffèrentes. Afin de comparer rigoureusement des simulations de modèles à une série d'observation (afin par exemple de mesurer l'efficacité d'un modèle en mode validation), il importe de transformer les simulations du modèle en vraie valeur de sortie puis en observation synthétique. Autrement dit, en ajoutant une erreur structurale et une erreur de sortie aléatoires tirées de leur modèle d'erreur correspondant.

Aussi, il apparait que la méconnaissance des conditions initiales ne joue pas un rôle majeur dans la calibration typique de GR2M. En effet, quelles que soit les valeurs choisies par le modèlisateur, les variables internes du modèle convergent relativement rapidement par rapport à la durée d'une série hydrologique typique. Autrement dit, pour une série de plus de trois ans, l'influence de notre méconnaissance des conditions initiales est marginale.

En examinant les distributions *a posteriori* des paramètres calibrés selon différents choix de modèles d'erreur, il apparait que les hypothèses concernant les erreurs sur les données influencent significativement la distribution des paramètres du modèle. Autrement dit, la distribution de paramètres ne peut être interprétée indépendemment des modèles d'erreurs utilisés pour la calibration.

Enfin, la méthode Bayésienne permet, par le truchement des vraies variables d'entrée et de sortie, de séparer et d'identifier *a posteriori* les erreurs structurales des erreurs de sortie et d'entrée. Cette possibilité mérite d'être exploré plus en profondeur puisqu'elle pourrait éventuellement permettre d'identifier et d'améliorer les modèles hydrologiques.

1.3.4 Sélection de modèle

Dans un autre ordre d'idée, il serait possible d'améliorer la sélection Bayésienne de modèles hydrologiques (*Marshall et al.*, 2005) en y incorporant le concept de *prior parent* développé dans le troisième article (*Huard et al.*, 2006). Élaboré dans le contexte de la sélection de la meilleure copule, l'idée du prior parent est simplement d'assigner aux paramètres de chaque modèle un prior ayant la même signification physique, quelle que soit la paramétrisation du modèle. Dans le cas des copules, le troisième article propose de définir le prior parent sur la base du τ de Kendall, une mesure d'association très utilisée dans la littérature sur les copules. Pour ce qui est des modèles hydrologiques, une mesure englobante et générale à tous les modèles semble plus difficile à définir. Quelques possibilités viennent toutefois à l'esprit: le temps de récession, le temps de concentration, le temps de montée ainsi que d'autres propriétés statistiques de l'hydrographe (*Yue et al.*, 2002). En définissant un prior

sur ces quantités et le transformant en prior pour les paramètres de chacun des modèles, la comparaison des modèles n'en serait que plus rigoureuse.

1.3.5 Pistes de recherche

Suite aux recherches effectuées, un certain nombre d'avenues sont apparues comme des cibles de recherche intéressantes. En voici un bref résumé.

Sélection de modèles d'erreurs informatifs

Le plus grand défi de la calibration est probablement la sélection de modèles d'erreurs informatifs, c'est-à-dire de modèles d'erreurs ne se limitant pas à décrire les erreurs sur les données, mais cherchant à améliorer l'inférence des vraies valeurs. À ce titre, l'intégration de sources d'informations additionnelles (gravimétrie, images radar, satellites) et même d'information historique ou paléontologique nous semble une avenue de recherche à considérer sérieusement. Par ailleurs, les corrélations entre erreurs paraissent un sujet important, puisque toute forme de dépendance d'une erreur sur une autre réduit le caractère aléatoire des erreurs et permet de mieux les inférer.

Validation et mesure de l'efficacité des modèles

Les critères de validation des hydrologues (RMSE ou Nash-Sutcliffe) comparent les sorties mesurées aux sorties simulées en assumant les données d'entrée exactes. Or, dans un contexte d'erreur sur les données d'entrée, ces critères de performance ne sont plus appropriés. En effet, en choisissant un modèle d'erreur d'entrée très vague, il devient possible d'obtenir artificiellement des NSE proches de un. Afin de

réellement mesurer la concordance entre les données et les simulations du modèle, il convient donc de comparer autant les différences sur les sorties du modèle que les différences entre les valeurs d'entrée mesurées et les valeurs d'entrée inférées lors de la calibration. La validation des modèles et des modèles d'erreur passe donc par une définition objective et générale de la performance qui dépendra à la fois du modèle hydrologique et des modèles d'erreur considérés.

Application aux modèles journaliers

L'application telle quelle de la méthode proposée aux modèles journaliers pose deux défis. Le premier, d'ordre conceptuel, est de définir des modèles d'erreurs réalistes sur les pluies et les débits. La situation semble plus compliquée qu'en mode mensuel en raison de la structure plus complexe des erreurs, particulièrement en ce qui a trait à l'auto-corrélation des erreurs. La probabilité finie de précipitations nulles mérite également d'être prise en compte, par exemple par certaines distributions mi-discrètes mi-jointes (distribution de Tweedie (*Dunn and Smyth*, 2005)). Le deuxième défi est d'ordre pratique et concerne l'échantillonnage simultané de milliers de paramètres. En effet, les séries journalières sont beaucoup plus longues que les séries mensuelles, et intégrer au-delà de 7000 paramètres de nuisance pour calibrer dix ans de données ne peut se faire à l'heure actuelle avec un ordinateur personnel en un temps raisonnable. Cette contrainte limite l'utilisation d'une telle méthode pour la pratique hydrologique courante, du moins à court terme. Toutefois, dans un contexte de recherche, la parallélisation du code permettrait de traiter le cas journalier et d'identifier des hypothèses simplificatrices permettant de réduire les temps de calcul.

Erreurs structurales

Le fait de ne pas agréger toutes les erreurs dans un seul terme permet de discerner les différentes sources d'erreurs. En effet, au lieu de chercher à estimer les paramètres du modèle, il est possible d'estimer les erreurs commises, autant sur les données d'entrée, de sortie que par le modèle lui-même. Une application intéressante de cette séparation des erreurs est d'analyser la structure et les caractéristiques de la série d'erreurs structurales. Cette analyse pourrait permettre d'identifier les conditions pour lesquelles le modèle performe mal. De plus, la série d'erreurs structurales pourrait servir à la mesure de la performance du modèle, indépendamment de la qualité des données. Évidemment, toutes ces applications sont fortement dépendantes de la qualité des modèles d'erreurs.

1.4 Conclusion

Dans cette thèse, un modèle d'incertitude inspiré de celui de *Kavetski et al.* (2003) est suggéré, intégrant les erreurs structurales en plus des erreurs sur les données d'entrée et de sortie. Utilisant l'analyse bayésienne, ce schéma est traduit sous forme d'une équation décrivant la forme de la distribution *a posteriori* des paramètres. Cette équation est résolu analytiquement pour le modèle hydrologique mensuel "abc", et numériquement en utilisant l'algorithme Monte Carlo Metropolis pour le modèle GR2M. L'étude des résultats permet de tirer plusieurs conclusions d'intérêt général sur la calibration des modèles hydrologiques en présence de multiples sources d'incertitudes, avec un focus centré sur les incertitudes des données d'entrée.

Parmi ces conclusions, il importe de souligner l'importance du prior sur les vraies données d'entrée. Les travaux effectués montrent en effet qu'un prior vague ne

permet pas d'inférer correctement la distribution a posteriori des paramètres. Plus précisément, la distribution ne converge pas vers la vraie valeur à mesure que la série de données s'allonge. Une autre conclusion importante concerne la sensibilité des résultats aux modèles d'erreurs. Les résultats de cette thèse montrent en effet que la distribution des paramètres est intimement liée aux hypothèses formulées au sujet des erreurs de mesure. De plus, sous l'hypothèse d'erreurs en entrée, l'interprétation des résultats de calibration ne peut plus se faire sur la base de la comparaison entre les sorties simulées et les sorties observées. La mesure de l'efficacité du modèle doit tenir compte à la fois des résidus de sortie, mais aussi des résidus en entrée : la différence entre données d'entrée mesurées et vraies variables d'entrée inférées. Enfin, la spécification explicite des erreurs d'entrée, de sortie et structurales permet leur identification a posteriori. Autrement dit, la méthode permet d'extraire la distribution inférée des erreurs. Ceci permet, entre autres, de distinguer les contributions des erreurs structurales des erreurs sur les données.

Une constatation générale est que la sélection des modèles d'erreur constitue le plus grand défi au développement de nouvelles méthodes de calibration. C'est l'information comprise dans le modèle d'erreur qui permet d'interpréter les données fournies à la méthode de calibration. Par conséquent, la validité des modèles d'erreur est essentielle au succès de la calibration, quelle que soit la méthode choisie (BATEA, SODA, filtre de particule). Parallèlement, il ne faut pas s'attendre à ce que des modèles d'erreur vagues permettent d'améliorer significativement l'efficacité des modèles calibrés. Le nerf de la guerre en calibration est la somme des informations disponibles. Un modèle d'erreur doit par conséquent chercher à amalgamer le plus de sources d'informations possible afin de raffiner son estimation de la *vraie* valeur. L'intégration de sources d'informations additionnelles apparaît donc comme un autre des grands axes de recherches menant à une analyse plus rigoureuse de

la performance des modèles pouvant mener à un diagnostic plus précis des erreurs structurales et, conséquemment, à une amélioration des modèles.

Les modèles d'erreurs sont également au coeur de la question de la pertinence d'utiliser une méthode de calibration tenant compte de multiples sources d'incertitudes. L'utilité de ces méthodes dépend évidemment de plusieurs facteurs, notamment la sensibilité du modèle aux erreurs, l'importance d'établir des intervalles de confiance réalistes sur les prédictions, la capacité à formuler des modèles d'erreurs informatifs mais aussi le temps alloué à la calibration. En vertu de ces considérations, il semble peu probable qu'à court terme la méthode de calibration décrites dans cette thèse rejoigne les hydrologues en dehors des institutions académiques. Dans un contexte de recherche toutefois, la possibilité d'extraire les différentes sources d'erreur *a posteriori* et d'analyser les erreurs structurales pourrait permettre d'améliorer la qualité des modèles hydrologiques.

Part I

Synthesis

CHAPTER 2

Introduction

An hydrological model is an algorithm allowing one to predict from experimental measurements the flow of water from a watershed to its outlet. These measurements can be climatic (rainfall, temperature, radiation), geographic (topography), pedologic (soil conductivity, aquifer levels) or socioeconomic (wells, soil usage, dam management). This diversity of data leads to a diversity in hydrological models, from simple ones with few parameters, to complex distributed models with tens of parameters describing subsurface flow and surface interactions at high spatial and temporal resolutions (*Singh*, 1995). Nevertheless, for the vast majority of models, a preliminary step before the model can actually be applied to a given catchment is calibration. Calibrating (or fitting) a model consists in finding the set of model parameters maximizing the concordance of simulated outputs with measured outputs. The calibration step is crucial because the estimated parameters are then used to run the model to predict the future behavior of the catchment under different scenarios. In this sense, a large part of the validity of a model is closely dependent on its calibration.

2.1 Standard calibration procedure

The best known method to calibrate a model was developed by Carl Friedrich Gauss about two hundred years ago: the standard least-squares method (SLS).

The principle is deceptively simple: find the parameters that minimize the sum of squared differences between simulated and measured outputs

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \sum_i [\tilde{y}_i - \mathcal{M}(\tilde{x}_i, \theta)]^2 \right\},$$

where $\mathcal{M}$ stands for any model, θ for the model's parameters, $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$ for the measured inputs and outputs respectively. This methods works admirably well when one condition is satisfied: the measured inputs can be considered exact. When input data is corrupted by errors, however, there is no guarantee that the estimated parameters are reliable. In fact, many studies point out that in such cases, the parameters tend to be systematically biased by the input errors (section 3.2). This is particularly problematic for environmental scientists, since experimental data is seldom error-free.

2.2 Impact of input errors

It is difficult to describe globally the effect of input errors on the calibration of models. There are however some guidelines. It is well known that fitting a straight line with errors on the input variable leads to the underestimation of the slope (*York*, 1966). This is explained thouroughly in *Zellner* (1971) and *Huard and Mailhot* (2006, section 3). In hydrology, the general concensus is that input errors tend to bias the parameters of hydrological models. That is, even if errors are distributed normally around an idealized true value, the model parameters estimated through SLS calibration are over- or under-estimated, and the parameter uncertainty does not reflect accurately the underlying data uncertainties (*Oudin et al.*, 2006; *Kavetski et al.*, 2003; *Troutman*, 1982). More precisely, the biases tend to average the model's behavior, *i.e.* underestimate the floods and overestimate the base flow (*Huard and Mailhot*, 2006).

2.3 Objectives

Although the sensitivity of hydrological models to input errors has been known for over three decades, the interest for these issues has surged only relatively recently, as evidenced by a number of papers proposing novel calibration methods (chapter 3). Part of this interest is probably due to the recent availability of the computational power these problems require, but also to the rising complexity of models and concerns about uncertainties and their impact on the reliability of predictions. In fact, *Gupta et al.* (2003) define in their conclusion *Two Important Issues That Need Addressing*, the first one being:

> "To be comprehensive, future strategies for model identification should seek to account explicitly for *all* of the following sources of uncertainty:
>
> a) input uncertainty
> b) state uncertainty
> c) structural uncertainty
> d) parameter uncertainty
> e) output uncertainty (observation error)
>
> which could (we think) be treated within a Bayesian framework..."

Such a strategy should hence manage to take into account multiple sources of errors such that that prediction uncertainties faithfully represent the underlying data and structural uncertainties. Despite the clear interest in uncertainty assessment methods, "a significant part of the community is still reluctant to embrace the estimation of uncertainty in hydrological and hydraulic modeling" (*Pappenberger and Beven*, 2006). According to these authors, the one acceptable reason for which uncertainty analysis is still not standard practice is the lack of guidance in this matter. The objectives of this thesis are to provide some insight into uncertainty analysis through

1) the development of a Bayesian framework dealing with multiple sources of uncertainty, and 2) it's application to hydrological models to understand the issues that appear when confronting multiple sources of uncertainty inhydrological modeling.

2.4 Methodology

The method is based on a conceptual framework linking errors, data and model (section 4.1). This framework is based on the work of *Kavetski et al.* (2003), accounting for output and input uncertainty, to which two other sources of uncertainty are added: structural and initial conditions uncertainty. The error generating processes are described by error models: probabilistic distributions describing the occurrences of input, output and structural errors. These error models are grafted to the model, and a formula for the model's parameters posterior distribution can be derived using Bayesian analysis (section 4.2).

This posterior can be solved analytically under certain conditions, namely the linearity of the model and Gaussian error models. This special case is explored in the first paper (*Huard and Mailhot*, 2006), where the method is applied to the hydrological model "abc" (section 4.3). This example allows a better understanding of the effect of input error on the calibration and the pitfalls associated with it. In general, however, analytical solutions for the posterior parameter distribution do not exist and one must revert to numerical solutions.

One such numerical solution for the Bayesian calibration of non-linear model is Markov Chain Monte Carlo (MCMC) sampling. A limitation of this approach, however, is that the time necessary to find a solution on a standard desktop computer becomes prohibitive for typical daily hydrological time series. Therefore, in the

second paper (*Huard and Mailhot*, 2007), calibration is performed for GR2M, a monthly nonlinear hydrological model (section 4.4). Different issues related to model calibration with multiple sources of uncertainties are discussed, namely the impact of initial state uncertainty, the treatment of structural uncertainty and the effect of error models on the parameter distribution. The paper also shows how, using a Bayesian framework, model errors can be separated from data errors.

This separation of errors is believed to be strongly dependent on the choice of error models. Whereas input and output error models can be defined based on the measurement protocol and accuracy, structural error models appear much more difficult to define rigorously. One way to improve the understanding of structural errors might be by comparing different models among themselves and analyzing their respective error series. This comparison must, however, be done with care. Indeed, all models are not equally plausible, and hence results from each one of them should be weighted by the model's own probability. One elegant way to estimate a model's probability is using Bayesian model selection theory (*Bretthorst*, 1996). One difficulty with Bayesian model selection is the selection of a prior distribution for each model's parameters. Since the parametrization of each model is different, choosing uniform priors for the parameters of every model can have completely different meaning in terms of hydrological behavior, and artificially bias the results of model selection. To avoid such biases, the third paper proposes the concept of *parent prior*, allowing the definition of priors with a consistent meaning across models. The concept is applied to copula selection and section 4.5 discusses how to apply the same ideas in an hydrological modeling context.

CHAPTER 3

Literature review

This section reviews the main sources of uncertainties in hydrological modeling as well as their impact on model efficiency. A number of calibration methods proposed to deal with multiple source of uncertainties are then presented.

3.1 Sources of errors in hydrological modeling

The basic data needed to fit an hydrological models are rainfall, potential evapotranspiration (PE) and discharge. A slew of other variables can be considered, such as aquifer level, water usage, vegetative cycles and various other land surface forcing. In this thesis, the models used are fairly simple conceptual models with low data requirements, limited to rainfall, PE and discharge. The next sections discuss the sources of error affecting each one of these three variables.

3.1.1 Rainfall uncertainty

In humid regions, the dominant forcing variable is rainfall, traditionally measured by automatic or manual rain gauges, whether tipping buckets or simple tanks or both. Other means of measuring rainfall are by vertical radar sensing, using a 2-D video

disdrometer or an optical rain gauge *Krajewski et al.* (1998). Each measuring method has its own shortcomings and related uncertainties. In the following, a review of the dominant sources of errors is presented for both point rainfall estimates and rainfall spatial variability.

Point rainfall estimates are obtained using gauge devices, whether tipping bucket gauges, optical gauges or tanks. According to *Adam and Lettenmaier* (2003), gauge measurements of rainfall are systematically underestimated due to the following effects: wind deformation above the orifice, wetting losses due to the adhesion of droplets on the surface of the gauge, evaporation of water in the gauge, splashing of raindrops outside the gauge, calibration of gauge and treatment of trace amounts of precipitaiton. Among those factors, the most important is the wind-induced undercatch, reaching between 2% and 10% for liquid water (*Yang et al.*, 2005; *Adam and Lettenmaier*, 2003). Indeed, wind blowing across an obstacle tends to create a zone of high pressure around the gauge, deflecting the droplets away from the funnel (see figure 3.1.) This undercatch increases with wind speed and is also dependent on the size of raindrops, smaller particles being more strongly deflected outside the gauge (*Krajewski et al.*, 1998). As wind speed increases with altitude, gauge set on top of buildings are more likely to suffer from this type of undercatch compared to ground stations. The situation is even worse in high latitudes regions, where the large proportion of solid precipitation magnifies the undercatch. Indeed, in Arctic regions, undercatch can reach 100% of total precipitation (*Yang et al.*, 2005, 1999; *Mekis and Hogg*, 1999; *Metcalfe et al.*, 1997).

Another important source of error is the calibration of tipping-buckets rainfall gauges (TPR). In a static TBR calibration, the volume of water needed to tip the bucket is assumed to be independent of the rainfall intensity. This type of calibration, however, often results in the underestimation of rainfall. Indeed, some

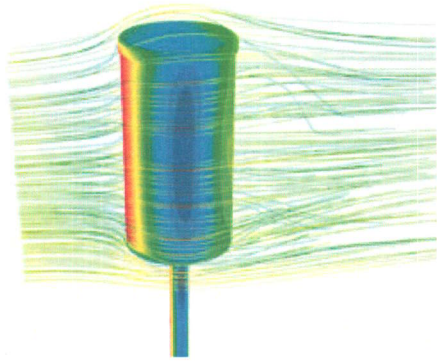

Figure 3.1: Streamlines colored by velocity magnitude with static pressure contours on the rain gauge (from *Newman and Kucera* (2005)).

water is lost after each tip as the bucket repositions itself. For intense convective storms, this may results in bias of 10 to 30% (*Humphrey et al.*, 1997). Static calibration is also problematic at small temporal scales (10-15 minutes) (*Molini et al.*, 2005), even for low rainfall intensity. Hence, it has been suggested to use dynamical calibration, in which the volume of water tipping the bucket is assumed to be dependent on the rainfall intensity (*Humphrey et al.*, 1997). Dynamic calibration is done by pumping water at a known rate to the TBR using a pump, and calibrating the tipping rate to the water flow for different intensities. This procedures allows the removal of mechanical errors systematically affecting TBR measurements.

Wetting losses result from raindrops or water films adhering to the gauge surface instead of filling a tank. They account for around 3% of error (*Adam and Lettenmaier*, 2003), with some variations depending on the evaporation rate and the measurement frequency; if a tank is emptied twice a day, it has less time to evaporate its content.

Other more pragmatic factors can influence data quality: debris obstructing instruments, displacement of the gauge station, replacement of station by a new model, change in manipulation protocol and personnel, changes in the station en-

vironment (growing vegetation, new buildings) and apparatus malfunction (mouses eating electric cables).

Now, even if all those sources of uncertainties were under control, there would still be the problem of representativity (or commensurability). Extrapolating the aerial rain distribution over a catchment from a handful of plate-size funnels can lead to over or under estimation depending on the location of the storm with respect to the stations (*Habib et al.*, 2001; *Blöschl*, 2001; *Willems*, 2001; *Morissey et al.*, 1995). For example, when comparing the estimates from a single gauge to a dense array of gauge (considered as the true value), *Wood et al.* (2000a) have shown that there is a considerable intra-grid variability, even for 2x2 km grids. In other words, at high temporal resolution, estimating rainfall from a single gauge leads, in this case, to errors of 30%. This situation is critical for distributed hydrological models, since they require a realiable description of the spatial variability of rainfall, preferably at the same scale as that used by the model. Without this spatial rainfall pattern, runoff estimation may be compromised, particularly for small catchments (*Chaubey et al.*, 1999; *Chaubey et al.*, 1999; *Lopes*, 1996). *Michaud and Sorooshian* (1994) have shown that under sampling the rainfall variability can cause errors of up to 58% for peak flow. In the province of Quebec, where many catchments are sparsely instrumented, the lack of this spatial information may be the major source of uncertainty in aerial rainfall estimates. However, this lack of representativity may be more critical for models operating at short time scales, where the timing of rainfall is important, than for monthly models, where one can expect representativity errors to average out, at least partially.

For high resolution distributed models, the spatial but also the temporal variability play a significant role (*Singh*, 1997). For instance, the direction of a storm with respect to the river flow influences the peak discharge; a storm moving along the flow

will result in higher peak flows than a storm moving counter to the flow. Similarly, different patterns for rainfall excess (constant, increasing, decreasing) directly affect the concentration time (*Singh*, 1997).

Although it is possible to install a dense network of rain gauges, installation and maintenance costs tend to be high and remote sensing methods are gaining popularity. Among those, radar rainfall measurements strive to improve the measurement of rainfall's spatial variablility. *Zawadski* (1973) studied the space-time correlation of storms and showed that although the studied storm is isotropic under 10km, at larger distances rainfall decorrelates nearly exponentially. In a study comparing a dense raingauge network to radar rainfall estimates on a 2 by 2km grid (Brue catchment, Somerset, England), (*Wood et al.*, 2000a,b) show that for a given grid cell, a single raingauge provide better estimate than the radar (over 15 minutes intervals). At the catchment scale, however, the radar performs better than a single raingauge. Radar data are however notoriously difficult to convert into rainfall estimates, and errors of 100% are not uncommon for light rain, while heavy rain show errors of around 30% (*Habib and Krajewski*, 2002).

Kriging methods are used extensively in geophysics and related fields to interpolate the value of some field at a given point from a set of point measurements (*Kyriakidis and Journel*, 1999). Generally, a kriging method will require the modeler to specify a number of measurements at point locations, some hypotheses about the mean and variance of the field under study and the covariance function (how the field changes with distance). Hence, kriging can be used to interpolate rain rates for instance, given gauge readings distributed over a watershed (*Sun et al.*, 2003). Kriging can even account for point measurement errors in the interpolation of the field (*Daley*, 1991). Kriging can also be used to combine radar data to raingauge

data (*Azimi-Zonooz et al.*, 1989), improving the spatial interpolation, especially in sparsely instrumented areas.

Willems (2001) proposes to describe the point rainfall errors due to the apparatus resolution, calibration errors and wind effects by a multiplicative normal error $\ln(R/\tilde{R}) \sim \mathcal{N}(0, \sigma)$, where R stands for the true rain and $\tilde{R}$ the measured rain. The variance σ^2 is given by

$$\sigma^2 = (.03)^2 + \left(\frac{.1\text{mm}}{R}\right)^2, \tag{3.1}$$

meaning a 3% relative error on total rainfall is assumed, along with a fixed error of 1 mm due to the apparatus resolution. *Willems* (2001) presents further variances for aerial rainfall based on the number and location of point rainfall estimates. A similar error model, also using a variance with fixed and proportional components, is used by *Weerts and El Serafy* (2006). *Petersen-Overleir* (2005) also uses a multiplicative normal error (bias and variance) to describe rainfall errors in the study of extreme rainfall.

The rainfall data used in the second paper is taken from the MOPEX database[1]. The source of this data is discussed in the second paper (section II) and more extensively in the MOPEX documentation (*Schaake et al.*, 2006a). The data, however, does not contain estimates of data accuracy.

3.1.2 Evapotranspiration uncertainty

Although a large part of the work presented in this thesis is concerned with errors on rainfall, the other input variable, potential evapotranspiration (PE) is not error-

[1] `nws.noaa.gov/oh/mopex/`

free. Since PE values are not directly measured but rather computed from models, PE errors depend on the model used and the variables it depends on. Some models estimate evapotranspiration from just the Julian day, radiation, or the minimum and maximum temperatures, where others include factors such as soil moisture, land use, vegetation type and growth season (*Xu and Singh*, 2002). In *Oudin et al.* (2005a,b), 27 different PE models are tested to assess the influence of PE estimates on model calibration. The conclusion of their study is that detailed EP description does not improve the efficiency for the models tested. In other words, most hydrological models are not sensitive enough to be affected by slight interannual changes in PE series. Error models for EP should hence focus not on individual PE values, but rather on the mean and intraannual variability of PE series.

3.1.3 Discharge uncertainty

Continuous discharge measurements are obtained using a stage measurement (the water level) and a rating curve. A rating curve is a function specific to a gauging station relating stage measurements to discharges. It is first defined by taking multiple measurements of stage and discharges under different conditions. The discharge is then plotted on a graph versus the stage and a functional relationship linking both variables is estimated, so that discharge can later be estimated directly from the stage. The reason why this is done is that while the stage can be measured automatically by an height gauge, estimating discharge is much more difficult. Indeed, discharge is measured by integrating the water velocity v over the river cross-section S (figure 3.2):

$$Q = \oint v \, \mathrm{d}S.$$

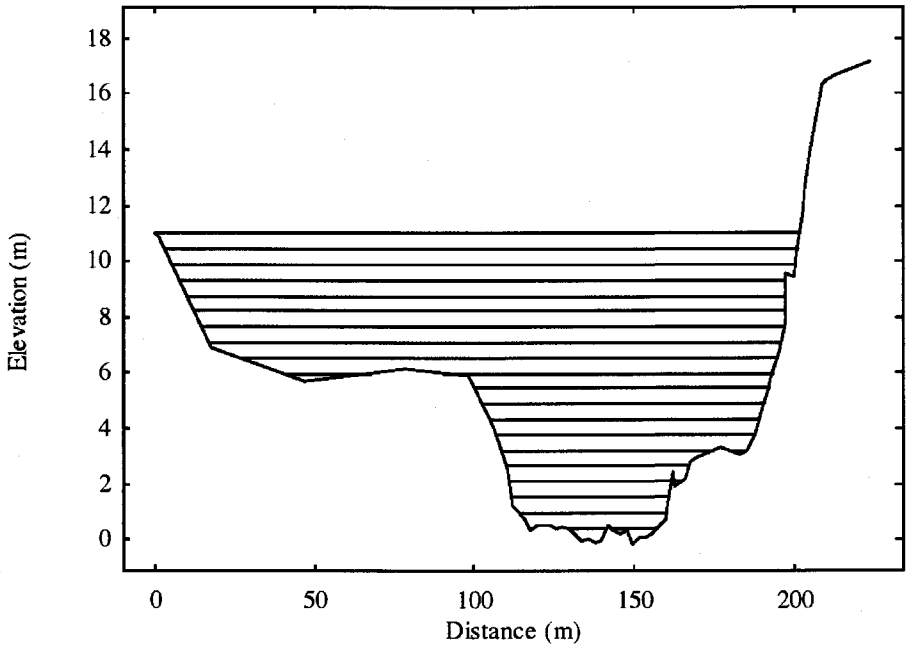

Figure 3.2: River cross section (from *Fenton and Keller* (2001)).

Measuring the velocity across the river at different depths can be a painstaking process, particularly during flood events when the weather is generally bad. Understandably, the rating curve uncertainty on rare and extreme events is generally greater than for small and frequent events. While Ultrasonic Doppler flow monitoring can be used to measure discharge more efficiently, there are a number of factors that introduce errors in the readings: lack of debris (no reflection), low water levels, multiple reflections, bias due to large reflectors, etc. (*Blake and Packman*, 2007).

Errors on the rating curves are caused by a number of factors (*Fenton and Keller*, 2001): flow unsteadiness, morphological changes in the river bed (sediment transport) or channel, changes in river bed roughness (vegetation, bed forms) and backwater effects from downstream obstacles. Flow unsteadiness in particular is the source of hysteresis in stage-discharge relationship. Indeed, from a theoretical point of view, discharge not only depends on stage but also on the river slope: water

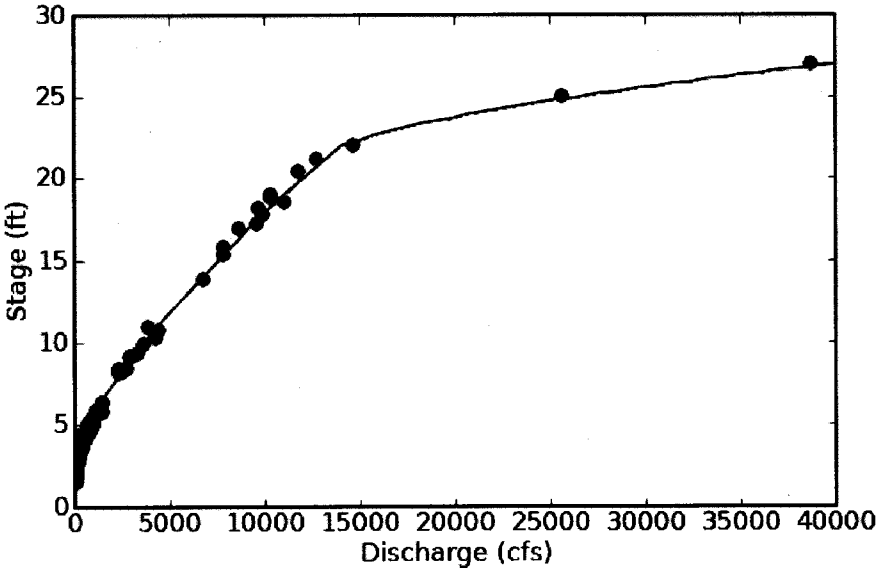

Figure 3.3: Rating curve of the Chunky River used by the USGS. The dots indicate individual stage-discharge measurements.

velocity is greater if the slope is high than if it low. Hence, before the peak of a flood, flow increases (raising limb), the slope is higher than in the steady state and the real discharge is higher than in the steady state. During flow recession, (falling limb), the slope is lower than in the steady state and the rating curve overestimates the discharge. In other words, during a flood event, the actual discharge would be given approximately by the loop shown in figure 3.4. Despite those sources of uncertainties, *Fenton and Keller* (2001) have suggested that in most cases, a rating curve neglecting the effect of the river slope provides reliable discharge estimates.

Formally, sampling uncertainty (finite number of stage/discharge measurements) and individual stage and discharge measurement uncertainties can be described individually to compute the overall discharge uncertainty (*Dymond and Christian*, 1982). This requires, however, considerable amount of information about the measurement apparatus and protocol and imposes constraints on the rating curve parametrization. Calibration of the rating curve also poses some difficulty, since hypotheses

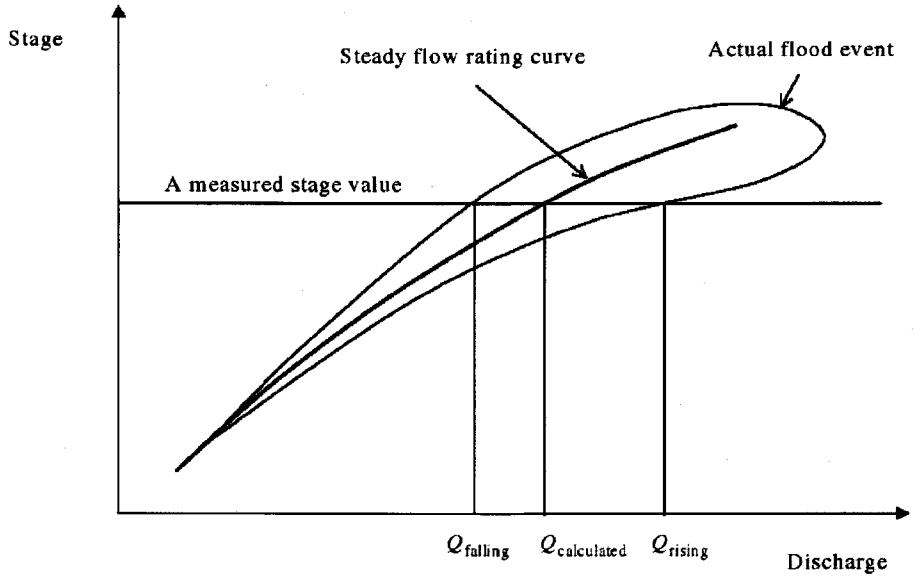

Figure 3.4: Difference in discharge caused by variations in river slope during a flood event, compared to the steady flow state approximated by the rating curve (from *Fenton and Keller* (2001)).

have to be made regarding the heteroscedasticity of discharge errors. *Petersen-Overleir* (2004) has shown that non-linear least squares implies a restricted type of heteroscedasticity that can lead to dubious values for estimated discharges. Bayesian methods have been used to calibrate the parameters of the rating curve (*Moyeed and Clarke*, 2005), providing uncertainty estimates for discharges consistent with data and parameter uncertainty.

3.1.4 Structural uncertainty

The most elusive source of errors is probably structural. Hydrological processes are extremely complex, mostly because they operate on multiple scales and because minor local disturbances can modify the global flow pattern (imagine a thin layer of clay drastically modifying the hydraulic conductivity of an entire soil column.)

Modeling those processes is a substantial challenge because gathering data is extremely labour intensive, and it is impossible to model all the processes at all scales. The errors arising from these compromises can be significant, no matter the calibration. Identifying the source of the dominant structural errors is probably one of the thoughest challenges hydrology has yet to face.

Even though structural errors largely depend on the actual model being used, there are still some hypotheses that should hold in general. The first one is that structural errors are probably heteroscedastic, that is, it seems reasonable to hold that the magnitude of errors depends on the magnitude of the event modeled. The second one is that structural errors are also most probably autocorrelated. Indeed, if for some reason the flow is underestimated at time t, it seems more probable that the flow at time $t + 1$ will also be underestimated rather than overestimated. This hypothesis is based on the fact that hydrological models have a "memory" ; the current state of the model depends on its last. Indeed, conceptual hydrological models contain internal variables, generally water store levels, that vary relatively slowly and buffer the effect of rainfall over time. If these conceptual internal stores are not "correct", the computed flow will contain slowly varying errors. In other words, model errors will be autocorrelated.

A popular technique to take structural errors into account in modeling is by running multiple models simultaneously and comparing their output. This idea is applied in ensemble forecasts and Bayesian model averaging (*Duan et al.*, 2007; *Tebaldi et al.*, 2005; *Hoeting et al.*, 1999). However, there is a potential pitfall here in using the term structural uncertainty to mean different things. Structural errors can be related to the intrinsic stochasticity of the process, deficiencies in the model or faulty boundary conditions. Running simulations with different initial conditions may represent faithfully errors related to boundary conditions (if chosen correctly)

and the model stochasticity, but it cannot account for model deficiencies (*Sivillo et al.*, 1997). Using genuinely independent model may give an idea of the errors related to modeling assumptions, but there is no guaranty that the set of models provides a complete or even realistic description of model errors.

3.2 Impact of errors on calibration

Although in hydrology input, output and structural errors are all present, traditionally, calibration methods have been designed to deal only with output errors. The hypothesis underlying this practice is that other sources of errors can be safely lumped into output errors. There is, however, growing evidence indicating that such an hypothesis does not hold for input errors. That is, that assuming that input data is exact when it is not does indeed influences the quality of calibration and subsequent predictions. More precisely, the estimated parameters are biased. In fact, for the case of a much simpler model, a straight line, this conclusion[2] has been known for a long time, at least since before the beginning of the last century (see *York* (1966) for references). Although methods exist to take input uncertainties into account(*Zellner*, 1971), they still go largely ignored.

The situation in hydrology is more complex. There are a number of studies about the sensitivity of hydrological models to input errors (rainfall and/or PE) (*Oudin et al.*, 2006, 2004; *Andréassian et al.*, 2004; *Kavetski et al.*, 2003; *Borga*, 2002; *Andréassian et al.*, 2001; *Høybye and Rosbjerg*, 1999; *Nandakumar and Mein*, 1997; *Paturel et al.*, 1995; *Michaud and Sorooshian*, 1994; *Xu and Vandewiele*, 1994; *Storm et al.*, 1988; *Troutman*, 1982). Although most of them indicate that models

[2]In the presence of input errors, the least-squares estimate of the slope of a straight line is systematically underestimated.

are sensitive to input errors, other show that the low-pass behavior provides a certain robustness to hydrological models(*Oudin et al.*, 2006, 2004). *Oudin et al.* (2006) attribute part of these conflicting results to the manner in which sensitivity is defined: in some cases, a reference series is computed for one set of parameters, then compared to output series generated with the same parameters but with corrupted inputs, in other cases, the model is re-calibrated with the corrupted data, allowing the model parameters to compensate for input errors.

The effect of errors on the parameters, whether they are input or output errors, is to increase the uncertainty on the parameters. This is also referred to as a decrease in parameter identifiability. From a calibration point of view, a large parameter uncertainty means that different parameter sets yield simulations equally compatible with observations. This concept has been given the name of equifinality by *Beven and Binley* (1992). When the parameter uncertainty is translated into prediction uncertainty, the confidence intervals over predictions may become so large as to make them uninformative (*Uhlenbrook et al.*, 1999; *Nandakumar and Mein*, 1997).

3.3 Calibration methods

In the next pages, a subset of known calibration methods are presented, along with references to application in hydrological sciences.

3.3.1 Standard optimization

The traditional way to calibrate hydrological models, and models in general, is to select an objective function describing the concordance between measurements

and observations and estimate the model parameters maximizing this concordance (*Beven*, 2006). Under this approach lies the hypothesis that all sources of uncertainties can be lumped into output errors, whose magnitude parameters seek to minimize. Understandably, the value of the calibrated parameters will depend on the choice of the objective function, that is the shape of the output error model. The literature contains various suggestions for the objective function (*Freer et al.*, 1996; *Gupta et al.*, 1998). One of the pervasive, however, is probably the normal distribution. The modeler has then only to specify the mean (usually zero) and variance of the distribution and estimate the parameters maximizing the residuals probability, which is equivalent to minimizing the sum of the squared residuals [3]. Although this may seem like a crude way to describe errors, there are a number of theoretical justifications for this assumption[4].

In hydrology, there is little evidence suggesting that residuals are normally distributed. Experience rather shows residuals are heteroscedastic(*Freer et al.*, 1996; *Sorooshian and Dracup*, 1980), that is, the magnitude of errors is not constant: large flows have larger error than small flows. Instead of modifying the objective function to take into this heteroscedasticity into account, modelers often apply a transformation to the flows before computing the objective function on the transformed flows (*Xu and Vandewiele*, 1994). Logarithmic or power transformations (often a square root) are examples of such transformations. Depending on the model's time resolution, residuals may also be found to be correlated. Objective functions based

[3]The Nash-Sutcliffe criteria, which is just the coefficient of determination F^2, is a variation on the same theme where the sum of squares is weighted by the variance of the observed flow series.

[4]Indeed, the central limit theorem states that the distribution of a sum of random variates, no matter their respective distribution, tends to a normal distribution. Thus, by lumping additive errors from different sources into a single error term, the resultant error can be assumed to be approximately normal. Another nice property of the normal distribution is that it maximizes entropy given a fixed mean and variance (*Jaynes and Bretthorst*, 2003). Entropy being a measure of the information content of a function, the normal distribution can be interpreted as the least informative distribution given a mean and variance.

on an hypothesis of autoregressive process can then be used to take this correlation into account (*Beven and Freer*, 2001; *Sorooshian and Dracup*, 1980).

A second hypothesis made by traditional calibration methods is that the model behavior can be accurately described by a single parameter set. This latter hypothesis is disputed among others by *Beven* (2006) on the basis that hydrological models are often over-parametrized with respect to the information content of observational data, leading to multiple instances of parameters giving similar simulations results (equifinality). This view is compatible with the Bayesian philosophy where the quantity of interest is not an optimum parameter but a distribution describing the probability of model parameters. In the following, five methods are presented that share this view, namely that a realistic model calibration should not be based on a single simulation run with a single parameter set.

3.3.2 GLUE

The Generalised Likelihood Uncertainty Estimation (GLUE) procedure (*Beven and Binley*, 1992) is a calibration method built on the equifinality principle: the idea that many different parameter sets yield equally sensible model outputs. Drawing from sensitivity analysis (*Hornberger and Spear*, 1981), it divides parameters between behavioral parameter sets and non behavioral parameters based on a objective function describing the data likelihood. Simply put, the GLUE methodology consists of three steps (*Beven and Freer*, 2001):

1. Select prior distributions for each model parameter, possibly including boundary conditions. Although any type of distribution could be selected, most applications use a uniform distribution on the range of physically realistic parameters (*Beven and Freer*, 2001; *Mantovan and Todini*, 2006).

2. Select a sampling method to draw the random parameters, typically Monte Carlo sampling (*Beven and Freer*, 2001).

3. Select a likelihood measure and an acceptance/rejection criteria. The likelihood measure can be simply the squared error, the coefficient of determination R (the Nash-Sutcliffe efficiency) or any other measure of agreement between simulations and observations. The acceptance criteria divides parameter into behavioral and non-behavioral sets. Parameter sets with a likelihood greater than the criteria are deemed behavioral while the others are discarded as non-behavioral.

A GLUE simulation will hence sample parameters, compute their likelihood and class them as behavioral or not. The set of behavioral parameters can then be used to make predictions about future behavior. The GLUE methodology is widely used in hydrology [5] but is also the target of a number of criticisms regarding non-formal likelihood measures, incoherent posterior predictive probabilities and inconsistencies between batch and recursive estimation (*Mantovan and Todini*, 2006).

In *Beven* (2006), the author suggests that input errors can be taken into account in a formal GLUE framework by increasing the variance of the likelihood measure, thus giving more weight to parameters predicting discharges farther from the measured discharge. While this widens the prediction bounds, it is equivalent to saying that input errors can somehow be lumped into output errors. Although this hypothesis is certainly justified for linear models, it is questionable for the typically non-linear hydrological models. There is no doubt, however, that the GLUE framework can accomodate input errors explicitly by treating them as stochastic parameters.

[5] See `www.es.lancs.ac.uk/hfdg/freeware/hfdg_freeware_gluepapers.htm` for a list of GLUE applications.

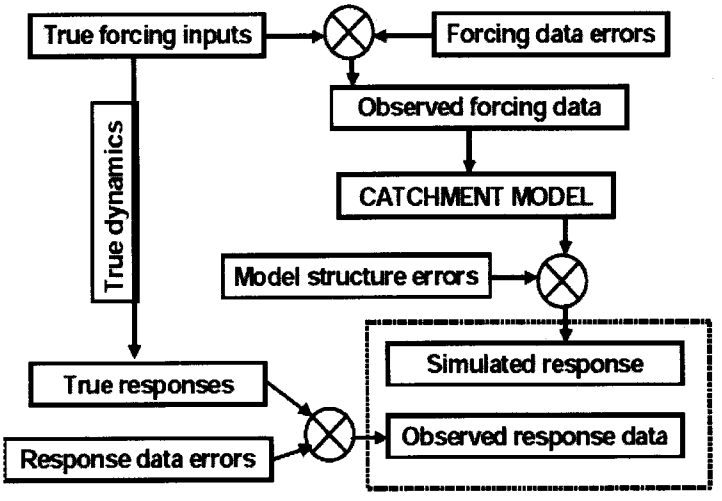

Figure 3.5: Conceptual uncertainty framework (from *Kavetski et al.* (2003)).

3.3.3 BATEA

The Bayesian framework proposed in the thesis is similar to BATEA, *Bayesian Total Error Analysis* (*Kavetski et al.*, 2003), whose name stems from Total Least Squares, a fitting method similar to SLS, but where both the output and input errors are considered (see the conceptual framework in figure 3.5). In BATEA, instead of assuming as in SLS that the input data is exact, the authors propose to estimate the true input values $\boldsymbol{x}$ along with the model's parameters $\boldsymbol{\theta}$, using the expression :

$$\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \mathrm{p}(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{x})\, \mathrm{p}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}). \tag{3.2}$$

Before any computation can take place, the distributions on the right hand side must be specified: the likelihood (or the response error model) $\mathrm{p}(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{x})$, the input error model $\mathrm{p}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})$ describing the relation between true input $\boldsymbol{x}$ and measured input $\tilde{\boldsymbol{x}}$, and the prior $\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x})$ for the model's parameters and the true input values (which could also safely be written as $\mathrm{p}(\boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{x})$.) For one of their applications,

the authors use a homoscedastic normal distribution for their response error model, and an input error model based on storm events. That is, they suppose that the hourly rainfalls during a storm are all equally corrupted. In other words, the storm depth is under or over estimated, but the intra storm variability is not affected by errors. Using these errors models, they generate synthetic data, add a random noise generated by the errors models, compute the posterior distribution using Markov Chain Monte Carlo sampling and compare the parameter distributions calibrated from the sample to the parameters true value. They show that when the error models correspond to the error generating distributions, the method finds the correct parameters, but when the input error models are vague or completely wrong, the estimated parameters are as bad as those found by SLS.

The initial goal of this thesis was to improve on the work of *Kavetski et al.* (2003) by integrating over the latent variables $\boldsymbol{x}$ instead of estimating them. It turns out, however, that both approaches, although theoretically different, are equivalent in practice. Indeed, by sampling over the joint posterior (3.2) and then computing histograms from the sampled parameters, the authors implicitly marginalize $\boldsymbol{x}$ from the posterior distribution, *i.e.* compute $\mathrm{p}(\boldsymbol{\theta}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \int \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})\, \mathrm{d}\boldsymbol{x}$. In any case, the intent of this work is to understand some of the issues related to multiple sources of uncertainties that have not been discussed by *Kavetski et al.* (2003): the effect of the true input prior, accounting for structural errors, dealing with initial state uncertainty, assessing the efficiency of models in the input uncertainty context and the relation between input and output error models.

3.3.4 Kalman filters

The Kalman filter (KF), named after Rudolf E. Kalman who proposed it at the beginning of the sixties (*Kalman*, 1960), recursively estimates the state of a dynamic system from noisy initial conditions and measurements. It is used in a wide variety of applications, from radar tracking to satellite navigation systems. The first implementation of the Kalman filter was in fact to estimate spacecraft trajectories for the Apollo program.

The Kalman filter is basically an equation to update the state of the system as new information (measurements) become available. In the case of a radar tracker, the state of the system would simply be the position and speed of an airplane. For an hydrological model, it could be the level of reservoirs. One constraint however of the original KF is that the state evolution of the system can be written as a linear equation. That is, if we label the state variables at time t by the vector ϕ_t, then the evolution of the system and its output should be written as

$$\phi_{t+1} = \mathsf{M}\phi_t + \mathsf{B}\tilde{\boldsymbol{x}}_t + \varepsilon_t, \quad \varepsilon \sim N(0, \Sigma_t^{str}) \tag{3.3}$$

$$y_t = \mathsf{H}\phi_t + \epsilon_t, \quad \epsilon \sim N(0, \sigma_t^{out}), \tag{3.4}$$

where M represents the state evolution, $\tilde{\boldsymbol{x}}$ the forcing variable with B describing its effect on the new state, ε the random structural error of covariance Σ_t^{str} , H the measurement matrix and ϵ is an output measurement error of standard variation σ_T^{out}.

What makes the KF especially useful is that it keeps track of the state uncertainty. Instead of assuming the initial state is known, it supposes that there is a distribution of initial states. Then, instead of updating each and every one of

those states individually, it assumes the states are normally distributed and only updates the mean and covariance describing the states distribution. It is a clever way to update the description of the states instead of the states themselves. More precisely, the KF first performs a forecast, predicting from ϕ_{t-1} an approximate value for ϕ_t and then compares the simulated output $\mathsf{H}\phi_t$ with the observation to correct (update) the state. The algorithm is described more precisely in table 3.1.

1. Forecasting	**2. Updating**
1.1) Forecast the mean system state: $$\phi_t^f = \mathsf{M}\phi_{t-1} + \mathsf{B}\tilde{\boldsymbol{x}}_{t-1}$$	2.1) Compute the Kalman gain: $$\mathsf{K}_t = \mathsf{P}_t^f \mathsf{H}^T \left[\mathsf{H}\mathsf{P}_t^f \mathsf{H}^T + \Sigma^{out}\right]^{-1}$$
2.2) Forecast the covariance: $$\mathsf{P}_t^f = \mathsf{M}\mathsf{P}_{t-1}\mathsf{M}^T + \Sigma_{str}$$	2.2) Correct the mean state estimate using measurement $\tilde{y}_t$: $$\phi_t = \phi_t^f + \mathsf{K}_t(\tilde{y}_t - \mathsf{H}\phi_t^f)$$
	3.3) Correct the covariance matrix: $$\mathsf{P}_t = (\mathsf{I} - \mathsf{K}_t\mathsf{H})\mathsf{K}_t^f$$

Table 3.1: Kalman filtering algorithm (from *Welch and Bishop* (2006)). Variables are defined in the text.

For complex large scale models (atmospheric, oceanic or distributed hydrologic models), the KF is very valuable as each simulation can take weeks to complete. On the other hand, the KF equations shown above assume that the model M and measurement operator H are matrices, meaning that the model and measurement operators are linear functions of the state variables, a rare occurrence in environmental sciences. To extend KF methods to nonlinear models written as

$$\phi_{t+1} = \eta(\phi_t, \tilde{\boldsymbol{x}}_t, \theta) + \varepsilon_t, \quad \varepsilon \sim N(0, \sigma_t^{str}) \tag{3.5}$$

$$y_t = h(\phi_t) + \epsilon_t, \quad \epsilon \sim N(0, \sigma_t^{out}), \tag{3.6}$$

the Extended Kalman Filter (EKF) computes the Kalman gain matrix and the covariance matrix using linear approximations for M and H. That is, estimates for M_t and H_t are found by computing the Jacobian matrix of η and h at their current value. The setback is that after being forecasted by a nonlinear model, the model state distribution loses its normality; it cannot be fully identified only from its mean and covariance. For highly nonlinear models, this neglect of higher order moments may lead to forecast errors (*Evensen*, 1994).

The Ensemble Kalman Filter (EnKF) solution suggested by *Evensen* (1994) (see *Evensen* (2003) for a complete introduction to the theory and practical implementation details) is to use an ensemble of states to represent the state distribution. That is, a number of initial conditions are defined and each one of them is forecasted individually. To update the states, the covariance matrix is computed from the ensemble statistics and used to update each state. This avoids computing the jacobian of the model at each time step and is numerically cheaper. Using this ensemble approach avoids making the assumption made by the EKF that the state distribution is normal (although P^f is still computed from a normal covariance approximation, see table 3.1), seems to simplify the overall implementation and allows EnKF to be used with nonlinear models without numerical instabilities. This is critical in hydrology since models are often highly nonlinear due to threshold structures, *i.e.* the behavior changes radically if one variable reaches some limiting value (reservoir gets full, rain rate exceeds infiltration rate, etc.)

The idea that EnKF can also be used to estimate parameters, instead of assuming them known, has been used recently by *Evensen* (2007), *Vrugt et al.* (2005) and *Young* (2002). *Evensen* (2007) derives a dynamic equation describing the evolution of the model parameters probability and solves it using EnKF. In the simultaneous optimization and data assimilation (SODA) approach, *Vrugt et al.* (2005) combine

an EnKF loop with a parameter estimation loop. For a given parameter set, the authors generate an output time series using EnKF, then compute the posterior probability of the series and use this probability to explore the parameter space in the search of a global optimum. This search algorithm is driven by the SCEM-UA method (*Vrugt et al.*, 2003) to rapidly converge toward the posterior distribution.

3.3.5 Sequential Monte Carlo

Sequential Monte Carlo methods (also called particle filters, bootstrap filter, condensation algorithm, interacting particle approximation and survival of the fittest) are methods designed to solve the Bayesian recursive estimation problem (*Doucet et al.*, 2000). That is, those methods approximate a posterior density π_t as it evolves each time new data become available. To do so, the posterior is approximated by a finite sample of particles, whose empirical distribution approaches the theoretical distribution as the number of particles approach infinity. Because particle filters are recursive method, they lend themselves to the real time calibration of models and are use extensively in target tracking problems[6].

In the particle terminology, particles $\theta^1, \theta^2, \ldots, \theta^N$ *target* a distribution π if and only if

$$\sum_i^N w^i g(\theta^i) \rightarrow \mathbb{E}_\pi(g), \tag{3.7}$$

where w^i are the particles weight and g is any measurable function such that its expectation value $\mathbb{E}_\pi(g)$ exists. In a recursive estimation problem, the posterior π_t, the target, evolves as new information is added. A particle filter algorithm is hence required to produce a particle system (particles and their weight) that consistently

[6]See the Sequential Monte Carlo Methods homepage at `www-sigproc.eng.cam.ac.uk/smc/` for a list of applications in different fields.

reproduces the target as it evolves. A general algorithm to generate such a particle system can be divided in three steps, mutation, correction and selection (*Chopin*, 2004):

1. **Mutation**: Draw particles for $i = 1, \ldots, N$ from a distribution k_t serving as a first approximation of the target π_t:

$$\theta_t^i \sim k_t(\theta_t \mid \hat{\theta}_{t-1})$$

   The samples drawn from distribution k_t are provisionary in the sense that they do not target π_t.

2. **Correction**: Assign a weight to each particle

$$w_t^i \propto \pi_t(\theta_t^i)/\tilde{\pi}(\theta_t^j),$$

   where $\tilde{\pi}(\theta) = \int \pi_{t-1}(\theta_{t-1}) k_t(\theta \mid \theta_{t-1}) \, \mathrm{d}\theta_{t-1}$. Once this step is accomplished, the particle system targets distribution π_t.

3. **Selection**: Resample the particle system so that the re-sampled particles have uniform weights:

$$(\theta_t^i, w_t^i) \rightarrow (\hat{\theta}_t^i, 1).$$

To give readers a more hands-on example, the following presents a simple application of particle filtering using an implementation called the sequential importance sampling algorithm (SIS) (see *Arulampalam et al.* (2002) for a tutorial), where there are only mutation and correction steps. First, let's express the evolution of a system's state by :

$$\phi_t = \eta(\phi_{t-1}, \tilde{\boldsymbol{x}}_t, \theta) + \omega_t, \tag{3.8}$$

where ϕ_t is the model's state at time t, θ are model's parameters, $\tilde{\boldsymbol{x}}_t$ the forcing variable at time t and ω a stochastic term describing a structural error. The output y_t from the model is described by a second equation:

$$y_t = h(\phi_t, \theta) + \nu_t, \tag{3.9}$$

where ν is the observational error. For the first iteration ($t = 0$), N_p particles (ϕ_0^i, θ_0^i), $i = 1, \ldots, N_p$ are sampled from a prior distribution and assigned uniform weights ($w_0^i = 1/N_p$). Then, the particles are propagated forward in time using the model $\eta(\cdot)$ and the current input variables (the mutation step)

$$\phi_t^i = \eta(\phi_{t-1}^i, \tilde{\boldsymbol{x}}_t, \theta_{t-1}^i) + \omega_{t+1}.$$

The particle's weight w are then updated by evaluating the likelihood of the output measurement $\tilde{y}_t$:

$$w_t^i \propto w_{t-1}^i \cdot p(\tilde{y}_t | h(\phi_t^i, \theta_t^i)),$$

where p is a function returning the likelihood of the observation.

The particle system steps forward and propagate the particles to the next time step, targeting each posterior distribution as it evolves. Using the particle system, it is straightforward to compute any expectation value of a function g using equation (3.7).

Although the basic idea is straightforward, there are some difficulties with its application. One such difficulty is that particles tend to degenerate: after a few time steps, most particles end up with very low weights, contributing weakly to the computation of expectation values, hence lowering the descriptive value of the system. One solution to this problem is to add a re-sampling step once in a while,

(*eg.* the SIR algorithm, (*Gordon et al.*, 1993)); particles with very low weights are discarded, and replaced by copies of particles with high weights. But then, this introduces another problem called sampled impoverishment, where a lot of particles become identical. The solution then is to perturb the copied particle's by adding a random noise using MCMC methods (*Doucet et al.*, 2000).

Particle filters methods have been applied to dynamic systems (*Liu and Chen*, 1998) and more precisely to calibrate an hydrological model by *Moradkhani et al.* (2005) in real and synthetic experiment. The authors show that in a synthetic experiment, the parameters converge toward their true value. For real experiments, the authors report good conformity between simulated and observed results, although some of the high floods lie outside the prediction uncertainty bounds. Particle filters have been compared to EnKF in an hydrological application in *Weerts and El Serafy* (2006). For a low number of ensembles/particles, the EnKF yields superior results in term of RMSE (in a synthetic experiment), while for intermediate number of ensembles/particles, particle filters perform better. However, the authors remark that EnKF is more robust to the choice of error models, and hence, more appropriate for real experiments. Particle filters have also been used recently for dynamic crop modelling (*Naud et al.*, 2007).

3.3.6 Gaussian processes

A Gaussian process is defined as a stochastic process generating samples such that any linear combination of those sample is normally distributed. For example, Brownian motion is a Wiener process, a particular instance of a Gaussian process, since the position of a particle subject to Brownian motion is normally distributed. *Kennedy and O'Hagan* (2001) use Gaussian processes to infer a function, the true process,

from data, a computer code (the model) and various assumptions about structural and data errors. That is, they assume there exists a true process $\zeta(\cdot)$, from which the data is a particular realization. The computer model $\eta(\boldsymbol{x}, \boldsymbol{\theta})$ is a simplification of the true process, neglecting the inherent stochasticity of the process as well as unobserved conditions (hidden variables). The relationship between observations y_i, the true process and the computer model is then assumed to be

$$y_i = \zeta(\boldsymbol{x_i}) + e_i = \rho\,\eta(\boldsymbol{x_i}, \boldsymbol{\theta}) + \delta(\boldsymbol{x_i}) + e_i, \tag{3.10}$$

where $\boldsymbol{x_i}$ is the input vector at time t_i, e_i is the measurement error, ρ a unknown parameter and $\delta(\cdot)$ a function describing structural errors, labeled by *Kennedy and O'Hagan* (2001) as the *inadequacy function* and assumed to be independent from the code output $\eta(\boldsymbol{x}, \boldsymbol{\theta})$.

Under the assumption that $\eta(\cdot)$ is a Gaussian process, it can be written as

$$\eta(\cdot) \sim \mathcal{N}(m(\cdot), c(\cdot, \cdot)),$$

where $m(\cdot)$ is the mean function and $c(\cdot, \cdot)$ the covariance function. The mean function m is then represented by a linear combination of different functions with unknown coefficients. For instance, if polynomials are used as a basis for m:

$$m(\cdot) = \sum_{i=1}^{p} \beta_i\, x^i = \boldsymbol{h}(\cdot)^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ are unknown coefficients. The covariance function is defined as a correlation function describing the smoothness of the function: $c(\cdot, \cdot) = \sigma^2 r(\boldsymbol{x} - \boldsymbol{x'})$. A choice for r could be the euclidean distance for example.

If the inadequacy function δ is also assumed to be a Gaussian process, and

if the residuals are assumed normally distributed, then it is possible to calibrate the model, that is to find the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\phi} \mid \boldsymbol{d})$ where $\boldsymbol{\phi}$ are hyperparameters describing prior assumptions and $\boldsymbol{d}$ the calibration data set using linear algebra. In the applications presented in *Kennedy and O'Hagan* (2001), the hyperparameters are estimated to a point value, and predictions are made by numerically integrating over the model parameters $\boldsymbol{\theta}$. The strong point of Gaussian processes is the ability it gives the modeler to infer the value of a function, the model, instead of being constrained to a deterministic set of equations. In terms of capturing the structural uncertainty, this is a remarkable improvement. One setback however is that the procedure described in *Kennedy and O'Hagan* (2001) is limited to problems with low dimensionality, although the authors imply that it would be possible to tackle higher dimensional problems using MCMC integration. Another issue is the implied normality assumptions for the model residuals, which may not be realistic in hydrology unless variables are appropriately transformed.

3.3.7 Comments

For all the methods presented above, assumptions about the nature of errors are made to simplify computations to a level where the methods can be applied to complex hydrological models. For instance, the particle filters does not account explicitly for input errors, *i.e.* , it assumes that input uncertainties can be described by parameter uncertainties. Kalman filters on the other hand can be designed to handle input uncertainties, but the state updating equation is based on the state covariance, imposing normal-like correlations among variables. Gaussian processes are an elegant proposition to describe different sources of uncertainty, but according to the authors, is not adapted for problems with many dimensions, limiting its use-

fulness in hydrological modeling. In BATEA, errors are assumed perfectly correlated among a given storm to reduce the number of latent variables.

In the following, a general Bayesian framework is proposed, theoretically imposing no constraints on the choice of errors models. In an actual calibration, however, this freedom in the selection of error models has a cost in terms of computation time, constraining the models for which the method can be applied in a reasonable amount of time.

CHAPTER 4

Discussion

This chapter presents an overview of the ideas and results presented in the papers included in Part II. For each paper, a synthesis of the objectives and methodology is provided, followed by comments on the results and their meaning. In the last section, some ideas about further work and applications are proposed.

The basis of this work is a conceptual framework describing how multiple errors relate to idealized true input and output variables, as well as to the true process. The framework is entirely general in the sense that it could be applied to any model. It is also not unique since the very definitions of structural, input and output errors are, to some extent, a matter of subjective choice.

Using Bayesian analysis, this uncertainty framework can be translated into a probabilistic form, allowing the derivation of an equation for the posterior parameter distribution. Under rather stringent assumption, this equation can be solved analytically. This special case is the topic of *Huard et al.* (2006) and discussed in section 4.3. The assumptions required for the existence of an analytical solution are the linearity of the model and the normality of the error models and the prior for the true input. These cases occur rarely, if ever, in practice, and the application to the linear model "abc" (*Fiering*, 1967) is a didactic exercise rather than an actual calibration

technique. However, the analyticity of the solution allows a deeper understanding of issues related to input errors, difficult to achieve with more complex models.

To dispense with the linearity and normality assumptions, the second paper uses Markov Chain Monte Carlo (MCMC) sampling to solve the parameter posterior equation (section 4.4). While this allows any model or error model to be chosen, it becomes numerically intensive as the data series become longer. For this reason, the method is applied to a monthly model, GR2M (*Mouelhi et al.*, 2006), rather than a daily model. Simulations are run with simple and more complex error models to highlight their impact on the parameter posterior distribution. Other issues are discussed, namely the initial state uncertainty and the identification of structural errors.

In these two first papers, even though structural errors are considered (implicitly or explicitly), the model is assumed known. In real applications, however, the first modeling step is often the selection of an appropriate model. Model selection is generally a subjective decision based on the experience of the modeler or simply trial and error. Recently, Bayesian model selection theory has been used to identify the most appropriate model or the weights for an ensemble of models (*Marshall et al.*, 2005). Such comparison of models, however, depends on the choice of priors for model parameters. To define priors having a consistent meaning across models, the third paper (*Huard et al.*, 2006) proposes the idea of *parent prior*, and applies it to copula selection. Copulas are multivariate distributions that enjoy a wide popularity in finance due to their flexibility in describing a wide range of dependence structure between variables, irrespective of their marginal distribution (*Genest and Favre*, 2007). Section 4.5 discusses how a parent prior idea might be used in hydrological model selection.

4.1 Conceptual framework

An inferential process basically strives to estimate what is unknown from what is known. In an hydrological calibration problem, the known elements are typically experimentally observed data series of rainfall, flow and other climatic variables requested and returned by the model. The unknowns are the parameters describing the model's behavior. Calibration hence amounts to the computation of the probability of those parameters knowing the input/output data and the model: $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \mathcal{M})$. The main problem with such a formulation is that models do not link directly input data to output data. Indeed, models assume they are forced with exact true inputs ($\boldsymbol{x}$), not experimental data ridden by input errors. Moreover, models are not exact, but merely approximations of the natural processes that occur in the real world. Hence, output of models are different from idealized true output variables, and this difference is defined here as the structural error. Finally, this idealized true output variable ($\boldsymbol{y}$) is not what is observed in the field since output errors corrupt the measurement ($\tilde{\boldsymbol{y}}$) of this true output variable. Overall, there is a gap between what is actually known ($\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$), what the model expects ($\boldsymbol{x}$), what it returns ($\widehat{\boldsymbol{y}}$) and what the "real" output variable is ($\boldsymbol{y}$).

To bridge this gap, relationships are needed between measured and true values, and between modeled and true processes. These relationships are called error models, and describe, probabilistically, the probability of occurrence of input, output and structural errors. Once those error models are defined, the next step is to merge them into a unified equation describing the posterior parameter probability.

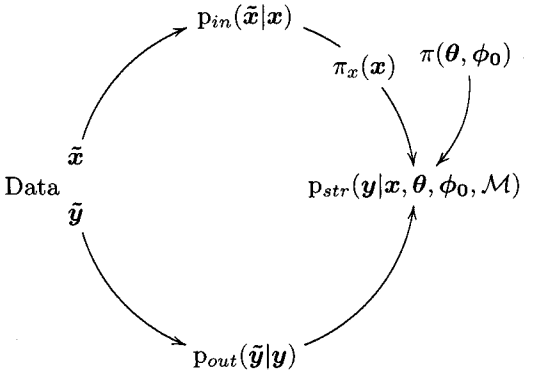

Figure 4.1: Uncertainty framework proposed and used in the thesis. The input error model p_{in} describes the probability of measuring $\tilde{x}$ knowing the true value of $\boldsymbol{x}$. Similarly, the output error model p_{out} describes the probability of measuring $\tilde{y}$ knowing the true value $\boldsymbol{y}$. The last element, p_{str} is the structural error model, and describes the probability of a true value given the model $\mathcal{M}$, the true forcing $\boldsymbol{x}$, the parameters $\boldsymbol{\theta}$ and initial conditions $\boldsymbol{\phi_0}$. The distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\phi_0})$ describes the *a priori* knowledge about parameters and initial conditions.

4.2 An equation for the posterior distribution

The probabilistic statement $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \mathcal{M})$ is not directly solvable. To bring it under a computable form, a number of steps have to be performed and assumptions made. The complete derivation is described in detail in section 2.2 of *Huard and Mailhot* (2007). Here, only the main assumptions are highlighted:

- Measurements are estimates of idealized true input and output variables.

Table 4.1: Error models linking known data and model to idealized true values and processes.

Input error model	$p_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})$
Output error model	$p_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})$
Structural error model	$p_{str}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi_0}, \mathcal{M})$

- The model is an approximation of some true process.
- Input and output errors are conditionally independent from each other given the true input and output, as well as conditionally independent from model parameters, initial conditions and the model.

Using those assumptions, the posterior parameter distribution can be written as

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \iiint \mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, \mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})\, \mathrm{p}_{str}(\boldsymbol{y} \mid \widehat{\boldsymbol{y}})\, \pi_x(\boldsymbol{x})\, \pi(\boldsymbol{\theta}, \boldsymbol{\phi_0})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\phi_0}, \tag{4.1}$$

where mention of the model $\mathcal{M}$ is omitted since it is always assumed known. There is also an implicit assumption, namely that all error models and priors are conditioned on I, the information on the watersheds characteristics (location, size and climatology). This assumption makes the separation of the prior $\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi_0})$ into two distinct priors $\pi_x(\boldsymbol{x})\, \pi(\boldsymbol{\theta}, \boldsymbol{\phi_0})$ a reasonable simplification. Indeed, while there are certainly correlations between rainfall and model parameters and model states, they are generally unknown a priori. In other words, it would be extremely difficult to describe an informative dependence structure between model parameters and rainfall before looking at the data. An intermediate view is to assign prior for the model parameters and state on the general climatology or previous studies on neighbouring bassins, information denoted here by I.

4.3 Linear case: "abc" model

The solution to equation (4.1) involves integrals over $\mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\phi_0}$. For a data series of length n, this implies an integration over more than $2n$ dimensions. Even for short time series, using standard integration algorithms (quadrature methods) is out of

the question, as their computation time grows almost exponentially with dimension. These algorithms are fit for 1D to 3D integrals, but useless for integrals over more than six dimensions, except in very specific cases (*Genz*, 1992). Fortunately, equation (4.1) has an analytical solution when certain conditions are met: linearity of the model and normality of the error models and of the true input prior. The first application of the method is hence concerned with the linear model "abc", and uses this analytical solution to compute the posterior distribution and evaluate the advantages of the method compared to traditional calibration methods, as well the implications of assuming the presence of input errors.

4.3.1 Objectives and methodology

The objective of the study is to see whether or not the bias caused by input errors in the estimated parameters fitted using SLS can be eliminated by taking input errors into account.

To do so, synthetic data is generated: a series of true rainfall is defined as the true rainfall and random *measurement* errors are added to create a synthetic rainfall measurement series. Output measurements are created similarly by choosing a set of model parameters as the true parameters, running the model with the true inputs so as to define a series of true outputs. These true outputs are then corrupted with normally distributed errors to generate a synthetic output measurement series. The model is calibrated against the synthetic data and the estimated parameters are compared with the true parameters as well as with the parameters estimated by SLS.

The model used is "abc", a pedagogical model designed to teach students the basic principle of hydrologic modeling (*Fiering*, 1967). It has three parameters

(a, b, c), describing the infiltration rate, evapotranspiration and the exchange between storage and discharge. The model is forced with aerial rainfall, and returns an aerial discharge. There is one state variable, the amount of water in storage. By stating that the model is linear, it is meant that it can be written under the form $\boldsymbol{y} = \mathsf{A}\boldsymbol{x} + \boldsymbol{b}$ (see *Huard and Mailhot* (2006, par. 50)). Under this form, and assuming the prior for the true input $\pi(\boldsymbol{x})$ and the error models are expressed by Gaussian functions ($\propto e^{\boldsymbol{z}^t \Sigma^{-1} \boldsymbol{z}}$), the solution to 4.1 is given by Equation (18) in *Huard and Mailhot* (2006).

4.3.2 Discussion

The results seem to indicate that using the method slightly improves the accuracy of the calibrated parameters. That is, the estimated parameters are closer to the true parameters than those estimated with SLS. As for predictions, they seem again to be slightly better than those obtained using SLS, but this conclusion is based only on visual cues. Due to numerical limitations, the simulations could not show that the parameter distribution converges toward the true parameters as the length of the calibration time series increases. However, the fact that this convergence actually takes place for a simpler model (a straight line fit) lends confidence in the hypothesis that the Bayesian method can actually overcome the bias in the "abc" parameters. Overall, the paper concludes that by using the proposed method, the calibrated parameter distribution is allowed to incorporate input uncertainties in a meaningful way, and provides confidence intervals that realistically depict the different uncertainties.

In retrospect, this conclusion appears a bit simplistic. First of all, saying that the parameter uncertainty incorporates the input uncertainty may be too vague. What

is really meant is that the parameter distribution considers the possibility of a true input being different from the input measurement. It does not mean, however, that running the model with parameters sampled from the posterior distribution reproduces the effect of input errors. In other words, although the parameter posterior distribution is representative of the uncertainties contained in the input and output data, parameter uncertainty cannot magically "account" for the other sources of uncertainties.

Secondly, the usefulness of the whole procedure strongly depends on the accuracy of the error models and of the prior for the true inputs. In this first application, since the errors are synthetic, defining the error models poses no problem, but in real life, it remains a complicated issue. Another important realization stemming from this research is that when each datum in the series is considered uncertain, the prior for the true input becomes critical for the correct evaluation of the parameters. That is, it appears that the parameter posterior distribution does not converge towards the true parameter value unless the prior for the true input is correctly defined.

Finally, a technical issue is the numerical cost to evaluate the posterior. For large time series ($n > 100$), the time needed to multiply and invert $n \times n$ matrices becomes prohibitive. One solution to reduce the numerical effort would be to make use of the properties of matrix A. In theory, the model output depends on all previous rainfall, but for practical purpose, only the last 20 or so values are relevant. This means that A can be approximated by a band matrix, with only the first 20 lower diagonals being non-zero. This simplification would allow the use of numerical routines specifically designed to deal with band matrices, and possibly allow the treatment of larger matrices, hence longer time series.

4.4 Non linear case: GR2M model

Most hydrological models are nonlinear with respect to their input variables. This means that the analytical solution used for "abc" cannot be applied in general, and a solution to equation (4.1) must be found elsewhere. In physics, a popular solution to solve large integrals is Metropolis Markov Chain Monte Carlo (MCMC) sampling. The MCMC theory is discussed in Bayesian textbooks (*Gelman et al.*, 1995) as well as more specific references (*Neal*, 1993). The basic idea is to draw stochastic parameters, and accept or reject them based on their posterior probability. If the sampled chain is long enough, the values drawn form a sample of the posterior distribution. The second paper uses the Metropolis algorithm to generate a Markov chain of samples empirically defining the parameter posterior distribution (4.1).

4.4.1 Objectives

The initial objective of the paper was to use the flexibility of MCMC sampling to assess the influence of different choices of error models on the posterior model parameters and the performance of the calibrated model. A number of difficulties, however, have hampered success and led to a complete rewrite of the paper. The main difficulty is that measuring model performance is not trivial. The standard criteria for hydrological model performance, the Nash-Sutcliffe efficiency, only measures differences between output observations and simulations. When taking input errors into account, input data is not considered exact anymore, therefore, using NSE as a performance criteria is misleading. Moreover, NSE is influenced by the variability found in the time series. In this sense, comparing stations among themselves is rather meaningless. Even comparing calibration and validation is risky since a

lower variability in the validation series can result in higher NSE in validation than calibration (*Schaefli and Gupta*, 2007). Overall, although the first version of the paper was an honnest attempt at understanding the effect of error models, its goal was too ambitious with respect to the tools that are available to achieve it.

The objective of the second version is rather to understand the calibration and validation issues that inevitably appear when multiple sources of uncertainty are included in the analysis. In other words, to present some of the issues that affected the realization of the first version. For instance, the fact that input errors are considered completely changes the way calibration and validation results are to be interpreted, and hence, how model performance is evaluated. The paper also highlights some of the possibilities offered by the approach, namely the *a posteriori* separation of the different sources of errors. On the whole, the exercise is to be interpreted as a way to reveal the issues that need to be addressed before attempting to apply the method in real scenarios, as well as underline the novel possibilities it opens.

4.4.2 Methodology

The solution to the parameter posterior equation is found using the Metropolis algorithm, where samples are generated through a random walk process. The Metropolis algorithm was chosen mostly for its ease of implementation, but other sampling algorithms may be employed in the future. As an example, assume parameter θ is to be sampled from a distribution $p(\theta)$. Starting from some value θ_t, a candidate θ' for the next step is drawn from a proposal distribution $q_t(\theta' \mid \theta_t)$. An acceptance ratio defined as :

$$a(\theta', \theta) = \frac{q_t(\theta' \mid \theta)p(\theta)}{q_t(\theta \mid \theta')p(\theta')}$$

defines the probability of this candidate forming the next state: the candidate θ' is accepted as the the new state θ_{t+1} with probability $\min(a(\theta', \theta), 1)$. Algorithmically, this is translated as:

1. A new candidate is drawn $\theta' \sim q_t(\theta \mid \theta_t)$,
2. A random uniform variate is drawn $u \sim U(0, 1)$,
3. If $a(\theta', \theta_t) > u$ then $\theta_{t+1} = \theta'$, otherwise $\theta_{t+1} = \theta_t$.

Repeating this algorithm for a large number of times ensures that the chain empirically reproduces the distribution $p(\theta)$.

The main obstacle to the MCMC solution is that the number of samples necessary to reach convergence, that is until the sample empirically reproduces the properties of its distribution, increases with the number of dimensions to cover. There are a number of criteria assessing the convergence of the sample, but none is entirely satisfactory by itsefl. Hence, visual inspection is used to determine at which point addind samples has no visible effect on the final distribution. Using daily models and assuming independent input, output and model errors, the calibration of five years of data would imply sampling over more than 3600 dimensions. Although this is certainly possible, it is a big step to take. A more cautious approach is to first tackle simpler monthly models with comparatively shorter time series (calibration over 20 years of data implies sampling over *only* 480 dimensions).

The model chosen for this application is hence GR2M (*Mouelhi et al.*, 2006; *Kabouya*, 1990; *Edijatno and Michel*, 1989), a parsimonious monthly hydrological model. Although GR2M has only two free parameters and two internal state variables, it has been shown to be sensitive to input errors (*Paturel et al.*, 1995).

The meaning of the model parameters is described in section 3.1 of the second paper, and a diagram of the model is shown in figure 4.2. The data used for calibration is taken from the Model Parameter Experiment (MOPEX) database (*Schaake et al.*, 2006b) and consists of aerial daily rainfall, potential evapotranspiration and discharge. The data is then aggregated to monthly values, a simple task since there are many stations with long stretches of data without missing values. Calibration is performed under different error model assumptions to understand the effect of these assumptions on the calibration results.

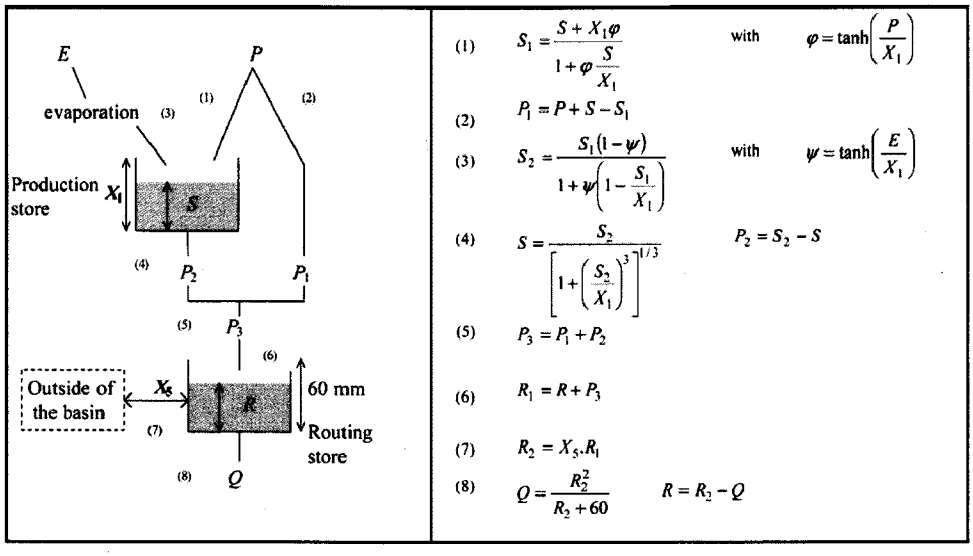

Figure 4.2: Diagram of GR2M (from *Mouelhi et al.* (2006)).

4.4.3 Discussion

Although this second paper deals with a simple monthly model, most of the issues it raises are independent from the modeling time step. One of the most important conclusions of the paper is the realization of the large effect error model assumptions have on model parameters. Assuming that input errors are large and output errors small and vice-versa leads to completely different parameter distributions

(see section 4.5 of *Huard and Mailhot* (2007)). This begs the question "What is the meaning of model parameters ?" For *Mantovan and Todini* (2006) and *Beven* (2006), parameters are dummy quantities, random values with a flimsy physical signification, on which all uncertainties are projected. This view, labeled by these authors as the "classical Bayesian view"[1], appears unproductive and slightly misleading. For *Jaynes and Bretthorst* (2003), the whole idea of Bayesian analysis revolves around the inference of unknown values. Whether these values are random or not is beside the point. Second, assuming parameters are dummy variables is detrimental because it dilutes the physical content of parameters. Ideally, model parameter, even for conceptual models, should have a physical interpretation, making it possible to define prior assumptions about them from the available hydrological, climatological and geographical knowledge. Treating parameters as dummy variables leaves little choice for modelers but to assign very vague priors to these parameters. Finally, stating that uncertainties are projected into the parameter distribution is an incomplete statement. The dispersion of the parameter distribution reflects the uncertainty on the inferred model parameter due to data and model errors. It does not in some magical way lump all sources of uncertainties. More explicitly, if the posterior parameter distribution is used to simulate output series, the dispersion of the outputs only represents the parameter uncertainty, and not the input and output uncertainties. These sources of uncertainties, as well as the model uncertainty, must be accounted for explicitly. If this is not done properly, confidence intervals run the risk of being too optimistic.

An interesting possibility offered by the uncertainty framework and briefly explored in *Huard and Mailhot* (2007) is the explicit separation and identification of

[1]The notion that Bayesian theory considers parameters like random values is rather the frequentist interpretation of Bayesian theory rather than the "classical" Bayesian paradigm as understood by Jeffreys (*Jeffreys*, 1983) or Jaynes (*Jaynes and Bretthorst*, 2003).

input, output and structural errors. Of course, what is obtained is a posterior distribution of these errors, and not definite values. Moreover, it is expected that these error distributions are closely dependent on the error models chosen to calibrate the model. Indeed, contrary to parameters that are inferred from the entire data series, input, output and structural error series are each as long as the data series: the ratio data/parameters is very low. Moreover, structural errors are linked to the data only through the input and output error models. In this sense, before analyzing the structural error distribution to find model inadequacies, the first steps would be to assess the validity of the error model assumptions and the sensitivity of the posterior structural error distribution to these assumptions. If the posterior structural error distribution radically changes for different but equally plausible structural error models, then there is not much to learn from the analysis of the inferred structural errors. If on the contrary structural errors features are relatively robust, analysis of the structural error time series could yield precious information about the contexts in which the model fails.

One thing that deserves to be mentioned is that for models expressed by a state equation and an observation equation:

$$\phi_{t+1} = \eta(\phi_t, \tilde{\boldsymbol{x}}_t, \theta) + \varepsilon_t \tag{4.2}$$

$$y_t = h(\phi_t) + \xi_t, \tag{4.3}$$

a state error ε could also be considered via a state error model. It is clear, however, that without some way to define an informative state error model, or some way to measure, at least indirectly, the state of the system, there is probably not much to gain by adding this additional source of error: it would simply be a fudge factor. This brings back the question about the meaning of model parameters. If the state

variables are "dummy" variables without physical meaning, it is difficult to assign an informative prior for the state and impossible to compare the model state with some physical property of the watershed, which could help calibrating and validating the model. On the other hand, if the state variables can be linked, even tenuously, with some observable, then it becomes possible to incorporate the measurements of this observable to the calibration process and improve the inference of the model's parameters via its state.

4.5 Model selection and the parent prior

In the proposed Bayesian framework, structural errors are considered by defining a structural error model, describing probabilistically the occurrence of structural errors. Another solution to describe structural uncertainties is to run simulations with different models. The underlying idea is that by assessing the output by a variety of models, the uncertainty over the modeling process is captured more effectively than by a unique model (*Duan et al.*, 2007). The question then arises: are all models equivalent, or are some models more probable than others and which ones ? Bayesian model selection can answer that question, but not without a number of assumptions. These assumptions concern both the prior probability for each model, but also the prior probability for the models' parameters. In the third paper, Bayesian model selection (*Bretthorst*, 1996) is used along with the proposed concept of the *parent prior*, allowing the definition of priors for model parameters that are consistent across models, *i.e.* have the same meaning. The paper applies those ideas to the selection of copulas. Explanations as to how a similar method could be applied in hydrology are then given.

4.5.1 Objectives

The objective of the third paper (*Huard et al.*, 2006) is to show how models can be selected based on formal hypotheses. These hypotheses concern 1) the prior probability of the different models, and 2) the probability of a property shared by all models. The proposed method is applied to select the "best" copula among a set to show the effectiveness of the method as well as its limitations.

4.5.2 Methodology

Although there is a rather large body of literature on the subject of prior selection for parameter estimation (*Kass and Wasserman*, 1996), the issue is relatively new in the context of model selection. Indeed, in such applications, each model has different parameters with different meanings. In order to compare the models rigorously, the priors for the parameters should ideally mean the same thing across models. If this is not the case, some models run the risk of being selected or rejected arbitrarily for parametrization reasons. To solve this problem, *Huard et al.* (2006) propose the concept of *parent prior*. A parent prior is a prior distribution for a property (significant in the context of the application) that is shared by all models. By variable substitution, this parent prior can be translated into priors for the parameters of each model.

The idea of the parent prior is elaborated and applied in the context of copula selection (see section 4.5). Copulas are multivariate joint distributions over the unit hypercube $[0,1]^n$ such that the marginal distributions are uniform over $[0,1]$ (*Nelsen*, 1999). Since the cumulative distribution function of any distribution is a uniform distribution over the unit interval, copulas can be used to model the

dependence between variables, irrespective of their marginal distribution. Hence in two dimensions, if $H(x, y)$ is a joint distribution with marginals $F(x)$ and $G(y)$, then there exists a copula C such that

$$H(x, y) = C(F(x), G(y)).$$

This type of construction yields much more flexibility than classical multivariate distributions where the marginal are all from the same family: multivariate normal, bivariate Gumbel, etc.

In the copula literature, the dependence between variables is often expressed by Kendall's τ, a non-parametric measure of association (see section 1 in *Huard et al.* (2006)). For the most popular copulas, this measure has a simple analytical expression in terms of the copula parameter. The widespread use of Kendall's τ and its simple interpretation makes it a logical choice for the property on which to base the parent prior. For the copula application, a uniform prior over Kendall's τ was used as a parent prior.

The third paper shows how this approach yields sensible results for copula selection. Using synthetic data generated from one copula, the selection method is applied to see how often the right copula is selected. The performance of the method depends closely on the length of the series, but also on the relative differences between copulas. If the copulas tested are all very much alike (*eg.* when the dependence is low), then the method finds the right copula less often than when the copulas are more contrasted.

4.5.3 Discussion

Traditionally, hydrological models are chosen either by expert knowledge or by comparing quantitative criterion of model accuracy among a set of plausible models (*Marshall et al.*, 2005). This method has the tendency, however, to bias model selection toward complex models since they typically are able to reach a better agreement with observations, at least during calibration. Some criteria like the Bayesian Information Criterion (BIC) (*Schwarz*, 1978) attempt to compensate this bias by penalizing a model according to its number of parameters, but have other weaknesses (*Weakliem*, 1999).

In *Marshall et al.* (2005), the authors propose to base the selection of an hydrological model on Bayesian model selection. One of the attractive features of Bayesian model selection is that for similar performances, it naturally discriminates simple against complex models. To understand how this Ockam's razor works, assume one wishes to compare two models: model $\mathcal{M}_1$ has no free parameter, and model $\mathcal{M}_2$ has one parameter a taking a value among $[1, 2, 3]$. It is assumed that the prior for a is uniform, that is: $\pi_a(i) = 1/3$ for $i = 1, 2, 3$. The likelihood of the first model is given simply by

$$\mathrm{p}(\mathcal{M}_1 \mid D) = \mathrm{p}(D \mid \mathcal{M}_1)\frac{\pi(\mathcal{M}_1)}{\mathrm{p}(D)}, \tag{4.4}$$

where D stands for data. The likelihood of the second model is obtained by considering the parameter as a latent variable, that is, by summing the likelihood over all the values taken by the parameter:

$$\mathrm{p}(\mathcal{M}_2 \mid D) = \sum_{i=1}^{3} \mathrm{p}(D \mid \mathcal{M}_2, a = i)\pi_a(i)\frac{\pi(\mathcal{M}_2)}{\mathrm{p}(D)} \tag{4.5}$$

$$= \frac{1}{3}\frac{\pi(\mathcal{M}_2)}{\mathrm{p}(D)}\sum_{i=1}^{3} p(D \mid \mathcal{M}_2). \tag{4.6}$$

Clearly, for the second model to reach a higher probability than the first, it has to "beat" the prior factor of 1/3. That it, if the model has a non-zero likelihood only for $a = 3$, then $\mathrm{p}(D \mid \mathcal{M}_2, a = 3)$ has to be three times greater than $\mathrm{p}(D \mid \mathcal{M}_1)$, otherwise the first model will be selected as the most probable. Hence it is not only the number of parameters that influence the probability of the model in a model selection problem, but also their prior distributions. Tightly constrained parameters do not discriminate complex models as much as parameters with vague priors. In this sense, the choice of the priors for the parameters is an important part of the model selection process. Of course, since the influence of priors decreases as data sets increases in length, the concern for suitable priors is mostly relevant for small data sets, or more generally, where the information content of data is relatively low, which is arguably the case where large data uncertainties are present.

In *Marshall et al.* (2005), a Bayesian model selection is performed on a series of models derived from the Australian Water Balance Model (*Boughton*, 2004) with two, three, four and ten water surface stores (the standard model has three). In this case, the choice of prior is relatively simple because there is effectively only one model. In a general case where different models are used, choosing consistent priors across model is not so straightforward and this is where the parent prior comes into play.

Finding a meaningful property common to all hydrological models of a set is not as straightforward as for copulas. For one, there is no unique hydrologically significant property representative of the model behavior. Secondly, such properties are generally not a simple function of the model parameters, but rather features of the hydrograph: time to peak, recession time, peak discharge or more global statistical properties (*Yue et al.*, 2002). Nevertheless, a parent prior could be defined using those hydrograph features, using the following algorithm:

1. Select an hydrograph property ϑ, for example, $\vartheta \equiv$ time to peak, function of the simulated flow $\widehat{\boldsymbol{y}}$, *i.e.* $\vartheta \equiv \vartheta(\widehat{\boldsymbol{y}})$.

2. Define a prior distribution for this property $\pi_\vartheta(\vartheta)$ based on expert knowledge, *eg.* $\pi(\vartheta) = \mathcal{N}(\mu = 10\text{hours}, \sigma = 1\text{hour})$.

3. Select a long input series $\boldsymbol{x}$.

4. For model parameters $\boldsymbol{\theta}^i$ chosen on a grid covering densely the parameters' domain, compute the hydrograph $\widehat{\boldsymbol{y}}^i = \mathcal{M}(\boldsymbol{\theta}^i, \boldsymbol{x})$ and the value of the property $\vartheta^i = \vartheta(\widehat{\boldsymbol{y}}^i)$.

5. Assign to the parameters $\boldsymbol{\theta}^i$ a weight w^i equal to the probability given by prior for the corresponding property $w^i = \pi_\vartheta(\vartheta^i)$.

6. Normalize the weights by integrating over the parameter grid.

7. Approximate $\pi(\boldsymbol{\theta})$ by splines or polynomials fitted to the normalized weights.

Applying this algorithm to each model yields a spline or polynomial representation of the parent prior in the model parameter space. Using theses priors for model selection ensures that model selection is independent from the arbitrariness of model parametrization. Similarly, those same priors could be used in Bayesian model averaging (*Hoeting et al.*, 1999) applications to make sure that the models weights are not biased by inconsistent priors.

4.6 Challenges for the future

This section discusses some of the issues that will need to be solved in order to really take advantage of calibration methods that take multiple sources of uncertainty into account, as well as some ideas to extend the work done during this thesis.

4.6.1 Defining informative error models

The results obtained from the method, and basically any calibration method, are as good as the underlying hypotheses. Consequently, there is no real advantage to the use of a Bayesian analysis combining all kinds of error sources if no care is given to the definition of the error models and priors. There are two faces to the problem of defining error models: 1. Defining realistic error models, describing as well as possible the errors affecting measurements, and 2. Upgrading error models by incorporating other sources of information. In other words, a realistic error model is good, but a precise inference of the true value is better yet.

To design a realistic rainfall error model, one would have to look at the number of stations, their locations, the gauge type used and the environmental conditions in which each gauge operates (*Willems*, 2001). The rigorous translation from point estimates to aerial estimate should be considered attentively, particularly with respect to extreme events. This could be achieved using the special statistical distributions discussed in *Kundu and Siddani* (2007). Moreover, gauge networks change over time so an accurate input error model would be split in different parts, each one dealing with observations taken during a given period.

Similarly, an accurate discharge error model would vary with new stage-discharge measurements and following updates to the rating curve. An elegant way to define the discharge error would be to use Bayesian analysis to calibrate a "rating curve model" based on streamflow principles (*Schmidt*, 2002), instead of empirically selecting the error model as is done in (*Huard and Mailhot*, 2007). These theoretical models are derived from assumptions about the general shape of the river banks and simplified flow equations. The calibration could be made in a recursive fashion, updating the parameter distribution as new stage-discharge measurements become

available. To illustrate this idea, here are the steps leading to such a discharge error model:

1. Define a stage error model $\mathrm{p}_s(\tilde{h} \mid h)$, the probability of committing an error when measuring the stage.

2. Define a discharge error model $\mathrm{p}_d(\tilde{q} \mid q)$ describing individual measurement errors on discharge.

3. Define a structural error model $\mathrm{p}(q \mid h, \Theta, \mathcal{R})$ describing errors due to the simplifying assumptions underlying the rating curve model (unaccounted variables, the intrinsic stochasticity of the process). $\mathcal{R}$ stands for the rating curve model and Θ for its parameters.

4. Using individual stage-discharge measurements $(\tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{H}})$, estimate the parameters posterior distribution

$$\mathrm{p}(\Theta \mid \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{H}}, \mathcal{R}) \propto \iint \mathrm{p}_d(\tilde{\boldsymbol{Q}} \mid \boldsymbol{Q}) \mathrm{p}(\boldsymbol{Q} \mid \boldsymbol{H}, \Theta, \mathcal{R}) \mathrm{p}_s(\tilde{\boldsymbol{H}} \mid \boldsymbol{H}) \pi(\boldsymbol{H}) \pi(\Theta) \, \mathrm{d}\boldsymbol{h} \, \mathrm{d}\boldsymbol{q}. \tag{4.7}$$

5. Given a stage measurement h, compute the probability of the true discharge q

$$\mathrm{p}(q \mid \tilde{h}, \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{H}}, \mathcal{R}) \propto \iint \mathrm{p}(q \mid h, \Theta, \mathcal{R}) \mathrm{p}(\Theta \mid \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{H}}, \mathcal{R}) \mathrm{p}(\tilde{h} \mid h) \pi(h) \, \mathrm{d}h \, \mathrm{d}\Theta. \tag{4.8}$$

The next step is to include this information about the true discharge into the calibration of the hydrological model. Introducing the true stage $\boldsymbol{h}$, the true rainfall $\boldsymbol{x}$, the true flow $\boldsymbol{q}$ and the parameters Θ as latent variables, the hydrological model

parameter posterior distribution can be written as

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{h}}, \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{H}}, \mathcal{M}, \mathcal{R}) \propto \iiint \mathrm{p}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\mathrm{p}(\tilde{\boldsymbol{h}} \mid \boldsymbol{h})\mathrm{p}(\boldsymbol{q} \mid \boldsymbol{x}, \boldsymbol{\theta}, \mathcal{M}, \boldsymbol{h}, \Theta, \mathcal{R}) \times \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}, \Theta, \boldsymbol{h} \mid \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{H}}, \mathcal{M}, \mathcal{R})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{h}\, \mathrm{d}\boldsymbol{q}\, \mathrm{d}\Theta. \quad (4.9)$$

Since the rating curve model and the hydrological model provide two independent assessments of the flow, the following hypothesis holds:

$$\mathrm{p}(\boldsymbol{q} \mid \boldsymbol{x}, \boldsymbol{\theta}, \mathcal{M}, \boldsymbol{h}, \Theta, \mathcal{R}) = \mathrm{p}(\boldsymbol{q} \mid \boldsymbol{x}, \boldsymbol{\theta}, \mathcal{M})\mathrm{p}(\boldsymbol{q} \mid \boldsymbol{h}, \Theta, \mathcal{R}), \quad (4.10)$$

meaning that the true flow can be predicted by both the stage measurement (through the rating curve model $\mathcal{R}$) and the rainfall (through the hydrological model $\mathcal{M}$). Using equation (4.10), equation (4.9) becomes

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{h}}, \mathcal{M}, \mathcal{R}) \propto \iiint \mathrm{p}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\mathrm{p}(\tilde{\boldsymbol{h}} \mid \boldsymbol{h})\mathrm{p}(\boldsymbol{q} \mid \boldsymbol{x}, \boldsymbol{\theta}, \mathcal{M})\mathrm{p}(\boldsymbol{q} \mid \boldsymbol{h}, \Theta, \mathcal{R}) \times \pi(\boldsymbol{\theta})\pi(\boldsymbol{x})\mathrm{p}(\Theta \mid \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{H}}, \mathcal{R})\pi(\boldsymbol{h})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{h}\, \mathrm{d}\boldsymbol{q}\, \mathrm{d}\Theta. \quad (4.11)$$

which can be solved by plugging equation (4.7) (the rating curve parameter distribution).

A more ambitious objective than realistically describing errors is to reduce them by incorporating additional independent information to the priors and error models. An example of such information is the measure of the local gravity field by the GRACE satellite as a proxy for the underground water level (*Wahr et al.*, 2004). Although the estimates obtained through remote observations may be crude, they nevertheless carry precious information that could help constrain or validate model behavior. For precipitation estimation, satellite images (visible and infrared) as well

as radar signatures could be part of an input error model that would allow more precise estimation of the true rainfall. Generally speaking, instead of simply using $\mathrm{p}(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})$ where $\tilde{\boldsymbol{x}}$ stands for gauge station measurements, a much more informative error model incorporating multiple sources of information could be defined:

$$\mathrm{p}(\boldsymbol{x} \mid \text{station}_A, \text{station}_B, \text{radar signature}, \text{satellite image}, \text{Climatic Regional Model}).$$

This approach has the advantage of modularity: the hydrological expects true inputs, and the input error model provides them by any available means.

4.6.2 Application to daily and hourly models

There are two main difficulties with the application of the proposed method to models working at daily or hourly time steps. The first one is the number of data involved in the computation. Indeed, for only one year of data, there are 365 true input latent variables and 365 true output latent variables to consider for a daily model. Adding the initial state variables, over 700 parameters must be sampled simultaneously. Sampling in such high dimensional spaces is feasible with MCMC techniques, but the computational burden becomes heavier than what most desktop computers can handle nowadays in a reasonable amount of time (less than a day). This difficulty is rather numerical than fundamental and could be solved, at least partially, by sampling multiple MCMC chains on parallel processors.

A more interesting issue is that at high time resolution, errors are expected to be partially correlated. In *Kavetski et al.* (2003), the proposed solution is to divide the time series into storms and assume that the rainfall errors within each storm are identical. A less drastic solution would be to assume errors are modeled by

an autoregressive process with a strong correlation coefficient or by a fractionally differenced process (*Smith and Harris*, 1987), better at capturing long range dependencies. In any case, assessing correlations in error series might be a partial solution to the dimensionality problem. Indeed, the assumptions that errors are dependent actually reduces the effective number of latent variables. Using a MCMC block sampling algorithm, these correlations can be exploited by the sampler to improve its convergence rate and reduce the number of steps necessary to reach stationarity (*Haario et al.*, 2001).

4.6.3 Structural error analysis

As shown in the second paper, a Bayesian framework allows the *a posteriori* separation of structural errors from data errors. The inference of the structural errors series is however likely to be strongly dependent on the choice of input, output and structural error models. Since the analysis of error series has the potential to diagnose model failures and help in the evaluation of the performance of models, improvement of error series inference seems like a fruitful endeavor.

One solution to improve the inference of error series could be to calibrate multiple models simultaneously. The main idea is that the true input and output variables are, in principle, independent from the model used. In other words, in a perfect world with accurate error models and sufficiently long series, different models would infer similar true input and output series. The only thing that would be different from model to model is the structural error series. What this means is that simultaneously calibrating different models sharing the same true input and true output variables should improve the inference of common variables, as well as of the individual structural error series.

Formally, say there are N models $\mathcal{M}_i$ along with their parameters $\boldsymbol{\theta}_i$ for $i = 1, \ldots, N$, then the inference of the error series can be carried out as follows:

$$\mathrm{p}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \mathcal{M}_1, \ldots, \mathcal{M}_N, \boldsymbol{x}, \boldsymbol{y} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})\pi(\boldsymbol{x}) \times \prod_{i=1}^{N} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_i, \mathcal{M}_i)\pi(\boldsymbol{\theta}_i)\pi(\mathcal{M}_i), \quad (4.12)$$

where $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_i, \mathcal{M}_i)$ is the structural error model of model i and $\pi(\mathcal{M}_i)$ its prior probability. $\pi(\boldsymbol{\theta}_i)$ stands for the prior for model's i parameters, whose shape could be defined using the parent prior concept. Equation (4.12) has been derived using the same assumptions described in section 4.4.2. Another assumption that has been made is that knowing the structural errors are of one model does not yield cogent information about the structural errors of another model. Although for an experienced modeler this assumption may not be true, translating this knowledge into a statistical distribution might prove difficult. Marginalizing out the various parameters $\boldsymbol{\theta}_i$ and the models $\mathcal{M}_i$, the probability for the true variables $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ can be used to yield the distribution of the input error series $\tilde{\boldsymbol{x}} - \boldsymbol{x}$ and output error series $\tilde{\boldsymbol{y}} - \boldsymbol{y}$. The distribution of the structural error series can be obtained for each model by computing $\boldsymbol{y} - \mathcal{M}_i(\boldsymbol{\theta}_i, \boldsymbol{x})$. Besides strengthening the inference of data errors, this approach would allow the comparison of structural error series from different models, which might help identifying model components responsible of its success or failure.

In the same spirit, the properties of the structural error series could be used to assess the performance (efficiency, reliability) of each model. Such a measure would have the advantage of being independent (in principle) from data errors. In other words, the ability of the model to reproduce physical processes could be examined, to some extent, irrespective of the data quality.

In some situations, a global assessment of the predictive power of a model is requested. In those cases, not only the model efficiency must be evaluated, but also the accuracy of the error models and priors used for calibration and predictions; in other words, the generalized model combining the model and the error models must all be evaluated jointly. This could probably be achieved by computing the probability of a validation output series:

$$\mathrm{p}(\tilde{\boldsymbol{y}}_+ \mid \tilde{\boldsymbol{x}}_+, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}).$$

A high probability would mean that the generalized model can predict with accuracy the output series given the input series under its error model assumptions. A low probability would indicate either model failure or inaccurate error models (too vague or wrongly defined).

CHAPTER 5

Conclusion

This thesis explores issues related to the calibration of hydrological models in presence of multiple sources of uncertainty. The problem of calibration is an inferential one and can be treated formally from a Bayesian standpoint. The uncertainty framework used to calibrate model is an extended version of the framework proposed by *Kavetski et al.* (2003). Apart from output and input uncertainty, this extended framework accounts for structural and initial state uncertainty. Each source of uncertainty is described in probabilistic terms by formal statistical distributions. The model's parameter distribution as well as predictive distributions can be computed using different techniques.

Under certain assumptions regarding the model and error models, the parameter posterior distribution can be computed analytically. This case is studied in a paper calibrating the hydrological model "abc", focusing on the effect of input uncertainty on calibration results.

The second paper uses MCMC techniques to solve the parameter distribution, removing constraints on the choice of model and error models. The time needed to compute the posterior distribution is however dependent on the length of calibration time series, limiting "desktop" simulations to monthly models. In this second paper, calibration of the GR2M model is performed under different assumptions regarding

error models to understand better the effect of these assumption on the calibration results. Structural errors are considered explicitly, allowing the *a posteriori* examination of inferred error time series.

The third paper introduces the concept of parent prior, used to define consistent priors among different models. The paper applies the idea to select the "best" copula, while the thesis discusses how it could be translated to hydrology. The use of the parent prior can improve the rigor of model selection applications and of Bayesian model averaging.

Although the applications of the Bayesian framework are limited in this thesis to monthly models, most of the general conclusions drawn from the simulations should apply to more complex models with higher time resolution. In fact, since daily models are expected to be more sensitive to input errors than monthly models, the importance of the issues highlighted in the thesis may be magnified.

The fundamental question, however, is whether or not the benefits provided by taking into account multiple sources of uncertainty into account are worth the necessary efforts. The answer to this question of course depends on the stakeholders requirements. In a case where uncertainty assessment is crucial (high project costs, potential lives at risk), a rigorous assessment of predictive uncertainty is primordial. The answer also depends on the available resources. On a standard desktop computer, the calibration of a daily model over ten years of data might take two weeks to complete. Since Monte Carlo simulations generally scale well on parallel architectures, this time could be substantially decreased using a modest cluster. Finally, the most important point to consider is whether or not the uncertainties can be described accurately. Without cogent information about the sources of errors, uncertainty assessment is futile; this thesis shows clearly that the predictive

uncertainty estimated by a Bayesian calibration method can only be as good as the assumptions regarding data and structural uncertainty.

Uncertainty assessment forces modelers to describe as faithfully as possible the errors corrupting their data. This focus on the data and its meaning is rich in insight into the processes affecting hydrological measurements and, from a scientific point of view, a sound practice. Hopefully, uncertainty assessment will become a routine step of hydrological modeling, improving understanding of watershed processes and the quality of hydrological forecasts.

APPENDIX A

Bayesian analysis

For many scientists, bayesian analysis is a side discipline of statistics, one subset of the toolbox developped in the last hundred years to solve probabilistic problems. This view is even shared by some people using bayesian analysis daily in their work. A confronting view is that bayesian analysis is THE theory of probability ; a complete algebra to manipulate probabilistic expressions (*Jaynes and Bretthorst*, 2003). It is the latter view that is used in this thesis.

The first and foremost difference with the usual frequentist notion of probability is the definition of a probability. The standard textbook definition of probability is as a frequency of occurrence. In other words, if an experiment is repeated a large number of times, the probability of an event is the number of times this event occurs over the total number of times the experiment is carried out. This definition is restrictive because many experiments cannot be repeated a large number of times. If we ask about the probability that a bridge will collapse under a given strain, we certainly do not intend to build many identical bridges to find out.

For Bayesians, a probability is a degree of personal belief in an hypothesis. It is by definition anthropocentric, liable to change from person to person. This inherent subjectivity is not unscientific in the sense that it forces one to clearly state the

hypotheses behind the reasoning, instead of hiding behind hypothesis tests and cookbook recipes.

A.1 Bayes' theorem

The manipulation of probabilistic statements can be done using three laws (table A.1). The notation used for logical operators is given in table A.2. Using the product rule, the famous theorem formulated by Reverend Thomas Bayes' can be easily derived by the symetry of the product rule:

$$\mathrm{p}(A, B \mid I) = \mathrm{p}(A \mid B, I)\mathrm{p}(B \mid I) = \mathrm{p}(B \mid A, I)\mathrm{p}(A \mid I), \tag{A.1}$$

where A, B and I are statements (*eg.* it will rain tomorrow, rain was observed today, etc). Bayes' theorem is obtained simply by reorganizing the terms of equation (A.1):

$$\mathrm{p}(A \mid B, I) = \frac{\mathrm{p}(B \mid A, I)\,\mathrm{p}(A \mid I)}{\mathrm{p}(B \mid I)} .$$

Bayes' theorem is often used to estimate the parameters of a model. Imagine for instance one wishes to estimate the parameters θ of a statistical distribution f from a dataset D given some information about the context I. The problem can be stated as:

$$\overbrace{\mathrm{p}(\theta \mid D, I)}^{\text{posterior}} = \frac{\overbrace{f(D \mid \theta)}^{\text{likelihood}}\,\overbrace{\mathrm{p}(\theta \mid I)}^{\text{prior}}}{\mathrm{p}(D \mid I)} .$$

where $\mathrm{p}(D \mid I) = \int f(D \mid \theta)\,\mathrm{p}(\theta \mid I)\,\mathrm{d}\theta$ is the normalization constant.

Table A.1: Laws of probability theory

Sum rule	$\mathrm{p}(A + \overline{A} \mid I) = 1$
Extended sum rule	$\mathrm{p}(A + B \mid I) = \mathrm{p}(A \mid I) + \mathrm{p}(B \mid I) - \mathrm{p}(AB \mid I)$
Product rule	$\mathrm{p}(AB \mid I) = \mathrm{p}(A \mid BI)\mathrm{p}(B \mid I)$

Table A.2: Logical operators

$\mathrm{p}(A$ AND $B \mid I)$	$\Leftrightarrow$	$\mathrm{p}(A, B \mid I)$
$\mathrm{p}(A$ OR $B \mid I)$	$\Leftrightarrow$	$\mathrm{p}(A + B \mid I)$
$\mathrm{p}(\mathrm{NOT}\ A \mid I)$	$\Leftrightarrow$	$\mathrm{p}(\overline{A} \mid I)$

A.2 Continuous variables

Although the formal rules of probability are only defined for a finite number of propositions, it is possible to consider continuous variables as well. However, certain problems may occur when going from the discrete to the continuous case. Indeed, for a continuous variable x, $\mathrm{p}(x) = 0$ is not a probability but rather probability density, a probabilities are rather obtained by $\mathrm{p}(x)\,\mathrm{d}x$. In the following, the usual convention is followed and the differentials are not explictly written. Exception made of cases where improper priors are used, that is, priors whose integral diverges, this simplification causes no problem.

A.3 Latent variables

Latent, or nuisance, variables are parameters with which we are not directly concerned, but are useful to define probabilistic statements. Imagine for instance that one is interested in the annual cost of owning a car. Let's write this as a probabilistic statement, $\mathrm{p}(C \mid I)$, the probability of the cost C knowing the prior information I. For the sake of simplicity, it is assumed that the cost depends only on the brand of

the car (B) and the distance travelled (l). Since the cost depends on those variables but their value is not of interest, B and l are called latent variables.

To compute the cost probability, the 30 more popular brands of car found on the road are considered ($B \in \{B_1, B_2, \ldots, B_{30}\}$) and it is assumed that the distance travelled lies somewhere between 0 and 60000km. Formally,

$$\sum_{i=1}^{30} \mathrm{p}(B_i) = 1$$
$$\int_0^{60000} \mathrm{p}(l)\,\mathrm{d}l = 1,$$

where $\mathrm{p}(B_i)$ is the prior probability for the brand, and $\mathrm{p}(l)$ the prior probability for the distance travelled. Introducing the variables B and l into $\mathrm{p}(C \mid I)$ is simply a matter of inserting statements of probability equal to one and then using the product rule :

$$\mathrm{p}(C \mid I) = \sum_{i=1}^{30} \int_0^{60000} \mathrm{p}(C, B_i, l \mid I)\,\mathrm{d}l = \sum_{i=1}^{30} \int_0^{60000} \mathrm{p}(C \mid B_i, l, I)\,\mathrm{p}(B_i)\,\mathrm{p}(l)\,\mathrm{d}l. \tag{A.2}$$

So although equation (A.2) depends on the car brand and the distance travelled, the final probability $\mathrm{p}(C \mid I)$ is completely independent from these nuisance parameters.

Numerical implementation

Simulations for the "abc" paper were done using Matlab™, a computing environment designed to deal effectively with linear algebraic problems. Although the interface is intuitive and the computing engine reasonably fast, larger projects tend to scale badly due to the requirement that each function is defined in a single file. Moreover, optimizations are hard to achieve since it is painfull to add C or Fortran extension to Matlab code.

For those reasons, the code written for the GR2M paper is mostly written in Python. Python is an interpreted language (no need for compilation and linking) with a intuitive syntax making it easy to learn. However, it is not as fast as C or Fortran, a serious issue since the simulations for GR2M take days to complete. Fortunately, bottleneck components can be written in Fortran and used natively within Python. Many libraries are available for scientific purposes (*Oliphant*, 2007), database interaction and graphics (*Hunter*, 2007) making it an efficient tool for scientific scripting (*Langtangen*, 2005).

The code for the calibration of GR2M makes extensive use of PyMC 2.0 (*Fonnesbeck et al.*, 2007), an open source package written in Python providing MCMC algorithms to which the author of the thesis has contributed. This second version of PyMC introduces a model definition syntax that makes it easy and intuitive to

define a Bayesian problem. Once the Bayesian model is defined in terms of custom parameter and data objects, launching the MCMC sampler is a matter of one line of code. The simulation output, the parameter samples, can be stored in different kind of databases such as SQL or HDF5 format. Put together, the features of PyMC makes it relatively easy to run millions of MCMC jumps on complex models and store the results effortlessly. With further efforts to optimize and parallelize the code, it seems likely that PyMC could help solve the open problems outlined in this thesis.

Bibliography

Adam, J. C., and D. P. Lettenmaier (2003), Adjustment of global gridded precipitation for systematic bias, *Journal of Geophysical Research Atmospheres*, *108*(D9), 4257, doi: 10.1029/2002JD002499.

Andréassian, V., C. Perrin, C. Michel, I. Usart-Sanchez, and J. Lavabre (2001), Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models, *Journal of Hydrology*, *250*(1-4), 206–223, doi: 10.1016/S0022-1694(01)00437-1.

Andréassian, V., C. Perrin, and C. Michel (2004), Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models, *Journal of Hydrology*, *286*, 19–35, doi: 10.1016/j.jhydrol.2003.09.030.

Arulampalam, S., S. Maskell, N. Gordon, and T. Clapp (2002), A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking, *IEEE transactions on Signal Processing*, *50*(2), 174–188.

Azimi-Zonooz, Krajewski, Bowles, and Seo (1989), Spatial rainfall estimation by linear and non-linear co-kriging of radar-rainfall and raingage data, *Stochastic Environmental Research and Risk Assessment*, *3*, 51–67, doi: 10.1007/BF01543427.

Beven, K. (2006), A manifesto for the equifinality thesis, *Journal of Hydrology*, *320*(1-2), 18–36, doi: 10.1016/j.jhydrol.2005.07.007.

Beven, K., and A. Binley (1992), The future of distributed models : Model calibration and uncertainty prediction, *Hydrological Processes*, *6*, 279–298.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, *249*, 11–29, doi: 10.1016/S0022-1694(01)00421-8.

Blake, J. R., and J. C. Packman (2007), Identification and correction of water velocity measurement errors associated with ultrasonic Doppler flow monitoring, *Water and Environment Journal*, (0), doi: 10.1111/j.1747-6593.2007.00089.x, in Press.

Blöschl, G. (2001), Scaling in hydrology, *Hydrological Processes*, *15*, 709–711.

Borga, M. (2002), Accuracy of radar rainfall estimates for streamflow simulation, *Journal of Hydrology*, *267*(1-2), 26–39, doi: 10.1016/S0022-1694(02)00137-3.

Boughton, W. (2004), The Australian water balance model, *Environmental Modelling & Software*, *19*(10), 943–956.

Bretthorst, G. L. (1996), An introduction to model selection using probability theory as logic, in *Maximum Entropy and Bayesian Methods*, edited by G. Heidbreger, pp. 1–42.

Chaubey, I., C. Haan, S. Grunwald, and J. Salisbury (1999), Uncertainty in the model parameters due to spatial variability of rainfall, *Journal of Hydrology*, *220*, 48–61.

Chaubey, I., C. T. Haan, J. M. Salisbury, and S. Gurnwald (1999), Quantifying Model Output Uncertainty Due to Spatial Variability of Rainfall, *Journal of the American Water Resources Association*, *35*, 1113–1123.

Chopin, N. (2004), Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference, *The Annals of Statistics*, *32*(6), 2385–2411.

Daley, R. (1991), *Atmospheric data analysis*, Cambridge University Press.

Doucet, A., S. Godsill, and C. Andrieu (2000), On sequential Monte Carlo sampling methods for Bayesian filtering, *Statistics and Computing*, *10*, 197–208.

Duan, Q. (2003), Global optimization for watershed model calibration, in *Duan et al.* (2003).

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Research*, *28*(4), 1015–1031.

Duan, Q., H. V. Gupta, S. Sorooshian, A. N. Rousseau, and R. Turcotte (Eds.) (2003), *Calibration of Watershed Models*, *Water Science and Applications*, vol. 6, AGU.

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

Dunn, P. K., and G. K. Smyth (2005), Series evaluation of Tweedie exponential dispersion model densities, *Statistics and Computing*, *15*, 267–280, doi: 10.1007/s11222-005-4070-y.

Dymond, J. R., and R. Christian (1982), Accuracy of discharge determined from a rating curve, *Hydrological Sciences Journal*, *4*(12), 493–504.

Edijatno, and C. Michel (1989), Un modèle pluie-débit journalier à trois paramètres, *La Houille Blanche*, *2*, 113–122.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research*, *99*(C5), 10,143–10,162.

Evensen, G. (2003), The Ensemble Kalman Filter: theoretical formulation and practical implementation, *Ocean Dynamics*, *53*, 343–367, doi: 10.1007/s10236-003-0036-9.

Evensen, G. (2007), The combined parameter and state estimation problem, *Submitted to Computational Geosciences.*

Fenton, J. D., and R. J. Keller (2001), The calculation of streamflow from measurements of stage, *Tech. rep.*, Cooperative research center for catchment hydrology.

Fiering, M. B. (1967), *Streamflow Synthesis*, Harvard Univ. Press, Cambridge, Mass.

Fonnesbeck, C., D. Huard, and A. Patil (2007), *PyMC 2.0 User's Guide*, code.google.com/p/pymc/.

Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resources Research*, *32*(7), 2161–2174.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, Texts in Statistical Science, Chapman & Hall/CRC Press, London.

Genest, C., and A.-C. Favre (2007), Everything you always wanted to know about copula modeling but were afraid to ask, *Journal of Hydrologic Engineering*, *12*(4), 347–368, doi: 10.1061/ASCE1084-0699.

Genz, A. (1992), Numerical computation of multivariate normal probabilities, *J. Comp. Graph Stat.*, *1*, 141–149.

Gordon, N. J., D. J. Salmond, and A. F. M. Smit (1993), Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *Proc. Inst. Electr. Eng.*, *140*(2), 107–113.

Gupta, H. V., S. Sorooshian, and P. Yapo (1998), Towards improved calibration of hydrologic models : Multiple and non-commensurable measures of information, *Water Resources Research*, *34*(4), 751–764.

Gupta, H. V., S. Sorooshian, T. S. Hogue, and D. P. Boyle (2003), Advances in automatic calibration of watershed models, in *Duan et al.* (2003).

Haario, H., E. Saksman, and J. Tamminen (2001), An adaptive Metropolis algorithm, *Bernoulli*, *7*(2), 223–242.

Habib, E., and W. F. Krajewski (2002), Uncertainty Analysis of the TRMM Ground-Validation Radar-Rainfall Products: Application to the TEFLUN-B Field Campaign, *Journal of Applied Meteorology*, *41*(5), 558–572.

Habib, E., W. F. Krajewski, and A. Kruger (2001), Sampling errors of tipping-bucket rain gauge measurements, *Journal of Hydrologic Engineering*, *6*(2), 159–166.

Hoeting, J. A., D. Madigan, A. Raftery, and C. Volinski (1999), Bayesian model averaging: A tutorial, *Statistical Science*, *14*(4), 382–417.

Hornberger, G. M., and R. C. Spear (1981), An approach to the preliminary analysis of environmental systems, *Journal of Environmental Management*, *12*, 7–18.

Høybye, J., and D. Rosbjerg (1999), Effect of input and parameter uncertainties in rainfall-runoff simulations, *Journal of Hydrologic Engineering*, *4*(3), 214–224.

Huard, D., and A. Mailhot (2006), A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc", *Water Resources Research*, *42*, W07416, doi: 10.1029/2005WR004661.

Huard, D., and A. Mailhot (2007), Calibration of hydrological model GR2M using Bayesian uncertainty analysis, *Submitted to Water Ressources Research.*

Huard, D., G. Évin, and A.-C. Favre (2006), Bayesian Copula Selection, *Journal of Computational Statistics and Data Analysis*, *51*(2), 809–822, doi: 10.1016/j.csda.2005.08.010.

Humphrey, M., J. Istok, J. Lee, J. Hevesi, and A. Flint (1997), A new method for automated dynamic calibration of tipping-bucket rain gauges, *Journal of Atmospheric and Oceanic Technology*, *14*(6), 1513–1519.

Hunter, J. D. (2007), Matplotlib: A 2D graphics environment, *Computing in Science & Engineering*, *9*(3), 90–95, doi: 10.1109/MCSE.2007.55.

Jaynes, E. T., and G. L. Bretthorst (2003), *Probability theory : The logic of science*, Cambridge University Press.

Jeffreys, H. (1983), *Theory of probability*, The International series of monographs on physics, 3rd ed., Clarendon Press, Oxford.

Kabouya, M. (1990), Modélisation pluie-débit au pas de temps mensuel et annuel en Algérie Septentrionale, Ph.D. thesis, Orsay-Paris Sud.

Kalman, R. (1960), A new approach to linear filtering and prediction problems, *Transaction of the ASME – Journal of Basic Engineering*, pp. 35–45.

Kass, R. E., and L. Wasserman (1996), The selection of prior distributions by formal rules, *Journal of the American Statistical Association*, *91*(435), 1343–1370.

Kavetski, D., S. W. Franks, and G. Kuczera (2003), Confronting input uncertainty in environmental modeling, in *Duan et al.* (2003), pp. 49–68.

Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, *Journal of the Royal Statistical Society Series B-Statistical Methodology, 63*, 425–464.

Krajewski, W. F., A. Kruger, and V. Nespor (1998), Experimental and numerical studies of small-scale rainfall measurements and variability, *Water Science and Technology, 31*(11), 131–138.

Kundu, P. K., and R. K. Siddani (2007), A new class of probability distributions for describing the spatial statistics of area-averaged rainfall, *Journal of Geophysical Research - Atmospheres, In press.*

Kyriakidis, P. C., and A. G. Journel (1999), Geostatistical Space-Time Models: A Review, *Mathematical Geology, 31*, 651–684, doi: 10.1023/A:1007528426688.

Langtangen, H. P. (2005), *Python scripting for computational science*, Springer.

Liu, J. S., and R. Chen (1998), Sequential Monte Carlo methods for dynamic systems, *Journal of the American Statistical Association, 93*(443), 1032–1044.

Lopes, V. L. (1996), On the effect of uncertainty in spatial distribution of rainfall on catchment modelling, *CATENA, 28*(1-2), 107–119, doi: 10.1016/S0341-8162(96)00030-6.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology, 330*, 368–381, doi: 10.1016/j.jhydrol.2006.04.046.

Marshall, L., D. Nott, and A. Sharma (2005), Hydrological model selection: A Bayesian alternative, *Water Resources Research, 41*, W10422, doi: 10.1029/2004WR003719.

Mekis, E., and W. D. Hogg (1999), Rehabilitation and analysis of Canadian daily precipitation time series, *Atmosphere-Ocean, 37*(1), 53–85.

Metcalfe, R., B. Routledge, and K. Devine (1997), Rainfall measurement in Canada: Changing observational methods and archive adjustment procedures, *Journal of Climate, 10*, 92–101.

Michaud, J. D., and S. Sorooshian (1994), Effect of rainfall-sampling errors on simulations of desert flash floods, *Water Resources Research, 30*(10), 2765–2776.

Molini, A., L. Lanza, and P. La Barbera (2005), Improving the accuracy of tipping-bucket rain records using disaggregation techniques, *Atmospheric Research, 77*(1-4), 203–217, doi: 10.1016/j.atmosres.2004.12.013.

Moradkhani, H., K.-L. Hsu, H. Gupta, and S. Sorooshian (2005), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resources Research*, *41*(5), W05012, doi: 10.1029/2004WR003604.

Morissey, M. L., J. A. Maliekal, J. S. Greene, and J. Wang (1995), The uncertainty of simple spatial averages using rain gauge networks, *Water Resources Research*, *31*(8), 2011–2018.

Mouelhi, S., C. Michel, C. Perrin, and V. Andréassian (2006), Stepwise development of a two-parameter monthly water balance model, *Journal of Hydrology*, *318*, 200–214, doi: 10.1016/j.jhydrol.2005.06.014.

Moyeed, R., and R. Clarke (2005), The use of Bayesian methods for fitting rating curves, with case studies, *Advances in Water Resources*, *28*(8), 807–818, doi: 10.1016/j.advwatres.2005.02.005.

Nandakumar, N., and R. G. Mein (1997), Uncertainty in rainfall-runoff model simulations and the implications for predicting the hydrologic effects of land-use change, *Journal of Hydrology*, *192*, 211–232, doi: 10.1016/S0022-1694(96)03106-X.

Naud, C., D. Makowski, and M.-H. Jeuffroy (2007), Application of an interacting particle filter to improve nitrogen nutrition index predictions for winter wheat, *Ecological Modelling*, *In Press*.

Neal, R. M. (1993), Probabilistic inference using Markov Chain Monte Carlo methods, *Tech. rep.*, Department of Computer Science, University of Toronto.

Nelsen, R. B. (1999), *An introduction to copulas*, Springer, New-York.

Newman, A. J., and P. A. Kucera (2005), Gauging rainfall, *Fluent News*, *Fall*, 12–13.

Oliphant, T. E. (2007), Python for scientific computing, *Computing in Science & Engineering*, *9*(3), 10–20, doi: 10.1109/MCSE.2007.58.

Oudin, L., V. Andréassian, C. Perrin, and F. Anctil (2004), Locating the sources of low-pass behavior within rainfall-runoff models, *Water Resources Research*, *40*, W11101, doi: 10.1029/2004WR003291.

Oudin, L., F. Hervieu, C. Michel, C. Perrin, V. Andréassian, F. Anctil, and C. Loumagne (2005a), Which potential evapotranspiration input for a lumped rainfall-runoff model?: Part 2–Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling, *Journal of Hydrology*, *303*(1-4), 290–306, doi: 10.1016/j.jhydrol.2004.08.026.

Oudin, L., C. Michel, and F. Anctil (2005b), Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 1 – Can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs?, *Journal of Hydrology*, *303*(275-289), doi: 10.1016/j.jhydrol.2004.08.025.

Oudin, L., C. Perrin, T. Mathevet, V. Andréassian, and C. Michel (2006), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *Journal of Hydrology*, *320*, 62–83, doi: 10.1016/j.jhydrol.2005.07.016.

Pappenberger, F., and K. J. Beven (2006), Ignorance is bliss: Or seven reasons not to use uncertainty analysis, *Water Resources Research*, *42*, W05302, doi: 10.1029/2005WR004820.

Paturel, J., E. Servat, and A. Vassiliadis (1995), Sensitivity of conceptual rainfall-runoff algorithms to errors in input data – case of the GR2M model, *Journal of Hydrology*, *168*, 111–125, doi: 10.1016/0022-1694(94)02654-T.

Petersen-Overleir, A. (2004), Accounting for heteroscedasticity in rating curve estimates, *Journal of Hydrology*, *292*(1-4), 173–181, doi: 10.1016/j.jhydrol.2003.12.024.

Petersen-Overleir, A. (2005), Misspecification of extreme rainfall population statistics and frequency inference due to measurement bias and variability, *Atmospheric Research*, *75*(4), 283–300, doi: 10.1016/j.atmosres.2005.01.004.

Schaake, J., S. Cong, and Q. Duan (2006a), U.S. MOPEX data set, *Tech. rep.*, National Oceanic and Atmospheric Administration.

Schaake, J., Q. Duan, V. Andréassian, S. Franks, A. Hall, and G. Leavesley (2006b), The model parameter estimation experiment (MOPEX), *Journal of Hydrology*, *320*(1-2), 1–2, doi: 10.1016/j.jhydrol.2005.07.054.

Schaefli, B., and H. V. Gupta (2007), Do Nash values have value?, *Hydrological Processes*, *21*(15), 2075–2080, doi: 10.1002/hyp.6825.

Schmidt, A. R. (2002), Analysis of stage-discharge relations for open-channel flows and their associated uncertainties, Ph.D. thesis, University of Illinois.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, *6*(2), 461–464.

Singh, V. P. (1995), *Computer Models of Watershed Hydrology*, Water Resources Publications.

Singh, V. P. (1997), Effect of spatial and temporal variability in rainfall and watershed characteristics on stream flow hydrograph, *Hydrological Processes*, *11*(12),

1649–1669, doi: 10.1002/(SICI)1099-1085(19971015)11:12⟨1649::AID-HYP495⟩3.0.CO;2-1.

Sivillo, J. K., J. E. Ahlquist, and Z. Toth (1997), An ensemble forecasting primer, *Weather and Forecasting*, *12*(4), 809–818.

Smith, W., and C. M. Harris (1987), Fractionally differenced models for water quality time series, *Annals of Operations Research*, *9*(1), 399–420.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resources Research*, *16*(2), 430–442.

Storm, B., K. Jensen, and J. Refsgaard (1988), Estimation of catchment rainfall uncertainty and its influence on runoff prediction, *Nordic Hydrology NOHYBB*, *19*(2), 77–88.

Sun, X., M. Manton, and E. Ebert (2003), Regional rainfall estimation using double-kriging of raingauge and satellite observations, *BMRC Research Report 94*, Bureau of Meteorology Research Centre.

Tebaldi, C., R. Smith, D. Nychka, and L. Mearns (2005), Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles, *Journal of Climate*, *18*(10), 1524–1540.

Troutman, B. M. (1982), An analysis of input errors in precipitation-runoff models using regression with errors in the independent variables, *Water Resources Research*, *18*(4), 947–964.

Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of conceptual rainfall-runoff models caused by problems to identify model parameters and structure, *Hydrological Sciences Journal*, *44*, 779–798.

Vrugt, J. A., H. Gupta, W. Bouten, and S. Sorooshian (2003), A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resources Research*, *39*(8), 1201, doi: 10.1029/2002WR001642.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resources Research*, *41*, W01017, doi: 10.1029/2004WR003059.

Wahr, J., S. Swenson, V. Zlotnicki, and I. Velicogna (2004), Time-variable gravity from GRACE: First results, *Geophysical Research Letters*, *31*, L11501, doi: 10.1029/2004GL019779.

Weakliem, D. L. (1999), A critique of the bayesian information criterion for model selection, *Sociological Methods Research*, *27*(3), 359–397.

Weerts, A. H., and G. Y. H. El Serafy (2006), Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models, *Water Resources Research*, *42*, W09403, doi: 10.1029/2005WR004093.

Welch, G., and G. Bishop (2006), An introduction to the Kalman Filter, *Tech. rep.*, Department of Computer Science University of North Carolina at Chapel Hill.

Willems, P. (2001), Stochastic description of the rainfall input errors in lumped hydrological models, *Stochastic Environmental Research and Risk Assessment*, *15*(2), 132–152.

Wood, S. J., D. A. Jones, and R. J. Moore (2000a), Accuracy of rainfall measurement for scales of hydrological interest, *Hydrology and Earth System Sciences*, *4*(4), 531–543.

Wood, S. J., D. A. Jones, and R. J. Moore (2000b), Static and dynamic calibration of radar data for hydrological use, *Hydrology and Earth System Sciences*, *4*(4), 545–554.

Xu, C.-Y., and V. P. Singh (2002), Cross comparison of empirical equations for calculating potential evapotranspiration with data from Switzerland, *Water Resources Management*, *16*, 197–219.

Xu, C.-Y., and G. Vandewiele (1994), Sensitivity of monthly rainfall-runoff models to input errors and data length, *Hydrological Sciences Journal*, *39*(2), 157–176.

Yang, D., S. Isida, B. E. Goodison, and T. Gunther (1999), Bias correction of daily precipitation measurements for Greenland, *Journal of Geophysical Research*, *104*(D6), 6171–6181.

Yang, D., D. Kane, Z. Zhang, D. Legates, and B. Goodison (2005), Bias corrections of long-term (1973-2004) daily precipitation data over the northern regions, *Geophysical Research Letters*, *32*, L19501, doi: 10.1029/2005GL024057.

York, D. (1966), Least-Squares fitting of a straight line, *Canadian Journal of Physics*, *44*, 1079–1086.

Young, P. C. (2002), Advances in real-time flood forecasting, *Philosophical Transactions of the Royal Society*, *360*, 1433–1450.

Yue, S., T. B. M. J. Ouarda, B. Bobée, P. Legendre, , and P. Bruneau (2002), Approach for describing statistical properties of flood hydrograph, *J. Hydrologic Engineering*, *7*(2), 147–153, doi: 10.1061/(ASCE)1084-0699(2002)7:2(147).

Zawadski, I. (1973), Statistical properties of precipitation patterns, *Journal of Applied Meteorology*, *12*, 459–472.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, Inc.

Part II

Articles

A bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc"

Le modèle "abc" est un modèle hydrologique mensuel à seulement trois paramètres généralement utilisés à des fins pédagogiques. En ce qui concerne ce travail, son avantage principal réside dans le fait que les sorties du modèle, les décharges, peuvent s'écrire comme une foncton linéaire des entrées (les précipitations). Cette propriété nous permet de procéder analytiquement à l'intégration des variables de nuisance et simplifie les calculs requis pour évaluer la distribution *a posteriori* des paramètres dans le cas où les données d'entrée sont incertaines. Cet article présente ainsi la dérivation de la solution à cette intégration et son application au modèle "abc". Des simulations sont effectuées afin de montrer l'influence des erreurs sur les données d'entrée sur la distribution de paramètres. Ces simulations semblent indiquer que l'utilisation de la méthode permet d'obtenir des distributions de paramètres qui réflètent de manière cohérente les incertitudes attitrées aux données d'entrée.

WATER RESOURCES RESEARCH, VOL. 42, W07416, doi:10.1029/2005WR004661, 2006

A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc"

David Huard[1] and Alain Mailhot[1]

Received 12 October 2005; revised 20 March 2006; accepted 29 March 2006; published 22 July 2006.

[1] The impact of input errors in the calibration of watershed models is a recurrent theme in the water science literature. It is now acknowledged that hydrological models are sensitive to errors in the measures of precipitation and that those errors bias the model parameters estimated via the standard least squares (SLS) approach. This paper presents a Bayesian uncertainty framework allowing one to account for input, output, and structural (model) uncertainties in the calibration of a model. Using this framework, we study the impact of input uncertainty on the parameters of the hydrological model "abc." Mostly of academic interest, the "abc" model has a response linear to its input, allowing the closed form integration of nuisance variables under proper assumptions. Using those analytical solutions to compute the posterior density of the model parameters, some interesting observations can be made about their sensitivity to input errors. We provide an explanation for the bias identified in the SLS approach and show that in the input error context the prior on the input "true" value has a significant influence on the parameters' posterior density. Overall, the parameters obtained from the Bayesian method are more accurate, and the uncertainty over them is more realistic than with SLS. This method, however, is specific to linear models, while most hydrological models display strong nonlinearities. Further research is thus needed to demonstrate the applicability of the uncertainty framework to commonly used hydrological models.

Citation: Huard, D., and A. Mailhot (2006), A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc," *Water Resour. Res.*, *42*, W07416, doi:10.1029/2005WR004661.

1. Introduction

[2] *Gupta et al.* [2003] identify two important issues that need to be addressed in order to improve the calibration of hydrological models: accounting for all sources of uncertainty (input, state, structural, parameter and output uncertainties), and basing model calibration on multiple noncommensurable measures of model performance. This article tackles the first issue using Bayesian analysis. We propose a theoretical framework in which all sources of uncertainty are accounted for. Using this framework ensures that the calibration of the model and its predictions remain coherent despite the underlying uncertainties. This article does not, however, present a general algorithm to use this framework. Indeed, the resolution method we use is tailor-made for the chosen application and would be inadequate for the vast majority of hydrological models. It allows, however, a deeper look into the obstacles that any such algorithm will have to face.

[3] Section 2 introduces the backbone of the method, the Bayesian uncertainty framework. We discuss the different sources of uncertainties occurring in hydrological models and describe how they fit into this framework. The standard Bayesian approach to calibration is then presented, and we show how it can be modified to take various sources of uncertainties into account. Familiarity of the reader with Bayesian analysis is assumed. Note that although input, output, model and state uncertainties can be treated in the proposed framework, the paper focuses on input uncertainty.

[4] Section 3 applies the method to the seminal problem of fitting a straight line to a data set. This topic has been discussed extensively and provides a computationally simple benchmark for our method. Importantly, it allows a simple and intuitive interpretation of the issues related to input uncertainties.

[5] Section 4 details the application of the uncertainty framework to "abc," a pedagogical hydrological model. Chosen for its analytical properties, "abc" is linear with respect to its input, while exhibiting a behavior similar to that of more complex hydrological models. Section 5 proceeds with the calibration of the "abc" model in different settings using numerical simulations. The analysis of these results relies in part on those from the straight line model. Section 6 summarizes the most interesting observations about the treatment of input errors. Finally, section 7 discusses the simplifications made in the paper and the issues that will have to be tackled in the future.

2. Calibration, Uncertainties, and Bayesian Analysis

[6] In order to use any model, whether it describes physical, biological or hydrological processes, its parame-

[1]Centre Eau, Terre and Environnement, Institut National de la Recherche Scientifique, Quebec, Quebec, Canada.

0043-1397/06/2005WR004661$09.00

ters must be specified. Indeed, models almost always rely on some set of parameters, allowing the user to tune the model to a particular setting. Ultimately, the objective is to find the "best" set of parameters. What best means is quite subjective and relative to the situation, and we use it here meaning the set of parameters leading to the most accurate predictions possible given the data at hand. Uncertainties in the estimation of parameters arise from data errors and modeling errors. Note the distinction between an error, a difference between true values and measurements, and the uncertainty, the incomplete state of knowledge about the true values.

2.1. Challenge of Calibration

[7] The estimation of a model's parameters, the calibration, goes from deceptively simple to highly complex depending on the problem at hand and the user's requirements. A typical calibration involves two broad steps: defining a function measuring the agreement of model output with the data, and finding the parameters that maximize this function. In most applications, the agreement is described by e^{-S}, with $S \equiv \sum(y_i - \tilde{y}_i)^2$, the sum of the squared differences between model output and measurements. Since in the common usage we are only interested in the best fitting parameters, the extreme value of the function, there is no harm in taking the logarithm and minimizing S instead of maximizing the exponential. This approach is called the standard least squares (SLS) approach, and has been applied with success over the years. There are other measures of agreement, a review of some used in hydrology is given by *Gupta et al.* [1998]. The second step, finding the maximum of the function, is relatively easy in one dimension, but can become a daunting task for high-dimensional cases due to the presence of local maxima. Novel algorithms exploit stochastic methods to explore the parameter space while avoiding staying trapped into local maxima. While there is no guarantee that the global maximum is found, *Vrugt et al.* [2003] report reliable results using an efficient Markov chain Monte Carlo sampler.

[8] In general, the SLS approach provides sound estimates as long as the input data is precisely known. However, if significant input errors are present and if the model is sensitive to those errors, the parameter estimates are biased and the confidence intervals are much too optimistic [*Kavetski et al.*, 2002]. In the last thirty years or so, a number of studies have highlighted the sensitivity of hydrological models to input errors [*Troutman*, 1982; *Andreassian et al.*, 2001; *Oudin et al.*, 2005], as well as their sensitivity to model errors [*Engeland et al.*, 2005] (see *Mein and Brown* [1978] for a dated but interesting review). Yet, despite these warnings, few calibration methods directly address the issue of input uncertainty. Known exceptions are the generalized likelihood uncertainty estimation (GLUE) methodology [*Beven and Binley*, 1992], Bayesian total error analysis (BATEA) [*Kavetski et al.*, 2002], particle filters [*Moradkhani et al.*, 2005], simultaneous optimization and data assimilation (SODA) [*Vrugt et al.*, 2005], and Gaussian processes [*Kennedy and O'Hagan*, 2001]. While these methods provide seducing solutions to the treatment of input uncertainty (and model uncertainty for some), their concern for numerical efficiency and practicality overshadows important theoretical issues. Our aim in this paper is to temporarily lay aside practical considerations and address the fundamental issues related to uncertainties. To do this, the first step is to define an uncertainty framework linking the different sources of uncertainties to the data and model.

[9] The most promising avenue to include all types of uncertainties in the calibration process is certainly Bayesian analysis. Bayesian analysis consists in the manipulation of probability statements about hypotheses via two logical rules, the sum rule and the product rule [*Jaynes and Bretthorst*, 2003]. It is worthwhile mentioning that in this context, a probability has the usual common sense meaning of a degree of confidence. This contrasts with the usual statistical probability, defined as a frequency of occurrence. The thriving literature feeding the feud between Bayesians and "frequentists" [*Efron*, 1986; *Clark*, 2005] will no doubt be of interest to fans of passionate debates. On a more serious note, reference textbooks usually cited are those of *Bernardo and Smith* [1994] and *Gelman et al.* [1995].

2.2. Uncertainty Framework

[10] The first step to design a method able to account for uncertainties is to lay down an uncertainty framework. This framework describes how errors occur and propagate through the physical model. It is based on an idealization of the sampling and modeling processes. Hence it should be viewed as an approximation of how "real" errors influence data and modeling.

[11] Although not identical, the framework we propose is very similar to the one implicitly used by *Vrugt et al.* [2005] in the SODA method, based on ensemble Kalman filters. Their focus, however, was not the theoretical issues related to input errors but rather the implementation of a practical algorithm to consider different sources of errors.

2.2.1. Error Models

[12] The framework we propose (see Figure 1) assumes the existence of true variables and a true process. It further assumes that, with the knowledge of the true inputs and the true process, it is theoretically possible to determine exactly the true outputs. However, as P.-S. Laplace noted, such determinism is only theoretical. In practice, we only have access to a finite number of imperfect measurements, data, and a more or less naive understanding of nature's behavior: the model. Indeed, models typically work at scales very different from the natural scales, simulate only aggregated input and output variables, account for a handful of effects, neglect all exterior influences too difficult to measure or simulate and are limited by our understanding of the physical laws and our computing power. Models are nevertheless useful to validate new physical laws, understand phenomena, predict events and give decision makers realistic scenarios to compare projects or costs. The idea behind our strategy is to give those models a boost by coupling them to comprehensive error models.

[13] We will consider three different types of errors that affect the modeling process: input, output and structural errors. It is worthwhile to define the nature and origin of these errors since they are at the core of the method. Input errors are defined as the difference between the input data and the true inputs. They originate from the inherent imprecision of measurements, as well as from their imperfect representativeness. For example, a lumped hydrological model may take as true input the total amount of precipitation over the watershed during the last month. A pluviometer, however, only averages rainfall over a few square centimeters, and rainfall over the whole catchment must be

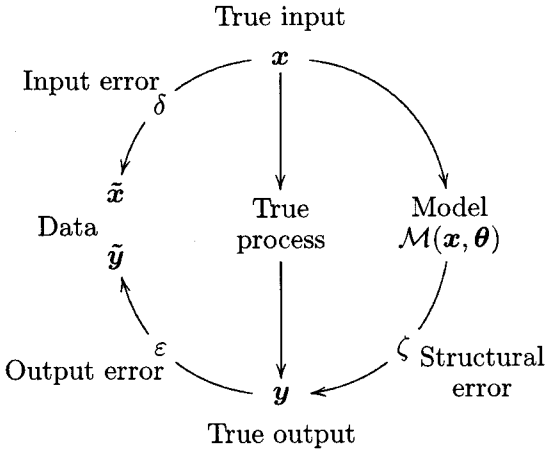

Figure 1. Idealized uncertainty framework.

extrapolated. Thus even with an infinitely precise pluviometer, there would still be a difference between the pluviometer readings and the theoretical input. This difference is what we call the input error.

[14] Output errors are defined analogously as the difference between true outputs and output measurements. Again, both the imprecision of measurements and the representativeness of the measures are sources of errors. In hydrological modeling, the output variable is typically the streamflow at the basin outlet. Extrapolated from the water height and a rating curve, streamflow measurements tend to lose accuracy as the flow increases.

[15] Structural errors are defined here as the difference between the true outputs and the model output using true inputs and true parameters. Structural errors can arise from incorrect modeling hypotheses or unmodeled processes [*Sorooshian and Dracup*, 1980]. They are stochastic by nature, due to the intrinsic variability of natural processes, but may nonetheless display distinct biases. For example, some hydrological models underestimate peak flow or overestimate base flow. Hence, in those cases, the distribution of structural errors is a function of the input variable.

[16] The errors affecting the data and model will be described by an error model, containing three elements: the input error model, the output error model and the structural error model. These three error models relate, probabilistically, the data taken to the true values. Denoting the input, output and structural errors by δ, ε and ζ respectively, the error models describe $f_\delta(\delta)$, $f_\varepsilon(\varepsilon)$ and $f_\zeta(\zeta)$, the error probability density functions (pdf).

[17] The two most obvious ways to relate errors and data are additively and multiplicatively. To allow the use of Gaussian distributions, we will work with additive errors:

$$\begin{aligned} &\text{Input} && \tilde{\mathbf{x}} = \mathbf{x} + \delta \\ &\text{Structural} && \mathbf{y} = \mathcal{M}(\mathbf{x}, \theta) + \zeta \\ &\text{Output} && \tilde{\mathbf{y}} = \mathbf{y} + \varepsilon, \end{aligned} \tag{1}$$

where $\mathbf{x}$, $\mathbf{y}$, θ stand for the vectors of true inputs, true outputs and true parameters respectively. Here and in the following, measurements are differentiated from true values by a tilde. Note that the structural error is defined by the difference between the true output and the model simulation using the true input $\mathbf{x}$ and true parameters θ.

2.2.2. Generalized Output Error Model

[18] Inspection of equation (1) suggests that the output error model can be made to include the structural error model

$$\tilde{\mathbf{y}} = \mathcal{M}(\mathbf{x}, \theta) + \varepsilon + \zeta.$$

Thus it is possible to replace the output error model and the structural error model by a "generalized" output error model, combining both errors. For the time being, however, we keep the two models apart since they are qualitatively and quantitatively different. Indeed, output errors are generally assumed to be unbiased, with homoskedastic or heteroskedastic variance. On the other hand, structural error are known to exhibit biases, due to incorrect or incomplete modeling of the underlying processes. Moreover, the magnitude and direction of this bias may be a function of the input variable.

2.2.3. Note on Parameter Uncertainty

[19] Our framework assumes the existence of input, output and structural uncertainties, but we have yet made no mention of the parameter uncertainty. The reason for this is that from our point of view, parameter errors are included in structural errors. Having a method that takes care of model errors, parameter errors lose their interest. We will, however, use the expression parameter uncertainty to describe our incomplete state of knowledge about the true parameters. Thus once the calibration is completed, there remains a parameter uncertainty owed to structural errors and limited supply of inexact data. By following the rules of probability, the resulting parameter uncertainty should incorporate the uncertainties about the model, inputs and outputs. Predictions can then be carried out by using not only the most probable parameter set, but the entire distribution. It is precisely the use of the whole parameter distribution that allows us to compute realistic confidence intervals about the predictions, incorporating the different sources of uncertainties.

2.3. Parameter Inference

2.3.1. Probability Inversion

[20] As stated earlier, the input, output and structural models define the probability of observing an error δ, ε and ζ. Yet, since the errors cannot be observed, it is more convenient to write the error models in terms of measured and true values:

$$\begin{aligned} &\text{Input} && p(\delta) \Leftrightarrow p(\tilde{\mathbf{x}}|\mathbf{x}) \\ &\text{Structural} && p(\zeta) \Leftrightarrow p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}) \\ &\text{Output} && p(\varepsilon) \Leftrightarrow p(\tilde{\mathbf{y}}|\mathbf{y}). \end{aligned} \tag{2}$$

The notation of probabilistic statements in this paper follows the general usage, in which p stands for the probability density. It should be remembered, however, that using the laws of probability on such expressions is a shortcut and that it can, in some cases, lead to paradoxes [*Jaynes and Bretthorst*, 2003].

[21] The next step is to combine the error models into a unified equation describing $p(\theta|\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathcal{M})$. To do so, we introduce $\mathbf{x}$ and $\mathbf{y}$ as nuisance variables and inverse the probability using Bayes' theorem:

$$\begin{aligned} p(\theta|\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathcal{M}) &= \iint p(\theta, \mathbf{x}, \mathbf{y}|\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathcal{M})\, d\mathbf{x}\, d\mathbf{y} \\ &= \iint p(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y}, \theta, \mathcal{M}) \\ &\quad \cdot p(\mathbf{x}, \mathbf{y}, \theta|\mathcal{M})\, d\mathbf{x}\, d\mathbf{y} \cdot \frac{1}{p(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}|\mathcal{M})} \\ &= \iint p(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y}) p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}) p(\mathbf{x}, \theta|\mathcal{M}) \\ &\quad \cdot d\mathbf{x}\, d\mathbf{y} \cdot \frac{1}{p(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})} \\ &= \iint p(\tilde{\mathbf{x}}|\mathbf{x}) \cdot p(\tilde{\mathbf{y}}|\mathbf{y}) p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}) p(\mathbf{x})\, d\mathbf{x}\, d\mathbf{y} \cdot \frac{p(\theta)}{p(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})} \end{aligned} \tag{3}$$

Equation (3) encloses the three error models, as well as a prior on the parameters $p(\theta)$, and a prior on the true input variable $p(\mathbf{x})$. We stress that this prior appears from the inversion of $p(\mathbf{x}|\tilde{\mathbf{x}})$ to $p(\tilde{\mathbf{x}}|\mathbf{x})$, the input error model defined above. As we will soon see, failure to distinguish both expressions leads to serious shortcomings in the estimation of parameters. Note that we used the simplifying assumption $p(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y}) = p(\tilde{\mathbf{y}}|\mathbf{y}) p(\tilde{\mathbf{x}}|\mathbf{x})$, that is, we supposed that the errors δ and ε were conditionally independent. While this is not mandatory, it simplifies the computations. Finally, the denominator $p(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ is simply a normalization constant, provided the integral in the numerator converges.

[22] Since this approach is derived from that of *Kavetski et al.* [2002], it is worthwhile to highlight the main difference between the two. In the work of *Kavetski et al.* [2002], the structural errors are not defined. Also, the nuisance variables $\mathbf{x}$ are not integrated but rather estimated, on the same footing as the model parameters. This approach makes sense since in their application, the number of nuisance variables is considerably lower than the sample size. In the case where there is one nuisance variable for each measurement, however, their approach would suffer from overparametrization.

2.3.2. Standard Shortcuts

[23] In Bayesian analysis as in standard least squares schemes, we implicitly assume that the input variables are exact, that is $p(\tilde{\mathbf{x}}|\mathbf{x}) = \delta(\tilde{\mathbf{x}} - \mathbf{x})$. Integrating equation (3) over $\mathbf{x}$ under this assumption, we obtain:

$$\begin{aligned} p(\theta|\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathcal{M}) &= \int p(\tilde{\mathbf{y}}|\mathbf{y}) p(\mathbf{y}|\tilde{\mathbf{x}}, \theta, \mathcal{M})\, d\mathbf{y} \cdot \frac{p(\theta)}{p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})} \\ &= \frac{p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta, \mathcal{M}) p(\theta)}{p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})}, \end{aligned}$$

the usual result of Bayes' theorem. Further assuming that the likelihood $p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta, \mathcal{M})$ is Gaussian and that the prior $p(\theta)$ is uniform, we obtain the SLS solution.

[24] It is worthwhile to stress that in this derivation of the usual Bayesian or SLS solution, there is no assumption about the form of the structural error model. Indeed, the likelihood $p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta, \mathcal{M})$ is rather a generalized likelihood, combining the output error model and the structural error model via the convolution of both distributions:

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta, \mathcal{M}) = \int p(\tilde{\mathbf{y}}|\mathbf{y}) p(\mathbf{y}|\tilde{\mathbf{x}}, \theta, \mathcal{M})\, d\mathbf{y}. \tag{4}$$

Hence once an output error model and a structural error model are specified, the minimization of equation (4) with respect to θ leads to an optimal solution, taking structural and output uncertainties into account. While apparently simple, this treatment of structural uncertainties hides a formidable challenge, namely, the adequate characterization of the structural error model.

[25] Our prime interest in this paper, however, lies not with structural errors but with the input errors and their impact on the parameters. Our focus on this subject stems from two realizations. The first one is that falsely assuming exact input causes parameter estimates to be biased. This bias is observed for hydrological models [*Kavetski et al.*, 2002] as well as for basic linear regressions [*York*, 1966], and can have a significant impact on the reliability of hydrological predictions. The second one is that despite a growing number of publications on this subject, the processes by which input errors affect model parameters are still obscure. Therefore this article focuses on basic issues related to input errors using simple, academic models and error models. These simplifications allow us to concentrate on the essential problems, instead of diverting our attention to technical and numerical issues.

3. First Application: The Straight Line Model

[26] Despite its apparent simplicity, the straight line model "when both variables are subject to error" hides complex difficulties and has been the focus of a large number of publications from various research areas: statistics [*Lindley and El-Sayyad*, 1968; *Kendall and Stuart*, 1983; *Fuller*, 1987; *Cheng and Ness*, 1994], econometrics [*Zellner*, 1971; *Erickson*, 1989], physics [*York*, 1966; *Reed*, 1989; *Gull*, 1989], and image reconstruction [*Werman and Keren*, 2001]. In the following, we will review the standard least squares approach (also called ordinary least squares) and its inherent bias in presence of input errors. We will then apply the Bayesian uncertainty framework to better understand the origin of this bias on the slope. Finally, we will study the impact that priors have on the results.

3.1. Standard Least Squares

[27] The standard least squares (SLS) solution to an optimization problem dates back from Gauss, who invented the method to identify the orbit of Ceres. SLS works under the assumption that the independent variable is known exactly. The optimal parameters are then simply the ones that make the sum of the squared errors as small as possible. Stated otherwise, the SLS parameters are those maximizing the probability of "drawing" the errors from a Gaussian distribution with zero mean. The wide success enjoyed by SLS has since been linked to the exceptional mathematical properties of the Gaussian distribution: the product of two Gaussian distributions is an unnormalized Gaussian distribution, the convolution of two Gaussians is another Gaussian, it is the only distribution whose maximum likelihood estimate is also the arithmetic mean, it is the maximum

entropy distribution for fixed mean and variance and the limiting distribution of additive errors from any distribution. In contrast with its huge success, SLS fails when errors affect both dependent and independent variables [*York*, 1966].

[28] Let's assume that our model is a straight line passing through the origin $y = \theta x$. Different measurements are taken, denoted by $\tilde{x}_i$, $\tilde{y}_i$. To simplify the example, we will suppose that there is no output error, that is, $\tilde{y}_i = y_i$. Now, if we assume, correctly, that the output data is exact, and that errors are only on the input data, the SLS method provides the following estimate for θ:

$$\hat{\theta} = \frac{\sum \tilde{y}_i^2}{\sum \tilde{x}_i \tilde{y}_i}.$$

This solution is correct, in the sense that it converges toward the true value as the sample size increases. However, if we falsely assume that the errors are on the output variable and the inputs are exact, we find instead

$$\hat{\theta} = \frac{\sum \tilde{x}_i \tilde{y}_i}{\sum \tilde{x}_i^2} \sim \frac{\sum \tilde{x}_i \tilde{y}_i}{\sum x_i^2 + \sigma^2}. \quad (5)$$

Of course, if $\sigma^2 = 0$, there is no uncertainty and both solutions are equal. However, as the error variance σ^2 increases, the slope gets underestimated. Moreover, this bias does not decrease as the number of samples increases.

[29] In most real life situations, input and output data contain errors, and the slope will systematically be under or overestimated depending on which variable we choose to be the independent one. Some authors have chosen to compute both slopes, and then take the average. Far more ingenuous algorithms have since been devised [*York*, 1966], based on the individual weighting of each data. Our objective, however, is not to review the various solutions to the straight line model but to understand the origin of the bias.

3.2. Density Mapping Effect

[30] For the sake of simplicity and pedagogy, we will consider a simple situation.

[31] 1. The model $\mathcal{M}$ is a straight line with intercept at zero, i.e., $y = \theta x$, with $\theta = 3$.

[32] 2. There are no output nor structural error, $p(\tilde{y}|x, \theta, \mathcal{M}) = \delta(\tilde{y} - \theta x)$.

[33] 3. The input error is Gaussian $p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|x, \sigma = 1)$.

[34] 4. The prior on θ is uniform on the interval $[\theta_a = 0, \theta_b = 5]$.

[35] 5. The prior on the true value $p(x)$ is uniform on the interval $[x_a = 0, x_b = 10]$.

[36] Let's imagine we have a single input measurement at $\tilde{x} = 2$. Knowing that the true value of x lies around the measurement, with a Gaussian probability

$$p(x|\tilde{x}) \propto p(\tilde{x}|x)p(x)$$

$$\propto \mathcal{N}(x|\tilde{x}, \sigma) \qquad x \in [x_a, x_b],$$

we want to compute the probability that $\tilde{y}$ is the true value knowing the slope θ. After a short analysis, we obtain

$$p(\tilde{y}|\tilde{x}, \theta, \mathcal{M}) \propto \frac{1}{\theta}\mathcal{N}(\tilde{y}/\theta|\tilde{x}, \sigma) \qquad \tilde{y} \in [\theta x_a, \theta x_b]. \quad (6)$$

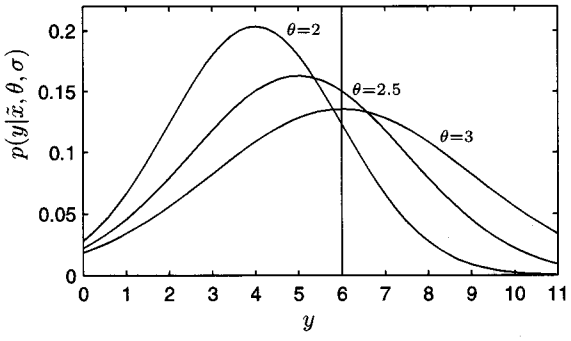

Figure 2. Probability density $p(y|\tilde{x}, \theta, \sigma)$ with $\tilde{x} = 2$, $\theta = 2$, 2.5, and 3, and $\sigma = 1$.

Figure 2 plots equation (6) for $\tilde{x} = 2$ and three values of θ: 2, 2.5 and 3. For small slopes, the y distribution is sharper than for bigger slopes. If we interpret function (6) as a mapping of the x elements onto the $\tilde{y}$ space, the values of $\tilde{y}$ are mapped more closely together with smaller slopes. Hence the pdf for $\tilde{y} = 6$ is higher with $\theta = 2.5$ than with $\theta = 3$, even though our intuitive guess, and the SLS estimate, would rather be $6/\tilde{x} = 3$.

[37] Suppose we now want to estimate the slope using measurements ($\tilde{x} = 2$, $\tilde{y} = 6$). Bayesian analysis tells us that

$$p(\theta|\tilde{x}, \tilde{y}, \mathcal{M}) \propto p(\tilde{y}|\tilde{x}, \theta, \mathcal{M})p(\theta), \quad (7)$$

but since $p(\theta) = \mathcal{U}(\theta_a, \theta_b)$, the probability for θ given $\tilde{y}$ is simply proportional to the probability of $\tilde{y}$ given θ. Thus, judging from Figure 2, $\theta = 2.5$ should be more probable than $\theta = 3$. Indeed, if we plot equation (7), we find that the most probable slope is around 2.5 (Figure 3).

[38] Figure 3 shows quite clearly how the uncertainty about x is asymmetrically mapped onto an uncertainty about θ. Of course, this asymmetry disappears as the uncertainty over x decreases. We will call the mapping of the input uncertainty onto the parameter uncertainty, the "density mapping" effect.

3.3. Straight Line Fitting

[39] A key question is whether or not this density mapping effect vanishes as the sample size increases. Unfortunately, it is difficult to find a simple answer based on analytical calculations. Hence we resort to numerical simulations to gain an insight into the problem. In this demonstration we use the first four assumptions of section 3.2. The only difference with the previous example is the number of samples and the prior on the true values, that will be modified to understand its impact on the parameter posterior pdf.

3.3.1. Synthetic Data

[40] We first generate a synthetic sample consisting of 500 couples $(\tilde{x}_i, \tilde{y}_i)$, shown in Figure 4. The true inputs x_i are randomly generated from a Gamma distribution, $x \sim \mathcal{G}(3, 1)$. The input measurements are synthesized by adding a Gaussian error to the true inputs: $\tilde{x}_i \sim \mathcal{N}(x_i, \sigma = 1)$. The output measurements are simply equal to the true outputs which are computed via the model and the true inputs: $\tilde{y}_i = y_i = \theta x_i$ with $\theta = 3$.

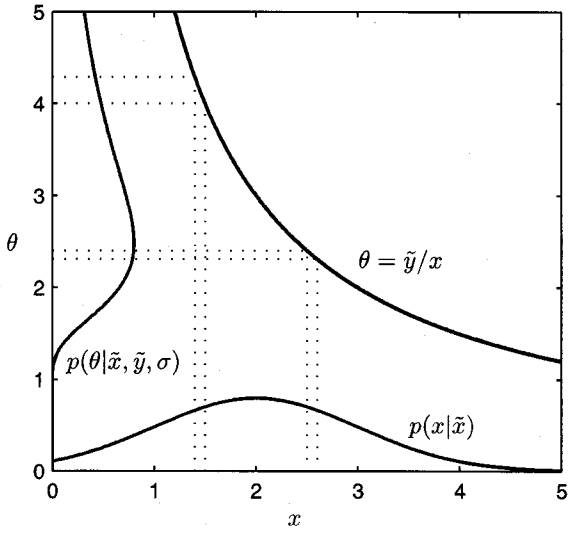

Figure 3. Probability densities $p(\theta|\tilde{x}, \tilde{y}, \sigma)$ and $p(x|\tilde{x})$ for $\tilde{x} = 2$, $\tilde{y} = 6$, and $\sigma = 1$. The hyperbola shows the relation between the slope and the true input for a given output $\tilde{y}$.

3.3.2. Error Models and Prior Hypotheses

[41] The parameter density is computed by recursively applying equation (3). That is, measurements are analyzed sequentially and update the posterior probability distribution $p(\theta_i|\tilde{x}_{1\ldots i}, \tilde{y}_{1\ldots i}, \mathcal{M})$. The input and generalized output error models are chosen correctly, that is, they correspond exactly to how the errors were applied to the true values:

$$p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|x, \sigma)$$

for the input error model and

$$p(\tilde{y}|x, \theta) = \delta(\tilde{y} - \theta x)$$

for the generalized output error model. The only remaining hypothesis concerns the prior for the true input values $p(x)$. To understand its impact on the calibration of θ, we look at three distinct cases, (1) Gamma, $p(x) = \mathcal{G}(x|3, 1)$; (2) Gaussian, $p(x) = \mathcal{N}(3, \sqrt{3})$; and (3) uniform, $p(x) = \mathcal{U}(0, 10)$, and compute the posterior pdf for the slope θ for each of these cases. We also compute the slope and the 90% confidence intervals using SLS.

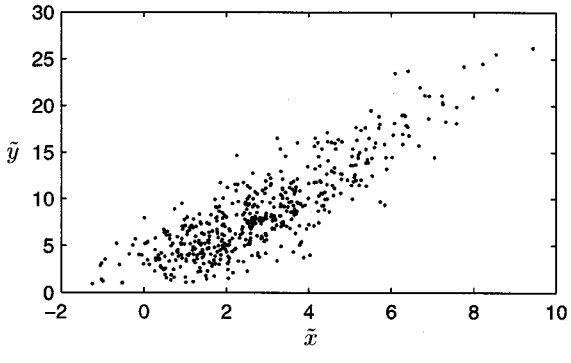

Figure 4. Synthetic data used to study the calibration of a straight line. The inputs $\tilde{\mathbf{x}}$ are computed using draws from a Gamma distribution to which a Gaussian noise is added. The outputs are given by $\tilde{y}_i = 3x_i$.

3.3.3. Results

[42] The results are shown in Figure 5. Slope estimates from the Gamma and Gaussian priors are very near to the true value, $\theta = 3$, whereas those of SLS and the uniform prior are significantly underestimated. The similarity between SLS results and the uniform prior is not a coincidence. Indeed, the parameter maximizing the posterior density using a uniform prior over the real domain is identical to the "major axis," a close cousin of SLS. That is, the slope for which the perpendicular distance from the points to the line is a minimum. Note that at this sample size, 500, the results do not vary much for different simulations.

[43] The most surprising observation about these results is that contrary to the Bayesian motto "the effect of the prior weakens with increasing sample size," the prior plays here a significant role even with a sample of 500 observations. Indeed, there is a significant difference between the posterior pdf using a uniform prior and a Gamma or Gaussian prior. Also note that the choice of the distribution parameters is also significant. A Gamma distribution with crude parameters may lead to worst results than a Gaussian with sensible ones.

3.3.4. Analysis

[44] The drastic effect of the prior can be explained as follows. For each input measurement, there is an unknown true input, whose value is inferred from the data using the error model and the prior on the true inputs. If the error model assumes a very small error, the effect of the prior is weak, whereas when the error model has a large variance, the prior plays an important role in the inference of the true input. Thus the effect of the prior does not weaken since with each new sample, another prior is added to infer the input true value. Moreover, if the prior is incorrect, each inference about the true value is flawed and the estimation of the parameters will not "converge" to the parameter's true value, even with an infinitely large sample.

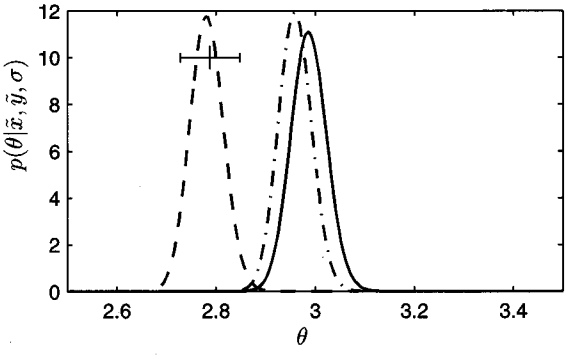

Figure 5. Probability density $p(\theta|\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \sigma)$ computed over a sample of 500 points using equation (3) with three different priors on the true values: Gamma (solid line), Gaussian (dash-dotted line), and uniform (dashed line). The cross indicates the SLS result as well as the 90% confidence interval. The true value of θ is 3.

[45] The importance of the prior on the results changes our perception of priors and how to choose them. There exists a wide variety of priors to choose from, an exhaustive review is given by *Kass and Wasserman* [1996]. In the light of the preceding results, we would recommend to avoid ignorance priors and rather look for additional evidence that would describe adequately the distribution of the true inputs. If a similar data set is available (other than the one under study), one solution is to fit a distribution to the data and use it as the prior. In the case where no such data sets exist, but where some descriptors such as the mean or variance are known (or can ge guessed), maximum entropy methods can be used to find the distribution with the highest entropy given those descriptors [*Jaynes*, 1983].

3.4. Highlights

[46] Before we go to the next section and begin our study of the hydrological model "abc," it is worthwhile to recall the highlights of this section: (1) The mapping of input uncertainties over the parameter space, the mapping density effect, is responsible for the "bias" observed in the parameter posterior density. (2) The influence of the mapping density effect decreases as data accumulate, but only as long as the prior on the true inputs allows reliable inference. That is, the estimated slope will only converge toward the true slope if the prior reflects the true input distribution.

4. Application to Hydrological Model "abc"

[47] We will now apply the Bayesian uncertainty framework to a simple hydrological model, the "abc" model. The "abc" model has been devised by Harold A. Thomas and introduced as a pedagogical tool by *Fiering* [1967]. Since then, it has been used as a benchmark for various computational methods [*Vogel and Sankarasubramanian*, 2003]. Its main advantages are that (1) it is linear with respect to the inputs, (2) it has only three parameters and one boundary condition, and (3) despite its simplicity, its calibration displays pathologies similar to more complex hydrological models.

[48] In the following, we first describe the "abc" model and how it can be written in a convenient form using linear algebra. Then, we describe the data set and justify the use of synthetic discharges in our numerical simulations. The next and last paragraphs discuss the error models chosen along with the prior distribution for the parameters and true values.

4.1. Description

[49] The "abc" model comprises two equations, one for the discharge Q_t and the other for water storage S_t, where t denotes time steps (annual or monthly). The model has three parameters (a, b, c) (hence the name) and a state parameter, S_t, the storage. The model is driven by the rainfall r_t:

$$\begin{aligned} Q_t &= (1-a-b)r_t + cS_t \\ S_{t+1} &= (1-c)S_t + ar_t \end{aligned} \tag{8}$$

Each parameter has a pseudophysical signification: a stands for the proportion of rainfall entering the storage, b is the proportion of rainfall lost to evapotranspiration and c is the percentage of water seeping from the storage to the basin outlet. All parameters take values in the interval [0, 1], and the conservation of water imposes an additional constraint on a and b, namely $a + b \leq 1$.

4.2. Matrix Formulation

[50] The main advantage of "abc" is that it can also be written elegantly in matrix form. To do so, we use induction to write the storage at the nth time step in terms of the initial storage and the rainfall history:

$$S_n = (1-c)^{n-1}S_1 + a\sum_{k=1}^{n-1}(1-c)^{n-k-1}r_k, \quad n > 1 \tag{9}$$

By defining the following vectors

$$\begin{aligned} \mathbf{x} &= [r_1, r_2, \ldots, r_n]^t \\ \mathbf{y} &= [Q_1, Q_2, \ldots, Q_n]^t \\ \mathbf{t} &= \left[c, c(1-c), c(1-c)^2, \ldots, c(1-c)^{n-1}\right]^t, \end{aligned}$$

and plugging equation (9) into (8), the model output $\mathbf{y}$ can be expressed in a convenient matrix formulation:

$$\mathbf{y} = \mathsf{A}\mathbf{x} + S_1\mathbf{t},$$

where matrix A is defined as

$$\mathsf{A} = \begin{bmatrix} (1-a-b) & 0 & 0 & \cdots & 0 \\ ac & (1-a-b) & 0 & & 0 \\ ac(1-c) & ac & (1-a-b) & & 0 \\ \vdots & & & \ddots & \vdots \\ ac(1-c)^{n-2} & ac(1-c)^{n-3} & ac(1-c)^{n-4} & \cdots & (1-a-b) \end{bmatrix}$$

4.3. Error Models Selection

[51] The selection of error models is restricted by a formidable requirement: closed form integrability. Indeed, equation (3) contains an integral over $2n$ variables. Standard numerical integration algorithms are not efficient with more than ten dimensions, and Markov chain Monte Carlo integration would take a huge amount of computing power to integrate over a mere 50 dimensions. Fortunately, Gaussian multivariate functions are readily integrable for any number of dimensions and are generally considered a valid depiction of a random error probability from an inferential point of view [*Jaynes and Bretthorst*, 2003]. For the particular context of "abc," we will thus limit our error models to multivariate Gaussian distributions in order to retain closed form integrability:

$$p(\tilde{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x}, \Sigma_\delta) \tag{10}$$

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \mathcal{N}(\tilde{\mathbf{y}}|\mathbf{y}, \Sigma_\varepsilon) \tag{11}$$

$$p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}) = \mathcal{N}\left(\mathbf{y}|A\mathbf{x} + S_1\mathbf{t}, \Sigma_\zeta\right), \tag{12}$$

where $\mathbf{x}$ and $\mathbf{y}$ are positively defined. Note that there are no constraints on the covariance matrices Σ_δ, Σ_ε and Σ_ζ. That

is, correlated and heteroskedastic errors can be assumed by assigning nonzero off diagonal terms and variances dependent on the input data. See *Sorooshian and Dracup* [1980] for examples of such matrices.

[52] At this point, we are ready to integrate equation (3) over **y**, effectively convoluting the structural and output error model into a generalized output error model:

$$\int_{\mathbb{R}} p(\tilde{\mathbf{y}}|\mathbf{y})p(\mathbf{y}|\mathbf{x},\theta,\mathcal{M})\,d\mathbf{y} = p(\tilde{\mathbf{y}}|\mathbf{x},\theta,\mathcal{M}) \tag{13}$$

Substituting functions (11) and (12) in (13) and integrating over the real domain, we find

$$p(\tilde{\mathbf{y}}|\mathbf{x},\theta,\mathcal{M}) = \mathcal{N}(\tilde{\mathbf{y}}|A\mathbf{x} + S_1\mathbf{t}, \Sigma_\epsilon), \tag{14}$$

where $\Sigma_\epsilon = \Sigma_\varepsilon + \Sigma_\zeta$. Note that since discharge is a purely positive quantity, the integral in equation (13) should only span the positive orthant. The integration over the real domain is an approximation, acceptable only if the means of $\mathbf{Ax} + S_1\mathbf{t}$ are three or four standard deviations away from 0. The same comment applies to the integration over **x**. Before we can proceed with the integration over the input variable, the prior $p(\mathbf{x})$ must first be defined.

4.4. Prior on the True Rainfall

[53] We choose to define the prior on the true rainfall as a historical prior, that is, $p(\mathbf{x})$ will reflect the distribution of monthly rainfall in the region. Details about how this is done are given in section 5.2. For now, it suffices to say that the prior is described by a sum of three Gaussian distributions with different means and variances. It is therefore possible to describe distributions very different in shape from a unique Gaussian, while respecting the requirement that all error models are to be Gaussian.

$$p(\mathbf{x}) = \sum_{j=1}^{3} f_j \mathcal{N}(\mathbf{x}|\mathbf{R}_j, \Sigma_{Rj}), \tag{15}$$

where $\sum f_j = 1$, $\mathbf{R}_j = R_j\,[1, 1, \ldots, 1]^t$ and $\Sigma_{Rj} = \sigma_{Rj}^2 I$, I standing for the $n \times n$ identity matrix.

4.5. Integration of the Input Nuisance Variables

[54] The integral

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}},\theta,\mathcal{M}) = \int_{\mathbf{R}} p(\tilde{\mathbf{y}}|\mathbf{x},\theta,\mathcal{M})p(\tilde{\mathbf{x}}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x} \tag{16}$$

can now be solved using the identity:

$$K = \int_{\mathbf{R}} \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i)\,d\mathbf{x} = \frac{(2\pi)^{\frac{n}{2}}|\widetilde{\Sigma}|^{\frac{1}{2}}}{\prod_i (2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left\{\frac{1}{2}\left(\widetilde{\mu}^t\widetilde{\Sigma}^{-1}\widetilde{\mu} - \sum_{i=1}^{n}\mu_i^t\Sigma_i^{-1}\mu_i\right)\right\}. \tag{17}$$

where

$$\widetilde{\Sigma} = \left(\sum_{i=1}^{n}\Sigma_i^{-1}\right)^{-1}$$

$$\widetilde{\mu} = \widetilde{\Sigma}\left(\sum_{i=1}^{n}\Sigma_i^{-1}\mu_i\right).$$

[55] Putting equations (10), (14) and (15) into (16), we find using the identity (17):

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}},\theta,\mathcal{M}) = \sum_{j=1}^{3} f_j K_j, \tag{18}$$

using covariance and means

$$\widetilde{\Sigma}_j = \mathbf{A}^t\Sigma_\epsilon^{-1}\mathbf{A} + \Sigma_\delta^{-1} + \Sigma_{Rj}^{-1}$$

$$\widetilde{\mu}_j = \widetilde{\Sigma}_j\left[\mathbf{A}^t\Sigma_\epsilon^{-1}(\tilde{y} - S_1\mathbf{t}) + \Sigma_\delta^{-1}\tilde{x} + \Sigma_{Rj}^{-1}\mathbf{R}_j\right].$$

The calculations are detailed explicitly in Appendix A.

4.6. Prior on the Parameters

[56] The prior on the parameters a, b and c remains to be defined. The prior density should reflect our degree of confidence in the value of the parameters. Since the parameters take values between 0 and 1, we will use the Beta distribution for the sake of generality

$$\mathcal{B}(x|r,s) = \frac{1}{\beta(r,s)}x^{r-1}(1-x)^{s-1} \qquad 0 \le x \le 1,$$

where $(r, s) > 0$. The prior on c will thus be $p(c) = \mathcal{B}(x|r_c, s_c)$. The prior on a and b is a little bit more involved due to the constraint $a + b \le 1$. Indeed, this constraint suggests a bivariate distribution. To define $p(a, b)$, we impose a dummy prior $\mathcal{B}(q|r_{ab}, s_{ab})$ on the direct runoff $q = 1 - a - b$, and using variable substitution, we compute the corresponding prior on (a, b):

$$p(a,b) = \frac{\mathcal{B}(1-a-b|r_{ab},s_{ab})}{(a+b)}.$$

The prior for the parameters is then completely specified by

$$p(\theta) = p(a,b)p(c).$$

[57] Now that all the ingredients are assembled, we are ready for numerical computations. The following section presents various results that highlight the impact of input errors on the parameters and the workings of the Bayesian uncertainty framework.

5. Results From "abc"

[58] The results from "abc" share traits similar to those of the straight line: influence of the density mapping and sensitivity to the prior $p(\mathbf{x})$. Other properties are also observed, such as model filtering and parameter averaging. Each one of these observations will be discussed in the following paragraphs. As we will see, the "abc" model is more complex than the straight line, due to the fact that it has more parameters and a memory of past events, via water storage. Hence past uncertainties propagate to blur the actual state of the system, regardless of the initial state uncertainty.

[59] The effective display of results is a troublesome issue; there seems to be no perfect way to visualize a 3-D parameter space at a glance. The solution adopted was to cut

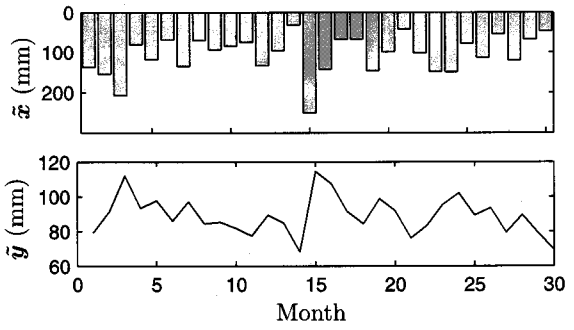

Figure 6. Inputs ($\tilde{\mathbf{x}}$) and outputs ($\tilde{\mathbf{y}}$) used in this study. The inputs are rainfall readings from a meteorological station on the Famine River. The outputs (discharges) were generated using the "abc" model with parameters $a = 0.6$, $b = 0.15$, and $c = 0.2$.

orthogonal slices in the parameter mesh at the posterior density's maximum (a_{max}, b_{max}, c_{max}). Figures will thus contain three plots, one for each slice. The color bar uses white to indicate a null probability and black for the maximum probability reached by the pdf.

5.1. Density Mapping Effect

[60] To demonstrate the effect of the density mapping on the parameters, we isolate this effect from other influences. We eliminate the effect of random noise by using exact data, specifying an input error model with a large variance, an output error model with a very small variance, a uniform prior on the true values and vague priors on the model parameters.

5.1.1. Data

[61] The true inputs $\mathbf{x}$ consist of monthly rainfall measurements from a station on the Famine River in the Beauce region, located south of Québec city. The true outputs $\mathbf{y}$ are the discharges simulated using the "true" rainfall and the "abc" model with parameters $a = 0.6$, $b = 0.15$ and $c = 0.2$ (the same parameters used by *Kavetski et al.* [2002]). The use of synthetic discharges allows us to check the consistency of the method and analyze the results more easily, that is, compare estimates to the "true parameters."

[62] In the following simulation, the sample ($\tilde{\mathbf{x}}$, $\tilde{\mathbf{y}}$) consists of 30 months of rainfall and stream flow. The data contain no errors ($\sigma_x = \sigma_y = 0$), thus ($\tilde{\mathbf{x}}$, $\tilde{\mathbf{y}}$) = ($\mathbf{x}$, $\mathbf{y}$) and the data fit perfectly the "abc" model with parameters $\theta = (0.6, 0.15, 0.2)$. Figure 6 shows a plot of the rainfall and stream flow used in this simulation.

5.1.2. Error Models and Prior Hypotheses

[63] Following are the modeling hypotheses, designed to bring out the density mapping effect and reduce other potential influences on the parameters posterior distribution: (1) homoskedastic input error model with a large standard error $\Sigma_\delta = 25^2 I$ mm^2 (almost 30% of the mean precipitation), (2) homoskedastic output error model with a negligible variance ($\Sigma_\epsilon = 1I$ mm^2), (3) uniform prior over the true rainfall ($p(\mathbf{x}) \propto 1$), and (4) vague prior for θ, that is, the prior described in section 4.5 with parameters $r_{ab} = s_{ab} = 1$ and $r_c = s_c = 1$.

5.1.3. Results and Analysis

[64] Computation of the parameters posterior density reveals a displacement of the cloud from the true parameters indicated by a cross (see Figure 7). This displacement amplifies as the assumed input uncertainty increases. Since all other influences have been curbed, we relate the pdf displacement from the true values to the density mapping effect. In "abc", its consequence is to displace the pdf density toward higher values of evapotranspiration b and infiltration a. In other words, the calibrated model has a lower percentage of direct runoff ($1 - a - b$) than the "real" model.

[65] The underestimation of the direct runoff can be explained using an analogy with the straight line model. We have seen in Figure 2 how smaller slopes map the outputs more densely than larger ones, with the consequence that the slope probability distribution has its maximum below the intuitive slope $\tilde{y}/\tilde{x}$. In the "abc" model, the direct runoff is akin to the slope: it specifies the instantaneous response of the model to the input. Models with high direct runoff display a great variability of discharges, and the output distribution is stretched over a wider range. Models with a low direct runoff, where storage plays the dominant part, exhibit more steady discharges, and the output distribution is more concentrated. Therefore, with a given discharge and an input uncertainty, we expect the most probable model to display a lower direct runoff than the "true" model.

[66] Now, one must realize that the density mapping effect is not an artifact of the Bayesian uncertainty framework but a very common phenomenon. When the input error is not accounted for, as in SLS for example, the effect is still present, biasing the results in an apparently uncontrollable way. The Bayesian framework merely provides the means to identify the effect, and as data accumulate, to reduce its effect on the parameters' pdf. The fact that the method accounts for the input uncertainties makes sure that the pdf is consistent with the data and reflects the input uncertainties' impact on calibration.

5.2. Prior $p(\mathbf{x})$ Impact

[67] As shown in section 3, the prior on the input true values plays a significant role in the estimation of the

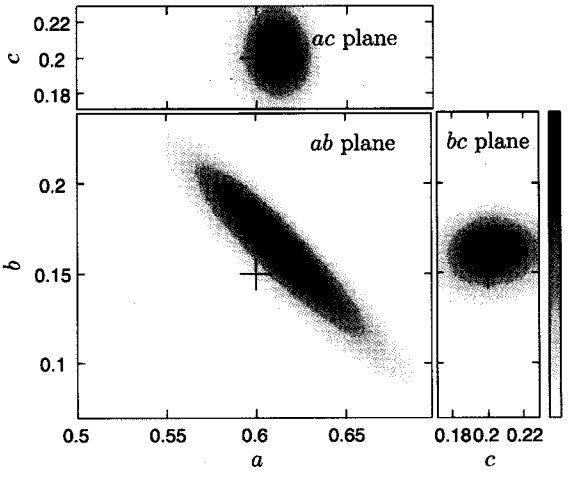

Figure 7. Density mapping effect on the parameters' posterior distribution. The slices are taken at the parameters most likely value [a, b, c] = [0.61, 0.16, 0.2]. True parameter values are indicated by crosses.

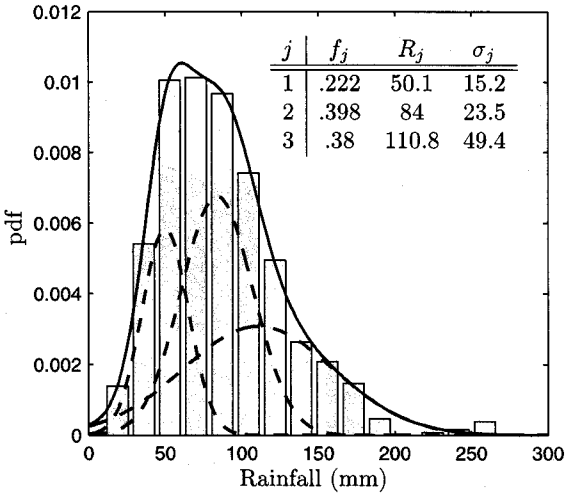

Figure 8. Histogram of St-Ephrem monthly rainfall based on data collected from 1929 to 2003. Superimposed (dashed lines) are the three Gaussian distributions whose sum (solid line) fits the histogram best and constitutes the historical prior $p(\mathbf{x})$ over the true rainfall.

parameters. In the last example, we used a uniform prior over the true rainfall to isolate the density mapping effect. We now use the same data and assumptions, except for a more realistic prior on the true inputs, an historical prior.

[68] To build this historical prior, we select a meteorological station (St-Ephrem) located near the watershed under study and plot an histogram of the monthly rainfall from 1929 to 2003. The best fitting distributions are Gamma, Weibull, and lognormal. However, if we choose to use any one of these distributions, the integral over $\mathbf{x}$ in equation (3) has no closed form solution. The trick is to fit the rainfall distribution by a sum of three Gaussian distributions (equation (15)), whose parameters are given in Figure 8.

[69] Although crude, this prior is adequate for our didactic purposes. In a real case study, however, it would be best to take seasonality into account. That is, specify a distinct prior for each month of the year. In the case where there is only one station on the watershed, the prior could also try to include the effect of spatially averaging the point rainfall input. The more relevant information that can be added to the prior, the better the calibration.

[70] The probability distribution of the parameters estimated using the historical prior is shown in Figure 9. Comparison with Figure 7 allows us to measure the difference made by the prior. The main change is in the storage parameter a, displaced to the left and achieving better agreement with the true parameters.

5.3. Model Filtering and Rainfall Smoothing

[71] An interesting observation from "abc" is that the model exhibits a form of filtering. By channeling rainfall in a storage compartment, the model is able to filter some of the noise imposed to the input data. That is, due to the time spent in storage, input errors are absorbed and averaged. This filtering makes the model relatively robust to nonbiased input errors. Indeed, the noise imposed to the precipitation must be considerable before its effect becomes apparent. Following this line of thought, a watershed with a very fast response (small size, flash floods) should be more sensitive to input errors than a large watershed with a long response time.

[72] Our last observation concerns the averaging of parameters due to the historical prior. When a historical prior is chosen for the true values of precipitation, and the input error model has a large variance, the prior plays an important role in determining the distribution of probable true rainfall. Indeed, as the variance of the input error model increases, the distribution of the true rainfall approaches the prior distribution. In other words, the intrinsic variability of rainfall embodied by measurements is discarded in favor of the historical knowledge about its distribution. By this process, the data are "smoothed" toward the prior's mode, reflecting more and more the average rainfall, and thus the average behavior of the model. Since the parameters a and c have, on average, no influence (they affect only the timing of discharges), they become increasingly difficult to estimate.

5.4. Comparison With SLS

[73] As can be seen, there are some pitfalls to avoid when using the Bayesian uncertainty framework, and the question is whether or not the quality of the calibration is worth the extra effort. To answer that question, we compare results obtained by our method assuming input and output uncertainties, with those obtained assuming only output uncertainties (this is effectively SLS grafted with priors on parameters).

5.4.1. Data

[74] The data is a series of 100 monthly rainfall measurements from the Famine River and the synthetic discharge computed using the "abc" model. In this case, a Gaussian noise is added both on the input and output data. The noise is homoskedastic and devoid of autocorrelation, with standard variations of $\sigma_x = 25$ mm and $\sigma_y = 15$ mm for the input and output variables respectively.

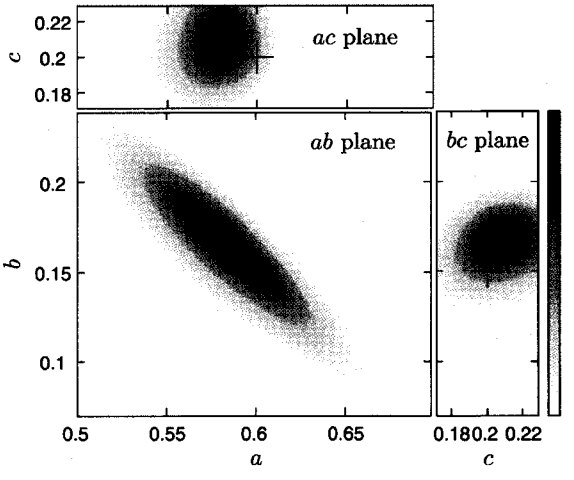

Figure 9. Impact of the prior for the true inputs. The slices are taken at the parameters most likely value $[a, b, c] = [0.59, 0.16, 0.2]$. The only difference with Figure 7 is the prior $p(x)$, now a historical prior instead of a uniform prior.

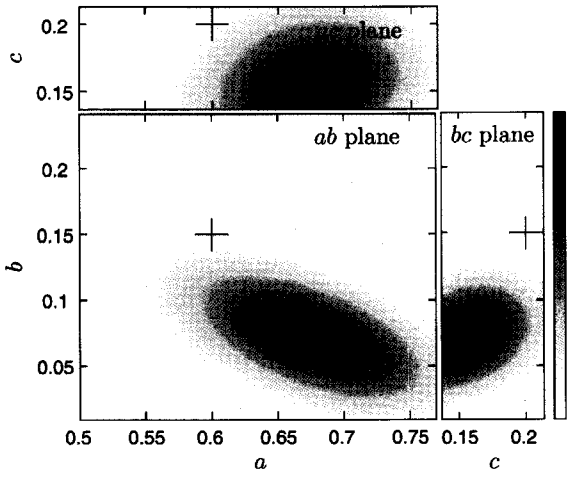

Figure 10. Parameters' posterior density, computed using only an output error model, thereby simulating a modified SLS algorithm. The slices are taken at the parameters most likely value [a, b, c] = [0.68, 0.07, 0.16].

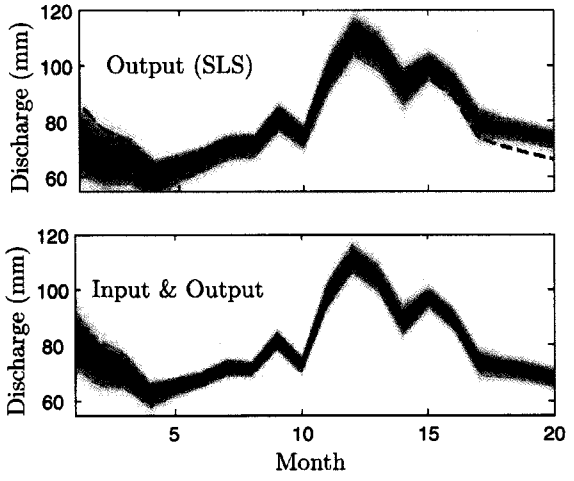

Figure 12. Comparison of predictions from two cases: (top) assuming only output errors and (bottom) assuming both input and output errors. The dashed line indicates the true synthetic discharge, and the solid line (SLS) indicates the discharge computed from the likeliest parameters.

5.4.2. Error Models and Prior Hypotheses

[75] The Bayesian method uses Gaussian error models with the correct covariance matrices ($\Sigma_\delta = 25^2 I$ mm^2 and $\Sigma_\epsilon = 15^2 I$ mm^2). Although the "SLS" method treats only the output error model, it is possible to include, partially, the effect of input errors on the parameters by defining an effective variance [*Orear*, 1982; *Lybanon*, 1984]. This effective variance is the sum of the output variance with the input variance multiplied by the model response: $\sigma^2_{SLS} = \sigma^2_\delta(1 - b)^2 + \sigma_\epsilon$. Here $1 - b$ is the average response of the model, so that using $b = 0.15$, we obtain $\Sigma_{SLS} = 26^2 I$ mm^2.

[76] The prior on the true inputs is the historical prior defined in section 5.2. The prior on the parameters (defined in section 4.5) is identical for the Bayesian and SLS calibrations.

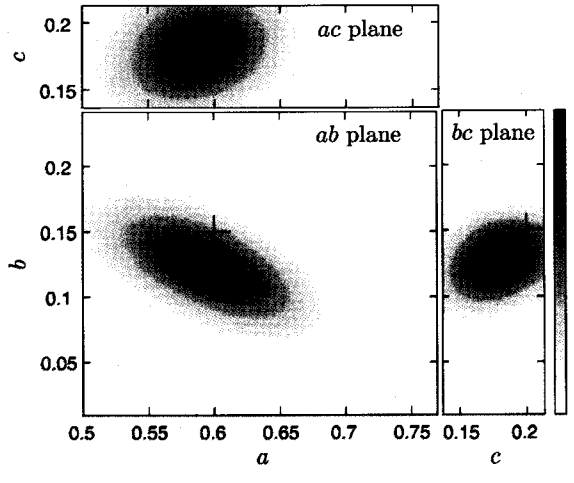

Figure 11. Parameters' posterior density, computed using an input and an output error model. The slices are taken at the parameters most likely value [a, b, c] = [0.59, 0.12, 0.18].

5.4.3. Results and Analysis

[77] The results are displayed in Figures 10 and 11. For this particular realization, the results obtained by the Bayesian method seem more accurate than those of SLS. Other simulations (not shown for lack of space) offer different pictures of the situation, but overall, the Bayesian density almost always presents superior results. What really matters, however, is not the parameters themselves but their prediction capability.

[78] To make predictions, we use as input 20 monthly rainfall values following the series used to calibrate the model. We uniformly sample the posterior density around 3000 times and simulate the discharge. Each discharge series is weighted by the probability of the parameter set, and a histogram of the discharge is computed. The results are linearly interpolated and displayed in Figure 12, where they can be compared to the true synthetic discharge (dashed line). The Bayesian predictions are closer to the synthetic discharge than those of SLS.

[79] The reader must be aware that these results constitute only one realization of a random Gaussian noise. The authors have conducted similar simulations where results are not so clear cut. In fact, when the noise is smaller, or when the data set is around $n = 30$, the predictions of SLS and of the Bayesian method are often very similar. Overall, it seems that Bayesian results are almost always better than those of SLS, the difference, however, may be negligible.

[80] Hence the decision of using the Bayesian method to account for input uncertainty depends upon two factors: (1) the impact of estimated input errors on model outputs and (2) the requirements of the end user. Since the difference between SLS and Bayesian estimates is directly related to the sensitivity of the model to input errors, the first step to assess the pertinence of a full Bayesian analysis is to evaluate the magnitude of input errors and their impact

on model predictions. If the impact is not significant with respect to output errors, then there is no need to include input uncertainties in the calibration process. If the impact is significant, the decision to use a Bayesian method relies on the needs of the end user. If the user only needs crude parameter estimates, there is no need to get fancy and a simple SLS will do. However, if the issue is sensitive and the reliable assessment of uncertainties plays a crucial role in the decision making process, using a method that considers input uncertainties would be preferable.

5.5. Comparison With *Kavetski et al.* [2002]

[81] Our study of the impact of input uncertainty was motivated in part by the article of *Kavetski et al.* [2002], on which we want to comment. The article presents a Bayesian method, BATEA, Bayesian Total Error Analysis, to account for input uncertainties. BATEA is applied to "abc", and the performances of BATEA and SLS are compared. The conclusion is that in the presence of input errors, SLS provides precisely wrong parameter estimates while BATEA gives probably right estimates. Indeed, the posterior pdf computed using BATEA are dead centered on the parameter's true values, a performance we could not replicate. We think this can be explained by the differences between the error models. In their case, the "abc" model is used with hourly time steps, where errors are likely to be highly correlated. Hence, instead of corrupting individual measurements, *Kavetski et al.* [2002] divide the rainfall history into storms, and multiply all the precipitations in a given storm by the same random factor. For example, for a time series of 1000 hourly precipitation, there may be five storm multipliers. These storm multipliers are then treated as parameters and estimated, rather than integrated, via a Bayesian analysis using MCMC algorithms.

[82] The method of *Kavetski et al.* [2002] is effective with large data sets and a relatively low number of nuisance variables. The method presented in this article deals with small data sets where each measurement has its own nuisance variable (the true value). In a sense, the two methods apply to different situations and cannot be easily compared in terms of performance or accuracy.

6. Summary

[83] The article describes a Bayesian framework to account for input, output and structural uncertainties in the calibration of models. The method is applied to two models: a simple straight line passing through the origin and the hydrological model "abc." In both cases, the main object of study is the impact of input uncertainties on the estimation of the model parameters. The analysis of those models has led us to some interesting observations concerning the calibration of models in the presence of input uncertainty.

[84] 1. The prior on the true inputs plays a major role. Contrary to the usual belief that the influence of the prior vanishes as data accumulate, the prior on the true inputs has a significative influence on the parameters estimation even for large samples. This is due to the fact that every datum added to the set comes with an additional prior, contributing to the shape of the posterior distribution.

[85] 2. Error models need to be carefully specified. The error models should be as realistic as possible, since underestimating the input error can bias the results, while overestimating the input error leads to the smoothing of inputs, averaging of the model behavior and difficulties in parameter estimation.

[86] 3. Biases are meaningful. The uncertainty on the parameter combines the output uncertainty and the input uncertainty via the model structure. Hence a bias on the parameter posterior density does not necessarily mean that the calibration is flawed, but rather that the "true" parameters are not those for which the measured outputs are the most probable. This displacement of the parameter posterior from the true values is related to how the model maps uncertain inputs onto outputs. In classical statistics, such displacements of the parameter's density maximum are called biases and regarded with suspicion. In the Bayesian framework, the origin of this bias is understandable, and we see no motivation to make corrections in order to obtain an "unbiased" estimator.

[87] 4. The entire posterior is significant. The fact that biases are commonplace in input error context underlines the importance of using the whole posterior distribution to make predictions, and not only the maximum of the distribution. By sampling the parameter's posterior to make predictions, we insure that the prediction uncertainty is faithful to the parameter uncertainty, and reflects the data and structural uncertainty.

7. Conclusion

[88] The uncertainty framework presented in the paper is general and theoretically, can be applied to any problem. The application to the "abc" model, however, made use of the fact that the model is linear in its input and that equation (3) could be integrated analytically. In general, hydrological models are nonlinear in their inputs, and such direct integration is impossible. The resolution method used here is thus inadequate for virtually all commonly used hydrological models. The usefulness of the uncertainty framework in applied hydrology is hence questionable until an algorithm is developed to apply it to nonlinear models. This is the subject of the authors' current research.

[89] The article bypasses the entire question of the hyperparameters. Indeed, in all simulations, we posed as known the error model parameters (noise variance, mean and correlation). In real cases, those hyperparameters must be estimated, or integrated as nuisance variable. The treatment of hyperparameters, computationally difficult, would have overshadowed the focus of the paper, treatment of input uncertainty. It is clear, however, that realistic applications of the method will have to tackle this issue.

[90] Another simplification concerns the state uncertainties. In the case of "abc", the model state is embodied by S_1, the initial storage. In our simulations, we took the initial storage as known, in order to simplify an already lengthy equation. A brief inspection, however, will convince the reader that it is relatively easy to integrate equation (18) with respect to S_1. Thus, for a uniform prior on the initial storage, an analytical formula for the parameters posterior density can be derived. The effect of our ignorance of the initial storage is then included into the uncertainty on the parameters.

[91] If the uncertainty framework can be successfully applied to nonlinear models, two issues will have to be addressed to make full use of its potential. The first one is

the specification of a rigorous structural error model. This is already a recognized open problem in hydrology. The second issue is the description of a detailed prior on the true values, taking scaling effects, seasonality and correlation into account. While these effects have been studied separately, their integration into a unique prior remains to be done.

[92] We hope by this article to promote the realistic assessment of uncertainties and their integration into model calibration methods. It is important to acknowledge the complexity of the calibration process and its importance for reliable predictions. The credibility of science is not built upon its vast knowledge, but upon its honesty in front of ignorance and uncertainty.

Appendix A: Calculations

[93] Here we detail the derivation of the posterior pdf for the parameters θ of the "abc" model. Starting with equation (16) and plugging the input and output error models along with the historical prior, we find

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}},\theta,\mathcal{M}) = \int_{\mathbf{R}} \mathcal{N}(\tilde{\mathbf{y}}|\mathbf{A}\mathbf{x}+S_1\mathbf{t},\Sigma_\epsilon)\mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x},\Sigma_\delta) \cdot \left[\sum_{j=1}^{3} f_j \mathcal{N}(\mathbf{x}|\mathbf{R}_j,\Sigma_{R_j})\right] d\mathbf{x} \quad \text{(A1)}$$

In order to apply the identity (17), we need to express the Gaussians as functions of $\mathbf{x}$:

$$\mathcal{N}(\tilde{\mathbf{y}}|\mathbf{A}\mathbf{x}+S_1\mathbf{t},\Sigma_\epsilon) = \mathcal{N}\left(\mathbf{x}|\mathbf{A}^{-1}(\tilde{\mathbf{y}}-S_1\mathbf{t}),\mathbf{A}^{-1}\Sigma_\epsilon\mathbf{A}^{-1'}\right)$$
$$\mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x},\Sigma_\delta) = \mathcal{N}(\mathbf{x}|\tilde{\mathbf{x}},\Sigma_\delta)$$

Substituting those last expressions into equation (A1) and writing it as the sum of a product of Gaussian, we can apply the identity for each element of the sum

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}},\theta,\mathcal{M}) = \sum_{j=1}^{3} f_j \int_{\mathbf{R}} \mathcal{N}(\mathbf{x}|\mu_1,\Sigma_1) \cdot \mathcal{N}(\mathbf{x}|\mu_2,\Sigma_2)\mathcal{N}\left(\mathbf{x}|\mu_{3j},\Sigma_{3j}\right) d\mathbf{x},$$

where

$$\mu_1 = \mathbf{A}^{-1}(\tilde{\mathbf{y}}-S_1\mathbf{t}) \quad \Sigma_1 = \mathbf{A}^{-1}\Sigma_\epsilon\mathbf{A}^{-1'}$$
$$\mu_2 = \tilde{\mathbf{x}} \quad \Sigma_2 = \Sigma_\delta$$
$$\mu_{3j} = \boldsymbol{R}_j \quad \Sigma_{3j} = \Sigma_{Rj}.$$

[94] **Acknowledgments.** The authors would like to thank the Associate Editor and reviewers for their insightful comments. David Huard is grateful for the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC Postgraduate Scholarships).

References

Andreassian, V., C. Perrin, C. Michel, I. Usart-Sanchez, and J. Lavabre (2001), Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models, *J. Hydrol.*, *250*(1–4), 206–223.

Bernardo, J., and A. Smith (1994), *Bayesian Theory*, John Wiley, Hoboken, N. J.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*, 279–298.

Cheng, C.-L., and J. W. V. Ness (1994), On estimating linear relationships when both variables are subject to errors, *J. R. Stat. Soc., Ser. B*, *56*(1), 167–183.

Clark, J. (2005), Why environmental scientists are becoming Bayesians, *Ecol. Lett.*, *8*(1), 2–14.

Efron, B. (1986), Why isn't everyone a Bayesian?, *Am. Stat.*, *40*(1), 1–5.

Engeland, K., C.-Y. Xu, and L. Gottshalk (2005), Assessing uncertainties in a conceptual water balance model using Bayesian methodology, *Hydrol. Sci. J.*, *50*(1), 45–63.

Erickson, T. (1989), Proper posteriors from improper priors for an unidentified errors-in-variables model, *Econometrica*, *57*(6), 1299–1316.

Fiering, M. B. (1967), *Streamflow Synthesis*, Harvard Univ. Press, Cambridge, Mass.

Fuller, W. A. (1987), *Measurement Error Models*, John Wiley, Hoboken, N. J.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, 526 pp., CRC Press, Boca Raton, Fla.

Gull, S. F. (1989), Bayesian data anaysis: Straight-line fitting, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, pp. 511–518, Springer, New York.

Gupta, H. V., S. Sorooshian, and P. Yapo (1998), Towards improved calibration of hydrologic models: Multiple and non-commensurable measures of information, *Water Resour. Res.*, *34*(4), 751–764.

Gupta, H. V., S. Sorooshian, T. S. Hogue, and D. P. Boyle (2003), Advances in automatic calibration of watershed models, in *Calibration of Watershed Models*, *Water Sci. Appl. Ser.*, vol. 6, edited by Q. Duan et al., pp. 9–28, AGU, Washington, D. C.

Jaynes, E. T. (1983), *E.T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, edited by R. D. Rosenkrantz, Hingham, Boston, Mass.

Jaynes, E. T., and G. L. Bretthorst (2003), *Probability Theory: The Logic of Science*, Cambridge Univ. Press, New York.

Kass, R. E., and L. Wasserman (1996), The selection of prior distributions by formal rules, *J. Am. Stat. Assoc.*, *91*(435), 1343–1370.

Kavetski, D., S. W. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modeling, in *Calibration of Watershed Models*, *Water Sci. Appl. Ser.*, vol. 6, edited by Q. Duan et al., pp. 49–68, AGU, Washington, D. C.

Kendall, M., and A. Stuart (1983), Functional and structural relationships, in *The Advanced Theory of Statistics*, vol. 2, 4th ed., chap. 29, Oxford Univ. Press, New York.

Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, *J. R. Stat. Soc., Ser. B*, *63*, 425–450.

Lindley, D. V., and G. M. El-Sayyad (1968), The Bayesian estimation of a linear functional relationship, *J. R. Stat. Soc., Ser. B*, *30*(1), 190–202.

Lybanon, M. (1984), A better least-squares method when both variables have uncertainties, *Am. J. Phys.*, *52*(1), 22–26.

Mein, R., and B. Brown (1978), Sensitivity of optimized parameters in watershed models, *Water Resour. Res.*, *14*(2), 299–303.

Moradkhani, H., K.-L. Hsu, H. Gupta, and S. Sorooshian (2005), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resour. Res.*, *41*, W05012, doi:10.1029/2004WR003604.

Orear, J. (1982), Least squares when both variables have uncertainties, *Am. J. Phys.*, *50*(10), 912–916.

Oudin, L., F. Hervieu, C. Michel, C. Perrin, V. Andreassian, F. Anctil, and C. Loumagne (2005), Which potential evapotranspiration input for a lumped rainfall-runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling, *J. Hydrol.*, *303*(1–4), 290–306.

Reed, B. C. (1989), Linear least-squares fits with errors in both coordinates, *Am. J. Phys.*, *57*(7), 642–646.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442.

Troutman, B. M. (1982), An analysis of input errors in precipitation-runoff models using regression with errors in independent variables, *Water Resour. Res.*, *18*(4), 947–964.

Vogel, R. M., and A. Sankarasubramanian (2003), Validation of a watershed model without calibration, *Water Resour. Res.*, *39*(10), 1292, doi:10.1029/2002WR001940.

Vrugt, J. A., H. Gupta, W. Bouten, and S. Sorooshian (2003), A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.

Werman, M., and D. Keren (2001), A Bayesian method for fitting parametric and nonparametric models to noisy data, *IEEE Trans. Pattern Anal. Mach. Intell.*, *23*(5), 528–534.

York, D. (1966), Least-squares fitting of a straight line, *Can. J. Phys.*, *44*, 1079–1086.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley, Hoboken, N. J.

D. Huard and A. Mailhot, Centre Eau, Terre et Environnement, Institut National de la Recherche Scientifique, Québec, QC, Canada G1K 9A9. (david.huard@ete.inrs.ca; alain.mailhot@ete.inrs.ca)

Calibration of hydrological model GR2M using Bayesian uncertainty analysis

La calibration des modèles hydrologiques en présence de multiples sources d'incertitude représente un des défis majeurs des sciences hydrologiques. Dans cet article, la calibration est considérée comme un problème d'inférence et est abordée à l'aide de l'analyse Bayésienne. Les sources d'incertitude affectant la modélisation hydrologique, incertitudes sur les données d'entrée et de sortie des modèles, incertitudes au sujet du modèle lui-même ainsi que les incertitudes liées à la méconnaissance des conditions initiales, sont décrites de manière probabiliste. La définition formelle de ces modèles d'erreur permet de dériver une équation décrivant la probabilité *a posteriori* des paramètres du modèle. Ce schéma d'incertitude est appliquée à la calibration du modèle GR2M, un modèle hydrologique mensuel. GR2M étant un modèle non linéaire, il n'existe pas de solution analytique permettant de calculer directement la probabilité des paramètres du modèle et l'algorithme Metropolis est utilisé pour générer une chaîne de Markov reproduisant empiriquement la distribution de ces paramètres. L'article s'intéresse à l'influence qu'ont les modèles d'erreurs sur le processus de calibration et de validation. Plus particulièrement, les thèmes étudiés sont le traitement des conditions initiales, l'influence des modèles d'erreurs sur la distribution de paramètres, l'interprétation des simulations en présence d'incertitude sur les données d'entrée et le traitement des erreurs structurales. Il est montré comment les erreurs structurales peuvent être, a posteriori, discernées des erreurs sur les données.

WATER RESOURCES RESEARCH, VOL. ???, XXXX, DOI:10.1029/,

Calibration of hydrological model GR2M using Bayesian uncertainty analysis

David Huard

Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Quebec (Qc), Canada

Alain Mailhot

Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Quebec (Qc), Canada

---

David Huard, Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Québec (Qc), Canada, G1K 9A9. (david.huard@ete.inrs.ca)

Alain Mailhot, Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Québec (Qc), Canada, G1K 9A9. (alain.mailhot@ete.inrs.ca)

Abstract. An outstanding issue of hydrological modeling is the adequate treatment of uncertainties in model calibration and prediction. The current paradigm is that the major sources of uncertainties, namely input, output and model uncertainty should be accounted for directly, instead of assuming they can be safely lumped into the output uncertainties. In this paper, Bayesian analysis is used to calibrate the conceptual hydrologic monthly model GR2M taking into account input, output, structural and initial state uncertainty through error models and priors. Calibration is performed under different error assumptions to study the influence of the initial state uncertainty, the consequences of large input errors, the impact of error assumptions on calibrated parameter posterior distributions and the definition of error models. It is shown how such an analysis can be used to separate, *a posteriori*, the different sources of errors, and in particular, to identify structural errors from data errors.

1. Introduction

Uncertainty analysis currently enjoys a considerable amount of attention in hydrology [*Beven*, 2006a]. The recent literature hosts spirited debates about the value of uncertainty analysis methods [*Todini and Mantovan*, 2007; *Beven et al.*, 2007; *Mantovan and Todini*, 2006], discussions about the necessity of validating such methods [*Hall et al.*, 2007] and also about the very definition of the word uncertainty [*Montanari*, 2007]. With this renewal of interest in uncertainty analysis, it is legitimate to raise the question of how uncertainty estimates may undermine the confidence of stakeholders in science when, for example, bounds are large. Not presenting the uncertainty estimates is certainly not worthy of a scientific approach and, as pointed out by *Beven* [2006b], may be, as well, a sure way of undermining the hydrological sciences. Once one recognizes the necessity of uncertainty analysis, one is forced to admit that its application requires more rigor and consistency [*Hall et al.*, 2007]. The efforts needed for uncertainty assessment are certainly important from an operational point of view, when one has to present results to stakeholder, but is also crucial when one wants to assess overall model performance.

The requirement for rigor and consistency can be achieved through Bayesian analysis. Within a Bayesian uncertainty assessment framework, a "generalized" model is put forward that grafts error models accounting for the various sources of uncertainties to the hydrological model itself. As the hydrological model encompasses our best understanding of hydrological processes occurring on a given watershed, error models should integrate and substantiate our knowledge of uncertainties affecting input and output data as well as those affecting the modeling process (structural uncertainties). The outcome of the

calibration procedure, namely the posterior distribution of model parameters and, consequently, the predictive uncertainty [*Todini and Mantovan*, 2007] is conditioned by the error models and priors. Hence, the overall "performance" or "reliability" of uncertainty assessment (which determine acceptability of model results [*Beven*, 2006b]) hinges upon our ability to define as precisely as possible adequate error models.

In a previous paper [*Huard and Mailhot*, 2006], a Bayesian framework inspired by the work of *Kavetski et al.* [2003] was used to calibrate "abc", a simple linear hydrological model [*Fiering*, 1967]. The linearity of the model and normality constraints on the error models allowed the derivation of an analytical solution to the parameter posterior distribution, allowing a better understanding of issues related to input errors. Linearity and normality constraints, however, made the resolution method unfit for common usage.

In this paper, the framework is extended by taking explicitly into account structural and initial state uncertainties. The parameter posterior distribution is evaluated using Markov Chain Monte Carlo (MCMC) sampling [*Kavetski et al.*, 2006a, b; *Kuczera and Parent*, 1998], thus removing linearity and normality constraints. A procedure is presented where error models and priors describing input, output, initial conditions and structural uncertainties are grafted to the hydrological model which is then calibrated. This approach enables the examination of issues related to: 1) the influence of initial state uncertainty (Under what conditions is it critical ?), 2) overconditioning due to input error assumptions (What does calibration mean in this context ?), 3) lumping of structural and output errors (When can it be done? What are the advantages ?), 4) definition of error models (What are their impact on the posterior parameter distribution ?) and 5) separation of model errors from data errors (How is it done ? How can it be used ?). The first four issues

are examined through simple examples using deliberately crude error model assumptions and the fifth one by a case study where more care is given to the specification of realistic error models.

Although this Bayesian uncertainty assessment framework is general and could be theoretically applied to any model, the numerical cost of evaluating the parameter posterior distribution is significant and grows with the length of the data series used for calibration. Hence, given the available computational resources, its application to daily or hourly models appears unwieldy. The method employed in the paper should hence be considered as a formal exercise, rather than an operational solution to the calibration of hydrological models (for such methods, see *Moradkhani et al.* [2005]; *Vrugt et al.* [2005]; *Kavetski et al.* [2003]; *Kennedy and O'Hagan* [2001] and *Beven and Binley* [1992]). In this spirit, the model chosen to illustrate the issues presented above is the parsimonious monthly model GR2M [*Mouelhi et al.*, 2006]. GR2M's nonlinear components make it sensitive to input errors [*Paturel et al.*, 1995], and hence relevant in an uncertainty analysis context. With only two free parameters, the posterior parameter distribution can be easily visualized and interpreted.

The model and data used for simulations are described in more detail in section 3, after the basic theoretical background and hypotheses underlying the proposed Bayesian uncertainty framework are explained in section 2. Section 4 discusses basic calibration and validation issues related to multiple sources of errors using simple error models. Section 5 presents a more realistic case study with carefully defined error models, along with a brief analysis of the results, including the *a posteriori* separation of the different types of

error. A brief discussion about the extension to daily model is included followed by the conclusion (section 6).

This paper relies heavily on the theory and methods of Bayesian analysis. The authors suggest the textbooks of *Gelman et al.* [1995] for an application focused point of view and *Jaynes and Bretthorst* [2003] for more theoretical discussions on the foundations of probability theory.

2. Bayesian uncertainty framework

This section introduces the proposed Bayesian uncertainty framework and derives equations to perform the calibration and validation of models in presence of multiple sources of uncertainties.

2.1. Basic concept

In a Bayesian calibration, the objective is to find the posterior distribution for the parameters $\boldsymbol{\theta}$ knowing the input data series $\tilde{\boldsymbol{x}}$ and output data series $\tilde{\boldsymbol{y}}$ as well as the model $\mathcal{M}$:

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \mathcal{M}). \tag{1}$$

Solving equation (1) may be very simple or extremely complex, depending on the modeler's assumptions about the sources of uncertainty that affect the calibration process. In hydrological modeling, many different sources of uncertainties must be tackled simultaneously. Indeed, according to *Beven* [2006b] "... we have unknown errors in input and boundary conditions that get processed non-linearly through a model that has structural errors and which is then compared with observations that have unknown measurement and commensurability error characteristics." Although accounting for all these sources

of errors realistically is a conceptual and technical challenge, not doing so can have undesirable consequences on modeling performance and the reliability of predictions [*Oudin et al.*, 2006; *Andréassian et al.*, 2004; *Kavetski et al.*, 2003; *Andréassian et al.*, 2001; *Nandakumar and Mein*, 1997; *Paturel et al.*, 1995; *Michaud and Sorooshian*, 1994; *Xu and Vandewiele*, 1994; *Troutman*, 1982]. In the following, a Bayesian approach is used to account for four different sources of uncertainty: the ignorance of the initial conditions, input data error, output data error and structural (model) error.

To describe the effect of those four types of uncertainties on calibration and validation, conceptual true variables describing the initial condition $\boldsymbol{\phi_0}$, the true input series $\boldsymbol{x}$ and the true output series $\boldsymbol{y}$ are introduced in eq. (1) as latent variables (idealized true variables $(\boldsymbol{x}, \boldsymbol{y})$ are what would be observed in the absence of data errors). Latent variables (or parameters) are variables useful to state a problem but whose effect is integrated out of the posterior of interest [*Jaynes and Bretthorst*, 2003]. Equation (1) then becomes

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \mathcal{M}) = \iiint \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi_0} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \mathcal{M}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{\phi_0}, \tag{2}$$

where it is assumed that the integrals span the admissible domain of each variable. Using these conceptual true values, hypotheses concerning the nature of errors can be translated probabilistically and equation (2) brought under a form where it can be solved computationally. Schematically, the inference process is displayed in figure 1: input and output data are linked to their respective true values by the input and output error models, and those true input and output variables are linked together by a structural error model, itself dependent on the model parameters and initial conditions.

2.2. Hypotheses, priors and error models

The first step is to use Bayes' theorem on eq. (2) so that data is conditioned on the true variables:

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \mathcal{M}) = \iiint \mathrm{p}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi_0}, \mathcal{M})\, \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi_0} \mid \mathcal{M})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\phi_0} \cdot \frac{1}{\mathrm{p}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})}. \tag{3}$$

Since the normalization constant $\mathrm{p}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ introduced by Bayes' theorem is of little relevance to the calibration process, it will be neglected in the following, with the equality relation replaced by proportionality sign.

For the next step, two hypotheses are made: 1. data errors cannot be inferred from the initial conditions, the model parameters or the model, and 2. input and output errors are conditionally independent given the true input and true output [*Kavetski et al.*, 2003]. The first hypothesis is significant, in that it defines data error models as entities independent from the model simulations. Note however that data errors are only meaningful with respect to the true variables, themselves dependent on the model's spatial and temporal scales. Input and output error models are hence defined only on the basis of the experimental protocol, apparatus leading to data acquisition and the relation between the observed variable and the true variable. Formally,

$$\mathrm{p}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi_0}, \mathcal{M}) = \mathrm{p}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \mid \boldsymbol{x}, \boldsymbol{y}) = \mathrm{p}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, \mathrm{p}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}), \tag{4}$$

where $\mathrm{p}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})$ is a statistical distribution defining the probability of measuring an input series $\tilde{\boldsymbol{x}}$ knowing the true input series $\boldsymbol{x}$. Similarly, $\mathrm{p}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})$ is a statistical distribution defining the probability of measuring $\tilde{\boldsymbol{y}}$ knowing the true output series $\boldsymbol{y}$. In the following, these particular distributions will be referred to as the input error model and the output error model, and denoted specifically by p_{in} and p_{out} as reminder of their meaning. The exact shape of these error models is up to the modeler and should capture the main sources

of data uncertainty. Defining realistic error models is a difficult task since data errors typically combine both measurement errors and commensurability errors (the difference between the variable the model expects and what is actually measured in the field) in ways that are site specific.

Using assumption (4), equation (3) becomes:

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \mathcal{M}) \propto \iiint \mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, \mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})\, \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\phi_0}, \mathcal{M})\, \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\phi_0} \mid \mathcal{M})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\phi_0}, \tag{5}$$

where $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\phi_0}, \mathcal{M})$ is a statistical distribution describing the probability of a true output series $\boldsymbol{y}$ knowing the simulated output series $\widehat{\boldsymbol{y}} \equiv \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\phi_0})$. In other words, this distribution describes the difference between model simulation and true output: the model error. It will hence be referred to as the structural error model, and denoted by $\mathrm{p}_{str}(\boldsymbol{y} \mid \widehat{\boldsymbol{y}}) \equiv \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\phi_0}, \mathcal{M})$ to shorten the notation.

The last step consists in splitting the prior $\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\phi_0} \mid \mathcal{M})$ into two independent priors $\pi_x(\boldsymbol{x} \mid \mathcal{M})\, \pi_{\theta,\phi_0}(\boldsymbol{\theta}, \boldsymbol{\phi_0} \mid \mathcal{M})$, where π is used, here and in the following, to denote priors (the subscript is dropped when there is no ambiguity about which prior is meant). While this separation is not mandatory, defining a joint multivariate prior distribution for the true input series, model parameters and state variables seems to be of a marginal interest given the length of the data series generally used in hydrology [*Mantovan and Todini*, 2006]. Equation (5) can then be written as

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \iiint \mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, \mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})\, \mathrm{p}_{str}(\boldsymbol{y} \mid \widehat{\boldsymbol{y}})\, \pi(\boldsymbol{x})\, \pi(\boldsymbol{\theta}, \boldsymbol{\phi_0})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\phi_0}, \tag{6}$$

where references to $\mathcal{M}$ were removed since the model is always assumed known.

Note that although $\pi(\boldsymbol{\theta}, \boldsymbol{\phi_0})$ and $\pi(\boldsymbol{x})$ are both priors, there is a notable difference between the two. Whereas $\boldsymbol{\theta}, \boldsymbol{\phi_0}$ are inferred from the whole data set, the inference on a

given true value x_i is based mostly on $\tilde{x}_i$ as well as $\tilde{y}_{j \geq i}$ through inferred parameters. If the errors are assumed large (vague input and output error models), then there is very little information in the data about x_i and the prior plays a major role in the inferential process [*Gull*, 1989]. The prior for the true input must hence be chosen with care, and modelers should avoid using abusively vague priors [*Kavetski et al.*, 2006a; *Huard and Mailhot*, 2006].

It is worthwhile to underline the conceptual differences between the different types of outputs that are defined in this Bayesian framework: simulations ($\widehat{\boldsymbol{y}}$) are the model outputs computed from inferred true inputs and parameters, output measurements ($\tilde{\boldsymbol{y}}$) are experimentally measured outputs, and true outputs ($\boldsymbol{y}$) are the idealized true values that would be measured if no measurement or commensurability errors were present. Another useful output that will be discussed later is the simulated output measurement ($\tilde{\boldsymbol{Y}}$), the output measurement predicted by the output error model given the inferred true output. In practice, different applications will require the use of different types of output. In water resource management contexts, decision makers will rather be interested in the inferred true outputs than the simulated outputs. In a model validation context however, field measurements should be compared to simulated output measurements to avoid mixing structural errors with measurement errors. Note that in many modeling studies, simulations are implicitly interpreted as the true values and compared directly with observations to assess model performance, with the ensuing risk of misinterpreting results.

2.3. Resolution method

In general, equation (6) has no analytical solution. One exception is when the model is linear with respect to the input variables and the error models as well as the true input prior $\pi(\boldsymbol{x})$ have a Gaussian shape. This particular case was studied in *Huard and Mailhot* [2006] using the "abc" model. However, typical hydrological models are nonlinear, eq. (6) cannot be integrated analytically and a different resolution method is needed. Unfortunately, numerical integration methods such as quadratures become inefficient for more than three or four dimensions, due to the curse of dimensionality: the computational cost increases exponentially with the dimension of the problem [*Novak and Ritter*, 1997]. Quadrature methods are thus inappropriate for this kind of application where the integral spans hundreds of dimensions (one for each true input and true output). In other words, the posterior probability cannot be evaluated directly. This is a frequent problem in Bayesian analysis, and a common solution is to sample parameters from the posterior distribution (6) using Monte Carlo Markov Chain (MCMC) algorithms [*Huang and Liang*, 2006; *Kavetski et al.*, 2006b; *Bates and Campbell*, 2001; *Kuczera and Parent*, 1998; *Gelman et al.*, 1995; *Neal*, 1993]. In this case the Metropolis algorithm is used, by which the parameter space is explored by making random steps. These steps are selected such that the distribution of the sampled parameters approaches the posterior distribution after a sufficient number of random steps (see *Chib and Greenberg* [1995] for an introduction to MCMC theory and more particularly Metropolis sampling). Hence, to solve equation (6), one samples $(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi_0})^i$ N times from $\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi_0} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$, and for N large enough (convergence can be assessed by different criteria [*Gelman et al.*, 1995]), the empirical distribution of the samples reproduces the probability distribution (superscripts

^{i} indicate sampled values). The integration over the latent variables is then performed by marginalizing over $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{\phi_0}$, that is, by considering the sampled $\boldsymbol{\theta}^i$ independently from the latent variables.

2.4. Prediction and validation

Hall et al. [2007] suggest that "without validation, calibration is worthless and so is uncertainty estimation.[...] Validation is also needed [...] because the data do not contain full information about how the catchment will respond in the future. The same argument applies to uncertainty estimation." In other words, validation checks not only how the calibrated model performs outside the calibration period, but also if the error model assumptions hold.

In a split-sample calibration/validation, a data series is split in a calibration series $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ and a validation series $(\tilde{\boldsymbol{x}}_+, \tilde{\boldsymbol{y}}_+)$. Calibration is performed under a set of assumptions defining the error models and priors, under which the posterior distribution $\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ for the parameters is computed. In validation, parameters can be seen as a source of uncertainty and thus treated as latent variables. Validation uses the entire posterior parameter distribution along with input $\tilde{\boldsymbol{x}}_+$ to infer the probable distribution of output measurements $\tilde{\boldsymbol{Y}}_+$, which can compared with the real observations $\tilde{\boldsymbol{y}}_+$ to validate the model and the calibration assumptions:

$$\begin{aligned}\mathrm{p}(\tilde{\boldsymbol{Y}}_+ \mid \tilde{\boldsymbol{x}}_+, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) &= \iiint \mathrm{p}(\tilde{\boldsymbol{Y}}_+, \boldsymbol{\theta}, \boldsymbol{x}_+, \boldsymbol{y}_+ \mid \tilde{\boldsymbol{x}}_+, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})\, \mathrm{d}\boldsymbol{x}_+\, \mathrm{d}\boldsymbol{y}_+\, \mathrm{d}\boldsymbol{\phi}\, \mathrm{d}\boldsymbol{\theta} \\ &= \iiint \mathrm{p}_{out}(\tilde{\boldsymbol{Y}}_+ \mid \boldsymbol{y}_+)\, \mathrm{p}_{str}(\boldsymbol{y}_+ \mid \widehat{\boldsymbol{y}}_+) \mathrm{p}_{in}(\tilde{\boldsymbol{x}}_+ \mid \boldsymbol{x}_+)\, \pi(\boldsymbol{x}_+) \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})\, \mathrm{d}\boldsymbol{x}_+\, \mathrm{d}\boldsymbol{y}_+\, \mathrm{d}\boldsymbol{\phi}\, \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{7}$$

The simulated output measurements $\tilde{\boldsymbol{Y}}_+^i$ can then be compared with the observed outputs $\tilde{\boldsymbol{y}}_+$ to assess the global efficiency of the model, that is, the combined accuracy of the model, error models and priors.

3. Model and data

3.1. GR2M

Simulations in this paper use the parsimonious hydrological lumped monthly model GR2M [*Mouelhi et al.*, 2006], an improved version of the model developed by *Edijatno and Michel* [1989] and *Kabouya* [1990]. Despite having only two free parameters, the model has been shown to perform well when compared to similar monthly models; on a benchmark test consisting of 410 basins throughout the world, GR2M shows the best performance among nine models, some of them counting five free parameters [*Mouelhi et al.*, 2006].

The two free parameters of GR2M are θ_1, the soil moisture storage maximum capacity, and θ_2, the water exchange term with neighboring catchments. The internal state variables consist of a soil moisture accounting store (S) and a quadratic reservoir (R). The model is forced by monthly rainfall ($\boldsymbol{r}$) and potential evapotranspiration ($\boldsymbol{e}$) and returns a monthly flow $\widehat{\boldsymbol{y}}$. Readers should note that θ_1, the soil moisture store capacity, controls the model's response to rain event, and to a certain extent, the variability of the simulated flow. As θ_1 increases, the simulated flow depends less on the current rainfall and more on the store level, itself dependent on past rainfall. For small θ_1, more rainfall is directed as excess rainfall and instantaneously routed as output flow. In other words, θ_1 controls the model's low-pass filter behavior (its "memory" of past events). In the following, the

GR2M model is denoted by $\widehat{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{\phi})$, with model parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, input series $\boldsymbol{x} = \{\boldsymbol{r}, \boldsymbol{e}\}$ and internal state variables $\boldsymbol{\phi} = \{R, S\}$.

Sensitivity analyzes have determined that GR2M is sensitive to white noise errors on rainfall, but comparatively robust to random errors on potential evapotranspiration (PE) [*Paturel et al.*, 1995]. That is, the Nash-Sutcliffe efficiency (NSE) decreases rapidly as the magnitude of random errors over rainfall increases, but much more slowly in the case of random errors over PE. Although the study of *Paturel et al.* [1995] dates back to a previous version of GR2M, a similar analysis was carried out and the same conclusions hold for the current version.

3.2. MOPEX data sets

The data used in this study is taken from the MOdel Parameter Estimation eXperiment (MOPEX) [*Schaake et al.*, 2006]. The database contains daily streamflow, rainfall, potential evapotranspiration, minimum and maximum temperatures. The rainfall data is produced by *Maurer et al.* [2002] and is the result of gridding point rainfall estimates from the United States, Canada and Mexico. The rainfall point estimates are daily totals taken from the National Oceanic and Atmospheric Administration (NOAA) Cooperative Observer (Co-op) stations. They are gridded using the synergraphic mapping system (SYMAP) [*Shepard*, 1984] as implemented by *Widmann and Bretherton* [2000]. The gridded data is then scaled to correct for complex topography in sparsly instrumented areas using long-term monthly averages from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) [*Daly et al.*, 1994, 1997]). The data is not corrected, however, for systematic gauge errors [*Maurer et al.*, 2002], which can significantly decrease the measured amount of rainfall.

Generally speaking, the major sources of systematic errors in rainfall measurements are wind-induced undercatch, wetting losses and evaporation losses. The undercatch due to wind has been estimated to vary from 2% to 20% for liquid precipitations to over 100% for solid precipitation, wetting losses from 3% up to 10% and evaporation losses are generally below 2% [*Metcalfe et al.*, 1997; *Yang et al.*, 1999]. These estimates are highly dependent on the type of gauge, the local climatology, altitude, topography and the apparatus calibration.

Another source of error is the "representativeness" of the data (commensurability). Lumped conceptual hydrological model assume the input variable is related to the areal rainfall over the entire catchment, whereas point rainfall estimates are often used [*Habib et al.*, 2001]. In the MOPEX database, the aerial adjustment is achieved by the spatial interpolation and catchment integration. However, it is difficult to discuss even roughly the relation between the MOPEX estimated rainfall and the "true" aerial rainfall, due to the complex manipulations separating the raw point data from the reanalysis results. Unless uncertainty estimates are provided, this is a setback to the use of reanalysis data for applications where uncertainty assessment is crucial.

As for evapotranspiration, the MOPEX values are derived from the NOAA Evaporation Atlas [*Farnsworth et al.*, 1982]. More specifically, they are computed by fitting a Fourier series with only an annual cycle to evaporation pan monthly averages. These averages were obtained by gridding the Evaporation Atlas maps with a resolution of 1/6 deg. The parameters of the series were then computed on this grid and averaged over each basin to estimate the basin average annual evapotranspiration cycle.

Since interannual random white noise PE errors do not seem to perturb hydrological models [*Andréassian et al.*, 2004; *Paturel et al.*, 1995], our concern is mostly with systematic or cyclic intraannual errors. One such cyclic source of error could be soil moisture availability. Indeed, PE strongly depends on the amount of moisture in the top soil layer, a dependence that is not taken into account by the MOPEX estimates. If the soil moisture has a strong cyclic component (high in spring, low in September for instance), the resulting PE may show an intraannual variability not captured by pan evaporation estimates, with potentially detrimental influence on model simulation.

Finally, MOPEX streamflow series are provided by the United States Geological Survey (USGS). Daily streamflows are taken at gauges unaffected by upstream regulation and where long time series are available. Stage measurements are converted into streamflow estimates through rating curves. Stage measurement errors are of the order of 3mm while single discharge measurement errors typically range around 5-10%, depending on the protocol (number of vertical velocity profiles, depth of measurements) and apparatus (velocimeter, Doppler profiler) [*Hirsch and Costa*, 2004]. Moreover, rating curves are not static and may vary due to changes in bed roughness, accumulation of debris following a flood, vegetation growth, presence of ice, stream bed scour and fill processes, bank erosion, etc [*Fenton and Keller*, 2001]. Some of these factors are accounted for by frequent updates to rating curves.

The simulations presented in section 4 are run for all eight stations presented in table 1 (although only results from the Chunky River watershed are shown). The basins chosen cover relatively dry to humid conditions found in the continental U.S.A. In the paper, units are systematically millimeters (mm) for rainfall (aerial) and monthly discharge (aerial).

4. Step by step Bayesian model calibration

Realistic calibration of a hydrological model in presence of data and model uncertainties using Bayesian analysis poses numerous conceptual and technical issues [*Kavetski et al.*, 2006b]. Among those, defining sensible error models and priors, integrating latent variables, validating the calibrated model and leveraging realism versus complexity are the most challenging. In a realistic calibration, many different issues are mingled and it is difficult to grasp how they interact and consequently, how to interpret the results. The objective of the current section is to isolate calibration issues in specific stand-alone examples, helpful in understanding the effects of calibration assumptions.

The first issue discussed is the initial state uncertainty and its relevance in different calibration contexts. Then the first case is presented, an application of Bayesian analysis where only output errors are considered, highlighting the conceptual difference between simulations and measurements. The second case shows a calibration where errors are assumed to be only on input data, and discusses how the input error assumption impacts the very definition of model performance. The third case introduces structural errors and discusses their lumping with output errors. From these three cases follows a brief analysis of the effect of error models assumptions on the posterior parameters distribution.

It should be noted that although the figures only show the case of the Chunky river watershed, identical simulations were run for all eight basins presented in table 1 with comparable results.

4.1. Initial state uncertainty

In GR2M, the internal state of the system is defined by the level of two reservoirs. Since they are conceptual reservoirs, their level cannot be measured experimentally and there

is, *a priori*, no way to define exactly their initial values. A convenient solution to this problem is to warm-up the model, *i.e.* to set the internal state to arbitrary values and let the model run for a fixed period [*Shelton*, 1985]. The rationale behind this strategy is that the effect of the initial state $\boldsymbol{\phi_0}$ on the simulated flow decreases rapidly. Indeed, at time t the internal state $\boldsymbol{\phi}_t$ depends on $\boldsymbol{\phi}_{t-1}$ and on x_t, the current input. By induction this is equivalent to saying it depends on $\boldsymbol{\phi_0}$, $x_1, x_2, \ldots, x_t$. Hence, if t is large enough, the effect of $\boldsymbol{\phi_0}$ becomes negligible compared with the effect of the input series $\boldsymbol{x}$. The question then is, how long should this warm-up period be to minimize the effect of the initial state uncertainty without sacrificing too much data ? For some catchments, rainfall data was logged long before flow measurements began, so the period where only rainfall data is available can be used as a warm-up period without concern about data expenditure. However, when this is not the case and only short rainfall and flow series are available, it may be worthwhile to minimize the warm-up period.

An estimate of a sensible warm-up period can be found by artificially perturbing the internal state and measuring how fast it converges back to its unperturbed state (figure 2). The rate of convergence depends closely on θ_1, the soil moisture store capacity; for large store capacity, the dependence of the model to its past state (its "memory") is stronger and internal state perturbations have a prolonged effect ($\sim$ 12 months). For small store capacities, the model has a short memory and perturbations vanish after as few as three months. Based on simulations on a number of catchments from the MOPEX database and different parameter sets, it was found that perturbations vanish with an average half-life of around 1.5 months. Warming-up the model during one year [*Paturel et al.*, 1995] hence leaves ample time to the internal state to reach its "steady" state. In the simulations

shown in cases 1,2 and 3, a warm-up period of two years was used since data availability was not an issue.

Using Bayesian analysis, the warm-up period can be avoided entirely. The solution is to define a prior for the initial state $\pi(\boldsymbol{\phi_0})$ and consider $\boldsymbol{\phi_0}$ a latent variable, as in equation (6). The effect of $\boldsymbol{\phi_0}$ on simulations can then be integrated out and parameters calibrated without sacrificing the first months to warm-up the model. Simulations (not presented here for brevity) suggest that the advantages of such an approach over the warm-up strategy depend on the relative information content [*Wagener et al.*, 2003] of the warm-up period. This information content (loosely defined as the diversity of hydrological conditions) is itself dependent on the length of the time series, the presence of extreme events and the prior for the initial conditions. For example, if a long calibration series is available, the relative information content of the first months is low compared to the rest of the series and using those first months in the calibration has only a weak effect on the parameter distribution. However, if those first months contain a very large flood, the like of which is not observed during the rest of the calibration series, then including them in the calibration series may have a significant effect on the posterior parameter distribution. Finally, if a vague prior for the initial state is defined, then the uncertainty over flow simulations for the first few months is dominated by the initial state uncertainty rather than the parameter uncertainty, and only a weak inference can be conducted on the model's parameters. Therefore, integrating the initial state uncertainty is advantageous only if a reasonably informative prior about the initial state can be defined.

4.2. Case 1: Assuming output errors only

In this first case, equation (6) is applied under the assumption that inputs are exact, or more pragmatically, that the effect of input errors can be safely lumped into output errors. This case is similar to a classical calibration, except that while a standard calibration aims at finding a unique set of parameters optimizing an objective function, a Bayesian analysis provides a parameter posterior probability distribution. Another difference is that objective functions can take any shape [*Gupta et al.*, 1998], whereas output error models are limited to formal statistical distributions (positive, piecewise continuous functions whose integral equals one). In this example, the calibration assumptions are given formally by:

$$\mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x}) = \delta(\tilde{\boldsymbol{x}} - \boldsymbol{x}) \tag{8a}$$

$$\mathrm{p}_{str}(\boldsymbol{y} \mid \widehat{\boldsymbol{y}}) = \delta(\boldsymbol{y} - \widehat{\boldsymbol{y}}) \tag{8b}$$

$$\mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) = \mathbf{L}(\tilde{\boldsymbol{y}} \mid \mu = \log \boldsymbol{y}, \sigma_{out}) \tag{8c}$$

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi_0}) = \mathbf{U}(\theta_1|0, 1000) \cdot \mathbf{U}(\theta_2|0, 2) \cdot \delta(\boldsymbol{\phi_0} - \widehat{\boldsymbol{\phi}}_{\mathbf{0}}) \tag{8d}$$

where δ stands for the Dirac delta, $\mathbf{L}$ for the lognormal probability density function (pdf, see eq. (A3)) and $\mathbf{U}$ for the uniform pdf. The hypothesis motivating the choice of a lognormal distribution to model output errors is that streamflow errors are roughly proportional to the value they affect: errors are multiplicative rather than additive [*Yapo et al.*, 1996] (see appendix A3 for details). This is a reasonable assumption if one considers that stage errors are transformed into discharge errors by a concave stage-discharge relationship (later explained in section 5.3). Note that the prior for the true input $\pi(\boldsymbol{x})$ is not defined here since it is just a constant (assumption (8a) implies that $\boldsymbol{x} = \tilde{\boldsymbol{x}}$). The

prior for the model parameters is given by a uniform prior over intervals covering a broad range of watershed behavior. Initial conditions $\boldsymbol{\phi_0}$ are fixed to $\widehat{\boldsymbol{\phi}}_{\mathbf{0}}$ by warming-up the model for a period of 24 months.

Plugging the error model assumptions (8a-d) into equation (6) and using the properties of the Dirac delta function to solve the three integrals yields a posterior distribution with a simple expression:

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \widehat{\boldsymbol{y}})\pi(\boldsymbol{\theta}), \tag{9}$$

where $\widehat{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{\theta}, \tilde{\boldsymbol{x}}, \widehat{\boldsymbol{\phi}}_{\mathbf{0}})$ and $\pi(\boldsymbol{\theta}) = \mathbf{U}(\theta_1|0, 1000) \cdot \mathbf{U}(\theta_2|0, 2)$. With all integrals solved, equation (9) can be computed directly and there is no explicit need to use MCMC sampling. Nevertheless for the sake of the demonstration, 100000 values of $\boldsymbol{\theta}$ are drawn from the posterior distribution (9), half of which is discarded to remove transient samples. The shape parameter of the output error model is set to $\sigma_{out} = .2$ (corresponding roughly to a 20% error on flow), and the calibration data $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ consists of 20 years of monthly rainfall, PE and flow from the Chunky River watershed (see table 1).

Figure 3 shows the results on a five years period. The lower plot shows the series of rainfall measurements from the MOPEX database. The middle plot shows the contoured probability density of the simulated flow along with the flow measurements from the MOPEX (full line). That is, the simulated flow $\widehat{\boldsymbol{y}}^i$ is computed for each $\boldsymbol{\theta}^i$ sampled, and a contour plot is drawn from the simulated flow histogram. Note that the variability of the simulated flow does not capture at all the variability seen in the observed flow. At first sight, it would appear that the parameter uncertainty is insufficient to account for the flow uncertainty. Although specifying a vaguer output error model (setting σ_{out} to some higher value) would increase the parameter uncertainty, it is not the real issue here and is

not advisable either [*Refsgaard et al.*, 2006]. Indeed, the difference in variability is mainly due to the fact that the plot compares two conceptually different variables: the measured flow $\tilde{\boldsymbol{y}}$ and the true flow $\boldsymbol{y}$ (equal in this case to the simulated flow $\widehat{\boldsymbol{y}}$). To be meaningful, the plot must compare measured flow and simulated *flow measurements*. These simulated flow measurements ($\tilde{\boldsymbol{Y}}$) are computed by drawing random variates from the output error model $\tilde{\boldsymbol{Y}}^i \sim \mathrm{p}_{out}(\tilde{\boldsymbol{Y}} \mid \widehat{\boldsymbol{y}}^i)$ with $\widehat{\boldsymbol{y}}^i$ computed for each $\boldsymbol{\theta}^i$ sampled from the posterior (9). The upper graph of figure 3 shows a contour plot of the probability density of simulated flow measurements compared with flow measurements (full line). It can be seen that the simulated measured flow captures better the variability of the observed flow series.

This conceptual difference between model simulation and measurements has more profound implications than capturing the simulation variability. Indeed, depending on the output error model chosen, it may lead to apparent paradoxes. For instance, if the output error model assumes that flow measurements systematically underestimate the true value, the flow simulated from the calibrated parameters will systematically overestimate the flow measurements. Therefore, a measure of calibration efficiency from a direct comparison of flow measurements and flow simulations would be meaningless and misleading. This conceptual difference between simulations and measurements has already been noted in Kalman filtering theory [*Burgers et al.*, 1998] and in the Bayesian recursive parameter calibration context [*Thiemann et al.*, 2001].

4.3. Case 2: Assuming input errors only

The second case deals with the opposite situation, namely that outputs are assumed exact and precipitations uncertain. Although this is a rather peculiar situation, it illustrates some potential benefits and setbacks associated with taking input errors into

account. Formally, the underlying hypotheses are given by:

$$\mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x}) = \mathbf{L}(\tilde{\boldsymbol{r}} \mid \mu = \log \boldsymbol{r}, \sigma_{in} = .2)\, \delta(\tilde{\boldsymbol{e}} - \boldsymbol{e}) \tag{10a}$$

$$\pi(\boldsymbol{x}) = \mathbf{W}(\boldsymbol{r} \mid \alpha, \kappa, \mu, \sigma) \tag{10b}$$

$$\mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) = \delta(\tilde{\boldsymbol{y}} - \boldsymbol{y}), \tag{10c}$$

where $\mathbf{W}$ stands for the exponentiated Weibull pdf (see eq. (A5)). The parameters of the exponentiated Weibull distribution $(\alpha, \kappa, \mu, \sigma)$ are estimated by fitting an historical rainfall series of 20 years, prior to the calibration period, from the same catchment the model is calibrated on. The exponentiated Weibull was chosen because it adequately captures the total monthly rainfall distribution over many catchments compared to other common semi-infinite distributions (see appendix A4). The structural error model and the prior for the parameters and internal state are identical to the last case (equations (8b) and (8d)).

To solve this case using MCMC sampling, the Dirac delta of the output error model is approximated by a lognormal distribution with a small shape parameter, that is, eq. (8c) with $\sigma_{out} \approx 0$. The integral over $\boldsymbol{x}$ must be performed in this case by Metropolis sampling of the following posterior

$$\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \widehat{\boldsymbol{y}})\, \mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, \pi(\boldsymbol{x}) \pi(\boldsymbol{\theta}). \tag{11}$$

The sampler now explores a space formed by the two model parameters $\boldsymbol{\theta}$ and the $20 \cdot 12 = 240$ true input values $\boldsymbol{x}$. However, as the shape parameter σ_{out} approaches zero, it becomes increasingly difficult to explore the $\boldsymbol{x}$-space because the values with a non-zero probability lie on an vanishingly small 240-dimensional volume, making random jumps very unlikely to be accepted by the Metropolis algorithm. With $\sigma_{out} = .02$, $3.0 \cdot 10^7$

samples had to be generated to obtain an approximately stable parameter distribution. That is, a change in the burn-in period has no effect on the mean of the distribution, but still has a small influence on its shape, even after ten million iterations. Ideally, after a given number of iterations, the shape of the distribution should show no dependence to the burn-in length.

The calibration results are presented in figure 4. The lower graph shows the probability density function (pdf) of the inferred true rainfall, while the upper graph shows the pdf of the simulated *measured rain*: true rain corrupted by random measurement errors drawn from the input error model. The simulated flow is not shown since it is, by definition, identical to the observed flow.

There are two observations that deserve to be highlighted. First, by assuming the measured flow is exact, the true rain $\boldsymbol{x}$ is constrained to values that reproduce exactly the measured output. In other words, equation (11) infers the value of the true rain, based on the knowledge of the measured flow. Although this may seem artificial, some interesting observations can be derived from such an analysis. Indeed, in a case not shown here, a clear annual trend in the ratio true rain/measured rain was noted; during the summer months, the true rain systematically underestimated the measured rain, suggesting that there was too much water input. And indeed, increasing the summer PE in those cases improved the overall model efficiency.

The second observation is related to the assessment of model's efficiency, often measured in hydrology by the Nash-Sutcliffe Efficiency criterion (NSE) (although *Schaefli and Gupta* [2007] point out that NSE does not provide a reliable basis for comparing results of different case studies). In this particular case, there is no difference between simulated

and measured flow, and the NSE reaches, by design, its maximum value, 100%. Obviously, any efficiency measure based on the concordance of measured and simulated output will yield exceedingly good, yet meaningless results. In this particular case, the efficiency of the model should rather be measured by the concordance between measured inputs and simulated inputs (or rather simulated *input measurements*) since this really is the quantity "predicted" by the model.

4.4. Case 3: Lumping output and structural errors

The last few years have seen an increasing number of strategies designed to account for structural errors in calibration or prediction [*Refsgaard et al.*, 2006], taking advantage of Bayesian model averaging [*Ajami et al.*, 2007], ensembles simulations [*Vrugt et al.*, 2005] or Gaussian processes [*Kennedy and O'Hagan*, 2001]. These methods capture the uncertainty related to the model either by running a number of different models (model averaging and ensemble simulations), or by describing the model error by a mixture of functions (Gaussian processes). The proposed framework is not a replacement for those methods, but rather a way to include formally a probabilistic description of structural errors into the calibration and prediction process.

Indeed, the common practice in hydrological modelling is to define an objective function which relates simulated values to observed ones, and assumes that uncertainties on input data are negligible. This objective function, even if it is not clearly stated, lumps together output and structural errors. This embedding of the structural error model into a "simple" objective function leads to inconsistent results (e.g. residuals do not comply with the assumed output distribution). In this example, a structural error model is formally lumped with an output error model, yielding a response error model to be incorporated to the

calibration process. Input errors are considered as well and the parameters posterior distribution is estimated.

Given output and structural error models, a response error model p_{res} lumping both output and structural errors can be defined simply as

$$\mathrm{p}_{res}(\tilde{\boldsymbol{y}} \mid \widehat{\boldsymbol{y}}) \equiv \int \mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})\, \mathrm{p}_{str}(\boldsymbol{y} \mid \widehat{\boldsymbol{y}})\, \mathrm{d}\boldsymbol{y}. \tag{12}$$

In some cases, this integral can be solved analytically, greatly reducing the numerical cost of sampling over the true output variables. Indeed, replacing the integration over $\boldsymbol{y}$ in the right hand side of equation (6) by equality (12) yields

$$\mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \int \mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, \mathrm{p}_{res}(\tilde{\boldsymbol{y}} \mid \widehat{\boldsymbol{y}})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \cdot \pi(\boldsymbol{\theta}), \tag{13}$$

which can be solved by sampling only the parameters $\boldsymbol{\theta}$ and true input values $\boldsymbol{x}$ from

$$\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x} \mid \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \mathrm{p}_{in}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\, \mathrm{p}_{res}(\tilde{\boldsymbol{y}} \mid \widehat{\boldsymbol{y}})\, \pi(\boldsymbol{x})\pi(\boldsymbol{\theta}) \tag{14}$$

and marginalizing over $\boldsymbol{x}$.

As an example, this third case assumes that both p_{out} and p_{str} are expressed in terms of the lognormal distributions:

$$\mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) = \mathbf{L}(\tilde{\boldsymbol{y}} \mid \ln \boldsymbol{y}, \sigma_{out}) \tag{15a}$$

$$\mathrm{p}_{str}(\boldsymbol{y} \mid \widehat{\boldsymbol{y}}) = \mathbf{L}(\boldsymbol{y} \mid \ln \widehat{\boldsymbol{y}}, \sigma_{str}). \tag{15b}$$

The lognormal distribution is chosen both for its simplicity and because it is heteroscedastic, a typical feature for hydrologic model residuals [*Xu*, 2001; *Yapo et al.*, 1996]. Inserting (15a,15b) into (12) and performing the integration yields

$$\mathrm{p}_{res}(\tilde{\boldsymbol{y}} \mid \widehat{\boldsymbol{y}}) = \mathbf{L}(\tilde{\boldsymbol{y}} \mid \ln \widehat{\boldsymbol{y}}, \sqrt{\sigma_{str}^2 + \sigma_{out}^2}), \tag{16}$$

a lognormal distribution lumping both output and structural errors. Again, the integration of the output and structural error models is by no means mandatory, and is done here out of optimization considerations. Given ample computing resources, the analytical integrability of the output and structural error models is not an issue.

Using the response error model (16) with $\sigma_{out} = .05$ and $\sigma_{str} = .1$, the input error model (10a) with $\sigma_{in} = .1$ and priors (10b, 8d), $1.0 \cdot 10^6$ samples $(\boldsymbol{\theta}, \boldsymbol{x})$ are drawn from equation (14) using MCMC sampling and the first half is discarded. The lower graph of figure 5 shows the simulated rain measurement probability density and the upper graph the simulated flow measurement probability density. Although this example has no pretense to realism, the overall calibration results appear satisfying, both for the inferred rain and flow. One thing deserving improvement though is the uncertainty over high flows, apparently overestimated by the response error model.

This example shows that from a conceptual point of view, taking structural uncertainty into account is relatively simple: define a structural error model and integrate, analytically or numerically, over the true output series $\boldsymbol{y}$. However, as Bayesians know too well, specifying a prior or an error model is not difficult, but specifying a good one is. The situation is further complicated since the proposed definition of structural errors $(\boldsymbol{y} - \widehat{\boldsymbol{y}})$ combines two types of structural errors: model inadequacies and the inherent stochasticity of the process (due to explaining variables not taken into account). Theoretically, with a long "error free" data set, it would possible to identify both types of structural errors, rectify the model to remove the inadequacies and describe probabilistically the remaining randomness of the process. Without such a perfect data set however, the expertise of hydrologists seems the best way to define sensible structural error models.

To validate the model and error models, the procedure detailed in section 2.4 is applied and the results shown in figure 6. Although the error models are rather crude, the resulting uncertainty for the flow (contour plot) appears consistent with the flow observations (full line).

4.5. Impact of error models on posterior parameter distribution

Values of calibrated parameters and their associated uncertainties are conditioned by the choice of the errors models. To illustrate this, the parameters calibrated from the three cases discussed above (with identical calibration data sets) are shown in figure 7. The sensitivity of parameters on error model assumptions can be explained by noting that parameters are estimated through the inferred true input and output values, which depend directly on the error models. In other words, the parameter distribution is inseparable from the choice of error models, and calibration results can only be understood by looking at the entire set of assumptions.

A feature of figure 7 deserving to be highlighted is that lumping all sources of errors on the output (Case 1) yields a larger value for θ_1 (the moisture store capacity) than when input errors are accounted for (Case 2 and 3). By increasing this parameter, the model reduces the potential for excess rainfall and, at the same time, the flow variability. In other words, calibration selects parameters that enhance the low-pass filter behavior of the model and reduce its sensitivity to input error.

Finally, readers may wonder about the significant difference between the parameters estimated with NSE and those obtained from case 1, since both methods assume that only the outputs are corrupted. The authors explanation for this difference is that it is largely due to the difference in heteroscedasticity assumed by the two methods. Gen-

erally speaking, a heteroscedastic error model tolerates large errors on large flows and estimates parameters that simulate well the low flows. On the other side, a homoscedastic error model considers all errors equal and estimates parameters that somehow capture the average behavior. For GR2M, this average behavior is obtained by a higher exchange coefficient θ_2 (less water is lost to neighboring catchments, increasing the base flow) and a higher soil moisture capacity θ_1, reducing the model's sensitivity to rainfall. In figure 7, the NSE parameters are estimated by maximizing the NSE criterion computed on the flows square root [*Xu and Vandewiele*, 1994], which is akin to a heteroscedasticity assumption [*Sorooshian and Dracup*, 1980]. However, this square root heteroscedasticity is weaker than the lognormal heteroscedasticity assumed by the output error model (Case 1), explaining why the NSE parameters are shifted to the upper right. If the parameters are estimated again by an NSE criterion computed this time directly on the flows (homoscedasticity assumption), one obtains an even greater shift with $\hat{\boldsymbol{\theta}} = (453, 0.88)$. This artificial favoritism of a low-pass filter behavior may have serious impacts if the model is used in a predictive mode as it can lead to underestimation of peak flows, as well as the associated uncertainties on these estimates.

One may suggest that shifts in parameters could possibly be traced back to distinct model failures. For instance, the large difference in the soil moisture capacity could be explained by the lack of an interception process in the model [*Savenije*, 2004] or an incorrect direct runoff parameterization. Although the compensation mechanisms of parameters for un- or mis-accounted sources of errors are interesting in their own right, it is not clear at this point whether such shifts can provide more than cursory evidence

to identify structural inadequacies. Moreover, for hydrological models counting dozens of parameters, such an analysis would lose much of its intuitiveness.

5. Case study

The objective of this section is to demonstrate the use of the Bayesian uncertainty framework under more realistic calibration assumptions. For each source of error, hypotheses are formulated in probabilistic terms and error models constructed on the basis of these hypotheses. Although the GR2M model may be too rudimentary for this application to be really realistic, it shows what can be done in terms of error models, and serves as a test case before tackling more complex daily models. In the following sections, errors models and priors are defined and the simulation setup is presented. Results are then discussed, along with the necessary modifications for an extension to daily models.

5.1. Rainfall error model

The rainfall error model is inspired by *Weerts and El Serafy* [2006] who describe hourly rainfall errors by a normal distribution with a variance $\sigma_{\delta r}^2$ that has both a proportional component (15%) and fixed one (.2 mm):

$$\sigma_{\delta r} = .15r + 0.2\text{mm},$$

where r stands for rainfall (mm). The fixed component is due to the finite resolution of rain gauges (for example the volume of the tipping bucket). For this application, the fixed component is set at .254 mm (.01 in), the default volume for tipping bucket gauges [*Habib et al.*, 2001]. It is assumed that this error affects hourly measurements, so that once scaled to monthly values, the standard variation due to the finite resolution of the gauge is estimated by $\sqrt{30 \cdot 24} \cdot .254\text{mm} \approx 7\text{mm}$. The proportional component roughly

represents commensurability (aerial representativity) errors and measurements errors such as evaporation and wetting losses as well as wind-induced undercatch. Once aggregated to monthly values, the proportional component reduces to about 7% (see appendix B). Since these measurement errors cause an underestimation of the true rainfall, the rainfall input error model is assumed to be also biased by around 7% [*Metcalfe et al.*, 1997]; that is, it is assumed that rainfall measurements are on average 93% of the true rainfall.

One should note that the smallest error standard deviation (7 mm) is of the order of the smallest monthly precipitation. Hence, there is a normalization issue due to the fact that for small rainfalls, the normal distribution assigns non zero probability to negative rainfall. To deal with this problem, a truncated normal distribution $\mathbf{N}_T$ (see eq. (A1)) is used instead of a standard normal distribution. Formally, the input error model is defined as :

$$\mathrm{p}_{in}(\tilde{\boldsymbol{r}} \mid \boldsymbol{r}) = \mathbf{N}_T(\tilde{\boldsymbol{r}} \mid .93\boldsymbol{r}, .07\boldsymbol{r} + 7\mathrm{mm}, 0, \infty). \tag{17}$$

5.2. PE error model

The error model for PE is based primarily on the assumption that white noise error on PE have only a weak influence on model performance [*Paturel et al.*, 1995]. It is also based on the authors observation that PE computations performed with the McGuinness-Bordne model [*McGuinness and Bordne*, 1972] yields values similar within a multiplicative factor to MOPEX PE series (results not shown here for the sake of brevity). In other words, the relevant uncertainty seems to be related mostly to intrannual rather than interannual variability. Therefore, it is assumed here that the true PE ($\boldsymbol{e}$) is obtained by multiplying the MOPEX PE series ($\tilde{\boldsymbol{e}}$) by a unique multiplier (ϵ): $\boldsymbol{e} = \epsilon\tilde{\boldsymbol{e}}$, thus rescaling the PE series. This hypothesis enormously simplifies computations as it replaces an integration over N

true PE values by an integration over only one multiplier ϵ. Since the multiplier along with measured PE completely determine the conceptual true PE, there is no need for a genuine PE error model. However, a prior for the multiplier still has to be defined:

$$\pi(\epsilon) = \mathbf{L}(\epsilon \mid 0, .5). \tag{18}$$

The lognormal is chosen because it is the most general hypothesis in terms of entropy (see appendix A3) that can be made for a positive variable of fixed mean and variance. The hyperparameters are set such that the multiplicative factor has a median of 1 and a standard variation of about .5.

5.3. Output error model

The output error model describes the probability of an error on the flow knowing the conceptual true flow. Flows $\tilde{y}$ are commonly obtained from stage measurements $\tilde{h}$ and a rating curve (stage-discharge) relationship $\tilde{y} = f(\tilde{h})$, defined from previous simultaneous stage and discharge measurements. The analysis of discharge errors has received substantial attention [*Shiklomanov et al.*, 2006; *Schmidt*, 2002; *Pelletier*, 1988; *Herschy*, 1985; *Dymond and Christian*, 1982] and multiple methods have been proposed to define the nature and structure of discharge errors. In the following, a simplified approach is chosen in order to reduce the complexity of the output error model.

Discharge errors originate from many different sources: stage measurement errors, single discharge measurements, rating curve sampling errors (a limited amount of data is available to fit the curve) and the effect of explaining variables other than stage (river slope, bed roughness, presence of ice, vegetation growth, etc). Although it is possible to describe different source of errors independently [*Dymond and Christian*, 1982], it requires

a considerable amount of information about single discharge measurements (number and type of velocity vertical profiles, apparatus, etc) and may be too complex in the case of aggregated monthly discharges. To reduce the complexity of the error model, it is assumed that all sources of errors can be lumped into a single term: a stage measurement error δ_h. Stage errors can then be converted into discharge errors, $\delta_y \approx f'\delta_h$, where $f(h)$ is the function relating stage h to discharge y. Visually, this error structure seems to capture the variability of stage-discharge measurements (see figure 8).

The rating curve f is defined using stage-discharge data, courtesy of the USGS (`waterdata.usgs.gov/nwis`). Following the suggestion of *Fenton and Keller* [2001], it is fitted by a polynomial computed over the square root of the flow. In many instance, a single curve does not capture reliably the behavior over the entire range, so piecewise polynomials are used. In the particular case of the Chunky river shown in figure 8, two piecewise polynomials of order 3 were used:

$$y = f(h) \equiv \begin{cases} \sum_{i=0}^{3} a_i \sqrt{h^i} & h \leq h_k \\ \sum_{i=0}^{3} b_i \sqrt{h^i} & h > h_k \end{cases}, \tag{19}$$

where h_k is the position of the kink separating both polynomials.

Formally, the output error model is based on the assumption that daily discharge errors are normally distributed with a variance dependent on stage:

$$\sigma_{y_d} = (10\text{mm} + .02h)f'(h). \tag{20}$$

Although the numerical parameters of 10mm and 2% were chosen empirically to capture the variability observed in the data (see figure 8), they agree reasonably well with physically realistic values. Indeed, the accuracy of a stage mesurement has been estimated at 9mm [*Schmidt and Garcia*, 2003] (although it is more commonly estimated at 3 mm

[*Hirsch and Costa*, 2004]). The accuracy of single discharges is known to vary between 5% and 10% [*Hirsch and Costa*, 2004], which correspond in the case of the Chunky river to stage accuracies between 1.2% and 2.4%.

To link the daily discharge error model with GR2M monthly flows, the total monthly discharge is converted into an average daily discharge. Assuming that daily errors can be approximated by a first-order autoregressive process with a fixed variance $\sigma_{y_d}^2$ and correlation between successive measurement given by ϕ, then the variance of the mean over n samples is given by

$$\sigma_y^2 = \frac{\sigma_{y_d}^2}{n}\frac{1+\phi}{1-\phi}\left(1 - \frac{2\phi(1-\phi^n)}{n(1-\phi^2)}\right). \tag{21}$$

Setting $n = 30$ and $\phi = .8$ in equation (21) yields a standard variation of $\sigma_y \approx .5\sigma_{y_d}$ for the monthly average daily streamflow error, with σ_{y_d} given by equation (20). Note that since the daily variance σ_{y_d} is not constant but depends on the stage h, equation (21) constitutes an approximation. However, the intramonthly variability of the stage is sufficiently low for this approximation to hold (see appendix B). The value of $\phi = .8$ is chosen to reflect the belief that daily streamflow errors are moderately correlated over a span of about one week.

Summing up, the output error model is defined by a normal distribution centered on the true discharge with a variance depending on the stage associated to the mean daily discharge and the derivative of the rating curve at that point:

$$\sigma_y = (5\text{mm} + .01h)f'(h). \tag{22}$$

5.4. Structural error model

Model residuals are often found to be heteroscedastic, auto-correlated and non-normal [*Yang et al.*, 2007; *Xu*, 2001]. Although it is custom to use Box Cox or similar transformations to stabilize the variance and reduce heteroscedasticity, using transformed variables may have undesirable consequences on model calibration [*Schaefli et al.*, 2007]. Hence, we follow *Beven and Freer* [2001] suggestion that residuals could be described by multiplicative autoregressive process and define the structural error model by a multivariate lognormal distribution, with covariance matrix Σ chosen to reproduce an autoregressive process of order one [*Sorooshian and Dracup*, 1980]. The log-likelihood of such a distribution is given by

$$L(\boldsymbol{y} \mid \boldsymbol{\mu} = \widehat{\boldsymbol{y}}, \rho = .6, \sigma = .15\text{mm})$$

with L defined by equation (A4) in appendix A3. The hyperparameters of the structural error model correspond roughly to a relative error of 15% and a moderate auto-correlation (for daily models, the auto-correlation is sometimes found to be around .9 [*Høybye and Rosbjerg*, 1999]).

5.5. Priors

The prior for the initial conditions is defined by fitting a bivariate normal distribution (see equation A2) to a series of "historical" internal states. This series of internal state is computed by running GR2M over 20 years of data prior to the calibration period using different sets of parameters. These parameter sets are chosen randomly around the classical estimate (minimizing NSE on the square root of discharges). The mean $\boldsymbol{\mu}_{\boldsymbol{\phi}}$ and covariance matrix $\Sigma_{\boldsymbol{\phi}}$ of the internal variables series are then computed and used as hyperparameters for the bivariate normal, thus defining the prior for the initial model

state:

$$\pi(\boldsymbol{\phi_0}) = \mathbf{N}(\boldsymbol{\phi_0} \mid \boldsymbol{\mu_\phi}, \Sigma_\phi). \tag{23}$$

The prior for the parameter is chosen to be a uniform distribution and is defined by

$$\pi(\boldsymbol{\theta}) = \mathbf{U}(\theta_1|0, 1000) \cdot \mathbf{U}(\theta_2|0, 2). \tag{24}$$

One reason for this choice is simplicity; with 20 years of calibration data, the prior for the parameter is not expected to play a significant role. The other motivation for a uniform distribution is to enable a fair comparison between the SLS and the Bayesian results.

The prior for the true rainfall $\pi(\boldsymbol{r})$ is defined as in eq. (10b), that is by an exponentiated Weibull distribution whose parameters are estimated by maximizing the likelihood of an "historical" precipitations series spanning 20 years prior to the calibration period (see section 4.3). The importance of this prior must not be underestimated, as it plays an important role in the inference process [*Huard and Mailhot*, 2006]. One way to improve the accuracy of this prior could be to define seasonal priors, *i.e.* different prior for each season or month when strong climatic seasonalities are expected.

5.6. Simulation setup

The posterior distribution is now given by:

$$\begin{aligned} \mathrm{p}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{r}}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{y}}) &= \iiint \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{y}, \epsilon, \boldsymbol{\phi_0} \mid \tilde{\boldsymbol{r}}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{y}}) \, \mathrm{d}\boldsymbol{r} \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\epsilon \, \mathrm{d}\boldsymbol{\phi_0} \\ &= \iiint \mathrm{p}_{in}(\tilde{\boldsymbol{r}} \mid \boldsymbol{r}) \mathrm{p}_{out}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) \mathrm{p}_{str}(\boldsymbol{y} \mid \widehat{\boldsymbol{y}}) \pi(\boldsymbol{\theta}) \pi(\boldsymbol{\phi_0}) \pi(\boldsymbol{r}) \pi(\epsilon) \, \mathrm{d}\boldsymbol{r} \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\epsilon \, \mathrm{d}\boldsymbol{\phi_0} \end{aligned} \tag{25}$$

derived under hypotheses similar to those presented in section 2.2. The integrand on the right hand side of equation (25) is again sampled using Markov Chain Metropolis

using 2.0×10^6 iterations and discarding the first half. Validation is performed using the procedure described in section 2.4 with the same error models used in calibration.

5.7. Results and analysis

The results of the calibration and validation of the Chunky watershed are now presented. Figure 9 shows contour plots of the posterior pdf for simulated flow, true flow and simulated flow measurements, along with the inferred true rain and simulated rain measurements (the plots show the last five of the twenty years series used for calibration to avoid cluttering the graph). Note that while the true flow and simulated flow measurements are very similar, the distribution of the simulated rain measurements shows more dispersion than the pdf for the true rain. Also, the agreement between true flow and observed flow (full line) appears very good. Both observations are direct consequences of the fact that the error models assume small flow errors and large rainfall errors. As discussed in section 4.3, the true rainfall is strongly conditioned on the observed flow, and model performance should not be judged on the agreement between output observations and simulations.

A clearer picture of the different errors is presented in figure 10, showing the posterior pdf for the rainfall errors ($\tilde{\boldsymbol{r}} - \boldsymbol{r}$), structural errors ($\boldsymbol{y} - \widehat{\boldsymbol{y}}$) and output errors ($\tilde{\boldsymbol{y}} - \boldsymbol{y}$). As discussed earlier, the output errors are the smallest since the output error model assumes relatively small measurement errors. Rainfall errors are almost always negative, reflecting the assumption that rainfall measurements underestimate the true value. Structural errors, on the other hand, are mostly positive, a feature that is not imposed by the structural error model. One possible explanation could be the underestimation of overland flow by the model, reducing the amount of instant runoff during particularly rainy periods.

This separation of errors into distinct components, especially the identification of structural errors, is expected to have useful applications both as a diagnostic tool to identify model deficiencies and as a measure of model performance. Indeed, model performance in calibration could be estimated, irrespective of data error, by measuring the dispersion and autocorrelation of the posterior structural error distribution. In terms of diagnosis, analyzing dependencies between structural errors and environmental factors might also reveal variables that are relevant to the modeling process but unaccounted for by the model. One should be careful, however, with conclusions drawn from inferred error series. Indeed, with inferred error series as long as data series, the influence of error model assumptions (statistical distribution, hyperparameters) on inferred errors is bound to be significant. Again, this calls for further research on the definition of accurate error models as well as their validation.

Figure 11 shows the probability distribution for the model parameters, compared to the parameters estimated by minimizing the NSE criteria computed on the flows square root (a variation of the standard least-squares (SLS)). When superposed with the parameter distributions of figure 7, the mode of the sampled parameters fall in between those of cases 1 and 3, but the dispersion is larger.

Figure 12 shows the model simulations during the validation period compared with the SLS estimate (dashed line). Although there is not much difference between the SLS prediction and the modes of the Bayesian distributions, the uncertainty assessment appears reasonable in the sense that observed flows fall within the high density regions of the pdf. It shall be enlightening to pursue this type of investigation for daily or hourly models to

see whether larger differences in prediction accuracy exist between standard optimization and a Bayesian analysis.

5.8. Extension to daily models

There is a certain contradiction of purpose in applying an elaborate uncertainty assessment method to a simple two parameter monthly model. On the one hand, this choice makes the application of the method and interpretation of the results certainly simpler than with complex daily models. On the other hand, the added value of uncertainty assessment is lower because simple monthly models are expected to be less sensitive to input errors than their daily or hourly counterpart. To apply the method to daily models, however, a number of issues have to be addressed.

The prior for the true input, for instance, has to deal gracefully with null rainfall. One possibility would be to use statistical distributions that have both continuous and discrete components. The Tweedy distribution is one of those, designed to yield a finite probability to null values and a probability density on the positive domain [*Dunn and Smyth*, 2005]. Another issue that will have to be dealt with is the proper description of autocorrelations in the rainfall, flow and structural error series. Instead of being a liability, the dependence between errors is an asset for a modeler because it reduces the effective number of degrees of freedom in the calibration problem. Hence, from an inferential point of view, error autocorrelation is a useful feature and should be considered more carefully. Whatsmore, reducing the effective number of degrees of freedom might help overcoming the main practical challenge, namely the increasing dimensionality of the problem. Indeed, as the number of dimensions increases, so do the number of samples necessary for the posterior distributions to reach convergence. The approach taken in this paper is adequate for

monthly series, but it is not yet clear how well it will scale for daily series, two order of magnitudes longer.

6. Conclusion

In calibrating hydrological models to empirical data, hydrologists face the challenge of identifying model parameters that reasonably reproduce a watershed's behavior. As this behavior is quantified by a limited set of field observations corrupted by errors, hydrologists run the risk of fitting the model to the noise instead of the real process. Accounting for this noise, however, requires assumptions about its nature, that is, a plausible description of input and output errors. Moreover, the model itself does not emulate all the processes that actually occur in reality, introducing discrepancies between input and output that also need to be described. In this paper, Bayesian analysis is used to merge hypotheses about input, output and model errors into the calibration and validation of a model. Using the parsimonious monthly hydrological model GR2M, simulations are run to answer a number of questions about the assessment of multiple sources of uncertainties.

The first question concerned the effect of the initial state uncertainty on calibration efficiency. Simulations show that for GR2M, the initial state uncertainty lasts only a few months, after which the internal state reaches a "steady state". Hence, for long time series, accounting for initial state uncertainty has only marginal effects. Input uncertainty, however, plays a far more important role. In fact, when input errors are considered large and output errors small, model calibration proceeds in "reverse", conditioning the input on the output. This results in simulations reproducing almost perfectly the output observations, rendering traditional efficiency measures (NSE) meaningless. Moreover, calibrated parameters are completely different from those obtained under an "output error

only" hypothesis. This emphasizes how important it is to analyze parameter distributions only in the light of the error model assumptions.

In a calibration where multiple sources of uncertainties are accounted for, validation plays the sensitive role of assessing the hydrological model efficiency as well as the accuracy and validity of the error model assumptions. With so many factors influencing predictions, it is difficult to identify possible causes of failure just from the comparison between output observations and simulations. By looking directly at the inferred pdf for the input, output and structural error series, however, model failures of error model inadequacies are easier to detect and diagnosis easier to perform. Moreover, analyzing structural error series might provide the means to evaluate model performance, without interference from data quality issues.

The obvious interest in accounting explicitly for different sources of uncertainty affecting hydrological modeling must be weighted against the interest of stakeholders in realistic uncertainty assessment and the efforts required to specify accurate error models. Data model errors should reflect as accurately as possible the errors introduced in the acquisition of data. Ideally, those error models would evolve with changes in the gauge station network (number, location and type of stations) and changes in rating curves following new discharge measurements. Structural error models appear yet more difficult to define rigorously, and progress might only be made through the prior/posterior Bayesian learning process. Hopefully, the knowledge gained from the next experiments in model calibration and the rapid increase in computing power will eventually make uncertainty assessment a routine exercise and an integral part of hydrological modeling.

References

Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resources Research*, *43*, W01403, doi:10.1029/2005WR004745.

Andréassian, V., C. Perrin, C. Michel, I. Usart-Sanchez, and J. Lavabre (2001), Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models, *Journal of Hydrology*, *250*(1-4), 206–223, doi:10.1016/S0022-1694(01)00437-1.

Andréassian, V., C. Perrin, and C. Michel (2004), Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models, *Journal of Hydrology*, *286*, 19–35, doi:10.1016/j.jhydrol.2003.09.030.

Bates, B. C., and E. P. Campbell (2001), A Markov chain Monte-Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resources Research*, *37*(4), 937–947.

Beven, K. (2006a), A manifesto for the equifinality thesis, *Journal of Hydrology*, *320*(1-2), 18–36, doi:10.1016/j.jhydrol.2005.07.007.

Beven, K. (2006b), On undermining the science?, *Hydrological Processes*, *20*, 3141–3146, doi:10.1002/hyp.6396.

Beven, K., and A. Binley (1992), The future of distributed models : Model calibration and uncertainty prediction, *Hydrological Processes*, *6*, 279–298.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, *249*, 11–29, doi:10.1016/S0022-1694(01)00421-8.

Beven, K., P. Smith, and J. Freer (2007), Comment on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology" by Pietro Mantovan and Ezio Todini, *Journal of Hydrology, 338*(3-4), 315–318, doi: 10.1016/j.jhydrol.2007.02.023.

Burgers, G., P. J. van Leeuwen, and G. Evensen (1998), Analysis scheme in the ensemble Kalman filter, *Monthly Weather Review, 126*, 1719–1724.

Chib, S., and E. Greenberg (1995), Understanding the Metropolis-Hastings algorithm, *The American Statistician, 49*(4), 327–335.

Daly, C., R. P. Neilson, and D. L. Phillips (1994), A statistical—topographic model for mapping climatological precipitation over mountainous terrain, *Journal of Applied Meteorology, 33*, 140–158.

Daly, C., G. H. Taylor, and W. P. Gibson (1997), The PRISM approach to mapping precipitation and temperature, in *10th Conference on Applied Climatology*, pp. 10–12, American Meteorological Society, Reno, NV.

Dunn, P. K., and G. K. Smyth (2005), Series evaluation of Tweedie exponential dispersion model densities, *Statistics and Computing, 15*, 267–280, doi:10.1007/s11222-005-4070-y.

Dymond, J. R., and R. Christian (1982), Accuracy of discharge determined from a rating curve, *Hydrological Sciences Journal, 4*(12), 493–504.

Edijatno, and C. Michel (1989), Un modèle pluie-débit journalier à trois paramètres, *La Houille Blanche, 2*, 113–122.

Farnsworth, R., E. S. Thompson, and E. Peck (1982), Evaporation atlas for the contiguous 48 United States, *Technical Report NWS 33*, NOAA, Washington, DC.

Fenton, J. D., and R. J. Keller (2001), The calculation of streamflow from measurements of stage, *Tech. rep.*, Cooperative research center for catchment hydrology.

Fiering, M. B. (1967), *Streamflow Synthesis*, Harvard Univ. Press, Cambridge, Mass.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, Texts in Statistical Science, Chapman & Hall/CRC Press, London.

Goodman, J. (1987), A comment on the Maximum Entropy principle, *Risk Analysis*, *7*(2), 269–272.

Gull, S. F. (1989), Bayesian Data Analysis : Straight-line fitting, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, pp. 511–518, Kluwer Academic Publishers.

Gupta, H. V., S. Sorooshian, and P. Yapo (1998), Towards improved calibration of hydrologic models : Multiple and non-commensurable measures of information, *Water Resources Research*, *34*(4), 751–764.

Habib, E., W. F. Krajewski, and A. Kruger (2001), Sampling errors of tipping-bucket rain gauge measurements, *Journal of Hydrologic Engineering*, *6*(2), 159–166.

Hall, J., E. O'Connell, and J. Ewen (2007), On not undermining the science: coherence, validation and expertise. Discussion of Invited Commentary by Keith Beven. Hydrological Processes 20, 3141-3146 (2006), *Hydrological Processes*, *21*(7), 985 – 988, doi: 10.1002/hyp.6639.

Herschy, R. W. (1985), *Streamflow measurement*, Elsevier.

Hirsch, R. M., and J. E. Costa (2004), U.S. stream flow measurement and data dissemination improve, *Eos*, *85*(21), 197,203.

Høybye, J., and D. Rosbjerg (1999), Effect of input and parameter uncertainties in rainfall-runoff simulations, *Journal of Hydrologic Engineering*, *4*(3), 214–224.

Huang, M., and X. Liang (2006), On the assessment of the impact of reducing parameters and identification of parameter uncertainties for a hydrologic model with applications to ungauged basins, *Journal of Hydrology*, *320*(1-2), 37–61, doi: 10.1016/j.jhydrol.2005.07.010.

Huard, D., and A. Mailhot (2006), A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc", *Water Resources Research*, *42*, W07416, doi:10.1029/2005WR004661.

Jaynes, E. T., and G. L. Bretthorst (2003), *Probability theory : The logic of science*, Cambridge University Press.

Kabouya, M. (1990), Modélisation pluie-débit au pas de temps mensuel et annuel en Algérie Septentrionale, Ph.D. thesis, Orsay-Paris Sud.

Kavetski, D., S. W. Franks, and G. Kuczera (2003), Confronting input uncertainty in environmental modeling, in *Calibration of Watershed Models, Water Science and Applications*, vol. 6, edited by Q. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau, and R. Turcotte, pp. 49–68, AGU Water Science and Applications Series.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resources Research*, *42*, W03407, doi:10.1029/2005WR004368.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resources Research*, *42*, W03408, doi:10.1029/2005WR004376.

Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *63*, 425–464.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm, *Journal of Hydrology*, *211*(1-4), 69–85, doi:10.1016/S0022-1694(98)00198-X.

Limpert, E., W. A. Stahel, and M. Abbt (2001), Log-normal distributions across the sciences: Keys and clues, *BioScience*, *51*(5), 341–352.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, *330*, 368–381, doi: 10.1016/j.jhydrol.2006.04.046.

Maurer, E. P., A. W. Wood, J. C. Adam, and D. P. Lettenmaier (2002), A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States, *American Meteorological Society*, *15*(22), 3237–3251.

McGuinness, J., and E. Bordne (1972), A comparison of lysimeter-derived potential evapotranspiration with computed values, *Technical Bulletin 1452*, Department of Agriculture.

Metcalfe, R., B. Routledge, and K. Devine (1997), Rainfall measurement in Canada: Changing observational methods and archive adjustment procedures, *Journal of Climate*, *10*, 92–101.

Michaud, J. D., and S. Sorooshian (1994), Effect of rainfall-sampling errors on simulations of desert flash floods, *Water Resources Research*, *30*(10), 2765–2776.

Montanari, A. (2007), What do we mean by 'uncertainty' ? The need for a consistent wording about uncertainty assessment in hydrology, *Hydrological Processes*, *21*, 841–845, doi:10.1002/hyp.6623.

Moradkhani, H., K.-L. Hsu, H. Gupta, and S. Sorooshian (2005), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resources Research*, *41*(5), W05012, doi:10.1029/2004WR003604.

Mouelhi, S., C. Michel, C. Perrin, and V. Andréassian (2006), Stepwise development of a two-parameter monthly water balance model, *Journal of Hydrology*, *318*, 200–214, doi:10.1016/j.jhydrol.2005.06.014.

Mudholkar, G., and A. Hutson (1996), The exponentiated Weibull family: Some properties and a flood data application, *Communications in Statistics: Theory and Methods*, *25*(12), 3059–3083.

Nandakumar, N., and R. G. Mein (1997), Uncertainty in rainfall-runoff model simulations and the implications for predicting the hydrologic effects of land-use change, *Journal of Hydrology*, *192*, 211–232, doi:10.1016/S0022-1694(96)03106-X.

Neal, R. M. (1993), Probabilistic inference using Markov Chain Monte Carlo methods, *Tech. rep.*, Department of Computer Science, University of Toronto.

Novak, E., and K. Ritter (1997), The curse of dimension and a universal method for numerical integration, in *Multivariate Approximation and Splines, ISNM*, edited by G. Nürnberger, J. W. Schmidt, and G. Walz, pp. 177–188, Birkhäuser.

Oudin, L., C. Perrin, T. Mathevet, V. Andréassian, and C. Michel (2006), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *Journal of Hydrology*, *320*, 62–83, doi:10.1016/j.jhydrol.2005.07.016.

Paturel, J., E. Servat, and A. Vassiliadis (1995), Sensitivity of conceptual rainfall-runoff algorithms to errors in input data – case of the GR2M model, *Journal of Hydrology*, *168*, 111–125, doi:10.1016/0022-1694(94)02654-T.

Pelletier, P. M. (1988), Uncertainties in the single determination of river discharge: A literature review, *Canadian Journal of Civil Engineering*, *15*, 834–850.

Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Advances in Water Resources*, *29*(11), 1586–1597, doi:10.1016/j.advwatres.2005.11.013.

Savenije, H. H. G. (2004), The importance of interception and why we should delete the term evapotranspiration from our vocabulary, *Hydrological Processes*, *18*(8), 1507–1511, doi:10.1002/hyp.5563.

Schaake, J., Q. Duan, V. Andréassian, S. Franks, A. Hall, and G. Leavesley (2006), The model parameter estimation experiment (MOPEX), *Journal of Hydrology*, *320*(1-2), 1–2, doi:10.1016/j.jhydrol.2005.07.054.

Schaefli, B., and H. V. Gupta (2007), Do Nash values have value?, *Hydrological Processes*, *21*(15), 2075–2080, doi:10.1002/hyp.6825.

Schaefli, B., D. B. Talamba, and A. Musy (2007), Quantifying hydrological modeling errors through a mixture of normal distributions, *Journal of Hydrology*, *332*(3-4), 303–315, doi:10.1016/j.jhydrol.2006.07.005.

Schmidt, A. R. (2002), Analysis of stage-discharge relations for open-channel flows and their associated uncertainties, Ph.D. thesis, University of Illinois.

Schmidt, A. R., and M. H. Garcia (2003), Theoretical examination of historical shifts and adjustments to stage-discharge rating curves, in *World Water and Environmental Resources Congress 2003 and Related Symposia 2003*, edited by P. Bizier and P. DeBarry, American Society of Civil Engineers.

Sharma, T. C. (1996), A Markov-Weibull rain-sum model for designing rain water catchment systems, *Water Resources Management*, *10*(2), 147–162.

Shelton, M. L. (1985), Modeling hydroclimatic processes in large watersheds, *Annals of the Association of American Geographers*, *75*(2), 185–202.

Shepard, D. S. (1984), *Spatial Statistics and Models*, chap. Computer mapping: the SYMAP interpolation algorithm, pp. 133–145, D. Reidel.

Shiklomanov, A. I., T. I. Yakovleva, R. B. Lammers, I. P. Karasev, C. J. Vorosmarty, and E. Linder (2006), Cold region river discharge uncertainty–estimates from large Russian rivers, *Journal of Hydrology*, *326*(1-4), 231–256, doi:10.1016/j.jhydrol.2005.10.037.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resources Research*, *16*(2), 430–442.

Thiemann, M., M. Trosset, H. V. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resources Research*, *37*(10), 2521–2535.

Todini, E., and P. Mantovan (2007), Comment on: 'On undermining the science?' by Keith Beven, *Hydrological Processes*, *21*, 1633–1638, doi:10.1002/hyp.6670.

Troutman, B. M. (1982), An analysis of input errors in precipitation-runoff models using regression with errors in the independent variables, *Water Resources Research*, *18*(4), 947–964.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resources Research*, *41*, W01017,

doi:10.1029/2004WR003059.

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrological Processes*, *17*, 455–476, doi:10.1002/hyp.1135.

Weerts, A. H., and G. Y. H. El Serafy (2006), Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models, *Water Resources Research*, *42*, W09403, doi:10.1029/2005WR004093.

Widmann, M., and C. Bretherton (2000), Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States, *J. Climate*, *13*, 1936–1950.

Wilks, D. S. (1989), Rainfall intensity, the Weibull distribution, and estimation of daily surface runoff, *Journal of Applied Meteorology*, *28*, 52–58.

Xu, C.-Y. (2001), Statistical analysis of parameters and residuals of a conceptual water balance model – Methodology and case study, *Water Resources Management*, *15*(2), 75–92.

Xu, C.-Y., and G. Vandewiele (1994), Sensitivity of monthly rainfall-runoff models to input errors and data length, *Hydrological Sciences Journal*, *39*(2), 157–176.

Yang, D., S. Isida, B. E. Goodison, and T. Gunther (1999), Bias correction of daily precipitation measurements for Greenland, *Journal of Geophysical Research*, *104*(D6), 6171–6181.

Yang, J., P. Reichert, and K. C. Abbaspour (2007), Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur River basin (Switzerland), *Water Resources Research*, *43*, W10401, doi:10.1029/2006WR005497.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, *Journal of Hydrology*, *181*(1-4), 23–48, doi:10.1016/0022-1694(95)02918-4.

Zhang, L., and V. P. Singh (2007), Bivariate rainfall frequency distributions using Archimedean copulas, *Journal of Hydrology*, *332*(1-2), 93–109, doi: 10.1016/j.jhydrol.2006.06.033.

Acknowledgments. David Huard is grateful for the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix A: Probabilistic distributions

A1. Truncated normal distribution

Assuming variable x is distributed normally but bounded to the interval $[a, b]$, then the probability density function of x is given by

$$\mathbf{N}_T(x \mid \mu, \sigma, a, b) = \frac{\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}, \tag{A1}$$

where ϕ and Φ stand for the probability density function and cumulative density function of the standard normal distribution.

A2. Multivariate normal distribution

The probability density function of the multivariate normal distribution is given by

$$\mathbf{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{|2\pi\Sigma|} \exp\{-(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})/2\} \tag{A2}$$

where $\boldsymbol{\mu}$ is the vector mean and Σ the covariance matrix.

A3. Lognormal distribution

Assuming y is normally distributed with mean μ and variance σ^2 and $x = \exp(y)$, then x has a lognormal distribution with a probability density function given by

$$\mathbf{L}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} \mathrm{e}^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}. \tag{A3}$$

A rationale for the use of the lognormal in science is given in *Limpert et al.* [2001], along with some of its properties. For instance, the lognormal is the distribution maximizing entropy on the semi-infinite interval given a fixed mean and variance [*Goodman*, 1987]. It is also the limiting distribution of the product of random variates, and hence, ideally suited to model multiplicative errors.

If y is generated by an auto-regressive process of order one: $y_i = \rho y_{i-1} + \epsilon_i$ with $\epsilon \sim N(0, \sigma)$ for $i = 1, 2, \ldots, n$ and $x_i = \exp(\mu_i + y_i)$, then the log-likelihood of $\boldsymbol{x}$ is given by

$$\begin{aligned} L(\boldsymbol{x} \mid \boldsymbol{\mu}, \rho, \sigma) = & -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\frac{\sigma^{2n}}{1-\rho^2} - \sum_{i=1}^{n} \ln x_i \\ & - \frac{1}{2\sigma^2}\left\{(1-\rho^2)(\ln x_1 - \mu_1)^2 + \sum_{i=2}^{n}\left[(\ln x_i - \mu_i) - \rho(\ln x_{i-1} - \mu_{i-1})\right]^2\right\}. \end{aligned} \tag{A4}$$

A4. Exponentiated Weibull distribution

The exponentiated Weibull distribution, introduced by *Mudholkar and Hutson* [1996] has the following pdf:

$$\mathbf{W}(x \mid \alpha, k, \mu, \sigma) = \frac{\alpha k}{\sigma} z^{k-1} \mathrm{e}^{-z^k}\left[1 - \mathrm{e}^{-z^k}\right]^{\alpha-1}, \quad z = \frac{x-\mu}{\sigma} \tag{A5}$$

with $z > 0$ and $k > 0$. It is a generalization of the Weibull distribution, with an additional shape parameter α. The Weibull distribution is frequently used in hydrology to model rainfall characteristics [*Zhang and Singh*, 2007; *Sharma*, 1996; *Wilks*, 1989].

Appendix B: Aggregation of proportional daily errors

Given a daily series d_i with $i = 1, \ldots, n$ with proportional errors $\delta_{d_i} = f_d d_i$, one wishes to determine the corresponding proportionality factor f_m for aggregated monthly values $m = \sum_{i=1}^{n} d_i$, such that

$$f_m^2 m^2 = \sum_{i=1}^{n} f_d^2 d_i^2. \tag{B1}$$

Isolating f_m and taking the expectation on both sides of equation (B1) yields

$$E[f_m] = f_d E\left[\frac{\sqrt{\sum_{i=1}^{n} d_i^2}}{m}\right]. \tag{B2}$$

This equation shows that the monthly proportionality factor increases with the intra-monthly variability of the series, in other words, that errors on large daily events strongly contribute to the error on the monthly totals. For example, using daily rainfall series from the Chunky watershed, taking $f_d = .15$ yields a monthly proportionality factor $f_m \approx .07$. With a daily rainfall equal to its mean value, *i.e.* no variability in the series, equation (B2) reduces to $f_m = f_d/\sqrt{n}$ and $f_m \approx .03$.

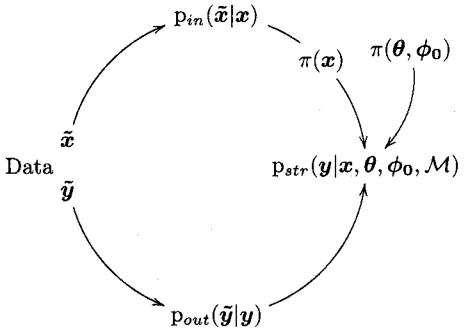

Figure 1. Inference process based on the theoretical uncertainty framework. Experimental input and output data are indicated by *tilded* variables $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$, and conceptual true values by non tilded variables: true input series $\boldsymbol{x}$; true output series $\boldsymbol{y}$; model parameters $\boldsymbol{\theta}$ and initial model state $\boldsymbol{\phi_0}$. The input, output and structural error models are indicated by p_{in}, p_{out} and p_{str} respectively, $\pi(\boldsymbol{\theta}, \boldsymbol{\phi_0})$ denotes the prior for the model parameters and the initial state, and $\pi(\boldsymbol{x})$ stands for the prior for the true input series.

Table 1. Properties of the eight watersheds used in the study. All stations span between 53 and 54 years of uninterrupted data with no null rainfall at the monthly scale. Rainfall, PE and flow data are areal values.

Code	Basin name	Area (km^2)	Mean monthly rainfall (mm)	Mean monthly PE (mm)	Mean monthly flow (mm)
05471500	South Skunk River near Oskaloosa, IA	4235	69.3	82.3	17.8
05570000	Spoon River at Seville, IL	4237	75.3	84.3	20.9
01562000	Raystown Branch Juniata River at Saxton, PA	1958	78.8	65.4	35.2
03011020	Allegheny River at Salamanca NY	4165	86.8	59.4	48.2
03303000	Blue River Near White Cloud, IN	736	93.8	73.4	68.2
01321000	Sacandaga River near Hope NY	1272	102.9	57.6	64.7
02475500	Chunky River near Chunky, MS	956	117.2	87.9	38.5
03443000	French Broad River at Blantyre, NC	767	161.2	70.1	99.1

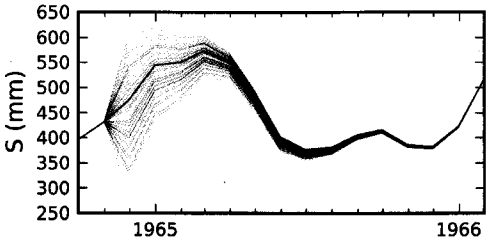

Figure 2. Once perturbed (thin lines), the internal state variable S describing the level (mm) of the soil moisture store returns to its unperturbed state (thick line). The other internal variable R (level of the routing store) displays a similar behavior. The figure shows results obtained with a large store capacity ($\theta_1 = 1000$), *i.e.* a long memory.

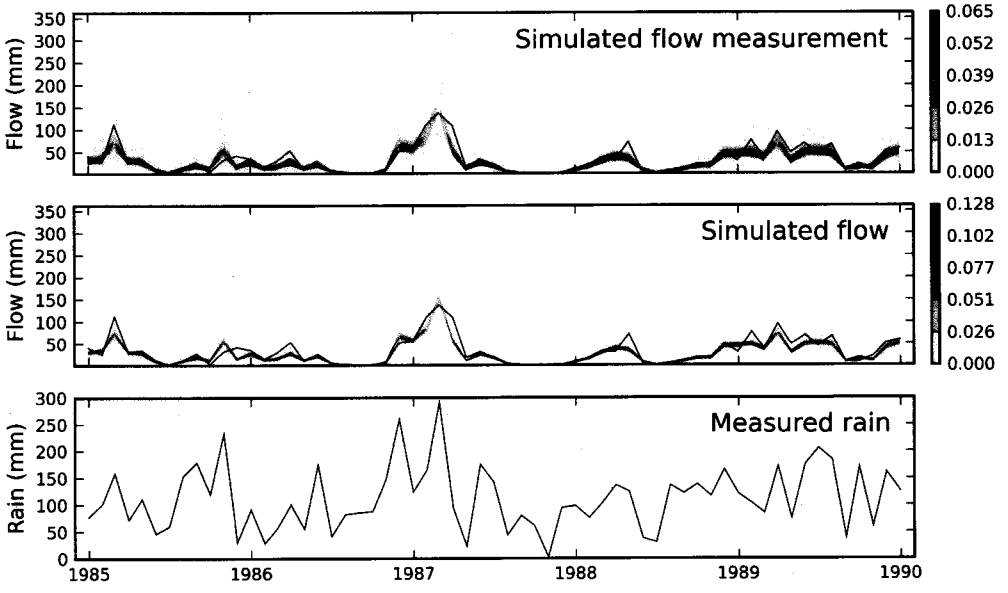

Figure 3. Case 1. Result of a calibration of the Chunky River watershed performed assuming that only output errors are present. The lower graph shows rainfall observations. The middle graph shows the pdf (contour plot) of the flow simulated using the parameters sampled during calibration compared with flow observations (full line). The upper graph shows the pdf of simulated *flow measurements*, including both parameter uncertainty and an explicit output observation error. Although calibration is performed over 20 years of data, only a subset of five years is shown here to avoid cluttering the figure. The scales on the right indicate the value of the pdf.

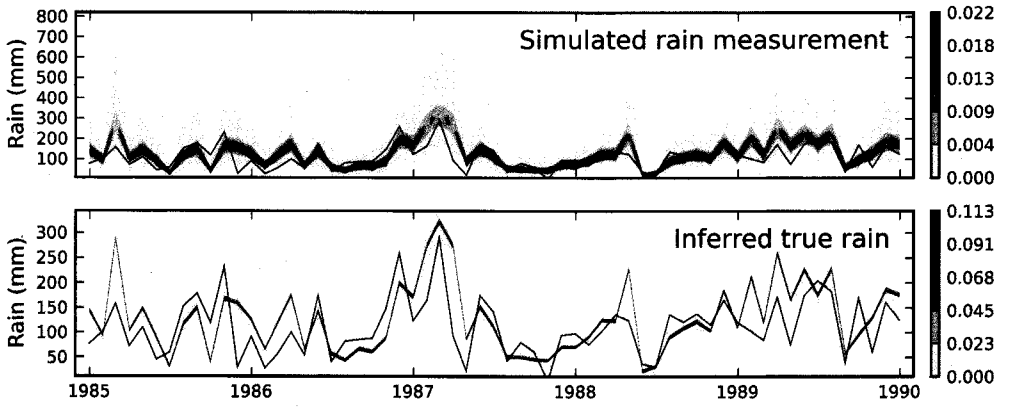

Figure 4. Case 2. Results of calibration of the Chunky River watershed under the assumption that only input errors are present. The lower graph shows a five years subset of the measured rainfall series (full line) used to calibrate the model, and a pdf (contour plot) of the sampled true rain. The upper graph shows the pdf of simulated *rain measurements*, including an explicit rain observation error. The scales on the right indicate the value of the pdf.

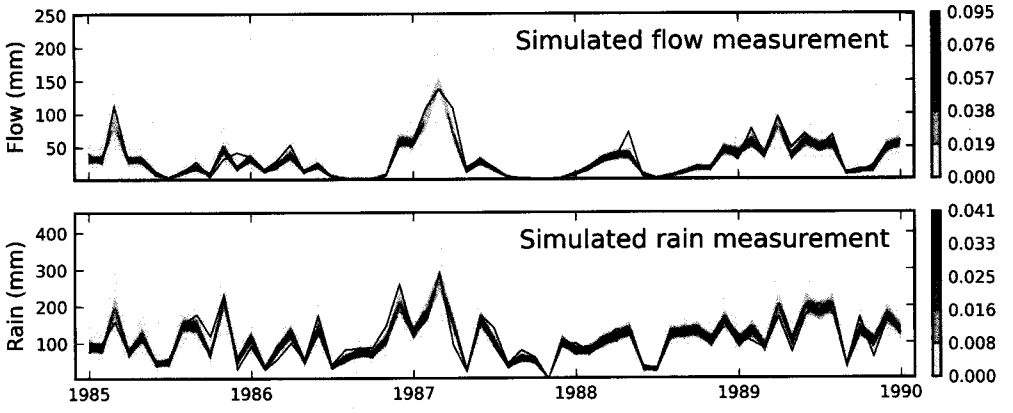

Figure 5. Case 3. Results from a calibration of the Chunky River watershed accounting for input, output and structural errors. The lower plot shows the pdf of simulated rain measurements (contour plot), compared with rainfall observations (full line), and the upper plot the pdf simulated flow measurement probability density (contour plot) compared with the observed flow series (full line). The scales on the right indicates the value of the pdf.

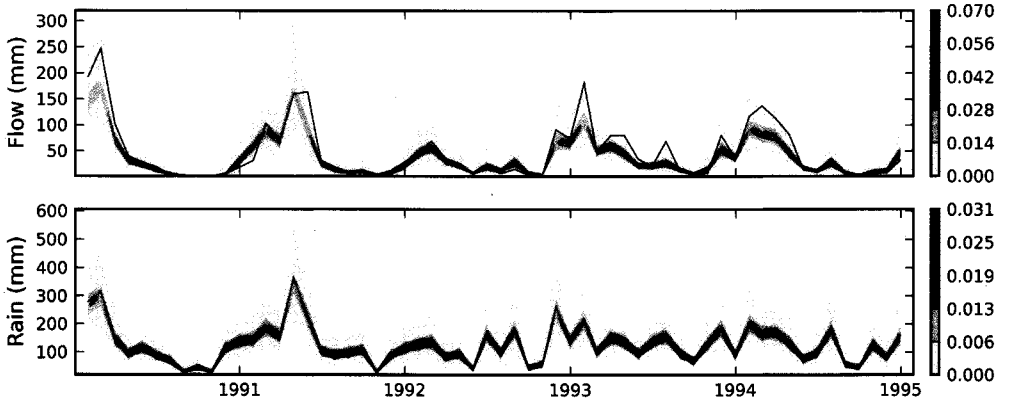

Figure 6. Validation of the model calibrated in case 3. The lower graph shows the pdf of the simulated measured rain $\tilde{\boldsymbol{X}}_+$ (contour plot) along with the observed rainfall (full line). The upper graph shows the pdf of the simulated measured flow $\tilde{\boldsymbol{Y}}_+$ (contour plot) compared with the observed flow series (full line) for the validation period. The scales on the right indicate the value of the pdf.

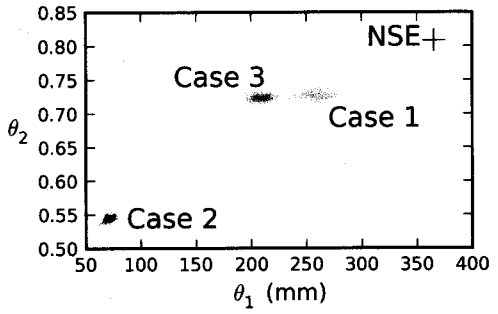

Figure 7. Parameter distributions sampled in the three cases presented. Parameter θ_1 describes the capacity (mm) of the soil accounting store and θ_2 the water exchange coefficient. As θ_2 decreases, more water is lost to neighboring catchments. The NSE parameter is computed by optimizing the NSE criteria computed over the square root transformed flow.

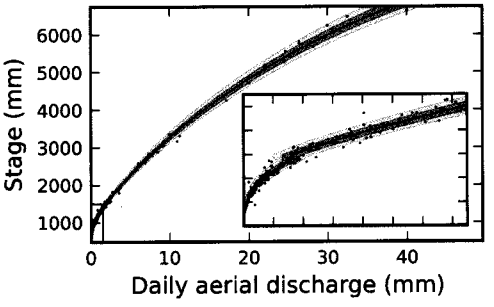

Figure 8. Daily output (discharge) error model for the Chunky River watershed. Dots indicate stage-discharge measurements taken by the USGS since 1938. The black line is a piecewise polynomial fit of the stage-discharge relation. The light gray area indicates the span of the first standard deviation (68% confidence) of the discharge error model and the darker gray the second deviation (95% confidence) for daily measurements. The rectangle in the lower-left corner shows the boundaries of the close-up view at the right.

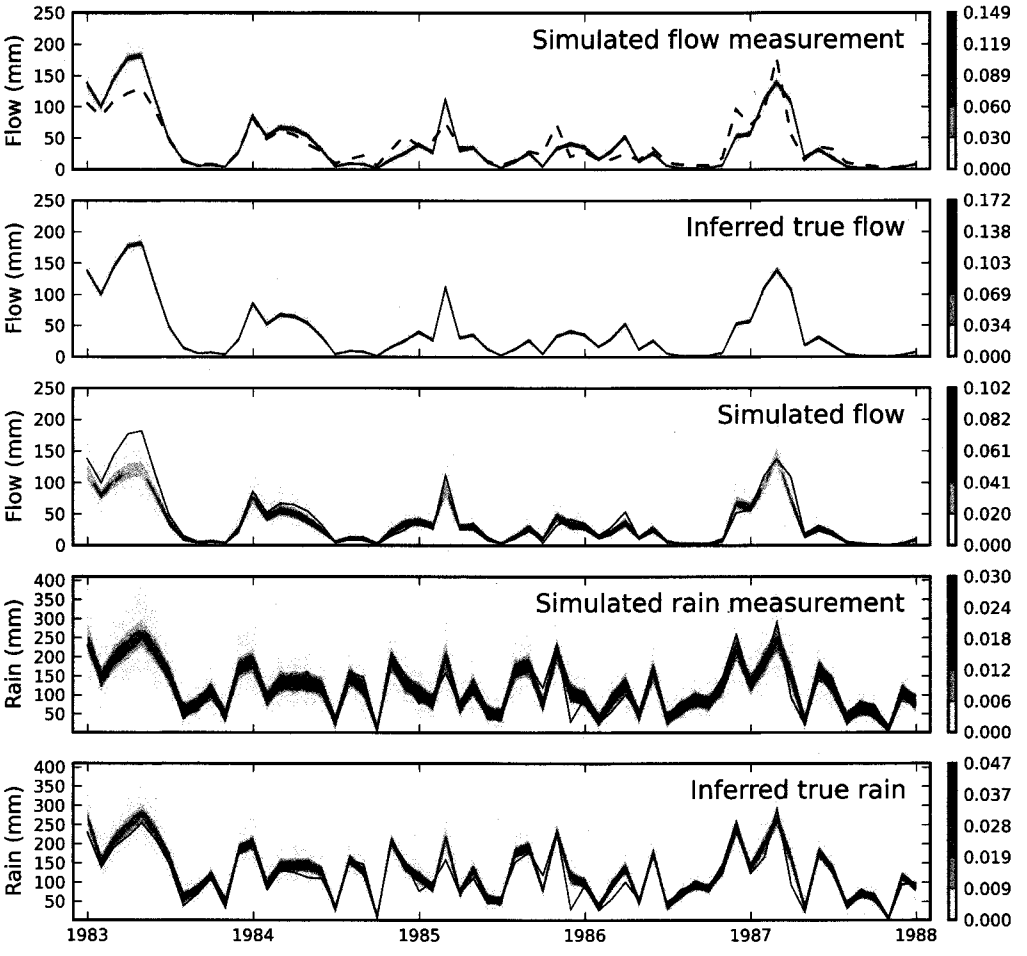

Figure 9. Calibration results for the Chunky River watershed using realistic priors and error model assumptions. Pdfs for each variable are shown as contour plots while observations are indicated by full lines. In the upper graph, the dashed line shows the flow simulated by parameters fitted using a least-square criteria (NSE) on the square root of the flows. The scales on the right indicate the value of the pdf.

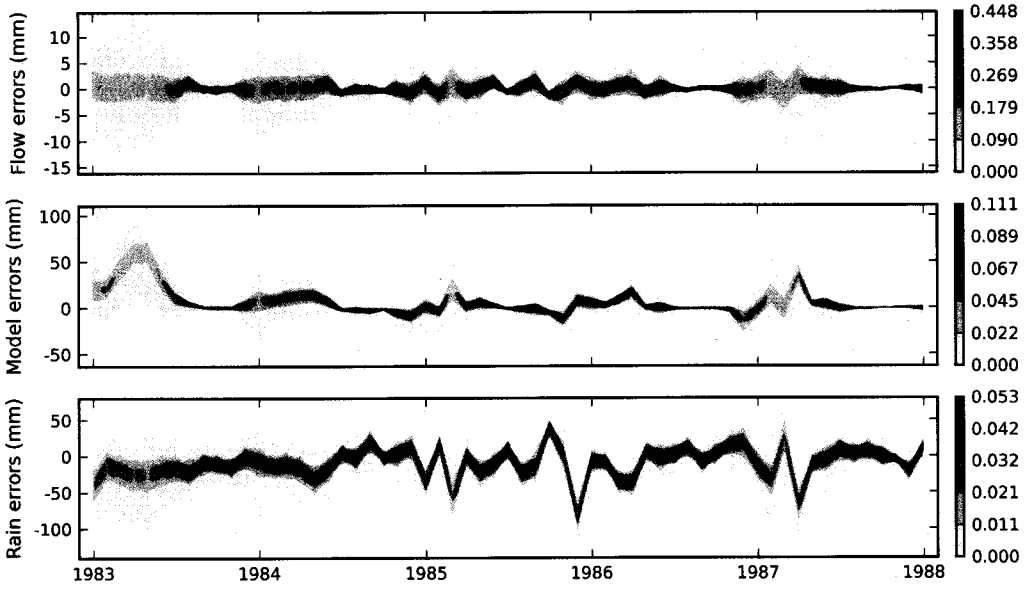

Figure 10. Inferred posterior pdf for the three main sources of errors: input (rainfall) errors (mm), structural or model errors (mm), and output (flow) errors (mm). The graphs show the last five years of the calibration period. The scales on the right indicate the value of the pdf.

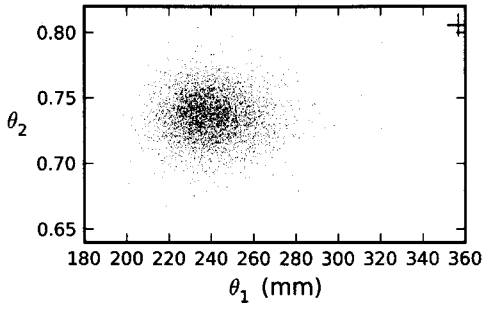

Figure 11. Parameters sampled during calibration under realistic error model assumptions. The cross indicates the optimal parameters found by maximizing the NSE criteria.

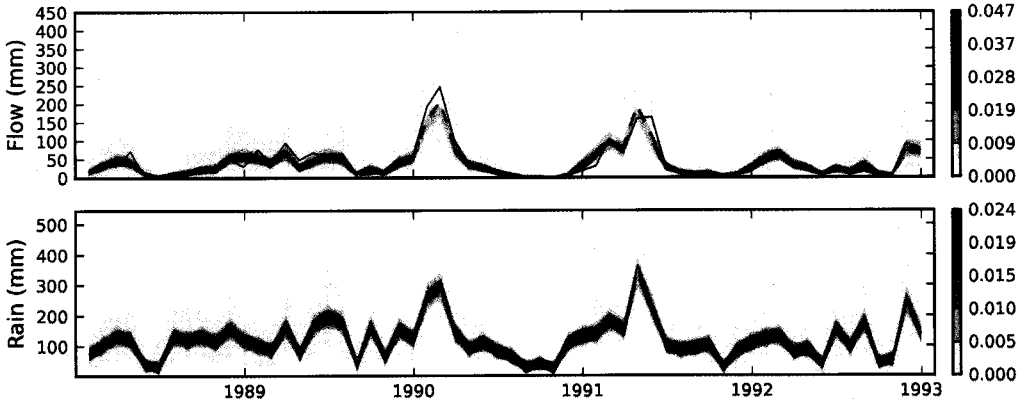

Figure 12. Flows simulated during the validation period. The lower graph shows a contour plot of the simulated rain measurements density, with actual measurements indicated by a full line. The upper plot displays the pdf of simulated flow measurements (contour plot), obtained by running the model over the parameters sampled during calibration and adding structural and measurement noise. The dashed line shows the flows estimated by the SLS method and the full line indicates flow observations. The scales on the right indicate the value of the pdf.

Bayesian copula selection

Les copules sont des distributions multivariées de lois marginales uniformes. Il en existe de nombreuses familles décrivant diverses structures de dépendances entre les marges. Toutefois, il n'existe pas de méthode satisfaisante permettant de sélectionner parmi ces familles la mieux adaptée à un jeu de données. Cet article présente donc une méthode Bayésienne qui, moyennant certaines hypothèses, permet d'attribuer un poids à chaque famille correspondant à la probabilité que celles-ci décrivent les données observées. La méthode se base sur l'idée du *prior parent*, un prior sur une quantité commune à toutes les familles de copules. La définition du prior parent permet d'attribuer un prior pour les paramètres des familles de manière à permettre la comparaison de celles-ci entre elles. Il est montré par le biais d'expériences synthétiques que la performance de la méthode, c'est-à-dire sa capacité à identifier la vraie copule, augmente avec la corrélation entre les données et la longueur de l'échantillon disponible pour effectuer la sélection.

Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 809–822
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Bayesian copula selection

David Huard, Guillaume Évin, Anne-Catherine Favre*

Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Qué., Canada G1K 9A9

Received 1 March 2005; received in revised form 23 June 2005; accepted 30 August 2005
Available online 6 October 2005

Abstract

In recent years, the use of copulas has grown extremely fast and with it, the need for a simple and reliable method to choose the right copula family. Existing methods pose numerous difficulties and none is entirely satisfactory. We propose a Bayesian method to select the most probable copula family among a given set. The copula parameters are treated as nuisance variables, and hence do not have to be estimated. Furthermore, by a parameterization of the copula density in terms of Kendall's τ, the prior on the parameter is replaced by a prior on τ, conceptually more meaningful. The prior on τ, common to all families in the set of tested copulas, serves as a basis for their comparison. Using simulated data sets, we study the reliability of the method and observe the following: (1) the frequency of successful identification approaches 100% as the sample size increases, (2) for weakly correlated variables, larger samples are necessary for reliable identification.

Keywords: Copulas; Model selection; Bayes' theorem; Goodness-of-fit test; Kendall's tau; Pseudo-likelihood

0. Introduction

In order to extrapolate extreme quantiles from data sets, or to generate random variables, it is usually necessary to select a distribution function matching the available data. The choice of the best distribution is not an exact science and relies on guesswork and testing of multiple hypotheses. Since each hypothesis comes with its particular test, the whole procedure is too complicated for end-users and generally left to experts, along with the interpretation of the results. Furthermore, existing methods cannot compare distributions without specifying an optimal parameter set for each one of them. The selection of the best distribution is thus intertwined with the estimation of parameters, a non-trivial problem itself.

The situation is even worse in the case of two-dimensional distributions, for which even more parameters need to be estimated. Fortunately, the elegant concept of copulas greatly simplifies matters. Copulas are multivariate distributions modeling the dependence structure between variables, irrespective of their marginal distribution. They allow to choose completely different margins, the dependence structure given by the copula, and merge the margins into a genuine multivariate distribution. The choice of the best bivariate distribution can then be done in two steps: choose the optimal margins, and then choose the optimal copula. In this paper, we introduce a simple Bayesian method to choose the "best" copula, given some bivariate data expressed by quantiles.

* Corresponding author. Tel.: +1 418 6543789; fax: +1 418 6542600.
E-mail addresses: david.huard@ete.inrs.ca (D. Huard), guillaume.evin@ete.inrs.ca (G. Évin), anne-catherine_favre@ete.inrs.ca (A.-C. Favre).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.08.010

The structure of the paper is as follows Section 1 introduces the main ideas of copula theory. Section 2 reviews existing approaches to select copulas and highlights salient features. Section 3 describes the proposed method and its derivation from Bayes' theorem. Results from numerical simulations are shown in Section 4, along with their analysis. Finally, we draw conclusions on the overall performance of the method and propose ideas for future work.

1. Copula theory

The concept of copula has been introduced by Sklar (1959) in the following way

Copula Definition. A copula is a joint distribution function of standard uniform random variables. That is,

$$C\left(u_1,\ldots,u_p\right)=\Pr\left\{U_1\leqslant u_1,\ldots,U_p\leqslant u_p\right\},$$

where $U_i\sim U(0,1)$ for $i=1,\ldots,p$.

For a more formal definition of copulas, the reader is referred to Nelsen (1999). Using the probability integral transformation, it is straightforward to see that a copula computed at $F_1(x_1), F_2(x_2),\ldots,F_p(x_p)$ is identical to the multivariate distribution function F evaluated at $(x_1,\ldots,x_p)$, i.e.,

$$C\left(F_1(x_1), F_2(x_2),\ldots,F_p(x_p)\right)=F\left(x_1,x_2,\ldots,x_p\right).$$

This last equality gives a first insight of the link between distribution functions and copulas, which is the content of Sklar's theorem.

Sklar's Theorem. *Let F be a p-dimensional distribution function with margins $F_1, F_2\ldots,F_p$, then there exists a p-copula C such that for all x in $\bar{\mathbb{R}}^p$,*

$$F\left(x_1,x_2,\ldots,x_p\right)=C\left(F_1(x_1), F_2(x_2),\ldots,F_p(x_p)\right),$$

where $\bar{\mathbb{R}}$ denotes the extended real line $[-\infty,\infty]$. If $F_1,\ldots,F_p$ are all continuous, then C is unique. Otherwise, C is uniquely determined on Ran $(F_1)\times$ Ran $(F_2)\times\cdots\times$ Ran (F_p), where Ran stands for the range.

According to Sklar's theorem, copulas separate marginal behavior, as represented by the F_i's, from the dependence structure. This constitutes one great advantage of copulas. In the usual representation of joint probabilities via multivariate distribution functions, the two cannot be separated. The general theory about copulas is summarized in Joe (1997), Nelsen (1999) or more recently in Cherubini et al. (2004). Copulas have been widely used in financial mathematics to determine the Value at Risk (see for example, Embrechts et al., 2002, 2003; Bouyé et al., 2000). Other fields of applications involve lifetime data analysis (Bagdonavicius et al., 1999), actuarial science (Frees and Valdez, 1998), and more recently, hydrology (De Michele and Salvadori, 2003; Favre et al., 2004).

Most copula applications are concerned with bivariate data. One reason for this is that relatively few copula families have practical p-dimensional generalization. The popular Archimedean 2-copulas (Genest and MacKay, 1986) for instance, have two known generalizations, both of them afflicted by serious shortcomings. Archimedean 2-copulas are defined as

$$C(u_1,u_2)=\begin{cases}\varphi^{-1}\left(\varphi(u_1)+\varphi(u_2)\right) & \text{if } \sum_{i=1}^{2}\varphi(u_i)\leqslant\varphi(0),\\ 0 & \text{otherwise}\end{cases}$$

with $\varphi(u)$ a $\mathscr{C}^2$ function satisfying $\varphi(1)=0$, $\varphi'(u)<0$ (φ is decreasing) and $\varphi''(u)>0$ (φ is convex) for all $0\leqslant u\leqslant 1$. $\varphi(u)$ is called the generator of the copula. The first generalization, termed symmetric (Joe, 1997), uses the same generator, thus the same dependence, for all variables

$$C\left(u_1,\ldots,u_p\right)=\varphi^{-1}\left(\varphi(u_1)+\cdots+\varphi\left(u_p\right)\right).$$

Since all variables are described by the same dependence, this generalization is too simplistic for most real life applications. The second generalization, termed asymmetric (Whelan, 2004), uses $(p-1)$ generators. For $p=3$, the

Table 1
Definition and parameter domain of the copulas used in this paper

Copula	$C(u, v\|\theta)$	$\theta \in \Omega$
Clayton	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$]0,\infty[$
AMH	$\dfrac{uv}{1-\theta(1-u)(1-v)}$	$[-1,1[$
Gumbel	$\exp\left\{-\left[(-\ln u)^{\theta}+(-\ln v)^{\theta}\right]^{1/\theta}\right\}$	$[1,\infty[$
Frank	$-\dfrac{1}{\theta}\ln\left(1+\dfrac{\left(\mathrm{e}^{-\theta u}-1\right)\left(\mathrm{e}^{-\theta v}-1\right)}{\mathrm{e}^{-\theta}-1}\right)$	$[-1,1]\mathbb{R}\setminus\{0\}$
Joe	$1-\left[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}\right]^{1/\theta}$	$[1,\infty[$
A12	$\left\{1+\left[\left(u^{-1}-1\right)^{\theta}+\left(v^{-1}-1\right)^{\theta}\right]^{1/\theta}\right\}^{-1}$	$[1,\infty[$
A14	$\left\{1+\left[\left(u^{-1/\theta}-1\right)^{\theta}+\left(v^{-1/\theta}-1\right)^{\theta}\right]^{1/\theta}\right\}^{-\theta}$	$[1,\infty[$
FGM	$uv+\theta uv(1-u)(1-v)$	$[-1,1]$
Gauss	$\int_{-\infty}^{\phi^{-1}(u)}\int_{-\infty}^{\phi^{-1}(v)}\dfrac{1}{2\pi\sqrt{1-\theta^2}}\exp\left(\dfrac{2\theta s\omega-s^2-\omega^2}{2\left(1-\theta^2\right)}\right)\mathrm{d}s\,\mathrm{d}\omega$	$[-1,1]$

formula is given by

$$C\left(u_1,u_2,u_3\right)=\varphi_1^{-1}\left(\varphi_1\circ\varphi_2^{-1}\left(\varphi_2\left(u_1\right)+\varphi_2\left(u_2\right)\right)+\varphi_1\left(u_3\right)\right). \tag{1}$$

Although $(p-1)$ different dependence structures are allowed instead of one, existence conditions must be satisfied for the copula defined in (1) to be valid. These conditions impose restrictions over the parameters of the generators, allowing only a limited range of correlations. All in all, the authors consider Archimedean p-copulas unpractical at best. Confronted with a multivariate sample, the usual approach is to analyze the data pair by pair using 2-copulas.

It is important here to say a few words about Kendall's tau as it plays a major role in the parametrization used in this paper (Section 3.1), as well as in the prior's choice (Section 3.2). Kendall's tau is a non-parametric measure of dependence between random variables commonly associated to copulas. It is defined as the probability of concordance minus the probability of discordance of two independent vectors (Kruskal, 1958). In other words, if (X_1, Y_1) and (X_2, Y_2) are independent and identically distributed random vectors, then

$$\tau=\Pr\left[\left(X_1-X_2\right)\left(Y_1-Y_2\right)>0\right]-\Pr\left[\left(X_1-X_2\right)\left(Y_1-Y_2\right)<0\right].$$

Kendall's tau may also be expressed simply in term of the copula function, as the expected value of the function $C(U, V)$ of uniform (0, 1) random variables U and V, whose joint distribution function is C, i.e.

$$\tau=4\int\int_{[0,1]^2}C(u,v|\theta)\,\mathrm{d}C(u,v|\theta)-1.$$

For Archimedean copulas, Kendall's tau can be written as $\tau=1+4\int_0^1\left(\varphi(t)/\varphi'(t)\right)\mathrm{d}t$.

The nine copula families used in this paper, chosen for their analytical properties, are defined in Table 1. Put together, they model a wide variety of dependence structure and cover most applications found in the literature. The first seven copulas (Clayton, Ali–Mikhail–Haq (AMH), Gumbel, Frank, Joe, A12, A14) belong to the Archimedean class. Note that A12 and A14 are coined from the order of appearance in Nelsen (1999). We also consider two non-Archimedean copulas: the Farlie–Gumbel–Morgenstern (FGM), a copula with a quadratic section, and Gauss copula, an elliptical copula. Note that the link between Gauss copula and the classical multivariate normal distribution is made explicit when the copula is expressed using the normal cdf: $C(u, v|\theta)=\Phi_\theta\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$. For more details about elliptical

copulas, one of the few families with practical p-dimensional generalizations, the reader is referred to Frahm et al. (2003).

2. Review of copula selection methods

The most commonly employed methods to select the best copula are based on a likelihood approach, which is used to define indicators of performance, as for example, the Akaike Information Criteria (AIC). Chen and Fan (2005) propose pseudo-likelihood ratio tests for selecting semi parametric multivariate copula models in which the marginal distributions are unspecified. For Archimedean copulas, Genest and Rivest (1993) proposed to compare the one-dimensional function

$$K_\theta(t) = \Pr(C(u, v|\theta) < t),$$

with its non-parametric estimation K_n, given by

$$K_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left(e_{jn} \leqslant t\right),$$

where $e_{jn} = (1/n)\sum_{k=1}^{n} \mathbb{1}\left(X_{1k} \leqslant X_{1j}, \ldots, X_{pk} \leqslant X_{pj}\right)$. The best copula is then the one for which the function K_θ is closest to K_n. Durrleman et al. (2000) suggested to choose the copula minimizing the distance (L^2-norm, Kolmogorov, etc.) from K_θ to the non-parametric estimation K_n. Using the same idea, they computed a distance based on the discrete L^2-norm. This distance is calculated between an empirical copula of Deheuvels (1979) and the tested copulas. For an exhaustive presentation of these procedures, we refer to the working paper of Durrleman et al. (2000).

In recent years, various authors developed goodness-of-fit tests (GOF tests) for copulas. Genest et al. (2005) proposed an user-friendly and powerful tool, a GOF test statistic with a non-truncated version of Kendall's process

$$\mathbb{K}_n(t) = \sqrt{n}\left\{K_n(t) - K_{\theta_n}(t)\right\},$$

where θ_n denotes a robust estimation of θ. The expression for the statistic is straightforward and the test has nice properties. For example, it sustains the prescribed error probability of the first kind under the null hypothesis, even with small sample sizes. Nevertheless, an explicit expression is needed for K_θ, which limits the set of copulas for which the GOF test statistic can be computed.

An easy way to construct GOF tests for copulas is to consider p-dimensional χ^2 tests. The methodology is presented in Pollard (1979). Dobrić and Schmidt (2004) recently used this method in a financial application. The main criticism about this approach concerns the arbitrary choice of the subsets that divide the p-dimensional space $[0, 1]^p$ (Kendall and Stuart, 1983). Also, the calculation of the empirical critical values proves to be troublesome.

Several authors employed the transformation of Rosenblatt (1952) to test whether the transformed random variables set is composed of independent uniformly distributed variables, as is the case under the null hypothesis (Justel et al., 1997). Using Rosenblatt's transformation, Chen et al. (2003) compared the kernel density estimation of their transformed random variables to the uniform density. As noted in Fermanian (2005), Rosenblatt's transformation may be a tedious preliminary task, especially for high dimensions. Therefore, Fermanian (2005) presents a GOF test based directly on the kernel density estimation of the original multivariate data. This test, however, requires some heavy numerical integration. Moreover, it was noted through simulations that with small sample sizes, it is difficult to sustain the prescribed error probability of the first kind.

The comments in this section are based on a survey (Evin, 2004) that compared the results of the GOF tests presented above. Numerous simulations led the authors to conclude that the test of Genest et al. (in press) is the only one to be unbiased, and by far the most powerful. All these tests, however, rely on previous estimation of an optimal parameter set. Strictly speaking, comparisons are made between copulas with given parameters, and not between copula families. We suggest that model selection methods should be independent of the parameter choice.

3. Selection of the right copula

We present in this section a Bayesian model selection method that attributes weights to copula families. It does not rely on parameter estimation and to our knowledge, may be applied to all known copulas.

Let $\mathscr{C}$ denote the set of all copulas. From this set, we select a finite subset $\mathscr{C}_Q \subset \mathscr{C}$ of copulas to be included in the proposed method. Each family of copulas in $\mathscr{C}_Q$ is identified by C_l, with $l = 1, \ldots, Q$. The proposed method consists simply in defining Q hypotheses:

H_l: The data come from copula C_l, $\quad l = 1, \ldots, Q$

and then computing $\Pr(\mathrm{H}_l|D)$, the probability of each hypothesis given the data D. We will suppose that the data set D is composed of n mutually independent pairs of quantiles (u_i, v_i), $i = 1, \ldots, n$. Note that the independence assumption may not hold if quantiles are computed empirically using ranks. In that case, the method must be thought of as approximative and used with caution. Applying Bayes' theorem, we get for each family:

$$\Pr(\mathrm{H}_l|D, I) = \frac{\Pr(D|\mathrm{H}_l, I)\Pr(\mathrm{H}_l|I)}{\Pr(D|I)}, \tag{2}$$

where $\Pr(D|\mathrm{H}_l, I)$ is the likelihood, $\Pr(\mathrm{H}_l|I)$ is the prior on the copula family, $\Pr(D|I)$ the normalization constant and I stands for any relevant additional knowledge. The "right" copula is then simply the copula with the highest $\Pr(\mathrm{H}_l|D, I)$.

3.1. Likelihood

The likelihood $\Pr(D|\mathrm{H}_l, I)$ in Eq. (2) is the probability of "drawing" data D from the lth copula. However, for most copulas, there exists no such explicit expression because the copula density depends on a parameter θ (cf. Table A.2). Note also that the parameterization is arbitrary, in the sense that we could choose any function $\beta = g(\theta)$ and replace $c(u, v|\theta)$ by

$$\tilde{c}(u, v|\beta) = c\left(u, v|g^{-1}(\beta)\right).$$

There is, of course, no wrong choice for the parameter, and for reasons that will soon be clear, we will choose Kendall's tau $\tau = g_l(\theta)$ to be the common parameter for all copulas in $\mathscr{C}_Q$ (cf. Table 2).

We introduce Kendall's tau in Eq. (2) as a nuisance variable:

$$\begin{aligned}\Pr(\mathrm{H}_l|D, I) &= \int_{-1}^{1} \Pr(\mathrm{H}_l, \tau|D, I)\, \mathrm{d}\tau \\ &= \int_{-1}^{1} \frac{\Pr(D|\mathrm{H}_l, \tau, I)\Pr(\mathrm{H}_l|\tau, I)\Pr(\tau|I)\,\mathrm{d}\tau}{\Pr(D|I)},\end{aligned} \tag{3}$$

where $\Pr(\mathrm{H}_l|\tau, I)$ is the prior on the family hypothesis and $\Pr(\tau|I)$ is the prior density on Kendall's tau. The likelihood $\Pr(D|\mathrm{H}_l, \tau, I)$ now depends on τ and, if all n data pairs are mutually independent, can be computed from the copula density:

$$\begin{aligned}\Pr(D|\mathrm{H}_l, \tau, I) &= \prod_{i=1}^{n} \Pr(u_i, v_i|\tau, l, I) \\ &= \prod_{i=1}^{n} c_l\left(u_i, v_i|g_l^{-1}(\tau)\right),\end{aligned} \tag{4}$$

where $c_l\left(u_i, v_i|g_l^{-1}(\tau)\right)$ is the density of the lth copula (cf. Table A.2). If quantiles are computed using ranks, mutual independence cannot be guaranteed and Eq. (4) rather describes a pseudo-likelihood. In that case, results obtained should be considered as approximations. Fortunately, this approximation is thought to improve with increasing sample size, as the dependence between ranks decreases.

Table 2
Kendall's tau and its domain of definition for the copulas used in this paper

Copula	$\tau = g(\theta)$	$\tau \in \Omega$
Clayton	$1 - \frac{2}{2+\theta}$	$[0, 1]\backslash\{0\}$
AMH	$1 - \frac{2}{3}\frac{\theta^2 \ln(1-\theta) - 2\theta \ln(1-\theta) + \theta + \ln(1-\theta)}{\theta^2}$	$[-0.181726, \frac{1}{3}]$
Gumbel	$1 - \theta^{-1}$	$[0, 1]$
Frank	$1 - \frac{4}{\theta}\left(1 - \frac{1}{\theta}\int_0^\theta \frac{t}{e^t - 1}\,dt\right)$	$[-1, 1]\backslash\{0\}$
Joe	No closed form	Discontinuities
A12	$1 - \frac{2}{3\theta}$	$[\frac{1}{3}, 1]$
A14	$1 - \frac{2}{1+2\theta}$	$[\frac{1}{3}, 1]$
FGM	$\frac{2\theta}{9}$	$[-\frac{2}{9}, \frac{2}{9}]$
Gauss	$\frac{2}{\pi}\arcsin\theta$	$[-1, 1]$

3.2. Priors

To select priors, a host of methods exists (Kass and Wasserman, 1996), some said objective, others said subjective, but the choice of the method itself remains subjective and open to debate. Our approach consists in stating desiderata precise enough to define uniquely the prior on τ as well as the prior on the family. The subjectivity of the prior's choice is then confined to these desiderata and put in evidence for criticism. Let us state these basic desiderata, the additional information denoted earlier by I:

(I_1) Kendall's tau belongs to the set Λ and each outcome of $\tau \in \Lambda$ is equally likely;
(I_2) for a given τ, all families satisfying $\tau \in \Omega_l$ are equally probable,

where Ω_l is τ's domain for the lth copula (see Table 2). The purpose of Λ is to give the user the possibility to include additional knowledge about the correlation between the variables. For example, if the correlation is known to be positive, we may assume that $\Lambda = [0, 1]$. In the case where no information is available, Λ is simply put equal to $[-1, 1]$.

Desideratum (I_2) determines the prior on the family. Indeed, since all families are equally probable with respect to a given τ (for $\tau \in \Omega_l$),

$$\Pr(\mathrm{H}_l|\tau, I_2) \propto \mathbb{1}(\tau \in \Omega_l). \tag{5}$$

Similarly, Desideratum (I_1) specifies the prior on τ:

$$\Pr(\tau|I_1) = \begin{cases} \dfrac{1}{\lambda(\Lambda)} & \tau \in \Lambda, \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

where $\lambda(\cdot)$ denotes the Lebesgue measure, here the width of the interval spanned by Λ.

These priors, chosen mainly for testing purposes, reflect complete ignorance of the correlation between the variables as well as no preference over the copula family. However, in real cases, cogent information would probably be available and should be included in the calculation via an informative prior. For example, if τ is known to lie around a certain

value, a beta distribution, extended to the range $[-1, 1]$, could provide an effective way to describe the prior on τ:

$$\Pr(\tau) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{1+\tau}{2}\right)^{\alpha-1}\left(\frac{1-\tau}{2}\right)^{\beta-1}.$$

To estimate the parameters, the easiest way would probably be to vary α, β until the shape of the distribution agrees with our intuitive perception of the probability. Information on the copula family may also be included, by modifying the weight given to each family. For example, if we knew beforehand that the type of data to analyse exhibits correlated extreme events, we could increase the weight given to copulas with strong tail dependence.

3.3. Normalization

Plugging Eqs. (4), (5) and (6) into (3), we find

$$\begin{aligned}\Pr(\mathrm{H}_l|D, I) &= \frac{1}{\Pr(D|I)}\int_{-1}^{1}\prod_{i=1}^{n} c_l\left(u_i, v_i|g_l^{-1}(\tau)\right)\cdot \mathbb{1}(\tau\in\Omega_l)\cdot\frac{\mathbb{1}(\tau\in\Lambda)}{\lambda(\Lambda)}\,\mathrm{d}\tau \\ &= \frac{1}{\Pr(D|I)}\frac{1}{\lambda(\Lambda)}\int_{\Omega_l\cap\Lambda}\prod_{i=1}^{n} c_l\left(u_i, v_i|g_l^{-1}(\tau)\right)\mathrm{d}\tau. \end{aligned} \tag{7}$$

In general, the normalization constant $\Pr(D|I)$ in Eq. (7) is computed using the sum rule (Jaynes and Bretthorst, 2003):

$$\Pr(D|I) = \sum_{l=1}^{Q}\Pr(D|\mathrm{H}_l, I)\Pr(\mathrm{H}_l|I). \tag{8}$$

However, the sum rule is only true if the hypotheses H_l are mutually exclusive and the set exhaustive. In our case, arguments invalidate both claims. First, if the data come from a copula not in $\mathscr{C}_Q$, the set of hypotheses is clearly not exhaustive. Second, if the set contains two or more copulas that are very similar, the hypotheses should not be considered completely exclusive. Solutions to insure exhaustiveness and take non-exclusivity into account are discussed in Section 4.3, their application, however, is beyond the scope of this article. We will hence limit the computation to the weights W_l:

$$W_l = \frac{1}{\lambda(\Lambda)}\int_{\Omega_l\cap\Lambda}\prod_{i=1}^{n} c_l\left(u_i, v_i|g_l^{-1}(\tau)\right)\mathrm{d}\tau. \tag{9}$$

Note that in the figures shown below, the weights are normalized for convenience.

4. Simulations and analysis

To assess the performance of the method, we select eight one-parameter copulas to form the subset $\mathscr{C}_Q$: Clayton, AMH, Gumbel, Frank, A12, A14, FGM and Gauss. Those copulas are chosen because analytical formulas exist both for the density (cf. Table A.2) and Kendall's tau (cf. Table 2). These copulas are then used to generate data sets of different sizes and correlations. We study the cases of small negative dependence $\tau=-0.2$ and small positive dependence $\tau=0.2$, using samples of sizes $n=30$, 300 and 600, and the cases of medium dependence $\tau=0.5$ and large dependence $\tau=0.7$, using samples of sizes $n=30$, 100 and 300. For each copula, for each n and for each τ, 1000 data sets are generated over which Eq. (9) for $l=1$ to Q is computed.

4.1. Main results

Once the weights are computed from (9), we count the number of times the right copula is chosen by the method, that is, the number of times it attains the highest weight among copulas from the set. The results are presented in

Table 3
Number of successful identifications over 1000 trials

Copula	τ	−0.2			0.2			0.5			0.7		
	n	30	300	600	30	300	600	30	100	300	30	100	300
Clayton					466	865	951	774	917	993	917	998	1000
AMH					127	495	734						
Gumbel					411	850	954	667	853	991	652	790	929
Frank		567	627	696	409	484	618	577	845	980	738	965	1000
A12								291	610	795	526	798	936
A14								216	521	786	287	597	903
FGM		52	478	637	52	438	635						
Gauss		549	750	832	330	576	760	380	682	934	474	750	958

Table 3. We note the following:

- As expected, the method becomes more accurate as the sample size grows.
- As τ approaches zero, larger samples are required for a successful identification.
- Some copulas are easier to identify than others.

Regarding the first comment, results show that the method converges to the right copula. That is, we suspect that as n increases, the probability of a successful identification approaches one.

The second comment may be explained by noting that certain copulas show a similar behavior when τ approaches zero. That is, copulas cluster into classes defined by an identical asymptotic density (Genest et al., 2005a). For example, FGM, AMH and Frank copula are part of the same class, whose asymptotic density is given by $c(u, v|\tau \to 0) \propto uv(1-u)(1-v)$. This similarity between copulas is more clearly seen in Fig. 2, displaying weights computed by the method, averaged over 1000 trials.

Concerning the third comment, the ease with which copulas are identified is related to the presence of similar copulas. Copulas that look alike obtain similar weights and hence, are difficult to identify. Conversely, copulas with peculiar densities stand out and are easily identifiable. The Clayton, for instance, was shown in Saïd (2004) to display a behavior different from other copulas. This fact is also pointed out in Genest and Verret (2005), where the authors showed that in the context of locally most powerful rank tests of independence, against alternatives expressed by copula models, Clayton differs from other families. This explains why in all cases, Clayton is the copula most often successfully identified.

To gain a better feeling of the reliability of the method, we compute the average weight obtained for each copula. Figs. 1 and 2 illustrate those weights, along with the weight obtained for the independent copula $c(u, v)=1$, as a measure of comparison. Fig. 2 shows clearly the connection between AMH, Frank and FGM. Fig. 1 suggests an unexpected connection between Gumbel and A14. Comparison of the two figures also shows how identification becomes easier when variables are strongly correlated. Note also that the independent copula obtains a relatively high weight for $\tau=0.2$ and $n=30$. In that case, the independent copula is selected around 50% of the time, signifying that for such a small τ, there is not sufficient data to distinguish from independence (see Table A.1).

4.2. Additional results and comments

In order to complete the analysis, further simulations are done to explore different questions. In particular, we inquire the impact of empirically computed quantiles on the reliability of the method, discuss what happens when the data come from an "unknown" copula, provide some directions about how to include copulas with multidimensional parameters and suggest a way to normalize the weights.

4.2.1. Empirical quantiles

As noted earlier, if $\vec{u}$, $\vec{v}$ are not known exactly, the method is approximate. To evaluate the effect of this approximation, we use pairs (u_i, v_i) generated from the copulas to compute (x_i, y_i) by $x_i = F_1^{-1}(u_i)$ and $y_i = F_2^{-1}(v_i)$, where F_1

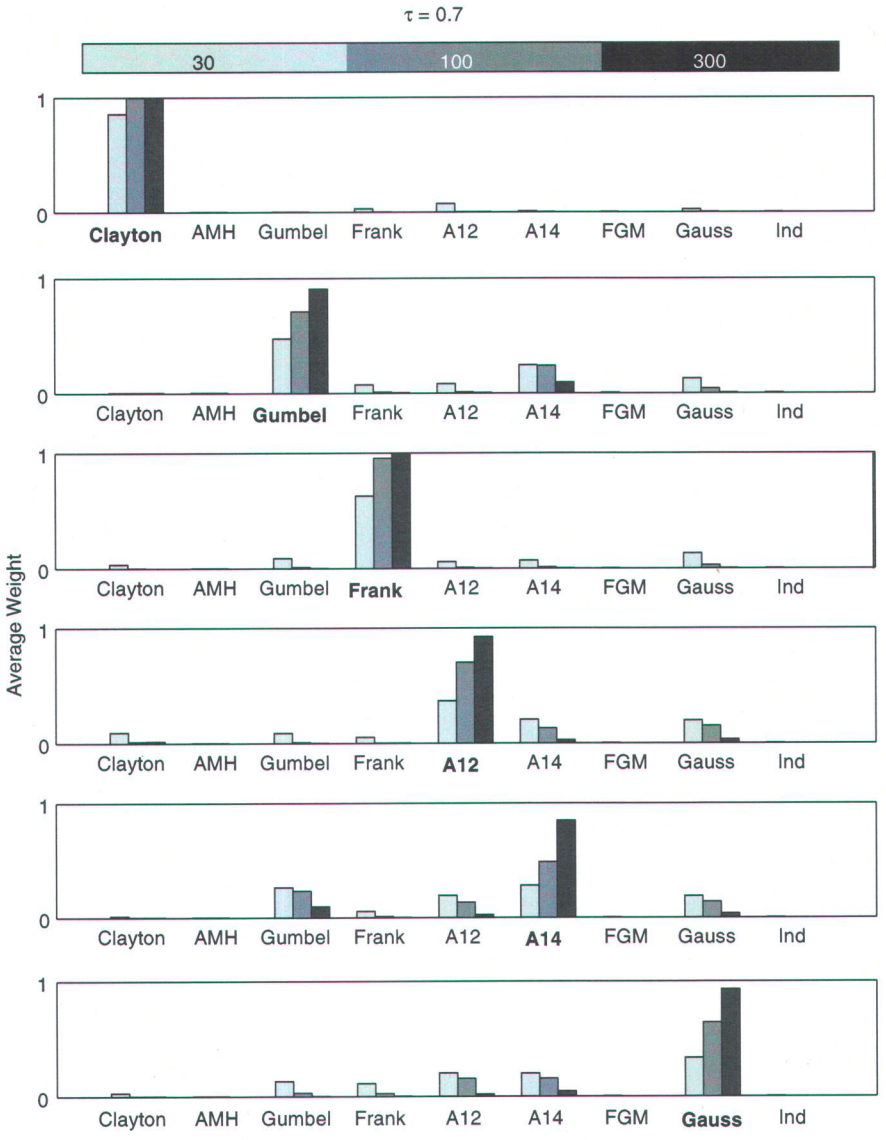

Fig. 1. Weights assigned to copulas, averaged over 1000 trials, for a Kendall's τ of 0.7 and samples sizes $n = 30$, 100 and 300. Bold font indicates the copula family that generated the samples.

and F_2 are asymmetrical Gamma distributions. Empirical quantiles $(\hat{u}_i, \hat{v}_i)$ are then estimated with $\hat{u}_i = 1/(n - 1)\#\{j \neq i : x_j \leqslant x_i\}$. The weights are computed using the empirical quantiles and compared to those obtained with the theoretical quantiles. Table 4 compares the number of successful identification in both cases. As expected, the uncertainty induced by the empirical quantiles reduces the number of successful identifications. The effect, however, weakens for large samples.

4.2.2. Unknown copula hypothesis

To understand what happens when a sample comes from a copula not in the set $\mathscr{C}_Q$, we run simulations using samples generated by Joe copula, with $\tau = 0.5$ and $n = 30$, 100 and 300. In the vast majority of cases, Gumbel copula obtains the highest weight. Although this result is hardly surprising since Joe and Gumbel have similar shapes, it highlights

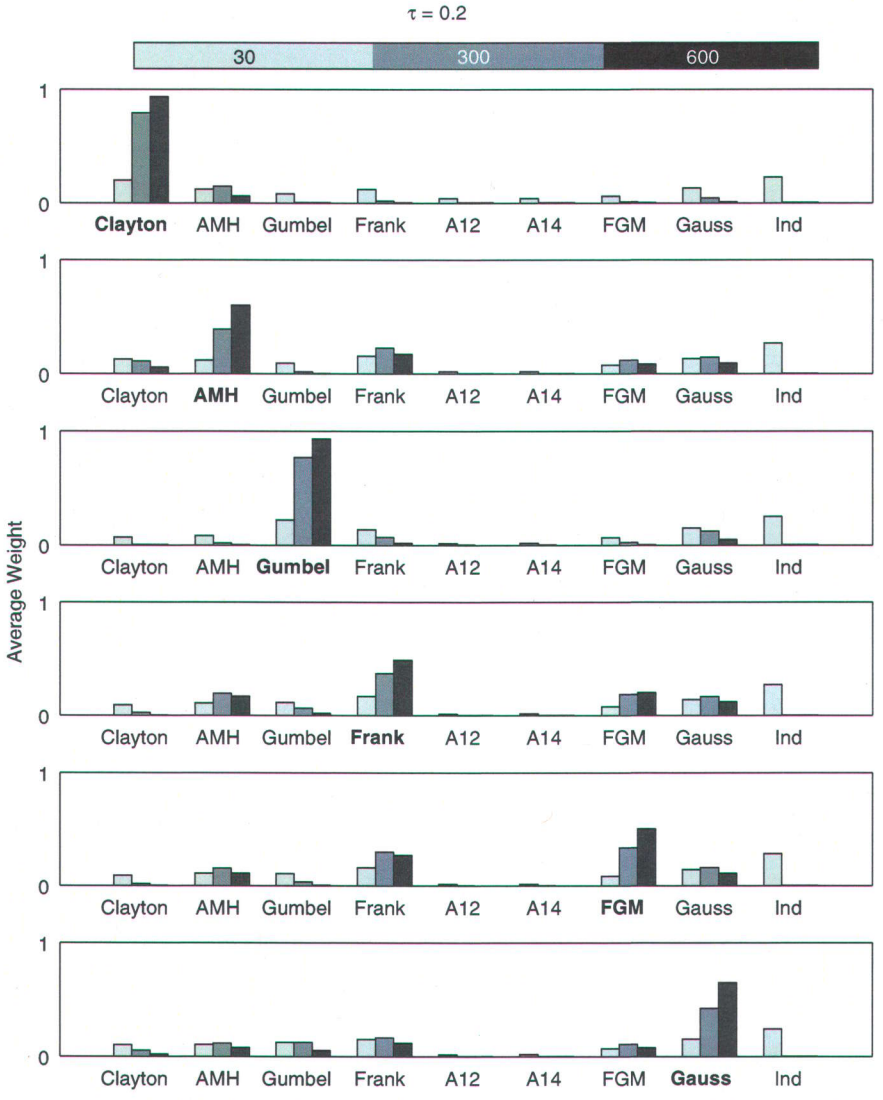

Fig. 2. Weights assigned to copulas, averaged over 1000 trials, for a Kendall's τ of 0.2 and samples sizes $n = 30$, 300 and 600. Bold font indicates the copula family that generated the samples.

the fact that no warning signal is given when the data are coming from a copula not in $\mathscr{C}_Q$. To rectify the situation, a solution would be to include an additional hypothesis:

H_{Q+1}: The data come from an " unknown" copula.

The difficulty is, of course, to define the density of an "unknown" copula. A solution proposed by Bretthorst (1996) in the context of radar target identification, is to describe the "unknown" hypothesis by an over-parameterized (o.-p.) model. This o.-p. model must have enough parameters to capture virtually every conceivable behavior. Due to its high flexibility, it should reach likelihoods equivalent to those of the right copula, but its priors on the extra parameters would reduce its overall posterior probability. Hence, it would obtain weights lower than the right copula, but higher than false copulas. The o.-p. model would then only be selected when the true model is not in $\mathscr{C}_Q$.

Table 4
Comparison of the number of successful identifications over 1000 trials for theoretical and empirical quantiles, using $\tau = 0.5$

Copula	Theoretical			Empirical		
n	30	300	600	30	300	600
Clayton	742	903	992	566	765	960
Gumbel	667	860	989	536	733	946
Frank	571	842	967	303	722	945
A12	322	604	823	144	394	675
A14	201	484	787	81	335	654
Gauss	404	685	917	230	471	840

4.2.3. Two parameters copula

In principle, the method can be applied to copulas with any number of parameters. In practice, however, difficulties appear since it implies high-dimensional integration and the definition of multidimensional priors. To see how it could be done, let us consider the case of a copula with two parameters $\vec{\theta} = (\theta_1, \theta_2)$. To apply the method, we parameterize the copula in terms of τ and κ, where κ is a relevant quantity, uniquely determined by $(\tau, \kappa) = g\left(\vec{\theta}\right)$. One such quantity could be the tail dependence, a measure of the probability of correlated extreme events (Juri and Wüthrich, 2003; Schmidt, 2002). A bivariate prior $\pi(\tau, \kappa)$ would then have to be specified, and the integration performed over τ and κ. Note that the computation of $g^{-1}(\tau, \kappa)$ might not be efficient. A better and equivalent solution consists in transforming the prior on (τ, κ) into a prior on $\vec{\theta}$ using the usual formula for variable substitution:

$$\pi_0\left(\vec{\theta}\right) = \pi_{\tau\kappa}\left(g\left(\vec{\theta}\right)\right)|J|,$$

where J is the determinant of the Jacobian of the transformation. This method has been used in the computation for this paper. It has the advantage that J generally has a closed form, which is often not the case for g^{-1}.

4.3. Normalization of the weights

As discussed in Section 3.3, the basic sum rule should not be applied since the set of hypotheses may not be exhaustive $\Pr\left(\mathrm{H}_1 + \cdots + \mathrm{H}_Q\right) \neq 1$ (+ stands for the logical OR operator), and some of the hypotheses might not be mutually exclusive $\Pr\left(\mathrm{H}_i, \mathrm{H}_j\right) \neq 0,\ i \neq j$. However, normalization is possible, although potentially tedious, if one introduces the unknown copula hypothesis and applies the extended sum rule. Indeed, adding the unknown copula hypothesis makes the set becomes exhaustive by definition, and the extended sum rule $\Pr(A + B) = \Pr(A) + \Pr(B) - \Pr(A, B)$ allows for non-exclusive hypotheses. If the set contains three hypotheses, for instance, we would reckon:

$$\begin{aligned}\Pr(D) &= \Pr\left(D, \mathrm{H}_1 + \mathrm{H}_2 + \mathrm{H}_3\right)\\ &= \Pr\left(D|\mathrm{H}_1\right)\Pr\left(\mathrm{H}_1\right) + \Pr\left(D|\mathrm{H}_2\right)\Pr\left(\mathrm{H}_2\right) + \Pr\left(D|\mathrm{H}_3\right)\Pr\left(\mathrm{H}_3\right)\\ &\quad - \Pr\left(D|\mathrm{H}_1, \mathrm{H}_2\right)\Pr\left(\mathrm{H}_1, \mathrm{H}_2\right) - \Pr\left(D|\mathrm{H}_1, \mathrm{H}_3\right)\Pr\left(\mathrm{H}_1, \mathrm{H}_3\right)\\ &\quad - \Pr\left(D|\mathrm{H}_2, \mathrm{H}_3\right)\Pr\left(\mathrm{H}_2, \mathrm{H}_3\right) + \Pr\left(D|\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\right)\Pr\left(\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\right).\end{aligned}$$

Thus, in principle, it is possible to compute genuine probabilities instead of weights.

5. Conclusion and future work

We presented a novel method to select the "right" copula given a data set. This method, built on a straightforward application of Bayesian analysis, provides interesting advantages over the commonly used statistical tests: it has a simple interpretation, it is independent of parameter choices, it is easy to implement numerically, and, as judged from our simulations, provides reliable identification, even for small samples. Also, it has the conceptual advantage of being

Table A.1
Number of times the right copula is identified compared with the number of times the independent copula is identified for $\tau = 0.2$ over 1000 trials

Copula	Right copula			Independent		
n	30	300	600	30	300	600
Clayton	311	865	951	451	0	0
AMH	19	495	734	553	0	0
Gumbel	266	849	954	495	1	0
Frank	197	483	618	547	1	0
FGM	0	437	635	585	2	0
Gauss	137	575	760	504	1	0

Table A.2
Density of the copulas used in this paper

Copula	$c(u, v\|\theta) = \dfrac{\partial^2 C(u, v\|\theta)}{\partial u \partial v}$
Clayton	$(1+\theta)u^{-1-\theta}v^{-1-\theta}\left(-1+u^{-\theta}+v^{-\theta}\right)^{-2-1/\theta}$
AMH	$\dfrac{\left[-1+\theta^2(-1+u+v-uv)-\theta(-2+u+v+uv)\right]}{[-1+\theta(-1+u)(-1+v)]^3}$
Gumbel	$[-\log(u)]^{-1+\theta}\left\{-1+\theta+[-\log(u)]^\theta+[-\log(v)]^\theta\right\}^{1/\theta}\left\{[-\log(u)]^\theta+[-\log(v)]^\theta\right\}^{-2+1/\theta}[-\log(v)]^{-1+\theta} \times \dfrac{1}{\exp\left\{[-\log(u)]^\theta+[-\log(v)]^\theta\right\}^{1/\theta}uv}$
Frank	$\dfrac{\theta\exp[\theta(1+u+v)][-1+\exp(\theta)]}{\{\exp(\theta)-\exp[\theta(1+u)]-\exp[\theta(1+v)]+\exp[\theta(u+v)]\}^2}$
Joe	$(1-u)^{-1+\theta}\left\{\theta-\left[-1+(1-u)^\theta\right]\left[-1+(1-v)^\theta\right]\right\}\left[(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta\right]^{-2+1/\theta}(1-v)^{-1+\theta}$
A12	$\left(-1+\frac{1}{u}\right)^\theta\left\{-1+\theta+(\theta+1)\left[\left(-1+\frac{1}{u}\right)^\theta+\left(-1+\frac{1}{v}\right)^\theta\right]^{1/\theta}\right\}\dfrac{\left[\left(-1+\frac{1}{u}\right)^\theta+\left(-1+\frac{1}{v}\right)^\theta\right]^{-2+1/\theta}\left(-1+\frac{1}{v}\right)^\theta}{uv(-1+u)(-1+v)\left\{1+\left[\left(-1+\frac{1}{u}\right)^\theta+\left(-1+\frac{1}{v}\right)^\theta\right]^{1/\theta}\right\}^3}$
A14	$\left(-1+u^{-1/\theta}\right)^\theta\left(-1+v^{-1/\theta}\right)^\theta\left[\left(-1+u^{-1/\theta}\right)^\theta+\left(-1+v^{-1/\theta}\right)^\theta\right]^{-2+1/\theta}\left\{1+\left[\left(-1+u^{-1/\theta}\right)^\theta+\left(-1+v^{-1/\theta}\right)^\theta\right]^{1/\theta}\right\}^{-2-\theta} \times \dfrac{\left\{-1+\theta+2\theta\left[\left(-1+u^{-1/\theta}\right)^\theta+\left(-1+v^{-1/\theta}\right)^\theta\right]^{1/\theta}\right\}}{\theta uv\left(-1+u^{1/\theta}\right)\left(-1+v^{1/\theta}\right)}$
FGM	$1+\theta(1-2u)(1-2v)$
Gauss	$\dfrac{1}{\sqrt{1-\rho^2}}\exp\left(\dfrac{\phi^{-1}(u)^2+\phi^{-1}(v)^2}{2}\right)\exp\left(\dfrac{2\rho\phi^{-1}(u)\phi^{-1}(v)-\phi^{-1}(u)^2-\phi^{-1}(v)^2}{2(1-\rho^2)}\right)$

a genuine model selection method, in the sense that it does not depend on the choice of an optimal parameter. The framework allows naturally for copulas with any number of parameters and for higher dimensional copulas. Furthermore, it can be applied to any copula, as long as the copula density and Kendall's tau can be computed numerically.

Future work will concern the following topics: theoretical Bayesian framework to demonstrate the convergence of the method, over-parameterized copula to model the "unknown copula" hypothesis, higher dimensional copulas and the development of a similar method to select margins.

Acknowledgements

Codes are available upon request. Thanks to Peter Perkins (The Mathworks) for his useful set of copula related functions and to Christian Genest and Alain Mailhot for helpful comments and suggestions. The authors are indebted to Eva-Maria Restle for her critical review of the paper, as well as to both referees for their scrutiny.

Appendix A

The number of times the right copula is identified compared with the number of times the independent copula is identified for $\tau = 0.2$ over 1000 trials is given in Table A.1.

The density of the copulas used in this paper is given in Table A.2.

References

Bagdonavicius, V., Malov, S., Nikulin, M., 1999. Characterizations and semiparametric regression estimation in Archimedean copulas. J. Appl. Statist. Sci. 8 (3), 137–153.

Bouyé, E., Durrleman, A., Nikeghbali, A., Riboulet, G., Roncalli, T., 2000. Copulas for finance—a reading guide and some applications. Technical Report, Groupe de Recherche Opérationnelle, Crédit Lyonnais.

Bretthorst, G.L., 1996. An introduction to model selection using probability theory as logic. In: Heidbreger, G. (Ed.), Maximum Entropy and Bayesian Methods, pp. 1–42.

Chen, X., Fan, Y., 2005. Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. La Revue Canadienne de Statistique 33 (3), 389–414.

Chen, X., Fan, Y., Patton, A., 2003. Simple tests for models of dependence between multiple financial time series, with applications to U.S. equity returns and exchange rates. Discussion Paper 483, Financial Markets Group, International Asset Management.

Cherubini, U., Luciano, E., Vecchiato, W., 2004. Copula Methods in Finance. Wiley Finance Series. Wiley, Chichester.

Deheuvels, P., 1979. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Académie Royale de Belgique, Bulletin de la Classe des Sciences, 5ème Série 65, pp. 274–292.

De Michele, C., Salvadori, G., 2003. A generalized Pareto intensity-duration model of storm rainfall exploiting 2-copulas. J. Geophys. Res. 108.

Dobrić, J., Schmidt, F., 2004. Testing goodness of fit for parametric families of copulas—application to financial data. Seminar of economic and social statistics, University of Cologne.

Durrleman, V., Nikeghbali, A., Roncalli, T., 2000. Which copula is the right one? Working document, Groupe de Recherche Opérationnelle, Crédit Lyonnais.

Embrechts, P., McNeil, A., Straumann, D., 2002. Correlation and dependence in risk management: properties and pitfalls. Risk Management: Value at Risk and Beyond. Cambridge University Press, Cambridge, pp. 176–223.

Embrechts, P., Lindskog, F., McNeil, A., 2003. Modelling dependence with copulas and applications to risk management. Handbook of Heavy Tailed Distributions in Finance. Elsevier, Amsterdam, pp. 329–384.

Evin, G., 2004. Choix de la meilleure famille de copule en hydrologie. Internship Report, École Nationale de la Statistique et de l'Analyse de l'Information.

Favre, A.-C., El Adlouni, S., Perreault, L., Thiémonge, N., Bobée, B., 2004. Multivariate hydrological frequency analysis using copulas. Water Resources Res. 40 (W01101),

Fermanian, J.-D., 2005. Goodness-of-fit tests for copulas. J. Multivariate Anal. 95, 119–152.

Frahm, G., Junker, M., Szimayer, I., 2003. Elliptical copulas: applicability and limitations. Statist. Probab. Lett. 63, 275–286.

Frees, E.W., Valdez, E.A., 1998. Understanding relationships using copulas. North American Actuarial J. 2 (1), 1–25.

Genest, C., MacKay, R.J., 1986. The joy of copulas: bivariate distributions with uniform marginals. Amer. Statist. 40, 280–283.

Genest, C., Rivest, L.-P., 1993. Statistical inference procedures for bivariate Archimedean copulas. J. Amer. Statist. Assoc. 88, 1034–1043.

Genest, C., Verret, F., 2005. Locally most powerful rank tests of independence for copulas models. J. Nonparametric Statist. 17, 521–535.

Genest, C., Quessy, J.-F., Rémillard, B., 2005a. Local efficiency of a Cramér–von Mises test of independence. J. Multivariate Anal., in press.

Genest, C., Quessy, J.-F., Rémillard, B, 2005b. Goodness-of-fit procedures for copula models based on the probability integral transformation. Scand. J. Statist., 32, in press.

Jaynes, E.T., Bretthorst, G.L., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK, New York.

Joe, H., 1997. Multivariate Models and Dependence Concepts. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Juri, A., Wüthrich, M., 2003. Tail dependence from a distributional point of view. Extremes 6 (3), 213–246.

Justel, A., Pena, D., Zamar, R., 1997. A multivariate Kolmogorov–Smirnov test of goodness of fit. Statist. Probab. Lett. 35, 251–259.
Kass, R.E., Wasserman, L., 1996. The selection of prior distributions by formal rules. J. Amer. Statist. Assoc. 91, 1343–1370.
Kendall, M., Stuart, A., 1983. The Advanced Theory of Statistics, fourth ed., vol. 2. Oxford University Press, New York.
Kruskal, W., 1958. Ordinal measures of association. J. Amer. Statist. Assoc. 53, 814–861.
Nelsen, R.B., 1999. An Introduction to Copulas. Lecture Notes in Statistics. Springer, New York.
Pollard, D., 1979. General chi-square goodness-of-fit tests with data-dependent cells. Z. Wahrscheinlichkeitstheorie und verwandte Gebiete 50, 317–331.
Rosenblatt, M., 1952. Remarks on a multivariate transformation. Ann. Math. Statist. 23, 470–472.
Saïd, M., 2004. Méthodes statistiques pour tester la dépendance entre les variables latentes pour des risques concurrents. Ph.D. Thesis, Université Laval.
Schmidt, R., 2002. Tail dependence for elliptically contoured distributions. Math. Meth. Oper. Res. 55 (2), 301–327.
Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Université de Paris 8, pp. 229–231.
Whelan, N., 2004. Sampling from Archimedean copulas. Quantitative Finance 4, 339–352.