Prévisions hydrologiques probabilistes dans un cadre multivarié : quels outils pour assurer fiabilité et cohérence spatio-temporelle ?

Joseph Bellier

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMUNAUTE UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Océan Atmosphère et Hydrologie**

Arrêté ministériel : 25 mai 2018

Présentée par

Joseph BELLIER

Thèse dirigée par **Isabella ZIN**
et codirigée par **Guillaume BONTRON**

préparée au sein de l' **Institut des Géosciences de l'Environnement**
et de l'**École doctorale Terre - Univers- Environnement**

Prévisions hydrologiques probabilistes dans un cadre multivarié : quels outils pour assurer iabilité et cohérence spatio-temporelle ?

Jury composé de :

M. Eric Blayo
Professeur, Université Grenoble Alpes, Grenoble, Président
M. Eric Gaume
Ingénieur général des Ponts, des Eaux et des Forêts, IFSTTAR, Nantes, Rapporteur
M. Massimiliano Zappa
Senior Researcher, WSL, Davos (Suisse), Rapporteur
M. Mathieu Vrac
Directeur de recherches CNRS, IPSL/LSCE, Paris, Examinateur
Mme. Maria-Helena Ramos
Chargée de Recherche, IRSTEA, Antony, Invité
Mme. Isabella Zin
Maître de conférences, Université Grenoble Alpes, Grenoble, Directeur de thèse
M. Guillaume Bontron
Ingénieur R&D, Compagnie Nationale du Rhône, Lyon, Co-Directeur de thèse

Remerciements

C'est paradoxalement au début du mémoire de thèse que prennent place les remerciements, alors que ce sont les derniers mots auxquels on réfléchit, une fois le précieux sésame en poche. Et avec eux, le risque d'oublier certaines personnes qui ont pu participer, d'une manière ou d'une autre, à ce travail de thèse. Qu'elles m'en excusent.

Je tiens tout d'abord à remercier chaleureusement mes deux directeurs de thèse, Isabella Zin et Guillaume Bontron, avec qui j'ai eu un grand plaisir à travailler. Bénéficier d'un œil avisé sur le monde académique et de l'autre sur le monde opérationnel fut pour moi très enrichissant. Isabella et Guillaume m'ont facilité la tâche en étant (souvent) d'accord sur les directions à suivre, tout en me laissant explorer les sujets qui me tenaient à cœur, malgré parfois leur éloignement des problématiques opérationnelles. Je les remercie pour leurs encouragements permanents, ainsi que pour leurs critiques systématiquement constructives. Finalement, l'art de toujours placer au bon endroit le délicat curseur du « *Demander plus pour produire plus* ».

Je voudrais ensuite remercier les membres de mon jury de thèse, Eric Gaume, Massimiliano Zappa, Eric Blayo, Mathieu Vrac et Maria-Héléna Ramos, d'avoir accepté d'évaluer mon travail, ainsi que pour les questions éclairées posées lors de ma soutenance. Merci également aux membres de mon comité de suivi, Renaud Marty, Matthieu Lafaysse, Charles Obled et Maria-Héléna Ramos, dont les remarques et avis émis lors de nos rencontres se sont révélés précieux.

Si j'ai pu faire cette thèse, c'est grâce au financement du Labex OSUG@2020 et de CNR, que je tiens à remercier. C'est également l'occasion d'adresser toute ma gratitude aux personnes de la gestion administrative et financière de l'IGE (Odette, Carméline, Manon, François, Valérie), dont j'admire la capacité à trouver des solutions aux problèmes rencontrés dans le monde impitoyable de l'administration.

De nombreuses personnes à CNR ont suivi mon travail de près ou de loin, et je tiens ici à les remercier pour les commentaires qu'ils ont pu m'adresser ainsi que pour leurs encouragements. Je pense notamment à Olivier Vannier, Étienne Dommanget (un grand merci pour la correction orthographique du mémoire !), Sébastien Legrand, Stanislas Siblot, Alexandre Falgon, Sabrina Célié, Aurélien Ben Daoud, François-Xavier Cierco et Muriel Haond. Au sein de l'IGE, j'adresse un remerciement particulier à Charles Obled, pour l'intérêt qu'il a porté à mon travail, et de manière plus générale pour son énergie dépensée à transmettre aux jeunes chercheurs ses connaissances en hydrologie, météorologie, glaciologie, ... la liste est longue ! Merci également à toutes les personnes de l'équipe

CYME, et plus généralement de l'IGE, avec qui j'ai pu partager de nombreuses et passionnantes réflexions scientifiques (mais pas que !).

Je remercie par ailleurs Samira Ahrouch, stagiaire dont j'ai eu le plaisir d'assurer l'encadrement. Si cette expérience peut avoir joué un rôle dans sa volonté de faire une thèse, j'en serais ravi.

Impossible enfin de ne pas remercier le socle d'amis qui m'a accompagné durant la thèse. Si j'ai autant apprécié ces trois et quelques années, c'est en grande partie grâce à eux. Je pense aux collègues doctorants de l'ex-LTHE (notamment François, Louise, Aude, Damien) avec qui j'ai échangé bien davantage que des templates Latex, ainsi qu'aux thésards côté glacio (Maria, Julien B l'un, Julien B l'autre, Jordi, Marion D, Lucas, Machi, Cédric, Jai, Marion R., Olivier, Etienne, Ilann), qui n'ont jamais renié (au contraire !) à accueillir un hydrologue dans leur rang. Les peaux de phoques, dégaines, et guidons de VTT encore chauds ne contrediront pas la réputation qu'ont les labos grenoblois d'héberger quantité de montagnards épanouis ! J'adresse également un petit mot à l'équipe de foot du LTHE pour les matchs inter-labo de juin sous 40° à l'UFRAPS, ainsi qu'à l'équipe de Dribble & Goal avec qui j'ai pu semer la terreur (ou pas !) dans la redoutable poule 1 FSGT. Un remerciement également adressé aux rescapés de l'ENSE3 à Grenoble (Damsou, Dago, Cécile, Carole, Minouche) ou à Lyon, Paris et ailleurs, avec qui j'ai passé des moments précieux. Et enfin, merci à la BimBimTeam (Etienne, Maëlle, Rémy et Fabien) pour les beaux raids courus ensemble, la belle parenthèse RIF 2017 dans l'Ardèche, et la grande aventure à venir à la Réunion.

Y por ultimo, muchas gracias au doux accent chilien de cette fin de thèse.

Résumé

Ce mémoire de thèse s'intéresse à la production de prévisions hydrologiques probabilistes à court/moyen terme, dans un contexte impliquant plusieurs bassins aux débits plus ou moins corrélés. Notre cas d'étude réel concerne différents affluents du Haut-Rhône français. Le travail a été mené autour de la mise en place d'une chaîne de prévision combinant des approches ensemblistes, à savoir la prévision d'ensemble météorologique et le multi-modèle hydrologique, avec des méthodes de correction statistique. Les approches ensemblistes permettent de générer de manière dynamique une incertitude propre à chaque situation, tandis que les corrections statistiques, appliquées sur les prévisions météorologiques (pré-traitement) et/ou hydrologiques (post-traitement), sont nécessaires pour garantir la fiabilité.

Chaque correction statistique, réalisée dans un cadre univarié, entraine la perte de la structure de dépendance spatiale et temporelle des prévisions. Nous nous sommes donc intéressés à son étape de reconstruction, en réalisant un diagnostic des méthodes existantes, notamment le Schaake shuffle et l'ECC. Des adaptations ont été proposées afin d'apporter une réponse aux limites constatées. Dans le cadre du pré-traitement, nous avons cherché à améliorer le conditionnement de la structure de dépendance à la situation météorologique. Pour le post-traitement, notre effort s'est porté sur le respect de l'autocorrélation des débits et le maintien de la fiabilité, notamment lors des phases problématiques de récession. La vérification des prévisions obtenues (météorologiques et hydrologiques) a été menée à l'aide d'outils univariés et multivariés, en portant une attention particulière à la fiabilité, grâce notamment au concept de stratification.

Nous avons enfin étudié les interactions entre les différents maillons de notre chaîne de prévision, en comparant plusieurs scénarios où certains maillons seulement étaient activés. Cette expérience a permis de fournir des indications concrètes sur les priorités à mettre en œuvre lors du déploiement ou de l'amélioration d'une chaîne opérationnelle de prévision hydrologique probabiliste.

Abstract

This dissertation adresses the production of short-to-medium range hydrological forecasts, in a context involving a number of basins with correlated streamflows. Our case study, based on real data, includes several tributaries of the upper Rhone river in France. Work has been conducted on implementing a forecasting chain that combines ensemble approaches, namely meteorological ensemble forecasting and hydrological multi-model, with statistical correction methods. Ensemble methods are able to dynamically generate an uncertainty that is case-specific, while statistical corrections, which are applied to meteorological (pre-processing) and/or hydrological (post-processing) forecasts, are needed to ensure forecast calibration.

Each statistical correction, performed in a univariate framework, induces the loss of the spatial and temporal dependence structure of the forecasts. We were therefore interested in reconstructing such a structure, by making a diagnosic study of existing methods, notably the Schaake shuffle and ECC. Adaptations were proposed in order to address the identified caveats. For pre-processing, we aimed at improving the conditioning of the dependence structure on the meteorological pattern. For post-processing, our effort focused on ensuring that streamflow forecasts respect the autocorrelation charateristics and preserve calibration, especially during the recession phases, which are problematic. Verification of the so-obtained (meteorological and/or hydrological) forecasts was conducted using univariate and multivariate tools, paying particular attention to calibration, using notably the concept of stratification.

Finally, we studied the interactions between the different modules of our forecasting chain, by comparing scenarios where only some of the modules were activated. This experiment allowed us to provide guidelines relative to the implementation or the upgrade of an operational probabilistic streamflow forecasting chain.

Table des matières

Avant-propos

Prévoir ce qui va se passer demain en fonction de ce que l'on sait aujourd'hui n'est pas une tâche aisée. Et celui qui s'y risque se place dans une position délicate, car sa prévision sera confrontée tôt ou tard à l'observation.

En hydrologie, prévoir signifie anticiper le débit qui s'écoulera dans un tronçon de rivière en un instant donné. Si la prévision hydrologique est souvent mise sur le devant de la scène lors d'épisodes de crues, ce n'est pas sa seule utilité. Par exemple, Compagnie Nationale du Rhône (CNR), qui assure l'exploitation hydroélectrique et l'aménagement du Rhône, a besoin de prévisions hydrologiques pour prévoir la production hydroélectrique mais également l'optimiser.

CNR s'est doté depuis 2002 d'une chaîne de prévision opérationnelle qui produit des prévisions de débit sur un ensemble de bassins versants du Rhône et de ses affluents, en alimentant des modèles hydrologiques à partir de prévisions météorologiques (précipitations et températures). Cette chaîne fonctionne de manière déterministe, tout en s'autorisant la production de scénarios « alternatifs » dès lors que la situation est particulièrement incertaine. Cependant, un travail est mené depuis peu pour basculer vers une approche probabiliste où l'incertitude de prévision est prise en compte de manière quantitative et systématique. L'objectif est alors d'être en mesure de fournir, en tout point du Rhône et pour diverses échéances, non plus une unique estimation du débit prévu mais une distribution de probabilité.

Même si elle n'est pas triviale, la question de la réalisation de prévisions hydrologiques probabilistes sur un bassin versant fonctionnant avec un régime pluie-débit a déjà été abordée. En revanche, le problème se complexifie dès lors que l'on s'intéresse aux prévisions sur un fleuve collecteur de différents affluents aux débits plus ou moins corrélés. Les distributions de probabilités pour les différents bassins et échéances ne peuvent alors pas êtres considérées indépendamment les unes des autres : il est nécessaire de prendre en compte leur *structure de dépendance* (spatiale et temporelle). En cas de réactions hydrologiques concomitantes sur plusieurs bassins, une modélisation correcte de cette structure de dépendance est primordiale.

Compte-tenu de la complexité des phénomènes physiques et du séquencement de leur modélisation, un traitement analytique des incertitudes est bien souvent impossible. Ainsi les distributions de probabilités sont communément représentées par un nombre fini de simulations dont la dispersion rend compte de l'incertitude : c'est l'approche *ensembliste*. Les prévisions d'ensemble, constitués de multiples scénarios météorologiques obtenus en perturbant légèrement l'état initial (et parfois la modélisation) d'un modèle atmosphé-

rique, sont l'exemple le plus connu. Leur utilisation en entrée des modèles hydrologiques s'est largement démocratisée dans le domaine de la prévision probabiliste des débits, notamment grâce à l'initiative de la communauté HEPEX (*Hydrologic Ensemble Prediction EXperiment*[1]), qui regroupe scientifiques, prévisionnistes et utilisateurs autour de la question de la prévision hydrologique d'ensemble.

Les approches ensemblistes et notamment les prévisions d'ensemble présentent un atout de taille : elles produisent par construction des prévisions dont la structure de dépendance est *cohérente* à l'échelle de plusieurs bassins, échéances et variables météorologiques. En d'autres termes, la structure de dépendance entre les différents membres de l'ensemble est automatiquement représentative de la situation en cours.

Cependant, des études récentes ont pointé la difficulté des chaînes de prévision basées uniquement sur des approches ensemblistes à rendre compte de l'incertitude complète de prévision. Ainsi, il est nécessaire d'avoir recours à des méthodes de correction statistique, afin de garantir une dispersion suffisante et corriger les biais systématiques. Cette correction statistique peut être réalisée sur les prévisions météorologiques et/ou les prévisions de débits.

Dans ce travail de thèse, nous cherchons donc à mettre en place une démarche permettant de combiner les approches ensemblistes avec celles de correction statistique, dans un contexte de prévision impliquant de multiples bassins et échéances.

Ce mémoire est organisé en trois parties :

La première partie est consacrée à la présentation du contexte dans lequel s'inscrit notre travail, des données et des outils utilisés par la suite. Au cours du chapitre 1, nous précisons pas à pas les caractéristiques d'une prévision hydrologique probabiliste dans un contexte tel que celui de CNR. Cela nous permettra, à l'issue de ce chapitre, de définir plus précisément les objectifs de la thèse. Le chapitre 2 présente les bassins choisis pour constituer notre zone d'étude, ainsi que les archives d'observations et de forçages météorologiques qui seront utilisées. Le chapitre 3 concerne les outils de modélisation hydrologique. Enfin, le chapitre 4 porte sur les méthodes d'évaluation de prévisions probabilistes. Nous présentons alors les outils utilisés dans un cadre univarié, avant de développer, sous la forme d'un article, le concept de stratification. Nous concluons par un aspect crucial : l'évaluation des prévisions dans un cadre multivarié.

La deuxième partie de cette thèse se concentre sur les forçages météorologiques. Ainsi, dans le chapitre 5 nous réalisons un diagnostic des déficiences des forçages bruts, puis appliquons une correction statistique. Cette correction s'appelle le « pré-traitement », car elle est préalable à la modélisation hydrologique. Réalisé dans un cadre univarié, le prétraitement requiert une étape postérieure de reconstruction de scénarios météorologiques cohérents à l'échelle de plusieurs bassins et échéances de prévision. C'est l'objet du chapitre 6, rédigé sous la forme d'un article.

Après le travail sur les forçages météorologiques, la troisième partie se focalise sur les prévisions de débit qui en découlent. Le chapitre 7 s'intéresse à la prise en compte

1. `https://hepex.irstea.fr/about-hepex/`.

de l'incertitude de modélisation hydrologique, en empruntant deux approches : le multi-modèle hydrologique et la correction statistique, appelée désormais « post-traitement ». De nouveau, le post-traitement est réalisé dans un cadre univarié, ce qui nécessite la reconstruction de scénarios hydrologiques cohérents. C'est l'objet du chapitre 8, rédigé sous la forme d'un article. Le travail mené jusqu'alors aura permis, pour chacun des maillons de la chaîne de prévision, d'identifier parmi celles étudiées la méthode qui mène aux meilleurs résultats. Ces maillons sont-ils pour autant tous nécessaires ? Pour clore ce travail de thèse, le chapitre 9 se propose de hiérarchiser leurs apports respectifs dans une chaine de prévisions ensembliste.

Première partie

Contexte, données et outils

Chapitre 1

Contexte et problématique

La prévision probabiliste des débits sur un fleuve collecteur d'affluents implique des problématiques variées, que nous proposons d'expliciter dans ce premier chapitre. Dans un premier temps, les notions essentielles qui se cachent derrière la « *prévision hydrologique opérationnelle* » sont introduites. Nous discutons ensuite des différentes sources d'incertitudes et listons les approches proposées dans la littérature pour permettre leur prise en compte, avant de présenter le principe de correction statistique. Nous abordons enfin un aspect dont la compréhension sera capitale pour la lecture du mémoire : le caractère *multivarié* des prévisions. A la fin de ce chapitre, nous serons en mesure de définir plus précisément les objectifs de notre travail de thèse.

1.1 La prévision hydrologique opérationnelle

1.1.1 Qu'est-ce c'est ?

De manière générale, émettre une prévision consiste à prévoir la valeur que va prendre une certaine variable à un instant donné dans le futur, en se basant sur notre connaissance de l'état initial ainsi que sur son processus d'évolution. L'intervalle de temps entre cet instant futur et le moment où la prévision est émise s'appelle l'*échéance* (de prévision). Dans cette thèse consacrée à la prévision hydrologique, la variable qui nous intéresse est le débit, exprimé en m^3/s, qui s'écoule dans une section de rivière donnée.

Classiquement, une prévision hydrologique concernera plusieurs échéances successives, de manière à prévoir non pas une unique valeur mais une série temporelle, à un pas de temps qui correspond à la différence entre deux échéances successives. Par ailleurs, nous définissons l'*horizon* (de prévision) comme l'échéance maximale de la prévision. L'hydrogramme est la manière la plus fréquente de représenter une prévision hydrologique (Figure 1.1).

On parle de prévision *opérationnelle* lorsque la production se fait en temps réel et sur une base régulière (hebdomadaire, quotidienne, etc.), en satisfaisant un certain nombre de contraintes comme un temps de calcul limité, ou encore la possibilité d'utiliser des données d'entrée en mode dégradé. La robustesse de la chaîne est alors une qualité requise (Pagano *et al.*, 2014).

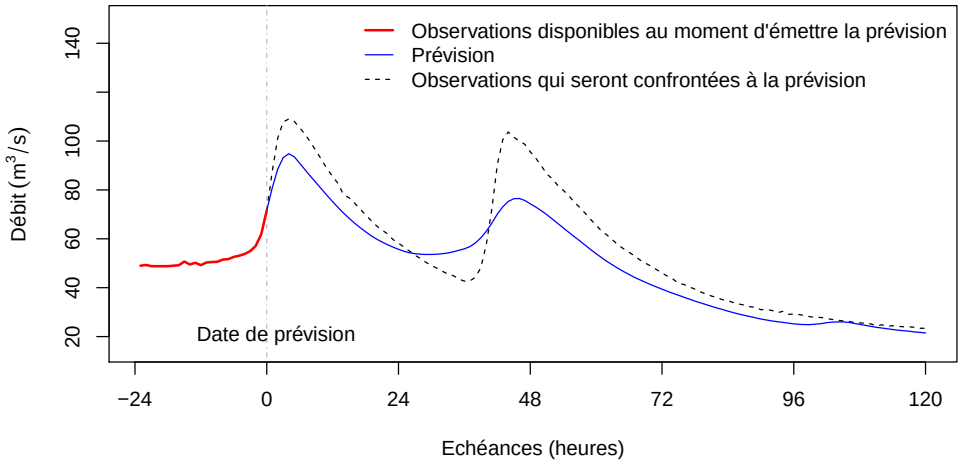

FIGURE 1.1 – Exemple d'une prévision hydrologique déterministe, pour un horizon de prévision de 120 h. Dans cet exemple, la prévision part d'un débit observé connu et disponible, ce qui n'est pas toujours le cas.

1.1.2 Pour quels usages ?

En hydrologie, les prévisions sont couramment utilisées pour la gestion de la ressource en eau et l'anticipation d'événements remarquables comme les crues ou les étiages. En France, la prévision des crues sur les grandes rivières est confiée à des organismes publics, les services de prévision des crues (SPC), qui sont appuyés par un service de coordination, le service central d'hydrométéorologie et d'appui à la prévision des inondations (SCHAPI). De ces prévisions découlent des actions prises par les pouvoirs publics, allant de la mise en vigilance de tronçons de rivières jusqu'à l'évacuation de zones inondables. D'autres organismes publics peuvent également être mandatés pour la prévision des étiages, desquels peuvent découler par exemple des mesures préventives sur les usages de l'eau.

Pour des gestionnaires d'aménagements hydroélectriques comme CNR, prévoir les crues permet de faciliter l'exploitation des ouvrages en anticipant certains contrôles, en s'assurant de la disponibilité des ressources nécessaires et en anticipant leur déploiement sur le terrain. Par ailleurs, la prévision et la gestion des étiages sur un fleuve comme le Rhône doit permettre de répondre à de forts enjeux que sont les prélèvements pour l'irrigation et le refroidissement des centrales nucléaires.

En situation courante (dite « énergétique »), les prévisions hydrologiques sont particulièrement utiles pour CNR. Elles permettent d'optimiser la production d'hydroélectricité en turbinant prioritairement sur certains créneaux (pour les aménagements avec capacité de stockage), d'assurer les obligations vis-à-vis de l'équilibrage du réseau électrique, et enfin de maximiser la vente sur les marchés de la production d'électricité.

1.1.3 Schéma général d'une chaîne de prévision

Une chaîne de prévision hydrologique sur un bassin fonctionnant avec un régime pluie-débit s'appuie sur deux maillons principaux : le forçage météorologique et le modèle hydrologique. Le premier alimente le second, qui est alors chargé de simuler les débits à

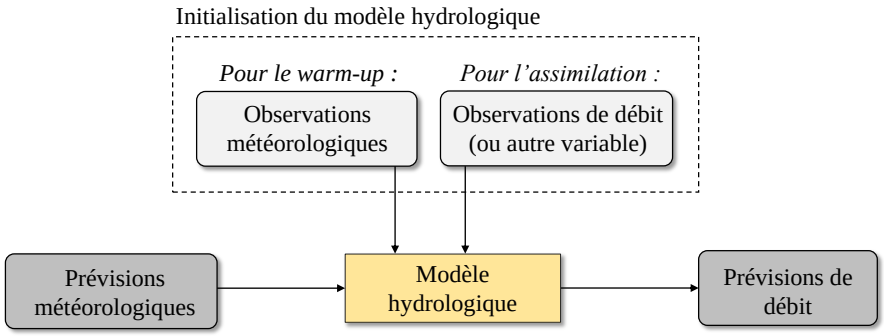

FIGURE 1.2 – Illustration schématique d'une chaîne de prévision hydrologique sur un bassin fonctionnant avec un régime pluie-débit.

l'exutoire en reproduisant le comportement hydrologique du bassin versant.

Le forçage météorologique. Dans un contexte de prévision opérationnelle, le forçage se limite souvent à deux variables : la précipitation et la température. Dès lors que l'horizon de prévision souhaité dépasse le temps de concentration du bassin, il n'est plus possible d'utiliser des observations, et par conséquent le forçage doit obligatoirement être une prévision météorologique. Dans ce mémoire, nous parlerons donc indistinctement de forçage météorologique ou de prévisions météorologiques. Classiquement, ces prévisions sont issues de modèles de prévision numérique du temps, qui sont développés et exécutés au sein de centres de prévisions météorologiques [1]. Le terme de *modèles météorologiques* sera utilisé à propos des modèles de prévision numérique du temps. Les systèmes opérationnels de prévision hydrologique mettent alors en place des accords avec ces centres pour recevoir, à des fréquences fixes, les forçages issus des modèles météorologiques.

Le modèle hydrologique. Il se présente sous la forme d'équations qui cherchent à représenter, de manière plus ou moins simplifiée, certains processus impliqués dans la génération des débits. Ces équations mettent en relation les forçages météorologiques et les débit par l'intermédiaire de *variables d'état* et de *paramètres*. Les variables d'état sont des grandeurs internes au modèle qui caractérisent un état du bassin (le taux d'humidité dans le sol ou la hauteur de neige, par exemple) ; elles varient donc au cours du temps. Les paramètres sont des grandeurs qui qualifient une caractéristique intrinsèque du bassin (son temps de concentration par exemple) ; par conséquent ils sont invariables dans le temps, et déterminés par des mesures ou, plus souvent, lors de la phase de calage. Un modèle hydrologique conçu à des fins de compréhension scientifique (du fonctionnement d'un bassin versant) présentera des variables d'état et des paramètres que l'on peut interpréter et relier à des processus physiques. En revanche, si l'objectif de modélisation est simplement de rendre compte des débits en sortie, alors leur interprétation physique n'est pas primordiale.

1. Nous ne parlerons pas ici de l'utilisation de modèles couplés atmosphère-hydrologie, qui est encore marginale dans le monde de la prévision hydrologique opérationnelle.

La modélisation hydrologique dans un contexte de prévision présente une particularité : il est nécessaire d'initialiser les variables d'état au moment de lancer le modèle hydrologique. Cette initialisation est généralement constituée de deux phases : le « *warm-up* » et l'*assimilation* (Figure 1.2).

La phase de warm-up consiste à exécuter le modèle sur une période qui précède la date de prévision, en utilisant comme forçage les observations météorologiques. Cela permet aux variables d'états d'être, au moment de lancer la prévision, représentatives de l'état initial du bassin versant. La période de warm-up doit être suffisamment longue (de plusieurs mois à une année hydrologique complète) pour que les variables d'état en fin de simulation soient indépendantes de celles utilisées au début.

Le warm-up permet donc d'aboutir à une représentation correcte de l'état initial du bassin, du moins tel que le modèle hydrologique le voit. Or on dispose souvent, au moment de lancer la prévision, de données observées récentes qui permettraient de compléter notre connaissance de l'état initial. Cela afin que la prévision produite ultérieurement par le modèle soit plus juste. La phase d'assimilation utilise (« assimile ») ces données observées pour mettre à jour l'état initial du modèle. Techniquement, cela consiste à ajuster la valeur de certaines variables d'état à la date d'émission de la prévision. Les derniers débits observés constituent un exemple de données régulièrement assimilées. Cela permet alors de s'assurer qu'il n'y ait pas de discontinuité entre les derniers débits observés et les premiers débits prévus, et donc de réduire les erreurs de prévision sur les premières échéances.

Dans cette thèse, nous considérons que la chaîne de prévision hydrologique s'arrête à la prévision des débits. Ainsi, nous ne nous intéresserons pas à d'autres maillons à l'aval de la chaîne qui pourraient exploiter ces prévisions de débit (par exemple, un modèle hydraulique pour prévoir les hauteurs d'eau).

1.1.4 Les différents horizons de prévision

L'horizon d'une prévision hydrologique est directement lié à celui de la prévision météorologique qui viendra forcer le modèle hydrologique. Ces horizons sont différents selon l'usage que l'on souhaite faire des prévisions, tout en gardant à l'esprit que l'on ne peut pas attendre les mêmes performances d'une prévision à un horizon de quelques heures que d'une prévision pour les mois à venir. En météorologie, et plus particulièrement pour la prévision des précipitations, nous distinguons :

- la *prévision immédiate* (on utilise parfois le terme anglais, *nowcasting*) : de 0 à 3 heures. Celle-ci ne fait généralement pas appel à la modélisation atmosphérique, mais plutôt à des méthodes exploitant l'imagerie radar, qui est capable de décrire de façon quasi immédiate les caractéristiques des cellules pluvieuses existantes. Grâce à des techniques d'advection, on peut alors extrapoler l'évolution spatiale et temporelle de ces cellules.

- la *prévision à courte échéance* : jusqu'à 3 jours. L'objectif est de simuler l'évolution de phénomènes à l'échelle synoptique (environ 1000 km) tels que les perturbations

frontales ou les systèmes convectifs de méso-échelle. Ces échéances de prévision font de plus en plus appel à des modèles météorologiques à aire limitée et à haute résolution (quelques km), capables de simuler les phénomènes convectifs.

– la *prévision à moyenne échéance* : de 3 à 15 jours. On cherche alors à simuler, grâce à des modèles globaux, la circulation atmosphérique à l'échelle supra-synoptique, l'objectif étant d'anticiper l'arrivée de dépressions ou d'anticyclones. À ces échéances, le caractère chaotique de l'atmosphère réduit fortement sa prédictibilité. La prévision déterministe n'étant alors plus guère performante, c'est le domaine des prévisions d'ensemble.

– la *prévision saisonnière* : jusqu'à plusieurs mois. Les modèles météorologiques ont atteint leur limite de prédictibilité, et par conséquent on cherche seulement à anticiper des anomalies par rapport à la normale climatologique, à l'aide de phénomènes basse fréquence (NAO ou ENSO [2] par exemple) détectés en des endroits du globe pouvant avoir une influence sur les conditions météorologiques du lieu concerné par la prévision.

Dans cette thèse, nous nous intéressons à la prévision des débits à un horizon de 5 jours (120 heures), ce qui correspond, dans le langage météorologique, à la prévision à courte/moyenne échéance.

1.1.5 Du déterministe au probabiliste

Une prévision déterministe correspond à la meilleure estimation possible de la variable à prévoir. C'est la prévision fournie par le prévisionniste qui utilisera les meilleures données d'entrée et outils à sa disposition puis, connaissant les limites de ces données/outils, ajustera sa prévision grâce à son expertise. Cependant, malgré les avancées scientifiques (meilleure compréhension des phénomènes physiques), techniques (outils informatiques plus puissants) et instrumentales (meilleure connaissance des états initiaux), la prévision déterministe est vouée à se tromper. En effet, ces avancées ne seront jamais suffisantes pour reproduire de manière exacte le comportement des systèmes que l'on cherche à modéliser (l'atmosphère, le bassin versant).

L'utilisateur d'une prévision déterministe, lorsqu'il prend une décision en se basant sur cette information, prend en compte consciemment ou inconsciemment le fait que la prévision est incertaine.

Prenons l'exemple d'un utilisateur qui subit des dommages entrainant une perte $\mathcal{L}$ dès lors que le débit d'un cours d'eau dépasse un certain seuil, par exemple 100 m^3/s. Il dispose néanmoins de la possibilité de se protéger en mettant en place des actions de protection, moyennant un coût $\mathcal{C} < \mathcal{L}$. Supposons qu'un évènement hydrologique se profile, et qu'il ait accès à la prévision déterministe représentée à gauche de la Figure 1.3. Il sait que le système produisant ces prévisions n'est pas parfait, mais il n'a pas une connaissance quantitative des erreurs. La décision est alors particulièrement délicate à prendre, car

2. NAO : *North Atlantic oscillation* ; ENSO : *El Niño–Southern oscillation.*

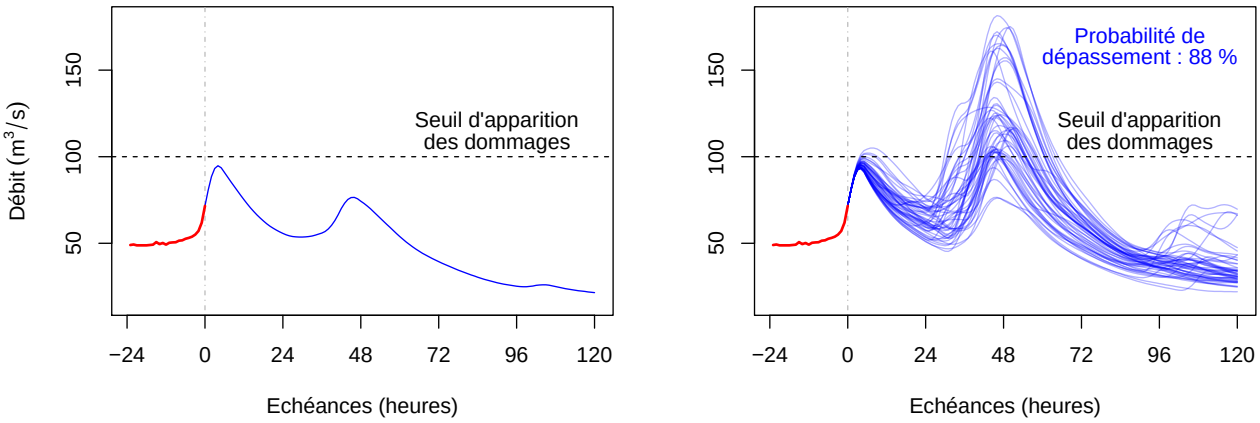

FIGURE 1.3 – Exemple d'une prévision déterministe (à gauche) et probabiliste (à droite), pour un même évènement.

d'un côté la prévision annonce que le débit maximal restera sous les 100 m^3/s, mais de l'autre il s'en approche dangereusement. Avec seulement cette information à disposition, l'utilisateur ne pourra pas prendre une décision de manière tout à fait rationnelle.

Imaginons maintenant qu'il ait accès à la prévision représentée à droite de la Figure 1.3, qui est composée d'un ensemble de 50 traces correspondantes à des scénarios équiprobables. L'utilisateur peut alors calculer la probabilité de dépassement du seuil : 88 %. Grâce à cette information ainsi qu'au ratio $\mathcal{C}/\mathcal{L}$, l'utilisateur pourra prendre rationnellement la meilleure décision, à savoir celle qui, sur le long terme (c'est-à-dire sur un grand nombre de cas), minimisera ses pertes. Nous comprenons, à travers ce bref exemple, l'intérêt d'estimer de manière quantitative les incertitudes de la prévision hydrologique.

Sur le plan théorique, les raisons pour émettre des prévisions non plus déterministes mais probabilistes sont multiples. Krzysztofowicz (2001) les résume ainsi :

- Cela fait preuve d'honnêteté, car on admet alors l'incertitude qui se développe dans notre chaîne de prévision,

- On fournit au décideur une information quantitative qui lui permet de prendre des décisions plus *rationnelles*. Des travaux pour le démontrer ont notamment été menés par Ramos *et al.* (2013a).

- Par ailleurs, cela aide à séparer la responsabilité du prévisionniste, qui s'appuie uniquement sur la science pour produire sa prévision, de celle du décideur, qui doit nécessairement prendre en compte des aspects autres. Ainsi, on ne peut en théorie jamais reprocher au prévisionniste de « manquer » sa prévision, car ce dernier fournit seulement des probabilités.

- Cela augmente la valeur économique des prévisions. Des études l'ont démontré pour les prévisions météorologiques (Buizza, 2008), mais également dans des contextes hydrologiques, à la fois pour la prévision des crues (Roulin, 2007; Verkade et Werner, 2011), la production d'hydro-électricité (Boucher *et al.*, 2012; Zalachori, 2013), ou

encore la gestion de réservoirs pour l'alimentation en eau et l'irrigation (Anghileri *et al.*, 2016).

Pour ces raisons, de nombreuses chaînes opérationnelles de prévision hydrologique produisent désormais des prévisions probabilistes. Thielen *et al.* (2009), Addor *et al.* (2011), Demargne *et al.* (2014) ou encore Bennett *et al.* (2014) décrivent de telles chaînes dans leur globalité, tandis que Cloke et Pappenberger (2009) proposent une revue de littérature. Le lecteur aura cependant l'occasion de se rendre compte, au travers des nombreuses références citées tout au long de ce mémoire, que l'adoption du cadre probabiliste dans la prévision hydrologique opérationnelle va bien au delà des quelques références susmentionnées.

1.2 Prise en compte des incertitudes de prévision

Produire des prévisions probabilistes implique la prise en compte quantitative et systématique des incertitudes de prévision. L'incertitude « finale » d'une prévision hydrologique, c'est-à-dire l'incertitude associée à la variable en sortie de la chaîne de prévision, peut être construite à posteriori en « habillant » la prévision déterministe d'une distribution de probabilité construite à partir des statistiques de ses erreurs passées. Cette approche, peu coûteuse, génère cependant des prévisions probabilistes dont la dispersion est indépendante de la situation hydrométéorologique. Or, certaines situations sont plus difficiles à prévoir que d'autres, et par conséquent il serait dommage de proposer une fourchette d'incertitude qui soit identique à chaque nouvelle prévision.

Une prévision hydrologique étant le résultat d'une chaîne de modélisation impliquant plusieurs maillons, l'approche communément adoptée par les systèmes de prévision consiste à identifier les différentes sources d'incertitudes, de manière à pouvoir les traiter de manière spécifique.

La majorité des études ayant cherché à quantifier l'incertitude dans la prévision hydrologique s'accordent sur les trois sources suivantes : le forçage météorologique, la modélisation hydrologique, et enfin les conditions initiales du bassin (Zappa *et al.*, 2011; Demirel *et al.*, 2013; Thiboult *et al.*, 2016). Dans cette section, nous discutons de chacune de ces sources, et mentionnons diverses approches qui ont été proposées pour les prendre en compte.

Quelle que soit la source d'incertitude visée, ces approches ont en commun l'idée de générer un ensemble de scénarios dont la dispersion représente l'incertitude : c'est la stratégie *ensembliste*. Cette stratégie s'impose du fait que les différentes étapes de modélisation se basent sur des modèles mathématiques complexes qui rendent impossible la formulation analytique des incertitudes.

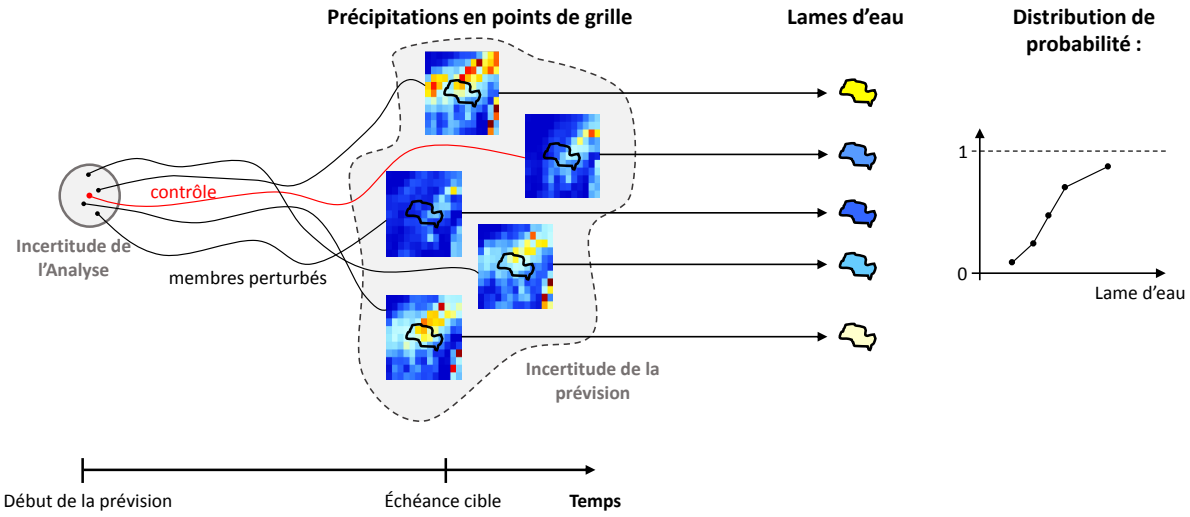

FIGURE 1.4 – Illustration du principe de la prévision d'ensemble, pour la prévision probabiliste de lames d'eau à une échéance donnée. Sur ce graphique, seule la perturbation de l'analyse est illustrée (mais pas une perturbation éventuelle du modèle).

1.2.1 Incertitudes sur les forçages météorologiques

1.2.1.1 La prévision d'ensemble météorologique, une incertitude « dynamique »

La stratégie la plus fréquemment adoptée pour introduire de l'incertitude sur le forçage météorologique est l'utilisation de *prévisions d'ensemble*, qui sont produites par les centres météorologiques. Le principe est le suivant.

L'atmosphère est un système dynamique fortement non linéaire, dont l'évolution dans le temps est extrêmement dépendante des conditions initiales ; on dit qu'elle est « chaotique » (Lorenz, 1963). Seulement, l'estimation de cet état initial dans les modèles météorologiques, appelé *analyse*, est incertaine. Certaines des erreurs sur l'analyse, aussi faibles soient-elles, peuvent alors s'amplifier avec les échéances, induisant ainsi des erreurs importantes sur la prévision. L'idée de la prévision d'ensemble est de générer plusieurs *membres* (i.e., scénarios), chacun correspondant à une simulation du modèle météorologique à partir d'analyses légèrement différentes mais néanmoins cohérentes au regard de l'incertitude sur notre moyen d'estimer l'état initial de l'atmosphère. Par ailleurs, les méthodes numériques utilisées pour simuler son évolution sont également imparfaites, si bien que la plupart des systèmes de prévision d'ensemble font désormais « varier » le modèle d'un membre à l'autre, de manière à ce qu'une partie de la dispersion de l'ensemble représente l'incertitude de modélisation météorologique.

Généralement, une prévision d'ensemble est constituée d'un ensemble de membres *perturbés* (dans l'analyse et la modélisation donc), ainsi qu'un membre non perturbé appelé *contrôle* (Figure 1.4). Afin de respecter les délais de mise à disposition des prévisions aux utilisateurs, les systèmes de prévision d'ensemble font généralement tourner le modèle météorologique à une résolution moindre que celle utilisée pour la prévision déterministe.

Ainsi, la prévision d'ensemble présente deux avantages principaux. D'une part, elle

permet d'étendre le potentiel informatif de la prévision à des échéances plus lointaines que la prévision déterministe. D'autre part, l'incertitude de prévision qu'elle propose est d'origine *dynamique*, et donc dépend de chaque situation météorologique.

Ce domaine de la météorologie fait l'objet depuis les années 90 d'un très grand nombre de publications dans la littérature. Citons ici les articles de Buizza *et al.* (1999), Leutbecher et Palmer (2008) et Bauer *et al.* (2015), qui permettent de comprendre de manière synthétique l'évolution de la prévision d'ensemble au cours des vingt-cinq dernières années. Nous présenterons dans la section 2.3.1 les différentes approches empruntées par chacun des systèmes de prévision d'ensemble étudié pour perturber l'analyse et la modélisation.

Pour conclure, il faut noter que l'utilisation de prévisions d'ensemble en entrée de la chaîne de prévision hydrologique requiert de s'adapter aux contraintes des modèles hydrologiques, qui bien souvent travaillent à des échelles spatio-temporelles plus fines que la taille de la maille des modèles météorologiques. Nous utiliserons dans cette thèse des modèles hydrologiques globaux (cf. chapitre 3), c'est-à-dire qui représentent le bassin comme une entité hydrologique unique. Ces modèles acceptent donc en entrée des prévisions de précipitations qui sont moyennées sur le bassin (on parle alors de *lame d'eau*). Il est alors nécessaire, dans ce cas, de transformer les prévisions émises en points de grille en prévisions de lame d'eau, à l'aide de méthodes de moyenne spatiale (voir Figure 1.4).

1.2.1.2 La prévision par analogie, une incertitude « statistique »

Nous présentons ici une alternative aux prévisions d'ensemble pour prendre en compte l'incertitude sur le forçage météorologique : la prévision par analogie. Cette approche part des hypothèses suivantes :

- Du fait de leur résolution limitée, les modèles météorologiques ne peuvent pas résoudre explicitement les équations régissant les phénomènes physiques en deçà d'une certaine échelle ; ces phénomènes doivent être modélisés via des schémas de paramétrisation. En conséquence, les effets locaux de variables de surface telles que la précipitation sont souvent mal représentés.

- En revanche, les variables décrivant la situation synoptique sont assez bien modélisées, car issues d'équations résolues explicitement et adaptées à l'échelle de la maille moyenne des modèles.

- Les archives d'observations, elles, représentent une source d'information intéressante sur les effets locaux des variables de surface, tant à l'échelle ponctuelle qu'à l'échelle de bassins de petite taille.

Le principe d'analogie est que deux situations synoptiques analogues devraient produire des effets locaux dont l'espérance et la variabilité sont similaires (Lorenz, 1969). L'idée est alors de construire une relation statistique non paramétrique qui lie les variables synoptiques, appelées *prédicteurs*, à la variable locale que l'on cherche à prévoir, appelée *prédictant*. Cette relation s'applique sur les prédicteurs et le prédictant considérés au même instant ; c'est donc intrinsèquement un modèle statistique d'adaptation et

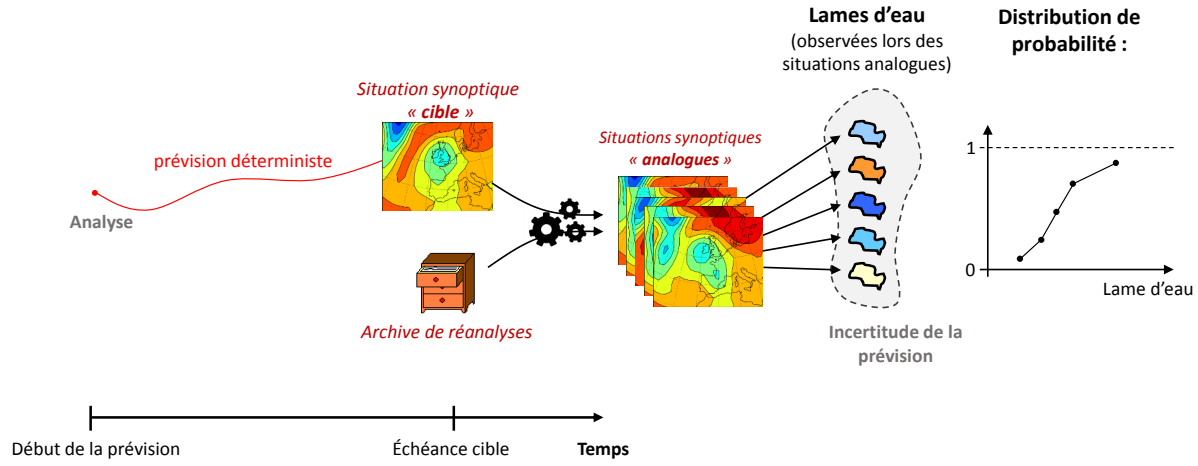

FIGURE 1.5 – Illustration du principe de la prévision par analogie, pour la prévision probabiliste de lames d'eau à une échéance donnée.

non pas de prévision. Dans un contexte de prévision, l'idée consiste alors à utiliser les prédicteurs prévus par un modèle météorologique.

Les différentes étapes, pour une échéance de prévision donnée, sont les suivantes (voir également Figure 1.5) :

1. La situation synoptique « cible » est caractérisée à l'aide des prédicteurs issus de la prévision déterministe d'un modèle météorologique.

2. On cherche ensuite, parmi une archive historique de situations « candidates », celles qui sont les plus similaires (ou *analogues*). Cela requiert la définition des critères d'analogie, ainsi que des domaines spatiaux sur lesquels ces critères sont calculés. Pour constituer l'archive historique des situations candidates, on fait appel aux *réanalyses*. Comme son nom l'indique, les réanalyses sont des analyses (c'est-à-dire des états de l'atmosphère modélisés à partir d'observations assimilées) produites rétrospectivement sur une période donnée, à l'aide d'un même modèle météorologique, et en utilisant des observations qui n'étaient pas forcément disponibles au moment de produire l'analyse opérationnelle.

3. Les valeurs des prédictants ayant été observées lors de ces situations analogues sont relevées, et permettent de construire une distribution de probabilité.

Un modèle de prévision par analogie est donc défini par les prédicteurs utilisés, les domaines spatiaux sur lesquels ils sont considérés, les critères d'analogie, et enfin le nombre d'analogues retenus en sortie. D'autre part, il est associé à un certain modèle déterministe mais également à un certain jeu de réanalyse. Dans cette thèse, nous étudieront les prévisions de précipitations produites à l'aide du modèle utilisé en opérationnel à CNR, qui provient des travaux successifs de Obled *et al.* (2002), Bontron (2004), Marty *et al.* (2012) et Ben Daoud *et al.* (2016). Plus de détails seront donnés dans la section 2.3.2.

Pour conclure, l'incertitude autour de la variable à prévoir dans la prévision par analogie n'est pas construite de manière dynamique, mais statistique. En effet, celle-ci rend

compte de l'incertitude qui règne dans la relation statistique entre prédicteurs et prédictant, du fait à la fois du nombre restreint de prédicteurs utilisés pour caractériser la situation synoptique, de la non-exactitude des réanalyses, ou encore de la profondeur limitée de l'archive historiques.

Des recherches ont été menées pour appliquer la technique d'adaptation par analogie non plus sur une prévision déterministe mais sur chacun des membres d'une prévision d'ensemble (Thévenot, 2004). L'objectif est alors de combiner l'incertitude d'origine dynamique de la prévision d'ensemble avec celle d'origine statistique de prévision par analogie.

1.2.2 Incertitudes sur la modélisation hydrologique

Le modèle hydrologique n'est qu'une représentation simplifiée du bassin versant, et par conséquent les débits qu'il simule sont entachés d'incertitudes (et ce même si l'on devait forcer le modèle avec les données météorologiques observées). De la même manière que pour le modèle météorologique, nous pouvons séparer l'incertitude liée à la modélisation de celle liée à l'état initial. Nous discutons ici de l'incertitude de modélisation, et présentons diverses stratégies ensemblistes qui ont été proposées pour l'introduire dans une chaîne de prévision.

Certaines font reposer l'ensemble des incertitudes de modélisation sur la valeur des paramètres du modèle hydrologique. Par exemple, la méthode GLUE (*generalized likelihood uncertainty estimation* ; Beven et Freer, 2001) rejette l'hypothèse d'un unique jeu de paramètre qui donnerait les meilleures performances de simulation, et propose de garder tous les jeux dont la performance est supérieure à un certain seuil. L'algorithme SCEM (*shuffled complex evolution Metropolis* ; Vrugt *et al.*, 2003) est quant à lui un algorithme d'exploration de l'espace des paramètres (cf. 3.4.4) qui, au lieu de converger vers un unique optimum (c'est-à-dire un unique jeu de paramètres), décrit de manière probabiliste une « zone » optimale. Ces deux approches, dont Blasone *et al.* (2008) ont d'ailleurs proposé une combinaison, permettant de construire une distribution de probabilité des paramètres. Pour transférer l'incertitude sur les débits, il suffit alors de tirer aléatoirement un ensemble de jeux de paramètres, puis de réaliser les simulations à partir de chacun de ces jeux.

Duan *et al.* (2007) utilisent également plusieurs jeux de paramètres, mais obtenus cette fois à partir de critères de calage différents. L'hypothèse est que chaque critère favorise la simulation de certaines phases de l'hydrodrogramme (hauts débits, bas débits, montée au pic, etc.), mais qu'il n'en existe pas un qui donne les meilleures performances sur toutes ces phases simultanément. En créant un ensemble où chaque membre provient d'un jeu de paramètres qui est optimal selon un critère donné, on espère qu'à chaque situation il y ait au moins un membre qui se rapproche de l'observation.

Enfin, d'autres ont cherché à faire reposer l'incertitude de modélisation non plus sur les paramètres, mais sur la structure du modèle. Clark *et al.* (2008) ont évalué un grand nombre de structures différentes de modèle, et montré qu'il n'y en avait aucune qui soit systématiquement meilleure. La stratégie « multi-modèle », qui consiste à créer un ensemble de simulations où chaque membre correspond à un modèle, paraît ainsi légitime. Plusieurs stratégies sont possibles pour constituer cet ensemble de modèles : retenir dif-

férents modèles « complets » parmi ceux proposés dans la littérature (Seiller *et al.*, 2012; Velázquez *et al.*, 2011), ou bien créer des modèles « hybrides » à partir de composants (fonction d'infiltration, de routage, etc.) provenant de plusieurs modèles « parents » (Clark *et al.*, 2015; Seiller *et al.*, 2017). Dans les deux cas, la difficulté consiste à s'assurer de la diversité de cet ensemble de modèle, car beaucoup de modèles ont des structures très proches. Certains ont ainsi cherché à quantifier cette diversité, en construisant des métriques qui évaluent la « distance » entre plusieurs modèles, que ce soit dans un espace théorique des modèles (Abramowitz et Gupta, 2008), ou bien vis-à-vis de leur réponse hydrologique (Seiller *et al.*, 2012).

1.2.3 Incertitudes sur les conditions initiales du bassin

L'état initial du bassin, et donc du modèle hydrologique, est également incertain. D'une part, les données utilisées lors de la procédure d'initialisation (warm-up et assimilation) sont incertaines, et d'autre part le modèle ne fournit qu'une représentation simplifiée de la réalité.

L'assimilation de données, qui permet de mettre à jour les états du modèle de manière à ce qu'ils soient davantage cohérents avec les dernières observations, a fait l'objet de recherches actives ces dernières années. Une revue détaillée a été réalisée par Liu *et al.* (2012). Néanmoins, encore très peu d'études se sont appuyées sur des méthodes ensemblistes permettant de produire un ensemble d'états initiaux du modèle. Nous pouvons citer DeChant et Moradkhani (2011) et Thiboult *et al.* (2016), qui ont utilisé des techniques séquentielles d'assimilation, respectivement le filtre particulaire et le filtre de Kalman d'ensemble.

Ces « filtres » mettent à jours les variables d'état du modèle de manière à ce que l'ensemble des états initiaux en sortie soit cohérent avec l'incertitude des données assimilées. Leur mise en place peut cependant s'avérer piégeuse (Thiboult et Anctil, 2015), et des recherches semblent nécessaires pour proposer des méthodes moins complexes et plus robustes qui puissent être utilisées plus facilement dans un contexte opérationnel. Par exemple, des techniques déterministes d'assimilation des derniers débits observés pourraient être exécutées de manière ensembliste, en perturbant les débits assimilés de manière à tenir compte de l'incertitude de mesure (qui peut être non négligeable, notamment dans les hauts débits).

1.3 Le post-traitement statistique

Si les approches ensemblistes permettent d'introduire de l'incertitude à un endroit spécifique de la chaîne de prévision, elles s'avèrent presque systématiquement insuffisantes pour la représenter entièrement. En d'autres termes, l'observation n'est pas suffisamment bien « capturée » par la dispersion des membres. On parle alors de défaut de *fiabilité*.

Pour illustrer le concept de fiabilité (sur lequel nous reviendrons en détail dans le chapitre 4), revenons à l'exemple illustré à la Figure 1.3, où la prévision annonce une

probabilité de dépassement de seuil de 88 %. S'il s'avère que, sur un grand nombre de prévisions passés où la probabilité était similaire, le seuil n'a été en réalité dépassé que dans 50 % des cas, alors il y a un défaut de fiabilité. La décision que prendra l'utilisateur sur la base de la prévision ne sera alors pas optimale.

Garantir la fiabilité des prévisions est donc crucial, car c'est seulement grâce à elle que l'utilisateur peut exploiter le caractère probabiliste des prévisions. Si ces dernières ne sont pas fiables, alors un post-traitement statistique est nécessaire.

1.3.1 Principe et grandes familles de méthodes

On définit généralement un post-traitement statistique comme n'importe quelle procédure qui corrige à posteriori les prévisions émises par un système en exploitant la relation entre les prévisions passées et leurs observations correspondantes (Wilks, 2018a). Un grand nombre de méthodes ont été développées durant les quinze dernières années. Dans cette section, nous caractérisons les grandes familles de méthodes, en se focalisant sur celles qui s'appliquent à des prévisions qui sont à la fois :

- univariées : la variable à prévoir est un scalaire,
- ensemblistes : la prévision est sous la forme d'un ensemble de valeurs possible, et non pas d'une distribution continue de probabilité.

L'ensemble de ces méthodes visent à créer, à partir de l'ensemble brut, une densité de probabilité qui représente de manière plus juste le comportement (incertain) de l'observation. On peut ainsi différencier les méthodes selon qu'une hypothèse est formulée ou non quant à ce comportement statistique. D'un côté, les méthodes *paramétriques* font appel à une loi de distribution donnée pour décrire le comportement de la variable à prévoir. De l'autre, les méthodes *non paramétriques* ne formulent aucune hypothèse : ce sont uniquement les données (l'archive des prévisions et observations passées) qui vont définir les nouvelles densités.

Méthodes paramétriques : Les deux méthodes les plus populaires sont l'EMOS (*ensemble model output statistics* ; Gneiting *et al.*, 2005) et la BMA (*Bayesian model averaging* ; Raftery *et al.*, 2005). En fait, l'EMOS et la BMA sont les représentants phares de deux familles auxquelles il est possible de rapprocher l'ensemble des méthodes paramétriques : les méthodes dites « de régression », et celles par « habillage des membres » (en anglais : *ensemble dressing*). Les méthodes de régression considèrent que la densité prédictive est décrite par une unique loi de distribution, dont les paramètres sont obtenus par régression à partir de caractéristiques de l'ensemble brut (moyenne, écart-type, etc.). En revanche, les méthodes par habillage des membres proposent une densité prédictive issue d'une combinaison de multiples fonctions de densité, chacune habillant un des membres de l'ensemble. Cette stratégie d'habillage se veut plus flexible, car elle autorise des formes de densités prédictives plus complexes (par exemple des densités multi-modales).

Méthodes non paramétriques : Certaines de ces méthodes visent à ajuster individuellement les membres de l'ensemble brut (*member-by-member post-processing* ; Van Schaeybroeck et Vannitsem, 2014) ou bien les quantiles de la distribution empirique brute (*quantile mapping*). D'autres méthodes cherchent à construire un modèle statistique entre des prédicteurs d'un côté, qui peuvent se limiter aux membres de l'ensemble brut mais parfois concerner des variables externes, et un prédictant de l'autre (la variable à post-traiter), dont on conservera plusieurs réalisations de manière à obtenir une distribution de probabilité. Ces modèles statistiques peuvent emprunter des approches très différentes, comme par exemple le principe d'analogie (Hamill et Whitaker, 2006) ou encore d'apprentissage statistique (Taillardat *et al.*, 2016). Contrairement aux méthodes paramétriques, les méthodes non paramétriques ne sont pas capables d'« extrapoler », c'est-à-dire d'aller au delà des valeurs ayant déjà été observées. Étant basées uniquement sur les données, c'est la richesse de ces données qui fera la performance de la correction.

Pour davantage de détails, nous dirigeons le lecteur vers deux revues récentes de littérature proposées par Li *et al.* (2017) and Wilks (2018a). Par ailleurs, les méthodes EMOS et BMA, utilisées dans cette thèse, seront détaillées dans les sections 5.2 et 7.3.

1.3.2 Pré-traitement et post-traitement

Jusqu'à maintenant, nous avons utilisé le terme de *post-traitement* pour parler de la correction statistique de quelques prévisions que ce soient : météorologiques (prévisions de précipitations ou de température) ou hydrologiques (prévisions de débit). C'est en effet le terme utilisé dans la communauté météorologique d'une part, mais également dans la communauté statisticienne. Dans le langage hydrologique en revanche, il est d'usage de nommer *pré-traitement* la correction des prévisions météorologiques, et *post-traitement* la correction des prévisions de débit. Cette appellation fait référence au positionnement de la correction statistique vis-à-vis du modèle hydrologique, qui est le maillon central de la chaîne de prévision (Figure 1.6). Dans la suite de ce mémoire, nous nous attacherons à différencier le pré-traitement du post-traitement. Le terme de *correction statistique* sera utilisé lorsque cela concerne l'un et l'autre.

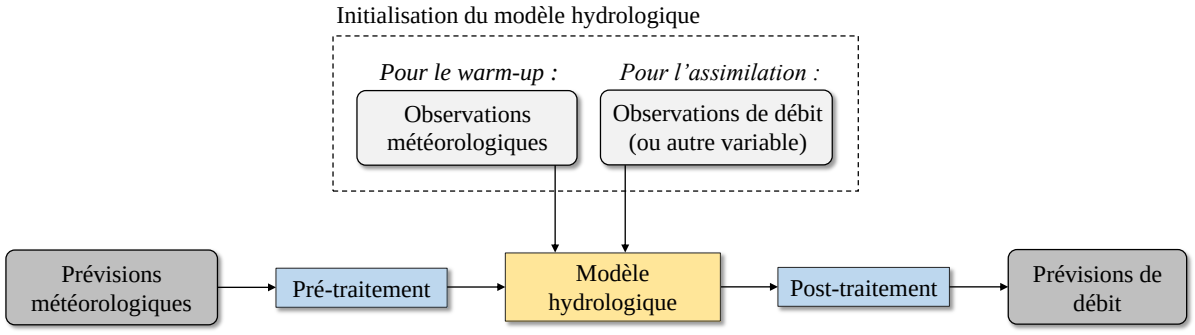

FIGURE 1.6 – Illustration schématique du positionnement des corrections statistiques *pré-traitement* et *post-traitement* au sein d'une chaîne de prévision hydrologique fonctionnant avec un régime pluie-débit.

1.3.3 D'une distribution continue à un ensemble

Les méthodes paramétriques de correction statistique aboutissent à une densité prédictive qui est, du fait de l'usage de lois de distribution, sous la forme continue. Ce serait satisfaisant si cette densité décrivait la variable finale nécessaire à l'utilisateur pour sa prise de décision : pour un producteur d'électricité par exemple, la production totale sur l'ensemble de son parc pendant une période donnée.

Cependant, les étapes de pré-traitement et post-traitement servent à corriger des prévisions qui ont vocation à être injectées dans d'autres outils de modélisation. Il est donc nécessaire de discrétiser ces densités prédictives de manière à obtenir de nouveau un ensemble de valeurs possibles : c'est l'étape d'*échantillonnage*. L'approche la plus intuitive consiste à sélectionner des quantiles équidistants de la densité prédictive, mais nous verrons dans le chapitre 8 que ce n'est pas la seule méthode possible.

1.4 Le besoin de travailler dans un cadre multivarié

La section précédente a abordé le principe de correction statistique s'appliquant à des prévisions *univariées*. Cependant, les outils de modélisation qui composent la chaîne de prévision ont cruellement besoin de prévisions *multivariées*, c'est-à-dire de prévisions qui concernent simultanément plusieurs localisations, échéances et variables (dans notre cas : précipitation et température pour les prévisions météorologiques, débit uniquement pour les prévisions hydrologiques). Cette section doit nous permettre de mieux appréhender la nécessité de travailler dans un cadre multivarié.

1.4.1 Exemple et définitions

Considérons trois prévisions d'ensemble `A`, `B` et `C` de précipitations journalières sur un bassin donné pour les journées de lundi et mardi. Ce sont donc des prévisions *multivariées* dont la *dimension* est égale à 2 : lundi et mardi. Supposons qu'elles aient été émises dans un contexte où un front pluvieux doit passer sur le bassin entre lundi et mardi, en déversant 40 mm/jour. L'incertitude des prévisions réside ici dans la date de passage de ce front. Par ailleurs, prenons un cas simplifié où ces prévisions sont constituées de 2 membres.

La prévision `A` considère ce passage aussi probable lundi que mardi, et contient par conséquent un membre $\{40, 0\}$ (lire : « 40 mm lundi puis 0 mm mardi ») et un membre correspondant au scénario alternatif $\{0, 40\}$. La prévision `B` estime en revanche le passage du front fortement plus probable mardi que lundi, et contient donc deux membres identiques, à savoir $\{0, 40\}$. Enfin, la prévision `C` contient les membres $\{40, 40\}$ et $\{0, 0\}$ (Tableau 1.1).

Dans les cas `A` et `C`, les prévisions *univariées* (ou *marginales*) sont identiques sur les deux dimensions (lundi et mardi) : un des membres prévoit 40 mm et l'autre 0 mm. En revanche, les prévisions univariées du cas `B` sont différentes.

Les prévisions multivariées `A` et `C`, bien qu'elles partagent les mêmes prévisions univariées, sont différentes, car leur *structure de dépendance* est différente. La structure de

TABLE 1.1 – Exemple de trois prévisions (`A`, `B` et `C`) multivariées de dimension deux (lundi et mardi), sous la forme d'ensembles constitués de deux membres. Les valeurs sont en mm/jour.

	Prévision `A`		Prévision `B`		Prévision `C`	
	lundi	mardi	lundi	mardi	lundi	mardi
Membre 1	40	0	0	40	40	40
Membre 2	0	40	0	40	0	0

dépendance d'une prévision multivariée désigne l'association des valeurs entre les différentes dimensions. Ainsi, dans le cas `A`, la valeur de 40 mm lundi est associée avec celle de 0 mm mardi, tandis que dans le cas `C` la valeur de 40 mm lundi est associée avec celle de 40 mm mardi.

Par ailleurs, les prévisions multivariées `A` et `B`, bien que différentes, sont toutes les deux *cohérentes* au regard de la situation météorologique : elles ne prévoient qu'un seul front pluvieux passant lundi ou mardi. En revanche, la prévision `C` n'est pas cohérente, car l'un des membres prévoit le passage de deux fronts successifs lundi et mardi. Cette cohérence joue un rôle crucial dans la modélisation hydrologique. Dans les cas `A` et `B`, le modèle simule, pour chaque membre, la réaction hydrologique du bassin à 40 mm de pluie. Seul le timing de cette réaction hydrologique change selon le membre, compte-tenu de l'incertitude sur la date de passage du front. Dans le cas `C` en revanche, le modèle simule, pour un des membres, la réaction hydrologique du bassin à 80 mm de pluie. Le débit maximal sera alors fortement différent.

Au travers de cet exemple, nous avons illustré la cohérence des prévisions entre plusieurs échéances : c'est la cohérence temporelle. Cependant, la cohérence spatiale (entre plusieurs localisations) et inter-variable (entre plusieurs variables) est tout aussi importante.

1.4.2 Reconstruire des prévisions multivariées cohérentes

Les prévisions météorologiques d'ensemble (cf. 1.2.1.1) sont par construction des prévisions cohérentes, car chaque membre est issu d'un run spécifique du modèle météorologique. L'étape de modélisation hydrologique conserve cette cohérence, chaque scénario météorologique étant transformé en un scénario hydrologique. De même, les approches ensemblistes pour prendre en compte les incertitudes sur la modélisation hydrologique (cf. 1.2.2) ou les conditions initiales du bassin (cf. 1.2.3) conservent également la cohérence, car chaque trace en sortie est associée à un scénario météorologique, un modèle hydrologique et un état initial donné.

En revanche, la correction statistique univariée « détruit » la cohérence des prévisions. En effet, après l'échantillonnage des distributions corrigées on obtient des ensembles qui sont indépendants pour chacune des dimensions du cadre multivarié. Ce problème est illustré à la Figure 1.7, pour un exemple de pré-traitement des prévisions de précipitation à l'aide de la méthode paramétrique EMOS dont nous parlerons plus tard. Les ensembles

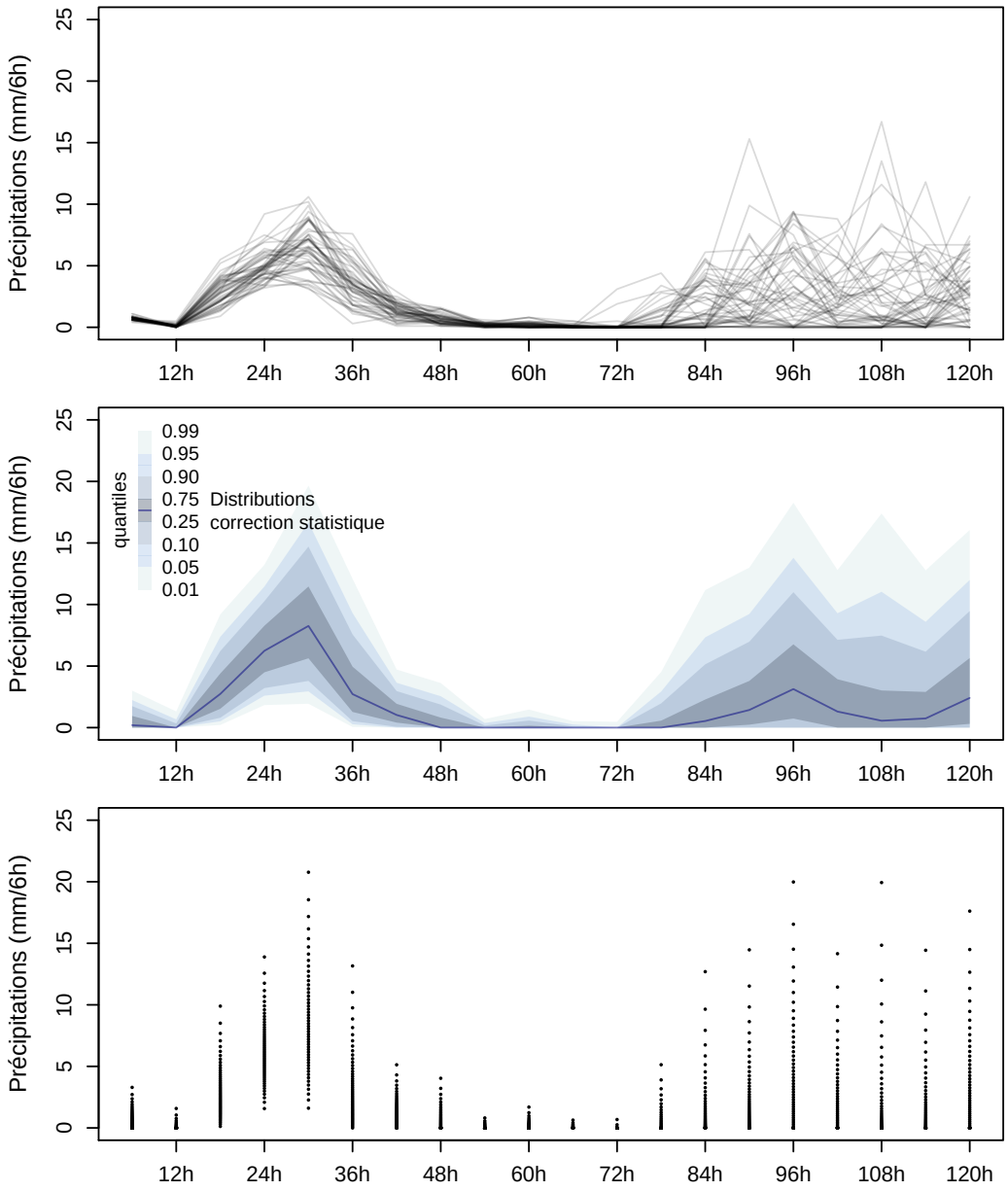

FIGURE 1.7 – Illustration du processus de correction statistique univariée à l'aide d'une méthode paramétrique, dans le cas de prévisions de précipitation. Les trois graphiques représentent les prévisions brutes (haut), les distributions corrigées, représentées par le biais de certains quantiles (centre), et les ensembles obtenus par échantillonnage (bas).

univariés obtenus après échantillonnage sont tracés sur la graphique du bas. Tels quels, ces ensembles ne permettent pas d'alimenter un modèle hydrologique.

Il est donc nécessaire de reconstruire, à partir des ensembles univariés obtenus après correction statistique, des prévisions multivariées cohérentes. Cette étape est cruciale, que ce soit après le pré-traitement ou bien après le post-traitement. Dans le premier cas, c'est un pré-requis à la modélisation hydrologique. Dans le second, c'est indispensable afin d'être en mesure de produire des prévisions sur le cours d'eau qui collecte différents affluents, et non pas simplement des prévisions sur chacun des affluents indépendamment.

1.5 Objectifs de la thèse

L'objectif de ce travail de thèse est de définir les méthodes à mettre en œuvre pour produire des prévisions hydrologiques probabilistes sur un fleuve collecteur d'affluents. Au cours des sections 1.2, 1.3 et 1.4 de ce chapitre introductif, nous avons identifié trois problématiques : la prise en compte des incertitudes en différents points de la chaîne, la correction statistique, et enfin la reconstruction de prévisions multivariées cohérentes.

Dans cette thèse, nous choisissons de reprendre des approches existantes concernant les problématiques de :

- la **prise en compte des incertitudes** : nous utiliserons les prévisions d'ensemble comme forçages météorologiques, et l'approche multi-modèle pour prendre en compte l'incertitude de modélisation hydrologique. En revanche, nous ne traiterons pas la prise en compte de l'incertitude sur les conditions initiales du bassin.

- la **correction statistique** : nous utiliserons l'approche EMOS pour le pré-traitement, et BMA pour le post-traitement.

L'objectif n'est pas d'améliorer ces composantes du point de vue méthodologique, mais plutôt de répondre à certaines questions qui se posent lors de leur intégration au sein d'une chaîne complète de prévision hydrologique. Comment caractériser la qualité et les défauts des forçages météorologiques ? Comment ces défauts impactent-ils la modélisation hydrologique ? Subsistent-ils après le pré-traitement ? Dans quelle mesure la stratégie multi-modèle hydrologique permet-elle d'améliorer les résultats ?

En revanche, nous nous intéresserons particulièrement à la **reconstruction de prévisions multivariées cohérentes**, qui est une composante essentielle de la prévision probabiliste sur un fleuve collecteur d'affluents. Nous tenterons alors d'apporter des développements méthodologiques à la littérature existante.

Enfin, nous nous risquerons à hiérarchiser les apports des différentes méthodes testées tout au long de ce travail de thèse, dans l'optique de fournir des indications concrètes sur les priorités à mettre en œuvre lors de la mise en place ou l'amélioration d'une chaîne opérationnelle de prévision hydrologique probabiliste.

Chapitre 2

Zone d'étude, archives d'observations et de prévisions météorologiques

L'objectif de ce travail de thèse est d'aller au delà de la prévision hydrologique probabiliste sur un unique bassin versant, en répondant entre autres à la problématique de la cohérence spatiale des prévisions. Il nous faut donc définir une zone d'étude comprenant plusieurs bassins, ainsi qu'un jeu de données associé à ces bassins.

Ces données sont de plusieurs natures :

- des archives *d'observations* des variables météorologiques (précipitation, température) et hydrologiques (débit) : ces données permettent de caler, valider et initialiser les modèles hydrologiques, mais servent également de référence lors de l'évaluation des prévisions en différents points de la chaîne.

- des archives de *prévisions*, qui concernent les variables météorologiques seulement, et qui serviront à forcer les modèles hydrologiques.

2.1 Zone d'étude

CNR exploite une série d'aménagements hydroélectriques implantés le long du Rhône, entre son entrée sur le territoire français et l'embouchure avec la mer Méditerranée. Le Rhône est un fleuve collecteur de nombreux affluents, dont les débits peuvent être plus ou moins corrélés. Il représente donc un cas d'étude intéressant pour les problématiques de cohérence spatiale et temporelle des prévisions probabilistes.

Historiquement, ce bassin versant, qui draine une surface de 88 000 km^2, a été découpée par CNR en 56 sous-bassins. Dans cette thèse, nous choisissons de travailler sur un sous-échantillon qui comprend les bassins du haut-Rhône, car ceux-ci nous permettent d'aborder les problématiques auxquelles CNR doit faire face, à savoir :

- la cohérence spatiale des prévisions entre des bassins plus ou moins corrélés,

- la modélisation des stocks nivaux.

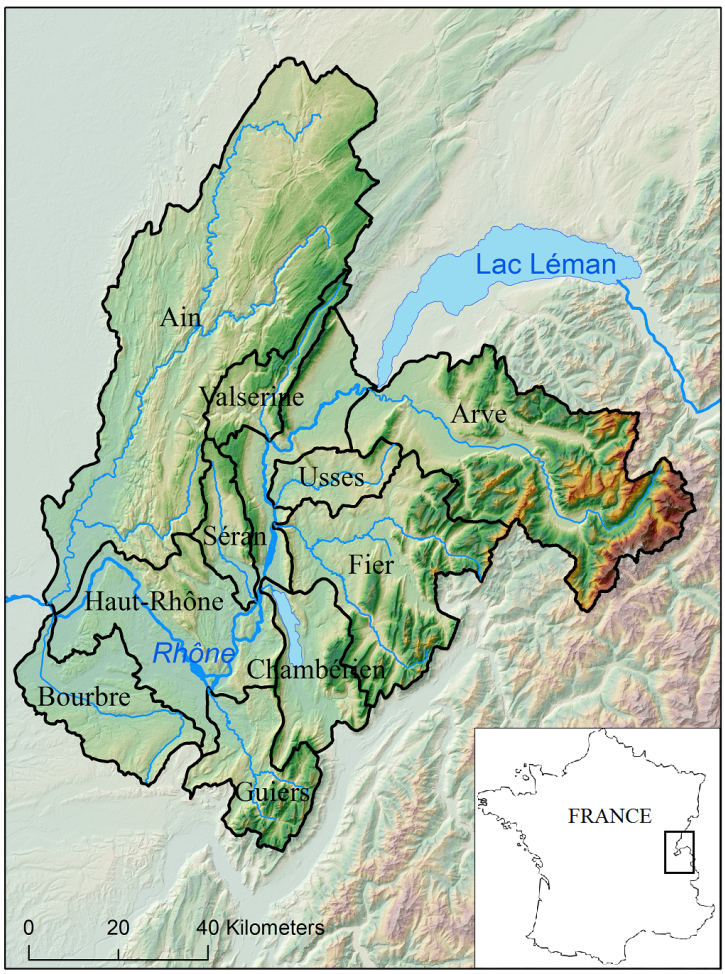

FIGURE 2.1 – Localisation et relief des 10 bassins de la zone d'étude.

Ces bassins, au nombre de 10, sont situés entre le Lac Léman et la confluence Ain-Rhône. La Figure 2.1 illustre leur localisation et leur relief. Par ailleurs, le Tableau 2.1 expose quelques caractéristiques morphologiques, calculées à partir d'un Modèle Numérique de Terrain (MNT) à une résolution de 100 m. Ces bassins abritent des massifs préalpins de moyenne altitude (sommets aux alentours de 2000 m), mis à part l'Arve dont la partie amont pénètre au cœur des Alpes et dont les sommets dépassent 4000 m. Ainsi, l'Arve, la Valserine, le Fier, le Séran et le Guiers ont une proportion non négligeable de surface de bassin où l'altitude dépasse les 1000 m (Tableau 2.1). La prise en compte du manteau neigeux dans la modélisation hydrologique est donc indispensable. Enfin, cet échantillon de bassin présente une grande disparité de surfaces. Nous distinguons ainsi des bassins comme l'Arve et l'Ain qui dépassent 2000 km^2, tandis que d'autres comme les Usses, la Valserine ou le Séran gravitent autour de 300 km^2.

Indépendamment de leur taille, ces 10 bassins revêtent un intérêt particulier pour CNR, car leurs débits alimentent le Rhône dans sa partie amont et, de fait, sont turbinés par l'ensemble de la chaîne d'aménagements hydroélectriques. Cependant, ils forment un système hydrographique non clos et fortement anthropisé dont le fonctionnement est complexe.

TABLE 2.1 – Caractéristiques morphologiques des 10 bassins de la zone d'étude.

Bassin	Surface topographique (km^2)	Altitude (m)	Pourcentage $<$ 1000 m	Pourcentage 1000-2000 m	Pourcentage $>$ 2000 m
Arve	2082	395-4798	40	42	18
Valserine	361	375-1727	52	48	0
Usses	309	278-1357	97	3	0
Fier	1375	270-2577	62	37	1
Chamberien	605	229-1864	82	18	0
Séran	290	228-1496	75	25	0
Guiers	609	210-2002	70	30	0
Bourbre	728	196-777	100	0	0
Ain	3764	189-1478	88	12	0
Haut-Rhône	1956	189-1724	93	7	0

Des débits provenant de la partie suisse du Rhône entrent dans le système environ 1 km avant la confluence Arve-Rhône. Ces débits n'obéissent pas au cycle naturel de l'eau, mais dépendent des opérations de régulation du Lac Léman et des opérations de production hydroélectrique par les Services Industriels de Genève. CNR, qui reprend l'exploitation du fleuve quelques km à l'aval, perturbe également l'écoulement naturel. En effet, les centrales CNR, bien que considérées comme « au fil de l'eau », disposent d'un marnage suffisant pour mener des opérations d'optimisation de la production hydroélectrique. Ces opérations se limitent à de la modulation infra-journalière, qui permet de turbiner davantage durant les heures de pointes, à l'exception du barrage de Génissiat dont les capacités de stockage plus importantes permettent de moduler la production à l'échelle hebdomadaire. En revanche, en période de crue les capacités de stockage de ces aménagements sont très largement négligeables devant les volumes d'eau écoulés. Leur exploitation est alors réalisée de telle façon qu'ils soient transparents vis-à-vis de l'écoulement des débits par rapport à la situation avant construction des aménagements.

Outre le Rhône lui-même, certains bassins de notre zone d'étude sont eux-mêmes aménagés. C'est notamment le cas de l'Ain, sur lequel des opérations de régulation, jusqu'à des échelles saisonnières, sont menées par EDF. Les débits de l'Arve et du Fier sont également influencés par des opérations d'EDF, mais dans une moindre mesure compte-tenu des plus faibles capacités de stockage. Enfin, le bassin Chambérien est particulier car il peut être soit un affluent du Rhône (c'est le fonctionnement « normal »), soit être alimenté par le Rhône (fonctionnement en crue). Ceci est permis grâce à un canal dans lequel l'écoulement peut se faire dans les deux sens, reliant le Rhône au lac du Bourget (Figure 2.1), lequel sert alors de stockage. Cette opération permet de réduire le débit de pointe du fleuve, limitant ainsi les risques de débordements à l'aval.

Ces diverses opérations, qui influent l'écoulement naturel des débits, sont prises en compte dans la chaîne de prévision opérationnelle de CNR (Bompart *et al.*, 2009). Cela requiert d'une part des outils de modélisation qui sont spécifiques à chaque bassin, donc peu généralisables, et d'autre part des données opérationnelles qui dépassent le champ de l'hydrométéorologie (disponibilité des groupes de production, prévisions des prix de

l'électricité, etc.).

Afin de rester dans un cadre de modélisation hydrologique qui soit généralisable à d'autres cas d'étude, nous avons choisi de ne considérer que les bassins du Haut-Rhône dont les débits sont suffisamment peu influencés pour qu'une modélisation pluie-débit soit performante. Ainsi, l'évaluation des forçages météorologiques sera conduite sur l'ensemble de ces 10 bassins, mais la modélisation hydrologique et l'évaluation des prévisions hydrologiques correspondantes ne concernera que les 6 bassins suivants : l'Arve, la Valserine, les Usses, le Fier, le Séran et le Guiers.

2.2 Archives d'observations

Au cours de cette section sont présentées les différentes données d'observations [1] qui ont été utilisées. Ces données nous donnent également l'occasion de compléter la présentation des bassins de la zone d'étude, qui s'est limitée jusqu'ici aux caractéristiques morphologiques.

2.2.1 Archives de précipitation

L'archive pluviométrique contient des mesures de précipitations *moyennées* par bassins. On parle alors également de « lames d'eau ». Ce sont des des mesures des précipitations totales, qui incluent les précipitations liquides et solides.

Cette archive, qui couvre la période 1992-2014, a été produite par Météo-France et fournie de manière opérationnelle à CNR. Les lames d'eau journalières sont d'abord calculées à partir du réseau de mesure dense à l'échelle journalière, puis celles-ci sont désagrégées au pas de temps 6 h au pro-rata des cumuls fournis par le réseau horaire, plus lâche. Ainsi, l'archive à disposition contient 4 mesures par jour, qui correspondent aux cumuls sur les créneaux de 6 h délimités par 00, 06, 12, 18 heures UTC. Cette archive ne remonte pas avant 1992, car le réseau pluviométrique horaire n'était pas assez développé pour que l'on puisse calculer des lames d'eau 6 h de façon satisfaisante.

Le Tableau 2.2 présente quelques caractéristiques pluviométriques des 10 bassins, calculées à partir de cette archive d'observations. Nous constatons qu'il n'y a pas de différences significatives en termes de cumul annuel, de fréquence de pluies nulles ou encore de valeurs de pluies décennales, mis à part les cumuls annuels sur la Bourbre et le haut-Rhône qui s'écartent de la moyenne.

2.2.2 Archives de températures

Si les précipitations sont l'élément déclencheur de la plupart des réactions hydrologiques, la température n'en joue pas moins un rôle important dans la modélisation hydrologique. Elle décide ainsi de la phase des précipitations (liquide ou solide), de la quantité

1. Le terme « observation » doit ici être compris comme « la meilleure estimation possible de ce qu'il s'est passé ». Ainsi, les réanalyses sont considérés comme des observations, bien qu'elles soient issues de modèles.

TABLE 2.2 – Caractéristiques pluviométriques des 10 bassins de la zone d'étude (F0 : fréquence des valeurs nulles ; P10 : valeur décennale sur 6 h, obtenue par ajustement de Gumbel sur les maximas annuels).

Bassin	Cumul (mm/an)	F0 (%)	P10 (mm/6h)
Arve	1391	64	32.3
Valserine	1657	62	41.8
Usses	1227	68	39.4
Fier	1430	65	38.5
Chamberien	1405	66	43.2
Séran	1455	67	44.2
Guiers	1497	67	48.9
Bourbre	1021	69	39.9
Ain	1553	59	35.2
Haut-Rhône	1956	63	38.0

d'eau restituée lors de la fonte nivale, ou encore de la quantité d'eau qui s'échappe du bassin sous forme d'évapotranspiration.

N'ayant pas à disposition une archive de températures issues de mesures de stations météorologiques, nous avons construit une archive d'observations à partir de la réanalyse ERA-Interim (Dee *et al.*, 2011) produite par le *European Centre for Medium-Range Weather Forecasts* (ECMWF). Ce jeu de données a été téléchargé sur le portail MARS de l'ECMWF, à un pas de temps de 6 h et une résolution spatiale de 0.75°, qui est la résolution archivée la plus proche de la résolution native du modèle (troncature T255, soit environ 80 km). Il couvre, comme l'archive pluviométrique, la période 1992-2014.

Cette archive de température a été construite de manière à être homogène à l'archive pluviométrique concernant les pas de temps et d'espace, afin de satisfaire les contraintes de la modélisation hydrologique globale (cf. chapitre 3). Nous avons donc cherché à ce que les valeurs représentent la température moyenne sur les créneaux 6 h délimités par 00, 06, 12 et 18 h UTC, mais également moyenne sur les bassin.

La température est une variable *instantanée*, contrairement aux précipitations qui sont *cumulées*. Ainsi, pour obtenir une température moyenne sur un créneaux 6 h, nous avons considéré la moyenne des valeurs instantanées des débuts et fin des créneaux.

Il a ensuite été défini que les températures de bassin contenues dans l'archive devaient correspondre aux températures pour les altitudes médianes des bassins, en anticipation de l'utilisation du module neige Cemaneige (Valéry, 2010). En effet, celui-ci extrapole la température sur 5 bandes d'altitudes à partir de la température prise à l'altitude médiane du bassin. Pour cela, Cemaneige applique un gradient altitudinal qui varie selon la saison (pour plus de détails, voir section 3.3.1). Nous proposons, pour chaque bassin, d'utiliser cette même méthode d'extrapolation pour ramener la température du point de grille ERA le plus proche du barycentre du bassin jusqu'à son altitude médiane. Ainsi, pour chaque créneau 6 h, la température $T_{med_{BV}}$ (en °C) est calculée via

$$T_{med_{BV}} = T_{ERA} + \frac{\theta_{altitude}}{100} \times (Z_{med_{BV}} - Z_{ERA}), \tag{2.1}$$

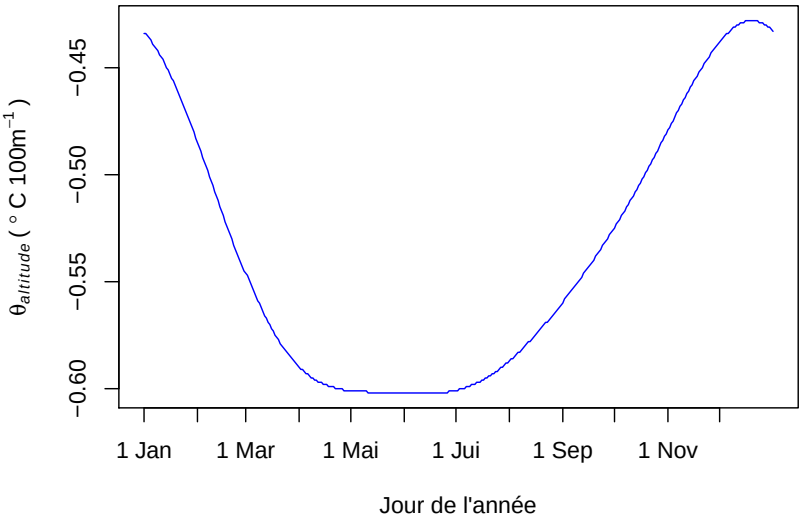

FIGURE 2.2 – Gradients altitudinaux de température utilisés pour construire l'archive des température de bassin à partir des réanalyses ERA-Interim.

où T_{ERA} est la température (en °C) issue de ERA-Interim au point de grille le plus proche du barycentre du bassin, Z_{ERA} est l'altitude (en m) de ce point de grille dans le modèle utilisé pour produire la réanalyse, $Z_{med_{BV}}$ est l'altitude médiane (en m) du bassin, et $\theta_{altitude}$ est un gradient altitudinal (en °C/100m).

Les valeurs qui ont été utilisées pour $\theta_{altitude}$ sont tracées à la Figure 2.2. Celles-ci sont tirées du package `airGR` (Coron *et al.*, 2017) de `R` au sein duquel est implémenté Cemaneige [2]. Ce sont des valeurs optimisées pour la France, issues des travaux de Valéry (2010).

2.2.3 Archives de débits

La dernière archive, débimétrique, contient des observations de débits à l'exutoire de chacun des 6 bassins modélisés, au pas de temps horaire. Ce sont des valeurs *moyennes* horaires. Les données ont été téléchargées depuis la base de données hydrométrique de CNR, Hydromet-Rhone. Nous n'avons conservé, pour cette archive, que la plus longue période commune à l'ensemble des bassins, à savoir 2003-2014.

La Figure 2.3 présente les débits moyens mensuels et le module de ces 6 bassins. Nous constatons que l'Arve est de loin le bassin modélisé qui contribue le plus aux débit du Rhône, avec un module proche de 70 m^3/s. Son régime hydrologique, qualifié de *nivo-glacio-pluvial*, reflète les influences multiples des différentes parties de son bassin dont les altitudes varient notoirement. La période de fonte s'étire ainsi de mars à août, avec un maximum en juin. Vient ensuite le Fier, avec un module proche des 30 m^3/s, puis le Guiers et la Valserine (autour de 15 m^3/s). Ces trois bassins, de moyenne montagne, présentent un régime *nivo-pluvial* avec une période de hautes eaux comportant deux pics de débit : le plus prononcé, au printemps, est dû à la fonte nivale tandis que le second, à la fin de l'automne, est lié aux précipitations. Enfin, le Séran et les Usses ont un module

2. Plus précisément, ces valeurs sont contenues dans la fonction `DataAltiExtrapolation_Valery`.

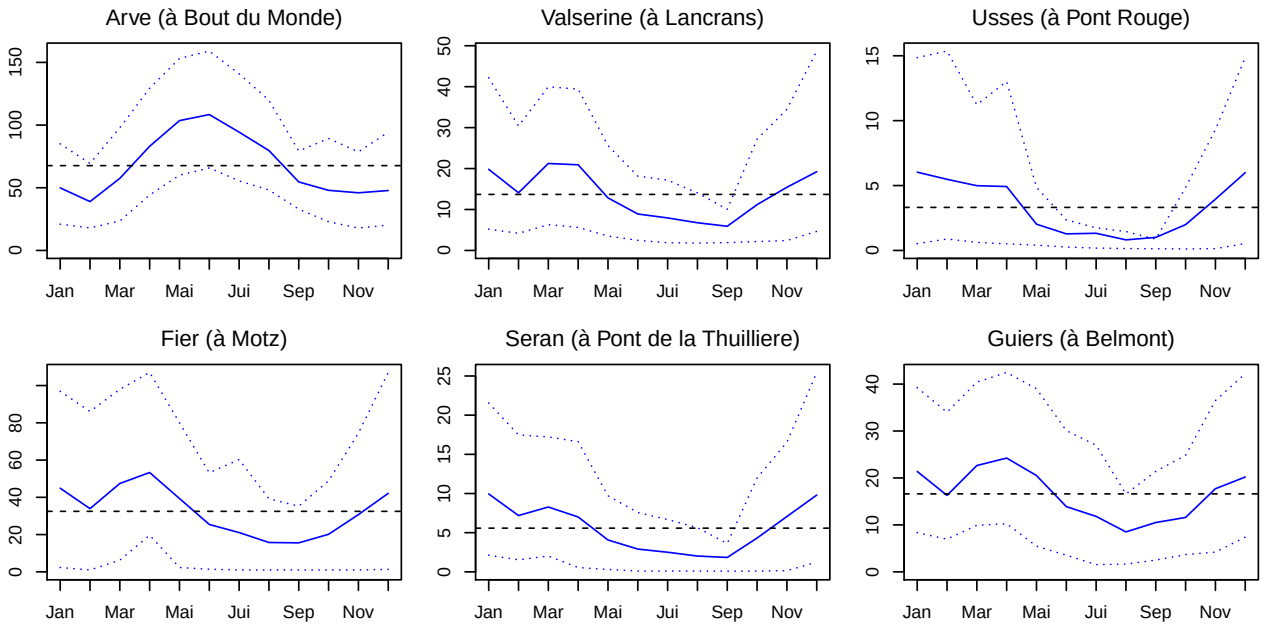

FIGURE 2.3 – Débits moyens mensuels (trait plein bleu) entourés de l'intervalle interquantile 10-90% (trait point bleu), et module (pointillé noir) des 6 bassins modélisés de la zone d'étude. Unité : m^3/s.

entre 3 et 6 m^3/s et présentent un régime *pluvio-nival*, du fait de leur pic de fonte moins marqué.

2.3 Archives des prévisions météorologiques

Dans le chapitre 1, nous avons entrevu différentes approches pour prendre en compte l'incertitude météorologique, et ainsi produire des forçages probabilistes. Nous présentons maintenant les caractéristiques de chacun des différents forçages étudiés. Ainsi sont distingués les prévisions d'ensemble d'une part, et les prévisions par analogie d'autre part.

Durant le diagnostic des performances de ces forçages et leur correction statistique, qui feront l'objet de la partie II, l'essentiel des efforts a été concentré sur la variable précipitation. Certains forçages, notamment les prévisions par analogie, ne fournissent pas la variable température. Par conséquent, ils ne seront pas considérés pour la modélisation hydrologique, mais seulement pour l'évaluation sur la variable précipitation.

2.3.1 Prévisions d'ensemble

Dans cette thèse, nous considérons des archives de prévisions d'ensemble issues de plusieurs systèmes, non pas dans l'objectif de comparer leur performances, mais plutôt de rassembler une variété suffisante de comportements ensemblistes pour être en mesure de vérifier la robustesse des méthodes statistiques développées par la suite (pré-traitement, post-traitement, vérification, etc.). Désirant se rapprocher au maximum d'un contexte de prévision opérationnelle, nous examinons des prévisions passées qui ont été émises au fur et à mesure par les centres de prévisions. Cette archive de prévisions passées est donc

différente des jeux de *reforecasts*, qui contiennent des prévisions générées à posteriori à partir d'un modèle météorologique fixe (Hagedorn, 2008; Hamill *et al.*, 2013).

2.3.1.1 Nos choix sur les systèmes de prévisions étudiés

Trois différentes archives ont été constituées, qui contiennent les prévisions d'ensemble provenant respectivement des centres météorologiques européen (ECMWF), américain (NCEP : *National Centers for Environmental Prediction*) et français (Météo-France).

Le choix d'inclure les prévisions du centre européen, que nous appellerons ECMWF-Ens, a été motivé par plusieurs études (Park *et al.*, 2008; Hamill *et al.*, 2008; Hagedorn *et al.*, 2012; Hamill, 2012, entre autres) ayant mené des inter-comparaisons où celles-ci se plaçaient en tête en termes de performances. Nous avons également décidé d'étudier les prévisions américaines GEFS (*Global Ensemble Forecast System*) qui, grâce à leur mise à disposition gratuite en temps réel, sont les prévisions les plus utilisées à travers le monde. Ces deux archives ont été constituées à partir de la base de données TIGGE (Park *et al.*, 2008). Cette base de données est issue de l'initiative éponyme qui a débuté en 2007 et regroupe, à des fins de recherche scientifique, les prévisions d'ensemble émises par 10 centres météorologiques à travers le monde. Ces prévisions sont archivées sur des grilles latitude/longitude à une résolution qui dépend des modèles. Elle est de 0.25° pour ECMWF-Ens, et de 1° pour GEFS. Nous avons téléchargé ces prévisions sur la période 2007-2014, pour les variables précipitations et température, en considérant le cycle de 00 h UTC uniquement. Seules les échéances jusqu'à 120 h sont considérées.

Il était par ailleurs intéressant de considérer les prévisions d'ensemble émises par Météo-France, nommées PEARP (Prevision d'Ensemble ARPEGE). Celles-ci sont en effet construites à partir d'un modèle météorologique global, ARPEGE (Action de Recherche Petite Echelle Grande Echelle), qui est optimisé sur la France. En revanche, leur résolution d'archivage dans TIGGE, qui est de 1.5°, a été jugée insuffisante vis-à-vis de la taille des bassins étudiés et de leur rapprochement géographique. Plusieurs bassins auraient alors eu des prévisions identiques, entravant ainsi nos objectifs méthodologiques quant à la prise en compte de la cohérence spatiale des prévisions. Par conséquent, les prévisions PEARP de notre archive nous ont été fournies par Météo-France à une résolution de 0.25° davantage compatible à notre contexte. Malheureusement, seules les prévisions pour la variable précipitation ont pu être obtenues. Le système PEARP n'ayant pas de cycle à 00 UTC, nous avons considéré le cycle de 18 UTC, dont les prévisions s'étendent jusqu'à un horizon de 108 h.

2.3.1.2 Caractéristiques des systèmes

L'objectif de cette section est de comprendre ce qui différencie ces trois archives (ECMWF-Ens, GEFS et PEARP), en décrivant les systèmes de prévision d'ensemble non pas dans leur version opérationnelle actuelle, mais sur la période étudiée, à savoir 2007-2014. Un lecteur pressé pourra se contenter des informations résumées dans le Tableau 2.3.

Nous avons vu en introduction que les systèmes de prévision d'ensemble, faisant l'hypothèse qu'une prévision déterministe ne pouvait être parfaitement juste, introduisaient des perturbations de manière à générer un ensemble de scénarios dont la dispersion représente l'incertitude de la prévision. Les systèmes diffèrent donc par leur modèle météorologique intrinsèque, mais également par leurs méthodes de perturbation, qui peuvent concerner l'analyse comme la modélisation.

Perturbation de l'analyse Bien que construite à partir d'observations, l'analyse n'est en fait elle-même qu'une modélisation (donc incertaine) de l'état initial de l'atmosphère. Plutôt que de la perturber au hasard, les systèmes de prévision d'ensemble cherchent à identifier ses zones (aussi appelées « directions », dans l'espace du modèle) qui sont dynamiquement sensibles, à savoir les zones où de faibles variations risquent d'engendrer d'importantes différences dans la modélisation de l'évolution future.

Face à ce même objectif, différentes approches ont été empruntées par les systèmes de prévision d'ensemble. ECMWF-ENS et PEARP ont historiquement employé la méthode des *Singulars Vectors* (SV ; Leutbecher et Palmer, 2008; Diaconescu et Laprise, 2012), tandis que GEFS adoptait l'approche des *Breeding Vectors* (BV ; Wei *et al.*, 2008). Schématiquement, les SV d'un système dynamique (en l'occurrence l'atmosphère) représentent les directions où les perturbations vont croître le plus rapidement, selon une norme donnée et durant un intervalle de temps donné. Pour ECMWF-Ens par exemple, cette norme est l'énergie totale du système, et l'intervalle représente les premières 48 h d'échéance. Les SV sont ainsi définis sur la base de la dynamique future de l'atmosphère. Les BV, en revanche, sont les directions qui ont été instables durant le cycle d'assimilation ; ils sont donc définis sur la base de la dynamique passée.

Par ailleurs, si judicieux soit-il d'identifier les zones dynamiquement sensibles, encore faut-il que celles-ci soient compatibles avec la méconnaissance de l'état initial, c'est-à-dire avec les erreurs de l'analyse. Jusqu'à 2010, l'ECMWF identifiait ces erreurs par le biais de SV complémentaires, obtenus durant les 48 h d'assimilation menant à l'analyse. En 2010, ils ont été remplacés par un module d'*Ensemble Data Assimilation* (EDA ; Buizza *et al.*, 2008), qui représente de manière plus fidèle les erreurs de l'analyse. PEARP a également adopté l'EDA en 2009 (Descamps *et al.*, 2014). En revanche, l'adoption par GEFS d'une méthode aux objectifs similaires, en l'occurrence l'*Ensemble Kalman Filer* (EnKF), ne s'est faite qu'en 2015, soit en dehors de notre période d'étude.

Perturbation du modèle Il est également possible de perturber la modélisation de l'évolution future de l'atmosphère, qui est incertaine. Dans les modèles météorologiques, cette modélisation est scindée en deux parties. La partie *dynamique* résout numériquement, via une discrétisation dans l'espace et dans le temps, les équations qui régissent mathématiquement la circulation atmosphérique à grande échelle. Certains phénomènes, en revanche, ne peuvent être résolus du fait de leur échelle inférieure à celle de la discrétisation : on parle de phénomènes « sous-mailles ». Ces derniers sont simulés via des schémas de paramétrisation ; c'est la partie *physique*.

Conceptuellement, si l'on considère une variable d'état X de l'atmosphère (tempéra-

TABLE 2.3 – Caractéristiques des archives de prévisions d'ensemble durant la période 2007-2014. Les changements significatifs de version des systèmes de prévision d'ensemble sont indiqués par « >mm/aa ».

Nom archive	**ECMWF-Ens**	**GEFS**	**PEARP**
Centre	ECMWF	NCEP	Météo-France
Système de prévision d'ensemble	Ensemble Prediction System (EPS)	Global Ensemble Forecast System (GEFS)	Prévision d'Ensemble ARPEGE (PEARP)
Modèle météorologique	Integrated Forecast System (IFS)	Global Forecast System (GFS)	Action de Recherche Petite Echelle Grande Echelle (ARPEGE)
Membres perturbés	50	<03/07 : 14 >03/07 : 20	>06/04 : 10 >12/09 : 34
Membre de contrôle	Oui	Oui	Oui
Représentation de l'incertitude sur les conditions initiales	<06/10 : SV (Leutbecher et Palmer, 2008) >06/10 : EDA (Buizza *et al.*, 2008) + SV	BV + ET with rescaling (Wei *et al.*, 2008)	>06/04 : SV with breeding >12/09 : EDA + SV (Descamps *et al.*, 2014)
Représentation de l'incertitude sur la modélisation	Stochastic Perturbed Physics Tendencies (SPPT ; Palmer *et al.*, 2009) + Spectral Backscatter Scheme (SPBS ; Berner *et al.*, 2009)	<02/10 : aucune >02/10 : Stochastic Total Tendency Perturbation (STTP ; Hou *et al.*, 2010)	<12/09 : aucune >12/09 : Multi-physique
Résolution horizontale (troncature et « équivalent » de grille)	<01/10 : T399 (50 km) >01/10 : T639 (32 km)	<02/10 : T126 (100 km) >02/10 : T190 (70 km) >02/12 : T254 (55 km)	<12/10 : T358C2.4 (23/134 km)[a] >12/10 : T538C2.4 (15.5/89 km)[a]
Résolution verticale	<11/13 : 62 niveaux >11/13 : 91 niveaux	<02/12 : 28 niveaux >02/12 : 42 niveaux	<01/08 : 41 niveaux >01/08 : 55 niveaux >12/09 : 65 niveaux
Résolution horizontale dans TIGGE	0.25°	1°	1.5° [b]
Cycle étudié	00 UTC	00 UTC	18 UTC
Horizon de prévision considéré	120 h	120 h	108 h

[a]La résolution horizontale d'ARPEGE est variable : maximale sur la France et minimale aux antipodes.

[b]Les prévisions PEARP ont été utilisées à la résolution de 0.25° (cf. 2.3.1.1).

ture, humidité, vent), on peut étudier son évolution dans le temps $\frac{\delta X}{\delta t}$, appelée « tendance » (en anglais : *tendency*). Les modélisateurs ont l'habitude de séparer dans la tendance le terme dynamique (D) du terme physique (P) :

$$\frac{\delta X}{\delta t} = D + P, \tag{2.2}$$

le terme P étant une sorte d'ajustement lié à la représentation des phénomènes sous-mailles.

Le système de prévision d'ensemble ECMWF-Ens, partant du principe que la circulation grande échelle est la partie la moins incertaine de la modélisation, ne perturbe pas explicitement le terme D. En revanche, il applique depuis 1998 une approche stochastique de perturbation, nommée *Stochastic Perturbed Physics Tendencies* (SPPT ; Palmer *et al.*, 2009), qui vise à représenter l'incertitude autant dans la structure des schémas de paramétrisation que dans les paramètres utilisés. Le terme P est alors perturbé via un bruit multiplicatif, qui respecte une certaine structure spatiale et temporelle de manière à ce que les perturbations soient cohérentes. Par ailleurs, ECMWF-Ens applique en parallèle la technique nommée *Spectral Backscatter Scheme* (SPBS ; Berner *et al.*, 2009). Cette technique, qui ne peut être schématiquement rapprochée à l'équation (2.2), vise à représenter les incertitudes de modélisation causées par des processus comme les systèmes convectifs de méso-échelle, qui ne sont ni suffisamment résolus dynamiquement, ni paramétrisés explicitement. De telles erreurs de modélisation peuvent être la source d'une dissipation trop rapide de l'énergie cinétique dans le modèle, qui est alors contrebalancée par le SPBS, via à une injection d'énergie cinétique de manière stochastique.

GEFS a commencé, à partir de 2010 seulement, à représenter l'incertitude de modélisation, via la technique *Stochastic Total Tendency Perturbations* (STTP ; Hou *et al.*, 2010). Celle-ci adopte également une approche de perturbation stochastique de la tendance, mais diffère de SPPT car elle vise à perturber la tendance totale $D + P$, et non plus le terme physique P uniquement.

Enfin, PEARP a emprunté, à partir de 2009, un chemin différent. L'approche choisie, dite « multi-physique » (Descamps *et al.*, 2014), cherche à introduire de l'incertitude sur la partie physique uniquement : certains phénomènes sous-mailles sont modélisés par différents schémas de paramétrisation, et chaque membre de l'ensemble choisit aléatoirement une combinaison de schémas. Cela garantit en théorie la cohérence des perturbations, mais ne permet pas d'en ajuster l'amplitude de manière à contrôler la dispersion des membres.

Autres différences Outre les méthodes appliquées pour perturber l'analyse ou la modélisation, les systèmes de prévision d'ensemble étudiés dans cette thèse diffèrent par leur résolution horizontale et verticale. Par ailleurs, celles-ci ont évolué au cours de la période 2007-2014 (cf. Tableau 2.3). Par example, la résolution horizontale de ECMWF-Ens est passée de 50 à 32 km, celle de GEFS de 100 à 55 km, et celle de PEARP de 23 à 15.5 km [3]. Cette résolution « native » ne doit cependant pas être confondue avec la résolution

3. Sur la France seulement, car PEARP utilise une résolution variable : maximale sur la France, et minimale aux antipodes.

à laquelle nous disposons des prévisions, à savoir 0.25°, 1° et 0.25° pour respectivement ECMWF-Ens, GEFS et PEARP, .

Enfin, les systèmes étudiés se différencient par le nombre de membres qui constituent les ensembles (membres perturbés + contrôle). A l'instar des caractéristiques discutées précédemment, le nombre de membres a évolué avec les changements de versions entre 2007 et 2014. Ainsi, PEARP est passé de 11 à 35 membres en décembre 2009, GEFS de 14 à 21 membres en mars 2007, tandis que pour ECMWF-Ens le nombre de membres est resté constant (51) au cours de la période d'étude. Il s'avère que plusieurs méthodes utilisées ou développées durant cette thèse (pour la correction statistique ou encore la reconstruction de la structure de dépendance) sont sensibles au nombre de membres des ensembles, et peuvent difficilement être évaluées sur une période qui inclut un changement. Par conséquent, nous avons décidé d'écarter les prévisions GEFS émises avant mars 2007, ainsi que les prévisions PEARP avant le décembre 2009.

2.3.1.3 Des prévisions en points de grille aux valeurs de bassins

Les prévisions d'ensemble météorologiques doivent être adaptées de manière à satisfaire les besoins de la modélisation hydrologique globale. Pour la variable précipitation, les prévisions en points de grille sont converties en prévisions de lame d'eau via la méthode de Thiessen (Tabios et Salas, 1985), qui effectue une moyenne spatiale des valeurs aux points de grille pondérée par leur pourcentage de surface représentée sur le bassin.

L'obtention des prévisions de température de bassin suit en revanche le protocole utilisé pour construire l'archive d'observations de température (cf. 2.2.2). Ainsi, la température au point de grille le plus proche du barycentre du bassin est ramenée à l'altitude médiane du bassin, en utilisant les mêmes gradients altitudinaux que ceux tracés Figure 2.2.

2.3.2 Prévisions de précipitation par analogie

Pour rappel (cf. 1.2.1.2), une archive de prévisions par analogie est associée à la fois à un modèle météorologique déterministe [4], qui fournit la prévision des prédicteurs, et un jeu de réanalyse, qui contient l'historique des situations passées (cf. 1.2.1.2). Dans cette thèse, nous avons considéré deux archives différentes qui utilisent les données météorologiques (modèle déterministe et réanalyses) provenant de deux centres différents, l'ECMWF et le NCEP. Ainsi, ces deux archives sont nommées ECMWF-Ana et NCEP-Ana.

Pour le modèle déterministe, ECMWF-Ana et NCEP-Ana exploitent le run de contrôle des prévisions d'ensemble ECMWF-Ens et GEFS, respectivement. Ces prévisions sont interpolées à une résolution commune de 0.5°. Pour les réanalyses, les deux archives se basent respectivement sur ERA-Interim et CFSR (*Climate Forecast System Reanalysis* ; Saha *et al.*, 2010), interpolées à la même résolution de 0.5°. Ces informations sont résumées dans le Tableau 2.4. Enfin, l'archive pluviométrique utilisée est celle présentée en 2.2.1.

4. Thévenot (2004) a montré l'intérêt qu'il pouvait y avoir à s'appuyer non plus sur les prédicteurs provenant d'une prévision déterministe, mais sur ceux provenant de plusieurs membres d'une prévision d'ensemble. Cependant, une telle approche n'a pas été étudiée ici.

Table 2.4 – Données utilisées par les archives ECMWF-Ana et NCEP-Ana de prévisions par analogie.

Nom archive	**ECMWF-Ana**	**NCEP-Ana**
Modèle déterministe	ECMWF-Ens (contrôle)	GEFS (contrôle)
Réanalyses	ERA-Interim	CFSR

Note : Toutes les données sont interpolées à la résolution de 0.5°.

C'est cette dernière, qui couvre 1992-2014, qui restreint la longueur de l'historique pouvant être exploité.

Le modèle « analogue » considéré pour ces deux archives est celui utilisé en opérationnel à CNR pour la prévision des précipitations. Il provient des travaux successifs de Obled *et al.* (2002), Bontron (2004), Marty *et al.* (2012) et Ben Daoud *et al.* (2016). Le prédictant est la lame d'eau par bassin cumulée sur 6 h. Trois niveaux de sélections sont mis en place, qui a chaque fois raccourcissent la liste des situations candidates. En d'autres termes, une situation écartées au niveau $n-1$ ne pourra être retenue au niveau n. Ces niveaux sont les suivants :

Niveau 0 : C'est une pré-sélection sur la température des situations candidates, de manière à garantir une certaine saisonnalité. Deux prédicteurs sont utilisés, la température à 850 hPa au début du créneau 6 h (T850-0h), et celle à 500 hPa à la fin du créneau 6 h (T500-6h). Ces deux prédicteurs sont considérés uniquement sur le point de grille le plus proche de chaque bassin. Le critère d'analogie est la RMSE, ou plus précisément la moyenne des deux RMSE calculées chacune sur un des prédicteurs. Le nombre d'analogues retenus est de 9000.

Niveau 1 : C'est le niveau le plus discriminant, qui retient les situations les plus similaires du point de vue de la circulation synoptique. Deux prédicteurs sont utilisés, le géopotentiel à 500 hPa au début du créneau 6 h (Z500-0h), et celui à 1000 hPa à la fin du créneau 6 h (Z1000-6h). Cette fois, ces prédicteurs sont considérés sur un large domaine spatial, qui est optimisé pour chaque bassin. Le critère d'analogie est le score S1 (ou score TWS ; Teweles et Wobus, 1954), qui quantifie la différence de forme entre deux champs de géopotentiels. Plus précisément, c'est la moyenne des deux scores S1 calculés chacun sur un des prédicteurs, qui est utilisée. Le nombre d'analogues retenus est de 175.

Niveau 2 : Ce dernier niveau vise à raffiner la sélection à l'aide d'une variable à la fois « locale » (contrairement aux géopotentiels qui caractérisent la situation « grande échelle ») et « physique » (le géopotentiel étant une variable davantage « dynamique »). Ainsi, deux prédicteurs sont considérés : le produit de l'humidité relative à 850 hPa et de la colonne d'eau précipitable, tout deux pris au début du créneau 6 h (RH850×TCW - 0h), et le même produit mais pris à la fin du créneau 6 h (RH850×TCW

-6h). De nouveau, ces prédicteurs sont considérés sur un domaine spatial qui est optimisé pour chaque bassin. Le critère d'analogie est la RMSE, ou plus précisément la moyenne des deux RMSE calculées chacune sur un des prédicteurs. Le nombre d'analogues retenus est de 40.

Pour construire les prévisions probabilistes de précipitation, il reste à aller chercher dans l'archive pluviométrique les valeurs de précipitation qui ont été observées sur les 40 dates analogues retenues, permettant ainsi de construire une distribution empirique.

Il est important de noter que ces distributions empiriques sont obtenues indépendamment pour chacun des bassins et chacune des échéances. Ainsi, contrairement aux prévisions d'ensemble, les prévisions par analogie ne produisent pas telles quelles des traces qui puissent alimenter un modèle hydrologique. Il faudrait pour cela reconstruire une structure de dépendance spatio-temporelle, problématique qui sera étudiée dans le chapitre 6.

Chapitre 3

Outils de modélisation hydrologique

Le modèle hydrologique est le maillon central de la chaîne de prévision hydrologique, car il permet de transformer les forçages météorologiques en débits à l'exutoire. Or c'est une source importante d'incertitude, du fait de son incapacité à reproduire exactement le comportement du bassin versant. Comme mentionné dans les objectifs de ce travail de thèse, nous proposons d'adopter une approche multi-modèle, de manière à prendre en compte une partie de cette incertitude.

Ce chapitre vise à présenter les outils de modélisation. Après une rapide typologie des modèles hydrologiques en section 3.1, nous présentons en section 3.2 les trois modèles utilisés dans cette thèse : le modèle ARX, qui est celui utilisé en opérationnel à CNR, puis TOPMODEL et GRP. Ces modèles sont couplés avec un module neige et une formulation d'évapotranspiration potentielle communs, qui sont présentés en section 3.3. Nous discutons ensuite en section 3.4 de la stratégie de calage qui a été adoptée, puis terminons en section 3.5 par la présentation des performances obtenues en calage et en validation.

L'objectif n'est pas de faire une description détaillée de toutes les composantes des modèles, mais de fournir une vision globale de leurs principes généraux de fonctionnement, de manière à appréhender la cohérence de l'approche multi-modèle. Ainsi, nous présentons les hypothèses principales présentes derrière chacun des outils de modélisation utilisés, et donnons au lecteur un certain nombre de références lui permettant d'aller plus loin s'il le souhaite.

3.1 Typologie des modèles hydrologiques

Nous avons vu dans le chapitre introductif qu'un modèle hydrologique se présentait sous la forme d'une série d'équations mettant en relation forçages météorologiques et débits par l'intermédiaire de *variables d'état* et de *paramètres*. Il existe une grande variété des modèles hydrologiques, et l'exercice de classification ne peut que dépendre des critères utilisés. Sans prétendre être exhaustif, nous avons sélectionné trois critères, jugés pertinents pour bien appréhender la description des modèles qui suivra.

Discrétisation spatiale Nous distinguons généralement les modèles *distribués* des modèles *globaux*. Un modèle distribué discrétise le bassin en mailles, qu'elles soient régulières

(pixels) ou « hydrologiques » (par exemple, des entités homogènes du point de vue de leur réponse hydrologique). Variables d'états et paramètres sont alors effectifs à l'échelle de la maille. A l'inverse, un modèle global représente le bassin comme une entité hydrologique homogène, ce qui exclut la prise en compte de la variabilité spatiale des processus. Variables d'états et paramètres caractérisent alors le bassin dans sa globalité.

Représentation des processus Différentes approches peuvent être adoptées pour relier les forçages météorologiques aux débits. Les modèles *à base physique* représentent les processus (infiltration, écoulements, etc.) par les lois physiques qui les régissent. Les processus hydrologiques couvrant une large gamme d'échelles, la pertinence de cette approche est fortement dépendante de l'échelle spatiale de représentation du bassin versant. Cherchant à s'affranchir de ces problèmes d'échelle, les modèles *conceptuels* tentent de représenter quelques processus dominants uniquement, sans forcément avoir recours explicitement aux lois physiques qui les régissent, mais bien souvent à des relations empiriques. Enfin, les modèles *statistiques* reproduisent les débits à partir uniquement de relations statistiques qui les relient aux forçages météorologiques. Ils sont ainsi construit à partir des données, et non pas à partir des caractéristiques (physiques, morphologiques, etc.) des bassins. Pour cette raison, les modèles statistiques sont parfois qualifiés de « boîtes noires ».

Utilisation L'utilisation d'un modèle hydrologique pour la prévision opérationnelle requiert la présence d'une procédure d'*initialisation* des variables d'état à la date de prévision. Certains modèles hydrologiques ont été conçus de pair avec cette procédure, tandis que d'autres ont été imaginés davantage dans un but de simulation continue. Dans le deuxième cas, la construction de la procédure d'initialisation est un préalable indispensable à une utilisation dans un contexte de prévision.

3.2 Présentation des modèles utilisés

Cette section présente les trois modèles hydrologiques qui ont été utilisés, à savoir ARX, TOPMODEL et GRP. Ces trois modèles fonctionnent ici au pas de temps horaire. Pour chacun, nous débutons par une description du modèle en simulation continue, avant de présenter la procédure d'initialisation des variables d'état.

Dans la section 3.3 qui suivra, nous décrirons deux composantes de la modélisation qui ont été « extraites » de la modélisation pluie-débit à proprement parler, et qui sont communes aux trois modèles hydrologiques. La première composante est la modélisation du manteau neigeux via le module Cemaneige, qui ajuste si besoin la lame d'eau en retranchant la partie solide et en ajoutant la lame d'eau de fonte (voir 3.3.1). La seconde est le calcul de l'évapotranspiration potentielle à partir des données de tempéarture, via la formulation d'Oudin (voir 3.3.2).

3.2.1 ARX

Le modèle ARX est le modèle hydrologique utilisé par CNR depuis 2004 pour la prévision opérationnelle. Son acronyme signifie *Autoregressive model with exogenous inputs*. C'est un modèle entièrement statistique, qui repose sur la relation d'autocorrélation des débits mais avec comme variables explicatives supplémentaires les lames d'eau qui s'abattent sur le bassin durant les pas de temps précédents. Plus de détails peuvent être obtenus dans Box *et al.* (2015) pour les modèles ARX en général, ou dans Remesan et Mathew (2015) dans un contexte de prévision hydrologique.

La variable de sortie, le débit Q^t au pas de temps t, est ainsi calculée via une équation de la forme :

$$Q^t = a + \sum_{i=1}^{B} b_i Q^{t-i} + \sum_{i=1}^{C} c_i P^{t-i}, \tag{3.1}$$

où P, la lame d'eau, est la variable exogène. Les coefficients de régression $(a, b_1, \ldots, b_B, c_1, \ldots, c_C)$ sont les paramètres du modèle [1], qui bien entendu ne sont pas interprétables physiquement. Segmentés selon les gammes de débit, les saisons et les conditions d'humidité du bassin, leur nombre peut largement dépasser la centaine.

La gamme de débit au pas de temps t est déterminée selon la valeur antérieure du débit, Q^{t-1}. Les conditions d'humidité sont déterminées selon la valeur antérieure d'une variable d'état du modèle, l'indice des précipitation antérieures (IPA), qui se calcule de façon récursive :

$$\text{IPA}^t = \alpha \ \text{IPA}^{t-1} + P^{t-1}, \tag{3.2}$$

avec $\alpha < 1$. Schématiquement, l'IPA s'apparente au niveau de remplissage d'un réservoir représentant l'humidité du sol du bassin (Hingray *et al.*, 2009). Ce réservoir est alimenté par les lames d'eau qui s'abattent sur le bassin, tandis qu'il se vidange au cours du temps via le coefficient α. Ce coefficient $\alpha < 1$ est une manière de représenter l'évapotranspiration, qui n'est pas prise en compte explicitement par le modèle ARX.

La Figure 3.1 illustre le fonctionnement en simulation continue du modèle ARX couplé avec Cemaneige, sur une année hydrologique. Ce graphique, qui sera également utilisé pour TOPMODEL et GRP, se lit de la façon suivante. Le premier graphique illustre le fonctionnement de Cemaneige (voir 3.3.1) via l'évolution du stock nival, qui est alimenté par les lames d'eau solides et déplété par les lames d'eau de fonte. Le second graphique présente les forçages du modèle hydrologique, qui ici ne concernent que la lame d'eau issue de Cemaneige. Les variables d'état du modèle, ici l'IPA seulement, sont représentées dans le troisième graphique. Enfin, le dernier graphique illustre les débits observés et simulés.

Bien que la simulation continue soit mathématiquement permise, les modèles ARX sont intrinsèquement conçus pour fonctionner en mode prévision. Les débits simulés sur les pas de temps qui suivent la date de prévision t_{prev} reposent en effet essentiellement sur les débits observés jusqu'à t_{prev}, grâce à la structure d'autocorrélation. Cette structure de modèle représente donc une certaine forme d'assimilation (des derniers débits observés), qui par construction met automatiquement le modèle en conformité avec la tendance ob-

1. La plupart des modèles ARX calés par CNR imposent $a = 0$, $B = 2$, et $b_1 + b_2 < 0$.

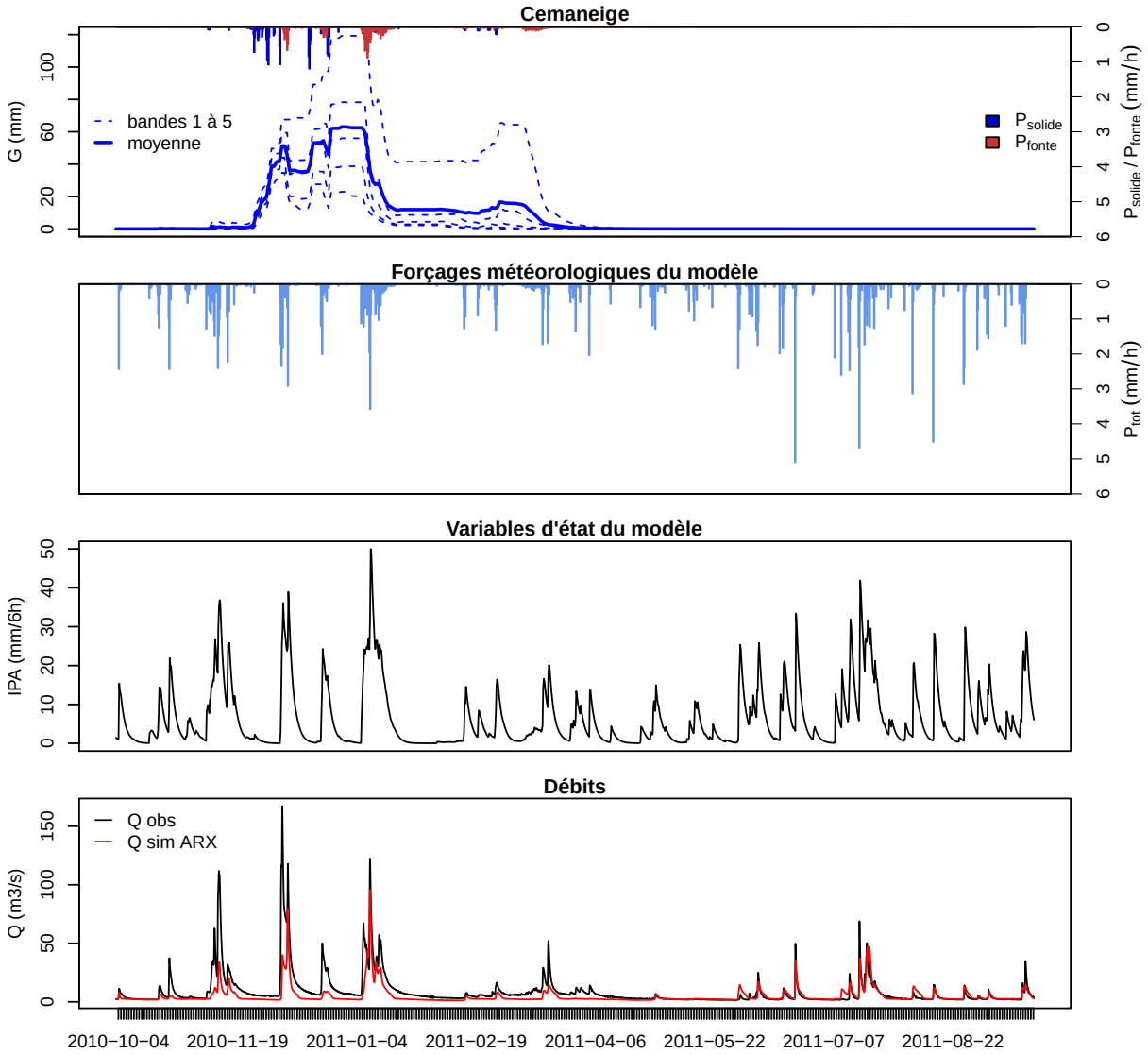

FIGURE 3.1 – Illustration du fonctionnement en simulation continue du modèle ARX couplé avec Cemaneige, sur une année hydrologique, pour le bassin de la Valserine.

servée (bien que cela ne mette aucune variable d'état à jour comme c'est communément le cas avec les modèles à réservoirs). Cette assimilation automatique est l'un des grands avantages de ce type de modèle en prévision opérationnelle. L'unique variable d'état du modèle ARX, l'IPA, est initialisée grâce à une période de warm-up qui exploite les précipitations observées sur la période précédant la prévision.

3.2.2 TOPMODEL

3.2.2.1 Simulation continue

Le second modèle, TOPMODEL (Beven et Kirkby, 1979; Beven *et al.*, 1995), est plus « classique », dans le sens où il sépare la réponse hydrologique à un évènement pluvieux entre une fonction de production et une fonction de routage. C'est un modèle conceptuel, qui se base sur les lois physiques pour représenter les écoulements de sub-surface. En réalité, TOPMODEL représente une famille de modèles qui peuvent différer par certains

aspects, mais partagent un principe commun que nous essayons dans cette section de résumer. Pour plus de détails, nous redirigeons le lecteur vers Obled et Zin (2004).

Dans TOPMODEL, la fonction de production est représentée par deux processus dominants :

- **L'écoulement de sub-surface**, ou écoulement hypodermique : c'est la contribution provenant de la lame d'eau infiltrée qui transite horizontalement par les couches de sol les plus conductives (proches de la surface), jusqu'à exfiltrer à la rencontre d'un chenal d'écoulement ; c'est la composante lente de l'hydrogramme simulé.
- **Le ruissellement sur les zones saturées** : c'est la contribution provenant de la lame d'eau qui s'abat sur des surfaces temporairement saturées, et qui ne peut donc pas s'infiltrer ; c'est la composante rapide.

Ces deux processus sont simulés via la modélisation d'une nappe dite « de versant », ou nappe superficielle, qui est la zone saturée au sein de laquelle se produit l'écoulement de sub-surface. Elle ne doit pas être confondue avec les nappes profondes, qui ne sont pas modélisées dans TOPMODEL.

La modélisation de cette nappe de versant s'appuie sur une discrétisation du bassin en pixels. Pour un pixel i et un pas de temps t, une équation d'équilibre, inspirée de la loi de Darcy, s'exprime ainsi :

$$A_i R_t = T_{i,t} \, \tan\beta_i, \tag{3.3}$$

où A_i est l'aire drainée par ce pixel i ; R_t est la lame d'eau au pas de temps t, supposée homogène sur tout le bassin ; $\tan\beta_i$ est la pente locale de la nappe, qui est supposée égale à la pente topographique du pixel i ; et enfin $T_{i,t}$ est la transmissivité du profil de sol au droit du pixel i. Un hypothèse forte de TOPMODEL est alors de considérer une transmissivité exponentiellement décroissante avec la profondeur du sol, à partir d'une valeur T_0 à saturation :

$$T_{i,t} = T_0 \, \mathrm{e}^{-\frac{d_{i,t}}{m}}, \tag{3.4}$$

où m est un coefficient de décroissance supposé (comme T_0) constant sur tout le bassin, et $d_{i,t}$ est la profondeur du niveau de la nappe. Une astuce de formulation consiste à exprimer ce niveau non pas en distance mais en lame d'eau. Appelé déficit local, il correspond à la lame d'eau supplémentaire qu'il faudrait infiltrer pour que la nappe affleure la surface et ainsi « sature » le pixel.

TOPMODEL fait intervenir des caractéristiques topographiques du bassin. Il propose en effet d'utiliser le MNT du bassin pour caractériser chaque pixel i par un indice topographique λ_i, directement dérivé des équations (3.3) et (3.4), et défini par

$$\lambda_i = \ln\left(\frac{A_i}{\tan\beta_i}\right). \tag{3.5}$$

Cet indice traduit la propension du pixel à se saturer plus ou moins facilement : A_i caractérise sa capacité à collecter de l'eau, tandis que $\tan\beta_i$ caractérise sa capacité à l'évacuer par gravité. La Figure 3.2 illustre, pour le bassin de la Valserine, le MNT ainsi que

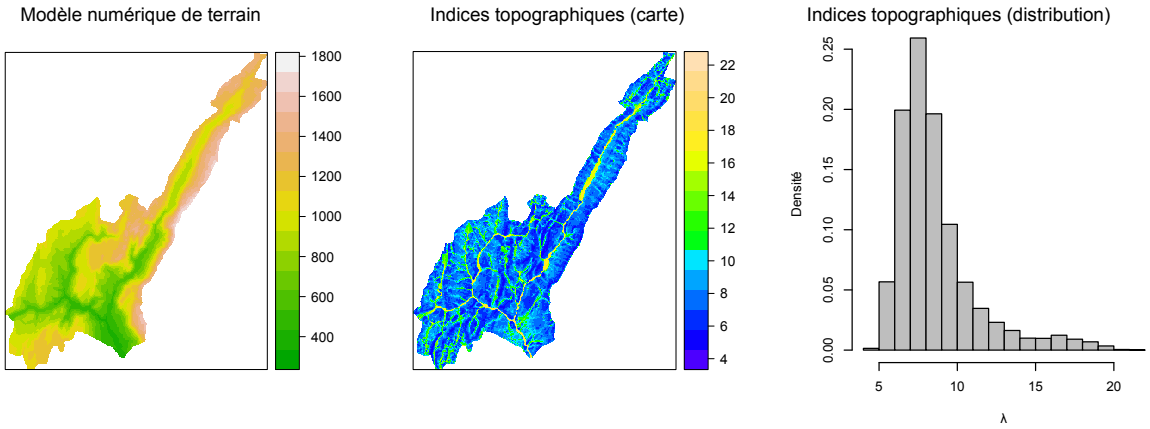

FIGURE 3.2 – MNT du bassin de la Valserine (gauche), et distribution spatiale (centre) et statistique (droite) des indices topographiques, obtenus par traitement SIG du MNT.

la carte des indices topographiques obtenus par traitement SIG. En réalité, TOPMODEL n'exploite pas la répartition géographique de ces indices topographiques, mais uniquement leur distribution statistique, également représentée dans la Figure 3.2. Ces indices sont invariables temporellement, et calculés une fois pour toute au moment de caler le modèle.

En développant les équations, il est possible, en faisant apparaître l'indice topographique λ_i, d'obtenir l'égalité

$$D_t - d_{i,t} = -m(\overline{\lambda} - \lambda_i), \tag{3.6}$$

où D_t est le déficit moyen au pas de temps t et $\overline{\lambda}$ l'indice topographique moyen du bassin. Cette équation, qui caractérise TOPMODEL, stipule que la variation du déficit local par rapport au niveau moyen est directement liée à la variation de l'indice topographique par rapport à la valeur moyenne du bassin. Le déficit moyen D est une variable d'état de TOPMODEL, qui quantifie le niveau de la nappe à l'échelle du bassin. Le terme global d'écoulement de subsurface, noté Q_{ss}, peut alors être calculé par l'équation :

$$Q_{ss_t} = A\ K_0\ m\ \mathrm{e}^{-\frac{D_t}{m}}\ \mathrm{e}^{-\overline{\lambda}}. \tag{3.7}$$

où A est la surface totale du bassin.

Décrivons maintenant la composante rapide de la fonction de production, qui correspond au ruissellement de la lame d'eau sur les zones saturées. L'équation (3.6) donne directement accès, connaissant le déficit moyen D_t, aux déficit locaux $d_{i,t}$ (les termes de droite étant invariables temporellement). On obtient alors, en sommant les pixels où le déficit est nul, la surface saturée du bassin, à partir de laquelle il est aisé de déterminer la contribution issue du ruissellement rapide, notée Q_r.

La Figure 3.3 illustre, sur une période de 2 mois, les contributions Q_{ss} et Q_r de la fonction de production de TOPMODEL. Les cartes de saturation, calculées à la fin d'une période de récession ainsi qu'au milieu d'une réaction hydrologique, illustrent ainsi la modélisation de la nappe de versant.

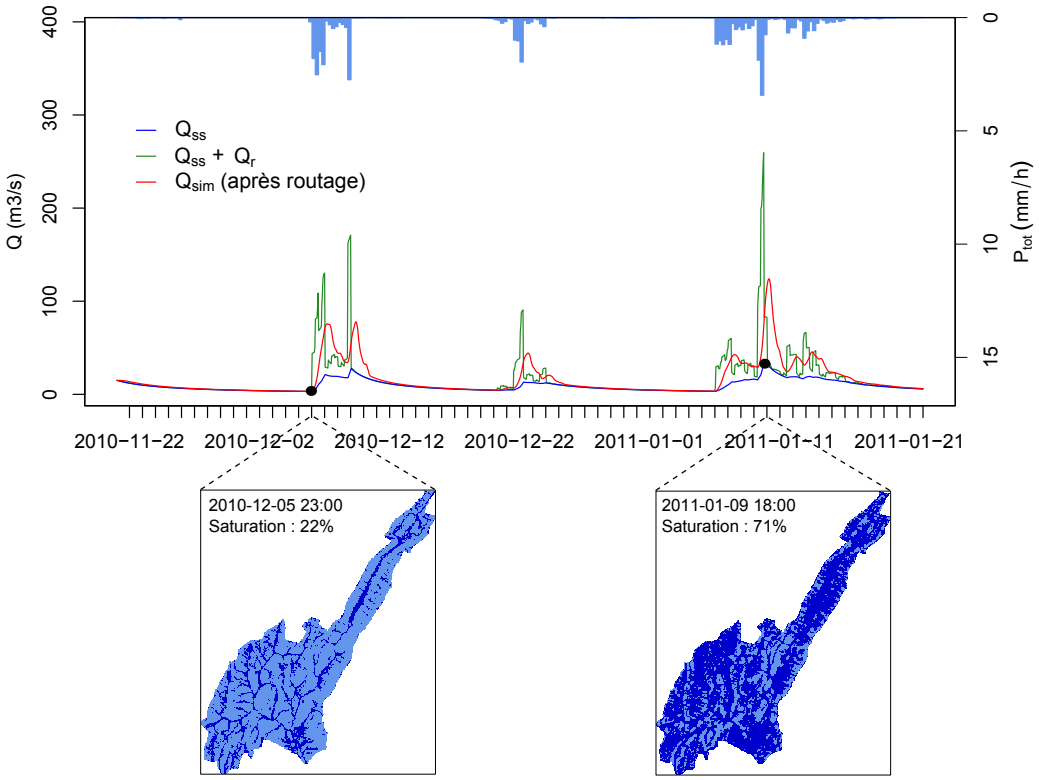

FIGURE 3.3 – Décomposition du débit simulé de TOPMODEL (rouge), à partir des contributions Q_{ss} (bleu) et Q_r (vert) de la fonction de production, sur une période de 2 mois. Ces contributions sont directement liées au niveau moyen de la nappe, dont une des résultantes est le pourcentage de saturation du bassin, illustré par les cartes de saturation pour deux exemples de dates.

Nous avons présenté, jusqu'à maintenant, le principe qui caractérise TOPMODEL. Autour de celui-ci s'ajoute généralement des routines supplémentaires, qui diffèrent selon les versions. Pour cette thèse, nous avons utilisé la version implémentée dans le package `topmodel` de `R`, qui correspond à la version Fortran 1995 de Beven *et al.* (1995), également décrite dans Beven (2012). Cette version inclut en amont de la recharge de la nappe un réservoir racinaire de profondeur $\mathrm{SR}_{\max}$, qui permet d'« encaisser » une certaine partie des lames d'eau avant toute réaction hydrologique. L'ETP peut alors puiser de l'eau dans ce réservoir racinaire. Son niveau SR, exprimé en terme de déficit (SR = 0 signifie que le réservoir est plein), constitue la seconde variable d'état du modèle. Enfin, le routage des contributions $Q_{ss} + Q_r$ est réalisé via une discrétisation du bassin par bandes d'égale distance à l'exutoire en suivant les chenaux d'écoulement (courbes isochrones). Un paramètre v_r représentant la vitesse d'écoulement dans les chenaux permet alors de répartir, dans le temps, l'arrivée à l'exutoire de $Q_{ss} + Q_r$. L'effet de ce routage est illustré en rouge sur la Figure 3.3. Il convient de noter que la version dans le package `topmodel` de `R` simule deux processus supplémentaires que nous avons neutralisé afin de conserver son caractère parcimonieux au modèle : le ruisselement hortonien (ou ruissellement par refus d'infiltration), ainsi que l'instauration d'un retard dans l'infiltration conduisant à la recharge de la nappe.

Pour résumer, TOPMODEL est un modèle conceptuel, mais qui s'inspire des lois

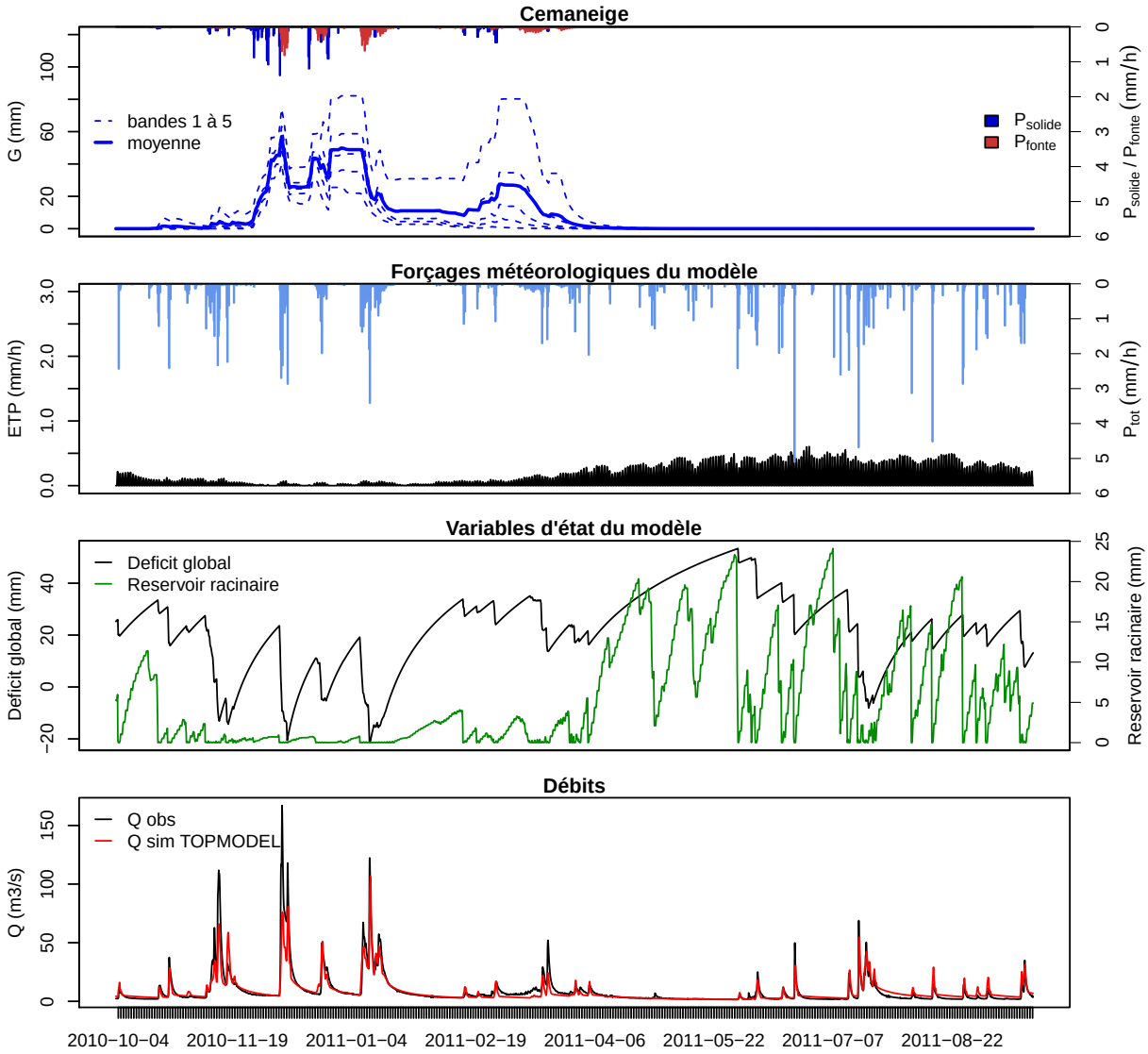

FIGURE 3.4 – Illustration du fonctionnement en simulation continue du modèle TOPMODEL couplé avec Cemaneige, sur une année hydrologique, pour le bassin de la Valserine.

physiques pour représenter certains processus (saturation des sols, écoulement de subsurface, routage). De plus, c'est un modèle global, car bien qu'il se permette d'utiliser des données discrétisées pour caractériser la morphologie du bassin, ses 4 paramètres (m, K_0, $\mathrm{SR}_{\max}$ et v_r) et ses 2 variables d'états (D et SR) sont représentatifs du bassin global. Le fonctionnement en simulation continue de TOPMODEL couplé à Cemaneige est illustré à la Figure 3.4.

3.2.2.2 Mode prévision

La version de TOPMODEL implémentée dans le package `topmodel` de `R` est conçue pour fonctionner en simulation continue, et ne dispose pas de procédure d'initialisation. Nous avons ainsi développé notre propre procédure, en se basant sur le schéma warm-up/assimilation.

Le warm-up, qui ne concerne que la routine du réservoir racinaire, permet d'initialiser

le déficit SR du réservoir racinaire. Des expériences ont montré qu'une durée de 90 jours s'avérait suffisante. Ensuite, l'assimilation permet de faire démarrer la prévision au niveau du dernier débit observé. Cette assimilation n'est cependant pas triviale, car il est complexe d'« inverser » le modèle de manière à calculer analytiquement les variables d'état qui produirait le débit simulé souhaité. En effet, ce débit simulé est le fruit de deux composantes non indépendantes de la fonction de production, qui sont ensuite réparties dans le temps par la fonction de routage. La procédure adoptée, qui vise à mettre à jour le déficit moyen D uniquement, se veut plus simple. Elle consiste à déterminer, par itérations, la valeur de D à $t_{\text{prev}} - 48$ h qui produit la simulation où le débit simulé à t_{prev} est égal au débit observé (à 0.5 m^3/s près). Les simulations réalisées dans les itérations, qui durent 48 h seulement, exploitent alors les forçages météorologiques observés. Cette durée de 48 h correspond à la borne haute des temps de concentration de bassins modélisés, et permet à la fonction de routage de se mettre entièrement en place (cette fonction « retardant » les contributions issues de la fonction de production). Ainsi, TOPMODEL en mode prévision démarre ses simulation non pas t_{prev} mais à $t_{\text{prev}} - 48h$, tout en garantissant que les débits prévues partent bien du dernier débit observé.

3.2.3 GRP

Le troisième modèle hydrologique est le modèle GRP, conçu pour la prévision hydrologique à court terme au pas de temps horaire. Son développement par l'IRSTEA a été axé sur la prévision des crues, grâce à l'expertise acquise sur les modèles GR (Génie Rural ; Perrin *et al.*, 2007), et notamment le modèle GR4J (Perrin *et al.*, 2003) dont il emprunte l'essentiel de la structure. C'est un modèle global, avec une structure à réservoirs, qui contient une fonction de production et une fonction de routage. Bien qu'il se présente à première vue comme un modèle conceptuel, son fonctionnement est davantage empirique. En effet, les modèles GR n'ont pas été conçus en ayant recours à priori à la physique des écoulements, mais plutôt de manière progressive en choisissant les opérateurs mathématiques qui se sont avérés simuler le mieux les débits à l'exutoire, grâce à des tests sur un très grand nombre de bassins versants. Des informations détaillées sur GRP peuvent être trouvées dans les travaux de Tangara (2005) et Berthet (2010).

Outre l'objectif de rendre la modélisation pluie-débit la plus parcimonieuse possible, GRP se différencie des autres modèles GR par la présence de routines spécifiques pour l'initialisation du modèle en mode prévision. Ces routines seront présentées en 3.2.3.2. Avant cela, nous présentons le fonctionnement du modèle en simulation continue.

3.2.3.1 Simulation continue

La structure de GRP est représentée dans la Figure 3.5. Nous décrivons ici brièvement les différentes opérations qui se succèdent, à partir de ses données d'entrées qui sont la pluie P et l'ETP.

La première étape de la fonction de production, l'interception, consiste à comparer l'ETP à P afin de déterminer la pluie nette et l'évapotranspiration nette. Ces flux interagissent avec un premier réservoir, nommé réservoir de production, qui représente de

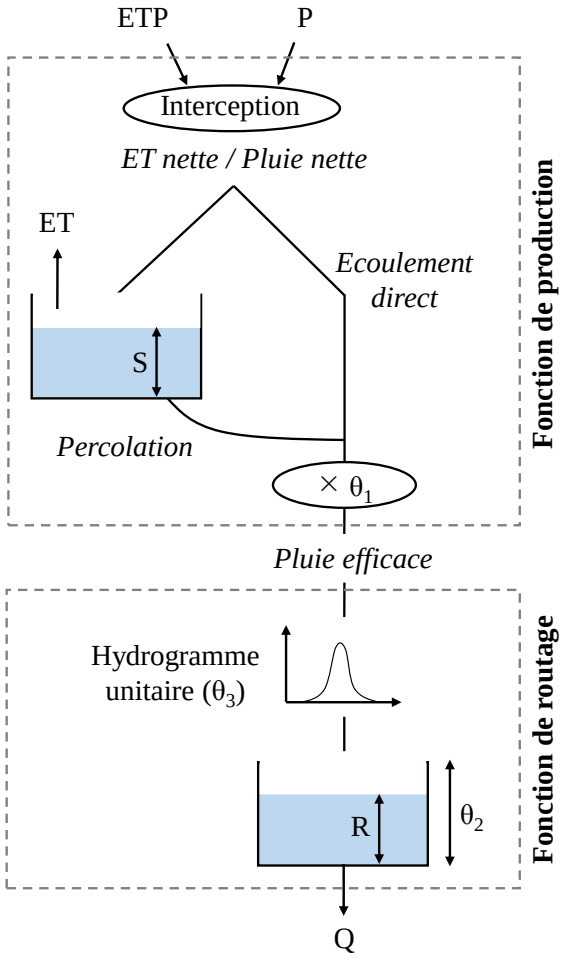

FIGURE 3.5 – Structure du modèle GRP.

manière empirique l'humidité du bassin. Son niveau de remplissage, S, détermine la fraction de la pluie nette (si elle est positive) qui sera stockée, tandis que la partie restante accède directement au routage. L'évapotranspiration nette, si elle positive, peut également venir ponctionner de l'eau dans ce réservoir. Pour rendre GRP plus parcimonieux, sa capacité a été fixée par ses concepteurs. Le réservoir de production se vidange selon une fonction puissance du niveau de remplissage, pour donner un terme (appelé percolation) venant s'ajouter à l'écoulement direct. Cette somme est ensuite ajustée par un coefficient multiplicatif θ_1 qui doit être calé. Cet ajustement empirique, qui rend GRP non conservatif, vise à prendre en compte d'éventuels échanges avec d'autres systèmes hydrologiques non modélisés (par exemple, des pertes vers des nappes profondes).

La pluie efficace, résultat de la fonction de production, pénètre ensuite la fonction de routage. Celle-ci comprend une étape linéaire puis une étape non linéaire. Tout d'abord, un hydrogramme unitaire réalise une convolution de la pluie efficace, introduisant ainsi un retard, tout en la répartissant sur différents pas de temps successifs. La forme de cet hydrogramme unitaire est fixée dans GRP, tandis que son temps de base θ_3 est un paramètre à caler. Ce paramètre est normalement relié aux caractéristiques géomorpho-

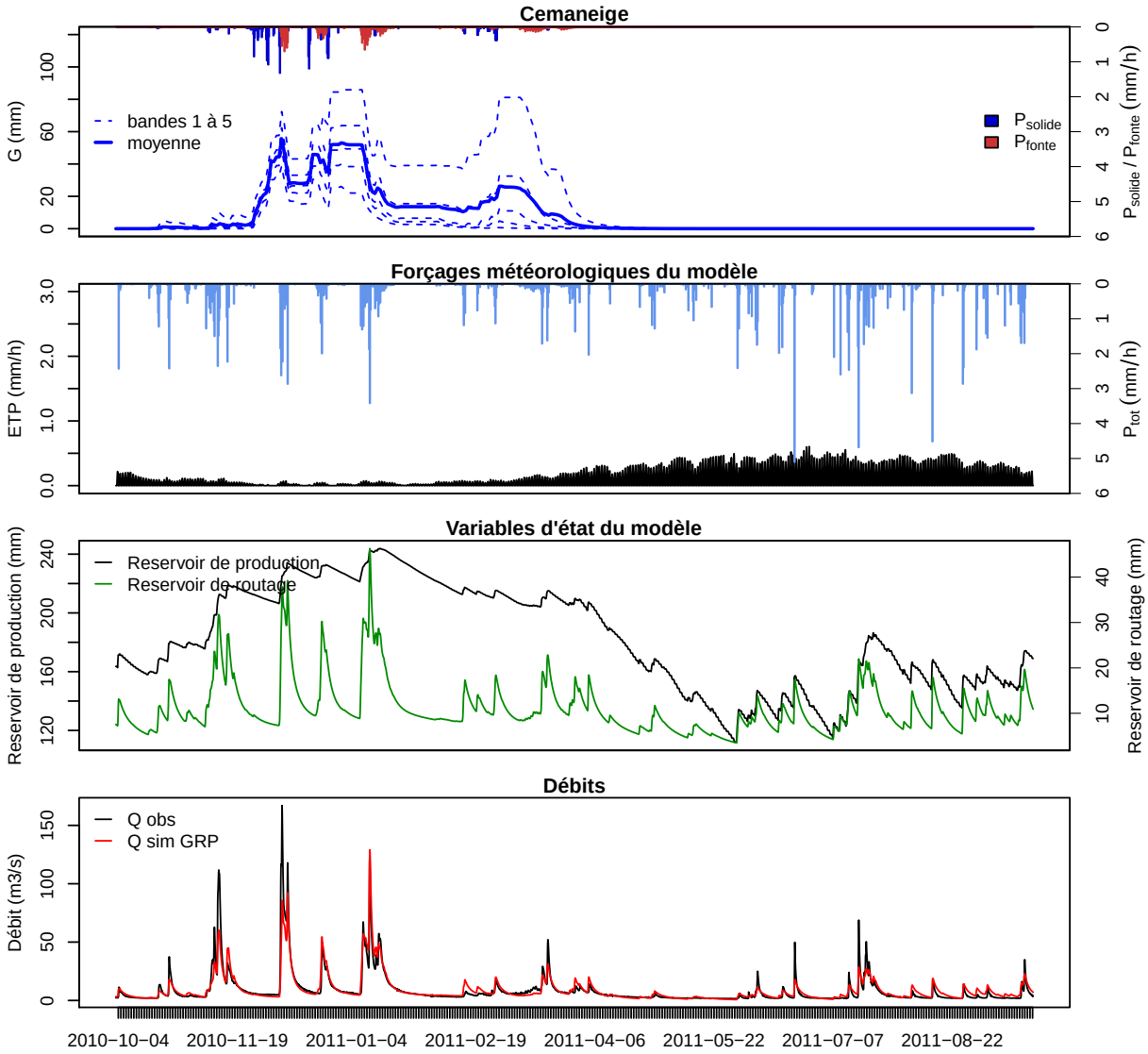

FIGURE 3.6 – Illustration du fonctionnement en simulation continue du modèle GRP couplé avec Cemaneige, sur une année hydrologique, pour le bassin de la Valserine.

logiques du bassin. La sortie de l'hydrogramme unitaire alimente ensuite un réservoir de routage, de remplissage R et de capacité θ_2 (paramètre à caler), dont la vidange suit une loi quadratique. Contrairement au réservoir de production qui réalise un suivi de l'humidité du bassin, et donc pour qui la notion de saturation a un sens, le réservoir de routage n'a pas vocation à se saturer. C'est seulement une manière empirique de représenter un compartiment de transfert en rivière.

En résumé, GRP dispose de 3 paramètres (θ_1, θ_2 et θ_3) et 2 variables d'états (S et R). Son fonctionnement en simulation continue, à l'instar des modèles GR dont il découle, s'appuie sur des opérateurs mathématiques qui ont été retenus non pas parce qu'ils représentent des processus physiques identifiés, mais parce qu'ils reproduisent correctement les débits simulés sur un très grand nombre de bassins versants. C'est donc, contrairement à TOPMODEL, un modèle bien davantage empirique que conceptuel, où l'interprétation physique de ses composants est souvent périlleuse. La Figure 3.6 illustre le fonctionnement

en simulation continue de GRP couplé à Cemaneige.

3.2.3.2 Mode prévision

Comme tous les modèles à réservoirs, GRP nécessite une période de warm-up pour initialiser, à la date de prévision t_{prev}, les niveaux S et R des réservoirs de production et de routage, respectivement.

En revanche, il se distingue (notamment des autres modèles GR) de par la présence d'une procédure spécifique d'assimilation. Celle-ci combine deux étapes, qui ont été proposées par Tangara (2005), puis testées et comparées à d'autres méthodes plus complexes par Berthet (2010). La première étape vise à mettre le modèle en conformité avec la situation hydrologique à t_{prev}. Le niveau R du réservoir de routage, à savoir la variable d'état la plus « en aval » du modèle, est mis à jour de manière à avoir un débit simulé à t_{prev} qui est strictement égal au débit observé. L'opération consiste à inverser la loi de vidange du réservoir de routage pour remonter au remplissage. Comme dans TOPMODEL (cf. 3.2.2.2), cette mise à jour repose uniquement sur le dernier débit observé et n'exploite pas les observations précédentes. Il n'est donc théoriquement pas impossible que modèle et observation partagent la même valeur à t_{prev} tout en présentant des dynamiques très différentes (par exemple, une récession pour l'un et une montée au pic pour l'autre).

La seconde étape vise à exploiter la dernière erreur de modélisation, en se basant sur le constat que cette erreur n'est pas un simple bruit blanc, mais une quantité qui présente certaines caractéristiques d'autocorrélation et d'hétéroscédasticité. GRP applique alors une correction auto-régressive multiplicative sur la prévision. L'effet de cette correction est non négligeable pour les premiers pas de temps, mais s'éteint beaucoup plus rapidement que celui de la mise à jour du réservoir de routage.

3.3 Présentation des modules communs aux 3 modèles

Les modèles ARX, TOPMODEL et GRP sont couplés avec deux outils communs : Cemaneige, pour la modélisation du manteau neigeux, et la formulation d'Oudin, pour le calcul de l'évapotranspiration potentielle.

3.3.1 Modélisation du manteau neigeux avec Cemaneige

Cemaneige est un module « neige » développé par Valéry (2010) pour fonctionner de pair avec un modèle hydrologique global, et ce dans le but d'améliorer la modélisation des débits à l'exutoire de bassins sur lesquels la neige joue un rôle non négligeable. Il fonctionne au pas de temps journalier, et est alimenté en entrée par des forçages de précipitation et de température.

Les précipitations doivent être cumulées sur 24 h et moyennées sur le bassin, tandis que les températures doivent être moyennées sur 24 h et correspondre à l'altitude médiane des bassins. Après avoir présenté le fonctionnement général de Cemaneige, nous décrivons les ajustements qui ont été apportés pour pouvoir le coupler avec nos modèles hydrologiques qui fonctionnent au pas de temps horaire.

3.3.1.1 Principe général de fonctionnement

Bien que ses forçages météorologiques soient moyennés sur le bassin, Cemaneige procède à une discrétisation en cinq bandes altitudinales d'égale surface (i.e., chaque bande représente 20 % de la surface du bassin). Les forçages globaux sont alors extrapolés à l'altitude médiane de chacune de ces bandes :

- pour les précipitations, via une correction multiplicative à l'aide d'un gradient altitudinal constant tout au long de l'année (Valéry, 2010, chapitre 6),
- pour les températures, via une correction additive à l'aide d'un gradient altitudinal qui dépend du jour de l'année (Valéry, 2010, chapitre 5). Leurs valeurs ont été tracées à la Figure 2.2, car déjà utilisées pour la construction de l'archive des températures observées à l'altitude médiane des bassins.

La suite de la modélisation s'effectue ensuite indépendamment sur chacune de ces bandes. La fraction solide/liquide des précipitations est d'abord estimée sur la base de la température : précipitations solides en dessous de 1°C ; liquides au dessus de 3°C ; mixte solides/liquides entre. La partie liquide participe à l'écoulement, tandis que la partie solide alimente le stock neigeux. Celui-ci est représenté de manière conceptuelle à l'aide d'un réservoir présentant deux états internes, le remplissage G et l'état thermique eT_G. Ce dernier permet éventuellement de retarder le déclenchement de la fonte si le manteau est trop froid. La fonte de ce stock neigeux est calculée à l'aide d'une méthode degrés-jour faisant intervenir deux paramètres à caler : le facteur de fonte, K_f, et le coefficient de pondération de l'état thermique du manteau, C_{T_G}. Le principe de l'approche degré-jour est de simuler une quantité de fonte journalière proportionnelle à l'écart entre la température de l'air et une température seuil dite « de fonte », fixée ici à 0°C.

Sur chacune des bandes d'altitude sont donc calculées deux quantités : la pluie liquide et la fonte. Une moyenne globale sur les cinq bandes est ensuite réalisée, avec des pondérations égales compte-tenu du fait que les bandes sont d'égales surfaces. Ces deux lames d'eau sont notées $P_{\text{liquide}_{24h}}$ et $P_{\text{fonte}_{24h}}$. Leur somme, notée $P_{\text{comp}_{24h}}$, est la lame d'eau complète sur la bassin qui participera à l'écoulement dans les modèles hydrologiques.

Pour davantage de détails concernant Cemaneige, nous invitons le lecteur à lire le chapitre 9 de Valéry (2010). La version du module qui a été utilisée est celle implémentée dans la package `airGR` (Coron *et al.*, 2017) de `R`.

3.3.1.2 Adaptations réalisées

Cemaneige fonctionne au pas de temps journalier, tandis que nos modèles hydrologiques fonctionnent au pas de temps horaire. Différentes adaptations permettant le couplage ont donc dû être apportées, qui sont schématisées à la Figure 3.7.

Agrégation journalière Soient P_{6h} et T_{6h} les forçages de précipitation et de température disponibles dans nos archives au pas de temps 6 h, et soient P_{24h} et T_{24h} les forçages acceptés par Cemaneige. Le passage au pas de temps 24 h est effectué par le biais d'une somme dans le cas des précipitations, et d'une moyenne dans le cas des température.

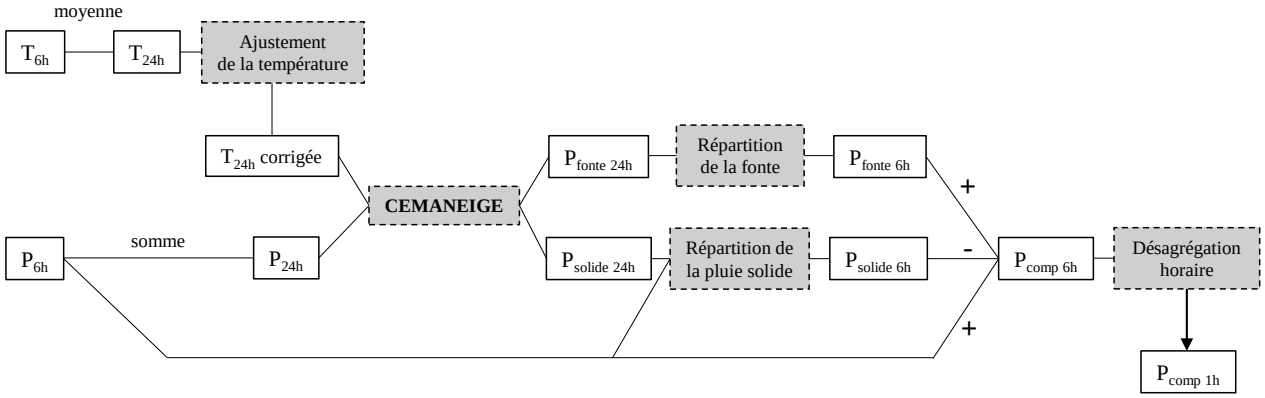

FIGURE 3.7 – Illustration schématique des adaptations réalisées pour utiliser Cemaneige au pas de temps horaire.

Ajustement de la température Avant d'alimenter Cemaneige, un traitement additionnel est réalisé sur la température T_{24h}. Il faut en effet rappeler que nos données de température (observations comme prévisions) sont des valeurs en points de grille, issues de modèles météorologiques, qui ont été extrapolées jusqu'à l'altitude médiane du bassin considéré (cf. équation (2.1)). Cette extrapolation, basée sur l'altitude du point de grille dans le modèle, peut mener à des biais non négligeables. Nous laissons ainsi la possibilité à notre couplage de Cemaneige d'ajuster T_{24h} de manière à corriger d'éventuels biais systématiques. Cet ajustement est additif seulement, et fait intervenir un paramètre supplémentaire à caler, noté Z_{pt}. Celui-ci qui peut être vu comme l'altitude corrigée du point de grille des données de température avant extrapolation.

Les adaptations suivantes concernent la désagrégation des sorties de Cemaneige du pas de temps 24 h au pas de temps horaire.

Répartition de la fonte Une première fonction se charge de la répartition de la lame d'eau de fonte issue de Cemaneige, $P_{\text{fonte}_{24h}}$, du pas de temps 24 h au pas de temps 6 h. Nous avons appliqué une répartition qui distribue 10, 20, 40 et 30 % de $P_{\text{fonte}_{24h}}$ sur les créneaux 00-06, 06-12, 12-18 et 18-00 h UTC, respectivement. Ces coefficients ont été choisis de manière à ce que les débits reproduisent au mieux les fluctuations journalières lors des cycles de fonte. La seule exception est le bassin de l'Arve sur lequel une répartition homogène sur tous les créneaux a été choisie. En effet, par sa surface et l'étalement de son altitude (cf. tableau 2.1), l'Arve produit une lame d'eau de fonte dont le signal journalier est extrêmement variable selon quelles parties du bassin participent à la fonte. Nous avons jugé la reproduction de ce signal trop complexe et en dehors du cadre de notre travail, d'où une répartition homogène.

Répartition de la pluie solide Une seconde fonction se charge ensuite de la répartition de la partie solide des précipitations, $P_{\text{solide}_{24h}}$, du pas de temps 24 h au pas de temps 6 h. Cette fonction, itérative, comprend les étapes suivantes :

1. On attribue $P_{\text{solide}_{6h}} = P_{\text{solide}_{24h}}/4$ à chacun des 4 créneaux.

2. On calcule le « trop-versé », c'est-à-dire la somme des différences $P_{\text{solide}_{6h}} - P_{6h}$ sur les 4 créneaux, puis on fixe pour chacun des créneaux $P_{\text{solide}_{6h}} = P_{6h}$ si $P_{\text{solide}_{6h}} > P_{6h}$. L'objectif ici est de ne pas allouer sur un créneau davantage de précipitations solides qu'il n'y avait de précipitations totales (avant Cemaneige).

3. On répartit de manière homogène le « trop-versé » sur les créneaux où $P_{\text{solide}_{6h}} < P_{6h}$.

4. Retour à l'étape 2, sauf si le « trop-versé » est égal à zéro.

Il convient de noter que cette approche pourrait être améliorée par l'exploitation des données de température T_{6h}.

Les lames d'eau complètes au pas de temps 6 h, notées $P_{\text{comp}_{6h}}$, peuvent alors être calculées via $P_{\text{comp}_{6h}} = P_{6h} - P_{\text{solide}_{6h}} + P_{\text{fonte}_{6h}}$. La dernière étape consiste alors à désagréger cette quantité au pas horaire.

Désagrégation au pas horaire Ne disposant pas de données permettant une désagrégation plus fine, nous supposons une répartition homogène de $P_{\text{comp}_{6h}}$ sur les 6 créneaux horaires.

3.3.2 Calcul de l'ETP horaire

Les modèles hydrologiques TOPMODEL et GRP doivent être forcés avec des données d'évapotranspiration potentielle (ETP) au pas de temps horaire. L'évapotranspiration ne joue pas un rôle crucial dans la prévision hydrologique à un horizon de quelques jours, car elle est généralement inférieure d'un ordre de grandeur aux précipitations qui causent des réactions hydrologiques. Cependant, elle est indispensable pour estimer les condition initiales du bassin lors de l'initialisation des modèles hydrologiques. Pour le calcul de ces forçages, nous avons utilisé la formule d'ETP journalière « de Oudin » (Oudin *et al.*, 2005), puis effectué une désagrégation au pas de temps horaire.

La formule de Oudin, qui utilise comme seule donnée d'entrée la température, est la suivante :

$$\text{ETP}_{24h} = \begin{cases} \frac{R_e}{28.5} \frac{T_{24h}+5}{100} & \text{si } T_{24h} + 5 > 0 \\ 0 & \text{sinon,} \end{cases} \tag{3.8}$$

où ETP_{24h} est l'évapotranspiration potentielle (en mm/j) et R_e est la radiation extraterrestre (en $\text{MJ/m}^2\text{/j}$), dont le calcul ne dépend que de la latitude et du jour de l'année (Morton, 1983, annexe C).

Pour obtenir les valeurs d'ETP au pas horaire, nous utilisons les coefficients de répartition qui sont utilisés dans la version publique du modèle GRP. Ces coefficients, tracés à la Figure 3.8, permettent d'avoir une ETP maximale autour de midi, et une ETP nulle pendant la nuit.

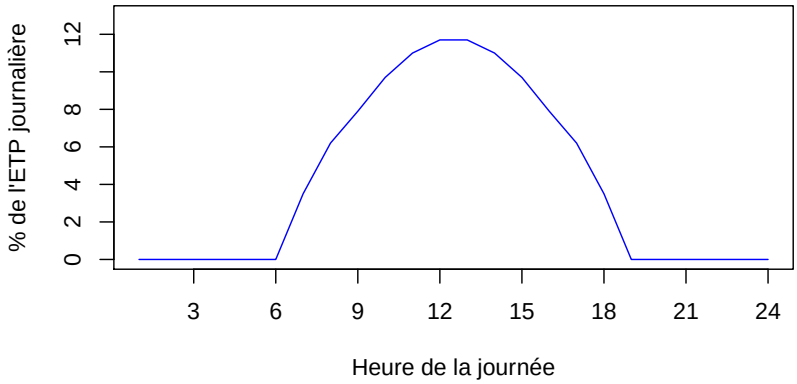

FIGURE 3.8 – Coefficients de répartition de l'ETP journalière au pas de temps horaire, qui proviennent de la version publique de GRP.

3.4 Stratégie de calage

Le calage d'un modèle hydrologique, qui plus est lorsqu'il est appelé à fonctionner en mode prévision, place le modélisateur devant un certain nombres de questions. Nous décrivons brièvement ici les choix qui ont été faits. Précisons que ces choix nous ont semblé être cohérents, mais ne prétendent aucunement être optimaux. En effet, notre problématique de thèse considère les modèles hydrologiques davantage comme des outils à disposition que des composants sur lesquels rechercher les meilleures performances. Néanmoins, cela ne nous empêche pas de discuter de certaines alternatives vers lesquelles le modélisateur peut espérer, selon nous, quelques gains.

3.4.1 Modèles couplés : calage itératif ou simultané ?

Chacun des trois modèles hydrologiques est couplé avec le module neige Cemaneige qui dispose, dans notre configuration, de 3 paramètres à caler (K_f, C_{T_G} et Z_{pt}). Faut-il alors caler ces paramètres en même temps que ceux du modèle hydrologique, ou bien les caler au préalable ? Du point de vue strictement pratique, le calage préalable des paramètres de Cemaneige requiert des données d'observations pouvant être rapportées aux sorties de Cemaneige. Par exemple, Riboust *et al.* (2018) apporte une modification à Cemaneige pour que celui-ci puisse exprimer la surface enneigée du bassin versant. Il expérimente alors le calage de Cemaneige en utilisant les observations satellitaires de couvert nival MODIS. Du point de vue théorique cependant, il semble judicieux de laisser ses paramètres « interagir » avec ceux du modèle hydrologique, car l'on attend du module neige qu'il améliore la prévision des débits. C'est donc la stratégie du calage simultané qui a été adoptée.

Il y a néanmoins une exception concernant le modèle ARX, dont la version à notre disposition sur les bassins d'étude a été calée par CNR, en utilisant un module neige différent de Cemaneige. Nous n'avons pas recalé ce modèle, et par conséquent les paramètres de Cemaneige ont été calés séparément (ce qui joue en sa défaveur par rapport aux deux autres modèles hydrologiques).

Une alternative possible, testée par Riboust *et al.* (2018), consiste à réaliser un calage simultané, mais avec un critère de calage qui est la moyenne pondérée d'un critère portant sur les débits et un critère portant sur la surface enneigé. Cela semble améliorer la simulation de l'enneigement par Cemaneige (le module est alors plus robuste), sans détériorer significativement les performances en débits. Cette piste n'a cependant pas été étudiée, car nous ne souhaitions pas rajouter une source de données supplémentaire.

3.4.2 Faut-il caler en simulation continue ou en mode prévision ?

Nous avons vu que dans un contexte de prévision, les modèles hydrologiques (couplés avec le module neige) pouvaient fonctionner soit en simulation continue, soit dans un mode prévision qui inclut une procédure d'initialisation. La question du mode de fonctionnement dans lequel réaliser le calage se pose alors.

Dans le cadre du développement de GRP, Tangara (2005) puis Berthet (2010) ont montré que l'assimilation de la donnée de débit contrôlait une grande partie de la prévision sur les premiers pas de temps, et donc qu'elle faisait partie intégrante du modèle. Ils conseillent alors de caler le modèle en mode prévision, les performances en validation étant légèrement meilleures. C'est d'ailleurs cette stratégie qui est implémentée dans la version publique de GRP. Cependant, elle implique le choix d'une échéance « cible » sur laquelle réaliser le calage. Le jeu de paramètres optimal est alors légèrement différent du jeu optimal issu du calage en simulation continue, car il s'adapte à l'assimilation pour l'échéance cible choisie. Est-il cependant adapté pour les autres échéances ? Berthet (2010) montre qu'une échéance cible choisie à 48 h permet d'obtenir, sur les échéances précédentes, des performances peu dégradées par rapport à si ces échéances avaient été considérées comme échéances cibles. Il n'étend cependant pas le test à des échéances plus lointaines.

Nous comprenons ainsi qu'un calage en mode prévision peut mener à de meilleures performances, mais que cela requiert un certaine vigilance sur la stabilité des performances au regard de la variété des échéances sur lesquelles les prévisions sont amenées à être utilisées. Cette remarque s'applique également à TOPMODEL, dont la procédure d'initialisation inclue une routine d'assimilation qui est fondée sur le même principe que GRP, à savoir la mise à jour d'un réservoir. Un calage en simulation continue risque peut-être de ne pas exploiter pleinement les routines d'assimilation, mais il a l'avantage, en écartant toute notion d'échéance, de ne faire aucune hypothèse quant à l'utilisation future des prévisions [2].

Nous avons donc opté pour la stratégie du calage en simulation continue. Nous réservons au module de post-traitement la tâche que de traiter les erreurs de modélisation hydrologique selon l'échéance.

Le modèle ARX fait exception, car sa structure lui impose un calage un peu particulier. En effet, l'espace des paramètres (cf. 3.4.4) d'un modèle ARX est extrêmement complexe,

2. Un autre avantage est que le calage en simulation continue est plus rapide que le calage en mode prévision, d'un facteur qui dépend de l'horizon (l'échéance maximale) et de la fréquence d'émission des prévisions ou, en d'autres termes du « chevauchement » des prévisions sur la période de calage.

à cause du processus autorégressif, mais surtout de la segmentation des paramètres selon, entre autre, la valeur du dernier débit. Il n'est alors pas possible, du point de vue numérique, de réaliser un calage autre qu'avec une échéance cible de 1 h. La stratégie de calage en simulation continue, simultanément sur les paramètres ARX et les paramètres Cemaneige, n'est donc pas possible. Nous avons ainsi conservé le calage du modèle ARX qui a été réalisé par CNR, et avons calé les paramètres Cemaneige séparément, en simulation continue.

3.4.3 Fonction-objectif

Le calage d'un modèle hydrologique vise à déterminer le jeu de paramètres qui lui permet de reproduire « au mieux » les débits observés. Derrière cette locution adverbiale se trouve nécessairement un critère quantitatif, que le calage cherche à optimiser (c'est-à-dire minimiser ou maximiser, selon les critères) ; c'est la fonction-objectif.

La plus utilisée en hydrologie est le critère de Nash, ou *Nash-Sutcliffe Efficiency* (NSE ; Nash et Sutcliffe, 1970), qui quantifie la somme des erreurs au carrés. Son équation est la suivante :

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{n}(Q_{\mathrm{sim},t} - Q_{\mathrm{obs},t})^2}{\sum_{t=1}^{n}(Q_{\mathrm{obs},t} - \mu_{\mathrm{obs}})^2} \tag{3.9}$$

où n est le nombre total de pas de temps, $Q_{\mathrm{sim},t}$ et $Q_{\mathrm{obs},t}$ sont respectivement les débits simulés et observés au pas de temps t, et μ_{obs} est la moyenne des débits observés sur les n pas de temps. Le fait de normaliser le terme d'erreur par la variance des observations n'est en rien utile au calage, mais permet au modélisateur de comparer entre elles des valeurs de NSE calculées sur des bassins dont l'amplitude des erreurs serait très différente. Une valeur de 1 signifie que le modèle reproduit parfaitement les observations, tandis qu'une valeur de 0 indique qu'il fait aussi bien qu'un modèle naïf qui simulerait systématiquement la moyenne μ_{obs} des observations. La procédure de calage cherche donc à maximiser ce critère.

Un des reproches principalement adressé au NSE est de favoriser la simulation des forts débits au détriment des bas débits. Du fait de l'élévation au carré des erreurs, le calcul du NSE est en effet contrôlé par à un faible nombre de pas de temps sur lesquels l'amplitude de l'erreur est grande, une situation qui concerne davantage les gammes de débits élevés (l'erreur étant hétéroscédastique). Une alternative consiste alors à calculer le NSE sur les débits transformés, de manière à réduire l'hétéroscédasticité. Le modélisateur s'intéressant en priorité aux bas débits pourra utiliser la transformation logarithme, tandis que celui cherchant une solution intermédiaire entre bas et haut débits pourra choisir la racine carrée. Oudin *et al.* (2006) ont montré que cette dernière solution était un compromis intéressant, et par conséquent nous avons réalisé le calage des modèles hydrologiques en maximisant le critère $\mathrm{NSE}_{\sqrt{Q}}$, défini par :

$$\mathrm{NSE}_{\sqrt{Q}} = 1 - \frac{\sum_{t=1}^{n}(\sqrt{Q_{\mathrm{sim},t}} - \sqrt{Q_{\mathrm{obs},t}})^2}{\sum_{t=1}^{n}(\sqrt{Q_{\mathrm{obs},t}} - \mu_{\mathrm{obs}_{\sqrt{Q}}})^2} \tag{3.10}$$

où $\mu_{\mathrm{obs}_{\sqrt{Q}}}$ est la moyenne de la racine carré des débits observés sur les n pas de temps.

Les travaux théoriques et expérimentaux de Gupta *et al.* (2009) ont mis en avant des limites à l'usage du NSE qui sont de nature différente. Par un développement de l'équation (3.9), ils montrent que ce critère peut être décomposé en trois termes, qui mesurent la corrélation linéaire, le biais et la variabilité des débits. Dis autrement, le NSE peut se voir comme une fonction-objectif multi-critère. Cependant le problème réside dans le fait que ces critères ne sont pas indépendants, et de plus ont des poids variables. Gupta *et al.* (2009) mettent en avant plusieurs conséquences. Ainsi, la maximisation du NSE entraîne (i) une sous-estimation de la variabilité des débits, qui entraine alors une sous-estimation des pics de débits (ce qui est contre-intuitif au regard de la réputation du NSE à favoriser les hauts débits), et (ii) des erreurs potentiellement importantes sur les volumes, causées par un terme de biais qui a peu de poids dans certains cas (celui-ci étant normalisé par la variabilité des débits observés, son poids diminue pour les bassins « nerveux »). Une approche multi-critère où ceux-ci sont indépendants et ont des poids stables peut alors permettre d'éliminer ces deux défauts. C'est en poursuivant cet objectif que le critère KGE (*Kling–Gupta Efficiency* ; Gupta *et al.*, 2009) a été proposé. Ce dernier n'a cependant pas été utilisé dans cette thèse.

3.4.4 Algorithme de calage

Après avoir défini le critère qui nous permet de comparer deux jeux de paramètres, nous discutons de la stratégie adoptée pour déterminer le jeu optimal, c'est-à-dire celui qui va être utilisé pour la prévision. Compte-tenu du nombre de paramètres mais surtout de leurs interactions mutuelles, la recherche manuelle des paramètres optimaux est une tâche fastidieuse et chronophage, qui par ailleurs demande une grande expérience de la part du modélisateur. Cette tâche est donc bien souvent confiée à un algorithme, dont la résilience au travail rébarbatif est bien supérieure !

Avant toutes choses, définissons l'espace des paramètres comme un espace multi-dimensionnel où les coordonnées de chaque point définissent un jeu de paramètres. Grâce à la définition d'une fonction-objectif, nous pouvons faire correspondre à cet espace une « surface de réponse » du modèle hydrologique. L'objectif de l'algorithme de calage est alors d'explorer cet espace jusqu'à converger vers l'optimum de la surface de réponse. Si cette définition semble trop abstraite, imaginons un modèle avec 2 paramètres X et Y, et soit $Z(X, Y)$ la fonction-objectif. La surface de réponse du modèle est alors analogue au relief d'un terrain, où l'altitude seraient représentée par $Z(X, Y)$, et les coordonnées géographiques par X et Y. Ce sont les coordonnées du point culminant de ce relief que l'algorithme de calage cherche à déterminer (dans le cas d'une maximisation de la fonction-objectif).

La première étape consiste à borner l'espace des paramètres en établissant pour chacun d'eux des plages de valeurs admissibles. Ensuite vient l'exploration de cet espace. Des stratégies naïves d'exploration sont la discrétisation régulière ou encore l'exploration aléatoire. Cependant, le nombre d'itérations nécessaires pour trouver l'optimum (de la surface de réponse) augmente exponentiellement avec le nombre de paramètres à caler, ce qui rend généralement ces stratégies inapplicables pour la calage des modèles hydrologiques. Des algorithmes ont alors été développés pour se déplacer dans l'espace des paramètres en

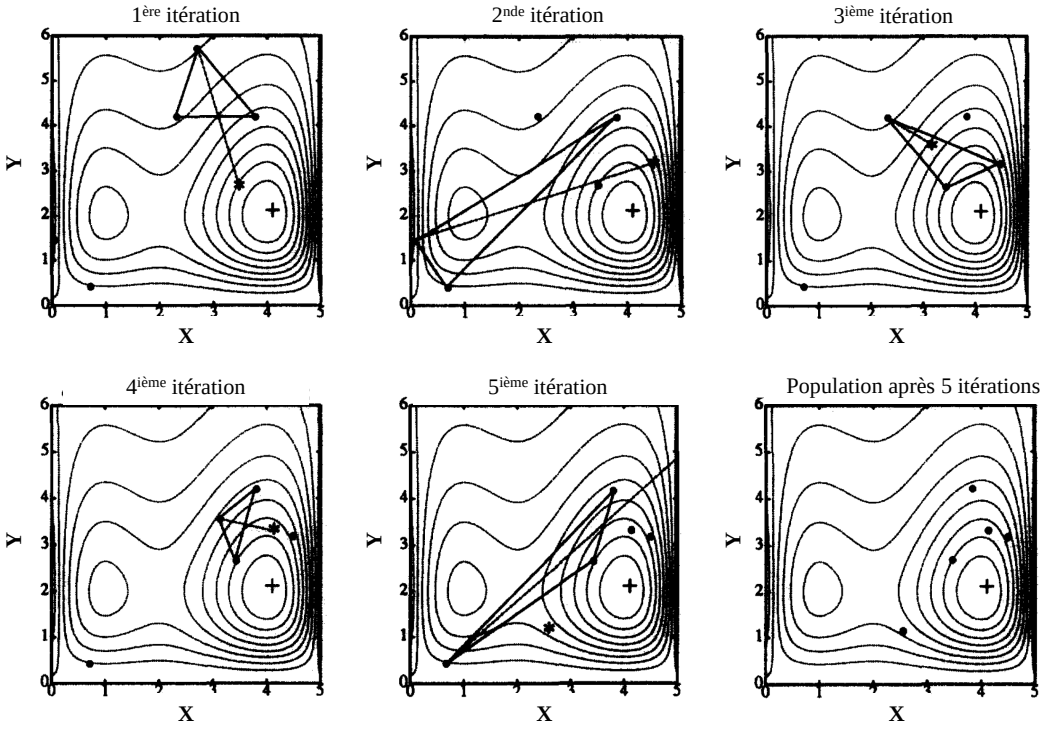

FIGURE 3.9 – Illustration de la méthode du simplex, qui fait évoluer une population de points en se rapprochant de l'optimum dans l'espace des paramètres (ici en dimension 2). Le symbole $*$ représente les nouveaux points crées lors de chaque itération. Figure adaptée de Duan *et al.* (1994).

suivant la direction d'amélioration de la fonction-objectif. On retrouve notamment la méthode du simplex (Nelder et Mead, 1965), illustrée à la Figure 3.9, ou encore la méthode « pas-à-pas » (Michel, 1989), qui est fréquemment utilisée dans le calage des modèles GR. Ces méthodes, dites *locales*, peuvent cependant être mises en défaut lorsque la surface de réponse du modèle est particulièrement complexe et présentant des pièges tels que par exemple des optimums locaux, ou encore des zones d'insensibilités.

Des méthodes dites *globales* ont été développées, qui sont plus coûteuses en calcul mais explorent une partie beaucoup plus grande de l'espace des paramètres. Cela peut consister, par exemple, à combiner une approche locale avec une stratégie multi-départ (Perrin, 2000). Dans cette thèse, nous utilisons la méthode du SCE-UA (*Shuffle Complex Evolution algorithm, University of Arizona* ; Duan *et al.*, 1994), considérée dans la littérature comme la plus efficace des méthodes globales. Cette méthode fait appel à des concepts d'évolution naturelle d'une population constituée d'individus. Nous définissons un individu comme un point dans l'espace des paramètres. Le SCE-UA débute en créant aléatoirement une population d'individus, puis la divise en plusieurs sous-groupes appelés complexes. Chacun de ces complexes évolue avec une méthode locale, la méthode du simplex (Figure 3.9). Ainsi, à chaque itération, certains individus sont remplacés par d'autres qui sont plus proches d'un optimum. Périodiquement, les complexes sont brassés de manière aléatoire (des individus passent d'un complexe à un autre), puis recommencent à évoluer de manière indépendante. Ce brassage permet de partager, entre les différents complexes, l'information sur les directions d'amélioration de la fonction-objectif. L'évo-

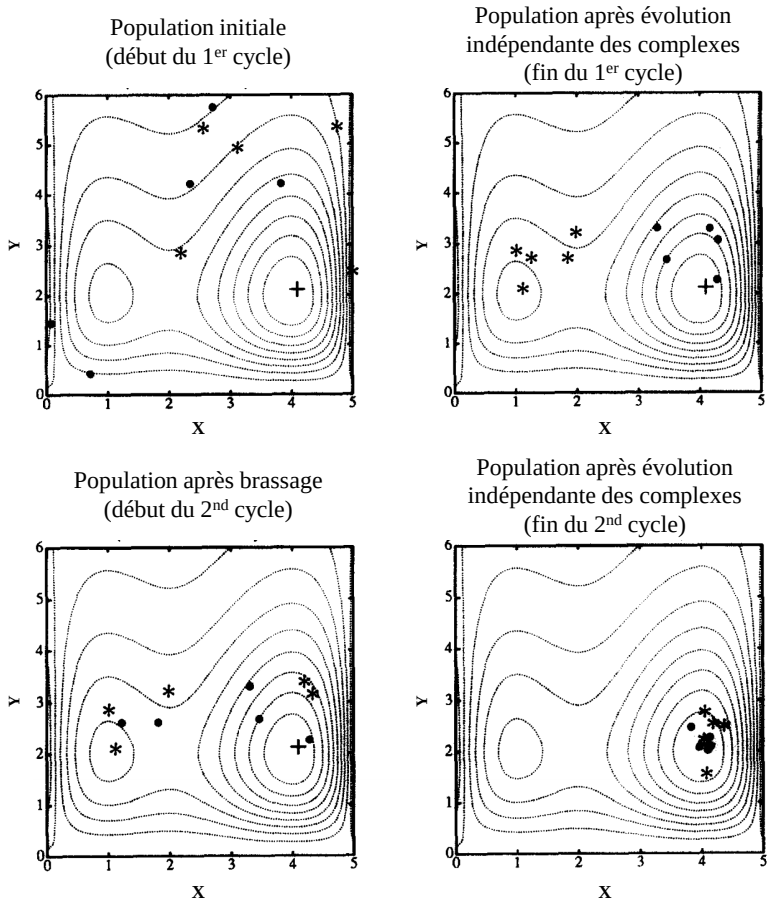

FIGURE 3.10 – Illustration de l'évolution d'une population dans la méthode SCE-UA, pour 2 complexes (représentés par les symboles $*$ et $\bullet$). Figure adaptée de Duan *et al.* (1994).

lution de la population dans la méthode SCE-UA est illustrée à la Figure 3.10, pour un exemple avec 2 complexes.

Nous avons utilisé l'algorithme de SCE-UA implémenté dans la fonction `SCEoptim` du package `hydromad` de `R`, avec les paramètres suivants :

- 5 complexes,
- $2N_{par} + 1$ individus par complexe,
- $2N_{par} + 1$ itérations avant brassage,
- un critère de convergence de 0.001, sur 5 étapes de brassage,

où N_{par} est le nombre de paramètres à caler. Ce paramétrage a conduit à environ 1500 à 2000 simulations pour le calage, en moyenne par modèle et par bassin.

La question de l'intérêt d'utiliser une méthode globale plutôt qu'une méthode locale peut néanmoins être posée, les méthodes locales étant bien moins gourmande en simulations. Mathevet (2005) a notamment réalisé une inter-comparaison de méthodes, en considérant plusieurs modèles hydrologiques de complexité différente et un très large échantillon de bassins. Il observe que des méthodes globales comme le SCE-UA ne donnent

des résultats de calage que marginalement meilleurs qu'avec la méthode locale pas-à-pas, et même parfois inférieurs en validation. l'auteur explique les résultats inférieurs en validation par la capacité des méthodes globales à aller chercher avec précision la position de l'optimum, alors que la position de celui-ci au sein d'une « zone optimale » dépend davantage des conditions hydrométéorologiques observées sur la période de calage. Ses travaux vont donc en faveur des méthodes locales, bien souvent capables d'atteindre cette zone optimale en un nombre bien plus faible d'itérations. Ainsi, de mauvais résultats pour un jeu de paramètres obtenus par une méthode locale doivent davantage inciter le modélisateur à revoir la structure de son modèle, plutôt que de changer son algorithme de calage.

3.5 Diagnostic des performances

Dans ce travail de thèse le calage des modèles hydrologiques a été réalisé simultanément avec Cemaneige (cf. 3.4.1), en simulation continue (cf. 3.4.2), selon le critère $NSE_{\sqrt{Q}}$ (cf. 3.4.3), et à l'aide de l'algotithme SCE-UA (cf. 3.4.4). Nous présentons désormais les performances obtenues sur les périodes de :

- **calage** : 1 septembre 2003 - 31 août 2006,
- **validation** : 1 septembre 2006 - 31 août 2009.

Dans la suite du manuscrit, nous utiliserons simplement le nom du modèle hydrologique (ARX, TOPMODEL ou GRP) pour parler du couplage de celui-ci avec les deux outils communs que sont Cemaneige et la formulation d'ETP d'Oudin.

3.5.1 En simulation continue

Les performances en simulation continue (Figure 3.11) placent TOPMODEL et GRP très proches, avec des valeurs de $NSE_{\sqrt{Q}}$ qui varient entre 0.6 et 0.9 selon les bassins, de manière relativement stable entre les périodes de calage et de validation. En revanche, les performances du modèle ARX sont très en deçà, avec des valeurs de $NSE_{\sqrt{Q}}$ qui varient entre 0 et 0.6. Pour mémoire, ARX est un modèle hydrologique intrinsèquement conçu pour fonctionner en mode prévision, et il n'est donc pas surprenant de le voir peu performant en simulation continue. De plus, contrairement à TOPMODEL et GRP, les paramètres de ARX ne sont pas calés simultanément avec ceux de Cemaneige, ce qui joue théoriquement en sa défaveur.

3.5.2 En mode prévision

Ce sont les performances en mode prévision qui vont intéresser prioritairement le prévisionniste. Chaque modèle hydrologique est alors évalué de pair avec sa routine d'initialisation, et par conséquent les performances sont désormais dépendantes de l'échéance. La Figure 3.12 présente les performances obtenues sur la période de validation, en considérant l'ensemble des simulations émises à 00 h UTC. Cette heure correspond au cycle

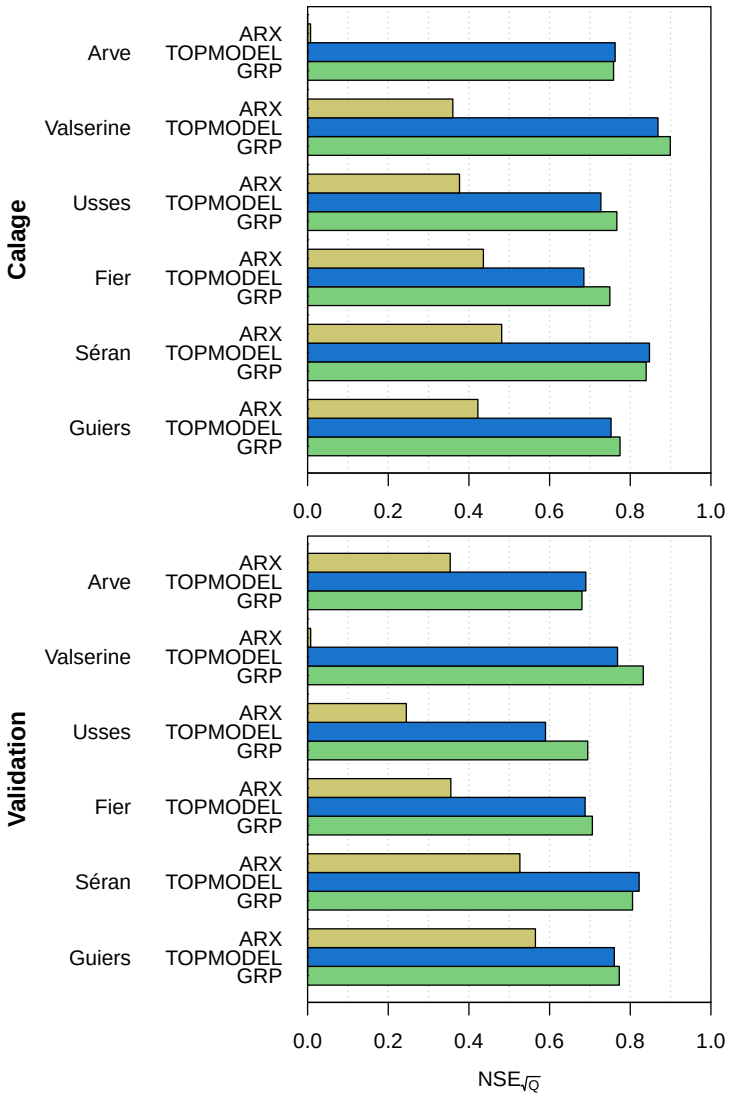

FIGURE 3.11 – Performances des modèles hydrologiques en simulation continue.

d'initialisation des modèles météorologiques qui seront utilisés pour forcer les modèles hydrologiques. On observe que les performances des modèles décroissent logiquement avec l'échéance, à mesure que l'effet de l'initialisation du modèle diminue. De manière générale, GRP est le modèle qui montre les meilleures performances. La différence avec TOPMODEL est très faible sur certains bassins comme les Usses ou le Séran, tandis qu'elle est plus significative sur d'autres bassins comme le Guiers. Enfin, sur l'ensemble des bassins les performances de ARX sont très proches de TOPMODEL et GRP (voir parfois supérieures) sur les premières échéances, c'est-à-dire jusqu'à 24/48h, ce qui correspondait aux échéances cibles pour CNR lorsque les modèles ARX ont été déployés. Les performances de ARX diminuent ensuite plus rapidement, pour tendre vers une valeur de $NSE_{\sqrt{Q}}$ en simulation continue qui est bien inférieure à celle des autres modèles.

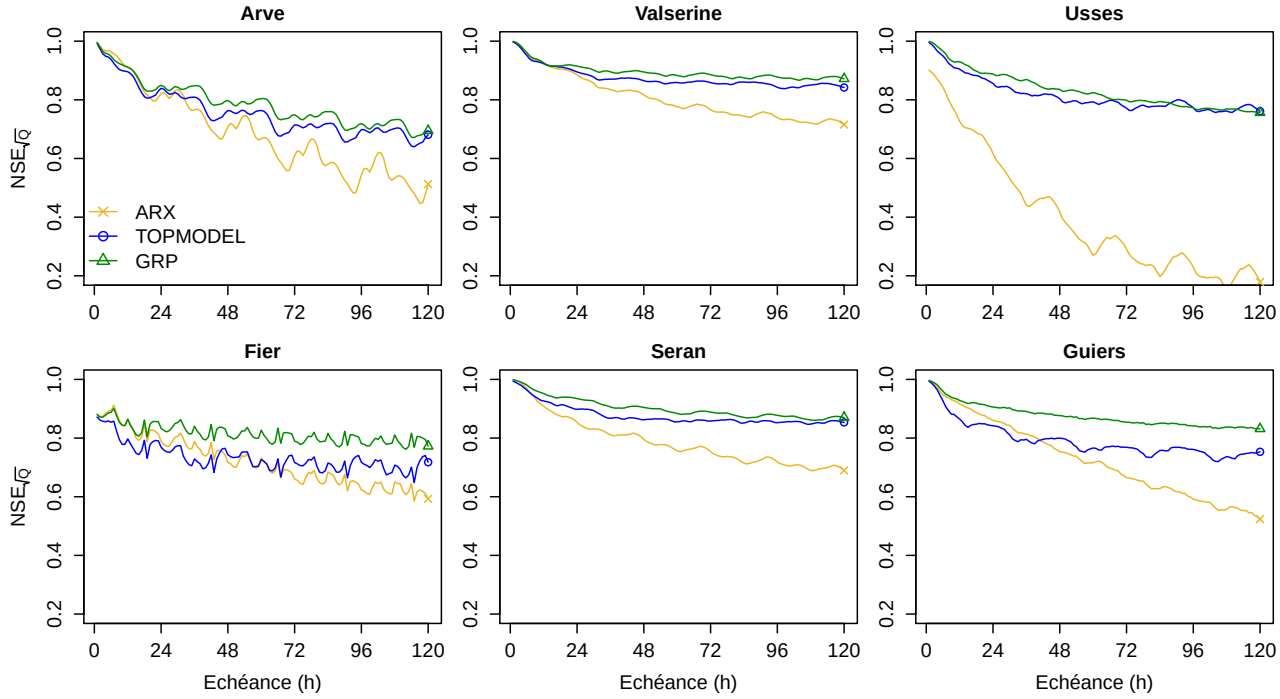

FIGURE 3.12 – Performances des modèles hydrologiques en mode prévision, sur la période de validation uniquement.

3.6 Synthèse

L'objectif de ce chapitre était de présenter les outils nécessaires à la modélisation hydrologique. Ils comprennent trois modèles hydrologiques (ARX, TOPMODEL et GRP), qui sont couplés avec un module neige (Cemaneige) et une formulation de l'évapotranspiration potentielle (celle de Oudin) communs. Ce sont des modèles globaux, qui fonctionnent au pas de temps horaire.

ARX est un modèle purement statistique, qui repose sur le processus d'autocorrélation des débits. TOPMODEL et GRP sont ensuite deux modèles conceptuels, mais qui empruntent des approches très différentes de modélisation de la relation pluie-débit. Le premier s'inspire de lois physiques, tandis que le second s'appuie sur des relations davantage empiriques.

Cette variété dans la structure des modèles sera exploitée dans la partie III, où nous adopterons une stratégie multi-modèle en préalable du post-traitement des prévisions de débit. Cela permettra alors de prendre en compte, de manière non statistique, une partie de l'incertitude de modélisation hydrologique.

Chapitre 4

Outils de vérification

Une prévision est nécessairement vouée à se confronter à l'observation, car c'est seulement ainsi que l'on peut juger de sa qualité. Cette confrontation est au centre d'un champ thématique particulier, la *vérification*. Cette appellation, née dans le domaine de la météorologie, provient du fait que les prévisionnistes « vérifient » ce qu'ils produisent.

Les éléments de ce chapitre sont fondamentaux, car ils permettent de retranscrire de manière objective ce que nous qualifions jusqu'à maintenant en des termes vagues : émettre de « bonnes » prévisions, ou encore « améliorer » les prévisions.

La section 4.1 aborde quelques grands principes de la vérification pour rapidement se recentrer sur les seuls aspects qui nous intéressent dans cette thèse. Ainsi, les deux outils que nous utiliserons le plus par la suite, le CRPS et l'histogramme de rang, sont présentés dès la section 4.2. La section 4.3 traite ensuite d'une étude concernant un aspect particulier de la vérification : la *stratification*. Cette étude, rédigée sous la forme d'un article (en anglais), est une contribution méthodologique aux travaux présents dans la littérature, tandis que le reste du chapitre vise davantage à présenter des outils de vérification déjà bien connus dans la communauté. La section 4.4 propose ensuite une extension au contexte multivarié, qui est fondamentale pour l'évaluation de la cohérence (spatiale, temporelle, et inter-variable) des prévisions. Enfin, la section 4.5 fait la synthèse de ce chapitre.

4.1 L'approche de vérification dans notre contexte

4.1.1 De l'intérêt d'évaluer les prévisions

La vérification vise à évaluer les performances d'un système de prévision. Cela peut être pour la personne qui conçoit ce système un moyen d'attester de son travail, comme pour l'utilisateur des prévisions une manière de choisir tel système plutôt qu'un autre. Le terme d'« utilisateur » n'est pas anodin ici, car il est synonyme de caractéristiques particulières qui peuvent influer la vérification. Ainsi, nous distinguons l'évaluation de :

- la **qualité intrinsèque** : son évaluation ne concerne que l'adéquation entre les prévisions et les observations correspondantes,
- la **utilité**, ou **valeur économique** : son évaluation intègre les prévisions dans un

contexte de prise de décision, qui est propre à chaque utilisateur. C'est donc ce que rapporte (ou coûte) la prévision à un utilisateur spécifique qui est évalué, d'où le terme « économique ».

L'approche basée sur la valeur économique ne peut s'appliquer qu'en bout de la chaîne, car c'est à partir des prévisions finales que seront prises les décisions. D'autre part, elle nécessite la caractérisation précise du problème de prise de décision, par exemple via la définition d'une fonction de coût. Dans cette thèse, nous ne cherchons pas à nous identifier à un utilisateur particulier, mais à améliorer les méthodes de prévision de manière à ce qu'elles conviennent au plus grand nombre. Pour cette raison, nous utiliserons uniquement des outils visant à évaluer la qualité intrinsèque des prévisions.

Par ailleurs, nous proposons de séparer les outils de vérification selon s'ils poursuivent des objectifs :

- de **quantification** (et donc de **classification**) : ces outils, appelés scores de vérification, mesurent de manière quantitative la qualité des prévisions. Ils peuvent être utilisés en validation, dans le but de comparer des systèmes de prévisions entre eux, mais également en calage, pour déterminer par exemple les paramètres optimaux d'une correction statistique.

- de **diagnostic** : ces outils, sous la forme de graphiques, cherchent à illustrer le comportement des prévisions, et ainsi aident à identifier d'éventuels défauts que l'on pourra essayer de corriger par la suite. Il est parfois possible de résumer un graphique par une unique valeur, auquel cas cela devient un score. Cependant, l'information est nécessairement réduite, et par conséquent le potentiel de diagnostic est moindre.

Nous avons fait le choix d'utiliser un score quantitatif, le CRPS (cf. 4.2.1), ainsi qu'un outil de diagnostic, l'histogramme de rang (cf. 4.2.2).

4.1.2 De quelle grandeur fait-on la prévision ?

La vérification peut porter sur des grandeurs de différentes natures. Cette grandeur définit une variable aléatoire dont on fait la prévision, et dont les observations seront des réalisations. Cette variable aléatoire peut être :

- **catégorielle** : elle ne peut alors prendre qu'un nombre fini de valeurs, qui représentent des catégories, ordonnées ou non. On parle de variable *nominale* lorsque les catégories ne sont pas ordonnées entre elles. Par exemple, si l'on s'intéresse à la phase des précipitations, les catégories peuvent être : précipitations liquides (catégorie A), solides (catégorie B), ou bien pas de précipitations (catégorie C). On parle de variable *ordinale* lorsque les catégories sont ordonnées entre elles. Un exemple est la prévision du dépassement (catégorie A) ou du non-dépassement d'un certain seuil (catégorie B) par le débit d'une rivière.

- **continue** : la variable peut alors prendre un nombre infini de valeurs, qui sont ordonnées entre elles. Par exemple, une prévision de débit de 32.7 m^3/s est une prévision continue.

La prévision quantitative de variables hydrométéorologiques (précipitations, température, débit) ne concerne généralement pas des variables catégorielles nominales. En revanche, on a le choix de faire porter la vérification sur des variables catégorielles ordinales ou bien sur des variables continues.

L'approche catégorielle requiert la définition d'un ou plusieurs seuils permettant de définir les catégories, ce qui introduit dans la vérification une donnée supplémentaire qui est propre à chaque utilisateur. Cette approche entre donc en contradiction avec notre objectif qui est d'écarter toute identification à un utilisateur particulier. Ainsi, nous privilégions la vérification de prévisions de grandeurs continues.

Pour le moment, nous nous plaçons dans un cadre univarié, où la grandeur sur laquelle porte la vérification est un scalaire. Nous verrons dans la section 4.4 qu'il est utile, pour l'évaluation de la cohérence des prévisions probabilistes, de basculer dans un cadre multivarié où la grandeur à prévoir n'est plus un scalaire mais un vecteur.

4.1.3 La forme des prévisions

Une prévision **déterministe** est constituée d'une seule valeur représentant la meilleure estimation possible de la variable aléatoire à prévoir. La prévision et l'observation sont alors de même nature, ce qui rend facilement appréhendable la notion de distance entre les deux.

Dans cette thèse cependant, nous nous intéressons à des prévisions **probabilistes**, qui associent une probabilité à n'importe quelle réalisation de cette variable aléatoire. La notion de distance entre prévision et observation est alors plus complexe, les deux objets étant de nature différente. Tandis qu'une prévision déterministe ne peut être par définition que juste ou fausse, une prévision probabiliste n'est ni juste ni fausse ; ce n'est qu'en accumulant un nombre de réalisation suffisamment élevé que l'on pourra juger de sa qualité.

Une prévision probabiliste peut prendre différentes formes. Tout d'abord, elle peut être continue, via une loi de probabilité à laquelle est rattachée une fonction de répartition $F(x)$ (ou bien une densité de probabilité $f(x)$, au choix). Dans le contexte de la prévision hydrométéorologique, la prévision probabiliste est bien plus souvent définie de manière discrète, sous la forme d'un ensemble $\mathbf{x} = x_1, \ldots, x_M$ auquel est rattaché la fonction de répartition empirique $\hat{F}(x)$:

$$\hat{F}(x) = \frac{1}{M} \sum_{m=1}^{M} H(x - x_m) \tag{4.1}$$

où H est la fonction de Heaviside, c'est-à-dire $H(u) = 0$ pour $u < 0$ et $H(u) = 1$ sinon. Le passage d'un format à l'autre est possible : nous pouvons ajuster une loi de distribution sur l'ensemble $\mathbf{x}$ pour passer de la représentation discrète à la représentation continue, tout comme il est possible d'échantillonner la distribution continue pour obtenir une ensemble discret.

Les outils de vérification peuvent s'avérer sensibles à cette discrétisation. C'est notamment le cas du CRPS dont nous parlerons plus loin (en 4.2.1), qui favorise les prévisions d'ensemble où le nombre de membres est élevé (Ferro *et al.*, 2008). Or, il est fréquent de vouloir comparer des prévisions d'ensemble de tailles différentes. La question est alors la suivante : faut-il chercher à éliminer l'effet du nombre de membres, ou bien considère-t-on le nombre de membres comme faisant parti de la prévision, auquel cas les prévisions doivent être évaluées dans leur forme d'origine ? Éliminer l'effet du nombre de membres peut consister, par exemple, à ajuster une distribution paramétrique sur chaque prévision d'ensemble puis calculer le score sur cette dernière, ou encore apporter un facteur correctif sur le score de manière à ce qu'il tienne compte du nombre de membres, comme le proposent Ferro *et al.* (2008).

Nous prenons le parti, dans cette thèse, d'évaluer les prévisions probabilistes dans le format où elles sont disponibles, c'est-à-dire sans chercher à éliminer l'effet du nombre de membres. Ainsi, des prévisions d'ensemble au nombre élevé de membres sont avantagées du strict point de vue statistique (vis-à-vis du calcul des scores), mais nous considérons cela comme une récompense pour leur capacité à mieux décrire les lois de probabilité prévues.

4.1.4 La fiabilité et la finesse, deux attributs essentiels

Lorsque l'on cherche à évaluer la qualité intrinsèque de prévisions, il est intéressant d'analyser spécifiquement différentes propriétés, appelées attributs. Dans notre contexte de prévision probabiliste de variables continues, Gneiting *et al.* (2007) ont proposé un cadre de vérification fondé sur deux attributs, la fiabilité et la finesse.

4.1.4.1 La fiabilité

Soit une prévision probabiliste représentée par une densité de probabilité f, et soit y l'observation correspondante, considérée comme une réalisation d'une variable aléatoire de densité de probabilité g. Gneiting *et al.* (2007) suggèrent de voir g comme le « mécanisme générateur de données » (en anglais : *data generating process*) que propose la nature pour ce cas de prévision. Nature dont on fait l'hypothèse qu'elle est omnisciente, dans le sens où le prévisionniste ne peut pas avoir accès à davantage d'information qu'elle.

La fiabilité (en anglais : *calibration*, ou encore *reliability*) correspond à la similitude entre les densités f et g (Jolliffe et Stephenson, 2003). Toute disparité entre ces deux densités équivaut à un défaut de fiabilité, les deux principaux correspondant à des écarts sur les deux premiers moments : la moyenne et la variance. Lorsque la moyenne de f est inférieure à celle de g, on parle de biais négatif, et dans le cas inverse de biais positif. Si la variance de f est inférieure à celle de g, on parle alors de sous-dispersion, et enfin dans le cas inverse de sur-dispersion. Ces quatre défauts caractéristiques de fiabilité sont illustrés à la Figure 4.1.

Pour une prévision f donnée, nous n'avons pas accès à la loi g, mais seulement à une unique réalisation y. Ainsi, par définition, la fiabilité ne peut être évaluée que sur un grand nombre de réalisations.

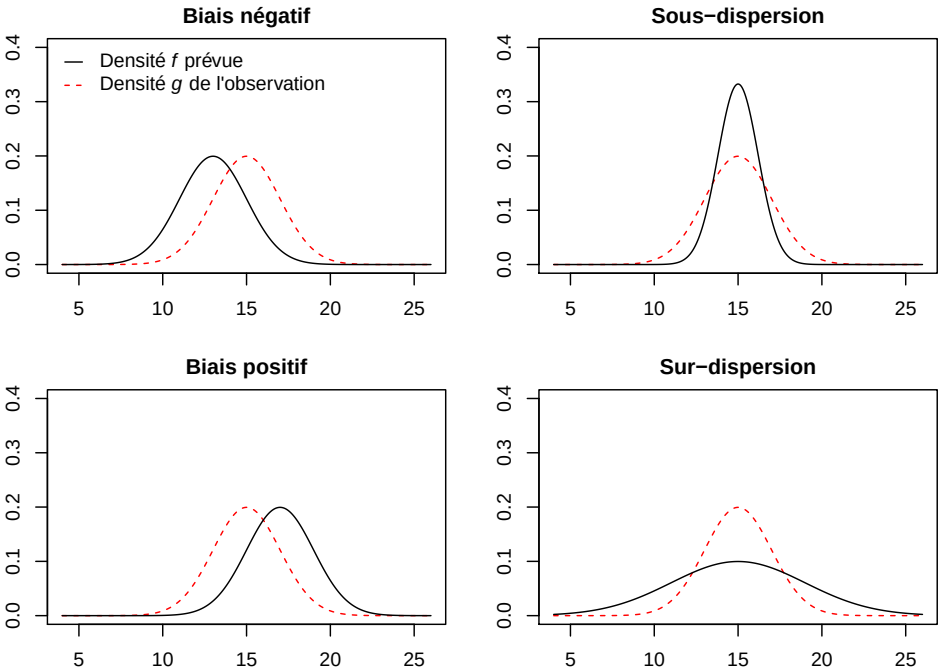

FIGURE 4.1 – Défauts caractéristiques de fiabilité concernant la moyenne (gauche) ou la variance (droite) de la prévision, dans le cas idéal où l'on connait la densité de l'observation. L'axe horizontal représente la variable aléatoire à prévoir, et l'axe vertical la densité de probabilité.

La difficulté réside alors dans le fait que les densités f prévues sont différentes pour chacune des réalisations, nous empêchant ainsi de collecter suffisamment d'observations y correspondantes à une même densité f pour pouvoir estimer la loi g et ainsi mesurer sa similitude. La plupart des outils d'évaluation de la fiabilité se basent alors sur une définition moins stricte de la fiabilité, en cherchant à caractériser le comportement moyen des réalisations y parmi une variété de densités prévues f. On dit alors que le système de prévision est fiable si les quantiles x % des prévisions f sont dépassés dans x % des cas par les observations y sur un grand nombre de réalisations.

Des auteurs ont proposé de faire le distinguo entre ces deux types de fiabilité. Murphy et Epstein (1967), Yates (1982) et Bontron (2004) parlent de fiabilité *à petite échelle* (en anglais : *in-the-small*) lorsque $f = g$ quelles que soient les prévisions f, et de fiabilité *à grande échelle* (en anglais : *in-the-large*) lorsque les quantiles x % sont effectivement dépassés dans x % des cas par les observations sur l'ensemble des réalisations. On peut démontrer trivialement que la fiabilité à petite échelle implique celle à grande échelle, mais que l'inverse n'est pas vrai.

Nous verrons que le concept de stratification, abordé dans la section 4.3, est une approche intéressante pour tenter de se rapprocher de l'évaluation de la fiabilité à petite échelle.

4.1.4.2 La finesse

La finesse (en anglais : *sharpness*) est directement liée à la dispersion des prévisions probabilistes. Elle représente, en quelque sorte, leur degré de certitude, indépendamment

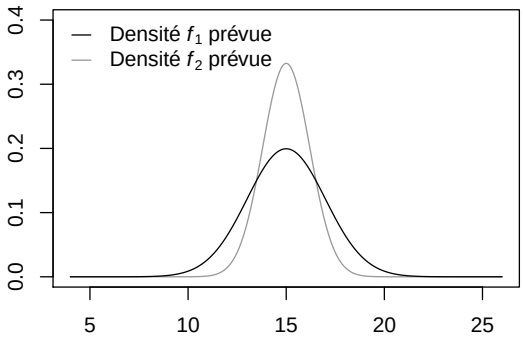

FIGURE 4.2 – Densités de probabilité de deux prévisions de finesse différentes : la prévision f_2 est plus fine que la prévision f_1.

des observations. Une finesse élevée correspond alors à une faible dispersion (Figure 4.2). Notez qu'il n'y a pas lieu, contrairement à la fiabilité, de parler de « défauts de finesse », car la finesse est un attribut qui dépend uniquement des prévisions.

4.1.4.3 Le paradigme de Gneiting *et al.* (2007)

Il est à priori désirable d'émettre des prévisions probabilistes qui soient les plus fines possibles, car cela signifie que l'incertitude sur la variable à prévoir est faible. Par exemple, dans un contexte de prise de décision où le dépassement d'un seuil implique l'émission d'une alerte, un utilisateur préférera des prévisions fines, car synonymes de probabilités de dépassement proches de 0 ou de 1. Cependant, des prévisions fines mais peu fiables ont peu d'intérêt, car l'utilisateur sait alors que la probabilité de dépassement émise ne correspond pas à la probabilité réelle de dépassement. A l'inverse, des prévisions parfaitement fiables mais peu fines, comme par exemple les prévisions climatologiques, sont d'un intérêt très relatif pour cet utilisateur, car elles ne lui apportent aucune information supplémentaire par rapport à ce qu'il sait déjà.

Nous comprenons donc que fiabilité et finesse doivent aller de pair que pour la qualité des prévisions soient élevée. Gneiting *et al.* (2007), au travers de leur paradigme « *of maximizing the sharpness of the predictive distributions subject to calibration* », stipulent que le but de la prévision probabiliste est de s'assurer de la fiabilité avant de chercher à maximiser la finesse.

C'est en s'inscrivant dans ce paradigme que nous définissons les contours de notre « boîte à outils » de vérification. Ainsi, nous avons besoin de :

1. un score quantitatif global qui évalue simultanément la fiabilité et la finesse des prévisions, pour pouvoir classifier différents systèmes de prévision, mais également pour déterminer le paramétrage d'un système lors d'éventuelles phases de calage. Ce score doit être capable, d'une part, de récompenser parmi deux systèmes de prévision fiables celui qui émet les prévisions les plus fines, mais également de pénaliser suffisamment un système qui rechercherait la finesse au détriment de la fiabilité.

2. un outil qui permette de diagnostiquer d'éventuels défauts de fiabilité dans les prévisions.

Il ne nous semble pas indispensable d'évaluer spécifiquement la finesse des systèmes de prévisions. En effet, le choix de la meilleure finesse ne peut être fait que parmi des systèmes diagnostiqués comme étant fiables, et auquel cas le score global est suffisant.

Maintenant que nous avons défini les besoins, nous présentons les outils retenus, à savoir le CRPS et l'histogramme de rang.

4.2 Un score quantitatif et un outil de diagnostic

Supposons un échantillon de vérification contenant N couples prévision-observation. Pour chaque réalisation $n \in \{1, \ldots, N\}$, l'observation est notée par y_n, tandis que la prévision est représentée par la fonction de répartition F_n dans le cas continu, et par l'ensemble $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,M})$ dans le cas discret. La taille M des ensembles est supposée constante.

4.2.1 Le Continuous Ranked Probability Score (CRPS)

4.2.1.1 Définition

Le *Continuous Ranked Probability Score* (CRPS ; Matheson et Winkler, 1976; Hersbach, 2000; Gneiting et Raftery, 2007) est une métrique qui mesure, pour chaque réalisation $n \in \{1, \ldots, N\}$ de l'échantillon de vérification, une « distance » entre la fonction de répartition de la prévision et celle de l'observation, qui est alors une fonction créneau (l'incertitude de l'observation étant considérée comme nulle [1]). Il est défini par :

$$\mathrm{CRPS}_n(F_n, y_n) = \int_{-\infty}^{+\infty} \left[F_n(x) - H(x - y_n)\right]^2 \, \mathrm{d}x \tag{4.2}$$

où H est la fonction de Heaviside, c'est-à-dire $H(u) = 0$ pour $u < 0$ et $H(u) = 1$ sinon.

Le CRPS peut s'interpréter graphiquement, comme l'illustre la Figure 4.3. Il correspond en effet à la somme de l'aire entre la fonction de répartition $F_n(x)^2$ et l'axe des abscisses pour $x < y_n$, et l'aire entre $1 - (1 - F_n(x))^2$ et la courbe d'ordonnée $F(x) = 1$ pour $x > y_n$.

Étant négativement orienté (une prévision parfaite a un CRPS nul), il est important de l'interpréter comme une « pénalité » que le prévisionniste cherchera à minimiser. Il s'exprime dans la même unité que l'observation, et par conséquent il peut être vu comme une généralisation de l'erreur absolue, vers laquelle il converge lorsque la prévision est déterministe [2].

1. L'introduction d'une incertitude autour de l'observation est un axe de travail fort intéressant, mais n'est pas abordé dans cette thèse.

2. Cette propriété l'autorise donc à servir de score de comparaison entre un système de prévision probabiliste et un système déterministe.

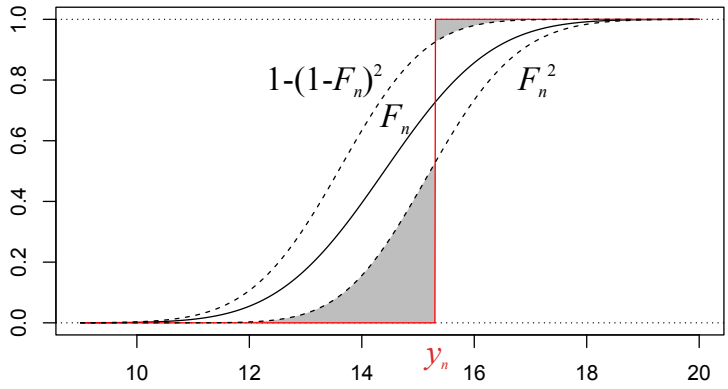

FIGURE 4.3 – Interprétation graphique du CRPS, qui correspond à l'aire grisée. L'axe horizontale représente la variable à prévoir.

Il est souvent écrit dans la littérature que « le CRPS est un score permettant d'évaluer la qualité globale des prévisions, car il mesure simultanément leur fiabilité et leur finesse ». Cette phrase fait deux raccourcis importants, que nous proposons d'expliciter ci-dessous.

Premièrement, une prévision probabiliste isolée n'est par définition ni bonne ni mauvaise, et par conséquent le CRPS n'a de réelle valeur que lorsqu'il est moyenné sur un large échantillon de prévisions. Ce CRPS moyen, noté $\overline{\text{CRPS}}$, est alors défini par

$$\overline{\text{CRPS}} = \frac{1}{N} \sum_{n=1}^{N} \text{CRPS}_n. \tag{4.3}$$

Ainsi, c'est bien la quantité $\overline{\text{CRPS}}$ qui est une mesure de la qualité globale des prévisions.

Le second raccourci concerne justement le CRPS comme outil de mesure de la qualité globale des prévisions, et notamment des attributs fiabilité et finesse. Le reste de cette section, qui pourra sembler quelque peu abstrait, vise à résumer les travaux qui ont permis de le démontrer. Un lecteur pressé qui serait prêt à nous croire sur parole peut ainsi passer directement à la partie 4.2.1.2.

Considérons y_n et F_n, présents dans l'équation 4.2, comme des réalisations de respectivement la variable aléatoire $y \in \Omega$ et de la fonction de répartition elle aussi aléatoire $F \in \mathcal{P}$, où Ω représente l'espace des valeurs possibles prises par l'observation, et $\mathcal{P}$ l'espace rassemblant toutes les distributions de probabilité qu'il est possible d'affecter sur Ω. Un score $\mathcal{S}$ (en anglais : *scoring rule* ; Gneiting et Raftery, 2007) est défini comme toute fonction $\mathcal{S} : \mathcal{P} \times \Omega \to \mathbb{R}$ qui assigne la valeur $\mathcal{S}(F, y)$ au couple $\{F, y\}$ contenant la prévision et l'observation. Le CRPS est un score (parmi d'autres) qui s'inscrit dans le cas où $\Omega = \mathbb{R}$. Il fait même partie d'une famille plus restreinte de scores, celle des scores strictement justes (en anglais : *strictly proper*).

Un score $\mathcal{S}$ est *juste* si, pour toute réalisation $F_n \in \mathcal{P}$, l'espérance de $\mathcal{S}(F_n, y)$ pour y suivant la loi de distribution définie par G_n, notée $\mathbb{E}_{G_n}[\mathcal{S}(F_n, y)]$, est minimisée pour $F_n = G_n$. Il est *strictement juste* si ce minimum est unique (Gneiting et Raftery, 2007). Cette propriété signifie donc qu'un prévisionniste qui connaîtrait la « vrai » distribution de l'observation n'a pas intérêt à émettre une prévision autre que cette même distribution.

En d'autres termes, cela l'encourage à traduire sa conviction réelle dans sa prévision, et non pas à opter pour une stratégie visant premièrement à optimiser le score.

Bröcker (2009) a étudié l'espérance mathématique des scores strictement justes, $\mathbb{E}[\mathcal{S}(F, y)]$, avec F et y aléatoires. Il a démontré qu'il était possible de la décomposer en trois termes :

$$\mathbb{E}[\mathcal{S}(F, y)] = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}, \tag{4.4}$$

où :

- Le terme UNC (pour *uncertainty*) dépend uniquement de la distribution non conditionnelle de y, appelée plus communément la climatologie. C'est en fait l'espérance du score que l'on obtiendrait si la climatologie était systématiquement considérée comme la prévision.

- Le terme REL (pour *reliability*), compare la distribution F avec la distribution de y conditionnelle à F. Il quantifie donc les défauts de fiabilité, et ce dans sa définition la plus stricte, c'est-à-dire la fiabilité à petite échelle (cf. 4.1.4.1). REL vaut zéro lorsque les prévisions sont parfaitement fiables.

- Le terme RES (pour *resolution*) est un terme orienté positivement, qui est une fonction croissante de la finesse de F sous la condition que les prévisions conservent leur fiabilité. Il est appelé ainsi car il quantifie la capacité du système produisant F à « résoudre » le problème de prévision par rapport à la climatologie. Pour un système qui émettrait systématiquement la prévision climatologique, qui est parfaitement fiable mais a une capacité de résolution nulle, nous aurions $\mathrm{REL} = 0$ et $\mathrm{RES} = 0$, et donc nous retrouverions $\mathbb{E}[\mathcal{S}(F, y)] = \mathrm{UNC}$.

Les travaux de Bröcker (2009) montrent donc que le CRPS, en tant que score strictement juste, est théoriquement capable de récompenser, parmi deux systèmes fiables, celui qui émet les prévisions les plus fines (via le terme RES), et par ailleurs de pénaliser (via le terme REL) les systèmes qui chercheraient la finesse au détriment de la fiabilité.

Dans la pratique, on dispose d'un échantillon contenant N réalisations $\{F_n, y_n\}$, et donc l'espérance $\mathbb{E}[\mathrm{CRPS}(F, y)]$ se traduit comme la moyenne sur les N réalisations, à savoir $\overline{\mathrm{CRPS}}$. Il s'avère alors que la décomposition de Bröcker (2009) est difficilement applicable, car elle requiert de disposer pour chaque prévision F_n de suffisamment d'observations y_n pour en estimer une distribution de probabilité. Nous en revenons ainsi au problème inhérent à l'évaluation de la fiabilité à petite échelle.

Hersbach (2000) a ainsi proposé une décomposition qui semble à première vue s'inscrire dans le schéma de Bröcker (2009), mais qui diffère quelque peu de manière à la rendre applicable en pratique. Cette décomposition, qui s'applique directement sur $\overline{\mathrm{CRPS}}$, a un terme REL qui quantifie non plus la fiabilité à petite échelle mais celle à grande échelle, dont la définition est moins stricte. Le terme RES n'en reste pas moins une fonction croissante de la finesse sous la condition que les prévisions conservent leur fiabilité (à grande échelle).

4.2.1.2 Calcul numérique de CRPS_n

La formulation (4.2) ne permet pas explicitement de calculer $\mathrm{CRPS}_n(F_n, y_n)$. Nous discutons désormais des différentes solutions pratiques. Tout d'abord, il est possible, lorsque la distribution F_n est paramétrique, de développer l'équation (4.2) de manière à obtenir une formulation analytique de $\mathrm{CRPS}_n(F_n, y_n)$ en fonction des paramètres de F_n. Ces développements ont par exemple été proposés par Gneiting *et al.* (2005), Friederichs et Thorarinsdottir (2012) et Scheuerer et Hamill (2015a) dans le cas des distributions Normale, GEV et Gamma, respectivement.

Lorsque les prévisions sont sous forme ensembliste en revanche, plusieurs solutions sont possibles. La première, que nous avons adopté dans cette thèse, consiste à discrétiser le calcul de l'intégrale dans l'équation (4.2). Nous utilisons la méthode de discrétisation proposée par Hersbach (2000) et implémentée dans la fonction `crpsDecompostion` du package `verification` de `R`. L'interprétation graphique qui lui est associée est donnée à la Figure 4.4.

L'alternative à cette première solution consiste à s'appuyer sur une formulation alternative du CRPS dont Gneiting *et al.* (2007) ont montré l'égalité avec la formulation (4.2) :

$$\mathrm{CRPS}_n(F_n, y_n) = \mathbb{E}_{F_n}[|X - y_n|] - \frac{1}{2}\mathbb{E}_{F_n}[|X - X'|] \tag{4.5}$$

où X et X' sont des variables aléatoires indépendantes de fonction de répartition F_n. Lorsque la prévision est sous la forme d'un ensemble $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,M})$, nous pouvons traduire les espérances en sommes discrètes sur les membres (alors considérés comme des réalisation indépendantes de la loi F_n) et ainsi calculer le CRPS via

$$\mathrm{CRPS}_n(\mathbf{x}_n, y_n) = \frac{1}{M}\sum_{m=1}^{M}|x_{n,m} - y_n| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}|x_{n,m} - x_{n,m'}|. \tag{4.6}$$

Nous verrons dans la section 4.4.2 que cette formulation permet de généraliser le CRPS à un contexte multivarié.

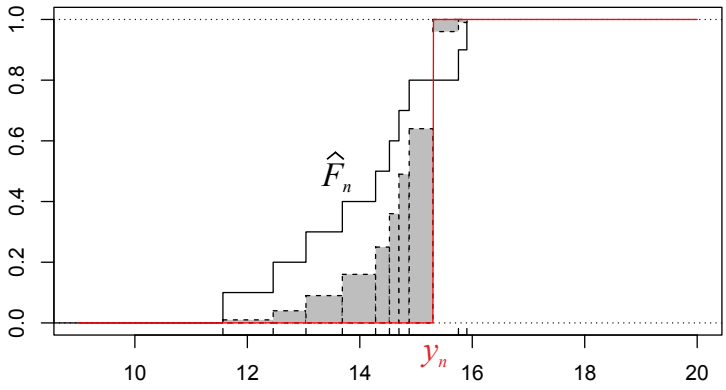

FIGURE 4.4 – Discrétisation du calcul du CRPS (aire grisée), lorsque la prévision probabiliste est sous forme non plus continue (Figure 4.3) mais discrète. Le nombre de membres est ici de 10.

4.2.1.3 Score de compétence

Le CRPS permet donc de comparer plusieurs systèmes de prévision probabiliste : celui ayant le plus faible $\overline{\text{CRPS}}$ sera celui produisant les prévisions de meilleure qualité. En revanche, l'interprétation d'une valeur de $\overline{\text{CRPS}}$ est peu évidente, d'autant plus qu'elle est fortement corrélée à l'amplitude des observations (Trinh *et al.*, 2013). Une autre limite est le fait que $\overline{\text{CRPS}}$ augmente sensiblement avec l'échéance, la qualité des prévisions se dégradant. Il est donc difficile d'interpréter sur un même graphique des valeurs de $\overline{\text{CRPS}}$ pour différentes échéances. Ainsi, il est plus aisé de ramener les performances de chaque système de prévision par rapport à celles d'un système de référence. C'est l'objectif visé par les scores de compétence (en anglais : *skill scores*).

Soit $\overline{S}$ le score moyen d'un échantillon de prévisions. Le score de compétence SS, vis-à-vis du score S, est défini par :

$$\text{SS} = \frac{\overline{S} - \overline{S}_{\text{ref}}}{\overline{S}_{\text{parfait}} - \overline{S}_{\text{ref}}} \tag{4.7}$$

où $\overline{S}_{\text{parfait}}$ est le score moyen obtenu avec des prévisions parfaites, et $\overline{S}_{\text{ref}}$ le score moyen obtenu avec des prévisions dites « de référence ».

Dans le cas du CRPS, le score de compétence associé, appelé CRPSS (*Continous Ranked Probability Skill Score*), s'exprime par

$$\text{CRPSS} = 1 - \frac{\overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}}, \tag{4.8}$$

car $\overline{\text{CRPS}}_{\text{parfait}} = 0$. Un CRPSS de 1 correspond ainsi à des prévisions parfaites, tandis qu'un CRPSS de 0 correspond à des prévisions tout juste aussi bonnes que les prévisions de référence. Un CRPSS négatif indique quant à lui que les prévisions sont de moins bonne qualité que les prévisions de référence.

Le choix des prévisions de référence conditionne donc les valeurs prises par le CRPSS. Si celles-ci sont trop « naïves », les valeurs de CRPSS seront artificiellement élevées. Pour construire cette référence, Pappenberger *et al.* (2015) distingue les approches basées sur la *climatologie*, la *persistance*, où l'*amélioration*. La première approche, qui utilise les distributions climatologiques de la variable à prévoir, est adaptée lorsque cette variable est peu corrélée aux dernières valeurs observées. C'est le cas des prévisions à longues échéances, ou plus simplement des prévisions pour des variables présentant une faible autocorrélation, comme les précipitations par exemple. En cas d'autocorrelation forte, des prévisions de référence fondées sur la persistance sont plus intéressantes, car plus difficiles à battre. Enfin, l'approche d'amélioration est adaptée lorsqu'il s'agit d'évaluer si les modifications apportées à un système de prévision (par exemple un post-traitement statistique) améliorent les prévisions.

Dans notre travail, nous utiliserons les prévisions climatologiques comme référence lors de l'évaluation des forçages météorologiques bruts. Ensuite, dès lors que nous évaluerons le gain apporté par telle ou telle procédure (pré-traitement, post-traitement, etc), nous

adopterons l'approche d'amélioration, et considérerons comme prévisions de référence les prévisions brutes (c'est-à-dire non « traitées » par la procédure en question).

4.2.2 L'histogramme de rang

L'histogramme de rang est un outil de diagnostic dont la finalité est d'évaluer la fiabilité des prévisions d'un échantillon. Il a été développé semble t-il de manière indépendante par Anderson (1996), Hamill et Colucci (1997) et Talagrand *et al.* (1997), d'où parfois l'appellation de « diagramme de Talagrand ». Hamill (2001) et Jolliffe et Stephenson (2003) ont ensuite contribué à le rendre extrêmement populaire dans la communauté de la prévision météorologique. Son attrait principal, en plus d'être non paramétrique, est d'évaluer les prévisions probabilistes dans leur forme la plus fréquente, à savoir la forme ensembliste.

L'histogramme de rang est fondé sur le principe suivant. Si un système de prévision est fiable (au sens le plus strict, c'est-à-dire à petite échelle) alors, pour chaque réalisation $n \in \{1, \ldots, N\}$ de l'échantillon de vérification, les membres $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,M})$ de l'ensemble et l'observation y_n peuvent être considérés comme des réalisations indépendantes d'une même loi de probabilité (qui est inconnue). Par conséquent, si l'on range ces $M+1$ valeurs dans l'ordre croissant, le rang de l'observation a autant de chance de prendre chacun des $\{1, 2, 3, \ldots, M+1\}$ rangs possibles.

Pour vérifier cette hypothèse, l'histogramme de rang se propose de relever, pour chacune des réalisations $n \in \{1, \ldots, N\}$, le rang prit par l'observation, puis d'en tracer l'histogramme. C'est donc un outil qui ne peut être construit que sur un échantillon suffisamment grand de couples prévision-observation. Il peut arriver (c'est même fréquent !) que l'affectation du rang de l'observation ne puisse se faire de manière unique, car sa valeur est égal à celle d'un ou plusieurs membres. Le rang est alors choisi aléatoirement parmi les valeurs possibles. Par exemple, soit $\mathbf{x}_n = (15.2, 13, 18, 9.9, 13)$ une prévision d'ensemble de taille $M = 5$, et soit $y_n = 13$ l'observation correspondante. Le vecteur des $M+1$ valeurs rangées dans l'ordre croissant, à savoir $(9.9, 13, 13, 13, 15.2, 18)$, contient donc trois valeurs égales. Le rang de l'observation sera alors tiré aléatoirement parmi les rangs $\{2, 3, 4\}$.

Si les N prévisions sont fiables, alors l'observation prendra environ $\frac{N}{M+1}$ fois chacun des $M+1$ rangs possibles, et par conséquent l'histogramme sera plat, mises à part les fluctuations dues à la taille finie de l'échantillon. Les défauts dans la fiabilité s'illustreront par des formes caractéristiques d'histogramme (Figure 4.5). En cas de sous-dispersion, les observations ont une probabilité plus forte de tomber dans les queues qu'au centre des distributions, et donc l'histogramme présentera une forme en $\cup$. A l'inverse, en cas de sur-dispersion l'histogramme présentera en forme de $\cap$. Enfin, les biais positifs et négatifs entraineront des histogrammes « penchés ».

Attention en revanche, l'histogramme de rang ne mesure pas la fiabilité à petite échelle mais seulement la fiabilité à grande échelle, du fait que les prévisions $\mathbf{x}_n$ sont différentes pour chaque réalisation $n \in \{1, \ldots, N\}$. Ainsi, si un défaut particulier de fiabilité implique une forme d'histogramme donnée (sous réserve que N soit suffisamment grand),

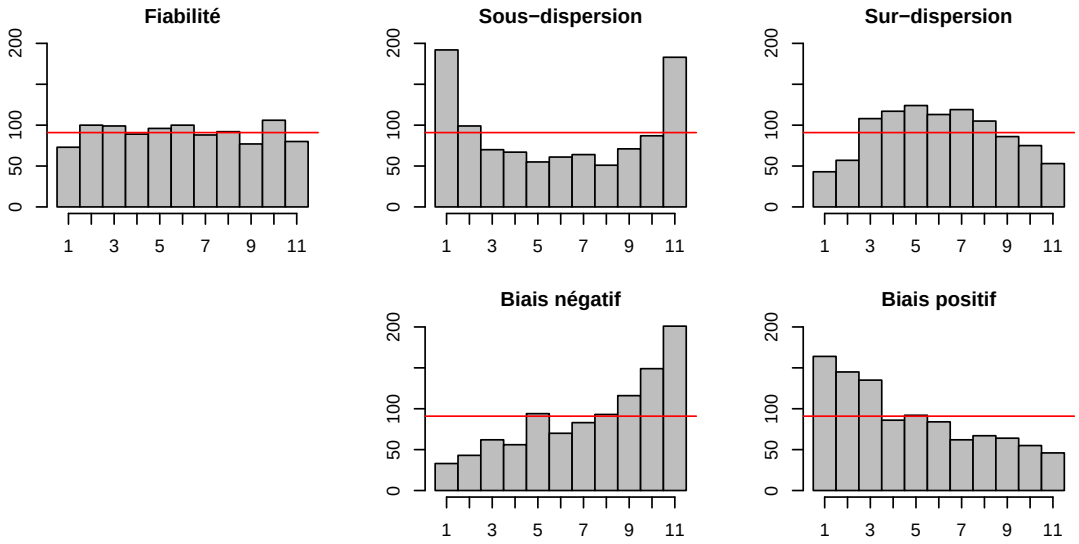

FIGURE 4.5 – Illustration de différentes formes caractéristiques d'un histogrammes de rang, avec $M = 10$ membres et $N = 1000$ réalisations. Sont indiqués en gras les défauts de fiabilité qui induisent ces histogrammes. La ligne rouge correspond à la fréquence $\frac{N}{M+1}$, qui est la fréquence attendue si les prévisions sont fiables.

l'inverse n'est pas vrai. Hamill (2001), dans un article fondateur quant à l'interprétation des histogrammes de rang, montre par exemple qu'un histogramme plat peut résulter de plusieurs défauts de fiabilité de natures différentes qui se compensent.

La section suivante de ce chapitre aborde le concept de *stratification* de l'échantillon de vérification, qui peut permettre à l'histogramme de rang, dans certains cas, de détecter de telles compensations. Nous précisons qu'une approche similaire a été discutée, dans un cadre de vérification de prévisions hydrologiques, par Bourgin (2014, chapitre 4). Cependant, nous n'avons eu connaissance de ces travaux qu'après l'écriture de l'article qui suit, et donc ce dernier n'y fait pas référence.

4.3 La stratification, avantages et écueils

Cette section correspond à un article publié dans la revue *Monthly Weather Review*, vol. 145, no 9, p. 3529-3544.

Sample Stratification in Verification of Ensemble Forecasts of Continuous Scalar Variables : Potential Benefits and Pitfalls

J. Bellier[1], I. Zin[1], and G. Bontron[2]

[1] Université Grenoble Alpes, Grenoble INP, CNRS, IGE, Grenoble, France
[2] Compagnie Nationale du Rhône, Lyon, France

(Manuscript received 29 Dec. 2016, accepted 23 May 2017, published online 4 Aug. 2017)

DOI: 10.1175/MWR-D-16-0487.1

Résumé

Dans le champ disciplinaire de la vérification, la stratification désigne le concept visant à découper l'échantillon de couples prévision-observation en plusieurs sous-échantillons qui soient les plus homogènes possibles, avec l'objectif de mieux comprendre le comportement des prévisions dans certaines conditions spécifiques. Dans cet article, nous présentons un cadre méthodologique de stratification s'appliquant aux prévisions ensemblistes de variables scalaires et continues. Nous distinguons trois approches possibles, qui diffèrent selon l'origine du critère de stratification : la prévision, l'observation, ou bien une information externe au couple prévision-observation. Sous ce formalisme sont revisitées les définitions de deux outils de vérification usuels que sont le CRPS et l'histogramme de rang. Nous proposons alors, pour chacun d'eux, des représentations graphiques qui synthétisent l'information acquise grâce à la stratification. Concernant l'histogramme de rang, il est démontré qu'une stratification selon un critère provenant de l'observation est à proscrire, car elle induit des perturbations sur l'histogramme qui faussent son interprétation. A l'inverse, une stratification selon un critère provenant de la prévision est judicieuse, car elle permet de se rapprocher de l'évaluation de la fiabilité dans sa définition la plus stricte. Des précédents travaux ayant montré l'apparition possible d'artefacts statistiques, nous proposons un test graphique qui permet de les détecter. Enfin, ces concepts sont mis en application sur un example réel impliquant deux archives de prévisions de précipitation d'origine différente : des prévision d'ensemble ainsi que des prévisions par analogie.

Abstract

In the verification field, stratification is the process of dividing the sample of forecast-observation pairs into quasihomogeneous subsets, in order to learn more on how forecasts behave under specific conditions. A general framework for stratification is presented for the case of ensemble forecasts of continuous scalar variables. Distinction is made between forecast-based, observation-based and external-based stratification, depending on the criterion on which the sample is stratified. The formalism is applied to two widely used

verification metrics: the CRPS and the rank histogram. For both, new graphical representations that synthesize the added information are proposed. Based on the definition of calibration, it is shown that the rank histogram should be used within a forecast-based stratification, while an observation-based stratification leads to significantly non-flat histograms for calibrated forecasts. Nevertheless, as previous studies have warned, statistical artefacts created by a forecast-based stratification may still occur, thus a graphical test to detect them is suggested. To illustrate potential insights about forecast behavior that can be gained from stratification, a numerical example with two different datasets of mean areal precipitation forecasts is presented.

4.3.1 Introduction

Probabilistic forecasts are nowadays widely used in the meteorological community, since they provide a useful estimate of the predictive uncertainty. In an operational context, these forecasts are generally in the form of ensembles representing possible scenarios. Despite progress in the verification field since their emergence, the complexity of their behavior still represents a great challenge for verification practitioners (Casati *et al.*, 2008). In few words, verification is the action of assessing the quality of the forecasts by comparing them to their corresponding observations (Jolliffe et Stephenson, 2003). Since a complete picture of the forecast quality cannot be obtained from a single measure, different verification measures have been proposed, which evaluate different attributes (i.e. aspects) of the forecast quality (Murphy, 1973). All measures, though, have in common to require a large number of forecast-observation pairs in the verification sample to be statistically robust. To help increasing the sample size, various forecasts may be pooled together (in the same sample), e.g. for different locations, for various ranges of predictands or from different model versions. However, computing a verification measure over an inhomogeneous sample faces the risk of having different forecast behaviors that average out. Stratification, as the process of partitioning the verification sample into different subsets, aims at conditioning the verification measure to specific conditions, so as to minimize this risk and lead to more insightful verification case studies.

It is difficult to trace back the origin of the term *stratification*, since the concept has probably emerged soon after first meteorological forecasts were verified. Indeed, authors very often present performance measures for different locations or seasons, which is an implicit way of stratifying the complete verification sample. Such an approach aims at making measures of forecast skill independent from the climatological frequency of events that have to be verified, which varies both in space and time (Hamill et Juras, 2006). Moreover, modellers are accustomed to conditioning verification case studies to specific meteorological conditions when improving numerical weather prediction models. However, it appears that the term *stratification* has been mostly used in the literature with the purpose of assessing the significance of different subsets in term of their contribution to the overall verification measure. Murphy (1995), in the first devoted paper, extended his general framework for forecast verification (Murphy et Winkler, 1987) to stratification along different meteorological conditions, in the case of probability forecasts of dichotomous events.

This article concentrates on the field of ensemble forecasts of continuous scalar variables. Hereafter, we consider an *ensemble* as a discrete approximation of a full forecast distribution. This definition encompasses forecasts issued by meteorological ensemble forecasting techniques (Buizza *et al.*, 1999) but also probabilistic forecasts issued by other forecasting techniques such as statistical adaptations, like the analogue method (Obled *et al.*, 2002; Hamill et Whitaker, 2006), or single-value forecast dressings (Schaake *et al.*, 2007). Two widely used verification measures for ensemble forecasts are the Continuous Ranked Probability Score (CRPS) (Matheson et Winkler, 1976; Hersbach, 2000; Gneiting et Raftery, 2007) and the rank histogram (Anderson, 1996; Hamill et Colucci, 1997; Talagrand *et al.*, 1997). As a measure of forecast calibration, the rank histogram has been subject to stratification in past studies, as advocated by Hamill (2001). He suggests stratification along a statistic of the ensemble in order to detect conditional biases that would be hidden when computing the rank histogram over the whole sample. As a stratification criterion, authors have used the mean and the standard deviation of the ensemble (Hamill et Colucci, 1997), or well-correlated quantities (Hamill et Colucci, 1998; Bröcker, 2008). A substantial contribution to the underlying theory has been made by Siegert *et al.* (2012), who expressed the risk of statistical artefacts that may affect the interpretation of rank histograms when stratifying along a statistic of a finite-size ensemble. Alternatively, Mullen et Buizza (2002) and, indirectly, Bellier *et al.* (2016), have stratified rank histograms along the observation. Although Siegert *et al.* (2012) have mentioned the risk of similar artefacts, theoretical aspects related to the latter approach has, to the knowledge of the authors, not been studied yet.

The rank histogram does not evaluate though how accurate is a forecast. Verification reports of ensemble forecasts very often include the average CRPS as a summary measure of the overall forecast accuracy. Previous contributions have mostly focused on its decomposition into different parts corresponding to specific attributes of the forecast (Hersbach, 2000; Bontron, 2004; Candille et Talagrand, 2005). Only few studies (Gneiting et Ranjan, 2011; Lerch *et al.*, 2017) have tackled the CRPS under the stratification approach, by studying the properties of the score when it is averaged over a restricted subset of the verification sample.

In this article, we propose a general stratification framework for ensemble forecasts of continuous scalar variables, and detail different ways of stratifying: along a function of the observation, of the forecast or of an external criterion. Within this framework, a new formulation of the average CRPS is derived. Concerning the rank histogram, the work done by Bröcker (2008) and Siegert *et al.* (2012) is extended to the problematic case in which stratification is made along a function of the observation, where it is shown that calibrated forecasts do not lead to flat histograms over each stratum. For both the CRPS and the rank histogram, new graphical representations are proposed that synthesize the information coming from stratification into low-complexity charts. To evaluate their potential benefits, two real datasets of probabilistic precipitation forecasts, having similar skills but different behaviors, are verified: ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) and analogue-derived forecasts, statistically adapted from the ECMWF control forecast.

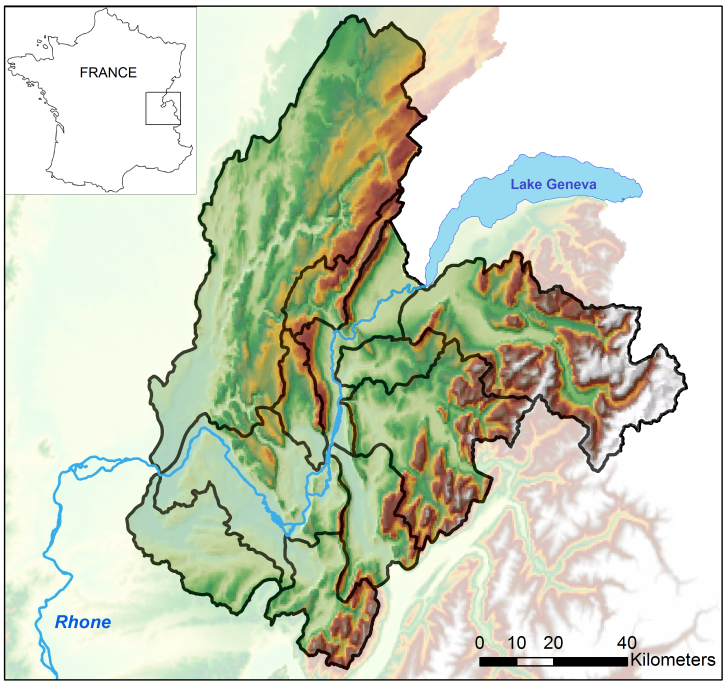

Figure 4.6 – Location of the 10 considered catchments in France (including the inner catchment that encloses the Rhone River section).

The article is organized as follows. Observation and forecast datasets are presented in section 4.3.2. Section 4.3.3 describes the stratification formalism, while sections 4.3.4 and 4.3.5 detail the application on the CRPS and the rank histogram, respectively. Section 4.3.6 presents results from a numerical example. Key points related to stratification are discussed in section 4.3.7. Section 4.3.8 concludes.

4.3.2 Observation and forecast data

Some aspects of the stratification framework benefit from illustrations based on real data. For ease of understanding, these data are presented first. The weather variable of interest, i.e. the predictand, is Mean Areal Precipitation (MAP) accumulated at 6-hour time step over hydrological catchments, in the perspective of hydrological forecasting. Please note that the formalism in sections 4.3.3, 4.3.4 and 4.3.5 applies to any other continuous scalar weather variables.

Fig. 4.6 shows the 10 considered catchments located in France just downstream from the Lake Geneva, with areas ranging from 290 to 3760 km^2. MAP observations used for verification were processed by Météo-France by kriging hourly and daily rain gauge data from the Météo-France network. The considered period is from 1 January 2010 to 31 December 2014.

Two forecast datasets are examined, both coming from the 0000 UTC cycle. The first dataset (labelled as ECMWF-Ens) contains 50-member ensemble forecasts produced by the ECMWF Ensemble Prediction System (EPS) (Buizza *et al.*, 1999) and downloaded from the TIGGE database (Park *et al.*, 2008). Only the perturbed members are consid-

ered here. Thiessen-based averaging (Tabios et Salas, 1985) has been used to transform grid-based forecasts to MAP forecasts. The second dataset (labelled as ECMWF-Ana) contains 40-member forecasts produced by statistical downscaling of the ECMWF control forecast using an analogue method developed successively by Obled *et al.* (2002), Bontron (2004), Ben Daoud *et al.* (2011), Marty *et al.* (2012) and Ben Daoud *et al.* (2016). In a nutshell, the synoptic forecast situation, characterized by means of large scale predictors (geopotential height, temperature and humidity), is taken from the control member. Then, most *analogue* synoptic situations are selected among an archive of reanalyses. Finally, MAP observations having been recorded on these dates are selected and constitute the forecast in the form of an ensemble. Although generally related to EPS, the terms *ensemble* and *member* are here used to describe analogue-based forecasts as well. More information about this dataset can be found in Bellier *et al.* (2016).

4.3.3 General stratification framework

4.3.3.1 Overview

Stratification, within the verification context, is defined as the action of dividing the sample of historical forecast-observation pairs into different subsets, according to a *stratification criterion*. The different subsets are called *strata* (singular: stratum). Hereafter, it is considered as implicit that this is done with the intent of computing performance measures over each of these subsets. The underlying objective is to partition the complete sample into strata that show different behaviors, in order to better understand strengths and weaknesses of the studied forecasting system. In this article, verification concerns forecasts of weather variables that are *continuous*, such as precipitation or temperature, and *scalar* (as opposed to multivariate), which presupposes a given location and time. It does not exclude, however, the possibility of pooling in the same sample forecasts from different locations and/or times if it is operationally justified. In previous dedicated studies (Bröcker, 2008; Siegert *et al.*, 2012), the formalism has been developed for stratification along a function of the forecast only. What follows is an extension where the criterion may depend on any characteristics of the forecast-observation pair.

Consider a continuous and scalar predictand which has a forecast Probability Density Function (PDF) f or Cumulative Distribution Function (CDF) F. Generally, an operational probabilistic forecast is available in the form of an ensemble of values instead of a full distribution. Thus, consider a finite-size ensemble drawn independently from f and sorted in ascending order:

$$\mathbf{x} = (x_1, \ldots, x_M)$$

where M is the size of the ensemble, assumed to be constant. An ensemble constructed this way is called a Monte-Carlo ensemble (Siegert *et al.*, 2012). We suppose that operational forecasts behave as such. Finally, let y denote the verifying observation.

Consider that y is a random variable, $\mathbf{x}$ a random vector and f a random distribution. Let denote by y_n, $\mathbf{x}_n$ and f_n their different *outcomes*. Therefore, one can constitute a verification sample:

$$T = \{(\mathbf{x}_n, y_n),\ n = (1, \ldots, N)\}$$

which contains N forecast-observation pairs with forecasts in the form of ensembles, supposed to be drawn from the latent (and unavailable) distributions f_n.

Consider further that for each forecast-observation pair within T can be defined a stratification criterion $\theta \in \mathbb{K}$, where $\mathbb{K}$ is the domain of θ. For example, $\mathbb{K}$ may correspond to $\mathbb{N}$ if θ is categorical, or to $\mathbb{R}$, $\mathbb{R}^k$ or $\mathbb{R}^{k\times l}$ if θ is a scalar, a vector of size k or a field of size $k \times l$, respectively. The different outcomes of θ are denoted by θ_n. Following Siegert *et al.* (2012)'s definition, the stratification function is the function

$$\Theta : \mathbb{K} \to (1, \ldots, S)$$

which maps the criterion θ to one of the S discrete indices corresponding to the different strata. After stratification, the s^{th} stratum contains all forecast-observation pairs satisfying $\Theta(\theta) = s$. The S strata are mutually exclusive but collectively exhaustive (i.e. every pair $(\mathbf{x}_n, y_n)$ belongs to one and only one stratum), meaning that $\sum_{s=1}^{S} N_s = N$ where N_s is the number of elements of the s^{th} stratum. In this section, the stratification function Θ is described as being either *observation*, *forecast* or *external*-based, depending on the origin of the data the criterion θ is taken as a function of. Possible reasons justifying each of the three approaches are suggested. For the first two, distinction is made between *statistic* and *meteorology*-oriented strategies: on the one hand, θ is a direct function of $(\mathbf{x}, y)$, while on the other hand θ represents meteorological covariates that are not strictly contained in $(\mathbf{x}, y)$. Consequently, the meteorology-oriented strategy theoretically permits same elements $(\mathbf{x}, y)$ that would occur on two different days (which is unlikely) to belong to two different strata. We are unaware of any other attempts to classify the different stratification approaches. The following one is a suggestion, which has been found appropriate to support the conclusions we provide about stratified CRPS and rank histograms.

4.3.3.2 Observation-based stratification

A forecaster may wonder: *how did the forecasts behave when specific events have occurred?* Such question raises the need of an observation-based stratification. Within the statistic-oriented strategy, the criterion is taken as the verifying observation, that is $\theta = y$. Then, the stratification function Θ is defined as $\Theta(\theta) = s$ if and only if $y \in \Gamma_s$, where Γ_s defines, for each stratum, an interval of $\mathbb{R}$ ($\mathbb{R}_+$ for precipitation). For example, suppose that one wants to better understand the forecast behavior when heavy rain events (say $\geq$ 30 mm day^{-1}) have occurred. The sample T will then be divided using $\Gamma_1 = [0, 30[$ and $\Gamma_2 = [30, +\infty[$, and the second stratum containing the elements $(\mathbf{x}, y)$ satisfying $y \in \Gamma_2$ will be carefully examined. Such a stratification is applied on the CRPS in the numerical example in section 4.3.6.

Within the meteorology oriented strategy, θ is taken as one or several meteorological covariate(s) of the observation. A typical example is the weather regime, defined as a large-scale spatial atmospheric pattern that has been identified among a finite set of possible ones (Michelangeli *et al.*, 1995; Vrac et Yiou, 2010). The information about the observed weather regime is not strictly contained in y, but close links may exist between both. For precipitation for instance, an observed anticyclonic weather regime is strongly associated

with outcomes where $y = 0$. In the easiest case, observed weather regimes have, for the N elements of T, already been identified and classified into one of the possible weather regimes. Then $\theta \in \mathbb{N}$ refers to a given weather regime and the stratification function can easily be defined. Otherwise, θ contains information about the weather regime, in the form of, for example, a vector of different meteorological variables or a spatial field of a given variable (e.g. geopotential height). In this case, stratification requires the definition of a distance metric, like the Euclidean distance if θ is a vector, or the S1 score (Teweles et Wobus, 1954) if θ is a field. Based on the computation of this distance over all couples of θ_n, one can classify each observed synoptic situation into a discrete number of classes using a clustering method.

4.3.3.3 Forecast-based stratification

If the question is now: *when given forecasts are issued, how do they behave?*, a forecast-based stratification is justified. Considering first the statistic-oriented strategy, one can consider the criterion $\theta = \kappa$ as a numerical statistic of the forecast PDF f from which the ensemble is supposed to be drawn. However, since the latent forecast PDF f is unknown, an estimation $\widehat{\kappa}$ from the finite-size ensemble $\mathbf{x}$ has to be used instead. Siegert *et al.* (2012) propose for $\widehat{\kappa}$ the mean, the median, the spread (as the standard deviation), the interquartile range or the total range between the smaller and the larger ensemble member. The stratification function can then be defined as follow: $\Theta(\theta) = s$ if and only if $\widehat{\kappa} \in \Gamma_s$ where Γ_s define intervals for each stratum.

Taking the criterion θ as a single statistic of $\mathbf{x}$ does not ensure, though, that forecasts are similar from the statistical perspective. For example, two ensemble forecasts can have similar spread but very different mean. Clustering techniques therefore constitute an alternative approach to gather into same stratum ensemble forecasts that have similar distributions according to a given distance metric. To measure how similar two CDF F_1 and F_2 are, we propose the use of the *integrated quadratic distance* (Thorarinsdottir *et al.*, 2013), defined as

$$d_{\mathrm{IQ}}(F_1, F_2) = \int_{-\infty}^{+\infty} \left[F_1(u) - F_2(u)\right]^2 \, \mathrm{d}u, \tag{4.9}$$

which satisfies all axioms of a metric. Its formulation can be seen as an extension of the CRPS as defined later in Eq. (4.10), where the distribution F_2 is no longer a Heaviside function. Discretization is necessary for computation since forecasts F_1 and F_2 are in the form of ensembles $\mathbf{x}_1$ and $\mathbf{x}_2$. Practically, d_{IQ} is computed over all couples of $\mathbf{x}_n$, and a clustering method is used to divide the sample T into different strata. Such a stratification is applied on the rank histogram in the numerical example in section 4.3.6.

Within the meteorology-oriented strategy, forecasts generated by similar meteorological situations are gathered into the same stratum. This is however more complicated than in the observation-based case where single meteorological situations are associated with each element of T. Here, if forecasts come from a meteorological EPS, each member is associated with a given meteorological situation. In other words, considering for example

that θ corresponds to the weather regime, there are possibly M different weather regimes associated with a given forecast $\mathbf{x}_n$. Different methods should be studied, e.g. considering as the stratification criterion, for each forecast at a given lead time, the most likely weather regime over the M members or the control's one. Nevertheless, none of these methods seems entirely satisfactory and this aspect is reserved for future studies.

4.3.3.4 External-based stratification

A last question one may wonder is: *in a given forecasting environment, how do the forecasts behave?* Here, the forecasting environment refers to information that is *external* to the forecasts and observations themselves (either the predictand values or meteorological covariates). As a consequence, this approach can be combined with any of the two previously presented. For example, θ can be taken as the location, if spatial disparities are suspected among forecasts for different locations, or as the month of the year, in order to detect seasonal biases in the forecasting model. Note that stratifying along the season is different from along the weather regime (either observed or predicted), the former considering only the forecast date while the latter is a flow-dependent approach. Furthermore, samples of operational forecasts can cover a period that includes one or several model upgrades. To assess their impact on verification measures, one can take θ as the model version. For any of these criterion, $\theta \in \mathbb{N}$ can be defined so as to easily construct the stratification function Θ.

As a concluding remark of this section, the classification of stratification approaches we propose can also be viewed under the perspective of the time at which the criterion θ is available. In an observation-based stratification, θ is unknown at the forecast time. In a forecast-based stratification, θ is known at the forecast time since it directly depends on the forecast (either the ensemble itself or some meteorological covariates). Finally, in an external-based stratification the criterion θ is known before the forecast time, as it does not depend on the forecast but only on the forecasting environment discussed above. This perspective is essential for an appropriate usage of stratification in the verification process, as will be discussed in section 4.3.7.

4.3.4 Application on the CRPS

The CRPS is a verification measure that evaluates the overall accuracy of a probabilistic forecast by estimating the quadratic distance between the CDF of the forecast and the observation (Matheson et Winkler, 1976; Hersbach, 2000; Gneiting et Raftery, 2007). It is defined, for an element n of the verification sample, as

$$\mathrm{CRPS}_n = \int_{-\infty}^{+\infty} \left[F_n(u) - H(u - y_n)\right]^2 \, \mathrm{d}u \tag{4.10}$$

where

$$H(u) = \begin{cases} 0 & \text{for } u < 0 \\ 1 & \text{otherwise} \end{cases}$$

is the Heaviside function. It is negatively oriented, meaning that smaller values are better. Since the forecast is in the form of an ensemble $\mathbf{x}_n$, the formulation (4.10) has to be discretized for computation, as proposed by Hersbach (2000). However, the way CRPS_n are computed does not influence the stratification process. In practice, the CRPS is averaged over a sufficiently large sample T, yielding what we refer to as the *overall* CRPS:

$$\overline{\mathrm{CRPS}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{CRPS}_n. \tag{4.11}$$

This quantity can be subject to stratification from two different perspectives.

4.3.4.1 Interpretation of the restricted CRPS

First, we review the interpretation of stratified CRPS from the *restriction* perspective. As suggested by Lerch *et al.* (2017), we define the *restricted* CRPS, denoted by $\overline{\mathrm{CRPS}}_s$, as the CRPS averaged over elements of the s^{th} stratum:

$$\overline{\mathrm{CRPS}}_s = \frac{1}{N_s} \sum_{n:\ \Theta(\theta_n)=s} \mathrm{CRPS}_n \tag{4.12}$$

where N_s is the number of elements in this stratum. It is indeed an appealing approach, especially in the evaluation of the forecast accuracy for extreme events, to compute and interpret $\overline{\mathrm{CRPS}}_s$ over small subsets of elements. However, Gneiting et Ranjan (2011) and Lerch *et al.* (2017) have shown that unwanted effects may appear. In particular, they have studied the *propriety* property of the restricted CRPS. As a desirable property, a verification score is *proper* if it rewards forecasters who issue forecasts that correspond to their true belief, and if it does not suggest any explicit hedging strategy (Gneiting et Raftery, 2007). Gneiting et Ranjan (2011) have shown that the restricted CRPS is improper under an observation-based stratification with $\theta = y$. Being aware of the stratification function that restricts the observation to a specific stratum, forecasters are encouraged to issue forecasts that emphasize this stratum and therefore differ from their true belief. Numerical examples that evidence this effect can be found in Lerch *et al.* (2017). However, they note that the restricted CRPS under a forecast-based stratification remains proper. We refer to the above references for more details.

4.3.4.2 Decomposition of the overall CRPS using stratification

In this paper, we rather suggest an approach from a *decomposition* perspective. Using the stratification function Θ with mutually exclusive but collectively exhaustive strata,

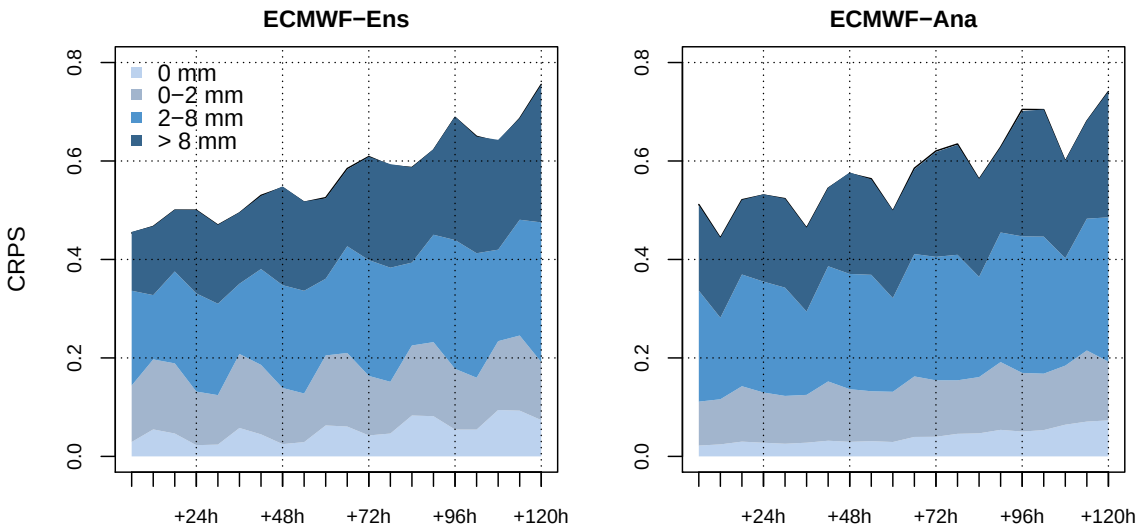

Figure 4.7 – Accumulated stratified CRPS for ECMWF-Ens (left) and ECMWF-Ana (right) forecasts, as a function of lead time. The stratification is done along the observation y.

one can derive

$$\begin{aligned}
\overline{\mathrm{CRPS}} &= \frac{1}{N}\sum_{s=1}^{S}\sum_{n:\ \Theta(\theta_n)=s}\mathrm{CRPS}_n \\
&= \frac{1}{N}\sum_{s=1}^{S} N_s\left(\frac{1}{N_s}\sum_{n:\ \Theta(\theta_n)=s}\mathrm{CRPS}_n\right) \\
&= \sum_{s=1}^{S}\frac{N_s}{N}\,\overline{\mathrm{CRPS}}_s. \qquad (4.13)
\end{aligned}$$

Eq. (4.13) is the underlying equation of the graphical representation we propose in Fig. 4.7, where the relative contributions $(N_s/N)\times\overline{\mathrm{CRPS}}_s$ of each stratum are coloured differently and summed up to $\overline{\mathrm{CRPS}}$. We define this representation as the *accumulated stratified* CRPS, which enables one to easily assess the significance of each stratum in terms of contribution to the overall CRPS.

Note that each stratification approach defined in section 4.3.3 can potentially be applied on this decomposition, since the equality (4.13) remains true irrespective of Θ. This statement can be extended to any other verification score that is defined for each individual forecast-observation pairs, since Eq. (4.12) and (4.13) are independent of the definition (4.10) of CRPS_n.

4.3.5 Application on the rank histogram

The Rank Histogram (RH) is a diagnostic tool aiming at assessing the *calibration* of the forecasts (Anderson, 1996; Hamill et Colucci, 1997; Talagrand *et al.*, 1997). Note that, in the literature, other words for calibration are sometimes used: *reliability* or *statistical consistency*. Unlike the CRPS, the RH is constructed on a collective basis, meaning over

a sufficiently large set of forecast-observation pairs.

4.3.5.1 Assessing calibration using the rank histogram

As defined by Jolliffe et Stephenson (2003), a forecasting system is *calibrated* if, and only if, the conditional probability distribution $p(y|f = f_n)$ of the observation, given any chosen forecast distribution f_n, is itself equal to f_n:

$$p(y|f = f_n) = f_n \tag{4.14}$$

for all possible f_n. Recall that n is the index of a possible outcome. Since forecasts are in the form of ensembles, Eq. (4.14) can be formulated as ensemble members $(x_1, \ldots, x_M)$ and the observation y being drawn from the same distribution f_n for each outcome n. In what follows, we express mathematically how the RH verifies this property.

Consider a given forecast distribution f_n from which ensembles $\mathbf{x}$ are drawn. Let's extend ensembles $\mathbf{x}$ with fictional bounding ensemble members x_0 and x_{M+1} such that $F_n(x_0) = 0$ and $F_n(x_{M+1}) = 1$. The key point here is that potentially different $\mathbf{x}_n$ can be drawn from the same f_n. Then, consider the observation y as a realization from $p(y|f = f_n)$, which is denoted hereafter by g_n. Let q_i for $i = \{1, \ldots, M+1\}$ be random variables corresponding to the conditional probability that the observation y falls between members x_{i-1} and x_i, given a specific $\mathbf{x}_n$ drawn from f_n:

$$q_i = \Pr\left(y \in]x_{i-1}, x_i] \mid \mathbf{x} = \mathbf{x}_n\right) \tag{4.15}$$

$$= \int_{x_{i-1}}^{x_i} g_n(u)\ \mathrm{d}u. \tag{4.16}$$

If forecasts are calibrated, g_n is equal to f_n. The observation y is therefore just one more draw from f_n which, over a large number of different $\mathbf{x}_n$ drawn from the same f_n, is equally likely to fall within each interval $]x_{i-1}, x_i]$ for $i = \{1, \ldots, M+1\}$. As a consequence, since there are $M + 1$ intervals, calibration implies

$$\mathbb{E}[q_i] = \frac{1}{M+1} \qquad \forall\ i = (1, \ldots, M+1) \tag{4.17}$$

where $\mathbb{E}[\cdot]$ denotes the expectation over different $\mathbf{x}_n$ drawn from the same f_n (Hamill, 2001). A graphical interpretation of q_i in case of calibration is proposed in the upper panel of Fig. 4.8, with a given 40-member ensemble $\mathbf{x}_n$ outcome. In this example, values q_4 and q_{29} (represented as shaded area) are different due to the fact that this given $\mathbf{x}_n$ is a Monte-Carlo ensemble and not a vector of quantiles. Nevertheless, both $\mathbb{E}[q_4]$ and $\mathbb{E}[q_{29}]$ are equal to 1/41.

The RH aims at verifying if the equality (4.17) holds by plotting

$$\mathrm{Freq}_i = \frac{\#(y_n \in]x_{n,i-1},\ x_{n,i}]\ ;\ n = (1, \ldots, N))}{N} \tag{4.18}$$

as a function of $i = \{1, \ldots, M+1\}$, where $\#(\cdot)$ denotes the number of cases the conditions

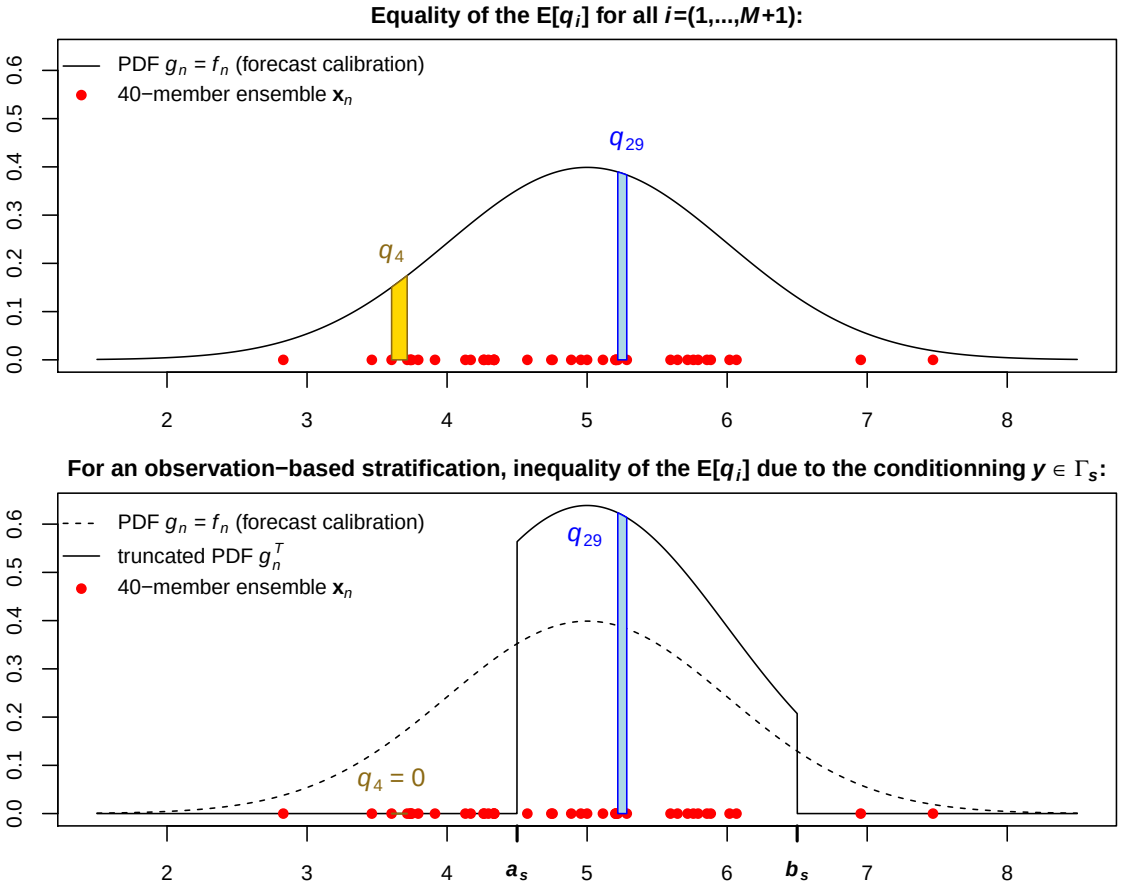

Figure 4.8 – **Upper panel**: Graphical interpretation of the conditional probabilities q_i that the observation falls between members x_{i-1} and x_i given the ensemble $\mathbf{x}_n$ drawn from f_n, in case of forecast calibration. **Lower panel**: The same, but in case of an observation-based stratification ($\theta = y$).

$\cdot$ is true. Intervals $]x_{n,i-1},\ x_{n,i}]$ are called *bins*, which are said to be populated when y_n falls in. When two or more members within $\mathbf{x}_n$ take the same value as y_n, corresponding bins are populated randomly. If forecasts are calibrated, i.e. the equality (4.17) holds, the so-obtained histogram should be flat, apart from fluctuations due the finite size N of the sample. We express such instance by

$$\text{Freq}_i \simeq \frac{1}{M+1} \qquad \forall\ i = (1, \ldots, M+1), \tag{4.19}$$

which tends to the strict equality as N approaches infinity. A significant non-flatness indicates a miscalibration. The appealing feature of the RH is that one can graphically learn, from the shape of the histogram, where the deficiencies in the forecasting system lie: $\cup$-shape and $\cap$-shape indicate under and over-dispersion, respectively, while $/$-shape and $\backslash$-shape indicate negative and positive bias, respectively (Hamill, 2001).

According to the definition (4.14) of calibration from Jolliffe et Stephenson (2003), the proper way to assess calibration would be to construct the RH on a sample containing only forecasts drawn from the same distribution f_n, and to repeat the process for all possible f_n. Obviously, this requirement is impossible to fulfil in an operational context. Instead,

RH are generally constructed over samples of forecasts in which distributions differ from each other. Such overall RH verifies if the equality (4.19) holds *on average*, while the strict definition of calibration would imply the equality holding *for each different distribution.* Murphy et Epstein (1967), Yates (1982) and Bontron (2004) have referred to the former definition as *in-the-large* calibration and to the latter as *in-the-small* calibration. It is important to highlight that in-the-small calibration implies in-the-large calibration, while the contrary is not true. Thus, flatness of the overall RH is a necessary but not sufficient condition for calibration, as first mentioned by Hamill (2001). If the assessment of in-the-small calibration is in practice infeasible since datasets hardly contain ensemble forecasts from the same distribution, an insight can be obtained with a forecast-based stratification which gathers into same stratum forecasts that are similar.

The present framework for forecast calibration differs from the theoretical framework proposed by Gneiting *et al.* (2007), although connections between both exist. Gneiting *et al.* (2007) defined several modes of calibration, namely *probabilistic*, *exceedance* and *marginal* calibration, with *strong* calibration when all three hold. Their probabilistic calibration is equivalent to the above-defined in-the-large calibration, and assessed by checking the flatness of the RH constructed over a non-stratified sequence of forecast-observation pairs. Furthermore, they introduced the concept of *completeness*: complete calibration (regarding one or several modes) is verified if the calibration mode(s) holds for any possible subsequences of forecast-observation pairs. This concept, loosely defined though, shares with the in-the-small calibration definition the idea that the assessment of calibration over a set of forecasts in which distributions differ from each other faces the risk of having different behaviors that average out. The other modes defined by Gneiting *et al.* (2007), namely exceedance and marginal calibration, are not considered in our present framework, but it seems reasonable to assume that in-the-small calibration should imply both exceedance and marginal calibration, at least we cannot think of a counter-example.

4.3.5.2 The concept of stratified rank histograms

After having defined theoretical aspects of the RH, consider a stratification along the criterion θ. Definition (4.15) can then be rewritten as

$$q_i = \Pr\left(y \in \left]x_{i-1}, x_i\right] \mid \mathbf{x} = \mathbf{x}_n,\ \Theta(\theta) = s\right). \tag{4.20}$$

A RH constructed over the strata s is then represented by

$$\mathrm{Freq}_{i,s} = \frac{\#\left(y_n \in \left]x_{n,i-1},\ x_{n,i}\right] \ ;\ n :\ \Theta(\theta_n) = s\right)}{N} \tag{4.21}$$

as a function of $i = \{1, \ldots, M+1\}$. For mutually exclusive but collectively exhaustive strata, one can write

$$\mathrm{Freq}_i = \sum_{s=1}^{S} \mathrm{Freq}_{i,s} \tag{4.22}$$

for $i = \{1, \ldots, M+1\}$, which is the underlying equation of the graphical representation we propose in Fig. 4.9. The *overall* RH is represented as the sum of S *stratified* RH coloured

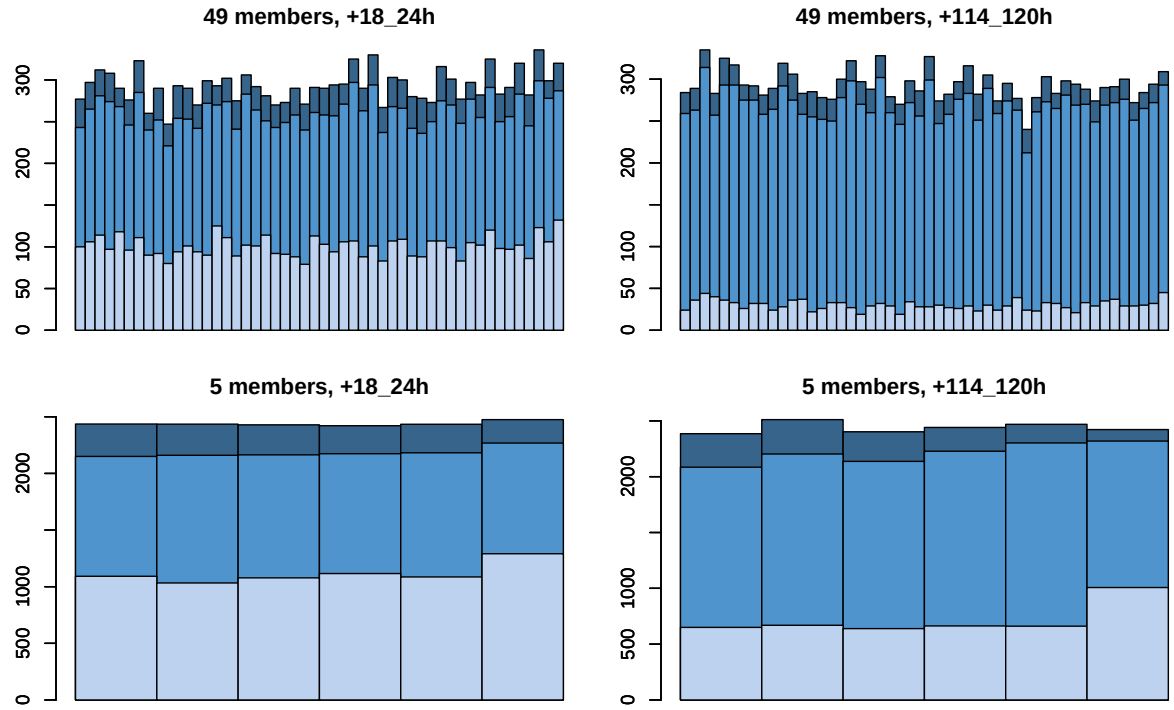

Figure 4.9 – Accumulated stratified rank histograms considering different lead times and ensemble sizes, for ECMWF-Ens forecasts under the perfect-model assumption (one random member is used as the verifying observation) in order to detect deviations from flatness due to a *forecast*-based stratification. The stratification is done along the ensemble mean, with $\Gamma_1 = [0]$ (light blue) ; $\Gamma_2 =]0, 3]$ (medium blue) and $\Gamma_3 =]3, +\infty]$ mm 6h^{-1} (dark blue).

differently. We define this representation as the *accumulated stratified RH*, which enables one to easily assess the contribution of each stratum to the overall RH.

Stratification of the RH is relevant provided that flatness is expected over each stratum for calibrated forecasts. If this condition is not satisfied, one could hardly infer from the shape of the histograms what comes from the stratification process and what is due to miscalibration of the forecasts. Flatness of stratified RH requires the equality (4.17) to hold for all $i = \{1, \ldots, M+1\}$. This is obviously true for any external-based stratification, since θ is independent of $(\mathbf{x}, y)$. However, several pitfalls encompass the forecast and observation-based stratification cases.

4.3.5.3 Forecast-based stratified rank histograms

Bröcker (2008) and Siegert *et al.* (2012) have shown that, under calibration, flatness is expected with a stratification criterion $\theta = \kappa$ or a covariate of the forecast. However, Siegert *et al.* (2012) have demonstrated that deviations may occur for some cases where $\theta = \widehat{\kappa}$, a statistic of f computed from the ensemble $\mathbf{x}$. This short paragraph aims at summarizing their findings, but more details and a mathematical demonstration can be found in their paper. The fact that $\widehat{\kappa}$ is only an estimation of the true statistic κ causes random sampling errors $(\kappa - \widehat{\kappa})$ that are turned into a systematic error when stratifying. To understand, let's imagine a dataset of calibrated ensemble forecasts all drawn from

the same distribution f_n (like e.g. climatological forecasts), and divided into two same-size strata according to their estimated mean $\widehat{\kappa}$. Because of the random sampling errors $(\kappa - \widehat{\kappa})$, the first stratum will contain more forecasts where $\widehat{\kappa} < \kappa$, and second stratum more forecasts where $\widehat{\kappa} > \kappa$. Since the verifying observation y is also drawn from f_n, the probability $\Pr(y > \widehat{\kappa})$ is higher (lower) than $\Pr(y < \widehat{\kappa})$ in the first (second) stratum. This will lead to a $\nearrow$-shape ($\searrow$-shape) histogram, although forecasts are perfectly calibrated. When other statistics $\widehat{\kappa}$ than the mean are considered, different forms of deviations from flatness in stratified RH are expected, but we refer to Siegert *et al.* (2012) for more details.

Two factors play a role in this undesirable artefact. The first one is the frequency at which forecasts overlap the bounds delineating the different strata. This is linked to their relative sharpness (i.e. their sharpness compared to their own climatology), as sharp forecasts are less likely to overlap the bounds of the strata. Therefore, the sharper ensembles are, the weaker is the artefact. In the above example, the artefact is maximized by the fact that all ensembles $\mathbf{x}_n$ are drawn from the same distribution f_n. Hence, f_n represents the forecast climatology, so the relative sharpness is null and consequently all $\mathbf{x}_n$ overlaps the bounds delineating the different strata. The second factor is the ensemble size M. The more members ensembles have, the smaller random sampling errors $(\kappa - \widehat{\kappa})$ are, and as a consequence the weaker the artefact is. To completely get rid of such artefact, Siegert *et al.* (2012) discuss the possibility of, for each forecast n, splitting randomly each ensemble $\mathbf{x}_n$ into two sub-ensembles. The first sub-ensemble would be used to compute $\widehat{\kappa}_n$. The second sub-ensemble would be considered for the RH and subject to stratification. The disadvantage of this method is that verification is made on forecasts containing less information than the raw forecasts, since only half of the members are considered.

As an alternative to trying to eliminate the artefact, this article proposes a graphical test to evaluate its impact so as to take it into account when interpreting histogram shapes. The objective is to construct stratified RH for the forecasts to verified at hand, although under a *perfect-model* assumption. The procedure is as follows: for each element of T, one member is randomly withdrawn from the ensemble forecast and considered as the new verifying observation. The so-obtained forecasts are perfectly calibrated regarding these *pseudo-observations* since both forecast members and the pseudo-observation are drawn each time from the same distribution. They also correctly respect the two characteristics of the original forecasts regarding the undesirable artefact. Indeed, the ensemble size is not much changed (only one member less), and the relative forecast sharpness should remain equivalent. Note though that this assumption is reasonable for large ensembles like ECMWF-Ens but might not hold for much smaller ensembles. The second step consists in applying the forecast-based stratification on this dataset. If stratified RH, for each stratum, do not show any significant deviation from flatness, then no undesirable artefact is likely to occur with the same stratification applied on the original forecast-observations pairs. If discrepancies appear for some strata, they have to be taken into account when interpreting stratified RH back to the original data. If necessary, one can even consider abandoning this stratification. Another interesting point of such graphical test is that considering the same sample size enables to graphically assess, *a priori*, how random fluctuations will affect the interpretation of RH shapes when considering the original

data.

Results of an experiment of such graphical test is given in the Fig. 4.9. Original forecasts are the 50-member ECMWF-Ens MAP forecasts that have been described in section 4.3.2. Stratification is done along the ensemble mean with $S = 3$ strata defined by $\Gamma_1 = [0]$; $\Gamma_2 =]0, 3]$ and $\Gamma_3 =]3, +\infty]$ (unit: mm 6h^{-1}). Vertically, the effect of the ensemble size is tested, with 49 and 5 (randomly selected) members. Horizontally, the effect of the relative sharpness is tested, by considering different lead times of the forecasts: 18_24 and 114_120 hours. Ensemble forecasts become indeed less sharp as lead time increases, because of the limited predictability of the atmosphere. As expected, all overall RH are flat, as a consequence of the perfect-model assumption. The upper-left histogram does not exhibit any visible deviation from flatness for any of the strata, meaning that the stratification applied on this forecast dataset is relevant regarding the artefact described above. However, one can detect slight slope compensations between the strata when reducing the ensemble size from 49 to 5, which is amplified for the 114_120 hour lead time. As a consequence, care must be taken in the interpretation of stratified RH when going back to the original data with such characteristics.

4.3.5.4 Observation-based stratified rank histograms

Although forecast-based stratified RH are justified for the assessment of *in-the-small* calibration, observation-based stratified RH look attractive to answer the question: *how did the forecasts behave when specific events have occurred?* In the following, we extend the work made by Bröcker (2008) and Siegert *et al.* (2012) to demonstrate, however, that calibrated forecasts do not lead to flat RH under an observation-based stratification.

Considering $\theta = y$, the stratification function is defined as $\Theta(\theta) = s$ if and only if $y \in \Gamma_s$, where Γ_s are intervals defining the S strata. Let's define $\Gamma_s =]a_s, b_s]$ for $s = \{1, \ldots, S\}$. Then, using the definition of truncated distributions, Eq. (4.20) becomes

$$q_i = \Pr\left(y \in]x_{i-1}, x_i] \mid \mathbf{x} = \mathbf{x}_n,\ y \in]a_s, b_s]\right) \tag{4.23}$$

$$= \int_{x_{i-1}}^{x_i} g_n^T(u)\ \mathrm{d}u \tag{4.24}$$

where $g_n^T(u)$ is the truncated PDF defined by

$$g_n^T(u) = \begin{cases} \frac{g_n(u)}{\int_{a_s}^{b_s} g_n(v)\ \mathrm{d}v} & \text{for } a_s < u \leq b_s \\ 0 & \text{otherwise.} \end{cases} \tag{4.25}$$

If forecasts are calibrated, g_n is equal to f_n. However, $\mathbb{E}[q_i]$ with q_i defined as such strongly depends on i, as evidenced by the graphical interpretation in the lower panel of Fig. 4.8. In this specific case, $q_4 = 0$ (because $f_n^T(x_4) = 0$ since $x_4 < a_s$) while q_{29} is higher than in the upper panel (because $f_n^T(x_{29}) > f_n(x_{29})$). One can then easily figure out that the $\mathbb{E}[q_i]$ are not equal for all $i = \{1, \ldots, M+1\}$, due to the fact that, in this case, f_n overlaps the bounds a_s and b_s delineating the strata s.

As a consequence, flat RH over the different strata are not expected with calibrated

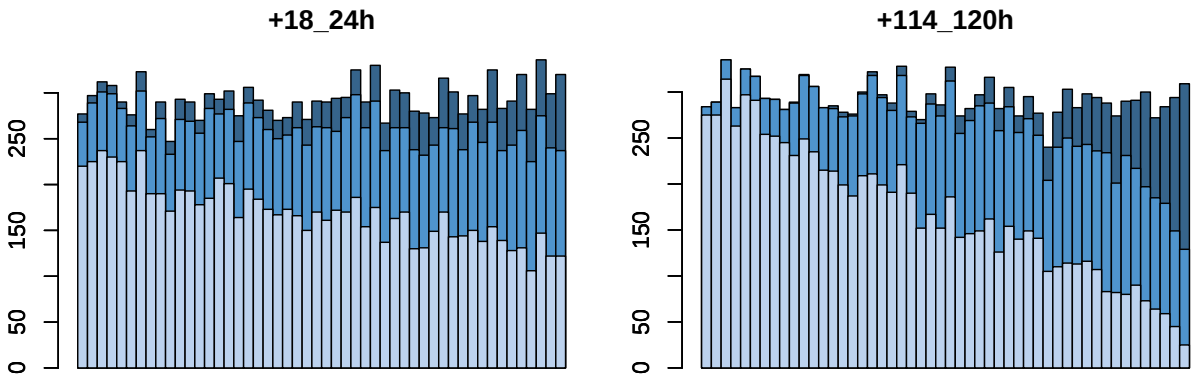

Figure 4.10 – Accumulated stratified rank histograms considering different lead times, for ECMWF-Ens forecasts under the same perfect-model assumption as in Fig. 4.9 in order to detect deviations from flatness due to an *observation*-based stratification ($\theta = y$). Strata are defined with $\Gamma_1 = [0]$ (light blue) ; $\Gamma_2 =]0, 3]$ (medium blue) and $\Gamma_3 =]3, +\infty]$ mm 6h^{-1} (dark blue).

forecasts when stratifying along the observation. The sharper ensembles are compared to the climatology of the observation, the weaker will be the deviations from flatness. This artefact vanishes as $F_n(a_s)$ goes to zero and as $F_n(b_s)$ goes to one (i.e. the ensemble forecast has no chance to overlap a_s and b_s). Otherwise, stratified RH will be impacted. Fig. 4.10 shows RH stratified along the observation for the same perfect-model forecast datasets as in Fig. 4.9, with 49 members (remind that one random member is used as the verifying observation). To illustrate the effect of the relative sharpness of the forecasts, 18_24 and 114_120 hour lead times are considered. We observe that deviations from flatness are well pronounced for both. As will be discussed in section 4.3.7, we therefore strongly advise against observation-based stratification when constructing RH.

4.3.6 Numerical example

In this section, we illustrate the potential benefits of stratification through the verification of the two forecast datasets presented in section 4.3.2. An *observation*-based stratification with $\theta = y$ is first conducted for the analysis of the CRPS, with the objective of characterizing the contribution of different ranges of the predictand to the overall CRPS. Fig. 4.7 shows the accumulated stratified CRPS of ECMWF-Ens and ECMWF-Ana forecasts, as a function of lead time. Forecast-observation pairs for the 10 catchments are pooled together. We observe that the two forecast datasets show a similar overall CRPS, both in terms of amplitude and diurnal cycle. Further insights can however be obtained from stratification. One can graphically assess from Fig. 4.7 the relative contribution of each precipitation range to the overall CRPS. For example, zero observed precipitation occurs very frequently ($N_s/N \simeq 65\%$, not shown in the paper), but their contribution (lower stratum) to the overall CRPS is small. On the contrary, occurrences with more than 8 mm 6h^{-1} are rare ($N_s/N \simeq 3\%$, not shown), but they contribute to about one third (higher stratum) of the overall CRPS. Moreover, one can obtain information about the origin of the diurnal cycle. Observed precipitation time series exhibit a small diur-

nal cycle, which has been found to be exacerbated by ECMWF-Ens members, especially when zero or low precipitation have occurred (not shown). As a consequence, most of the CRPS diurnal cycle in the left chart comes from the two lower strata. Alternatively, ECMWF-Ana forecasts do not amplify the diurnal cycle, since they bypass the thermodynamic process related to the precipitation generation in the atmospheric model. The CRPS cycle in the right chart is therefore mostly explained by the observation cycle, which is stronger for the two higher precipitation strata.

Then, a *forecast*-based stratification is carried out for the assessment of calibration using RH. The 42_48 hour lead time is here considered. As a preliminary step for both datasets, forecast-observation pairs for the 10 catchments were pooled together since they were found to behave similarly (according to stratified RH within an external-based stratification along the catchments, not shown). This enables to enlarge the size of the sample T. In this example, $N = 18200$ pairs. Then, T is stratified using a clustering technique, with $S = 6$ strata, towards the objective of gathering into same stratum forecasts that are similar with regards to their entire distribution, not only their mean, spread or any other statistic. For clustering, a hierarchical cluster analysis has been conducted, using the *integrated quadratic distance* (cf. Eq. (4.9)) as the metric for the dissimilarity between two distributions and the *Ward* distance (Murtagh et Legendre, 2014) as the distance between two clusters. As all other data handling, this process has been done within the `R` environment (R Core Team, 2014). The `hclust` function from the `stats` package has been used. Bottom panels of Fig. 4.11 and 4.12 show the distributions populating each strata of ECMWF-Ens and ECMWF-Ana forecasts, respectively. To detect if this stratification is subject to a statistical artefact affecting the interpretation of the forecast-based stratified RH, the graphical test proposed in section 4.3.5.3 has been applied (not shown) and no significant deviations from flatness due to stratification are expected.

In the upper panel of Fig. 4.11 and 4.12 is represented the accumulated stratified RH. For the sake of ease of interpreting stratified RH shapes, individual RH for each stratum are also plotted in the middle panel. Several insights about forecast behavior can be obtained from this stratification. First, recall that when several members take the same value as the verifying observation, corresponding bins are populated randomly. It occurs very frequently when dealing with variables such as precipitation that have a point mass in zero. Stratum (a) represents forecasts with all members equal to zero. Non-zero observations can therefore populate the $(M+1)^{\text{th}}$ bin only. This stratum collects 29.1% of the forecasts for ECMWF-Ens, but only 4.9% for ECMWF-Ana, which tend to very often have at least few members different from zero. As a significant strength of this analogue model though, ECMWF-Ana forecasts are calibrated over (a) while ECMWF-Ens forecasts exhibit a $(M+1)$ bar higher than it should. Moreover, studying the stratum (a) enables to appreciate the potentially significant fraction, as for ECMWF-Ens forecasts, of the RH that comes from random process due to zero precipitation. From strata (b) to (f) in Fig. 4.11, one can conclude that ECMWF-Ens forecasts are generally under-dispersive for the 42_48 hour lead time in our example. Note that the overall RH already showed a $\cup$-shape, but only the forecast-based stratification ensures that it really indicates under-dispersion and not a combination of $\diagdown$-shape and $\diagup$-shape. In addition, we observe

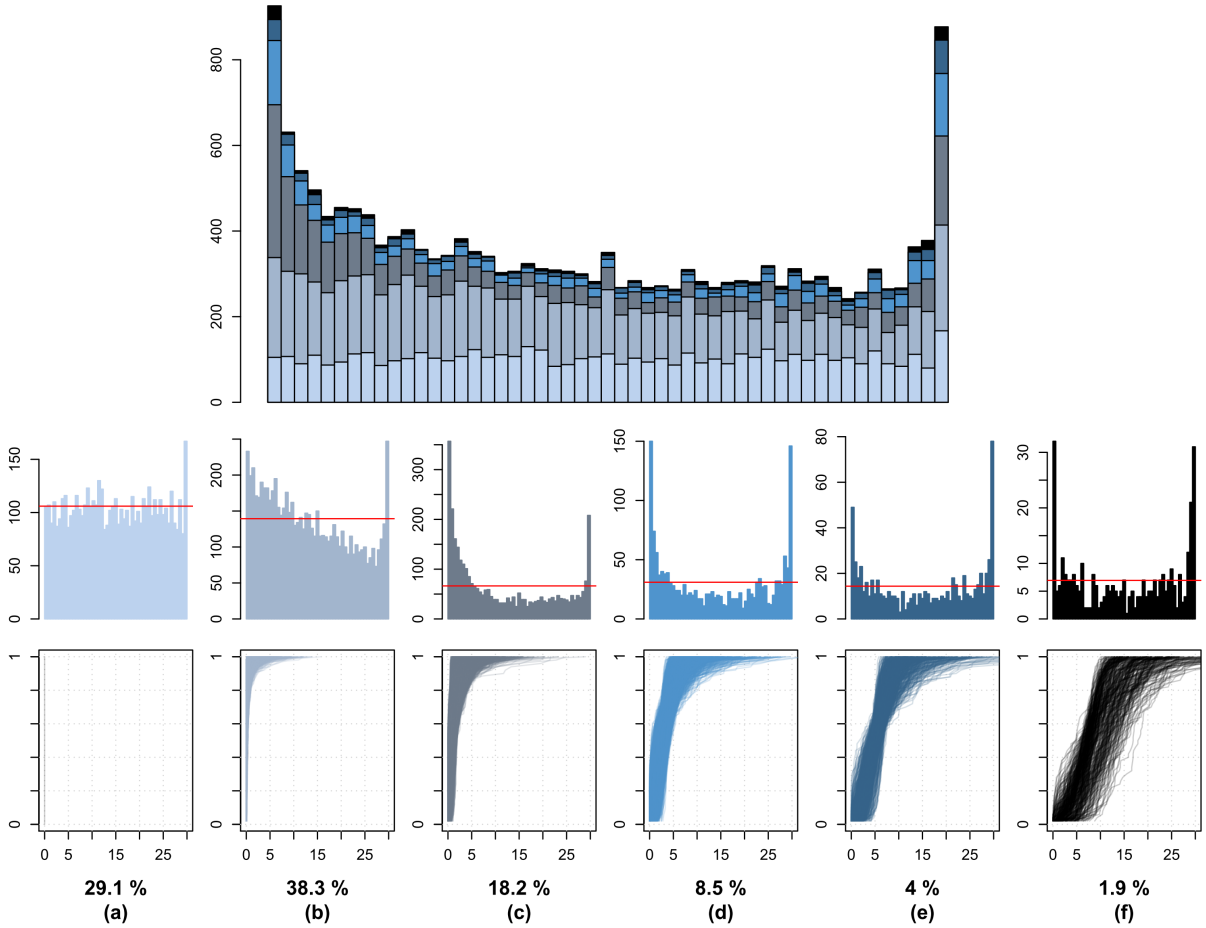

Figure 4.11 – **Upper pannel**: Accumulated stratified rank histogram for ECMWF-Ens forecasts, for the 42_48 hour lead time, under a forecast-based stratification using clustering (cf. section 4.3.3.3). **Middle panel**: Individual stratified rank histograms. **Lower panel**: Plots of all forecast distributions populating each stratum. The x-axis is in mm $6h^{-1}$. The different strata are denoted by letters (a) to (f), with their percentage of the total sample size.

that strata (b) and (c) also exhibit a positive bias ($\diagdown$-shape) which corresponds to an over-estimation of low precipitation, while (d), (e) and (f) do not. Concerning ECMWF-Ana forecasts in Fig. 4.12, significant differences between the overall RH shape and the shapes of the different strata show up, which can only be observed after stratification. On the one hand, strata (b) and (c) display a positive bias ($\diagdown$-shape) in the right side of the RH, which corresponds to the part of the distributions with non-zero members. In other words, the observation is more often as it should equal to zero when forecasts from (b) and (c) are issued. This illustrates a deficiency of this analogue model, which has difficulties to generate forecasts with a probability of precipitation equal to zero (few non-zero members tend to always be kept). On the other hand, strata (e) and (f) show a tendency to over-dispersion, coupled with a negative bias ($\diagup$-shape). This illustrates the fact that, for high precipitation, this analogue method tends to still conserve few low-precipitation members.

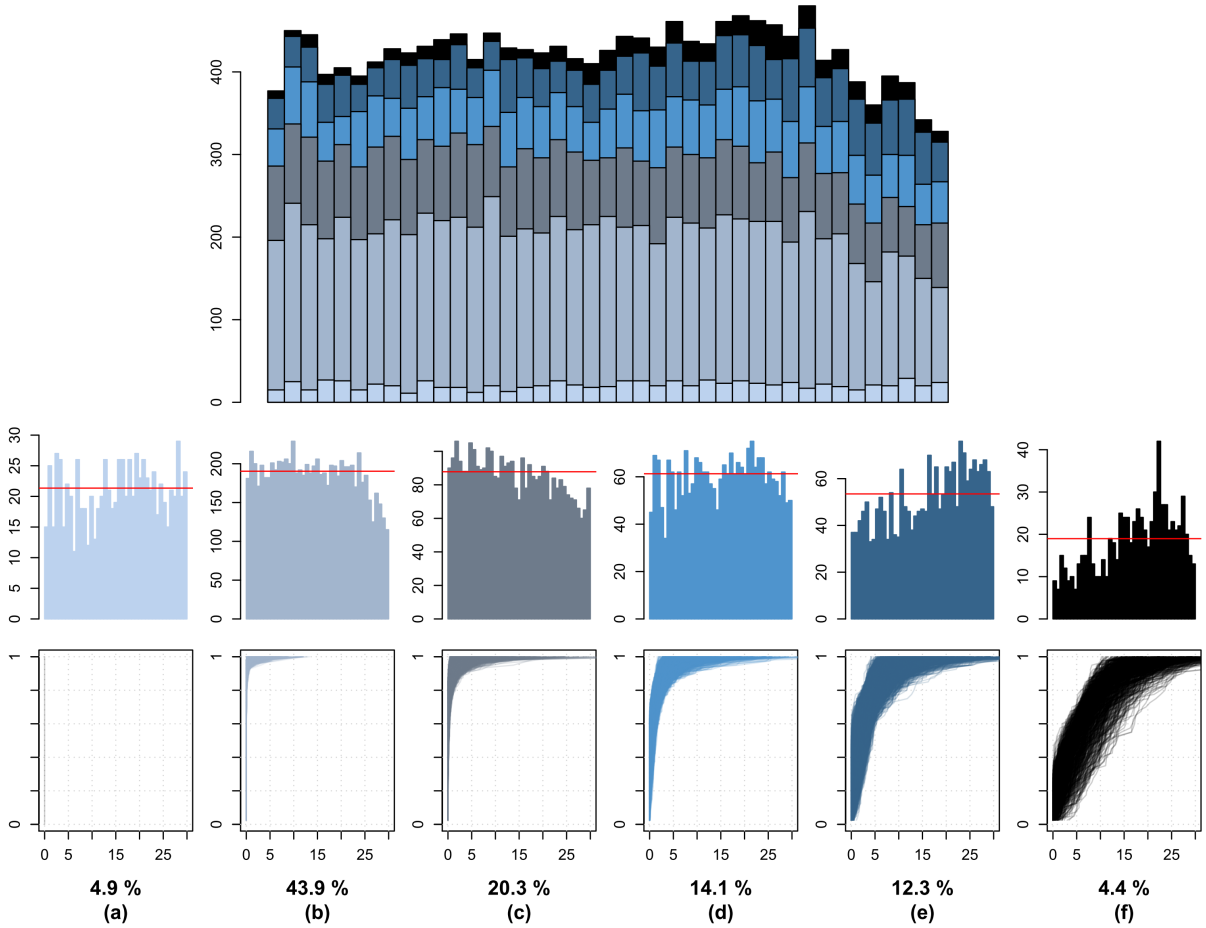

Figure 4.12 – Same as Figure 4.11, for the ECMWF-Ana forecasts.

4.3.7 Discussion

4.3.7.1 Should we consider observation-based stratification?

As discussed in the section 4.3.4 dedicated to the CRPS, previous studies have shown that observation-based stratification is problematic if forecasters need to compute restricted CRPS over specific strata and to interpret them individually from one stratum to another. This would consist, for instance, in using the restricted CRPS for ranking different forecasting systems, or as the objective function of an optimization process within the forecast post-processing step. For such needs, forecast-based stratification is recommended instead, as the restricted CRPS remains proper. Note also that weighted versions of the CRPS (Gneiting et Ranjan, 2011) that emphasize, inside the integral of Eq. (4.10), a specific region of the predictand's range are alternative possibilities.

The second approach discussed in section 4.3.4 that decomposes the overall CRPS into the contributions coming from the different strata is nonetheless free of theoretical barriers to any stratification strategies. We remind that it is a way to better understand the sensitivity of the overall CRPS to specific subsets of the verification sample, but not to evaluate the forecast accuracy over each subset individually. Possible reasons for advocating an observation-based rather than a forecast-based approach are, for instance, the desire to learn more about the CRPS behavior on climatological forecasts (widely

used as reference forecasts in skill scores), or to ensure same sample sizes in strata when comparing different forecast datasets.

The case of the RH is intrinsically different as it is, unlike the CRPS, constructed and interpreted on a collective basis. It has been shown in section 4.3.5 that the stratification process can impact such interpretation, as evidenced by artefacts yielding non-flat RH constructed with forecasts under the perfect-model assumption. Both observation and forecast-based stratification approaches are concerned, although to different extents. In the former case, the artefact comes from a misuse of the RH as a way to assess calibration. Calibration (or miscalibration) is indeed a forecast property that one wants to be aware of before observations occur. This is the underlying principle of post-processing, where forecast biases can be identified and conditioned to forecasts or external characteristics so as to be corrected at forecast time. The assessment of calibration has therefore no reason to be conditioned on the future. This would face the risk of drawing erroneous conclusions about forecast behavior. In a past study within an hydrological forecasting context, Bellier *et al.* (2016) have constructed RH over a sample containing high-flow events selected according to observed peak flow values. Strong ⟋-shape was found, but misinterpreted as being a symptom of under-estimation bias of the forecasting system while it was mainly caused by the observation-based stratification, which conserved only the high-flow (observed) events. We therefore strongly advise against any observation-based stratification, neither statistic nor meteorology-oriented, when assessing forecast calibration using the RH.

Instead, a forecast-based stratification is perfectly justified as it tends to approach the "true" assessment of forecast calibration by gathering into same stratum forecasts that are similar. The potential artefact in forecast-based stratified RH is purely statistical and results from the fact that ensemble forecasts have a finite number of members. Its strength thus strongly reduces for large ensembles. Moreover, the graphical test we propose, based on the perfect-model assumption, enables one to assess *a priori* whether or not the stratification is reasonable. We therefore advocate, as long as care is taken, for forecast-based stratification when computing RH.

4.3.7.2 Connection between CRPS and rank histogram

Hersbach (2000) has proposed a decomposition of the CRPS into a reliability part and a resolution/uncertainty part. The reliability part is closely connected to the rank histogram. For each bin i, the squared difference between the average frequency that the observation falls below the middle of the bin and the corresponding forecast probability i/M is quantified. The sum of these components yields the reliability part, which should for calibrated forecasts tend to zero as the size N of the sample approaches infinity. It is essential to note, however, that this decomposition does not apply to individual CRPS_n but to the average value $\overline{\mathrm{CRPS}}$, as defined respectively in Eq. (4.10) and (4.11). Hence, applying such a decomposition on a stratified dataset faces the risk of drawing erroneous conclusions about forecast calibration. Particularly, a similar reasoning as in section 4.3.5.4 easily shows that significantly positive values for the CRPS reliability part are expected with observation-stratified forecasts under the perfect-model assumption. As for

the RH, we recommend to avoid observation-based stratification when applying Hersbach (2000)'s CRPS decomposition. Concerning a forecast-based stratification, the pattern discussed in section 4.3.5.3 is likely to play a role in case of small forecast ensembles, but we reserve for future studies the quantification of this potential impact.

4.3.7.3 The issue of sample size

As mentioned earlier, the verification of ensemble forecasts requires a *sufficiently* large sample of forecast-observation pairs. Otherwise, the average CRPS will fluctuate for each pair being added to or withdrawn from the sample and RH's bins will not be populated enough to correctly interpret the shape. Quantitatively, what *sufficiently* means is not in the scope of this paper, yet it has been tackled by several authors. Not exhaustively, Candille *et al.* (2007) propose for the CRPS to account for sample size with bootstrap methods (Efron et Tibshirani, 1994). Goodness-of-fit tests for RH flatness exist (Elmore, 2005; Jolliffe et Primo, 2008), and Bröcker (2008) also suggests to plot each RH on a probability paper in order to give quantitative information as whether deviations from flatness are due to sample size or indicate a systematic bias. Nevertheless, it is important to highlight the fact that sample size is the major constraint to the stratification process. It is especially true for the assessment of calibration using forecast-based stratified RH, which would in theory require a large number of ensemble forecasts drawn from the same distribution. Therefore, a compromise has to be found between the need of strata large enough for a robust verification and the desire to learn more on how forecasts behave. For example, the forecast-based stratification in the numerical example was constrained to 6 strata due to sample size. With such a restricted number of strata in the case of precipitation, it hardly differs from a stratification along the ensemble mean. Nevertheless, the authors have found interesting to present this sophisticated method which can be potentially more worthwhile in other cases.

4.3.8 Conclusions

In this article, a general framework for stratification was described for the verification of ensemble forecasts of continuous scalar variables, in the pursuit of a better understanding of forecast behavior. Distinctions were made, on the one hand, between *observation*, *forecast* and *external*-based approaches, depending on where the stratification criterion comes from, and on the other hand between *statistic* and *meteorology*-oriented strategies, as whether the criterion is function of quantitative outcomes or of meteorological covariates related to physical processes.

The stratification formalism was applied to two widely used verification tools for continuous scalar variables, namely the CRPS and the rank histogram. For the CRPS, a technique that enables to easily assess the contribution of each stratum to the overall CRPS has been proposed, which can potentially be applied with any of the above-cited stratification approaches. However, simply restricting the computation of the average CRPS to a specific subset of the verification sample is problematic in case of an observation-based stratification, as the CRPS is rendered improper. For the rank histogram, past related

studies have been extended to the observation-based stratification case, where a mathematical and graphical demonstration showed that a flat histogram over each stratum is not expected with perfectly calibrated forecasts. Therefore, the authors strongly advise to avoid any observation-based stratification when assessing forecast calibration using the RH. Instead, a forecast-based stratification should be preferred, as it tends to approach the “true" assessment of forecast calibration. Past studies brought to light the risk of a statistical artefact affecting the interpretation of forecast-based stratified RH. We proposed a graphical test, based on the idea of perfect-model assumption, to detect if the user's targeted stratification can override such artefact.

The numerical example enabled to expose insights that can be potentially gained about forecast behavior. In particular, the assessment of calibration has been conducted under a forecast-based stratification using a clustering technique. For the 42_48 hour lead time studied, mean areal precipitation forecasts from the ECMWF ensemble prediction system over the 2010-2014 period were found generally under-dispersive, which is a well-known feature of the ECMWF ensemble prediction system for short lead times. Forecasts generated using an analogue method were found much more calibrated, although some bias compensations were observed.

This article is a contribution to the issue of sample stratification, which we believe should be considered more often in the verification process, as a way to limit the risk of missing key aspects of forecast behavior that would average out otherwise. For future work, we encourage other verification tools than the CRPS and the rank histogram to be studied under the stratification framework. Moreover, practical and quantitative guidances about the issue of sample size under stratification are required.

Acknowledgments

This work has been supported by a grant from Labex OSUG@2020 (Investissements d'avenir – ANR10 LABX56) and Compagnie Nationale du Rhône. Ensemble forecasts from TIGGE were supplied from ECMWF's TIGGE data portal. The authors thank Michael Scheuerer for helpful discussion about the CRPS. They also thank Stefan Siegert and an anonymous reviewer for their meticulous reviews that greatly improved the quality of the article.

4.4 Extension aux prévisions multivariées

Cette section se propose d'étendre nos outils de vérification, jusqu'à présent restreints au contexte univarié, à un cadre multivarié. Soit L la dimension de ce cadre multivarié, l'indice $l \in \{1, \ldots, L\}$ correspondant à différentes combinaisons de bassins, échéances ou variables. La grandeur à prévoir n'est plus un scalaire, mais un vecteur de dimension L.

Deux approches peuvent être entreprises : utiliser des outils qui s'appliquent directement aux prévisions multivariées, ou bien agréger les prévisions multivariées de manière à les ramener à un contexte univarié dans lequel des outils comme le CRPS ou l'histogramme de rang peuvent être utilisés. Ce sera l'objet des parties 4.4.2 et 4.4.3. Avant cela, nous discutons en 4.4.1 des attributs fiabilité et finesse dans un cadre multivarié.

Le contenu de cette section ne sera mis en application qu'à partir du chapitre 6. Le lecteur pourra ainsi décider d'y revenir à ce moment là.

4.4.1 Fiabilité et finesse dans un cadre multivarié

4.4.1.1 La fiabilité

Soit une prévision probabiliste multivariée représentée par une densité multivariée f, et soit $\mathbf{y}$ l'observation correspondante, considérée comme une réalisation de la loi de densité multivariée g. La fiabilité correspond alors à la similitude entre les densités multivariées f et g.

L'égalité $f = g$ impose tout d'abord l'égalité des densités marginales. Si l'on note respectivement $f^1, \ldots, f^L$ et $g^1, \ldots, g^L$ les L densités marginales des densités multivariées f et g, la fiabilité impose $f^1 = g^1, \ldots, f^L = g^L$. D'éventuels disparités entre ces densités marginales équivalent à des défauts de fiabilité dont les caractéristiques sont celles du cas univarié : biais négatif/positif si il y a un écart dans la moyenne des densités marginales, et sous/sur-dispersion si l'écart concerne la variance. Ces quatre défauts sont illustrés dans les deux premières colonnes de la Figure 4.13, dans un cadre multivarié de dimension $L = 2$.

L'égalité des densités marginales n'est cependant pas une condition suffisante de la fiabilité ; il faut également que f et g aient la même structure de dépendance. Dans un cadre de dimension $L = 2$, des disparités dans la structure de dépendance peuvent prendre la forme de sous-estimation ou de sur-estimation de la corrélation. Ces défauts sont illustrés dans la colonne de droite de la Figure 4.13, où les densités marginales sont identiques mais les ellipses d'iso-densité multivariée sont plus ou moins aplaties en fonction de la corrélation.

Comme pour le cas univarié, nous parlerons de fiabilité *à petite échelle* lorsque $f = g$. Cependant, l'évaluation de cette fiabilité est rendue impossible en pratique, par le fait que nous n'avons pas accès à la densité g mais seulement à une unique réalisation $\mathbf{y}$. L'évaluation doit donc nécessairement se baser sur une définition moins stricte de la fiabilité. Nous n'avons cependant pas connaissance d'auteurs ayant clairement formulé une telle définition dans un contexte multivarié. Plusieurs outils d'évaluation ont été

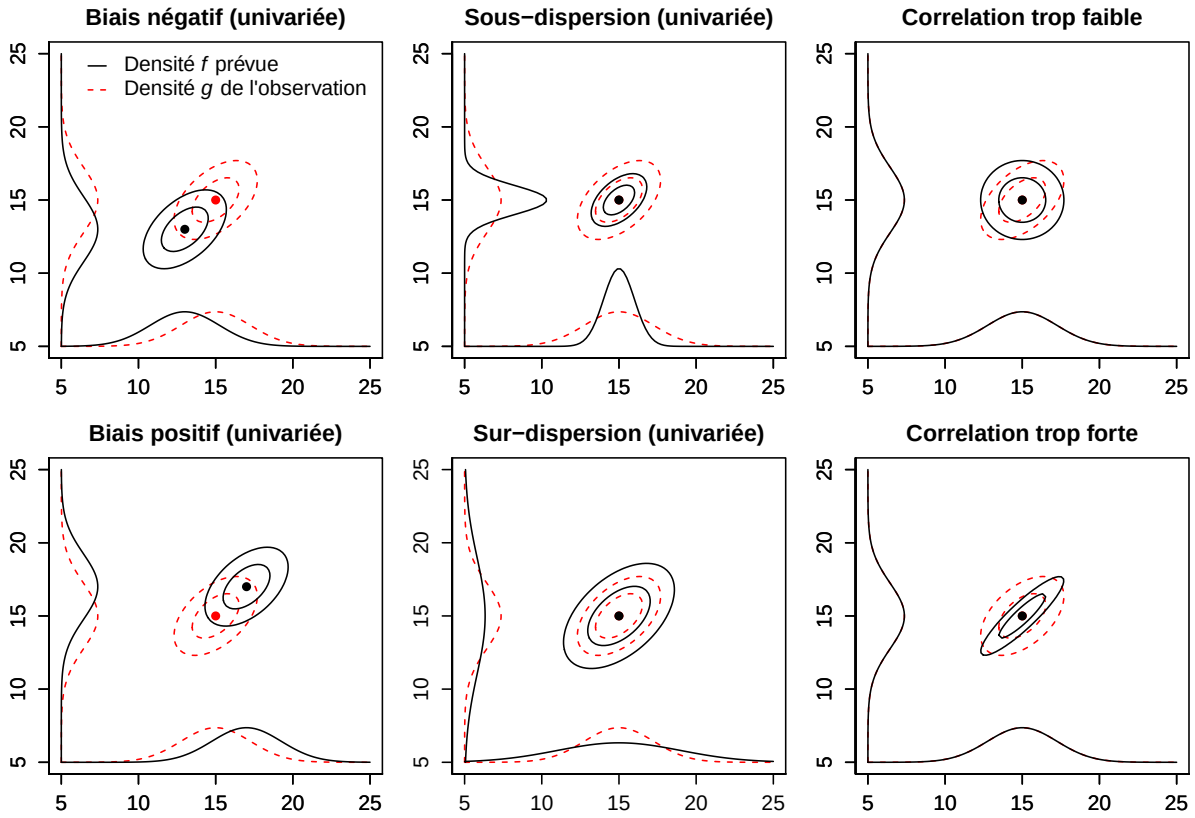

FIGURE 4.13 – Défauts caractéristiques de fiabilité concernant la moyenne (gauche), la variance (centre) ou la corrélation (droite) d'une prévision probabiliste multivariée en dimension 2, dans le cas idéal où l'on connait la densité de l'observation. Les axes horizontaux et verticaux représentent les dimensions 1 et 2 du vecteur aléatoire à prévoir. Les densités multivariées sont représentées au centre des cadres par deux ellipses d'iso-densité, tandis que les densités marginales sont tracées en bordure. Cette figure est l'extension de la Figure 4.1 au contexte multivarié.

proposés (Gneiting *et al.*, 2008; Thorarinsdottir *et al.*, 2016; Wilks, 2018b), qui s'appuient sur des prévisions multivariées étant uniquement sous forme ensembliste. L'objectif de ces outils est alors de caractériser le comportement moyen des observations $\mathbf{y}$ sur un grand nombre de réalisations, et de voir s'il est différent de celui des membres de l'ensemble. Nous proposons donc d'appeler cela la fiabilité *à grande échelle*.

4.4.1.2 La finesse

La finesse d'une prévision multivarié représente l'« étendue » du spectre des valeurs possibles. Elle est donc fortement liée à la finesse des densités marginales : plus les densités marginales seront fines, plus la densité multivarié sera fine. Comme dans le cas univarié, plus la densité est fine, moins l'incertitude est élevée, ce qui est désirable sous condition de fiabilité (selon le paradigme de Gneiting *et al.* (2007)).

4.4.2 Scores multivariés

Notons $\mathbf{X} = (X_1, \ldots, X_L)$ le vecteur aléatoire à prévoir (de dimension L), et supposons un échantillon de vérification contenant N couples prévision-observation. Pour

chaque réalisation $n \in \{1, \ldots, N\}$, l'observation est notée $\mathbf{y}_n = (y_n^1, \ldots, y_n^L)$, tandis que la prévision est représentée par l'ensemble $\mathbf{x}_n = (\mathbf{x}_{n,1}, \ldots, \mathbf{x}_{n,M})$, où chaque membre $\mathbf{x}_{n,m} = (x_{n,m}^1, \ldots, x_{n,m}^L)$ pour $m \in \{1, \ldots, M\}$ est un vecteur de dimension L.

4.4.2.1 Energy score (ES)

Le premier outil multivarié que nous considérons est l'*Energy Score* (ES), proposé par Gneiting *et al.* (2007). Il est défini, lorsque les prévisions sont sous forme ensembliste, par :

$$\mathrm{ES}_n(\mathbf{x}_n, \mathbf{y}_n) = \frac{1}{M} \sum_{m=1}^{M} \|\mathbf{x}_{n,m} - \mathbf{y}_n\| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \|\mathbf{x}_{n,m} - \mathbf{x}_{n,m'}\|, \tag{4.26}$$

où $\|\cdot\|$ est la norme euclidienne. L'ES peut être considéré, par son analogie avec la formulation (4.6), comme l'extension en dimension L du CRPS, vers lequel il converge lorsque $L = 1$. Il mesure ainsi une « distance », en dimension L, entre la prévision et l'observation. A noter que l'ES est un score strictement juste (cf. 4.2.1, où la définition donnée d'un score juste s'étend trivialement au cadre multivarié), et donc selon Bröcker (2009) il permet bien d'évaluer simultanément la fiabilité et la finesse des prévisions. De la même manière que le CRPS, l'ES a vocation à être moyenné sur l'ensemble de l'échantillon de vérification :

$$\overline{\mathrm{ES}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{ES}_n. \tag{4.27}$$

De plus, nous pouvons lui associer un score de compétence, l'*Energy skill score* (ESS), qui est défini par :

$$\mathrm{ESS} = 1 - \frac{\overline{\mathrm{ES}}}{\overline{\mathrm{ES}}_{\mathrm{ref}}}, \tag{4.28}$$

où $\mathrm{ES}_{\mathrm{ref}}$ correspond à l'ES moyen obtenu avec des prévisions de référence. L'interprétation de l'ESS est similaire à celle du CRPSS (cf. 4.2.1.3).

Pinson et Tastu (2013) et Scheuerer et Hamill (2015a) ont étudié le pouvoir de discrimination de $\overline{\mathrm{ES}}$, c'est-à-dire sa capacité à donner des scores différents pour des prévisions de qualité différente. En introduisant artificiellement des défauts dans la fiabilité des prévisions, ils ont montré que ce score était très sensible à la moyenne, un peu moins à la variance, et enfin assez peu sensible à la structure de dépendance.

4.4.2.2 Variogram score (VS)

La faible capacité de discrimination de l'ES à des défauts dans la structure de dépendance a conduit Scheuerer et Hamill (2015a) à proposer une alternative à l'ES, fondée sur le concept de variogramme. L'idée est alors de comparer le variogramme empirique de la prévision à celui de l'observation.

Le variogramme d'ordre p du vecteur aléatoire $\mathbf{X} = (X^1, \ldots, X^L)$ peut être défini par la quantité

$$\gamma_p(l, l') = \frac{1}{2}\mathbb{E}[|X^l - X^{l'}|^p], \tag{4.29}$$

pour $l, l' \in \{1, \ldots, L\}$. Considérant la prévision $\mathbf{x}_n$ comme un ensemble de M réalisations de $\mathbf{X}$, on peut calculer le variogramme empirique de la prévision en traduisant l'espérance en somme sur les membres :

$$\widehat{\gamma}_p(l, l')_{\text{prev}} = \frac{1}{2M} \sum_{m=1}^{M} |x_{n,m}^{l} - x_{n,m}^{l'}|^p, \tag{4.30}$$

pour $l, l' \in \{1, \ldots, L\}$. En ce qui concerne l'observation, le vecteur $\mathbf{y}_n$ est la seule réalisation de $\mathbf{X}$ disponible, et donc la meilleure estimation de son variogramme est :

$$\widehat{\gamma}_p(l, l')_{\text{obs}} = \frac{1}{2} |y_n^l - y_n^{l'}|^p, \tag{4.31}$$

pour $l, l' \in \{1, \ldots, L\}$. A partir de ces deux définitions, Scheuerer et Hamill (2015a) ont proposé le *Variogram score* (VS), défini par :

$$\text{VS}(\mathbf{x}_n, \mathbf{y}_n) = \sum_{l=1}^{L} \sum_{l'=1}^{L} w_{l,l'} \left(\left| y_n^l - y_n^{l'} \right|^p - \frac{1}{M} \sum_{m=1}^{M} \left| x_{n,m}^l - x_{n,m}^{l'} \right|^p \right)^2, \tag{4.32}$$

où $w_{l,l'}$ pour $l, l' \in \{1, \ldots, L\}$ sont des poids non négatifs.

Ces poids $w_{l,l'}$ permettent de mettre l'accent sur certaines composantes des variogrammes, c'est-à-dire certaines combinaisons de dimensions. Selon Scheuerer et Hamill (2015a), une plus forte pondération des combinaisons de dimensions les plus corrélées permet d'augmenter le rapport « signal sur bruit » du score vis-à-vis de sa capacité à discriminer des prévisions aux structures de dépendances différentes. Cette pondération peut se calculer lorsque ces combinaisons de dimensions sont associées à une forme de distance : spatiale si l et l' se réfèrent à des bassins différents, temporelle si ce sont des échéances différentes. Dans notre cas cependant, le contexte multivarié inclut à la fois des bassins et des échéances différentes, ainsi que éventuellement des variables différentes (précipitations et température, dans le cas des forçages météorologiques). Afin d'éviter de fixer les poids de manière arbitraire, nous décidons de ne pas faire de pondération, et ainsi fixons $w_{l,l'} = 1$ pour $l, l' \in \{1, \ldots, L\}$. Concernant l'ordre p du variogramme, nous observons que Scheuerer et Hamill (2015a) obtiennent une meilleure capacité de discrimination du VS avec $p = 0.5$, et par conséquent nous adoptons cette valeur.

Le score moyen $\overline{\text{VS}}$ et le score de compétence associé VSS peuvent être calculés de la même manière que pour l'ES (cf. équations (4.27) et (4.28)).

Le VS est construit de manière à mieux détecter que l'ES les défauts dans la structure de dépendance. Ses auteurs montrent également des résultats où il est performant vis-à-vis des défauts de dispersion. En revanche, il est par construction totalement insensible à un défaut systématique qui concernerait la moyenne. Du fait de cette limite, le VS est un score juste et non pas strictement juste. C'est pourquoi Scheuerer et Hamill (2015a) recommandent de considérer plusieurs scores lors de la vérification de prévisions multivariées, en complément d'une analyse préalable des performances univariées.

Par leur développement relativement récent, les score multivariés tels que l'ES et le

VS n'ont fait l'objet que de peu d'études spécifiques telles que celles citées précédemment. Ainsi, leur comportement est encore mal connu, et ce particulièrement lorsque la dimension L du contexte multivarié est élevée.

4.4.3 Agrégation des quantités multivariées

Une alternative à l'usage des scores multivariés consiste à s'appuyer sur des *fonctionnelles* $\varphi : \mathbb{R}^L \to \mathbb{R}$ qui, appliquées sur chacun des membres, permettent de transformer les ensembles multivariés en ensembles univariés. Il est alors possible d'évaluer ces ensembles en revenant à des scores univariés tels que le CRPS, dont le comportement est mieux connu.

Pour qu'une telle approche de vérification soit pertinente, il est nécessaire que ces fonctionnelles soient sensibles aux caractéristiques qui définissent la distribution multivariée du vecteur aléatoire $\mathbf{X}$ à prévoir, à savoir les distributions marginales mais également la structure de dépendance.

Nous proposons dans cette thèse d'utiliser deux fonctionnelles, la somme et le maximum, qui d'une part respectent la contrainte susmentionnée (nous le montrons plus bas), et d'autre part semblent judicieuses dans un contexte de prévision de débit, car elles sont respectivement associées au concept de volume et de pic de débit. Ces deux fonctionnelles sont définies par :

$$\varphi_{\text{sum}}(\mathbf{X}) = \sum_{l=1}^{L} X_l \tag{4.33}$$

$$\varphi_{\max}(\mathbf{X}) = X_{(L)}, \tag{4.34}$$

où $X_{(L)}$ est la statistique d'ordre L, autrement dit le maximum parmi les composants $X_1, \ldots, X_L$.

Nous montrons ici que les variables aléatoires $\varphi_{\text{sum}}(\mathbf{X})$ et $\varphi_{\max}(\mathbf{X})$ sont bien sensibles aux distributions marginales et à la structure de dépendance de $\mathbf{X}$, pour un cas où $\mathbf{X}$ suit une loi normale multivariée $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ avec $\boldsymbol{\mu} = (\mu, \ldots, \mu)$ et :

$$\Sigma = \begin{pmatrix} \sigma & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & \sigma \end{pmatrix}.$$

L'expérience consiste à estimer la sensibilité de la densité (univariée) des variables $\varphi_{\text{sum}}(\mathbf{X})$ et $\varphi_{\max}(\mathbf{X})$ à des variations dans les paramètres μ, σ et ρ de la loi dont $\mathbf{X}$ est issu. La Figure 4.14 illustre ces densités pour différentes combinaisons de paramètres, dans un cas où $L = 20$. Elles ont été estimées empiriquement à partir de 10000 tirages, et tracées à l'aide de la fonction `density` de `R`. On observe qu'elles sont bien sensibles à μ, σ et ρ, même si cette sensibilité prend parfois des formes différentes pour $\varphi_{\text{sum}}(\mathbf{X})$ et $\varphi_{\max}(\mathbf{X})$. Par exemple, lorsque ρ augmente (i.e., la corrélation entre les composants de $\mathbf{X}$ est plus forte), la densité de $\varphi_{\text{sum}}(\mathbf{X})$ s'aplatit, tandis que celle de $\varphi_{\max}(\mathbf{X})$ s'aplatit mais également se décale vers des valeurs plus faibles.

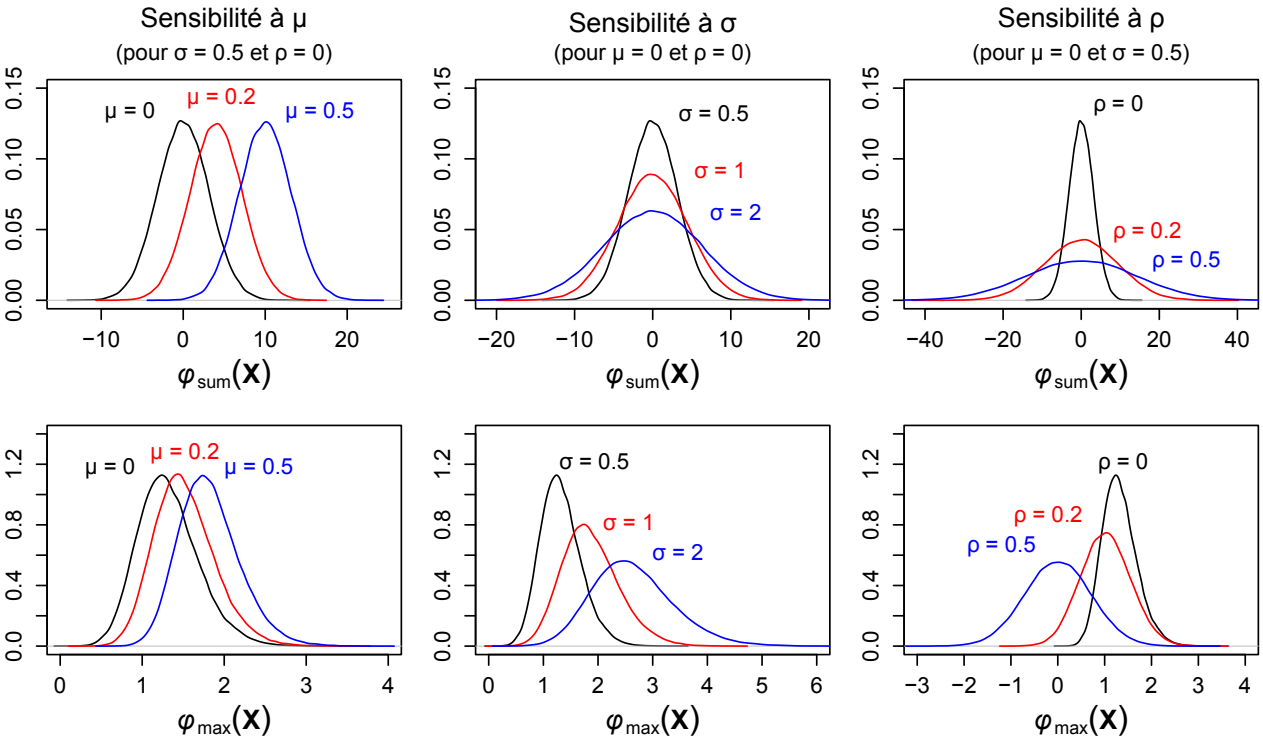

FIGURE 4.14 – Sensibilité de la densité des variables aléatoires $\varphi_{\text{sum}}(\mathbf{X})$ et $\varphi_{\max}(\mathbf{X})$ à des variations dans les paramètres μ, σ et ρ de la loi normale multivariée dont $\mathbf{X}$ est issu.

A partir de ces deux fonctionnelles, nous proposons ainsi les scores CRPS_{sum} et $\text{CRPS}_{\max}$, qui sont définis par :

$$\text{CRPS}_{\text{sum},n}(\mathbf{x}_n, \mathbf{y}_n) = \text{CRPS}_n\left((\varphi_{\text{sum}}(\mathbf{x}_{n,1}), \ldots, \varphi_{\text{sum}}(\mathbf{x}_{n,M})), \varphi_{\text{sum}}(\mathbf{y}_n)\right) \quad (4.35)$$

$$\text{CRPS}_{\max,n}(\mathbf{x}_n, \mathbf{y}_n) = \text{CRPS}_n\left((\varphi_{\max}(\mathbf{x}_{n,1}), \ldots, \varphi_{\max}(\mathbf{x}_{n,M})), \varphi_{\max}(\mathbf{y}_n)\right). \quad (4.36)$$

Ces scores seront occasionnellement utilisés en alternative à l'ES et le VS pour l'évaluation de prévisions multivariées. Comme précédemment, nous considérons les scores moyens $\overline{\text{CRPS}_{\text{sum}}}$ et $\overline{\text{CRPS}_{\max}}$, ainsi que les scores de compétence associés $\text{CRPSS}_{\text{sum}}$ et $\text{CRPSS}_{\max}$.

Par ailleurs, nous utiliserons également parfois les fonctionnelles φ_{sum} et $\varphi_{\max}$ pour évaluer la fiabilité (à grande échelle) de prévisions multivariées à l'aide de l'histogramme de rang.

4.5 Synthèse

La vérification, dans le domaine de l'hydrométéorologie, est le processus de confronter les prévisions aux observations afin d'en évaluer la qualité. Au cours de ce chapitre, nous avons présenté les outils de vérification qui seront utilisés dans la suite de la thèse. Il nous semble important de retenir que :

– La qualité de prévisions probabilistes de variables continues (précipitations, température, débit) peut se traduire en deux attributs que sont la fiabilité et la finesse.

Le but de la prévision est de s'assurer de la fiabilité avant de chercher à maximiser la finesse.

– Dans le cas univarié, le CRPS sert à quantifier la qualité globale (fiabilité et finesse) des prévisions, et l'histogramme de rang à détecter les éventuels défauts de fiabilité.

– Dans le cas multivarié, deux stratégies sont possibles. La première repose sur l'utilisation des scores multivariés que sont l'Energy score (ES) et le Variogramme score (VS). La seconde s'appuie sur des fonctionnelles telles que la somme et le maximum pour replacer les prévisions multivariées dans un contexte univarié où le CRPS et l'histogramme de rang peuvent être utilisés.

Ce chapitre a également été l'occasion de développer des aspects méthodologiques concernant un point particulier de la vérification, la stratification. Celle-ci consiste à diviser l'échantillon de vérification en différentes strates de manière à mieux comprendre le comportement des prévisions, par exemple en dévoilant certains biais qui autrement se compensent. Les principaux résultats sont les suivants :

– Les représentations graphiques sous forme de « strates accumulées » permettent au CRPS et à l'histogramme de rang d'illustrer visuellement les contributions relatives de chacune des strates.

– L'histogramme de rang ne doit pas être stratifié selon un critère qui dépend de l'observation, car cela fait artificiellement croire à un défaut de fiabilité.

– A l'inverse, il est judicieux de stratifier l'histogramme de rang selon un critère qui dépend de la prévision, car cela nous rapproche de l'évaluation stricte de la fiabilité.

Deuxième partie

Fiabilité et cohérence du forçage météorologique

Introduction à la Partie II

Dans cette partie, l'objet d'étude est le forçage météorologique (i.e., les prévisions météorologiques), que nous cherchons à améliorer dans la perspective de la modélisation hydrologique. Ces forçages concernent deux variables, la précipitation et la température, ainsi que plusieurs bassins et échéances. Sous forme ensembliste, ils permettent de quantifier l'incertitude de la prévision météorologique. Nous proposons, pour les améliorer, une approche en deux temps.

Premièrement, les forçages sont corrigés statistiquement de manière à éliminer d'éventuels défauts dans la fiabilité. Cette correction est univariée, c'est-à-dire que les prévisions sont corrigées indépendamment pour chaque combinaison de variable météorologique, bassin et échéance. Les météorologues appelleront cette étape le « post-traitement », tandis que les hydrologues (auxquels nous nous revendiquons) l'appellerons « pré-traitement », car préalable à la modélisation hydrologique. Ce sera l'objet du chapitre 5.

Du fait de son caractère univarié, cette correction détruit la structure de dépendance que contenaient les prévisions brutes. Pour obtenir des prévisions multivariées cohérentes, il faut alors reconstruire cette structure de dépendance. Ce sera l'objet du chapitre 6.

Chapitre 5

Pré-traitement univarié

Différents systèmes de prévision météorologique probabiliste ont été présentés dans le chapitre 2, qu'ils soient basés sur la prévision d'ensemble (ECMWF-Ens, GEFS et PEARP) ou bien sur la prévision par analogie (ECMWF-Ana et NCEP-Ana). Dans le premier cas, nous pouvons craindre des défauts de fiabilité, car les méthodes de perturbation utilisées pour générer de la dispersion dans la prévision ne s'appliquent pas directement aux variables de surface telles que la précipitation et la température, mais bien en amont dans la modélisation. Dans le second cas en revanche, les prévisions sont issues d'une relation statistique calée par minimisation d'un CRPS, ce qui laisse présumer d'une meilleure fiabilité.

La section 5.1 de ce chapitre se propose de vérifier ces considérations émises à priori, en réalisant un diagnostic de la fiabilité des différents forçages bruts (section). Pour mémoire, la fiabilité est une condition nécessaire pour que le caractère probabiliste des prévisions puisse être exploité. Pour pallier aux éventuels défauts, nous réalisons une correction statistique basée sur l'approche EMOS, qui est décrite en section 5.2. Les résultats de ce pré-traitement sont présentés dans la section 5.3. Enfin, la section 5.4 propose une synthèse. Pour l'ensemble de ce chapitre, nous nous plaçons dans un cadre statistique univarié.

Il convient de rappeler que notre objectif n'est pas d'apporter des développements méthodologiques concernant le pré-traitement univarié, mais plutôt d'analyser les performances, sur notre cas d'étude, de méthodes existantes.

5.1 Diagnostic des forçages bruts

La vérification est réalisée sur la période 2011-2014, qui est commune à toutes les archives (ECMWF-Ens, GEFS, ECMWF-Ana, NCEP-Ana, et PEARP). Toutes sont évaluées sur la variable précipitation, tandis que pour la variable température seules le sont ECMWF-Ens et GEFS (les autres archives ne comprennent que la variable précipitation).

Rappelons que ces forçages sont émis à 00 h UTC, au pas de temps 6 h et avec un horizon de prévision de 120 h. Seul PEARP fait exception, les prévisions étant émises à 18 h UTC et jusqu'à un horizon de 108 h.

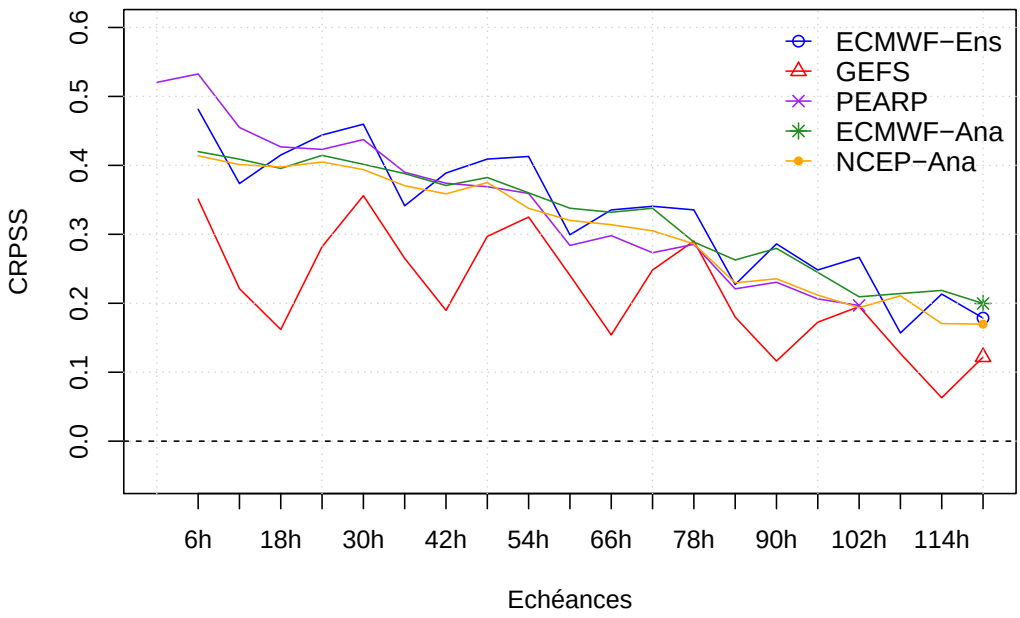

FIGURE 5.1 – CRPSS des prévisions brutes de précipitations, moyennés sur les 10 bassins. Les prévisions de référence sont les distributions climatologiques annuelles par bassin.

5.1.1 Précipitation

Intéressons-nous tout d'abord aux performances globales des prévisions de précipitation. La Figure 5.1 présente le CRPSS des prévisions de précipitation moyennés sur les 10 bassins de la zone d'étude, avec comme prévisions de référence les distributions climatologiques annuelles par bassin [1]. Les résultats montrent que ECMWF-Ens, PEARP, ECMWF-Ana, NCEP-Ana ont des performances très proches, tandis que GEFS présente des performances bien moindres. Les performances se dégradent logiquement avec l'échéance, mais restent toujours meilleures que celles des prévisions climatologiques. Enfin, on remarque que les prévisions d'ensemble présentent un cycle journalier de CRPSS bien plus marqué que les prévisions par analogie. L'origine de ce cycle a été discutée dans la section 4.3.6, à l'occasion de l'exemple numérique sur le potentiel de la stratification du CRPS.

Il est intéressant, avant de mettre en place une correction statistique, d'analyser la fiabilité des prévisions brutes, car c'est seulement s'il y a des défauts à corriger qu'un gain peut être espéré. La Figure 5.2 illustre les histogrammes de rang obtenus pour les échéances 12, 48 et 108 h, en regroupant les prévisions des 10 bassins de la zone d'étude au sein du même échantillon [2]. Ces histogrammes sont stratifiés selon la prévision moyenne de l'ensemble : égale à 0, entre 0 et 2, supérieure à 2 (unité : mm/6h).

1. Des tests avec des distributions climatologiques plus sophistiquées (par exemple une climatologie par fenêtre glissante, avec distinction des créneaux 00-06, 06-12, 12-18, 18-00 h UTC) ont conduits à des CRPSS quasiment identiques. La variable précipitation a donc une climatologie trop peu marquée pour que cela influe sensiblement le CRPSS.

2. Une stratification préalable selon le bassin a permis de montrer qu'il n'y avait pas de différences significatives de comportement des prévisions de précipitation entre les 10 bassins de la zone d'étude.

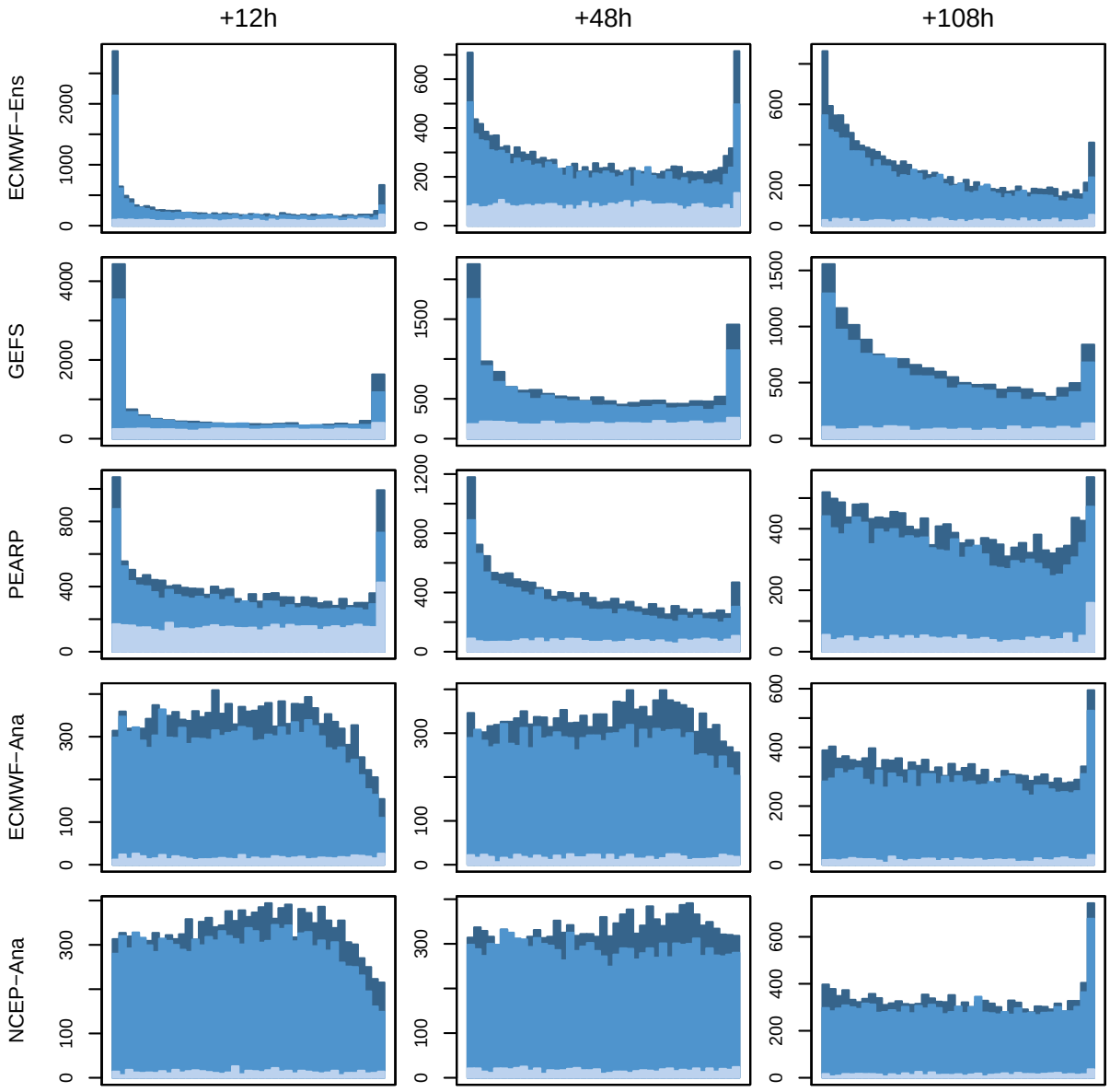

FIGURE 5.2 – Histogrammes de rang des prévisions brutes de précipitations pour différentes échéances. Les 10 bassins sont mélangés. Les strates coloriées en bleu claire, bleu et bleu foncé représentent respectivement les cas où la prévision moyenne est égale à 0, entre 0 et 2, et supérieure à 2 mm/6h.

Il y a une nette différence de comportement entre les prévisions d'ensemble et les prévisions par analogie. D'un coté, on note une sous-dispersion généralisée des prévisions d'ensemble, qui est très marquée pour les échéances courtes, puis s'amenuise progressivement mais reste présente jusqu'aux dernières échéances. Nous observons également une tendance à la sur-estimation (i.e., biais positif) lorsque les précipitations prévues sont faibles (strate intermédiaire).

De l'autre côté, les prévisions par analogie semblent relativement fiables. Il subsiste malgré tout certains défauts, qui diffèrent selon les échéances. Pour l'échéance 12 h par exemple, l'histogramme de la strate intermédiaire décroit pour les derniers rangs, ce qui témoigne de l'observation tombant trop rarement dans la partie haute de la distribution. Comme nous l'avons montré précédemment avec la Figure 4.12, ce phénomène est dû à des cas où l'observation est nulle mais où l'algorithme de prévision par analogie a néanmoins retenu quelques analogues pluvieux. La faible épaisseur de la strate inférieure témoigne en effet de la difficulté de l'algorithme à annoncer une probabilité d'occurrence égale à 0, c'est-à-dire à ne retenir aucun analogue pluvieux. Pour les précipitations plus fortes (strate supérieure), le problème s'inverse et l'algorithme éprouve des difficultés à annoncer une probabilité d'occurrence égale à 1, car quelques analogues secs subsistent fréquemment. L'observation est dans ces cas-là bien souvent non nulle, et par conséquent les rangs les plus à gauche de l'histogramme sont sous-peuplés. On observe en revanche des comportements différents pour les échéances plus lointaines, comme par example 108 h. À de telles échéances, la situation synoptique prévue est un peu moins conforme à celle observée, et donc les analogues retenus peuvent parfois être très différents de ceux qui auraient été retenus en prévision synoptique parfaite. Cette erreur « synoptique » semble alors se traduire par une dispersion trop faible pour capter correctement les précipitations faibles mais non-nulles, comme le montre la strate intermédiaire où le dernier rang est sur-peuplé.

Pour conclure sur la variable précipitation, il semble qu'il y ait davantage de gains à attendre de la correction statistique pour les prévisions d'ensemble que pour les prévisions par analogie. Nous verrons dans la section 5.3.1 si cela se confirme.

5.1.2 Température

Pour mémoire, seules les archives ECMWF-Ens et GEFS contiennent des prévisions de température. La Figure 5.3 présente en trait plein le CRPSS moyennés sur les 10 bassins, avec comme prévisions de référence les distributions climatologiques construites un utilisant une fenêtre glissante de 2 mois et en distinguant les créneaux 00-06, 06-12, 12-18, 18-00 h UTC [3]. On constate que les performances sont plus faibles que pour la précipitation, et par ailleurs présentent un cycle journalier plus important.

Ces performances limitées s'expliquent par le fait que la température des bassins correspond à la température du point de grille le plus proche de chaque bassin, qui est ensuite extrapolée pour la faire correspondre à l'altitude médiane du bassin (cf. 2.3.1.3). Cette

3. Cette climatologie plus sophistiquée semble judicieuse pour la température, celle-ci présentant un cycle annuel et journalier bien plus important que pour la précipitation.

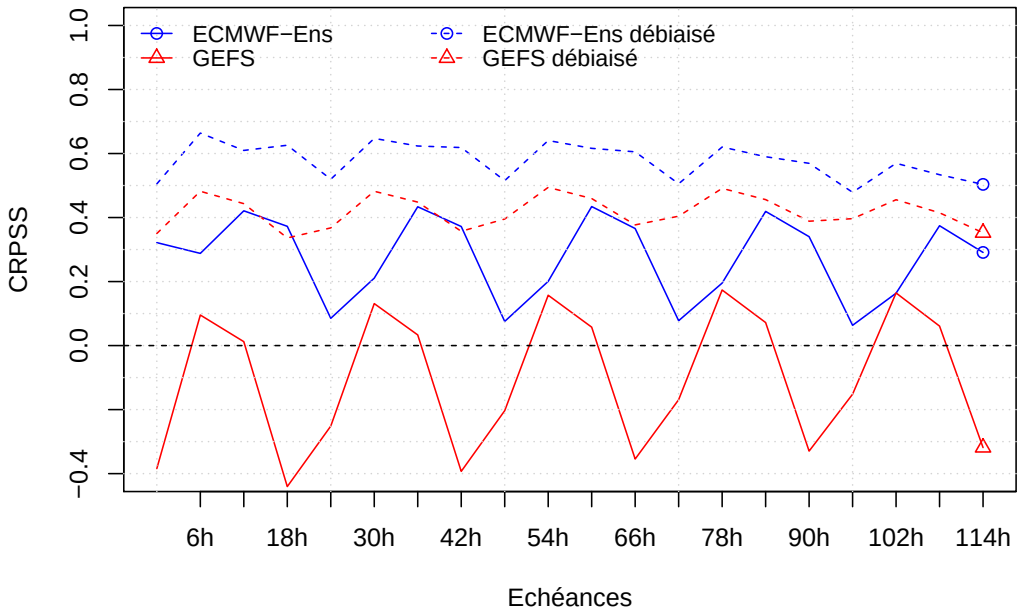

FIGURE 5.3 – CRPSS des prévisions de température brutes et corrigées du biais systématique, moyennés sur les 10 bassins. Les prévisions de référence sont les distributions climatologiques construites par fenêtre glissante de 2 mois, avec distinction des 4 créneaux journaliers.

extrapolation s'appuie sur des grilles différentes pour les prévisions et les observations, ce qui peut introduire un biais systématique dans les prévisions. La Figure 5.4 présente les histogrammes de rang pour les échéances 12, 48 et 108 h, pour le bassin de la Valserine [4]. On observe en effet un biais systématique fort, les observations étant, dans le cas de la Valserine, systématiquement sous-estimées (c'est l'inverse pour certains autres bassins).

Pour mieux appréhender la qualité réelle des prévisions de température, il est pertinent de réduire ce biais systématique, qui provient davantage de notre contexte d'utilisation que des modèles de prévisions eux même. Nous proposons pour cela une correction additive simple, définie par :

$$T_{\text{debiaisée}\,j} = T_{\text{brute}\,j} + a_j, \tag{5.1}$$

où $j \in \{1, \ldots, J\}$ dénote les différents bassins. Cette correction est indépendante du jour calendaire, du créneaux 6 h de la journée ou encore de l'échéance de prévision. Le coefficient a_j est obtenu pour chaque bassin j en minimisant le biais sur l'année civile précédente. Par exemple, les prévisions émises en 2013 sont débiaisées à partir du coefficient calculé sur 2012. Le CRPSS obtenu avec ces prévisions débiaisées est tracé en trait pointillé sur la Figure 5.3. Les performances sont alors significativement meilleures, pour ECMWF-Ens comme pour GEFS. En terme de fiabilité, la Figure 5.5 montre que le biais systématique a été corrigé. En revanche les prévisions demeurent fortement sous-dispersives, y compris jusqu'aux dernières échéances.

4. Les prévisions des différents bassins ne sont cette fois pas regroupées au sein du même échantillon pour la construction des histogrammes de rang, car leur climatologie est différente du fait de leur altitude différente.

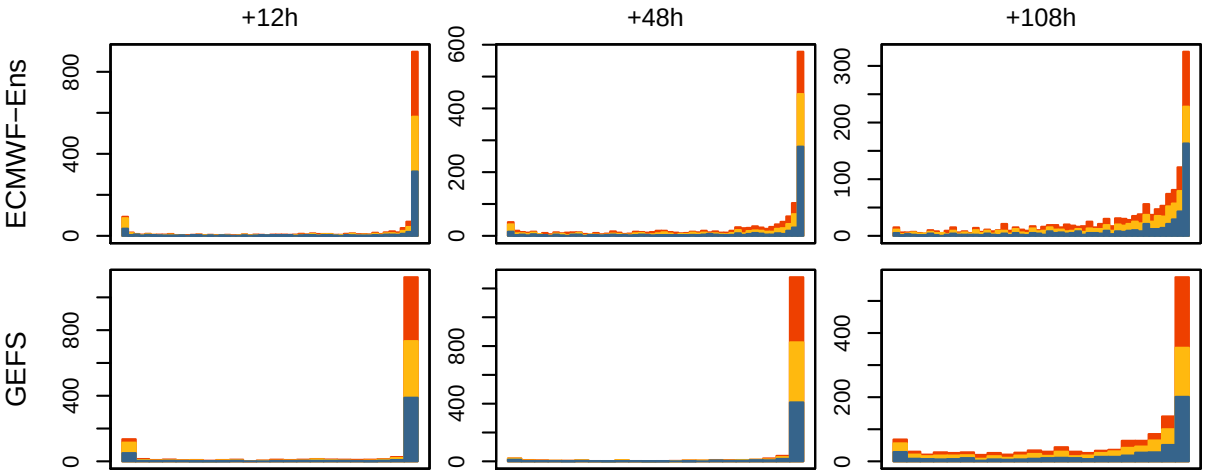

FIGURE 5.4 – Histogrammes de rang des prévisions brutes de température pour différentes échéances, sur le bassin de la Valserine. La stratification est cette fois réalisée de manière à obtenir le même nombre de prévisions dans chaque strate. Le critère de stratification reste la prévision moyenne.

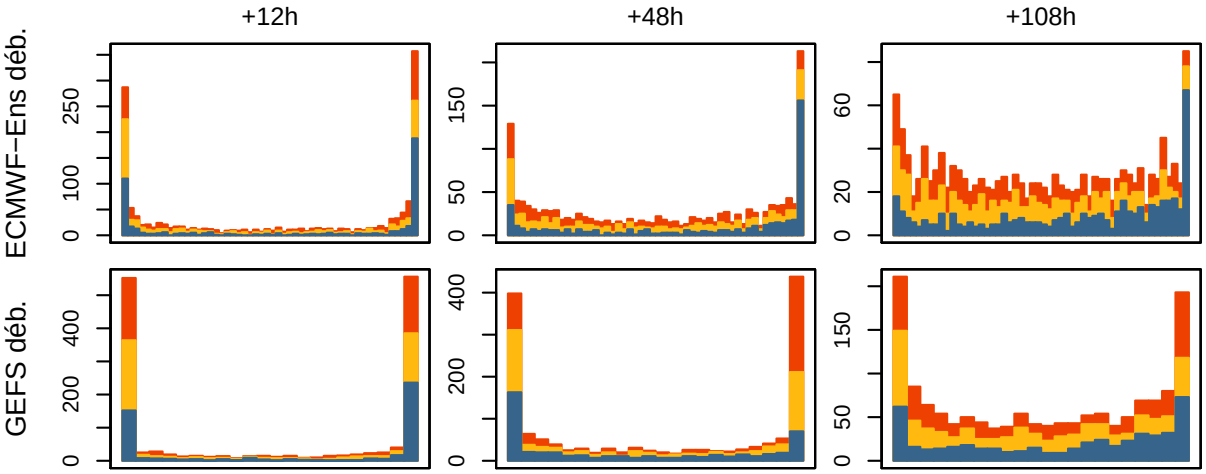

FIGURE 5.5 – Idem que la Figure 5.4, mais après correction du biais systématique.

5.2 Pré-traitement via l'EMOS

Ainsi, nos prévisions d'ensemble de précipitation comme de température présentent des défauts de fiabilité importants, et notamment une sous-dispersion notable. Leur correction statistique, qui vise à corriger la fiabilité dans le but d'augmenter la qualité globale, constitue l'étape de pré-traitement.

D'une manière générale, la correction statistique de prévisions probabilistes (ici sous la forme ensembliste) vise à construire un modèle statistique qui relie les prévisions aux observations. Plus précisément, ce modèle décrit la densité de probabilité $p(y|x_1, \ldots, x_M)$ de l'observation y sachant la prévision $x_1, \ldots, x_M$. L'étape de calage, qui s'appuie sur un échantillon historique de couples prévision-observation, permet de déterminer les paramètres du modèle. Une fois calé, celui-ci est alors en mesure de fournir la densité de l'observation sachant chaque nouvelle prévision. Cette densité conditionnelle, appelée densité prédictive, remplace alors la prévision « brute ».

L'approche EMOS (*ensemble model output statistic* ; Gneiting *et al.*, 2005) considère que la densité prédictive $p(y|x_1, \ldots, x_M)$ est une densité unique, qui est décrite par une loi de distribution dont les paramètres sont reliés à des statistiques de l'ensemble brut (moyenne, écart-type, etc.) via des équations de régression. Pour cette raison, certain auteurs préfèrent parler de méthodes de régression (Wilks, 2018a), en opposition aux approches par habillage des membres, où la densité prédictive est une densité de mélange issue des densités qui « habillent » chacune un membre de la prévision brute [5]. Cette dernière catégorie de méthodes inclut la BMA (*Bayesian model averaging*), qui sera utilisée pour le post-traitement des prévisions de débit (cf. 7.3).

Ainsi, la performance d'un modèle d'EMOS, c'est-à-dire le gain en qualité des prévisions corrigées par rapport aux prévisions brutes, dépend principalement de :

- l'adéquation de la loi de distribution au comportement de l'observation,
- la capacité des équations de régression à conditionner les paramètres de la loi de distribution aux statistiques de l'ensemble brut (et donc à proposer des distributions prédictives qui s'écartent de la distribution climatologique).

Nous décrivons tout d'abord l'EMOS dans sa version standard, qui sera utilisé pour corriger les prévisions de température. La section 5.2.2 présentera ensuite une variante proposée par Scheuerer et Hamill (2015b) pour la correction des prévisions de précipitation.

5.2.1 L'EMOS-normal pour la température

La première version d'EMOS a été proposée par Gneiting *et al.* (2005), pour la correction de prévisions d'ensemble de variables gaussiennes comme la température ou la

5. L'auteur partage l'avis de Wilks (2018a). En effet, l'appellation EMOS provient de « ensemble-MOS », sachant que le MOS peut être considéré comme le synonyme de post-traitement dans le cadre déterministe (Glahn *et al.*, 2009). Sous cet angle, les méthodes par habillage des membres sont également des approches « ensemble-MOS ». Dans ce mémoire cependant, nous conserverons l'appellation EMOS, car c'est actuellement sous ce nom que la méthode proposée par Gneiting *et al.* (2005) est la plus connue.

pression. Ainsi, ce modèle stipule que l'observation sachant la prévision suit une loi normale paramétrée par la moyenne μ et la variance σ^2 :

$$y|x_1, \ldots, x_M \backsim \mathcal{N}(\mu, \sigma^2). \tag{5.2}$$

Il s'agit ensuite de choisir les statistiques qui « résumeront » l'ensemble brut $x_1, \ldots, x_M$. Dans cette version d'EMOS, que nous appelons EMOS-normal, ces statistiques sont la moyenne et la variance estimée :

$$\bar{x} = \frac{1}{M} \sum_{m=1}^{M} x_m \tag{5.3}$$

$$s^2 = \frac{1}{M+1} \sum_{m=1}^{M} (x_m - \bar{x})^2. \tag{5.4}$$

Enfin, ces statistiques sont reliées aux paramètres de la loi normale par le modèle de régression suivant, à 4 coefficients :

$$\mu = a_0 + a_1 \bar{x} \tag{5.5}$$

$$\sigma^2 = b_0 + b_1 s^2, \tag{5.6}$$

avec b_0 et b_1 non négatifs.

Ainsi, la moyenne μ et la variance σ^2 des distributions prédictives sont des fonctions linéaires des moyennes $\bar{x}$ et variances s^2 de l'ensemble brut, respectivement. Les coefficients a_0 et a_1 permettent de corriger des biais systématiques dans la moyenne, tandis que les coefficients b_0 et b_1 permettent d'ajuster la dispersion de manière à correctement représenter l'incertitude de prévision.

Ces coefficients de régression a_0, a_1, b_0 et b_1, qui sont les paramètres du modèle d'EMOS, sont estimés à partir d'un échantillon de calage contenant suffisamment de couples prévision-observation. Afin d'adapter au mieux le modèle d'EMOS aux caractéristiques des erreurs de prévision, cet échantillon de calage est propre à chaque bassin, mais également à chaque échéance. La dépendance des caractéristiques de l'erreur aux échéances de prévision est quelque chose que l'on retrouve quasi systématiquement dans les systèmes ensemblistes dont la prévision d'ensemble météorologique fait partie. En effet, il est extrêmement complexe de générer une dispersion des membres qui soit adaptée aux erreurs de prévision pour l'intégralité des variables, localisations et échéances. Les Figures 5.2 et 5.5 le confirment, avec des histogrammes de rang montrant des défauts de fiabilité qui diffèrent d'une échéance à l'autre.

La technique classique d'estimation des paramètres dans ce type de problème consiste à maximiser la fonction de vraisemblance (ou plus généralement son logarithme) des données de calage. Cependant, Gneiting *et al.* (2005) montrent que, dans le cas de l'EMOS, cette approche mène à des prévision sur-dispersives. Ils suggèrent à la place une approche de minimisation du CRPS : on cherche alors le jeu de coefficients qui donne le plus faible CRPS global (cf. 4.2.1.1) sur l'échantillon de calage. Cette technique est rendue accessible

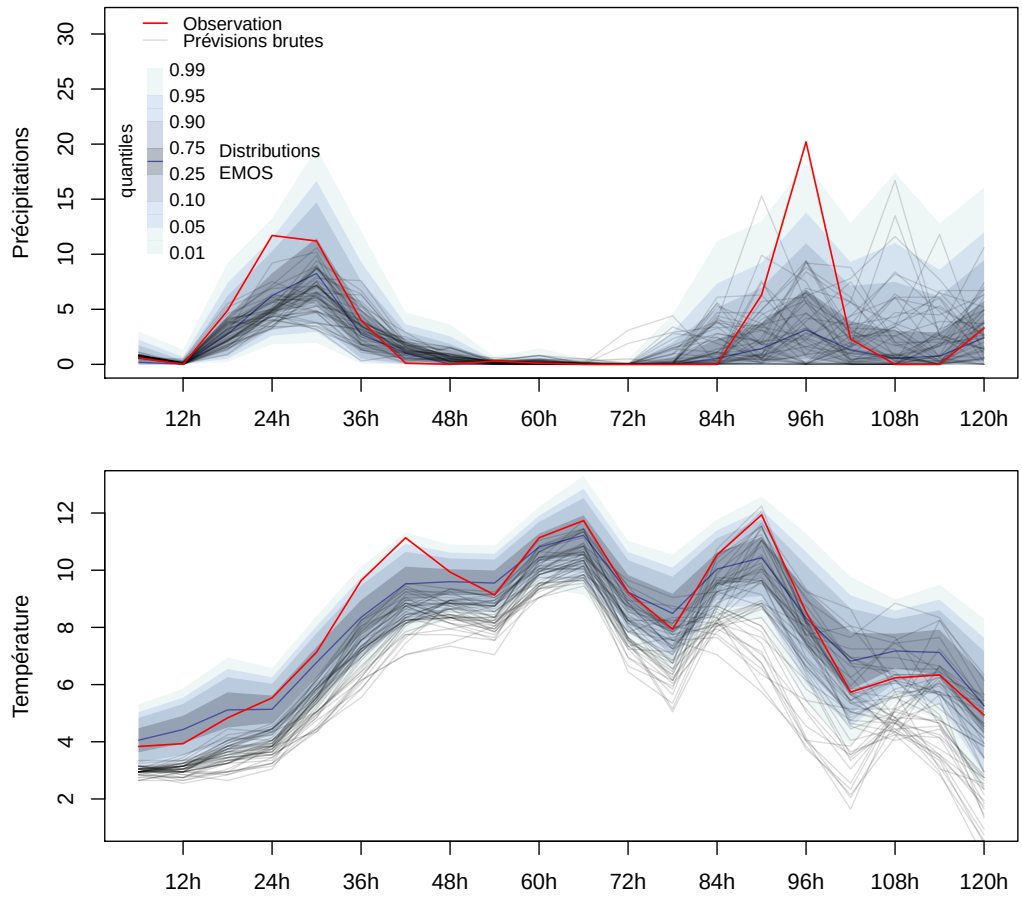

FIGURE 5.6 – Illustration des distributions prédictives calculées par l'EMOS-CSG (précipitations) et l'EMOS-normal (température), à partir des prévisions ECMWF-Ens brutes sur le bassin de la Valserine et émises le 5 novembre 2013.

numériquement par le fait qu'une formulation analytique du CRPS peut être obtenue lorsque la prévision est une loi normale (Gneiting *et al.*, 2005, équation 5). Wilks (2018b) recommande néanmoins d'être vigilant lorsque l'on cale une correction statistique en minimisant le CRPS, car cela ne garantit pas obligatoirement la fiabilité des distributions prédictives. Il propose alors de réaliser la minimisation en adjoignant au CRPS une pénalité pour les défauts de calibration. Cette alternative n'a pas pu être testée, bien qu'elle semble tout à fait cohérente avec notre ligne de conduite qui est de s'assurer de la fiabilité avant de chercher à maximiser la finesse.

Pour plus de détails sur cette méthode EMOS, nous invitons le lecteur à lire l'article original de Gneiting *et al.* (2005). Dans la suite, nous utilisons la version implémentée dans le package `ensembleMOS` de `R`, et l'appliquons sur les prévisions de température ECMWF-Ens et GEFS brutes, c'est-à-dire sans la correction additive de l'équation (5.1). Nous laissons en effet le soin au pré-traitement EMOS que de réaliser lui-même la correction de biais. Un exemple des distributions prédictives obtenues est illustré par la Figure 5.6 (graphique du bas).

Concernant la longueur de la période de calage, une analyse de sensibilité balayant l'ensemble des échéances a montré qu'une période de 50 jours (c'est-à-dire 50 couples

prévision-observation) donnait les meilleurs résultats dans notre contexte. Ainsi, la prévision de chaque date d sur la période 2011-2014 est corrigée en exploitant les prévisions émises durant la période allant des dates $d-55$ à $d-6$ (les prévisions de $d-5$ à $d-1$ n'ayant pas encore les observations disponibles pour toutes les échéances, en supposant un contexte opérationnel). Une si courte période de calage est possible grâce au fait que la méthode EMOS est paramétrique : la relation entre prévision et observation pour des températures modérées est extrapolée à des températures plus basses ou plus hautes grâce au modèle de régression. Cette extrapolation est rendue stable par le fait les erreurs de prévision de température sont homoscédastiques, c'est-à-dire qu'elles ne dépendent pas de la température elle même. En d'autres termes, tous les couples prévision-observation de l'échantillon de calage apportent une information utile pour la correction d'une prévision donnée. Nous verrons que ce qui n'est pas le cas pour la variable précipitation, qui requiert une bien plus grande période de calage.

5.2.2 L'EMOS-CSG pour les précipitations

L'EMOS-normal s'appuie sur une distribution prédictive sous la forme d'une loi normale, loi qui n'est cependant pas adaptée à la variable précipitation. Cette dernière est en effet asymétrique, non négative, et présente un point d'accumulation en zéro. Nous proposons donc d'utiliser une variante de l'EMOS issue des travaux de Scheuerer et Hamill (2015b), qui s'appuie la loi Gamma, notée $\Gamma(k,\theta)$ où k est le paramètre de forme et θ le paramètre d'échelle. La loi Gamma est intéressante pour la modélisation des précipitations car d'une part ses valeurs sont non négatives, et d'autre part sa densité est asymétrique, avec une queue plus longue dans les valeurs fortes. Pour se rapprocher de l'EMOS-normal, les auteurs préfèrent paramétrer la loi Gamma en utilisant ses deux premiers moments μ et σ^2 plutôt que k et θ, grâce aux relations :

$$\mu = k\theta \tag{5.7}$$

$$\sigma^2 = k\theta^2. \tag{5.8}$$

Prise telle quelle, la loi Gamma ne permet cependant pas de modéliser le point d'accumulation en zéro. Ils proposent donc de décaler d'un paramètre δ la densité vers la gauche, puis de censurer à zéro (Figure 5.7). Cette censure, qui revient à remplacer les valeurs négatives par zéro, permet la modélisation simultanée de la probabilité d'occurrence et de la quantité de précipitation. La distribution ainsi décalée, notée $\tilde{\Gamma}(\mu,\sigma^2,\delta)$, est appelée distribution CSG (*censored, shifted Gamma*).

La variante d'EMOS proposée pour les précipitations, appelée EMOS-CSG, stipule donc que l'observation sachant la prévision suit une loi CSG :

$$y|x_1,\ldots,x_M \backsim \tilde{\Gamma}(\mu,\sigma^2,\delta). \tag{5.9}$$

Avant de modéliser ces distributions prédictives, Scheuerer et Hamill (2015b) proposent d'ajuster une loi CSG sur les seules observations de l'échantillon de calage. Les paramètres μ_{cl}, σ^2_{cl} et δ_{cl} de cette loi, obtenus par une méthode d'ajustement, décrivent alors la

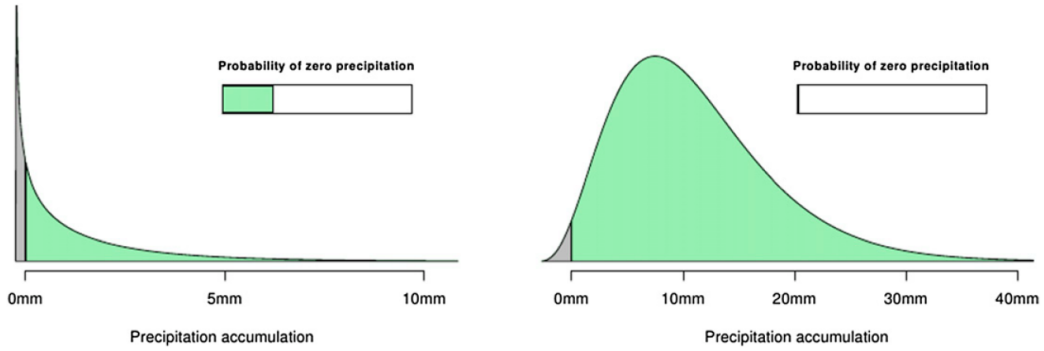

FIGURE 5.7 – Exemples de densités issues de la loi Gamma décalée et censurée (loi CSG). La fraction de densité pour les valeurs négatives est transformée en probabilité de non-occurrence. Figure tirée de Scheuerer et Hamill (2015b).

distribution climatologique de l'observation. Nous verrons plus loin que le modèle de régression permet de faire converger les distributions prédictives (conditionnelles) vers cette distribution climatologique (non conditionnelle) dès lors que les prévisions brutes n'ont plus de compétence, par exemple pour les échéances lointaines.

Pour conditionner la distribution prédictive à l'ensemble brut, les auteurs proposent de résumer cet ensemble brut par les statistiques suivantes :

$$\bar{x} = \frac{1}{M}\sum_{m=1}^{M} x_m, \tag{5.10}$$

$$\mathrm{MD} = \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} |x_m - x_{m'}|, \tag{5.11}$$

$$\mathrm{POP} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}_{\{x_m>0\}}, \tag{5.12}$$

où $\mathbb{1}_{\{\cdot\}} = 1$ si $\{\cdot\}$ est vraie et 0 sinon. Moyenne et dispersion sont représentées par les statistiques $\bar{x}$ et MD, respectivement, tandis que la troisième statistique, POP, mesure la probabilité d'occurrence de précipitation.

Ces statistiques sont ensuite reliées aux paramètres μ, σ^2 et δ de la distribution prédictive via le modèle de régression à 6 coefficients suivant :

$$\mu = \frac{\mu_{cl}}{a_1}\ \mathrm{log1p}\left[\mathrm{expm1}(a_1)\left(a_2 + a_3\mathrm{POP} + a_4\frac{\bar{x}}{\bar{x}_{cl}}\right)\right] \tag{5.13}$$

$$\sigma = b_1\ \sigma_{cl}\sqrt{\frac{\mu}{\mu_{cl}}} + b_2\ \mathrm{MD} \tag{5.14}$$

$$\delta = \delta_{cl}, \tag{5.15}$$

où $\mathrm{log1p}(u) = \log(1+u)$ et $\mathrm{expm1}(u) = \exp(u)-1$. La normalisation dans l'équation (5.13) de $\bar{x}$ par sa moyenne $\bar{x}_{cl}$ sur l'échantillon de calage permet de stabiliser le coefficient a_4 face à une période de calage qui serait très différente de celle de validation. D'une manière similaire à l'EMOS-normal, les coefficients a_1, a_2, a_3, a_4, b_1 et b_2 de l'EMOS-CSG sont

estimés par minimisation du CRPS sur l'échantillon de calage. Pour cela, une formulation analytique du CRPS pour la loi CSG a été proposée (Scheuerer et Hamill, 2015b, équation 10).

Ce modèle de régression, complexe au premier abord, est intéressant sur plusieurs aspects :

- Il permet aux distributions prédictives de converger vers la distribution climatologique si les prévisions brutes n'ont plus de compétence. En effet, on retrouve $\mu = \mu_{cl}$ et $\sigma = \sigma_{cl}$ si $a_3 = a_4 = b_2 = 0$ et $a_2 = b_1 = 1$ (le choix de a_1 important peu car $\text{log1p}(\text{expm1}(a_1)) = a_1$).

- Il introduit, via le premier terme de l'équation (5.14), une dépendance de σ par rapport à μ, et donc une hétéroscédasticité explicite. Ainsi, l'incertitude prédictive augmente avec la quantité de précipitations prévue. Cette dépendance, absente dans l'équation (5.6) de l'EMOS-normal, est particulièrement intéressante pour la correction des prévisions fortement sous-dispersives, où la dispersion MD n'est pas toujours suffisamment corrélée à l'erreur de prévision.

- Il permet de réduire la croissance de la moyenne μ avec celles des prédicteurs POP et $\bar{x}$, grâce au logarithme (et au paramètre a_1) dans l'équation (5.13). Les auteurs montrent en effet qu'une croissance linéaire n'est pas toujours appropriée, notamment pour les prévisions de faible compétence (échéances lointaines par exemple) où des cumuls élevés peuvent s'avérer très peu fiables.

Pour plus d'informations sur l'EMOS-CSG, nous invitons le lecteur à se reporter à l'article original de Scheuerer et Hamill (2015b). La version que nous utilisons et avons présenté ci-dessus est une version simplifiée de la version proposée par ces auteurs. Premièrement, la version originale exploite, pour le calcul des statistiques de l'ensemble brut de chaque bassin, les prévisions de plusieurs points de grille alentours et externes au bassin, via un système de pondération. Selon Scheuerer et Hamill (2015b) mais également Scheuerer (2014), l'exploitation des prévisions sur ce voisinage permet des gains substantiels de performance, en réduisant l'effet des « erreurs de position » (des cellules pluvieuses) que peuvent potentiellement faire les modèles météorologiques. Dans notre cas cependant, nous avons dès le début de la thèse calculé les prévisions de précipitation par bassin, et il était chronophage de remonter aux prévisions en points de grille pour pouvoir appliquer cette procédure. C'est cependant une piste intéressante d'amélioration. Ensuite, la version originale inclut une procédure de *quantile mapping* qui vise à ajuster, préalablement à l'EMOS, les membres des prévisions de manière à ce qu'ils aient la même climatologie que l'observation. D'après Scheuerer (communication personnelle), cette procédure est surtout nécessaire lorsque les prévisions sur le voisinage du bassin sont exploitées, car les points de grille de ce voisinage peuvent avoir des climatologies très différentes, notamment en zone de montagne. N'exploitant pas les prévisions sur ce voisinage, nous avons écarté cette procédure. Enfin, il faut noter qu'une version encore simplifiée de l'EMOS-CSG, issue des travaux de Baran et Nemoda (2016), est depuis peu disponible dans le package `ensembleMOS` de `R`.

Concernant le calage, il s'avère que l'EMOS-CSG requiert une période bien plus longue que celle nécessaire à l'EMOS-normal pour la température, pour trois raisons. Premièrement, le nombre de coefficients du modèle de régression, c'est-à-dire le nombre de paramètres à caler, est plus important (6 au lieu de 4). Ensuite, la prévision de la variable précipitation étant fortement héteroscédastique, le modèle de régression n'est stable que si l'échantillon de calage contient des prévisions suffisamment variées pour balayer l'étendue du spectre des valeurs possibles de précipitation. Enfin, les échantillons de calage contiennent un grand nombre de paires prévision-observation qui sont égales à zéro, paires qui contiennent alors une information très limitée pour le calage du modèle.

Scheuerer et Hamill (2015b) ont quantifié la perte de performance en passant de 11 années de calage à 3 puis à 1, à chaque fois avec des modèles d'EMOS-CSG mensuels qui utilisent, pour le calage, les prévisions des mois précédents, courant et suivants (i.e., une fenêtre glissante de 91 jours). Les tailles des échantillons de calage sont donc respectivement 11×91, 3×91 et 1×91. Ils observent une dégradation très importante des performances dans le cas d'une unique année de calage, tandis que cette dégradation est moins importante avec 3 années de calage, mais reste significative.

Face à ces résultats et compte-tenu de la longueur de nos archives de prévision, nous décidons de corriger nos prévisions de précipitation en exploitant pour le calage les prévisions émises durant les 4 années antérieures, tout en conservant l'idée de la fenêtre glissante (afin de conserver une certaine saisonnalité). Cela nous permet de corriger les prévisions ECMWF-Ens, GEFS et ECMWF-Ana sur la période 2011-2014, grâce à la longueur des archives qui remontent jusqu'en 2007. En revanche, ce choix nous empêche de corriger les prévisions PEARP, dont l'archive débute en 2010. Nous avons conservé la fenêtre de 3 mois (91 jours), ce qui conduit à des échantillons de calage de taille $4 \times 91 = 364$. Des tests avec des fenêtres glissantes de 4 mois (121 jours) et 6 mois (182 jours) n'ont pas permis d'améliorer les performances.

5.3 Résultats

Nous évaluons maintenant les prévisions de précipitation et de température corrigées à l'aide des méthodes EMOS, en faisant appel au CRPSS pour quantifier le gain de qualité globale. Il est important de noter que les prévisions corrigées sont désormais sous la forme de distributions continues (paramétriques), tandis que les prévisions brutes sont sous la forme ensembliste. Nous exploitons cette forme continue en calculant les CRPSS (ou plus exactement les CRPS) des prévisions corrigées de manière analytique, conformément à notre choix d'évaluer les prévisions dans la forme où elles sont disponibles [6].

La vérification de la fiabilité à l'aide de l'histogramme de rang impose en revanche que les prévisions soient sous forme ensembliste. Nous procédons alors à un échantillonnage des distributions continues, en sélectionnant les quantiles équidistants définis par les niveaux $\alpha_m = m/(M+1)$ pour $m = 1, \ldots, M$. La taille M de ces ensembles est prises égale à

6. En discrétisant les prévisions pré-traitées, nous aurions obtenu des CRPS plus élevés, et donc des CRPSSS (i.e., des performances) moindres.

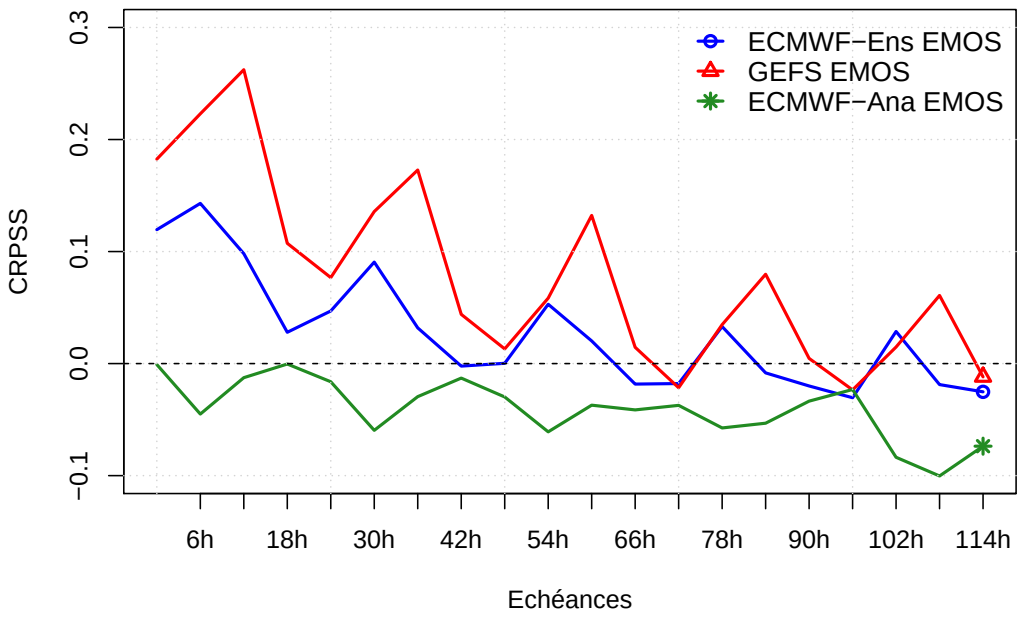

FIGURE 5.8 – CRPSS des prévisions pré-traitées de précipitations, moyennés sur les 10 bassins. Les prévisions de référence sont les prévisions brutes (les 3 systèmes n'ont donc pas les mêmes prévisions de référence).

celle de l'ensemble brut, de manière à ce que les histogrammes de rang soient directement comparables.

5.3.1 Précipitation

La Figure 5.8 présente le CRPSS des prévisions de précipitations ECMWF-Ens, GEFS et ECMWF-Ana, chacun calculé avec ses propres prévisions brutes comme prévisions de référence. Le CRPSS le plus élevé est obtenu avec les prévisions GEFS, suivi de ECMWF-Ens. Pour ces deux systèmes, le CRPSS reste positif jusqu'à environ 3-4 jours d'échéance. Dit autrement, le pré-traitement pour ces deux systèmes de prévision d'ensemble est bénéfique jusqu'à environ 3-4 jours, tandis que l'on observe une dégradation pour les échéances plus lointaines. Par ailleurs, on remarque que le cycle diurne est en opposition de phase avec celui de la Figure 5.1, ce qui signifie que le pré-traitement apporte davantage de gain pour les créneaux 6 h qui avaient le moins de compétence.

Ces gains en qualité globale étaient espérés pour les prévisions d'ensemble, compte-tenu des défauts de fiabilité diagnostiqués à la Figure 5.2. La Figure 5.9 permet d'évaluer la fiabilité après leur pré-traitement, en traçant les histogrammes de rang pour les mêmes échéances. Les différentes strates sont désormais relativement plates pour GEFS comme pour ECMWF-Ens ce qui témoigne d'une bonne fiabilité. Il y a néanmoins une exception pour les échéances courtes (12 h sur la figure), où ECMWF-Ens tend à conserver un biais positif et GEFS un biais négatif. Comportement qui n'apparaissait pas sur les histogrammes avant pré-traitement (Figure 5.2), et que nous ne sommes pas parvenus à expliquer.

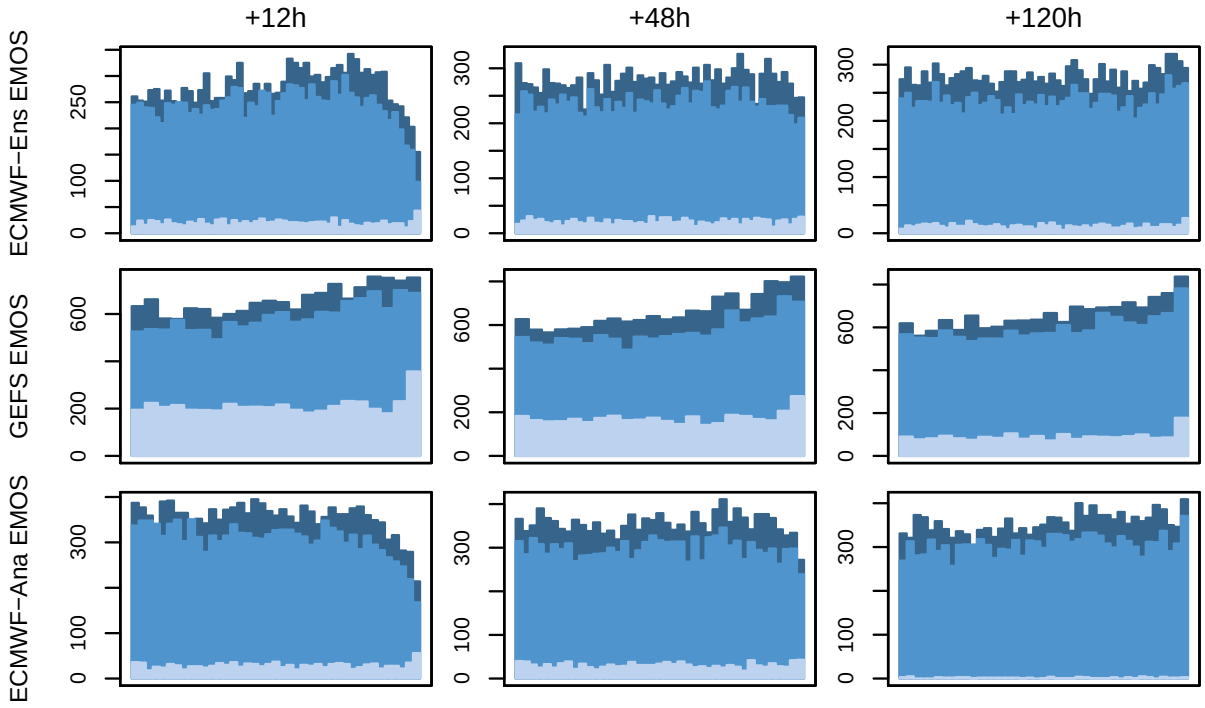

FIGURE 5.9 – Histogrammes de rang des prévisions pré-traitées de précipitations, pour différentes échéances. Les 10 bassins sont mélangés. La même stratification qu'à la Figure 5.2 est appliquée.

La Figure 5.8 montre par ailleurs que le pré-traitement via la méthode EMOS-CSG dégrade les prévisions par analogie ECMWF-Ana. Le diagnostic des prévisions brutes avait souligné la fiabilité déjà correcte des prévisions par analogie. Prévisions qui sont issues, faut-il le rappeler, d'un modèle statistique déjà construit (par minimisation du CRPS) sur la relation entre les prévisions (synoptiques, certes) et l'observation. Il est ainsi peu surprenant qu'un modèle statistique supplémentaire, en l'occurrence l'EMOS-CSG, ne parvienne pas à améliorer les performances. Les histogrammes de rang sur la Figure 5.9 témoignent bien d'un gain en fiabilité pour les échéances plus lointaines (108 h sur la figure), mais il semble que ce gain ait été contrebalancé par une perte importante de finesse.

La Figure 5.10 trace désormais le CRPSS des prévisions pré-traitées avec les prévisions climatologiques comme prévisions de référence. Considérer la même référence pour les trois systèmes de prévision permet maintenant de déterminer lequel d'entre eux présente la meilleure qualité globale. Ainsi, malgré le gain plus important du pré-traitement sur les prévisions GEFS, les prévisions ECMWF-Ens pré-traitées restent au final celles qui ont le CRPSS le plus élevé. Néanmoins, nous observons qu'au fur et à mesure que l'échéance augmente, l'écart de CRPSS entre les prévisions ECMWF-Ens pré-traitées et ECMWF-Ana brutes tend à se réduire. Les prévisions par analogie restent même plus performantes pour les échéances les plus lointaines (au delà de 96 h). Cela confirme donc l'intérêt de la méthode, notamment dans des contextes de mise en alerte aux inondations, où l'on recherche des prévisions performantes pour des échéances lointaines.

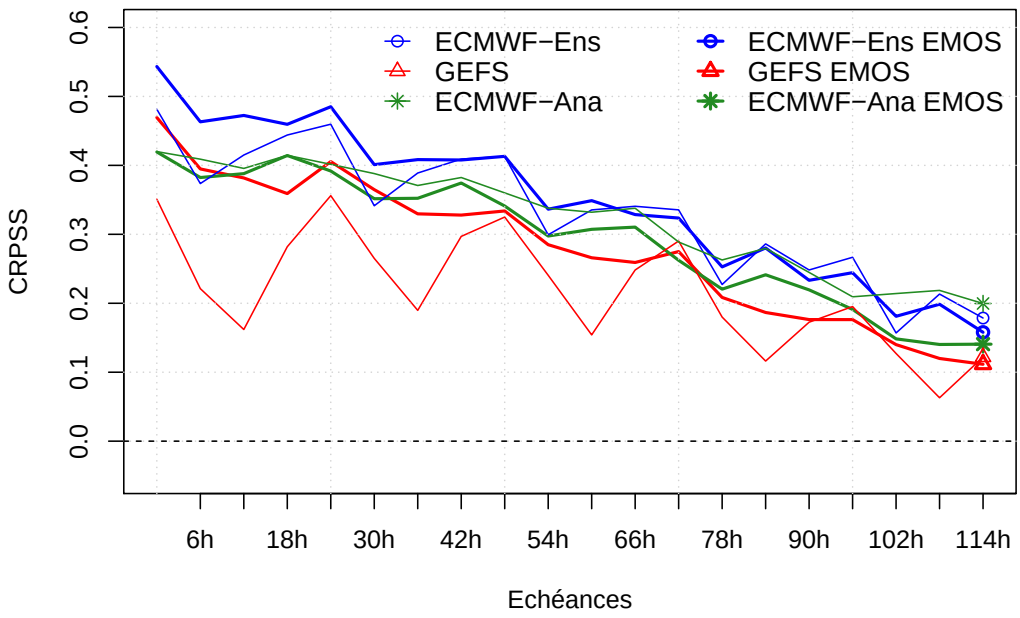

FIGURE 5.10 – Idem que Figure 5.8, mais avec les prévisions climatologiques comme prévisions de référence.

5.3.2 Température

Concernant la température, les Figures 5.11 et 5.12 présentent les CRPSS des prévisions corrigées via l'EMOS-normal, avec dans le premier cas les prévisions brutes comme prévisions de référence, et dans le second cas les prévisions climatologiques (la même référence, donc). Sur chacun des graphiques est également reporté le CRPSS des prévisions simplement débiaisées (cf. 5.1.2). Les résultats montrent que, pour ECMWF-Ens comme pour GEFS, l'EMOS-normal parvient à fournir des prévisions plus performantes que les prévisions débiaisées. En prenant la climatologie comme référence, on constate que les prévisions ECMWF-Ens pré-traitées ont de meilleures performances que les prévisions GEFS pré-traitées.

La Figure 5.13 confirme que les prévisions pré-traitées ont une meilleure fiabilité que les prévisions brutes et débiaisées. Cependant, cette fiabilité n'est pas parfaite : il semble subsister une tendance à la sous-dispersion, qui est fortement accentuée pour les valeurs faibles (en témoigne le premier rang surpeuplé de l'histogramme). Cela s'observe pour ECMWF-Ens comme pour GEFS, sur les échéances courtes comme les échéances longues. Nous ne sommes pas parvenus à identifier l'origine de ce défaut, qui s'avère également être présent dans les histogrammes de Baran *et al.* (2014, figure 7) de prévisions de température corrigées via la même méthode d'EMOS-normal. Ce défaut persistant conforte l'intérêt de l'approche de Wilks (2018b), qui est d'ajouter au CRPS lors du calage une pénalité qui tienne compte des défauts de fiabilité.

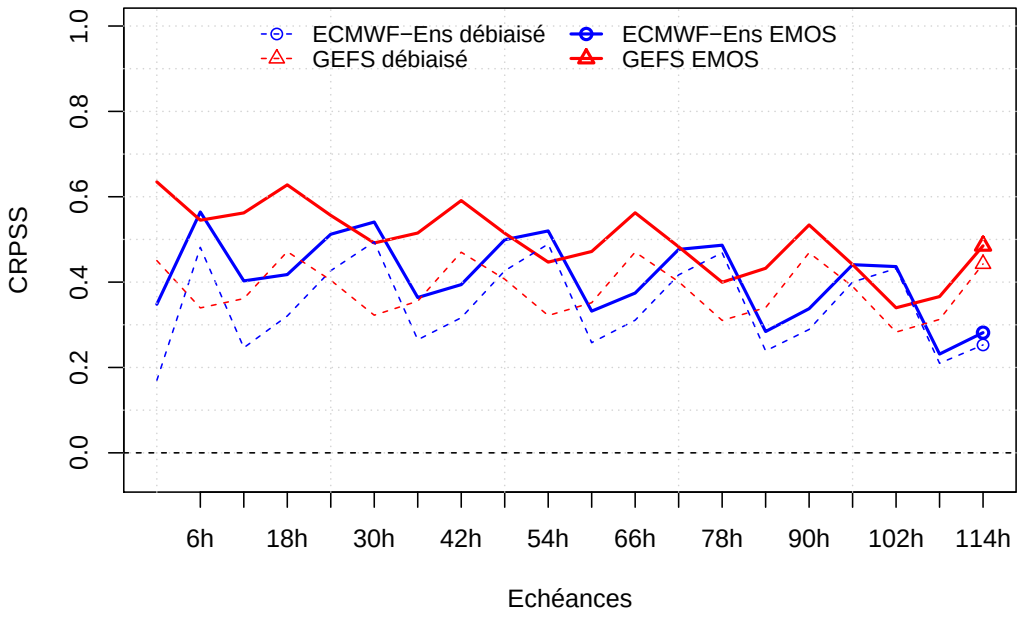

FIGURE 5.11 – CRPSS des prévisions pré-traitées de température, moyennés sur les 10 bassins. Les prévisions de référence sont les prévisions brutes (les courbes bleues et rouges n'ont donc pas les mêmes prévisions de référence).

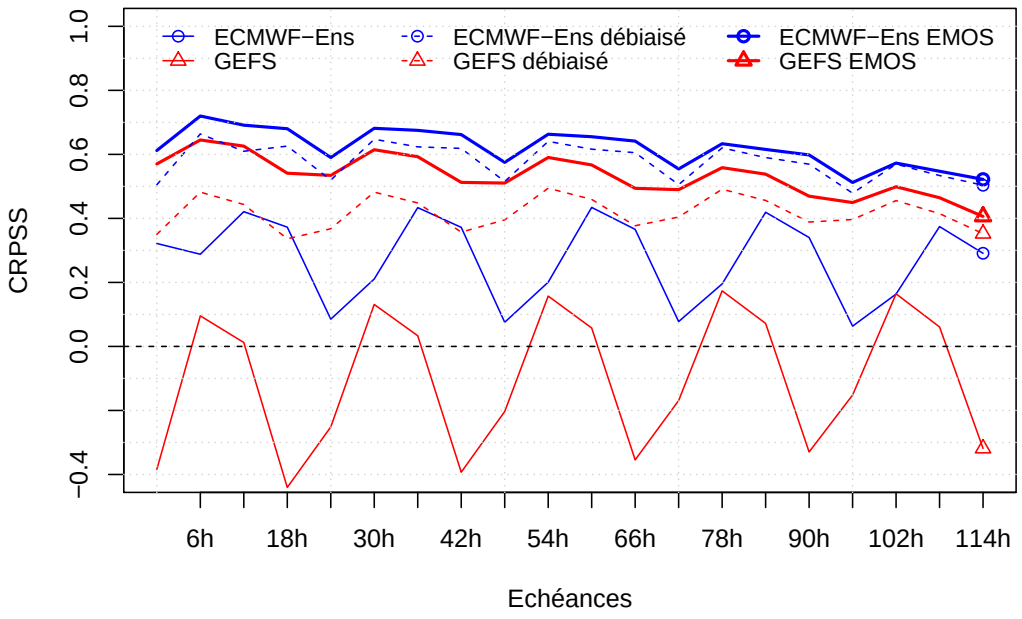

FIGURE 5.12 – Idem que Figure 5.11, mais avec les prévisions climatologiques comme prévisions de référence.

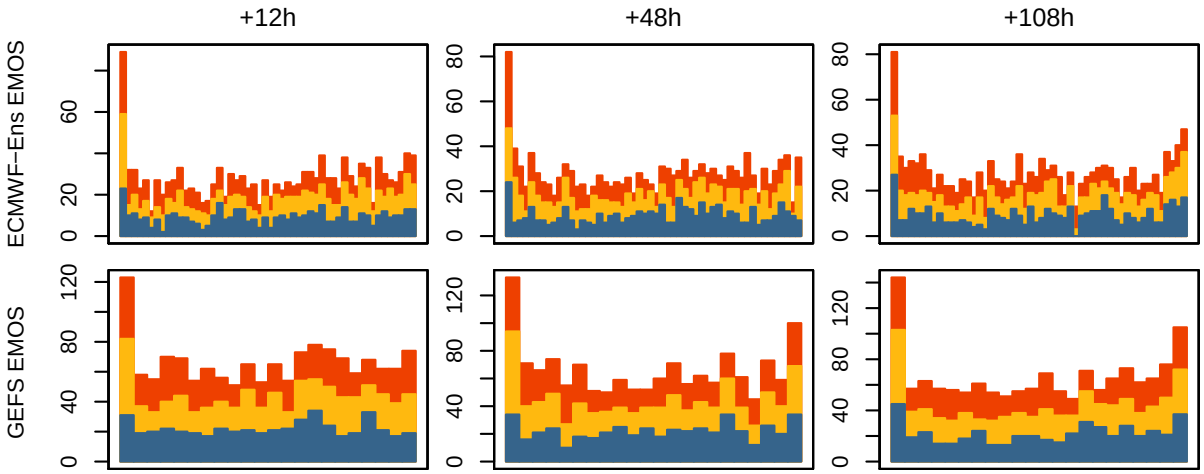

FIGURE 5.13 – Histogrammes de rang des prévisions pré-traitées de température, pour différentes échéances, sur le bassin de la Valserine. La même stratification qu'à la Figure 5.4 est appliquée.

5.4 Synthèse

L'objectif poursuivi dans ce chapitre était d'améliorer les forçages météorologiques, en corrigeant si besoin les défauts de fiabilité. Le diagnostic des performances des forçages bruts a confirmé le manque de fiabilité des différentes prévisions d'ensemble (ECMWF-Ens, GEFS et PEARP), défaut qui s'est exprimé au travers d'une forte sous-dispersion concernant à la fois la variable précipitation et température. En revanche, les prévisions par analogie, disponibles pour la variable précipitation seulement, ont fait preuve d'une fiabilité correcte.

Pour palier à ces défauts, nous avons testé une approche de correction statistique univariée basée sur la méthode EMOS. Cette méthode consiste à remplacer la distribution empirique de l'ensemble brut par une distribution continue dont les paramètres sont reliés à des statistiques de l'ensemble brut (moyenne, écart-type) via des modèles de régression.

Ce pré-traitement est parvenu à corriger de manière globalement satisfaisante la fiabilité des prévisions d'ensemble de précipitation et de température, tout en gardant une finesse suffisante pour gagner en qualité globale par rapport aux prévisions brutes. En revanche il s'est avéré inefficace pour améliorer les prévisions par analogie.

Seuls les forçages ECMWF-Ens et GEFS contiennent à la fois les prévisions de précipitation et de température qui sont nécessaires à la suite de notre travail. Les forçages ECMWF-Ens pré-traités se révélant globalement fiables et présentant de meilleures performances que GEFS pré-traitées, nous les utiliseront dans les chapitres 6, 7 et 8.

Arrivés ici, nous nous retrouvons avec des prévisions univariées certes de meilleure qualité, mais où la notion de « traces » (spatiale, temporelle et inter-variable) a disparu, ce qui se révèle être une barrière à la modélisation hydrologique. Le chapitre suivant doit nous permettre de reconstruire, à partir de ces prévisions météorologiques univariées, des prévisions multivariées cohérentes.

Chapitre 6

Reconstruction de forçages multivariés cohérents

Les méthodes paramétriques de pré-traitement telles que l'EMOS permettent de corriger les défauts de fiabilité des forçages météorologiques, mais entraînent la perte de la structure de dépendance. Elles produisent ainsi des prévisions univariées, qui de plus sont sous forme continue. Cependant, la modélisation hydrologique requiert des prévisions multivariées cohérentes, et ce sous forme ensembliste.

Il est nécessaire, pour passer de distributions prédictives continues à des prévisions ensemblistes, de réaliser un échantillonnage. Cette étape a déjà été entrevue au chapitre précédent, lorsqu'il a fallu évaluer à l'aide de l'histogramme de rang les prévisions pré-traitées via l'EMOS. Nous avons alors appliqué un échantillonnage dit « systématique », ou encore échantillonnage « quantile », qui consiste à générer des quantiles équidistants. C'est la méthode que l'on retrouve le plus fréquemment car les prévisions ensemblistes qui en découlent obtiennent les meilleures performances du point de vue univarié (Bröcker, 2012). Ainsi, nous assumons dans ce chapitre un échantillonnage quantile. Nous verrons cependant au chapitre 8 qu'il soulève certains problèmes lorsqu'il est appliqué à des variables très autocorrélées telles que le débit.

Si l'échantillonnage permet de répondre à la question du passage du continu au discret, il reste la question de la reconstruction de prévisions météorologiques multivariées cohérentes à partir de prévisions pré-traitées univariées. Cette étape, appelée « réarrangement » (en anglais : *reordering*), constitue la problématique de ce chapitre.

La section 6.1, qui est écrite sous la forme d'un article (en anglais) et qui constitue la majeur partie du chapitre, se concentre sur le réarrangement des prévisions de précipitation. Il est notamment proposé deux nouvelles méthodes qui font usage des analogues pour améliorer le conditionnement du réarrangement à la situation météorologique. Dans la section 6.2, nous discutons de l'extension du réarrangement à la cohérence inter-variable, c'est-à-dire à la cohérence entre précipitation et température. Cette aspect sera peu étudié, en raison de la qualité des données observées de température. Enfin, la section 6.3 fait la synthèse de ce chapitre.

6.1 Usage des analogues pour le réarrangement des prévisions de précipitation

Cette section correspond à un article publié dans la revue *Water Resources Research*, Vol. 53, no 11, p. 10085-10107.

Using Meteorological Analogues for Reordering Postprocessed Precipitation Ensembles in Hydrological Forecasting

J. Bellier[1], G. Bontron[2], and I. Zin[1]

[1] Université Grenoble Alpes, Grenoble INP, CNRS, IGE, Grenoble, France
[2] Compagnie Nationale du Rhône, Lyon, France

(Manuscript received 1 Jun. 2017, accepted 31 Oct 2017, published online 1 Dec. 2017)

DOI: 10.1002/2017WR021245

Résumé

Il est fréquent que les prévisions météorologiques d'ensemble présentent des défauts de fiabilité, auquel cas une correction statistique est nécessaire. Ces méthodes de correction sont essentiellement univariées, c'est-à-dire qu'elles s'appliquent indépendamment pour chaque combinaison de bassin, échéance et variable météorologique. Les ensembles corrigés doivent alors être réarrangés de manière à reconstruire une structure de dépendance multivariée qui soit cohérente. Le *Schaake shuffle* et l'*ensemble copula coupling* sont les deux méthodes les plus populaires dans la littérature. Dans cet article, nous nous intéressons à la reconstruction de la structure spatio-temporelle de prévisions de précipitation, et proposons deux variantes basées sur l'utilisation d'analogues météorologiques. Ces variantes, ainsi que les méthodes originales susmentionnées, sont évaluées sur des prévisions réelles issues de l'ECMWF, au travers d'une démarche de vérification comportant plusieurs étapes. Sont d'abord évaluées les prévisions multivariées de précipitation issues du réarrangement, puis les prévisions univariées de débits qui résultent de la modélisation hydrologique. Les résultats montrent que les méthodes se classent différemment selon si la vérification porte sur l'une ou l'autre variable. Le *Schaake shuffle* s'avère notamment peu performant dès lors que la vérification porte sur les débits. Au travers de cette étude, nous mettons en avant le rôle crucial que joue la structure spatio-temporelle des prévisions d'ensemble de précipitation dans la prévision hydrologique.

Abstract

Meteorological ensemble forecasts are nowadays widely used as input of hydrological models for probabilistic streamflow forecasting. These forcings are frequently biased and have to be statistically post-processed, using most of the time univariate techniques that apply independently to individual locations, lead times and weather variables. Post-processed ensemble forecasts therefore need to be reordered so as to reconstruct suitable multivariate dependence structures. The Schaake shuffle and ensemble copula coupling

are the two most popular methods for this purpose. This paper proposes two adaptations of them that make use of meteorological analogues for reconstructing spatio-temporal dependence structures of precipitation forecasts. Performances of the original and adapted techniques are compared through a multi-step verification experiment using real forecasts from the European Centre for Medium-Range Weather Forecasts. This experiment evaluates not only multivariate precipitation forecasts but also the corresponding streamflow forecasts that derive from hydrological modeling. Results show that the relative performances of the different reordering methods vary depending on the verification step. In particular, the standard Schaake shuffle is found to perform poorly when evaluated on streamflow. This emphasizes the crucial role of the precipitation spatio-temporal dependence structure in hydrological ensemble forecasting.

6.1.1 Introduction

Decision makers use hydrological forecasts in a wide range of activities such as flood prevention, operation of hydropower facilities and water supply. These are issued, in most operational settings, by forcing a hydrological model with meteorological forecasts. In a probabilistic framework, which has been shown to enable more rational decision making (Krzysztofowicz, 2001; Ramos *et al.*, 2013a), uncertainty in the forcing data is generally tackled using ensemble forecasts (Cloke et Pappenberger, 2009). These comprise multiple runs of a numerical weather prediction model with slightly altered initial conditions and sometimes model assumptions (Buizza *et al.*, 1999). In hydrological modeling, ensemble forecasts pertain to different locations, lead times and weather variables (typically precipitation and temperature) and are thus considered as multivariate ensembles, where each dimension corresponds to a specific combination of the three. Despite great progress over the last decades, ensemble forecasts for surface weather variables such as precipitation and temperature very often fail to represent a true estimate of the forecast uncertainty (Park *et al.*, 2008; Verkade *et al.*, 2013). They consequently have to be statistically post-processed [1], using techniques such as ensemble model output statistics (EMOS) (Gneiting *et al.*, 2005), Bayesian model averaging (BMA) (Raftery *et al.*, 2005), or adaptations thereof (e.g., Sloughter *et al.*, 2007; Scheuerer, 2014; Scheuerer et Hamill, 2015b). We refer to Li *et al.* (2017) for a recent review. These methods have been shown capable of yielding calibrated univariate distributions from which an ensemble of the desirable size can be drawn. However, most of the post-processing techniques are univariate, meaning that forecasts are corrected for biases along each dimension independently, while potentially strong multivariate dependencies are ignored. Work has been done to adapt EMOS or BMA techniques to account for multivariate dependencies within the post-processing itself (e.g., Berrocal *et al.*, 2007; Baran et Möller, 2015, 2017), but as they consider parametric multivariate aspects these adaptations have been so far restricted to low-dimensional settings. Hence, this paper concentrates on methods for reconstructing, independently from the univariate post-processed ensembles, a multivariate ensemble where the trajectories

1. Cet article concernant uniquement la correction des forçages météorologiques, il n'y a pas lieu de différencier le *pre-processing* (pré-traitement) du *post-processing* (post-traitement). Ainsi, nous utiliserons tout au long de l'article le terme *post-processing*, malgré le fait qu'il s'applique ici aux forçages météorologiques.

(i.e., members) are physically realistic.

For this purpose, non-parametric techniques have emerged where post-processed ensemble values within each dimension are permuted in a way that duplicates the rank dependence structure of a suitable dependence template (Wilks, 2015). Due to the process of giving ensemble values a certain rank, such methods are referred to as reordering methods, or empirical copula methods (Clark *et al.*, 2004; Schefzik *et al.*, 2013). While the permutation idea remains unchanged, the different methods distinguish themselves by the dataset used for specifying the dependence template. The Schaake shuffle, as the first approach proposed by Clark *et al.* (2004), uses historical (i.e., observed) trajectories to specify the template. Its main drawback however is the incapacity to condition the dependence structure on atmospheric states. For example, precipitation forecasts at two different locations may be strongly correlated in case of large-scale weather fronts, while being much less correlated in convective situations. Adaptations have recently been proposed to make the Schaake shuffle flow-dependent. In the approach suggested by Schefzik (2016a), historical trajectories that are selected for the dependence template start at past dates for which operational raw ensemble forecasts were similar to the one issued at the current forecast date. The downside though is that the observation dataset available for historical trajectories is restricted to the period where past forecasts are available. Rare hydrological events where dependence structures depart from climatology may thus suffer from an insufficient number of candidate historical trajectories. To relax this constraint, Scheuerer *et al.* (2017) propose a version where the selected historical trajectories have marginal distributions that are similar to those of the post-processed forecast at all locations and all lead times. However, high dimension contexts would require a very long observation dataset for a correct match along all dimensions. Additionally, a correct match regarding marginal distributions does not ensure that meteorological situations, and therefore dependence structures, are similar. In contrast to the Schaake shuffle and above-cited adaptations, the ensemble copula coupling (ECC) approach of Schefzik *et al.* (2013) uses the raw ensemble forecast to specify the template, thus capturing the atmospheric model flow dependence structure. Consequently, ECC performances are closely related to the ability of the model to correctly represent the covariability between the different dimensions. In particular, the covariability between locations may be missed due to a coarse spatial resolution of the model compared to the spatial scale of the forecasting problem at hand.

In the context of hydrological forecasting, meteorological forecasts may concern a large number of locations and lead times whose covariability is of great importance. This is particularly true for precipitation, as the weather variable to which hydrological models are, in many regions, most sensitive. Severe large-scale hydrological events such as floods indeed occur in case of concomitant events on several basins, due to remarkable spatio-temporal structures on precipitation. Besides, the fact that precipitation has a point mass at zero makes the reordering process very sensitive to the number of zeros present in the dependence template. Reordering techniques have been applied to precipitation forecasts in several past studies (e.g., Clark *et al.*, 2004; Schaake *et al.*, 2007; Wu *et al.*, 2011) and many authors have stressed its crucial role for hydrological forecasting (e.g., Voisin *et al.*,

2010; Verkade *et al.*, 2013; Demargne *et al.*, 2014). However, very few, with the exception of Scheuerer *et al.* (2017), have compared the performances of different reordering techniques per se by evaluating the resulting streamflow forecasts.

This paper has two objectives. First, we investigate the use of meteorological analogues for reordering post-processed precipitation ensemble forecasts, as a way to preferentially select past observations that have similar atmospheric states than the forecast ones (Obled *et al.*, 2002; Bontron, 2004; Marty *et al.*, 2012; Ben Daoud *et al.*, 2016). Two new methods are proposed, which make use of the analogues in different ways. Secondly, we compare these methods to existing ones through a multi-step verification experiment, thus quantifying the impact of the precipitation spatio-temporal dependence structure in a hydrological forecasting context. In particular, we investigate the extent to which different verification approaches, focusing either on precipitation or on the resulting streamflow, yield consistent results about the performance of the different reordering methods.

Section 6.1.2 presents a description of the data set that is used later in the case study and based on real forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). In section 6.1.3 we first review the theory underlying the reordering process and then describe two different techniques based on analogues. The first technique, as a variant of the Schaake shuffle, uses as spatio-temporal template historical trajectories that are temporally centered around analogue dates for a specific lead time. The second technique combines the Schaake shuffle with ECC using, on the one hand, independent spatial templates from analogue dates at each lead time and, on the other hand, the temporal template from the raw ensemble forecast. Results from the verification case study are presented and discussed in section 6.1.4. Section 6.1.5 concludes.

6.1.2 Study basins and data

6.1.2.1 Study basins, observational and reanalysis data sets

In this paper, the variable to which reordering applies is mean areal precipitation (MAP) accumulated at 6-hour time step over hydrological basins. Five basins located in the upper Rhone River region in France are considered: Valserine, Usses, Fier, Séran and Guiers. Figure 6.1 displays their location and a picture of the relief. Their geographic proximity within the northern part of the Prealps mountain range ensures a significant spatial correlation in precipitation, but situations where spatial patterns depart from climatology are not rare, especially in the summer period where local convective precipitation are more likely to occur. MAP observations over these basins were processed by Météo-France by kriging hourly and daily rain gauge data from the Météo-France network. This dataset covers the 1992-2014 period. Hourly streamflow records at each outlet, used for hydrological model calibration and initialization, were obtained from the Compagnie Nationale du Rhône (CNR)'s gauging stations. Records start between 1990 and 2004 depending on basins. Table 6.1 lists basin areas as well as means and 90-percentile of streamflow records. Finally, the archive of atmospheric predictors used for searching meteorological analogues comes from the ECMWF Reanalysis (ERA) -Interim dataset (Dee *et al.*, 2011) publicly available on the ECMWF data portal. These predictors are

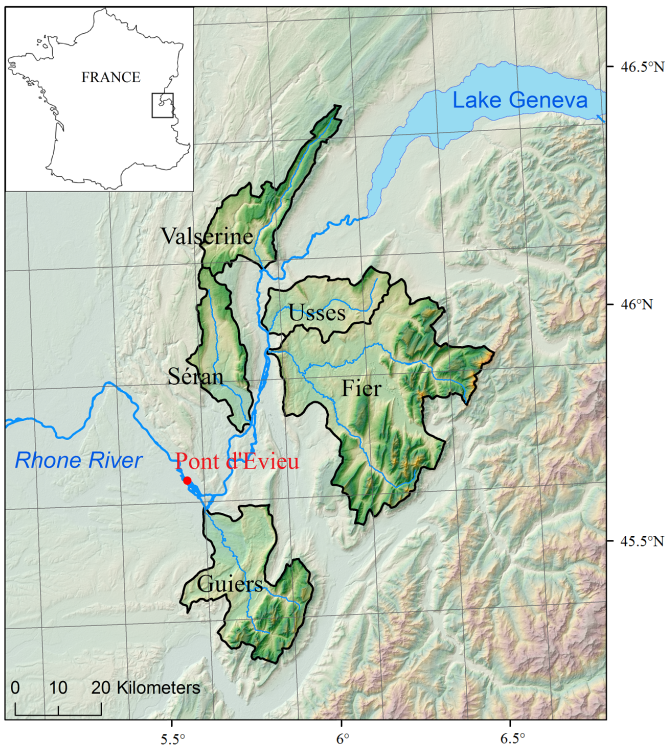

FIGURE 6.1 – Location of the study basins. The grid in grey represents the regular $0.25° \times 0.25°$ grid at which ECMWF ensemble precipitation forecasts are available in the TIGGE archive.

temperatures (T) at 850 and 500 hPa, geopotential heights (Z) at 1000 and 500 hPa , relative humidity (RH) at 850 hPa and total column water (TWC). Predictor fields were extracted on a regular $0.5° \times 0.5°$ grid (the native resolution is approximately 79 km), over the 1992-2014 period so as to match that of the MAP dataset.

6.1.2.2 Hydrological model

Lumped models of ARX type (auto-regressive model with exogeneous input ; Box *et al.* (2015)) are used for streamflow simulation at the outlet of the five basins. They are part of the integrated forecasting chain developed at CNR over the entire Rhone basin and operationally used for hydrological forecasting (Bompart *et al.*, 2009). The output variable at a given lead time k, here the streamflow Q^k, is calculated with an equation of

TABLE 6.1 – Characteristics of the five considered basins

Basin	Area (km^2)	Stream gauge	MQ ($m^3\ s^{-1}$)[a]	Q90 ($m^3\ s^{-1}$)[a]
Valserine	361	Lancrans	14	32
Usses	309	Pont Rouge	3	9
Fier	1375	Motz	33	80
Séran	290	Pont de la Thuilliere	6	14
Guiers	609	Belmont	17	36

[a]MQ (mean streamflow) and Q90 (90-percentile streamflow) are estimated over the 2004-2014 period.

the form:

$$Q^k = a + \sum_{i=1}^{B} b_i Q^{k-i} + \sum_{i=1}^{C} c_i \mathrm{MAP}^{k-i}, \tag{6.1}$$

where MAP is the exogenous input. Coefficients $(a, b_1, \ldots, b_B, c_1, \ldots, c_C)$ vary according to streamflow ranges, soil moisture conditions and seasons. Modeling is performed at hourly time step, under the assumption that precipitation are evenly distributed over the 6-hour accumulation periods. Hydraulic propagation along the Rhone River, from the outlet of each basin to the global outlet of Pont d'Evieu (see Figure 6.1), is carried out using propagation-time functions. Note that this study is conducted using lumped hydrological models but, as several basins are involved, it is not much different from using spatially distributed models that would consider grid-based precipitation forecasts as forcing instead of MAP forecasts. Indeed, the challenge of reconstructing coherent spatial dependence structures remains unchanged. An example of reordering that applies to grid-based forcings can be found in Vrac et Friederichs (2015), in the context of reanalysis bias-correction.

In the operational setting at CNR, a snow module uses temperature forecasts to adjust MAP values by eventually adding snow melt or withdrawing solid precipitation. However, the present study concentrates on precipitation, studying how their spatio-temporal dependence structure may affect streamflow forecasts. For this reason, hydrological modeling was conducted without the snow module. Since forecast streamflow are compared to simulated (i.e., output of the model with observed forcings) and not to observed streamflow, not taking into account temperature forecasts should not impact the conclusions of the study. We reserve for future works the assessment of the role played by the precipitation-temperature covariability in forcing data.

6.1.2.3 Forecast datasets

Precipitation forecasts come from the ECMWF ensemble prediction system (EPS) and have been downloaded over the 2007-2014 period from the TIGGE archive (Park *et al.*, 2008). They consist of the 51-member ensemble forecasts that were operationally issued during this period and stored on a regular $0.25° \times 0.25°$ grid. Forecasts starting at 00 UTC with lead times from 6 to 120 hours have been considered here. Note that forecasts before 2007 were not archived in TIGGE. The grid-based to MAP transformation has been conducted using Thiessen-based averaging (Tabios et Salas, 1985). Hereafter, the "raw forecast" mention refers to these MAP forecasts. Lastly, forecasts of the atmospheric predictors (the same as for ERA-Interim) needed for the search of meteorological analogues are here provided by the control (i.e., non-perturbed) run of the ECMWF ensemble forecasts. These forecasts were also extracted from TIGGE, while on a regular $0.5° \times 0.5°$ grid to match that of the reanalysis dataset. This resolution is a trade-off between the native resolution of the reanalysis dataset and the resolution at which forecasts are available in TIGGE.

6.1.2.4 Univariate post-processing and sampling

MAP forecasts were bias-corrected over the 2011-2014 period using a variant of the EMOS (Gneiting *et al.*, 2005) approach. In a nutshell, EMOS is a univariate post-processing technique that consists in fitting a regression model between training observations and statistics of the ensemble forecasts, yielding parametric forecast distributions. For this study we followed the approach proposed by Scheuerer *et al.* (2017), which is itself a simplified variant of the method proposed by Scheuerer et Hamill (2015b). Ensemble statistics are the ensemble mean, the ensemble mean difference (as a dispersion measure) and the probability of precipitation. Forecast distributions are censored, shifted Gamma distributions. We refer to the above-cited references for more details. As a notable difference though, ensemble statistics were directly computed from MAP forecasts without using neighbourhood information from grid-based forecasts. Moreover, we used a simple multiplicative, preliminary adjustment of the forecasts instead of the more complex quantile mapping approach suggested in section 4a of Scheuerer et Hamill (2015b). Regression models were fitted monthly, using running past 4 years of forecasts as training. The objective was to reproduce an operational setting where the only forecasts available for training are those prior to the current emission date. Due to the forecast archive starting in 2007, the length of 4 years has been chosen as a trade-off between the need of a training period long enough for an effective post-processing and a verification period long enough for results to be robust. The gain in forecast quality due to univariate post-processing was measured in terms of the continuous ranked probability skill score (CRPSS ; see 4.2.1.3) and the uniformity of rank histograms. The results are provided in Figures A.1-A.3 in the supporting information to this paper. It is found that post-processed forecast distributions are now fairly calibrated, while the skill has been improved up to 3 days.

From each marginal distribution representing a specific combination of basin and lead time, a sample was drawn to construct a univariate ensemble of the desired size M. We used the systematic sampling strategy by choosing equidistant quantiles with levels $\alpha_m = m/(M+1)$ for $m = 1, \ldots, M$. The ensemble size M after post-processing was chosen equal to the size of the raw ensemble, i.e., 51. The reason for this is twofold. First, the metrics used in the verification experiment are sensitive to the ensemble size, meaning that using a larger M may yield better results, due to the finer depiction of each marginal distribution. Consequently, for a fair comparison between raw and post-processed ensembles, M must be set to 51. The second reason, which becomes more clear in section 6.1.3, is that some reordering methods are based on the raw ensemble, meaning that they cannot apply to post-processed ensembles which size is not equal to that of the raw ensemble.

6.1.3 Reordering methods

6.1.3.1 Overview

After the univariate post-processing and sampling steps, the task remains to reconstruct adequate multivariate trajectories. For the formal description of the process, we adopt the formalism of Schefzik *et al.* (2013) and Schefzik (2016a). Let $j \in \{1, \ldots, J\}$ be

a basin (or more generally, a location) and $k \in \{1, \dots, K\}$ a lead time. Only one weather variable, MAP, is considered here. The dimension of the multivariate setting is then $L = J \times K$. Adjectives univariate, J-variate, K-variate and L-variate qualify trajectories (or ensembles of trajectories) that contain data for one basin at one lead time, all basins at one lead time, one basin at all lead times and all basins at all lead times, respectively. Within ensembles, regular subscripts denote member labelling while parenthetical subscripts indicate sorted data.

Let the M-member raw ensemble forecast, with each member comprising a L-variate trajectory, be denoted by

$$\mathbf{x} = (\mathbf{x}^{1,1}, \dots, \mathbf{x}^{J,K}) = ((x_1^{1,1}, \dots, x_M^{1,1}), \dots, (x_1^{J,K}, \dots, x_M^{J,K})), \tag{6.2}$$

where j and k increment independently. The post-processing and sampling steps yield univariate post-processed ensembles $\tilde{x}_1^{j,k}, \dots, \tilde{x}_M^{j,k}$ for each dimension (j, k). Let us now define a template

$$\mathbf{z} = (\mathbf{z}^{1,1}, \dots, \mathbf{z}^{J,K}) = ((z_1^{1,1}, \dots, z_M^{1,1}), \dots, (z_1^{J,K}, \dots, z_M^{J,K})) \tag{6.3}$$

as a dataset comprising M physically realistic L-variate trajectories. Once $\mathbf{z}$ is defined, the reordering process consists in making, for each dimension (j, k), permutations between the components $\tilde{x}_1^{j,k}, \dots, \tilde{x}_M^{j,k}$ such that the rank dependence structure matches that of $\mathbf{z}^{j,k} = (z_1^{j,k}, \dots, z_M^{j,k})$. Practically, it proceeds as follows, as formalized by Schefzik (2016a):

1. For each dimension (j, k), the univariate components of $\mathbf{z}^{j,k} = (z_1^{j,k}, \dots, z_M^{j,k})$ are ranked such that $z_{(1)}^{j,k} \leq \dots \leq z_{(M)}^{j,k}$, and the permutations

$$\pi_{j,k}(m) = \text{rank}(z_m^{j,k}) \tag{6.4}$$

   are derived for $m \in (1, \dots, M)$, with ties resolved as random.

2. For each dimension (j, k), the univariate components $\tilde{x}_1^{j,k}, \dots, \tilde{x}_M^{j,k}$ are ranked such that $\tilde{x}_{(1)}^{j,k} \leq \dots \leq \tilde{x}_{(M)}^{j,k}$. Using the permutations $\pi_{j,k}$ from step 1, the reordered post-processed ensemble $\tilde{\mathbf{x}}^{j,k}$ is obtained by

$$\tilde{\mathbf{x}}^{j,k} = \left(\tilde{x}_{(\pi_{j,k}(1))}^{j,k}, \dots, \tilde{x}_{(\pi_{j,k}(M))}^{j,k}\right). \tag{6.5}$$

3. The L reordered ensembles $\tilde{\mathbf{x}}^{j,k}$ are aggregated so as to generate the L-variate post-processed ensemble forecast

$$\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}^{1,1}, \dots, \tilde{\mathbf{x}}^{J,K}). \tag{6.6}$$

We refer to Clark *et al.* (2004) and Wilks (2015) for concrete numerical examples of the process, or to Schefzik *et al.* (2013) for a mathematical interpretation under the view of empirical copulas.

6.1.3.2 Existing methods

We remind that the different methods distinguish themselves by the dataset used for specifying the template $\mathbf{z}$. Three existing methods, namely the Schaake shuffle, ECC and the minimum divergence Schaake shuffle, are described here and considered in the verification experiment. In addition to the formal description that follows, Figure 6.2 illustrates these methods in a two-dimensional setting. We here let j correspond either to the Fier (horizontal axis) or the Valserine (vertical axis) basin, while k refers invariably to the +72h lead time. The forecast to be reordered has been issued on 30 January 2013. The left panels show scatter plots of the template $\mathbf{z}$, from which the rank dependence structure, represented on the middle panels in the form of Latin squares, are derived. Scatter plots of the post-processed forecast $\tilde{\mathbf{x}}$ are finally displayed on the right panels. We observe on these plots that the individual member values on the vertical axis have been associated with those on the horizontal axis in a way that replicates the rank dependence structure of the template. In margins of the scatter plots of $\mathbf{z}$ and $\tilde{\mathbf{x}}$ are also drawn the marginal empirical cumulative distribution functions (CDF).

In the ECC approach (Schefzik *et al.*, 2013), the template is the raw ensemble, that is $\mathbf{z} = \mathbf{x}$. The post-processed forecast therefore inherits the dependence structure predicted by the model, assuming that it is capable of representing spatio-temporal patterns adequately. As $\mathbf{z} = \mathbf{x}$, ECC is restricted to post-processed ensembles which size M equals that of the raw ensemble. Schefzik *et al.* (2013) distinguish between ECC-Q, ECC-R and ECC-T, the last letter referring to a specific sampling scheme. In this study though, we consider sampling as part of the post-processing step, and not as part of the reordering step. Given that we use an equidistant quantiles sampling scheme (see section 6.1.2.4), the method referred hereafter to as ECC corresponds to the ECC-Q method in Schefzik *et al.* (2013).

In the Schaake shuffle approach (Clark *et al.*, 2004), the dependence structure is provided by M historical trajectories randomly selected within the observation dataset. Formally, let t_0 be the initialization date of the ensemble forecast and $t_1, \ldots, t_M$ be M dates in the past for which J-variate observations are available. These observations are denoted by $y^{j,t_1}, \ldots, y^{j,t_M}$ for $j \in \{1, \ldots, J\}$. The template is constructed by taking the observations that followed over time, defining $\mathbf{z}^{j,k}$ for each dimension (j, k) as

$$\mathbf{z}^{j,k} = \left(y^{j,t_1+k}, \ldots, y^{j,t_M+k}\right), \tag{6.7}$$

where y^{j,t_m+k} denotes the observation for basin j at the date of k lead times after t_m, for $m \in \{1, \ldots, M\}$. For instance, considering 6-hour accumulation periods, $t_m + 5$ with $t_m =$ 1997-04-22 00:00 corresponds to the date of 1997-04-23 06:00. By aggregating, one can finally obtain the complete L-variate template

$$\mathbf{z} = (\mathbf{z}^{1,1}, \ldots, \mathbf{z}^{J,K}). \tag{6.8}$$

In the original implementation of the Schaake shuffle in Clark *et al.* (2004), the dates $t_1, \ldots, t_M$ are randomly selected from past years in the historical record so as to lie within 7 days before and after the calendar day of t_0. However, in the present case study we have

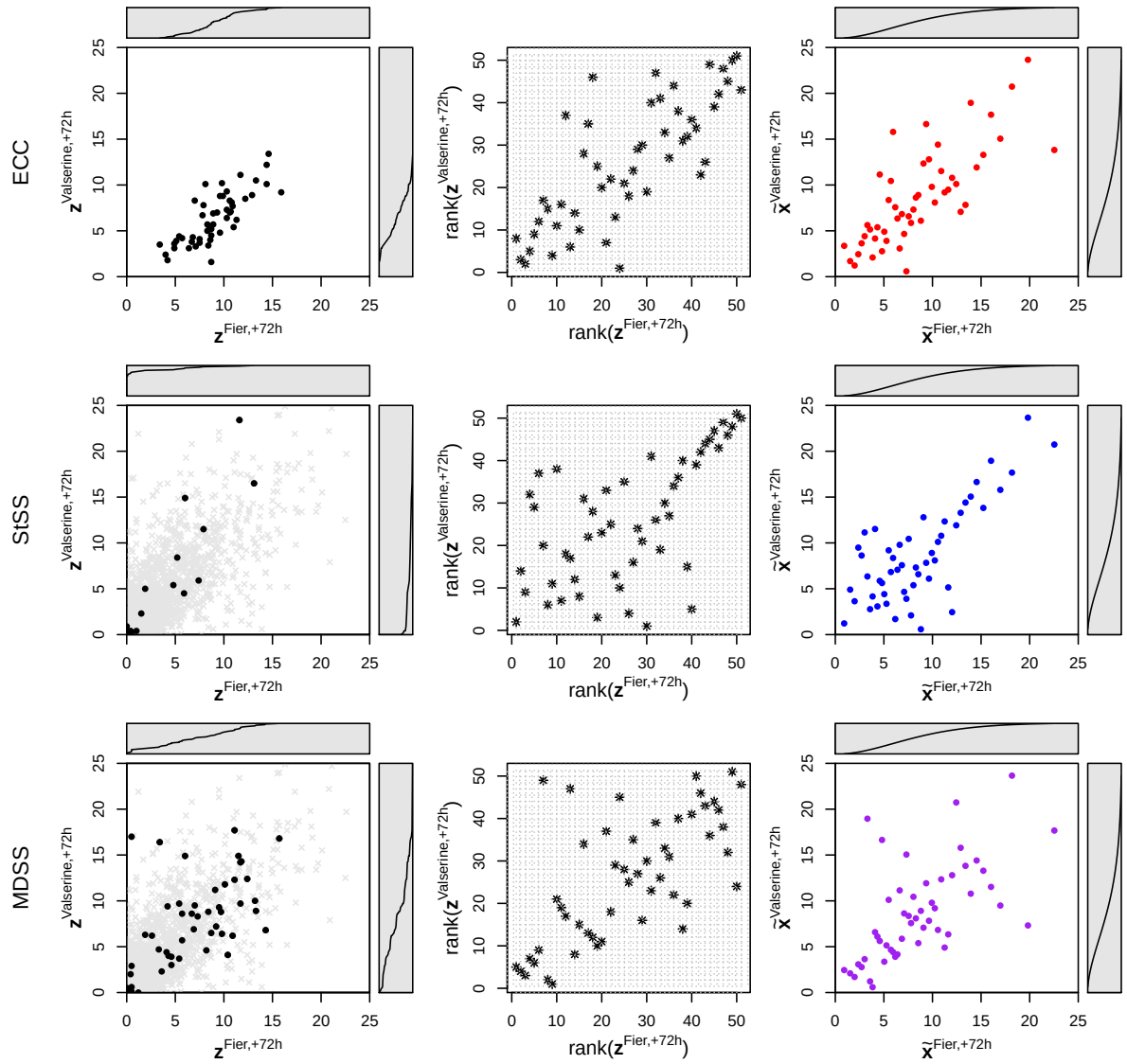

FIGURE 6.2 – Illustration in a two-dimensional setting of the ECC (top row), StSS (middle) and MDSS (bottom) reordering methods, based on the MAP forecast issued on 30 January 2013. Ensembles comprise 51 members. (left column) Scatter plots of the template $\mathbf{z}$, where the $M = 51$ trajectories are indicated by black dots. For StSS and MDSS, trajectories potentially available for selection are represented by grey crosses. (middle) Latin squares representing the rank dependence structure of the template $\mathbf{z}$. (right) Scatter plots of the reordered post-processed forecast $\tilde{\mathbf{x}}$. Marginal empirical CDFs are plotted in shaded areas. MAP is in mm 6h^{-1}.

considered this 14-day window as too restrictive regarding the annual climatology of the atmospheric conditions responsible for precipitation. In the approach referring hereafter to as the standard Schaake shuffle (StSS), this constraint is relaxed by using a 3-month window centered on the calendar day of t_0. Either way, the core principle of the Schaake shuffle applies, that is, the post-processed forecast inherits the spatio-temporal dependence structure of randomly selected observations from the climatology. The 1992-2014 period where observations are available is here considered. The middle-left panel of Figure 6.2 shows, for the forecast of 30 January 2013, the 3-month-window filtered set of historical observations, indicated by grey crosses, among which only a subset, indicated by black dots, has been randomly selected so as to set up the template $\mathbf{z}$. As it is entirely based on observations, StSS theoretically allows arbitrarily large post-processed ensemble sizes M, although it is not tested in this study. Large ensembles are of interest for end-users focusing on extremes, as it is the only way of accessing the tails of the distributions during the sampling step that follows post-processing. The main caveat of StSS, however, is that it assumes the independence of the spatio-temporal dependence structure to the ensemble forecast. The marginal CDFs of $\mathbf{z}$ have consequently no reasons to be similar to those of $\tilde{\mathbf{x}}$. In particular, $\mathbf{z}$ is likely to contain a large number of zero precipitation values, which are then ranked randomly. This issue will be further discussed in section 6.1.4.3.

Scheuerer *et al.* (2017) have proposed an adaptation referred to as the minimum divergence Schaake shuffle (MDSS), which replaces the random selection of historical trajectories by a more sophisticated procedure. Let $F_z^{j,k}$ and $F_{\tilde{x}}^{j,k}$ denote, for a given dimension (j,k), the empirical CDF of $\mathbf{z}^{j,k}$ and $\tilde{\mathbf{x}}^{j,k}$, respectively. The similarity between these two CDFs can be quantified by studying the divergence

$$\Delta_z^{j,k} = \int_{-\infty}^{+\infty} \left(F_z^{j,k}(u) - F_{\tilde{x}}^{j,k}(u) \right)^2 \mathrm{d}u. \tag{6.9}$$

The procedure of MDSS consists in choosing the set $\mathbf{z}$ of historical trajectories that minimizes the total divergence $\Delta_z^{tot} = \sum_{j,k} \Delta_z^{j,k}$. We refer to Scheuerer *et al.* (2017) for more details about the algorithms implemented for computing and minimizing the quantity Δ_z^{tot}. Note that in the present study, the same 3-month-window filtering as in StSS is applied before the minimum divergence algorithm starts. The bottom row of Figure 6.2 illustrates how MDSS manages to construct a template $\mathbf{z}$ that has marginal CDFs similar to those of $\tilde{\mathbf{x}}$, even though the minimization procedure was applied on the total divergence Δ_z^{tot} over all dimensions (j,k), and not simply on the two divergence $\Delta_z^{Fier,+72h}$ and $\Delta_z^{Valserine,+72h}$. The method would however reach its limits in cases where the number of dimensions is high (i.e., large number of basins and/or lead times), and/or the pool of available historical trajectories is limited (i.e., short observation dataset).

6.1.3.3 Preferential selection of past dates using an analogue method

As mentioned earlier, the StSS approach assumes the independence of the multivariate dependence structure to the ensemble forecast to be reordered. The MDSS technique addresses this issue from a statistical perspective, by selecting the subset $\mathbf{z}$ of observations that forms the desirable marginal distributions. In this paper, we address the selection

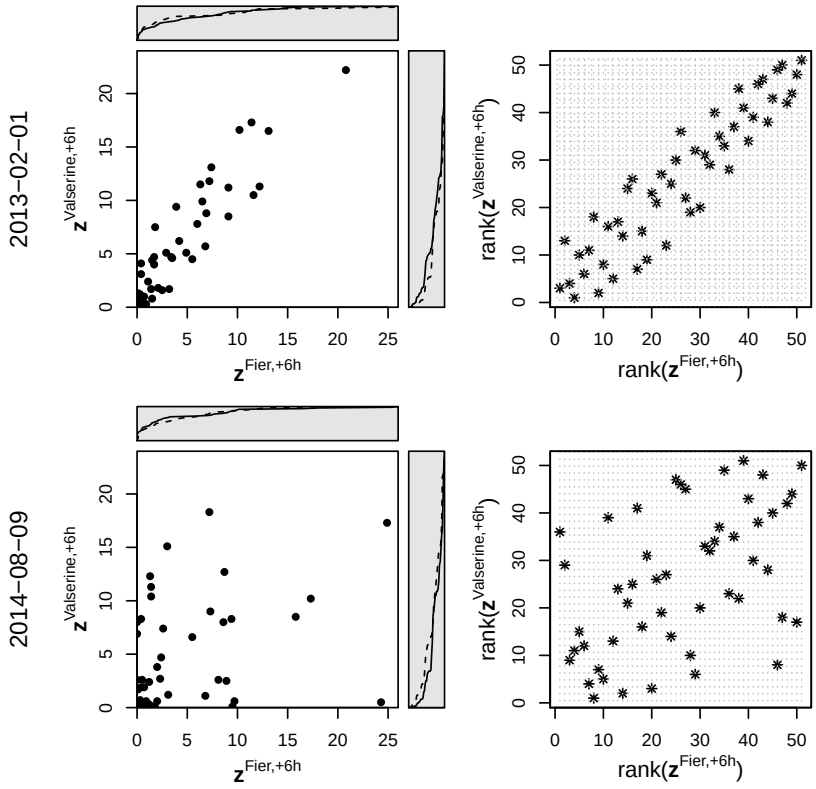

FIGURE 6.3 – (left) Two-dimensional templates made of meteorological analogues of two different forecast dates. (right) Corresponding rank dependence structures.

procedure from a meteorological perspective, by using an analogue method to preferentially select observations according to atmospheric states similarities. This is motivated by the fact that similar precipitation amounts may result from situations showing different atmospheric patterns, and thereby different spatio-temporal dependence structures. As an example, Figure 6.3 depicts two-dimensional spatial templates that are made of meteorological analogues obtained for two different forecast dates, using the method we describe below. Despite the similarity of the marginal distributions, the first set of analogues suggests a strong spatial correlation, as frontal precipitation systems usually lead to, while the second set shows a low correlation, as it occurs in some convective situations.

The basic idea of analogue techniques is that analogue synoptic (i.e., large-scale) situations should lead to similar local effects, the term "analogue" standing for states of the atmosphere that resemble each other closely (Lorenz, 1969). The general procedure consists in characterizing the current forecast situation over the area encompassing all considered basins, then searching analogue situations (simply referred to as analogues hereafter) among an archive of candidate situations, and finally selecting the observations corresponding to the analogues. Two archives (over the same period) are therefore needed: one for candidate synoptic situations and another for local observations. A possible approach, followed among others by Hamill et Whitaker (2006) for post-processing ensembles, Delle Monache *et al.* (2013) for constructing ensembles or Schefzik (2016a) for reordering post-processed ensembles, consists in searching analogues among the archive of past forecasts. It thereby forces analogue dates to lie within the period in which

Table 6.2 – Setting of the Analogue method

Level	Predictors[a]	Criterion[b]	Spatial domains (lat × lon)[d]	Number of analogues
L0	T850 (0h) T500 (6h)	RMSE	2° × 2°	7600
L1	Z500 (0h) Z1000 (6h)	S1 score[c]	10° × 20°	200
L2	RH850 (0h) × TCW (0h) RH850 (6h) × TCW (6h)	RMSE	2° × 2°	$51 = M$

[a]Two predictor fields for each level. (0h) refers to fields taken at the beginning of the accumulation period, (6h) at the end.
[b]Averaged over the two predictor criterion values.
[c]Score that compares gradients between grid points (Teweles et Wobus, 1954).
[d]Correspond to the domain over which criteria are computed. These are centered on the basin's zone.

past forecasts are available. We have chosen an alternative strategy that considers the archive of ERA-Interim meteorological reanalysis for candidate situations, thus leveraging the 1992-2014 period where both meteorological reanalyses and precipitation observations are available. We largely follow the method developed successively by Obled *et al.* (2002), Bontron (2004), Marty *et al.* (2012) and Ben Daoud *et al.* (2016), with the notable difference though that observations corresponding to analogues are not used to estimate a forecast probability distribution but instead used as template for the reordering process.

This paragraph aims at summarizing the most important features of the analogue method we have used, but as it is not the core of our paper we refer to the above-cited references for more details. The forecast atmospheric situation for a specific 6-hour accumulation period is provided by the control run. The selection of analogues, which are common to all basins, follows a 3-level procedure: a pre-selection, based on temperature fields, aims at constraining analogues to respect the seasonality (level L0). The level L1, as the most discriminant one, selects the analogues with most similar atmospheric circulation patterns, according to geopotential height fields. Synoptic circulation indeed plays a major role in the generation of precipitation. Nonetheless, additional variables at a more local scale, and representative of the physical processes responsible for precipitation, can efficiently refine the characterization of the atmospheric situation. A last selection level thus retains the M analogues with most similar air mass humidity patterns (L2). Table 6.2 provides details about predictors, analogy criteria, spatial domains and numbers of selected analogues relative to each selection level.

As an important point to remind, this analogue method provides M dates, corresponding to analogues, for each lead time independently. We denote these dates by $t_1^k, \ldots, t_M^k$ for $k \in \{1, \ldots, K\}$. Note that, for each lead time k, the labelling $m \in \{1, \ldots, M\}$ comes randomly and does not reflect in any way the strength of the analogy. The two reordering methods suggested in this paper, to which the next subsections correspond, make use of these analogue dates in different ways.

6.1.3.4 The analogue Schaake shuffle with center of gravity (AnSS-G)

The first method is a direct adaptation of StSS that, similarly to MDSS, conditions historical trajectories on the forecast case. The difference with MDSS relies on the conditioning on atmospheric states instead of on marginal distributions, by using analogue dates at a single specific forecast lead time. An appropriate choice of this lead time, which is not fixed but differs for each forecast, is essential. The idea is to search for the lead time that represents the center of gravity of the post-processed forecast regarding MAP amounts. For this purpose, we first define a spatial averaging function

$$S : (\mathbb{R}_0^+)^J \to \mathbb{R}_0^+ \ , \ (u_m^{1,k}, \ldots, u_m^{J,k}) \mapsto s_{um}^{\ k} = \omega_1 u_m^{1,k} + \ldots + \omega_J u_m^{J,k}, \tag{6.10}$$

where $k \in \{1, \ldots, K\}$ and $m \in \{1, \ldots, M\}$, which transforms MAP values of J nearby basins to a single averaged value. The variable u in equation (6.10) represents any quantity to which S potentially applies, that is x or $\tilde{x}$. Coefficients ω_j for $j \in \{1, \ldots, J\}$ are non-negative weights such that $\sum_{j=1}^{J} \omega_j = 1$. We suggest, in the view of hydrological forecasting, taking them as the relative basin areas. Indeed, MAP amounts being equal, one may want to assign in the reordering process more influence to large basins than to small basins, since their streamflow contribution at the global outlet are expected to be more important. Let further $\overline{s_u{}^k}$ be ensemble averages defined by $\overline{s_u{}^k} = \frac{1}{M} \sum_{m=1}^{M} s_{um}^{\ k}$ for $k \in \{1, \ldots, K\}$.

We can now define the center of gravity of the MAP forecast, denoted by k_g, as the lead time that divides in two equal parts the total cumulated sum of predicted precipitation $\sum_{k=1}^{K} \overline{s_{\tilde{x}}{}^k}$. Practically, this corresponds to searching for the lead time k that minimizes the function

$$G(k) = \left| \sum_{k'=1}^{k} \overline{s_{\tilde{x}}{}^{k'}} - \frac{1}{2} \sum_{k'=1}^{K} \overline{s_{\tilde{x}}{}^{k'}} \right|. \tag{6.11}$$

Using the so-obtained k_g, we can finally define $\mathbf{z}^{j,k}$, for each dimension (j, k), as

$$\mathbf{z}^{j,k} = \left(y^{j,t_1^{k_g}+(k-k_g)}, \ldots, y^{j,t_M^{k_g}+(k-k_g)} \right) \tag{6.12}$$

and aggregate them so as to generate the L-variate template

$$\mathbf{z} = (\mathbf{z}^{1,1}, \ldots, \mathbf{z}^{J,K}). \tag{6.13}$$

We refer to this method as the analogue Schaake shuffle with center of gravity (AnSS-G). As in the StSS approach, the AnSS-G method provides physically realistic spatial dependence structures since the dates for each lead time are common to the different basins. Temporal coherence is also ensured, as these dates form temporal sequences from a lead time to another. The improvement over StSS stems from the conditioning of these temporal sequences on the forecast atmospheric situation. Figure 6.4 shows the result of the AnSS-G reordering for the same forecast example as in Figure 6.2. It is not shown here that the center of gravity corresponds to the +66h lead time, while forecasts on both the horizontal and vertical axes concern the +72h lead time. In spite of this, the template

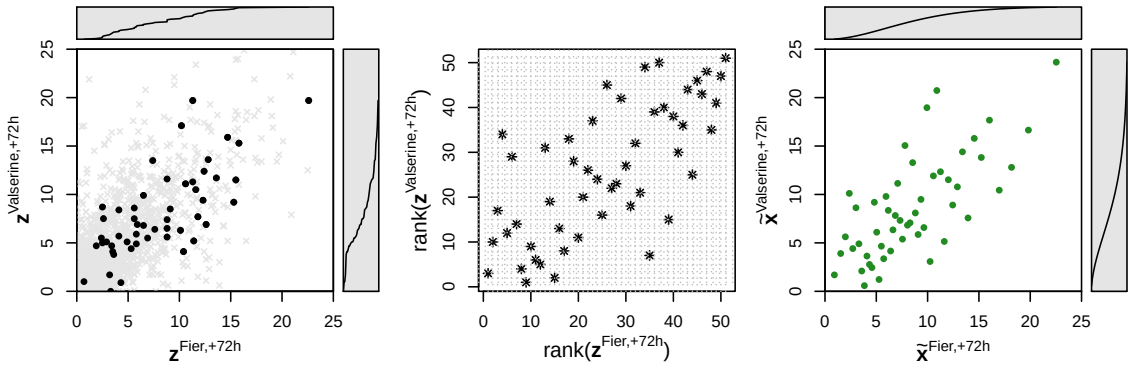

FIGURE 6.4 – Same as Figure 6.2 for the AnSS-G method.

$\mathbf{z}$ for the +72h lead time largely deviates from the climatology, meaning that it has been considerably conditioned by the forecast situation characterized at +66h.

Similarly to StSS and MDSS, the AnSS-G technique allows arbitrarily large post-processed ensemble sizes. It however presents some limitations. First, the dependence structures are less analogue to the forecast situation as the distance to the center of gravity increases. The improvement over StSS should thus be significant for short-term forecasts, where $|k - k_g|$ is for any $k \in \{1, \ldots, K\}$ limited to a couple of days at most. For longer-lead forecasts though, $|k - k_g|$ is for some $k \in \{1, \ldots, K\}$ likely to exceed the mean duration of the weather patterns characterized by the analogues, meaning that dependence structures are no more conditioned by the forecast. Secondly, AnSS-G uses analogues dates that are common to the J basins. If these are spread over a large geographical area, the analogue method may fail to find good analogues, as the spatial domains over which analogy criteria are computed must be large accordingly. To summarize, the performances of AnSS-G are expected to tend to those of StSS as K increases and/or the geographical distances between the basins increase.

6.1.3.5 The approach combining analogue Schaake shuffle and ECC (AnSS-ECC)

This paper proposes an alternative approach that combines the Schaake shuffle, constrained by analogues, with ECC. We hereafter refer to this method as AnSS-ECC. The basic idea is twofold: on the one hand, spatial dependence structures from observations taken at analogue dates can address the ECC caveat of performing poorly when the spatial resolution of the raw ensemble is coarse regarding basin sizes. On the other hand, the raw ensemble forecast is assumed to provide a coherent temporal dependence structure along all lead times. Formally, the method proceeds as follows.

1. The template $\mathbf{z} = (\mathbf{z}^{1,1}, \ldots, \mathbf{z}^{J,K})$ is constructed from the J-variate observations corresponding to the analogue dates for each lead time independently, that is

$$\mathbf{z}^{j,k} = (y^{j,t_1^k}, \ldots, y^{j,t_M^k}) \tag{6.14}$$

for all dimensions (j, k). We recall that the dates $t_1^k, \ldots, t_M^k$ for each lead time

$k \in \{1, \ldots, K\}$ are common to the J basins, and specified by the analogue method described in section 6.1.3.3.

2. We proceed to the general reordering process following steps 1-3 in section 6.1.3.1.

At this point, the so-obtained reordered post-processed forecast, that we denote by $\widehat{\mathbf{x}}$, has coherent spatial dependence structures for each lead time individually, but temporal ones are not accounted for since the analogue dates for the different lead times do not form temporal sequences. Let us now consider a specific way of partitioning the L-variate ensembles $\mathbf{x}$ and $\widehat{\mathbf{x}}$. One can write $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^K)$ where $\mathbf{x}^k = (\mathbf{x}_1^k, \ldots, \mathbf{x}_M^k)$ for $k \in \{1, \ldots, K\}$ are J-variate ensembles made of J-variate trajectories $\mathbf{x}_m^k = (x_m^{1,k}, \ldots, x_m^{J,k})$ for $m \in \{1, \ldots, M\}$. Same partitioning applies for $\widehat{\mathbf{x}}$. The second part of AnSS-ECC consists in making permutations within elements of $\widehat{\mathbf{x}}^k$ so as to duplicate the temporal dependence structure of the spatial averages of the raw ensemble $\mathbf{x}$, leaving spatial structures unchanged. This is described by steps 3-5 below, which are very similar to steps 1-3 in section 6.1.3.1, although ranking J-variate quantities instead of univariate quantities:

3. For each lead time k, the J-variate components of $\mathbf{x}^k = (\mathbf{x}_1^k, \ldots, \mathbf{x}_M^k)$ are assigned ranks such that $s_{x_{(1)}}^{\ k} \leq \ldots \leq s_{x_{(M)}}^{\ k}$, where $s_{x_m}^{\ k} = S(\mathbf{x}_m^k)$ for $m \in \{1, \ldots, M\}$ are spatial averages computed using the function S defined in equation (6.10). The permutations
$$\pi'_k(m) = \text{rank}(s_{x_m}^{\ k}) \tag{6.15}$$
for $m \in \{1, \ldots, M\}$ are then derived, with ties resolved as random.

4. For each lead time k, the J-variate components of $\widehat{\mathbf{x}}^k = (\widehat{\mathbf{x}}_1^k, \ldots, \widehat{\mathbf{x}}_M^k)$ are assigned ranks such that $s_{\widehat{x}_{(1)}}^{\ k} \leq \ldots \leq s_{\widehat{x}_{(M)}}^{\ k}$ where likewise $s_{\widehat{x}_m}^{\ k} = S(\widehat{\mathbf{x}}_m^k)$ for $m \in \{1, \ldots, M\}$. Using the permutations π'_k from step 3, reordering yields
$$\tilde{\mathbf{x}}^k = \left(\widehat{\mathbf{x}}^k_{(\pi'_k(1))}, \ldots, \widehat{\mathbf{x}}^k_{(\pi'_k(M))}\right). \tag{6.16}$$

5. The K reordered ensembles $\tilde{\mathbf{x}}^k$ are aggregated so as to generate the final L-variate post-processed ensemble forecast
$$\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^K). \tag{6.17}$$

The steps 1-2 (spatial reordering) and 3-5 (temporal reordering) above are illustrated in the top and bottom rows of Figure 6.5, respectively, using the example forecast of 30 January 2013. The spatial part is illustrated for the +66h lead time, while the temporal part considers the +60h and +66h lead times. Note that in the temporal part the basin indices have been dropped, as the reordering concerns spatial averages.

For the specific case in which $J = 1$, AnSS-ECC reduces to the standard ECC. Otherwise, the method uses the concept of multivariate ranking: in equation (6.15), J-variate components $\mathbf{x}_1^k, \ldots, \mathbf{x}_M^k$ are assigned ranks, via the computation of spatial averages $s_{x_1}^{\ k}, \ldots, s_{x_M}^{\ k}$ to which ordinary ranking can then be applied. Note that the parenthetical

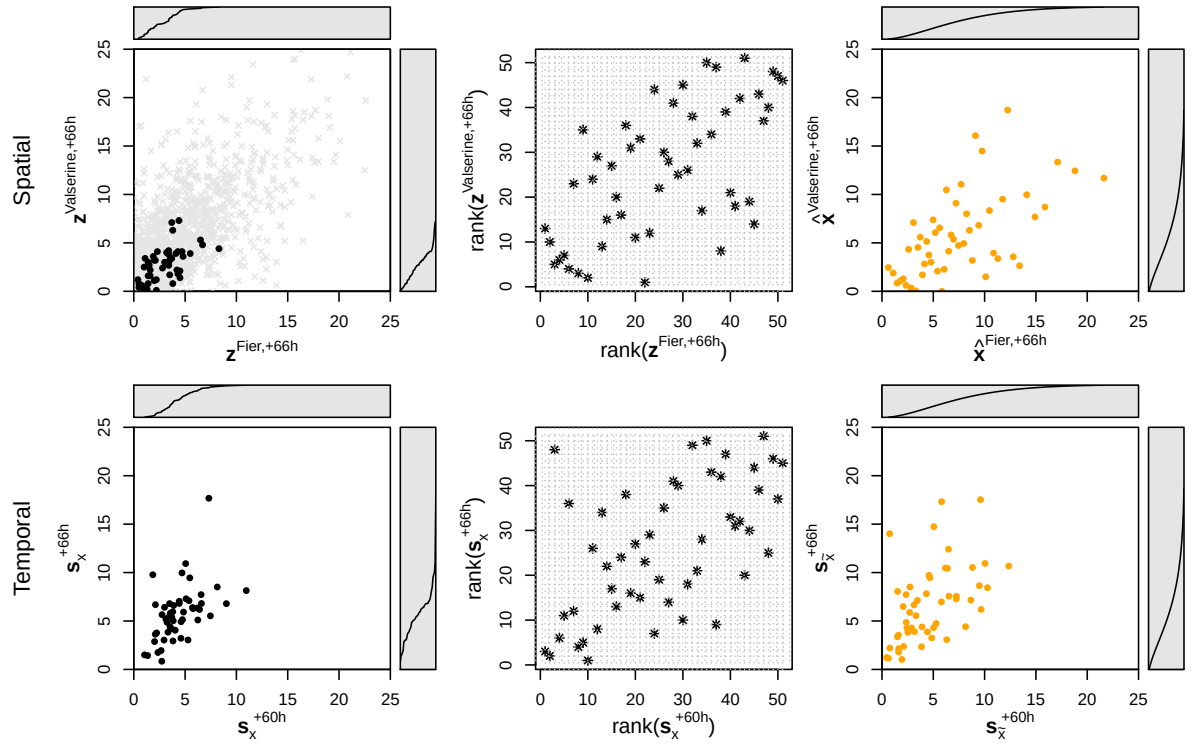

FIGURE 6.5 – Same as Figure 6.2 for the AnSS-ECC method, although spatial and temporal reordering operate in different ways. (top) Illustration of the spatial reordering, which applies to MAP ensembles at the same lead time, here +66h. (bottom) Illustration of the temporal reordering, which applies to MAP spatial average ensembles at two different lead times, here +60 and +66h.

subscript notation in equation (6.16) indicates sorted data according to this multivariate ranking. The idea of multivariate ranking has also been tackled by Schefzik (2016b), who suggests alternatives to spatial averages for multivariate ranking: the multivariate pre-rank, the average pre-rank or the signed Euclidean norm. Band-depth pre-ranks (Thorarinsdottir *et al.*, 2016) or minimum spanning tree pre-ranks (Wilks, 2004) are also alternative possibilities. We refer to the corresponding articles for more details.

To conclude, the AnSS-ECC technique can be seen as a form of downscaling of the spatial dependence structure provided by the raw ensemble. It is based on the assumption that this raw ensemble cannot represent correctly the spatial covariability, hence the method goes back to atmospheric state information predicted by the model to derive finer spatial dependence structures, using the analogues. The temporal dependence structure of the raw ensemble is however preserved, assuming that it is less impacted by the spatial model resolution. In this paper, we have considered that all basins are sufficiently nearby to allow the spatial reordering to use the same analogue dates. In the case where a large number of basins would cover a vast geographic area, AnSS-ECC can straightforwardly be extended so as to consider sub-groups on which the spatial reordering is applied independently, using different sets of analogue dates. A limitation of AnSS-ECC though, which is common to ECC, is that the size of the post-processed ensemble must equal that of the raw ensemble.

6.1.4 Verification results and discussion

We now evaluate the AnSS-G and AnSS-ECC reordering techniques, as well as ECC, StSS and MDSS for comparison, on the case study described in section 6.1.2 with post-processed ECMWF forecasts. As a naive benchmark, we also include in the evaluation the approach referred to as Random, which assumes independence between all dimensions and thereby performs the reordering in a random way. It is worth reminding that the post-processed forecasts reordered using the aforementioned methods share the same univariate ensemble values but have a different multivariate dependence structure, while forecasts referred to as Raw retain the univariate ensemble values before post-processing. Thus, one can compare the performances of ECC, StSS, MDSS, AnSS-G and AnSS-ECC to those of Random for evaluating the effect of reordering, or to those of Raw for evaluating the effect of the complete post-processing/reordering process.

The case study described in section 6.1.2 involves $J = 5$ basins and $K = 20$ lead times, leading to a $L = 100$-dimensional setting. It reproduces an operational hydrological forecasting context over the 2011-2014 period, where ECMWF 6-hour MAP forecasts are post-processed using forecasts from the past 4 years for training and reordered using historical observations within the 1992-2010 period (if historical observations are required, as for AnSS, MDSS, AnSS-G and AnSS-ECC). The verification experiment is made of three steps to which the next subsections correspond.

6.1.4.1 Verification of pairwise correlations in MAP forecasts

First, we evaluate the ability of the different methods to generate MAP forecasts that reproduce the observed correlations. This step aims at detecting if the reordering process induces a systematic under or over-estimation of the forecast member correlations across the different dimensions. It is the verification strategy that was undertaken in Clark *et al.* (2004) to evaluate the performance of the Schaake shuffle.

For estimating the dependence between two series of precipitation data, we use the Kendall's τ rank correlation coefficient, on series containing only pairs of both non-zero values. This is suggested by Yoo et Ha (2007) who found out that considering elements with zeros in either one or both values yields abnormally high estimates of correlation coefficients. This result was confirmed by Serinaldi (2008), who in addition recommends the use of the Kendall's τ instead of the Pearson'r product-moment correlation coefficient, since the latter is not meaningful when data series show substantial departures from the bivariate normal distribution, as it is the case with precipitation. Given a case of dimension $L = 100$, there are possibly $L(L-1)/2 = 4950$ pairwise dimension combinations for which Kendall's τ can be computed. However, preliminary works have shown that MAP observations are significantly correlated across the 5 basins at the same time step or on two consecutive time steps, but not more. We therefore consider more restricted frames of dimension 10 that include MAP series for all basins over two consecutive time steps. For evaluating correlations in forecasts, we consider the frames as moving where the two consecutive time steps correspond to two consecutive lead times. For example, the frame labelled as "6/12h" includes MAP series comprised of forecast values valid at first and

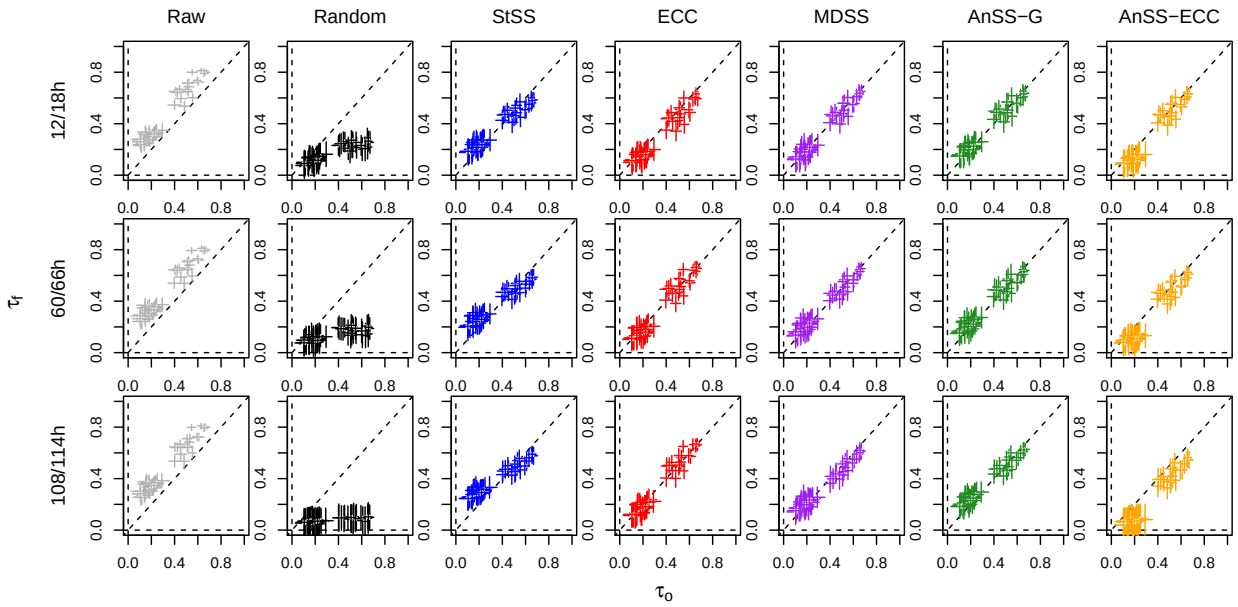

FIGURE 6.6 – Forecast Kendall's rank correlation coefficients τ_f versus observed ones τ_o for each pairwise combinations in the 10-dimensional frames including the 5 study basins and 2 consecutive lead times. Vertical and horizontal bars represent 90% confidence intervals computed using bootstrapping. Results are shown for 3 frames involving different lead times.

second lead times. For each frame, there are consequently 45 pairwise combinations for which Kendall's τ for forecasts (τ_f) and observations (τ_o) are compared. Note that the τ_f are taken as the mean of the 51 Kendall's τ obtained for each individual member of the ensemble. Figure 6.6 compares observed and forecast correlations by plotting the τ_f of the 45 pairwise combinations versus corresponding τ_o, over 3 different frames involving different lead times. Unbiased forecasts regarding spatial and temporal correlations should have all pairwise points lying on the diagonal. One can discern in all plots two distinguishable groups of points. The one with $\tau_o \in [0.4, 0.8]$ displays spatial correlations (i.e., pairwise combination of different basins at the same lead time) while the one with $\tau_o \in [0, 0.4]$ corresponds to temporal correlations (i.e., either same or different basins, but at two consecutive lead times). Confidence intervals at the 90% level around Kendall's τ coefficients are computed using bootstrapping (Efron et Tibshirani, 1994) and represented with vertical and horizontal bars.

We observe in Figure 6.6 that Raw overestimates spatial and temporal correlations, while ECC does not, although both share same rank dependence structures. This indicates that the analysis of correlations in multivariate forecasts is not independent of the univariate distributions. In this case, increasing the forecast dispersion by post-processing has weakened the correlations, pulling ECC points towards the diagonal. As expected, Random is not able to reproduce spatio-temporal correlations, even though the τ_f are significantly positive for short lead times due the small dispersion of the forecasts compared to longer lead times. Indeed, the smaller is the dispersion of the forecasts, the less crucial is an appropriate reordering. Concerning StSS, MDSS, AnSS-G and AnSS-ECC, all seem able to reproduce spatio-temporal correlations fairly well. Nonetheless, one can notice that reordered forecasts with observation-based temporal structures, as StSS, MDSS and

AnSS-G, tend to slightly overestimate temporal correlations, while ECC and AnSS-ECC with model-based temporal structures tend to underestimate them.

6.1.4.2 Multivariate verification of MAP forecasts

By comparing forecast pairwise correlation coefficients to observed ones, we verify whether a given reordering method is able to generate precipitation forecasts that are physically realistic. However, the adequacy between each forecast and its corresponding observation is not taken into account. Thus, the ECC, MDSS, AnSS-G and AnSS-ECC techniques cannot be rewarded for their main strength with respect to StSS, namely their capacity to condition dependence structures on forecast cases. Consequently, we now consider verification metrics that evaluate the joint behavior of the precipitation forecasts and their corresponding observations.

We employ the energy score (ES) and the variogram score (VS) introduced by Gneiting *et al.* (2007) and Scheuerer et Hamill (2015a), respectively, which quantify the overall quality of multivariate ensemble forecasts. Their definitions are given in sections 4.4.2.1 and 4.4.2.2. The ES and VS are the scores that have been most popular in recent studies evaluating multivariate forecasts (e.g., Schefzik *et al.*, 2013; Keune *et al.*, 2014; Hemri *et al.*, 2015; Schefzik, 2016a, 2016b; Ben Bouallègue *et al.*, 2016; Scheuerer *et al.*, 2017). One of the most common criticism of the ES is its limited capacity to discriminate forecasts with different multivariate dependence structures, whereas it nicely detects misspecification of univariate means and variances (Pinson et Tastu, 2013; Scheuerer et Hamill, 2015a). In theory, the VS constitutes an appropriate alternative, as it is designed to be more sensitive to multivariate dependence structures. Note however that this aspect should not be overstated here, since all forecasts to be evaluated (except Raw) share the same univariate ensemble values. In other words, any ES or VS variations between methods, albeit modest, are entirely attributable to dependence structures and can be interpreted accordingly. The same 10-dimensional frames as in section 6.1.4.1 are considered. Results are presented reporting energy skill scores (ESS) and variogram skill scores (VSS). Skill scores, whose general definition is given in section 4.2.1.3 , relate the performances of the forecasts to those of a reference forecast dataset. Values between 0 and 1 indicate that the forecasts outperform the reference forecasts (with 1 corresponding to perfect forecasts), while negative values indicate lower performances than the reference forecasts ones. For this study, Raw forecasts are taken as reference.

Figure 6.7 depicts the ESS and VSS as a function of the lead time pairs that are considered in each 10-dimensional frame. First of all, we note that all post-processed forecasts, regardless the reordering, outperform raw forecasts at short lead times (until +72h according to the ESS and +48h according to the VSS), while a sophisticated reordering is necessary to preserve skill until +120h. Furthermore, the improvement of ECC, StSS, MDSS, AnSS-G or AnSS-ECC over Random is apparent, all the more as lead times increase. Again, it shows that reordering becomes increasingly important as the dispersion of ensemble forecasts increases. Figure 6.7 then shows that ECC outperforms StSS for all lead times (except at 6/12h for the VSS). However, ranking the three other methods, namely MDSS, AnSS-G and AnSS-ECC, is lead time and score dependent. We further

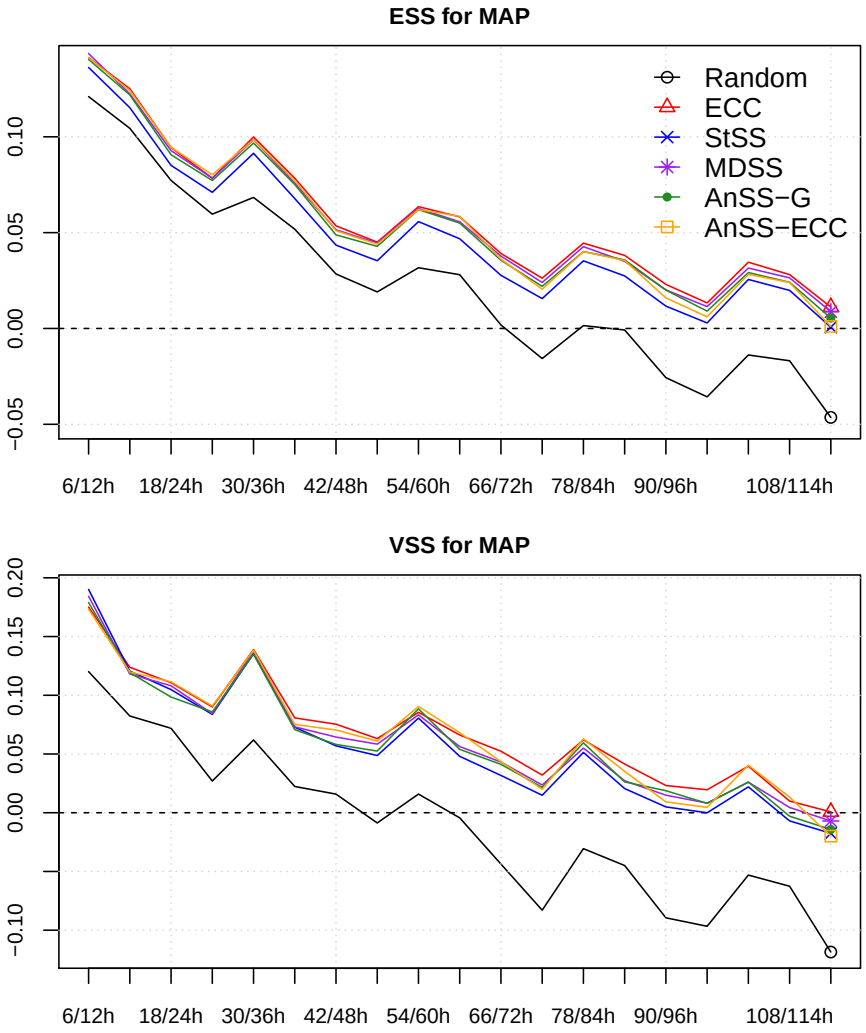

FIGURE 6.7 – ESS and VSS of post-processed MAP forecasts, computed over 10-dimensional frames including the 5 study basins and 2 consecutive lead times. The horizontal axis represents these lead time pairs. Raw forecasts serve as reference.

observe that, according to the ESS, their skill is close to ECC for shorter lead times, while tending to StSS for longer lead times. This finding is not fully supported by the VSS though. In addition, VSS discrepancies between methods are surprisingly not much different from ESS ones, which contrasts with results of Scheuerer et Hamill (2015a) but is in accordance with those of Schefzik (2016b).

6.1.4.3 Univariate verification of streamflow forecasts

Sections 6.1.4.1 and 6.1.4.2 have focused on MAP forecasts. The last verification step of this experiment consists in evaluating the resulting streamflow forecasts, which are daily issued at 00 UTC, at an hourly time-step. The evaluation is made on the added contributions of the 5 study basins at the global outlet of Pont d'Evieu, using the models described in section 6.1.2.2. The so-obtained streamflow do not correspond to any measurable quantity as the records on the Rhone River at the Pont d'Evieu gauging station include other inflows coming from upstream as well as from lateral drainage areas. Nonetheless, streamflow forecasts are evaluated against simulated streamflow, as the result of the hydrological modeling chain forced by observed precipitation, and not against observed streamflow. Hydrological errors are therefore not taken into account. Considering lead times independently, hydrological modeling has reduced the dimension of the verification problem to one. We can thus employ univariate verification metrics to assess the forecast quality, which is here separated in two different attributes: calibration and sharpness. Calibration is a joint property of the forecasts and the observations and refers to the statistical consistency between the two, while sharpness refers to the dispersion of the forecasts, hence it is a property of the forecasts only.

Although the verification of probabilistic forecasts requires a large sample for assessing calibration, the visual inspection of individual hydrographs is undoubtedly useful for a better understanding of how forecasts behave, regardless the position of the simulated streamflow in the forecast spread. Thus, let us first take a closer look at the forecast of 30 January 2013, which has been used in Figures 6.2, 6.4 and 6.5. Figure 6.8 shows the different streamflow forecasts that were obtained using the MAP forecasts derived from each reordering method. Note that additional examples can be found in Figures A.4-A.7 in the supporting information to this paper. Different behaviors can be observed. First, the Random hydrograph exhibits a small dispersion, as a result of MAP forecast values within the different basins and lead times that are associated randomly. The low probability of associating several high values of MAP in the same trajectory induces a "smoothing effect" that yields a large majority of average streamflow trajectories which have, in this example, peak values around 350 m^3s^{-1}. Such a smoothing effect can also be observed on the lowest trajectories (the ones around 200 m^3s^{-1} at peak) of the StSS hydrograph, as an illustration of the problematic issue that concerns the number of zeros present in the spatio-temporal dependence template $\mathbf{z}$ used for reordering the MAP forecast. Recall that StSS uses as template past observed trajectories from the climatology that are not conditioned on the forecast. Due to the predominance of dry days in the climatology, template ensembles $\mathbf{z}^{j,k}$ for each dimension (j,k) are likely to contain a large number of zeros for which ranks will be assigned randomly. On the other hand, non-zero values will

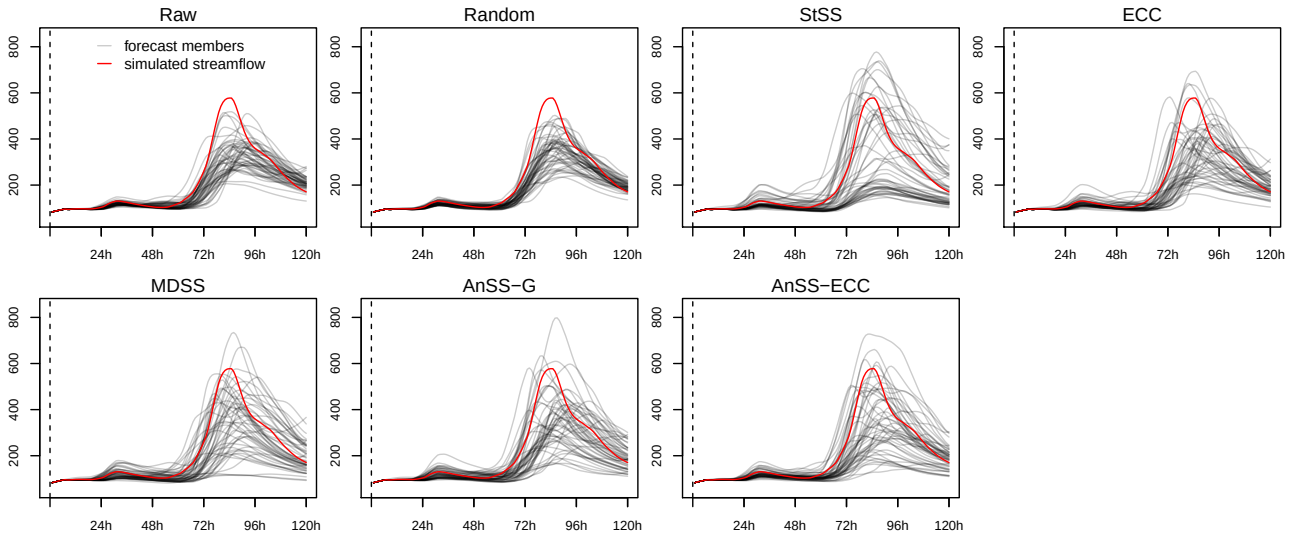

FIGURE 6.8 – Examples of streamflow forecasts at Pont d'Evieu issued on 30 January 2013 at 00 UTC. The vertical axis unit is m^3s^{-1}.

be ranked normally, thus reproducing the correlation of the rainy days of the climatology. This is clearly visible in the middle row of Figure 6.2, where the Latin square of StSS that displays the rank dependence structure exhibits a "random square" within which the dots corresponding to members ranked 1st to approximately 40th are randomly spaced. The dots corresponding to the other members lie in the upper-right part of the plot, indicating a strong positive correlation. By construction, StSS rank dependence structures for precipitation will systematically include such a random square, its size depending on the proportion of dry days in the climatology. When applying StSS to rainy forecasts where post-processed ensembles $\tilde{\mathbf{x}}^{j,k}$ contain almost only non-zero values, as it is the case in Figure 6.2, the lowest (yet positive) values are associated together but in a random way, while the highest values are associated together according to the strong positive correlation. Figure 6.9 displays the same streamflow forecast for StSS as in Figure 6.8, although each forecast trajectory $m \in \{1, \dots, M\}$ is here coloured with respect to the percentages of zeros in the corresponding MAP template trajectory $\mathbf{z}_m = (z_m^{1,1}, \dots, z_m^{J,K})$. Low-reactive trajectories are clearly distinguishable, with percentages close to 100. The MDSS, AnSS-G and AnSS-ECC techniques also use observations for specifying the template, but their conditioning on the forecast case greatly weakens this phenomenon.

These behaviors are confirmed by the study of rank histograms (Hamill, 2001) constructed from the streamflow forecasts for individual lead times. The rank histogram aims at evaluating forecast calibration, but its qualitative assessment is also highly informative for better understanding how forecasts behave over a large number of cases. To minimize the risk of different forecast behaviors that average out in the rank histogram, we use the accumulated stratified representation suggested in Bellier *et al.* (2017a), which decomposes the overall rank histogram (i.e., computed over the complete verification sample) in different strata. Each strata represents the part of the rank histogram that originates from a subset of the complete sample. The subdivision is here made according to the streamflow forecast ensemble mean, considering the same number of forecasts within each

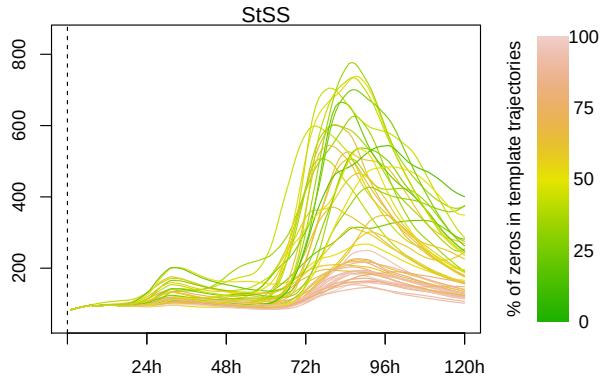

FIGURE 6.9 – Same as Figure 6.8 but for StSS only, where each forecast trajectory is coloured with respect to the number of zeros in the corresponding template trajectory.

strata. The interpretation of the shape of each strata is the same as for the overall rank histogram: flatness indicates a good calibration, ∪-shape and ∩-shape indicate under- and over-dispersion, respectively, while ⟋-shape and ⟍-shape indicate negative and positive bias, respectively.

Figure 6.10 depicts such stratified rank histograms for the 24, 72, and 120 hour lead times. One can observe that the calibration of streamflow forecasts is strongly impacted by the spatio-temporal dependence structures of MAP forecasts. Streamflow forecasts referred to as Random are strongly under-dispersive as shown by the ∪-shapes, the gain in calibration from MAP post-processing being not translated to streamflow because of the random dependence structures. For StSS, the problematic behavior discussed above appears clearly, in particular in the subset containing the highest streamflow forecasts (dark blue stratum), which corresponds to the most rainy events. The histograms for this stratum show right-shifted ∩-shapes, indicating that the reference does not fall often enough within the low-reactive trajectories (i.e., in the left part of the histogram), but too often just above. In other words, these low-reactive trajectories form an unrealistic mode in the predictive distributions that makes them abnormally right-skewed and over-dispersive. This lack of calibration is still visible for MDSS and AnSS-G but to a lesser extent, thanks to the conditioning of the template on the forecast case. Among the two reordering methods that make use of analogues, it appears that AnSS-G's behavior shows resemblances with StSS, while AnSS-ECC tends to behave more like ECC. This is an indication that temporal dependence structures in precipitation forecasts play a major role in the rainfall-runoff process, as will be discussed later on. Lastly, one can notice that the first bin of the rank histogram of AnSS-ECC becomes too highly populated as lead time increases, reflecting the reference falling too many times below the lowest member. This means that AnSS-ECC faces difficulties with generating low streamflow trajectories. We assume that it derives from temporal correlations that become under-estimated by AnSS-ECC as lead times increase, as Figure 6.6 shows.

We now evaluate the overall quality of the streamflow forecasts, using the continuous ranked probability score (CRPS) whose definition is given in section 4.2.1.1 . While rank histograms account for calibration only, the CRPS assesses calibration and sharpness simultaneously. Results are presented reporting skill scores (CRPSS ; see 4.2.1.3). Its

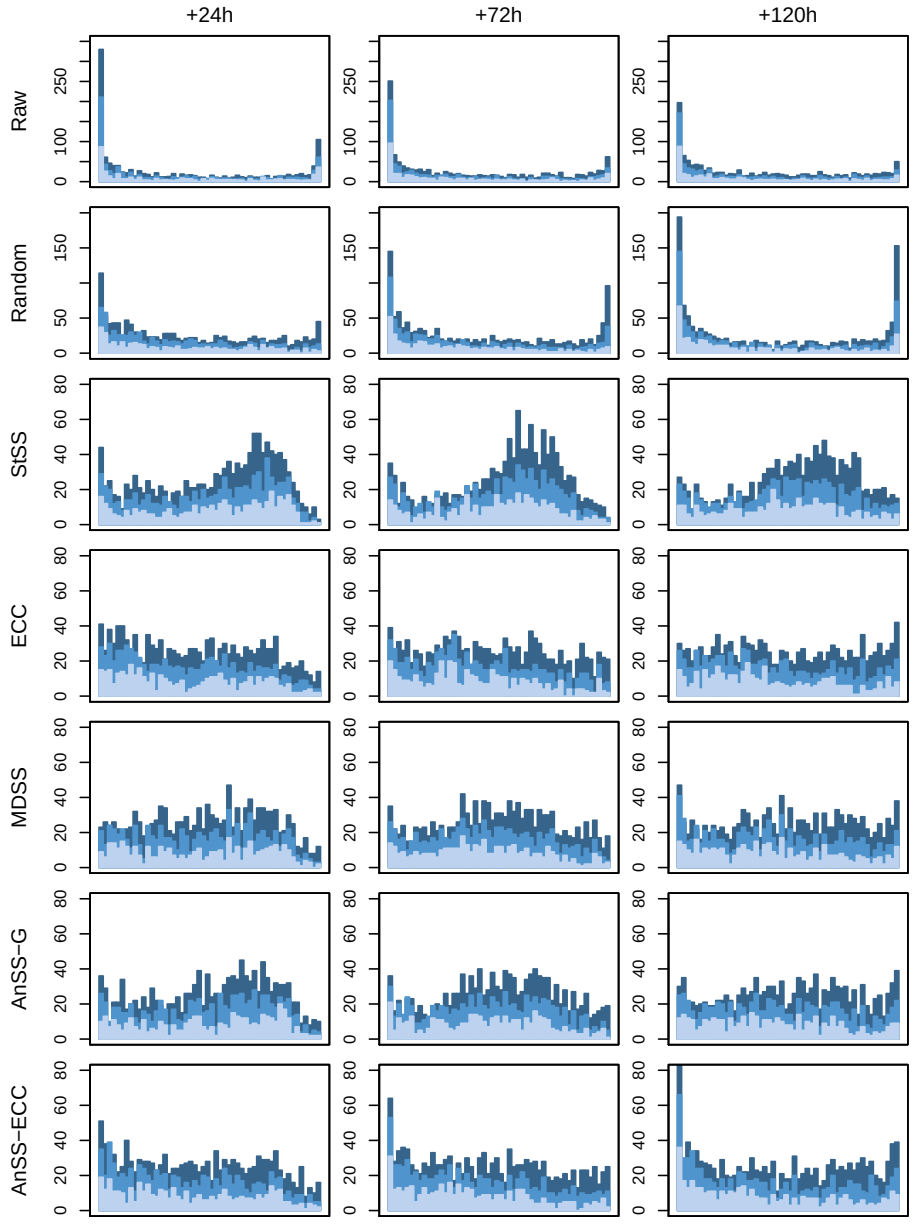

FIGURE 6.10 – Accumulated stratified rank histograms of streamflow forecasts for lead times 24, 72 and 120 hours. The stratification criterion is the forecast ensemble mean. Numbers of forecasts within each strata are equal. Light, medium and dark-blue strata contain lowest, intermediate and highest streamflow forecast, respectively.

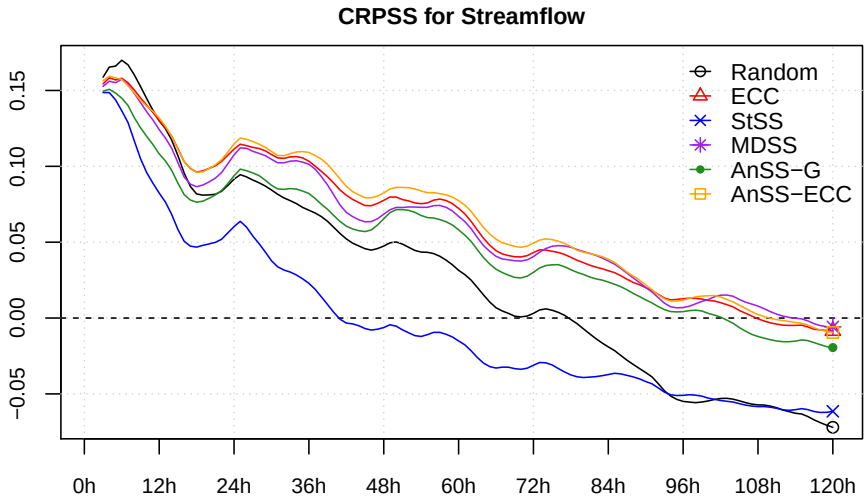

FIGURE 6.11 – CRPSS for streamflow forecasts at Pont d'Evieu, as a function of lead time. Raw forecasts serve as reference.

interpretation is similar as for ESS and VSS. Raw forecasts are again taken as reference.

Figure 6.11 displays the CRPSS as a function of lead time. We observe that the different reordering methods are ranked differently when compared to the results of section 6.1.4.2. AnSS-ECC now outperfoms AnSS-G, and, to a lesser extend, ECC and MDSS. The improvement of AnSS-ECC over ECC demonstrates that spatial structures from the model are worth replacing by ones from the analogues. In turn, the lower skill of AnSS-G, especially compared to MDSS of which it is a direct variant, indicates that the conditioning of the observed trajectories at a single lead time is not adapted to such a 120-hour forecast horizon. Figure 6.11 also shows that the CRPSS of StSS is noticeably lower than ECC, MDSS, AnSS-G and AnSS-ECC ones. More surprisingly, even Random outperforms StSS until +96h. Explanations are provided by the Figure 6.12, which compares the average ensemble standard deviation obtained with the different methods, as a measure of the forecast sharpness. We observe that Random exhibits the best sharpness (i.e., forecasts are, in average, the least dispersive), as a consequence of the smoothing effect discussed earlier that pulls all streamflow trajectories towards the mean. At the opposite, StSS shows the worst sharpness, due to the combination of the low-reactive trajectories affected by the smoothing effect and the trajectories that reproduce the correlation of the climatology. In overall, despite a strong miscalibration, Random ends with a sufficient sharpness for its CRPSS to exceed that of StSS, at least until +96h. This result, based on the CRPSS only, does not mean that Random should be preferred over StSS, as the independence between basins and lead times is meaningless for hydrological applications, while StSS reproduces observed correlations. Rather, it demonstrates that StSS has limitations, and encourages the use of flow dependant techniques such as ECC, MDSS, AnSS-G or AnSS-ECC.

These findings illustrate the difficulty in evaluating the spatio-temporal dependence structure of MAP forecasts. Our verification experiment has shown that the ESS/VSS-based results from section 6.1.4.2 and the CRPSS-based results presented above conduct to inconsistent results, i.e., the ranking of the methods is different, although both ap-

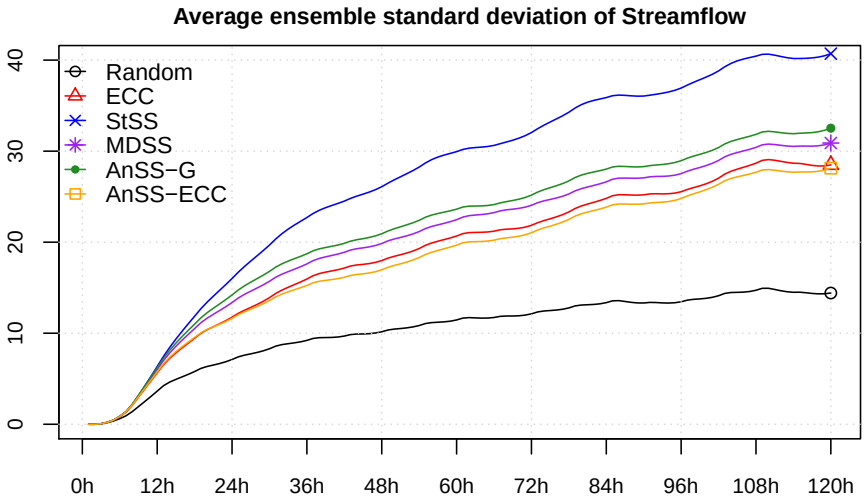

FIGURE 6.12 – Average ensemble standard deviation of streamflow forecasts at Pont d'Evieu, as a function of lead time. The vertical axis unit is m^3s^{-1}.

proaches evaluate, in certain ways, multivariate MAP forecasts. This can be explained by the following reasons. First, the ESS and VSS computed on the 10-dimensional frames are less sensitive to misspecification of temporal dependence structures than of spatial ones, since the former are generally less pronounced than the latter. However, temporal structures are of greater importance in hydrological modeling. The basin has a "memory effect" that makes sequences of precipitation over multiple days crucial for soil moisture conditions. Rainfall-runoff modeling is moreover highly non-linear, since different processes such as infiltration or overland flow may be triggered or not depending on antecedent precipitation. The second reason is that hydrologically-oriented verification is case-specific. Basin sizes, concentration times or morphological characteristics are taken into account and may lead to a balance between basins that differs from that in the ESS/VSS-based verification where all basins have equal weights. This illustrates the benefit of a hydrologically-oriented verification of meteorological forecasts, as advocated by Pappenberger *et al.* (2008b). In addition, in perspective of further verification studies it is worth mentioning the use of spatial verification metrics, which aim at assessing how forecast fields reproduce various spatial characteristics of observation fields (see Gilleland *et al.* (2009) for a review). However, such techniques are essentially tailored to gridded forecasts, hence they have not been tested in this study.

6.1.5 Conclusions

The first objective of this paper was to propose new techniques that use meteorological analogues for reordering post-processed precipitation ensemble forecasts in view of hydrological forecasting, as alternatives to the popular Schaake shuffle and ECC methods, or to the recent MDSS technique. The AnSS-G method was first described, as an adaptation of the Schaake shuffle that uses the analogues to condition historical trajectories on predicted atmospheric states. The conditioning is accomplished at the lead time that represents the center of gravity of the forecast regarding precipitation amounts. It

is designed for short-term forecasts, i.e., a couple of days ahead, and over a relatively small spatial area, since the analogues are common to all the basins. A second method referred to as AnSS-ECC was proposed, which uses the analogues for spatial reordering while retaining the temporal structure of the raw ensemble. The reordering is therefore ensured to be flow-dependent from short to longer lead times. Straightforward adaptations can allow the method to apply over large geographic areas. These two techniques, which are based on observations for setting the dependence template, are particularly relevant with respect to ECC when the atmospheric model at hand has a coarse spatial resolution compared to the basin scales. Also, these are promising alternatives to MDSS in regions where precipitation events may arise from various atmospheric situations to which different spatio-temporal patterns correspond.

The second objective was to quantify the impact of the precipitation spatio-temporal dependence structure on hydrological forecasting. For that purpose, the AnSS-G and AnSS-ECC techniques, as well as the existing Schaake shuffle, MDSS and ECC ones, have been evaluated through a multi-step verification experiment. The first step consisted in verifying the ability of the reordering methods to generate physically realistic precipitation forecasts, i.e., which reproduce the observed spatio-temporal correlations. The five methods were found satisfactory regarding this aspect. The second step consisted in evaluating each multivariate precipitation forecasts individually via the energy and the variogram scores. Again, all techniques showed similar skills, and clearly outperformed the forecast dataset that assumed a random reordering. In the last verification step, streamflow forecasts resulting from hydrological modeling were evaluated against simulated streamflow (as a result of models forced by observed precipitation). Very different results from the two previous steps were found out, in particular the standard Schaake shuffle being less skilful than the other reordering techniques, but more surprisingly, than the random reordering too. These findings thus evidenced the sensitiveness of hydrological modeling to the accuracy of the spatio-temporal dependence structure in precipitation ensemble forecasts. By adding a case-specific filter, the hydrologically-oriented verification gives an important weight to certain aspects of the dependence structure to which the hydrological model is highly sensitive.

Probabilistic hydrological forecasting systems in operational contexts are generally made up of several cascaded components. Our setting has included the following: meteorological forcings, post-processing, reordering and hydrological modeling. This paper is a contribution to the study of the reordering component, where an inter-comparison experiment has been conducted leaving all other components unchanged. Future works are nevertheless needed to verify whether our conclusions are reached within more integrated approaches. In addition, precipitation has been the only weather variable considered, but in hydrological modeling the temperature is likely to play a role as well, which would imply multiplying by two the dimension of the multivariate setting. Accounting for temperature is not a big issue when applying ECC, since the model provides both precipitation and temperature outputs for each of the ensemble members. However, Schaake-shuffle-based techniques that preferentially select the historical trajectories to specify the template would require assumptions to be made on which variable is the most important in the se-

lection process, as in Scheuerer *et al.* (2017) where precipitation prevails. The covariability between the two variables should therefore be further assessed. Finally, issuing probabilistic streamflow forecasts that are calibrated also requires a post-processing component that accounts for hydrological model errors. The reconstruction of coherent hydrological scenarios is thus likely to call for an additional reordering process, as studied by Hemri *et al.* (2015). More works are therefore required to investigate how multivariate aspects should be treated within complete hydrological forecasting systems.

Acknowledgments

Joseph Bellier's research is supported by grants from Labex OSUG@2020 (Investissement d'avenir – ANR 10 LABX56) and Compagnie Nationale du Rhône. The authors gratefully acknowledge Michael Scheuerer for sharing his R codes of EMOS post-processing and MDSS reordering. They also thank the three reviewers for their helpful comments and suggestions. Forecast data used in this paper are freely available on the TIGGE archive hosted on the ECMWF data portal (https://www.ecmwf.int/en/research/projects/tigge). Data extraction has been made on May 2015. Precipitation and streamflow observation data were provided by Météo France and Compagnie Nationale du Rhône, respectively, and can be obtained on request for research purposes.

6.2 Et la température dans tout ça ?

Nous avons considéré jusqu'à maintenant le réarrangement des prévisions de précipitation dans un cadre spatio-temporel, sans inclure la cohérence inter-variable. Il est cependant aisé de rajouter au cadre théorique de la section 6.1.3.1 un indice supplémentaire $i \in \{1, \dots, I\}$ qui se réfère à la variable météorologique. La dimension du contexte multivarié devient donc $L = I \times J \times K$. Dans notre contexte de prévision hydrologique, les forçages météorologiques concernent les variables précipitation et température seulement, et donc $I = 2$.

Comme mentionné en conclusion de l'article, la méthode ECC est tout à fait adaptée à la cohérence précipitation-température car l'ensemble brut, supposé cohérent, contient les prévisions pour ces deux variables. L'extension du Schaake shuffle standard (StSS) à un tel contexte est également trivial, à condition que les dates aléatoirement sélectionnées dans l'archive des observations comprennent les deux variables. En revanche, les adaptations du Schaake-shuffle présentées dans l'article, où les dates ne sont plus sélectionnées aléatoirement mais conditionnées à la prévision, sont sujettes à discussion car elles ont été pensées d'abord pour la cohérence spatio-temporelle des précipitations.

Les auteurs de la méthode MDSS, Scheuerer *et al.* (2017), considèrent que le conditionnement de la structure de dépendance spatio-temporelle des précipitations prévaut sur le conditionnement des autres structures (structure spatio-temporelle de la température, et structure précipitation-température). Ils suggèrent alors de réaliser un pré-filtrage des dates en fonction de la prévision de température, puis d'appliquer l'algorithme de minimisation de la divergence totale (cf. équation 6.9) sur les prévisions de précipitation seulement. La méthode AnSS-G suit implicitement la même hypothèse, car l'algorithme de sélection des dates analogues, bien que conçu pour être pertinent vis-à-vis des précipitations, réalise néanmoins un pré-filtrage en fonction de la température (cf. tableau 6.2). Ainsi, ces deux approches permettent de conditionner de manière sophistiquée la structure des précipitations, tout en assurant une cohérence « climatologique » des températures. Il serait possible de chercher à améliorer ces méthodes en donnant davantage de poids à la température : pour MDSS, en incluant la température dans le calcul de la divergence totale ; pour AnSS-G, en modifiant l'algorithme de sélection des analogues de manière à ce que ce dernier soit davantage pertinent vis-à-vis de la température. Des travaux dans ce sens ont été récemment menés par Caillouet (2016). Cependant, cela vaut-il la peine de dégrader le conditionnement des structures de précipitation pour améliorer celui des températures ?

La méthode AnSS-ECC est différente, car elle mixe les analogues (pour le réarrangement spatial) et l'ensemble brut (pour le réarrangement temporel), avant de coupler les deux sources de données grâce à la fonction de moyenne spatiale (cf. équation 6.10). Or cette fonction ne s'applique qu'aux précipitations, ce qui signifie qu'il n'est pas possible de confier aux analogues le réarrangement spatial des prévisions de température. Le réarrangement des prévisions de température doit donc se baser entièrement sur l'ensemble brut. Quel est l'impact sur les débits lorsque cet ensemble brut a une résolution spatiale trop large par rapport à la taille des bassins ?

Il serait nécessaire, pour apporter des réponses à ces questions, de réaliser une étude plus approfondie sur l'impact de la structure de dépendance des prévisions de température (spatiale, temporelle, mais également sa dépendance avec les précipitations). Cependant, une telle étude n'a pas pu être menée durant cette thèse, à cause de la fragilité de nos données d'observation de température. Pour mémoire, ces données ne proviennent pas de réelles observations, mais de la réanalyse ERA-Interim dont la résolution spatiale native est de 0.75°. Cette résolution, très basse comparée à l'échelle des bassins étudiés, est un obstacle à l'estimation fiable de la covariabilité des températures entre les bassins. Plutôt que de tirer des conclusions fragiles, nous avons préféré réserver cet aspect à des travaux ultérieurs.

Nous proposons donc, dans la suite de cette thèse, de contourner la problématique de l'impact du réarrangement des prévisions de température en adoptant la méthode ECC, qui a montré de bons résultats lors de l'évaluation dans la section 6.1.4, et qui s'applique naturellement au contexte précipitation-température.

6.3 Synthèse

Les prévisions météorologiques d'ensemble se révélant peu fiables, nous leur avons appliqué une correction statistique, appelée pré-traitement. Réalisée dans un cadre univarié, cette correction a entrainé la perte de la structure de dépendance (spatiale, temporelle et inter-variable). Structure qu'il a fallu reconstruire à l'aide d'un procédure de « réarrangement », afin d'obtenir des prévisions multivariées qui puissent alimenter un modèle hydrologique. Dans ce chapitre, nous nous sommes tout particulièrement intéressé à la structure de dépendance spatio-temporelle des prévisions de précipitation. Les principaux résultats obtenus sont les suivants :

- L'inter-comparaison de différentes méthodes de réarrangement ne donne pas toujours les mêmes résultats selon que la vérification porte sur les forçages météorologiques eux-mêmes ou sur les débits après modélisation hydrologique.

- Le Schaake shuffle, qui s'appuie sur la climatologie pour construire la structure de dépendance, s'avère peu performant lorsque la vérification porte sur les débits. Il est même dans ce cas moins performant qu'une procédure proposant un réarrangement aléatoire. Les variantes consistant à conditionner la structure à la situation de prévision s'avèrent nettement plus performantes. On retrouve notamment la méthode de Scheuerer *et al.* (2017) de similarité des marginales (MDSS), ainsi que l'approche proposée dans ce chapitre qui conditionne la structure via les analogues pour une des échéances (AnSS-G).

- La méthode ECC, qui reproduit la structure de dépendance du forçage brut, donne par ailleurs de bons résultats. Son couplage avec les analogues (AnSS-ECC) permet de légèrement améliorer les performances, en affinant encore la structure de dépendance spatiale.

Bien que ce soit la méthode AnSS-ECC qui ait été la plus performante dans notre expérience, nous utiliserons dans la suite de la thèse la méthode ECC, qui d'une part présente l'avantage de s'appliquer naturellement au contexte précipitation-température, et d'autre part fait preuve de simplicité tout en ne nécessitant aucune autre source de données que le forçage brut.

Troisième partie

Fiabilité et cohérence des prévisions hydrologiques

Introduction à la Partie III

Le forçage météorologique comme objet d'étude ayant fait l'objet de la partie II, nous nous penchons désormais sur les prévisions hydrologiques qui en découlent. La variable étudiée est donc le débit au pas de temps horaire. Nous limitons ainsi notre zone d'étude aux seuls bassins pour lesquels la modélisation hydrologique a été entreprise, à savoir l'Arve, la Valserine, les Usses, le Fier, le Séran et le Guiers. Pour améliorer les prévisions hydrologiques sur ces bassin, nous empruntons une approche en deux temps qui rappellera celle entreprise sur les forçages.

Ainsi, nous étudions dans le chapitre 7 la prise en compte de l'incertitude hydrologique, à l'aide d'une approche ensembliste, le multi-modèle, et d'une correction statistique univariée, appelée désormais « post-traitement » car postérieure à la modélisation hydrologique.

Le chapitre 8 sera ensuite consacré à la reconstruction de prévisions multivariées cohérentes, afin de prendre en compte la covariabilité cruciale qui existe entre les différents bassins et échéances. Si les principes peuvent sembler très similaires, il s'avère que la variable débit présente plusieurs spécificités qui justifient un traitement séparé.

Nous terminerons ce travail de thèse dans le chapitre 9, en prenant du recul sur l'ensemble des maillons de la chaine de prévision. Nous adopterons alors une approche intégrée afin d'évaluer l'intérêt de chacun des ces maillons, et tenterons de dresser d'éventuelles priorités quant à leur mise en œuvre au sein d'une chaîne opérationnelle de prévision hydrologique probabiliste.

Chapitre 7

Multi-modèle et post-traitement univarié

Le travail mené jusqu'à présent a permis de disposer de forçages météorologiques à la fois fiables et cohérents. Nous nous posons désormais la question de la capacité de la modélisation hydrologique à transférer cette fiabilité dans les prévisions de débit. Cette étape de modélisation étant par nature incertaine, il y a fort à parier que la simple propagation de forçages météorologiques fiables ne soit pas suffisante pour produire des prévisions hydrologiques fiables. Nous nous attacherons à vérifier cette hypothèse, avant de se pencher sur deux stratégies d'amélioration, l'une ensembliste, à savoir le multi-modèle, et l'autre statistique.

Compte-tenu des résultats précédemment acquis, nous considérons la chaîne de prévision de référence suivante :

- **forçages météorologiques** : ECMWF-Ens. Ces forçages contiennent 51 membres, et couvrent la période 2011-2014.
- **pré-traitement** : correction univariée par l'EMOS, puis réarrangement par l'ECC.
- **modèle hydrologique** : GRP, couplé avec Cemaneige et la formulation d'Oudin pour l'évapotranspiration potentielle.

Après le diagnostic des prévisions sortant de cette chaîne dans la section 7.1, nous évaluons dans la section 7.2 l'intérêt de l'approche multi-modèle, qui doit permettre de prendre en compte de manière non statistique une partie de l'incertitude hydrologique. La section 7.3 présente ensuite les éléments méthodologiques de la méthode BMA utilisée pour la correction statistique. La section 7.4 est consacrée aux résultats, où nous comparons notamment la BMA avec la méthode EMOS déjà utilisée pour le pré-traitement. Enfin, la section 7.5 fait la synthèse de ce chapitre.

7.1 Diagnostic des prévisions en mono-modèle hydrologique

Faut-il le rappeler, les prévisions probabilistes ne peuvent être évaluées que sur un échantillon rassemblant un grand nombre de réalisations. Par conséquent, des conclusions quant à la qualité des prévisions ne peuvent être tirées de cas isolés. Cependant, ces cas isolés n'en restent pas moins illustratifs, et permettent parfois de mieux appréhender certains comportements des prévisions.

C'est pourquoi nous proposons de débuter ce chapitre par l'illustration (en Figure 7.1) de 12 prévisions issues de notre chaîne de référence, pour des cas concernant le bassin de la Valserine et sélectionnés aléatoirement sur la période 2011-2014. Les débits simulés (c'est-à-dire les débits obtenus avec les forçages observés) du modèle GRP sont tracés en bleu, de manière à appréhender l'erreur de modélisation hydrologique.

Parfois, le modèle hydrologique semble fournir des simulations de bonne qualité (cas **a**, **e** et **h** par exemple), tandis que d'autre fois il propose des simulations moins appropriées (cas **c** et **d** par exemple). Ces cas concernent aussi bien des hauts débits que les bas débits. Par ailleurs, il est fréquent que la dispersion des prévisions soit nulle ou faible (cas **a**, **c**, **j**, **i** et **l** par exemple), en conséquence de forçages météorologiques qui ne sont pas suffisamment dispersifs pour provoquer des réactions (du modèle hydrologique) différentes d'un membre à l'autre. L'observation est, dans ces cas là, bien souvent en dehors du spectre des valeurs prévues. Le besoin de correction statistique semble donc, à première vue, bien réel !

Les histogrammes de rang, tracés à la Figure 7.2 pour l'ensemble des bassins, confirment le bien fondé de ces soupçons. En effet, les résultats mettent en lumière une sous-dispersion généralisée, dans des proportions qui varient peu selon les bassins, mais également peu selon les échéances. Ce dernier point témoigne d'une incertitude structurelle pour l'instant non prise en compte : l'incertitude de modélisation hydrologique.

Il existe différentes méthodes pour prendre en compte de manière ensembliste une partie de l'incertitude de modélisation hydrologique. Nous avons mentionné, dans la section 1.2.2 en introduction du mémoire, la possibilité d'utiliser plusieurs modèles hydrologiques, plusieurs jeux de paramètres pour un même modèle hydrologique, ou encore plusieurs procédures d'initialisation [1]. Nous proposons, dans cette thèse, d'expérimenter l'approche multi-modèle. N'ayant pas testé les autres alternatives, nous n'affirmerons pas que le multi-modèle est la meilleure approche. Ce n'est pour nous qu'un outil, qui nous permet de nous intéresser à l'usage des prévisions hydrologiques d'ensemble. Nous laissons à d'autres le soin d'explorer plus en détail ce champ d'étude.

1. Pour être exact, l'utilisation de plusieurs procédures d'initialisation relève de la prise en compte des conditions initiales du bassin et non pas de la modélisation hydrologique.

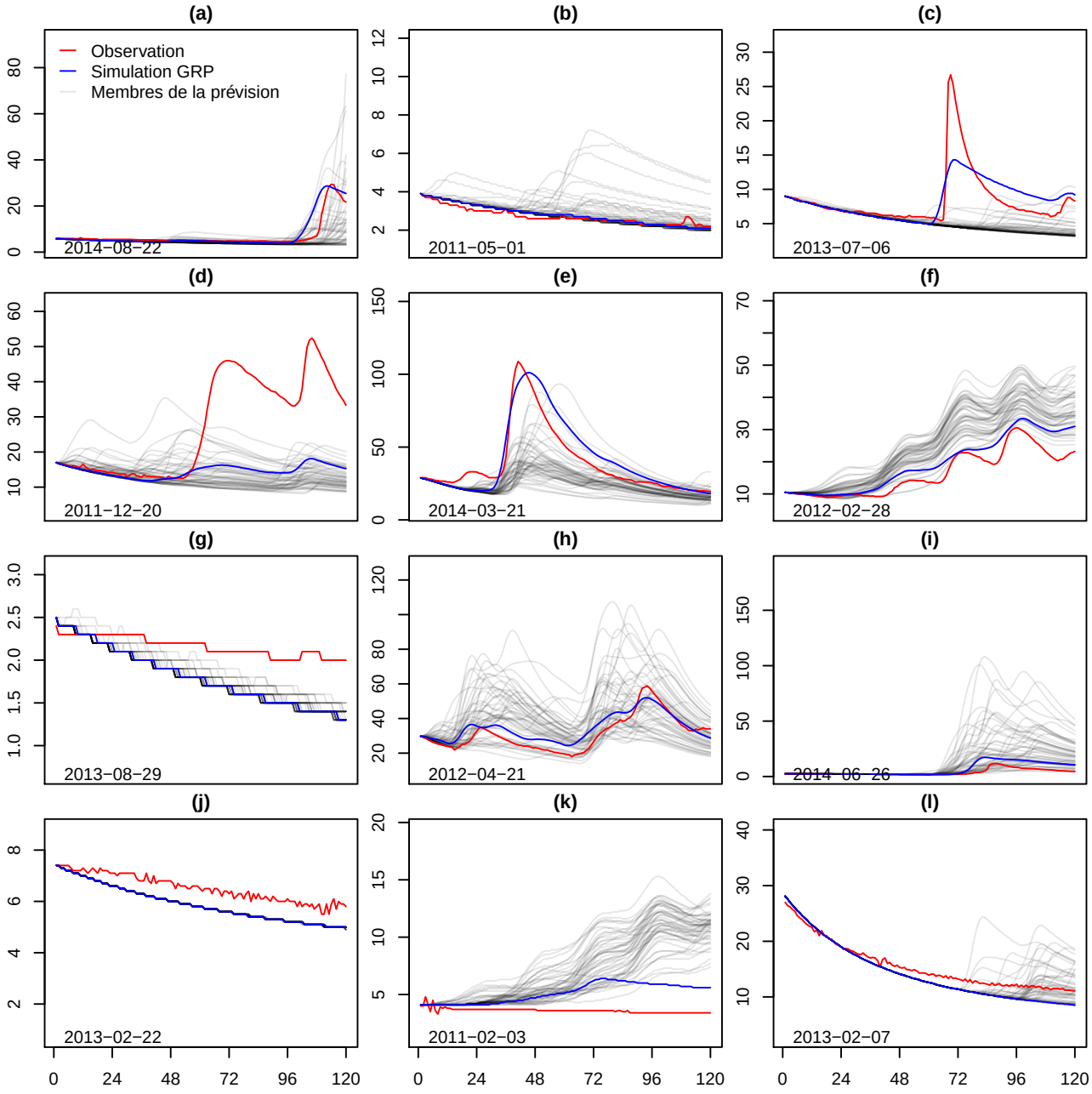

FIGURE 7.1 – Exemples de prévisions sur la Valserine, sélectionnées aléatoirement sur la période 2011-2014. L'axe vertical est en m^3/s, et l'axe horizontal en heures.

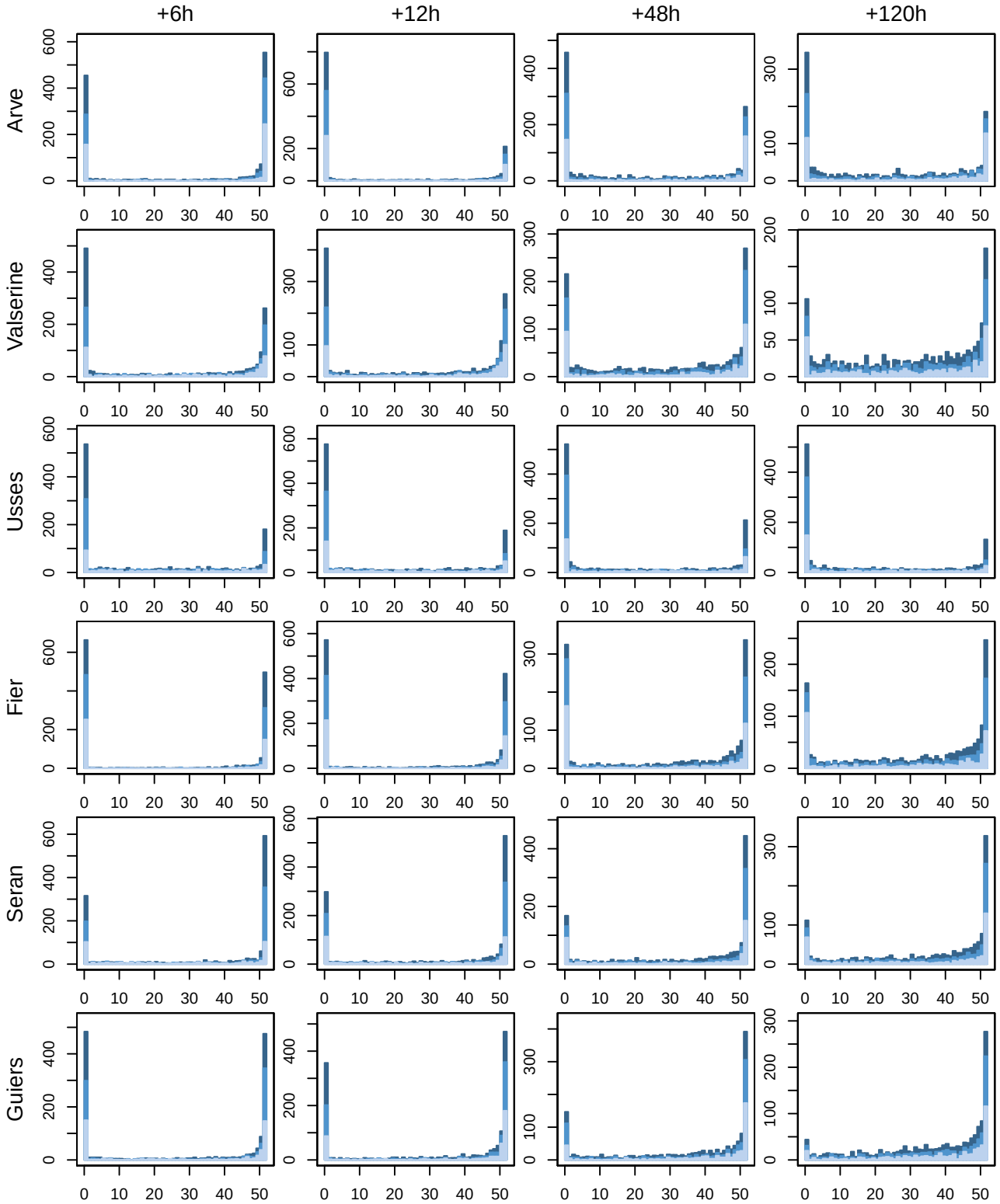

FIGURE 7.2 – Histogrammes de rang des prévisions de débit en sortie du modèle GRP, pour chacun des bassins et pour différentes échéances. Les histogrammes sont stratifiés selon la prévision moyenne, avec un nombre égal de cas dans chaque strate.

7.2 De l'intérêt d'une approche multi-modèle

L'approche multi-modèle consiste à ne pas considérer un unique modèle hydrologique mais plusieurs modèles, qui sont idéalement de structure différente. Dans notre cas nous disposons, en plus de GRP, des modèles ARX et TOPMODEL, ce qui forme un jeu de trois modèles hydrologiques de conception et de structure différentes (cf. 3.2). Le nombre de membres en sortie est ainsi multiplié par trois.

La Figure 7.3 reprend les mêmes 12 cas de prévision que précédemment, en illustrant désormais les prévisions multi-modèles. Les débits simulés des trois modèles hydrologiques sont également reportés sur les graphiques, de manière à illustrer les différences de modélisation hydrologique. Il peut arriver que les trois modèles soient très proches (cas **d** ou **h** par exemple), mais bien souvent au moins un des modèles s'écarte assez significativement des autres, ce qui augmente mécaniquement la dispersion des prévisions multi-modèles par rapport aux prévisions obtenues avec GRP seulement. Cependant, la non-réaction des modèles hydrologiques aux forçages météorologiques reste très fréquente sur les premiers pas de temps, ce qui entraine des regroupements des traces en autant de « paquets » que de modèles hydrologiques (cas **c**, **j** et **l** par exemple).

Si l'on généralise à l'ensemble de l'échantillon, cela se traduit par des histogrammes de rang aux formes inhabituelles, comme en témoigne la Figure 7.4 qui montre un très grand nombre de cas où l'observation tombe entre ces paquets de membres. La fiabilité est-elle améliorée pour autant ? S'il est difficile d'en juger, il ne fait en revanche pas de doute que notre approche multi-modèle est insuffisante pour prendre en compte l'incertitude hydrologique complète.

Regardons néanmoins si cela permet d'améliorer la qualité des prévisions. La Figure 7.5 présente le CRPSS des prévisions multi-modèles avec comme référence les prévisions obtenues avec le modèle GRP seulement. Pour l'ensemble des bassins, on constate que l'approche multi-modèle apporte un gain significatif. La raison pour laquelle le gain est plus important pour certains bassins (le Séran par exemple) que pour d'autres (comme la Valserine) n'a pas été identifiée. Il semble en tout cas que ces gains soient décorrélés des performances de simulation hydrologique de chaque modèle. En effet, la Figure 3.12 présentées dans le chapitre sur les modèles hydrologiques montre des performances pour les modèles GRP, TOPMODEL et ARX très similaires sur les bassins de la Valserine et du Séran. En revanche, il est possible que les trois modèles hydrologiques soient davantage complémentaires sur le Séran qu'ils ne le sont sur la Valserine. Cette hypothèse reste cependant à vérifier.

Ainsi, notre approche multi-modèle permet d'améliorer la qualité globale des prévisions, mais est insuffisante pour capturer entièrement l'incertitude hydrologique. Une étape de post-traitement statistique est donc nécessaire. Par ailleurs, les membres des prévisions de débit ne sont plus équiprobables car les performances de simulation de chacun des modèles ne sont pas équivalentes. Ainsi, nous proposons dans la section qui suit une méthode de post-traitement particulièrement adaptée aux prévisions multi-modèles, la BMA.

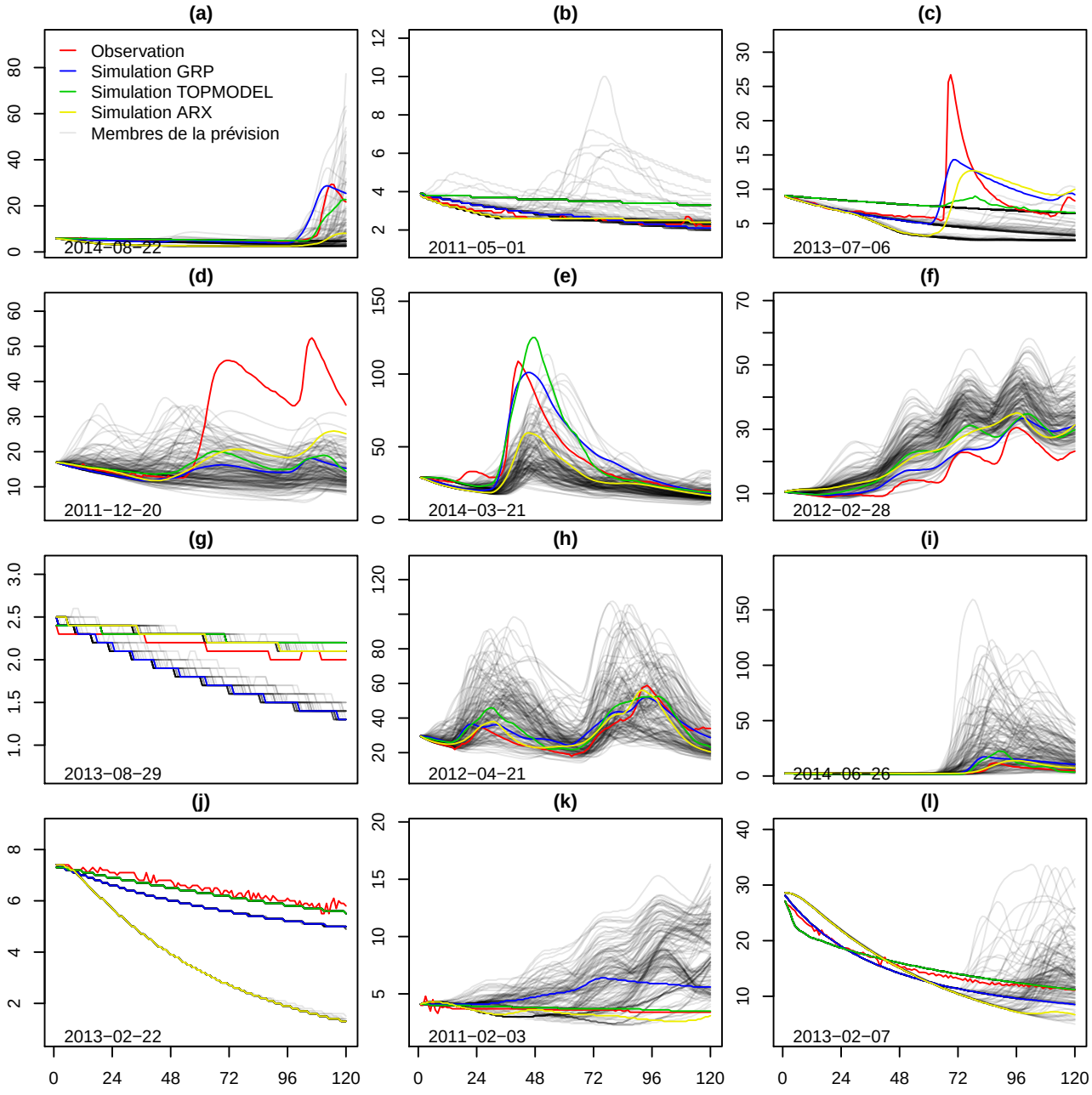

FIGURE 7.3 – Mêmes exemples de prévision qu'à la Figure 7.1, mais avec l'approche multi-modèles.

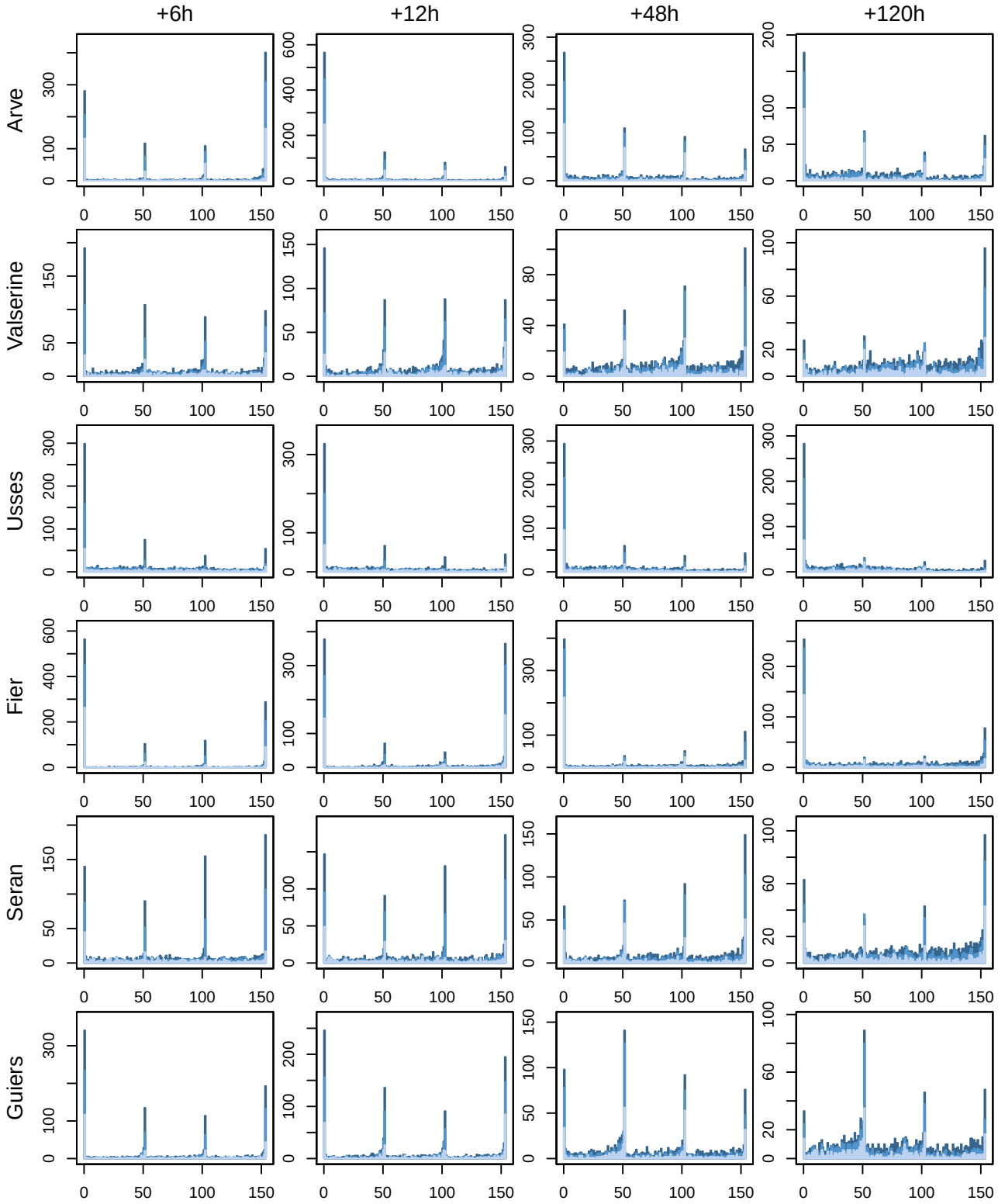

FIGURE 7.4 – Histogrammes de rang des prévisions de débit multi-modèles.

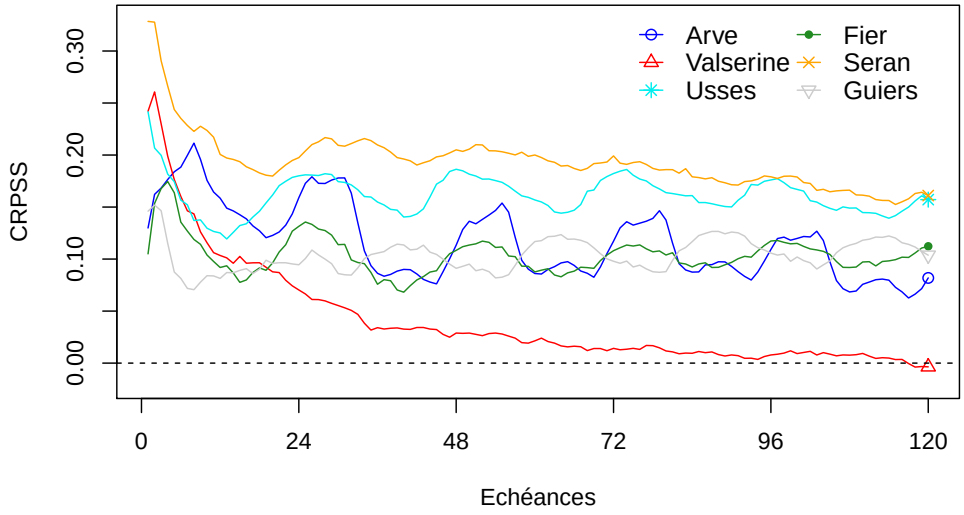

FIGURE 7.5 – CRPSS des prévisions de débit multi-modèles pour chacun des bassins. Les prévisions de référence sont les prévisions obtenues en sortie du modèle GRP seulement.

7.3 Post-traitement via la BMA

7.3.1 Présentation

La BMA (*Bayesian Model Averaging*, Raftery *et al.*, 2005) fait partie des méthodes de correction statistique par habillage des membres. Contrairement à l'EMOS où la densité prédictive est une densité paramétrique unique, les méthodes par habillage des membres proposent une densité prédictive sous la forme d'une densité de mélange (en anglais : *mixture density*). Cette dernière combine plusieurs densités paramétriques, appelées noyaux, qui « habillent » chacune un membre de l'ensemble. Il convient de noter que les méthodes par habillage des membres permettent des densités prédictives multi-modales, chose que ne permet pas l'EMOS.

Roulston et Smith (2003), Wang et Bishop (2005) et Fortin *et al.* (2006) ont travaillé sur l'approche dite du « meilleur membre », qui consiste à déterminer les caractéristiques des noyaux (notamment leur variance) à partir des erreurs du meilleur membre de l'ensemble, c'est-à-dire celui qui se trouve le plus proche de l'observation pour un cas de prévision donné. La méthode BMA proposée par Raftery *et al.* (2005) s'en inspire, mais plutôt que de chercher le meilleur membre, elle pondère les membres selon leur probabilité d'être le meilleur. C'est cette pondération qui la rend attractive pour la correction des prévisions multi-modèles où les membres ne sont plus échangeables.

Décrivons maintenant son fonctionnement plus en détail. Soit $x_1 \ldots, x_{\widehat{M}}$ une prévision d'ensemble de $\widehat{M}$ membres non échangeables. Nous omettons ici les indices du bassin et de l'échéance, le lecteur devant garder à l'esprit que la correction est univariée. La BMA considère que la densité prédictive $p(y|x_1, \ldots, x_{\widehat{M}}, D)$ de l'observation y sachant la prévision $x_1 \ldots, x_{\widehat{M}}$ et les données de calage D peut s'exprimer, selon la formule des

probabilités totales, par :

$$p(y|x_1,\dots,x_{\widehat{M}},D)=\sum_{m=1}^{\widehat{M}}p(x_m|D)\ p(y|x_m,D). \tag{7.1}$$

Le terme $p(x_m|D)$ est la vraisemblance du membre x_m étant correct, sachant les données de calage D. Il reflète la performance du membre x_m tel que $\sum_{m=1}^{\widehat{M}}p(x_m|D)=1$, et peut donc être assimilé à un terme de pondération, généralement noté ω_m. Le terme $p(y|x_m,D)$, appelé noyau, est la densité à posteriori de y sachant le membre x_m et les données de calage D.

Ce noyau s'exprime à l'aide d'une loi paramétrique, qui doit être capable de représenter de manière adéquate le comportement attendu de l'observation. Dans notre cas, la variable à prévoir est le débit horaire, qui prend des valeurs non négatives et présente une assymétrie forte avec une queue plus allongée dans les valeurs fortes. Plutôt que de chercher la loi adéquate comme nous l'avons fait pour l'EMOS, nous adoptons l'approche classiquement utilisée en hydrologie qui consiste à transformer la variable débit de manière à la rendre approximativement normale. Cette transformation sera détaillée dans la section 7.3.2.

Pour le moment, faisons l'hypothèse que cette transformation a déjà été réalisée. Par conséquent, nous supposons que :

$$y|x_m,D\sim\mathcal{N}(\mu_m,\sigma_m^2). \tag{7.2}$$

La variance σ_m^2 pour $m\in\{1,\dots,\widehat{M}\}$ est un paramètre qui, comme ω_m, reflète la performance du membre x_m sur les données de calage D. La moyenne μ_m est quant à elle fixée de manière à corriger un éventuel biais systématique du membre x_m. Dans notre cas, nous choisissons une correction à l'aide d'un terme additif :

$$\mu_m=a_m+x_m, \tag{7.3}$$

où a_m pour $m\in\{1,\dots,\widehat{M}\}$ est un paramètre. Les versions de BMA communément rencontrées dans la littérature proposent bien souvent une correction du biais à l'aide d'une régression linéaire de type $\mu_m=a_m+b_mx_m$. Nous avons cependant obtenu, comme Hemri *et al.* (2013), de moins bons résultats avec ce type de correction, et par conséquent nous conservons la correction seulement additive.

Plaçons-nous maintenant dans un contexte de prévision hydrologique multi-modèle. Soit I le nombre de modèles hydrologiques, et M le nombre de membres dans le forçage météorologique. Les $\widehat{M}=I\times M$ membres de la prévision multi-modèles sont notés $x_{1,1},\dots,x_{I,M}$ [2]. Au sein de cet ensemble, les membres $x_{i,1},\dots,x_{i,M}$ sont considérés comme échangeables car provenant d'un même modèle $i\in\{1,\dots,I\}$. La formulation de la densité

2. Pour faciliter la lecture, nous préférons cette notation à celle plus rigoureuse $x_{1,1},\dots,x_{1,M},x_{2,1},\dots,x_{2,M},\dots,x_{I,1},\dots,x_{I,M}$ où les indices i et m incrémentent séparément.

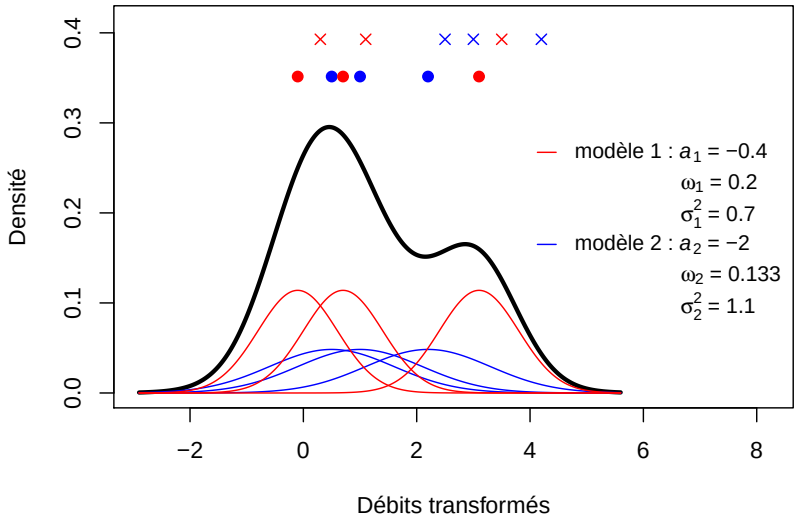

FIGURE 7.6 – Exemple d'une densité prédictive issue de la BMA (en noir) dans un contexte où $I = 2$ et $M = 3$. Les $2 \times 3 = 6$ membres de la prévision brute sont représentés par les croix, tandis que les membres débiaisés sont représentés par les ronds pleins. Les noyaux pondérés qui « habillent » ces membres débiaisés sont tracés en dessous. Ici, le modèle 1 est supposé meilleur que le modèle 2, car $\omega_1 > \omega_2$ et $\sigma_1^2 < \sigma_2^2$.

prédictive de la BMA, proposée par Fraley *et al.* (2010), est alors la suivante :

$$p(y|x_{1,1}, \ldots, x_{I,M}) = \sum_{i=1}^{I} \sum_{m=1}^{M} \omega_i \, \mathcal{N}(a_i + x_{i,m}, \sigma_i^2), \tag{7.4}$$

avec $\sum_{i=1}^{I} \sum_{m=1}^{M} \omega_i = 1$. La Figure 7.6 illustre un cas pour $I = 2$ et $M = 3$.

Les paramètres a_i pour $i \in \{1, \ldots, I\}$ servent à corriger les biais systématiques, et sont estimés de manière à avoir $\mathbb{E}[y - (a_i + \frac{1}{M} \sum_{m=1}^{M} x_{i,m})] = 0$ sur les données de calage D. Les poids ω_i et variances σ_i^2 des noyaux pour $i \in \{1, \ldots, I\}$ sont ensuite estimés par maximisation du logarithme de la fonction de vraisemblance sur les données de calage D. Pour cela, Raftery *et al.* (2005) ont proposé l'utilisation de l'algorithme d'espérance-maximisation (EM) (en anglais : *expectation-maximization*), qui est toujours utilisé dans la plupart des versions implémentées de la BMA. Nous ne rentrerons pas dans les détails de cette procédure, et invitons le lecteur à se référer aux articles mentionnés ci-dessus. Notons que l'estimation des paramètres de la BMA par minimisation du CRPS (comme dans l'EMOS) serait très couteuse en temps de calcul, compte-tenu du fait que la densité prédictive est une densité de mélange et donc qu'une formulation analytique du CRPS ne peut être obtenue. Dans la suite, nous utilisons la version implémentée dans le package `ensembleBMA` de `R`.

Du fait de la transformation des données de débit nécessaire à l'usage de noyaux Gaussiens, les distributions prédictives issues de la BMA sont exprimées dans l'espace normal. Par conséquent, les quantiles qui en sont extraits doivent subir la transformation inverse, qui les ramènera dans l'espace réel au sein duquel ils s'expriment en m^3/s. La Figure 7.7 illustre, après transformation inverse, les distributions prédictives issues de la

BMA pour les mêmes 12 cas de prévisions qu'aux Figures 7.1 et 7.3.

Le calage a été réalisé bassin par bassin, échéance par échéance. Par ailleurs, nous avons choisi de ne pas utiliser, comme pour le pré-traitement, uniquement des données antérieures à chaque date de prévision. En effet, cela aurait encore raccourci la période de validation, qui pour les forçages était de 4 années (2011-2014). Ici, nous adoptons une démarche de validation croisée, en calant un modèle de BMA sur 3 années, puis en vérifiant les prévisions sur l'année retirée du calage. De plus, une segmentation saisonnière est réalisée de manière à différencier les périodes de hauts et bas débits. Par simplicité, cette segmentation est unique pour l'ensemble des bassins. Elle distingue les périodes comprenant les mois de novembre à février (NDJF), de mars à juin (MAMJ), et enfin de juillet à octobre (JASO). Ainsi, les prévisions correspondant par exemple à la période JASO 2013 sont issues du modèle de BMA calé à partir des couples prévision-observation sur les périodes JASO 2011, JASO 2012 et JASO 2014. En procédant ainsi pour toutes les saisons et années, nous conservons la période 2011-2014 pour la validation. Certes, cette démarche ne correspond pas à un réel contexte opérationnel, cependant nous l'avons jugé comme étant acceptable compte-tenu du fait que l'archive des prévisions en sortie de modélisation hydrologique est supposée relativement homogène [3].

Cette méthode BMA sera comparée en 7.4 à la méthode EMOS, et plus précisément la méthode EMOS-normal qui a déjà été décrite en 5.2.1. En effet cette dernière génère des distributions prédictives sous la forme d'une loi normale, et peut donc parfaitement être utilisée pour le post-traitement de prévisions de débit dès lors que les données ont été transformées. C'est une stratégie qui a notamment été entreprise par Hemri *et al.* (2015).

7.3.2 Transformation de la variable débit

Détaillons maintenant la transformation qui permet de rendre la variable débit approximativement normale. Une transformation de variable ne consiste ni plus ni moins qu'à re-exprimer les données dans un espace différent, avec les unités qui lui sont propres. Deux transformations sont classiquement utilisées en hydrologie pour transposer les données dans un cadre Gaussien.

La première est la transformation Box-Cox (Box et Cox, 1964), qui est une généralisation de la transformation logarithmique :

$$t_{\text{boxcox}}(u) = \begin{cases} \frac{u^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(\text{u}) & \text{si } \lambda = 0, \end{cases} \tag{7.5}$$

où u est la variable à transformer dans l'espace normale, et λ est un paramètre à estimer. De nombreux auteurs l'ont utilisé en préalable d'un post-traitement hydrologique, par exemple Duan *et al.* (2007); Hemri *et al.* (2013, 2015); Engeland et Steinsland (2014); Madadgar et Moradkhani (2014); Qu *et al.* (2017).

3. En effet, d'une part les forçages météorologiques sont pré-traités, ce qui réduit la non-homogénéité initialement causée par les changements fréquents de version des modèles atmosphériques, et d'autre part les modèles hydrologiques sont identiques sur toute la période.

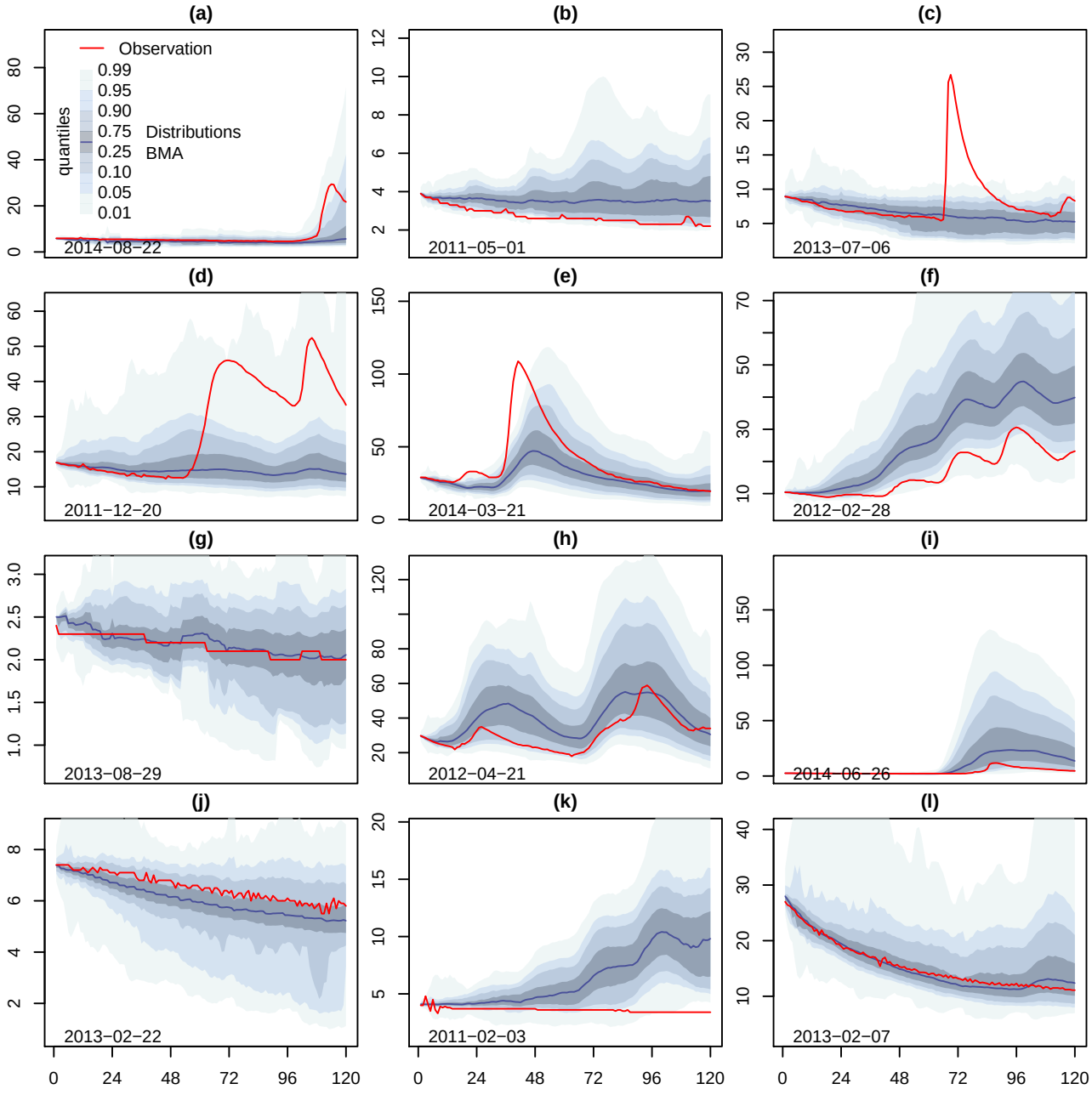

FIGURE 7.7 – Distributions prédictives obtenues par la BMA à partir des prévisions multi-modèles en entrée, pour les mêmes exemples de prévision qu'aux Figures 7.1 et 7.3.

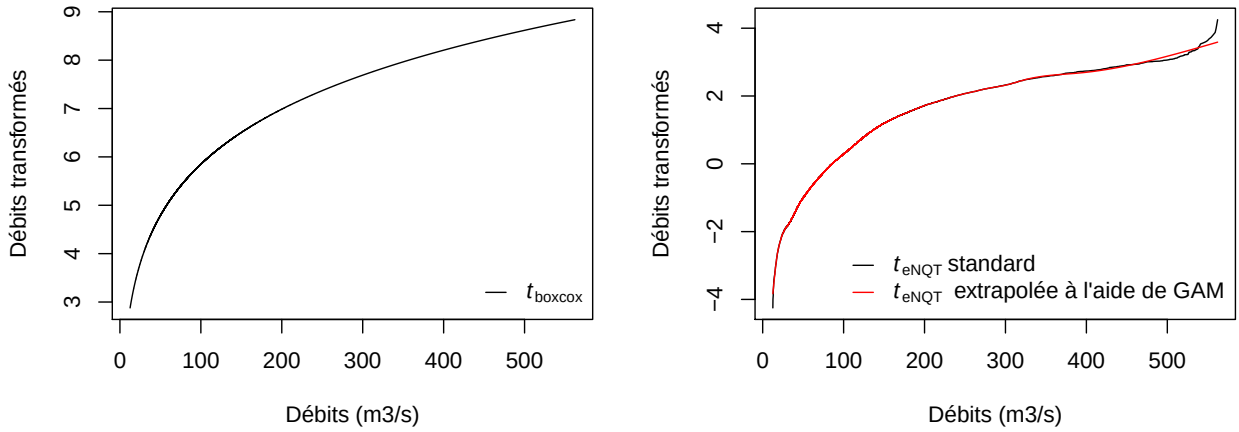

FIGURE 7.8 – Transformations Box-Cox (à gauche) et eNQT (à droite) sur le bassin de l'Arve pour la saison MAMJ. Les paramètres de ces deux transformations, respectivement λ et F_{emp}, sont estimés sur la période MAMJ 2003-2010.

La seconde est la transformation *empirical Normal quantile transform* (eNQT), proposée par Krzysztofowicz (1997). Celle-ci consiste à appliquer à la variable u sa propre fonction de répartition empirique, notée F_{emp}, puis la fonction de répartition inverse de la loi normale centrée réduite, Φ^{-1} :

$$t_{\text{eNQT}}(u) = \Phi^{-1}(F_{\text{emp}}(u)). \tag{7.6}$$

Cette transformation a été utilisée par de nombreux auteurs pour normaliser des débits en vue d'un post-traitement. Nous pouvons citer par exemple Krzysztofowicz (2002); Todini (2008); Reggiani *et al.* (2009); Bogner et Pappenberger (2011).

Les paramètres de ces deux transformations, λ et F_{emp} [4] sont propres à chaque bassin mais également à chaque saison, compte-tenu de notre segmentation saisonnière des paramètres de la BMA. Nous les estimons ici à partir de l'archive des observations, sur la période 2003-2010. Ces valeurs sont par conséquent indépendantes de l'échéance de prévision, et donc la même valeur sera utilisée pour chacune des échéances lors de la transformation des prévisions.

La Figure 7.8 trace en noir les transformations t_{boxcox} et t_{eNQT} obtenues sur le bassin de l'Arve pour la période MAMJ. Leur forme est similaire, à savoir une pente d'abord forte puis qui diminue à mesure que les débits augmentent, de manière à ramener à une amplitude constante des variations de débit qui croissent avec le débit lui-même. Nous observons cependant, pour t_{eNQT}, un comportement problématique dans les hauts débits, qui s'exprime ici par une augmentation soudaine de la pente. Cette « rupture » de t_{eNQT} se révèle très sensible aux quelques valeurs extrêmes mesurées durant la période de calage (de la transformation). De plus, le fait que la transformation ne soit définie que sur l'intervalle des valeurs observées durant le calage est particulièrement problématique lors de la transformation inverse, où les valeurs de débit de retour dans l'espace réel

4. Techniquement, F_{emp} est une fonction plus qu'un paramètre. Cependant, c'est bien elle qui définit la transformation eNQT, et par conséquent il n'est pas illogique de l'appeler « paramètre ».

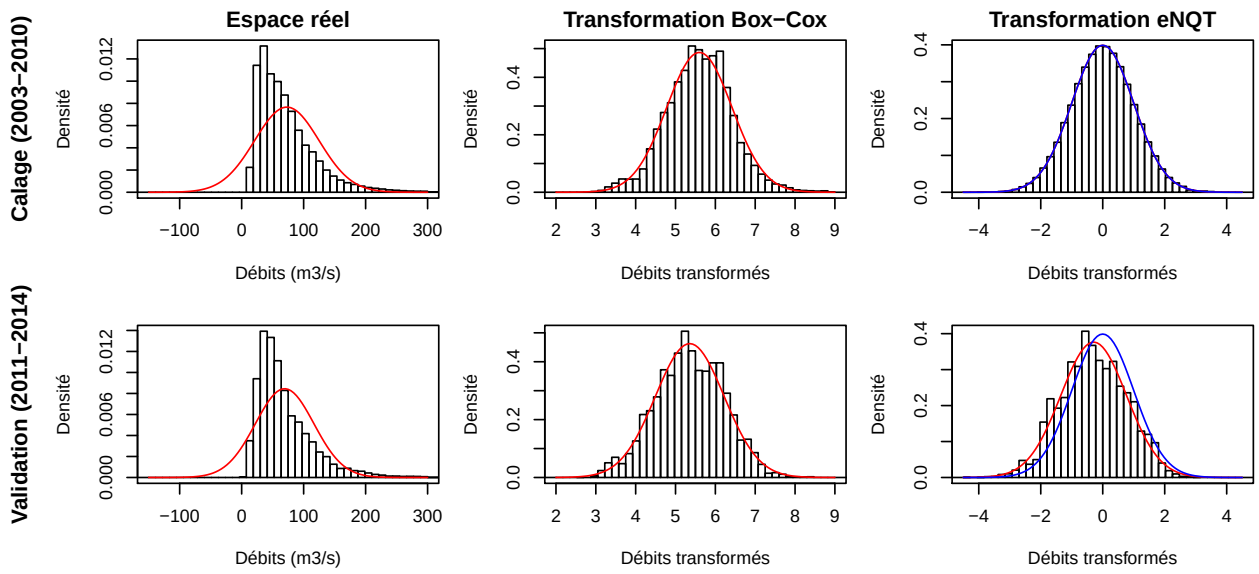

FIGURE 7.9 – Histogrammes des débits observés sur le bassin de l'Arve pour la saison MAMJ, dans l'espace réel (à gauche), transformé via Box-Cox (au centre) et transformé via eNQT (à droite), pour la période 2003-2010 sur laquelle les paramètres λ et F_{emp} sont calés (en haut), ainsi que la période 2011-2014 qui sert de validation (en bas). Les courbes rouge représentent la densité de la loi normale ajustée empiriquement sur l'échantillon de chaque histogramme. Les courbes bleues pour eNQT représentent la loi normale centrée réduite.

sont contraintes d'appartenir à cette gamme des valeurs observées. Ces problèmes ont été largement discutés par Bogner *et al.* (2012). Ces derniers proposent, pour contourner ces problèmes, d'extrapoler la fonction t_{eNQT} à l'aide de différentes méthodes, et notamment les modèles additifs généralisés (*generalized additive models* - GAM ; Hastie et Tibshirani (1987)) que nous proposons ici d'utiliser. Pour les détails techniques concernant cette extrapolation, nous nous référons à Bogner *et al.* (2012). La fonction t_{eNQT} extrapolée à l'aide de GAM est tracée en rouge sur le Figure 7.8. En plus de l'extrapolation, nous observons que la partie instable de la transformation a été lissée, et par conséquent sa robustesse à la période de calage a été améliorée.

Sur la Figure 7.9, qui trace les histogrammes d'un échantillon d'observations avant et après transformation, nous observons que le pouvoir de normalisation des transformations Box-Cox et eNQT (avec l'extrapolation GAM) est relativement stable en validation.

Laquelle des deux transformations choisir ? D'un côté, la transformation Box-Cox a l'avantage d'être paramétrique, ce qui l'exempte d'une extrapolation ad hoc comme pour l'eNQT. De l'autre, la transformation eNQT présente l'avantage de normaliser mais également de centrer-réduire les données de débit. Nous verrons plus tard en quoi cela peut être intéressant (cf. 8.4). Néanmoins, centrer-réduire des données déjà normalisées est trivial, et par conséquent cette propriété de l'eNQT n'est pas suffisante en soi pour que l'on se tourne vers cette transformation. Afin d'être en mesure de choisir, nous avons testé la BMA combinée avec l'une et l'autre transformation.

7.4 Résultats

Nous évaluons désormais les performances des prévisions de débit multi-modèles post-traitées, et les comparons à celles des prévisions brutes issues de GRP seulement ainsi que du multi-modèle. Par ailleurs, nous comparons également la méthode de post-traitement BMA à celle de l'EMOS, ainsi que la méthode de transformation des données eNQT à celle de Box-Cox.

Les résultats nous permettant de répondre à l'ensemble de ces questions sont présentés à la Figure 7.10, sous forme de CRPSS pour chacun des bassins et échéances. On constate tout d'abord que la BMA donne de meilleures performances que l'EMOS sur la majorité des bassins. Sur l'Arve, les résultats sont identiques pour les deux méthodes, tandis que le Fier est le seul bassin où l'EMOS se montre meilleur sur certaines échéances. Par ailleurs, contrairement à l'EMOS, la BMA apporte un gain de performances (par rapport au multi-modèle seul) sur la totalité des bassins. Cependant, on peut observer deux types de comportement différents. Sur la Valserine, les Usses et le Séran, ce gain tend à augmenter avec les échéances (il est même « négatif » sur les échéances inférieures à environ 24 h). Sur le Fier en revanche, ce gain diminue avec les échéances. Nous ne sommes pas parvenus à expliquer ces deux comportements.

La fiabilité des prévisions obtenues avec la BMA et l'EMOS est illustrée dans les Figures 7.11 et 7.12, respectivement. En sortie de BMA, les prévisions apparaissent comme assez fiables, exception faite des premières échéances sur le bassin des Usses. Il semble ici que le problème proviennent des très bas débits qui s'observent parfois sur ce bassin. Outre l'incertitude qui peut régner sur la mesure, ces bas débits compliquent l'étape de normalisation des données, et donc par conséquent l'étape de post-traitement. La fiabilité des prévisions en sortie de l'EMOS semble en revanche moins bonne. En effet, les prévisions sont sous-dispersives pour l'ensemble des bassins et des échéances.

Enfin, concernant les transformations, les performances de eNQT et Box-Cox s'avèrent très proches (Figure 7.10). Selon le bassin et la méthode de post-traitement, il se peut que l'une ou l'autre donne de meilleures performances ; cependant les résultats ne permettent pas d'établir un classement objectif. Ainsi, nous pouvons conclure que le choix de la transfromation eNQT ou Box-Cox a un impact très faible sur les performances des prévisions post-traitées.

Au regard de ces résultats, nous proposons d'utiliser, dans la suite de la thèse, la méthode BMA en combinaison de la transformation eNQT.

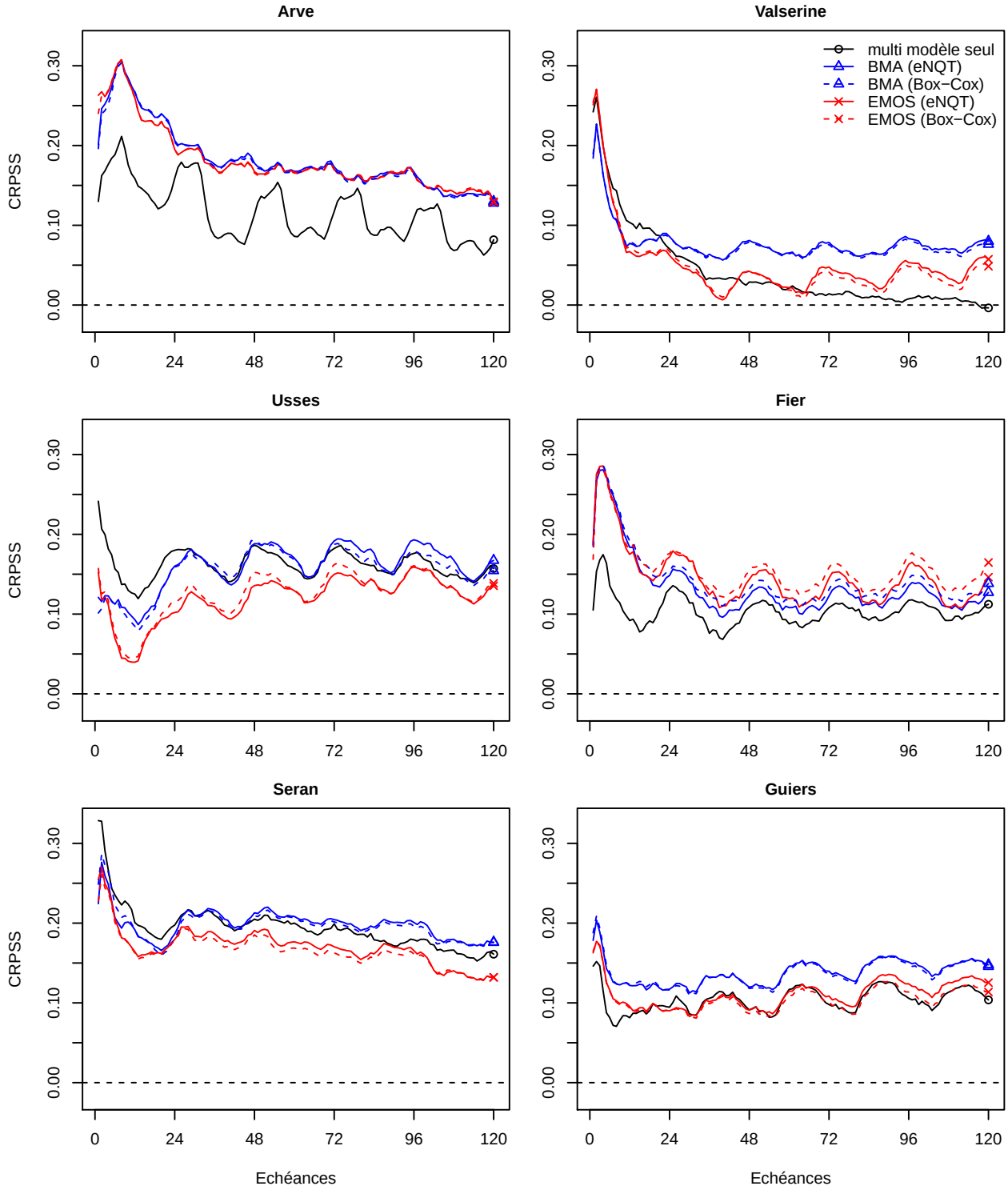

FIGURE 7.10 – CRPSS des prévisions de débit multi-modèles post-traitées à l'aide des méthodes BMA (en bleu) et EMOS (en rouge), en combinaison des méthodes de transformation eNQT (en trait plein) ou Box-Cox (en trait pointillé). Les prévisions de référence sont les prévisions brutes obtenues en sortie du modèle GRP seulement. Sur chacun des graphiques est également reporté le CRPSS des prévisions multi-modèles sans post-traitement (en trait plein noir).

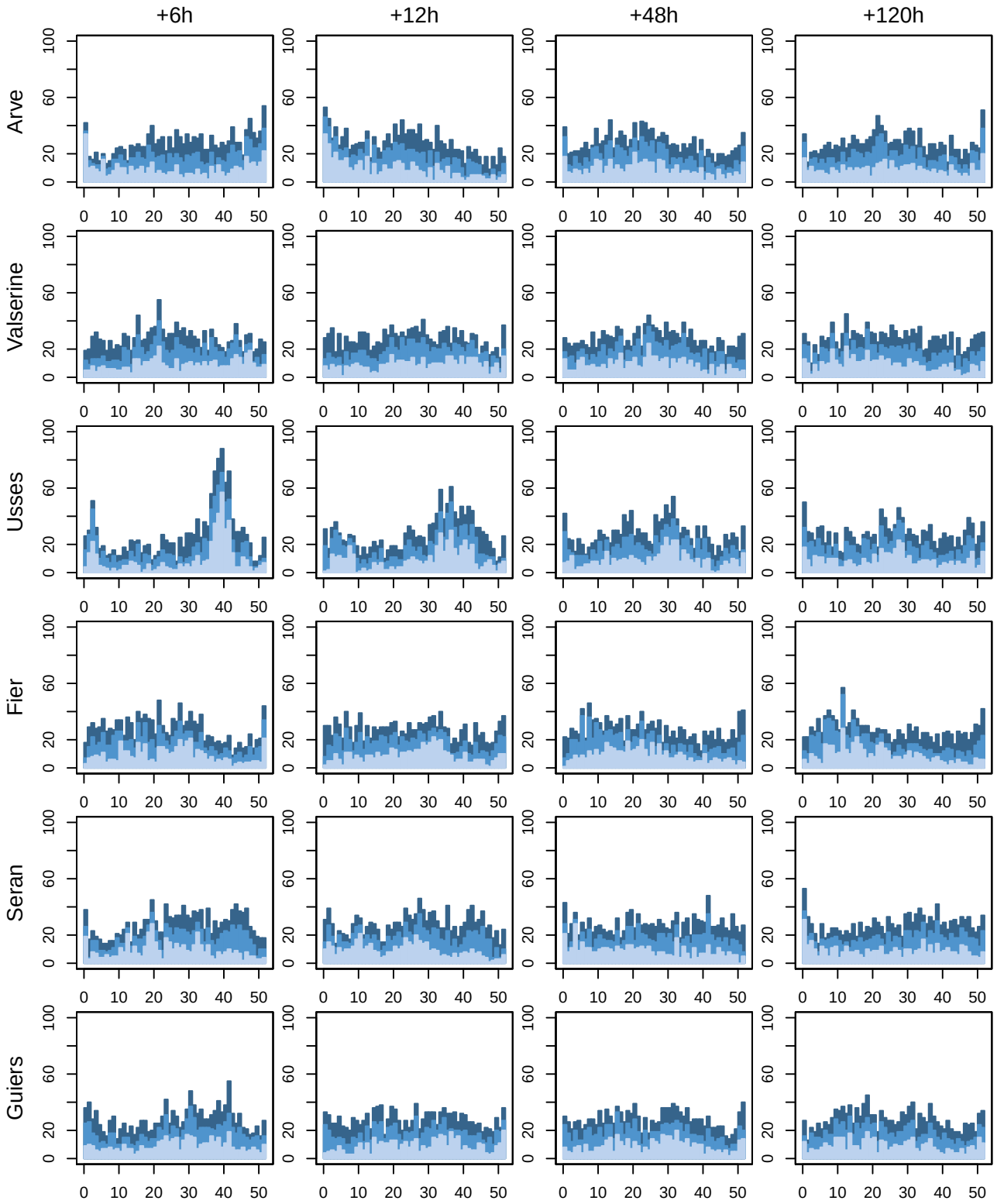

FIGURE 7.11 – Histogrammes de rang des prévisions de débit multi-modèles post-traitées à l'aide de la BMA, en combinaison de la transformation eNQT.

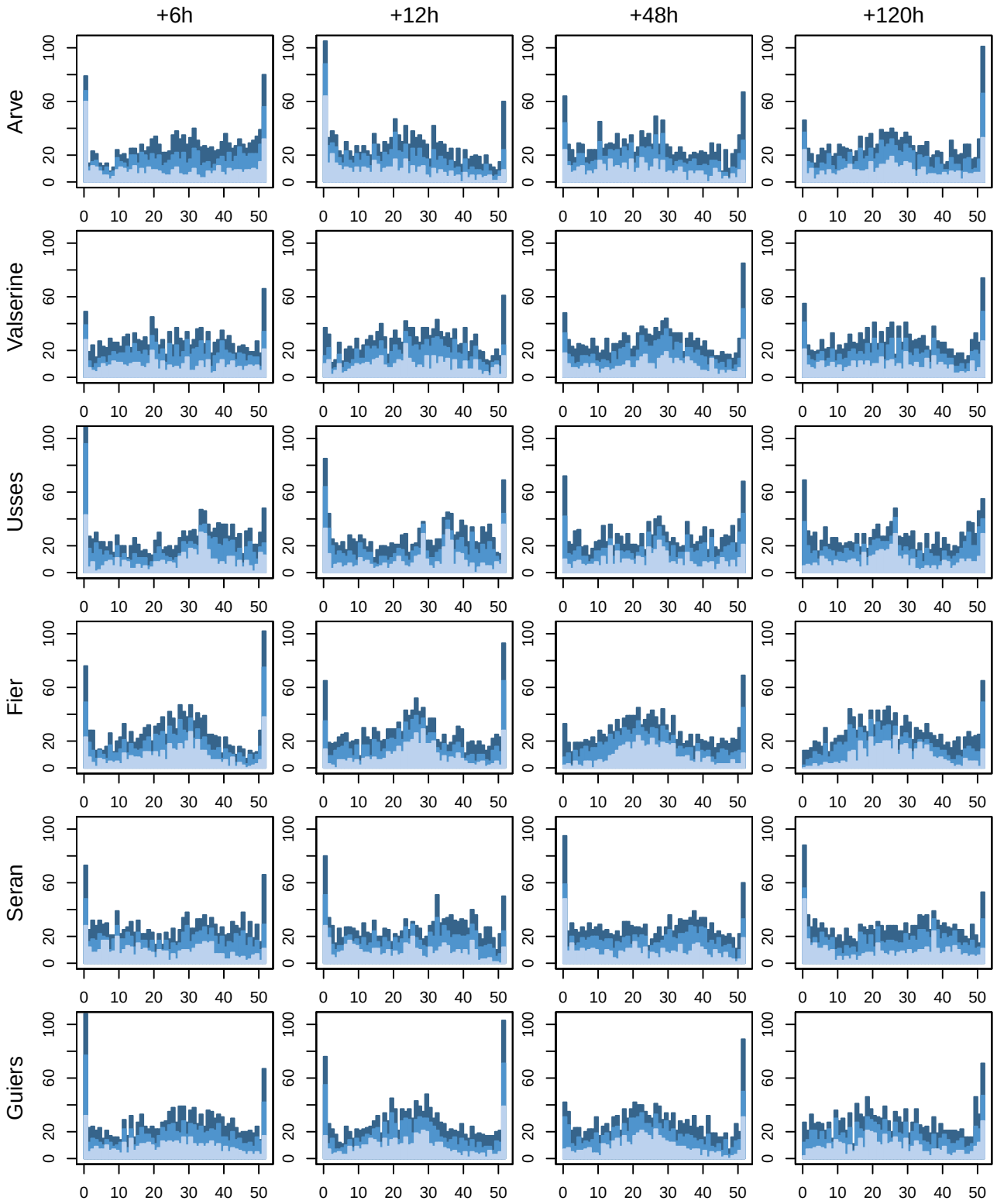

FIGURE 7.12 – Histogrammes de rang des prévisions de débit multi-modèles post-traitées à l'aide de l'EMOS, en combinaison de la transformation eNQT.

7.5 Synthèse

Le premier objectif de ce chapitre était de diagnostiquer les performances des prévisions de débit obtenues en propageant simplement des forçages météorologiques pré-traités dans le modèle hydrologique. Le second était d'étudier deux stratégies de prise en compte de l'incertitude de modélisation hydrologique, l'une ensembliste (le multi-modèle) et l'autre statistique (le post-traitement univarié). Les principales conclusions à retenir sont :

- Malgré le pré-traitement des forçages météorologiques, les prévisions de débit obtenues à l'aide d'un seul modèle hydrologique sont fortement sous-dispersives. La fiabilité des prévisions météorologiques n'entraine donc pas la fiabilité des prévisions hydrologiques.

- L'approche multi-modèle améliore les performances en prenant en compte, de manière non statistique, une partie de l'incertitude hydrologique. Cependant, un post-traitement statistique reste nécessaire pour obtenir des prévisions fiables.

- Pour ce post-traitement, la méthode BMA offre globalement de meilleures performances que l'EMOS, et produit des prévisions plus fiables.

Nous nous retrouverons à présent dans une situation similaire à celle évoquée en conclusion du chapitre 5 : en appliquant une (nécessaire) correction statistique, nous avons d'une part troqué des prévisions ensemblistes contre des prévisions exprimées sous le forme de densités continues, et d'autre part perdu la structure de dépendance spatio-temporelle que contenaient les prévisions brutes. Le chapitre suivant doit nous permettre de reconstruire des prévisions qui puissent être exploitées dans un contexte multivarié.

Chapitre 8

Reconstruction de prévisions hydrologiques cohérentes

Comme le pré-traitement, le post-traitement a été réalisé d'une part dans un cadre univarié, c'est-à-dire indépendamment pour chaque combinaison de bassin et d'échéance, et d'autre part à l'aide d'une méthode paramétrique, qui génère des distributions sous forme continue. Il est alors nécessaire de reconstruire, à partir de ces distributions marginales, des prévisions d'ensemble multivariées de débit qui soient cohérentes.

C'est l'objet de ce chapitre, qui correspond à un article publié dans la revue *Water Resources Research*, Vol. 54, no 8, p. 5741-5762.

Generating Coherent Ensemble Forecasts After Hydrological Postprocessing: Adaptations of ECC-based Methods

J. Bellier[1], I. Zin[1], and G. Bontron[2]

[1] Université Grenoble Alpes, Grenoble INP, CNRS, IGE, Grenoble, France
[2] Compagnie Nationale du Rhône, Lyon, France

(Manuscript received 19 Jan. 2018, accepted 4 May 2018, published online 28 Aug. 2018)

DOI: 10.1029/2018WR022601

Résumé

Les prévisions d'ensemble de débit ont généralement besoin d'être corrigées statistiquement de manière à rendre compte de l'incertitude de prévision totale. Dans la majorité des cas, ces méthodes de post-traitement sont paramétriques d'une part, et univariées d'autre part, dans le sens où elles ne tiennent pas compte des dépendances entre les bassins et les échéances. Il est alors nécessaire d'appliquer une procédure ad hoc appelée « échantillonnage-réarrangement », qui génère des ensemble multivariés cohérents à partir des distributions marginales issues du post-traitement. Déjà populaire dans le domaine du post-traitement météorologique, l'approche *ensemble copula coupling* (ECC) est attractive pour la prévision hydrologique, car elle préserve la structure de dépendance de l'ensemble brut supposé cohérent. Cependant, les manières existantes de mettre en œuvre

l'ECC montrent des lacunes lorsqu'elles s'appliquent à des variables telles que le débit au pas de temps horaire, du fait de l'auto-corrélation forte d'une part, et de l'absence fréquente de dispersion des prévisions d'autre part. En partant de ce diagnostic, cet article se propose d'explorer différentes variantes de l'ECC, et notamment l'addition d'une perturbation à l'ensemble brut de manière à le rendre dispersive, ainsi que le lissage des trajectoires de l'ensemble post-traité afin de les rendre plus réalistes. L'évaluation porte sur un cas d'étude de prévision hydrologique impliquant différents bassins du Haut-Rhône sur sa partie française. Les résultats montrent que les nouvelles variantes proposées améliorent les prévisions, tout en évitant une inflation de la complexité de la procédure d'échantillonnage-réarrangement.

Abstract

Hydrological ensemble forecasts are frequently miscalibrated, and need to be statistically postprocessed in order to account for the total predictive uncertainty. Very often, this step relies on parametric, univariate techniques that ignore the between-basins and between-lead times dependencies. This calls for a procedure referred to as sampling-reordering, which generates a coherent multivariate ensemble from the marginal postprocessed distributions. The ensemble copula coupling (ECC) approach, which is already popular in the field of meteorological postprocessing, is attractive for hydrological forecasts as it preserves the dependence structure of the raw ensemble assumed as spatially and temporally coherent. However, the existing implementations of ECC have strong limitations when applied to hourly streamflow, due to raw ensembles being frequently non-dispersive and to streamflow data being strongly autocorrelated. Based on this diagnosis, this paper investigates several variants of ECC, in particular the addition of a perturbation to the raw ensemble to handle the non-dispersive cases, and the smoothing of the temporal trajectories to make them more realistic. The evaluation is conducted on a case study of hydrological forecasting over a set of French basins. The results show that the new variants improve upon the existing ECC implementations, while they remain simple and computationally inexpensive.

8.1 Introduction

Hydrological forecasts are of great interest in a wide range of applications such as flood warning, hydropower forecasting or inland waterway transport operation. Over the last decades, the knowledge of the predictive uncertainty related to these forecasts has been shown to lead to more rational decision-making (Krzysztofowicz, 2001; Ramos *et al.*, 2013a). Operational systems are thus increasingly moving from deterministic toward probabilistic forecasting, employing most of the time the ensemble approach (Cloke et Pappenberger, 2009; Demeritt *et al.*, 2010). The typical procedure in many hydrological forecasting centers consists in forcing a hydrological model with meteorological ensemble forecasts. These comprise multiple runs of a numerical weather prediction model with slightly altered initial conditions and sometimes model assumptions (Buizza *et al.*, 1999). In addition to this meteorological uncertainty, systems may eventually account for errors

in initial catchment conditions via ensemble data assimilation techniques (DeChant et Moradkhani, 2011; Thiboult *et al.*, 2016), but also for errors in model parameters and structures, using for example multi-model approaches (Seiller *et al.*, 2012; Thiboult *et al.*, 2016; Velázquez *et al.*, 2011). However, it is highly frequent that the dispersion of the ensemble members is not sufficient to account for the total uncertainty, potentially in addition to systematic biases. Therefore, streamflow ensemble forecasts must generally be statistically postprocessed.

Examples of technique for statistical postprocessing are ensemble model output statistics (EMOS ; Gneiting *et al.*, 2005) or Bayesian model averaging (BMA ; Raftery *et al.*, 2005). Initially designed for meteorological postprocessing, EMOS and BMA have been applied to streamflow forecasts in several past studies (Duan *et al.*, 2007; Hemri *et al.*, 2013, 2015; Madadgar et Moradkhani, 2014). A broader overview of hydrological postprocessing methods can be found in Ramos *et al.* (2013b). In a large majority, these methods are parametric, in that the ensembles to be corrected are transformed into parametric probability distributions, and univariate, as they apply individually to each combination of lead time and basin (or, more generally, location). The postprocessing step thus conducts in a collection of independent marginal distributions. However, end-users may need to draw from these marginal distributions a finite ensemble of multivariate trajectories that accounts for between-basins and between-lead times dependencies. This is required, for instance, to drive hydrodynamic models, to estimate accumulated volumes across several lead times, or to compute probabilistic forecasts at the outlet of complex water systems involving a number of basins. The so-obtained ensemble must preserve the univariate skill gained through postprocessing, but also capture the multivariate dependence structure adequately. In addition, the individual trajectories must be realistic regarding the strong autocorrelation of streamflow, especially at the hourly time step. While hydrological postprocessing has been extensively studied in the literature, very few studies (Engeland et Steinsland, 2014; Hemri *et al.*, 2013, 2015) have concerned the generation of coherent multivariate streamflow ensembles. This multivariate extension, referred to as sampling-reordering, is the core subject of this paper.

The challenge can be addressed by modeling a parametric correlation function that links all the marginal distributions and creates a multivariate distribution within which one may sample trajectories. We refer for instance to the Gaussian copula approach (GCA) proposed by Pinson et Girard (2012) for wind power forecasting, and applied to streamflow forecasts by Hemri *et al.* (2015). Meanwhile, non-parametric methods have emerged that consist in sampling ensemble values independently from each marginal distribution, before reordering them such that they respect the rank dependence structure of a specified template. We distinguish the ensemble copula coupling (ECC, Schefzik *et al.*, 2013) technique, in which the dependence template is the raw (i.e., unprocessed) ensemble forecast, from the Schaake shuffle (Clark *et al.*, 2004) approach and adaptations thereof (Bellier *et al.*, 2017b; Schefzik *et al.*, 2013; Schefzik, 2016a; Scheuerer *et al.*, 2017; Wu *et al.*, 2018) where past observations are used to specify the template.

The ECC approach, which adapts the dependence structure to the current meteorological pattern, is particularly attractive to the field of streamflow postprocessing. Indeed, the

raw streamflow ensemble trajectories are assumed to be spatially and temporally coherent, as a result of meteorological forcings being themselves coherent, which is a prerequisite for hydrological modeling. A distinction is made by Schefzik *et al.* (2013) between ECC-Q, ECC-R and ECC-T, the last letter referring to different sampling schemes: equidistant quantiles, random draws, or transformations, respectively. They have found ECC-Q to give the best results, due to a better depiction of the marginal distributions. However, the method yields unrealistic jumps in the trajectories when applied to strongly autocorrelated data such as hourly streamflow, which has led Hemri *et al.* (2015) to use ECC-T instead. As a simpler solution, the procedure of smoothing the trajectories of ECC-Q has, to the knowledge of the authors, not been studied yet. Furthermore, neither ECC-Q nor ECC-T handles satisfactorily the cases of non-dispersive raw ensembles, although it occurs very frequently in streamflow forecasting when precipitation forcings have not been sufficient for the hydrological model to react. Alternatively, Ben Bouallègue *et al.* (2016) have proposed a version referred to as dual-ECC that extents ECC to account for autocorrelation of the forecast errors, but there is no reference in the literature concerning an application to streamflow forecast.

We therefore focus in this paper on ECC-based methods for generating coherent multivariate streamflow ensemble forecasts, from any collection of distributions that results from univariate hydrological postprocessing. In our case study, these distributions result from BMA, but any other univariate postprocessing method such as EMOS could be used instead. We define ECC-Q as the standard method, and explore several variants on the basis of specific shortcomings. They include the aforementioned ECC-T and dual-ECC, but also novel procedures such as trajectory smoothing and template perturbation. These are evaluated on a real case study of streamflow forecasting over a set of French basins. The parametric GCA is considered as the benchmark, in order to relate ECC-based methods to an approach that assumes a stationary multivariate dependence structure.

The paper is organized as follows. Section 8.2 presents the data and the setup for this study. In section 8.3, we discuss the requirements that a sampling-reordering method should fulfil, and suggest corresponding verification metrics. The GCA benchmark and the different ECC variants are developed in section 8.4. Section 8.5 presents the results of their evaluation, and section 8.6 concludes.

8.2 Data and setup of the study

Throughout the paper, let the index $j \in \{1, \dots, J\}$ refer to basins, $k \in \{1, \dots, K\}$ to lead times, $m \in \{1, \dots, M\}$ to meteorological members and $i \in \{1, \dots, I\}$ to hydrological models. Products of the postprocessing step are denoted using the tilde symbol.

8.2.1 Study basins and streamflow data

We consider in this study $J = 6$ basins, namely the Arve, Valserine, Usses, Fier, Séran, and Guiers basins, located in the upper Rhone River region in France. Figure 8.1 displays their location and a picture of the relief, and Table 8.1 provides some hydrological cha-

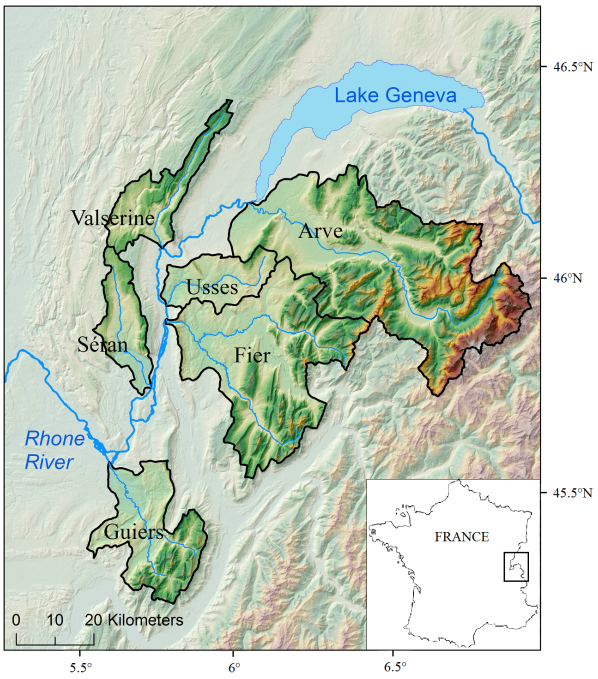

FIGURE 8.1 – Location of the study basins.

racteristics. They all contribute to the Rhone River, on which Compagnie Nationale du Rhône (CNR) operates a series of hydropower plants, and has developed a forecasting chain (Bompart *et al.*, 2009) for the need of both electricity production forecasting and industrial and public safety. As these basins are part of a complex water system, it is crucial to correctly model the spatial and temporal dependencies in the forecasts. Hourly streamflow records at each of the gauging stations indicated in Table 8.1 were obtained from CNR. The 2003-2010 period was used for calibrating and validating the hydrological models, while the 2011-2014 period is set for evaluating the sampling-reordering methods.

8.2.2 Hydrological modeling

The archive of hourly streamflow forecasts was set up as follows. Ensemble forecasts of mean areal precipitation (MAP) and temperature (MAT) come from the European

TABLE 8.1 – Characteristics of the study basins

Basin	Area (km^2)	Stream gauge	MQ ($m^3\ s^{-1}$)[a]	Q90 ($m^3\ s^{-1}$)[a]
Arve	2082	Bout du Monde	68	124
Valserine	361	Lancrans	14	32
Usses	309	Pont Rouge	3	9
Fier	1375	Motz	33	80
Séran	290	Pont de la Thuilliere	6	14
Guiers	609	Belmont	17	36

[a]MQ (mean streamflow) and Q90 (90-percentile streamflow) are estimated over the longest common period, that is the 2003-2014 period.

Centre for Medium-Range Weather Forecasts (ECMWF) and have been downloaded from the TIGGE archive (Park *et al.*, 2008). They comprise $M = 51$ members, which have been considered in the study as exchangeable. The 00 UTC cycle is considered, with lead times up to 120 h ahead. These forecasts were postprocessed over the 2011-2014 period using, for MAP, a simplified version of the EMOS variant based on censored, shifted Gamma distributions (Scheuerer et Hamill, 2015a) and, for MAT, the standard EMOS technique (Gneiting *et al.*, 2005). The ECC-Q technique (Schefzik *et al.*, 2013) was then applied to recover coherent multivariate ensemble forecasts. More information relative to this meteorological postprocessing step can be found in Bellier *et al.* (2017b).

Streamflow forecasts were then obtained by driving a modeling system that comprises a snow accounting model, Cemaneige (Valéry *et al.*, 2014), coupled with $I = 3$ different hydrological models. The first model is the well-known variable contributing area TOPMODEL (Beven, 2012). The second one, GRP (Berthet *et al.*, 2009), is a parsimonious reservoir model, designed specifically for flood forecasting. Finally, ARX (Remesan et Mathew, 2015) is a fully statistical model that estimates streamflow values based on regression equations with previous streamflow and MAP as inputs, and with coefficients segmented by model state parameters. As implemented here, the three models are driven by MAP and MAT forcings, and assimilate the latest streamflow observations in order to reduce forecast errors on the first lead times. Forecasts are launched at 00 UTC, with an hourly time step and a maximum lead time of $K = 120$ h. Models were calibrated over the 2003-2006 period and validated over the 2007-2010 period. Validation Nash et Sutcliffe (1970) efficiency (NSE) coefficients under the perfect forecast mode (i.e., with observed forcings and with the assimilation) can be found in the supporting information Figure B.1. It is found that GRP outperforms TOPMODEL and ARX for almost all basins and all lead times. Nonetheless, the multimodel hydrological system runs systematically the three models, as a way to tackle a part of the hydrological uncertainty by considering diverse modeling assumptions and structures, without prior assumptions on which model performs best under specific conditions. Streamflow members that originate from the same model are considered as exchangeable.

Complete multivariate streamflow ensemble forecasts are denoted by $\mathbf{f} = (\mathbf{f}^{1,1}, \ldots, \mathbf{f}^{J,K})$, where each univariate ensemble $\mathbf{f}^{j,k} = (f^{j,k}_{1,1}, \ldots, f^{j,k}_{I,M})$ comprises I groups each having M exchangeable members. An example of such a forecast is given in Figure 8.2a, for the date of 26 October 2013 and the Guiers basin. This example will be used throughout the paper, and simply referred to as the forecast example. It has been selected as it illustrates a hydrological event with an interesting temporal structure (two peaks seem plausible), preceded by a series of lead times where the ensemble is non-dispersive. Supporting information Figures B.2-B.6 propose additional examples for randomly selected dates.

8.2.3 Streamflow postprocessing

For streamflow postprocessing, we applied the Bayesian model averaging (BMA) approach, which was introduced by Raftery *et al.* (2005) and adapted by Fraley *et al.* (2010) to deal with groups of exchangeable members. BMA has been considered for streamflow

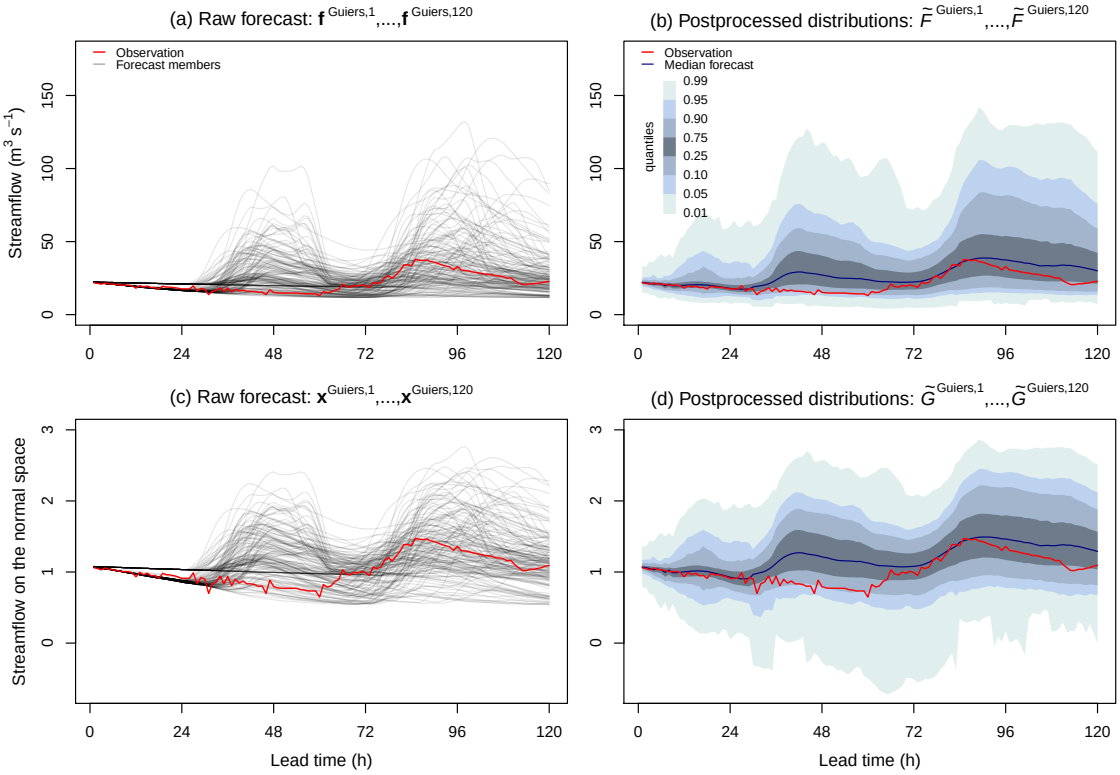

FIGURE 8.2 – Forecast example for the Guiers basin and the date 26 October 2013. (a) Raw forecast on the original space and (c) on the normal space, after transformation. The ensemble size is $3 \times 51 = 153$. (d) Quantiles of the BMA postprocessed distributions on the normal space and (b) on the original space, after backtransformation.

postprocessing in several past studies (Duan *et al.*, 2007; Hemri *et al.*, 2013; Madadgar et Moradkhani, 2014; Qu *et al.*, 2017). The next paragraphs give a brief description of the procedure, but as it is not the core of the paper we encourage interested readers to read the above references.

First of all, the implementation of BMA is more convenient under the assumption of Gaussianity. Since streamflow data are certainly non-Gaussian, a widely used strategy consists in transforming the streamflow data from their original space to a normal space prior to BMA. For this purpose, we employ the empirical Normal quantile transformation (eNQT, Krzysztofowicz, 1997) defined as

$$t(u) = \Phi^{-1}\left(F_{\text{emp}}(u)\right) \tag{8.1}$$

with Φ^{-1} denoting the inverse of the cumulative distribution function (CDF) of the standard normal distribution, and F_{emp} the empirical CDF of a given variable $u \in \mathbb{R}$, estimated over a sample of realizations. Here, the transformation $t(\cdot)$ is defined for each basin by estimating F_{emp} from their respective observation records, and applied prior to BMA on both the raw forecasts and the observations. Setting $x_{i,m}^{j,k} = t(f_{i,m}^{j,k})$ for all $\{j, k, i, m\}$, the raw streamflow forecasts after transformation on the normal space can be denoted by $\mathbf{x} = (\mathbf{x}^{1,1}, \ldots, \mathbf{x}^{J,K})$ with $\mathbf{x}^{j,k} = (x_{1,1}^{j,k}, \ldots, x_{I,M}^{j,k})$. This transformed forecast is displayed in Figure 8.2c for the forecast example.

As BMA applies individually to each combination of basin and lead time, let us drop temporarily the indices j and k. Consider a univariate variable of interest, y, and let $x_{1,1}, \ldots, x_{I,M}$ be the corresponding ensemble forecast. Assuming that y is Gaussian, the BMA predictive probability distribution function (PDF) $p(y|x_{1,1}, \ldots, x_{I,M})$ is a weighted average of normal PDF centered on the bias-corrected forecast members $x_{i,m}$, that is:

$$p(y|x_{1,1}, \ldots, x_{I,M}) = \sum_{i=1}^{I} \sum_{m=1}^{M} \omega_i \mathcal{N}(a_i + x_{i,m}, \sigma_i^2), \tag{8.2}$$

where $\mathcal{N}(a_i + x_{i,m}, \sigma_i^2)$ denotes the PDF of a normal distribution with mean $a_i + x_{i,m}$ and variance σ_i^2. The formulation (8.2) is suggested by Fraley *et al.* (2010) to accommodate with groups of exchangeable members. The weights ω_i and variances σ_i^2 reflect how well forecasts from the model i fit the training data, and are estimated by maximum likelihood using the Expectation-Maximization algorithm. The bias-correction parameter a_i is estimated such that $\mathbb{E}[y - (a_i + \frac{1}{M}\sum_{m=1}^{M} x_{i,m})] = 0$ over the training data. Note that, as in Hemri *et al.* (2013), we have considered a simple additive bias correction, as was found to outperform the linear correction suggested in Gneiting *et al.* (2005).

Finally, the BMA approach applied to all the combinations of basins and lead times yields a collection of independent marginal CDF denoted by $\{\tilde{G}_{1,1}, \ldots, \tilde{G}_{J,K}\}$, which are on the normal space. These are represented in Figure 8.2d for the forecast example, via specific quantiles. Their counterparts after backtransformation to the original space are denoted by $\{\tilde{F}_{1,1}, \ldots, \tilde{F}_{J,K}\}$, and similarly represented in Figure 8.2b. To leverage in the most effective way the 2011-2014 period during which the streamflow forecasts are available, we considered the following cross-validation strategy. Each year was divided into three seasons (defined as the 4-month periods NDJF, MAMJ and JASO), and the forecasts over a specific year and season were postprocessed using as training the forecasts over the other three years, but in the same season. This way, the verification period remains 2011-2014.

When applying the inverse eNQT function $t^{-1}(\cdot)$ to backtransform the streamflow data into the original space, it is possible that few sampled quantiles fall outside the range of the historical observations used for estimating F_{emp}. Bogner *et al.* (2012) have discussed this problem, and we followed the approach they suggest which consists in extrapolating the functions t^{-1} of each basin using Generalized additive models (GAM). It must be noted that GAM have been necessary in a couple of cases only, due to the relatively long observation series used for specifying the distributions F_{emp}. As an alternative to the eNQT, the Box-Cox transformation (Box et Cox, 1964) has been used by Duan *et al.* (2007), Hemri *et al.* (2013, 2015), Madadgar et Moradkhani (2014) and Qu *et al.* (2017), but was found to perform slightly worse than eNQT in our case study.

This concludes the section describing the setup of the study. As the core subject of the paper, we now focus in generating postprocessed streamflow ensemble forecasts that account for between-basins and between-lead times dependencies, calling for methods referred to as sampling-reordering. The methodological aspects that follow can therefore be understood regardless of the technique used for postprocessing, as long as it ends with

a collection of continuous and univariate distributions.

8.3 Requirements and verification tools

In this section, we discuss the requirements that a sampling-reordering method applying to streamflow forecasts should fulfil. We suggest three criteria, with corresponding verification tools:

Autocorrelation criterion: Each individual trajectory of the multivariate ensemble must ensure adequate representation of serial dependence (i.e., across lead times). Besides the desire of trajectories to be visually realistic, this is of particular interest if forecasts aim at forcing a hydrodynamic model. Unrealistic streamflow gradients may indeed cause numerical issues, but also lead the model to trigger some processes such as river bank overflow that strongly affect the river stage modeling. For verification, we suggest to compare the autocorrelation function (ACF) of the streamflow forecast trajectories against a reference ACF. The notion of ACF is based on the assumption of stationarity of the time series at hand, which is questionable when dealing with streamflow series. We therefore suggest a process known in time series analysis as differencing, which consists in considering the series of lag-one differences, with the aim of stabilizing the mean and thus eliminating trends and seasonality (Hyndman et Athanasopoulos, 2012). As we deal with multiple series samples (M trajectories $\times$ the number of forecast cases within the verification period), we concatenate the different samples while ensuring that pairs from different samples are excluded from the ACF calculation. It is still possible that the stationarity assumption of lag-one differenced streamflow series is still violated to a certain extent. However, we consider that it as acceptable as long as we restrict to compare ACF of series that cover the exact same period, and differ to a small extent compared to the streamflow seasonal variability. An important question concerns the data used for setting the reference ACF. The most logical choice would be to consider observed time series, but it happens that these series fluctuate to a certain extent due to gauging uncertainty, as the red line in Figure 8.2 shows. On the one hand, it is likely that hydrodynamic models are implemented such that they handle these fluctuations, but on the other hand it is meaningless to design sampling-reordering methods for reproducing these fluctuations. Thus, we consider for setting the reference ACF the streamflow forecasts before their postprocessing, assuming that the hydrological model generates realistic streamflow trajectories.

Univariate criterion: Consider now forecasts for a given basin and lead time. The goal of postprocessing is to produce well calibrated and yet sharp predictive distributions (Gneiting *et al.*, 2007). Calibration is a joint property of the forecasts and the observations and refers to the statistical consistency between the two, while sharpness refers to the dispersion of the forecasts, hence it is a property of the forecasts only (Jolliffe et Stephenson, 2003). In the case of parametric postprocessing, the available predictive distributions are continuous, and thus drawing discrete ensembles within these distributions leads inevitably to a loss of information (Zamo et Naveau, 2018). Therefore, for a fixed ensemble size, different ensembles although drawn from the same distribution will result in different forecast performances when verifying. The sampling-reordering methods we present in section 8.4 may use different sampling schemes, and eventually modify the

values afterward to respect autocorrelation constraints. This criterion thus aims at evaluating the resulting univariate ensembles regardless of multivariate aspects. The assessment of the overall quality (i.e., calibration and sharpness simultaneously) is conducted using the continuous ranked probability score (CRPS) and its skill score version (CRPSS). The definitions are given in sections 4.2.1.1 and 4.2.1.3 , respectively. Skill scores relate the forecast performances to that of reference forecasts. Values between 0 and 1 indicate that the forecasts outperform the reference forecasts (with 1 corresponding to perfect forecasts), while negative values indicate lower performances than the reference forecasts ones. For this study, the raw streamflow forecasts $\mathbf{f}$ are taken as reference. Univariate calibration is assessed using the rank histogram. Flatness of the histogram indicates a good calibration, $\cup$-shape and $\cap$-shape indicate under and over-dispersion, respectively, while $\diagup$-shape and $\diagdown$-shape indicate a negative and positive bias of the central tendencies, respectively. In this paper, rank histograms are represented under the stratified accumulated form (Bellier *et al.*, 2017a), with the objective of detecting whether different histogram shapes (and thus different forecast behaviors) average out. Each stratum represents the part of the histogram that originates from a subset of the complete verification sample. The subdivision is here made according to the streamflow forecast ensemble mean.

Multivariate criterion: Streamflow forecasts are likely to be used for forecasting quantities such as maxima or accumulated volumes across several lead times, eventually at the outlet of complex water systems involving a number of basins. It is therefore crucial for sampling-reordering methods to account for the dependencies between basins and lead times. Different scores can be used to evaluate ensemble forecasts of multivariate quantities. First, we consider the CRPS of univariate quantities that are sensitive to the dependence structure, namely the sum and the maximum, as they relate to the concepts of volume and peak flow. This verification strategy has been undertaken by Hemri *et al.* (2015) and Scheuerer *et al.* (2017), among others. In addition, we use multivariate scores that apply directly to multivariate ensemble forecasts, namely the energy score (ES) and the variogram score (VS), which have been introduced by Gneiting et Raftery (2007) and Scheuerer et Hamill (2015b), respectively. In a nutshell, the ES nicely detects misspecifications of univariate means and variances, but shows a limited capacity to discriminate forecasts with different multivariate structures (Pinson et Tastu, 2013; Scheuerer et Hamill, 2015b). The VS is an appropriate complement, since it better discriminates the differences in the dependence structure while it is insensitive to the univariate mean. The definition of the ES and VS are given in sections 4.4.2.1 and 4.4.2.2, respectively. Considering skill scores (see 4.2.1.3) with raw forecasts $\mathbf{f}$ as reference, we end up with the four following metrics: $\mathrm{CRPSS}_{\mathrm{sum}}$, $\mathrm{CRPSS}_{\mathrm{max}}$, ESS and VSS. A first series of these scores is computed separately for each basin from the forecast vectors of the lead times $1, 2, 3, \ldots, 24$ h, as a way to spot serial dependence issues regardless of the spatial correlations. The average skill score over all basins is reported. A second series (without $\mathrm{CRPSS}_{\mathrm{max}}$) is computed from the forecast vectors of the 6 basins and the lead times $6, 12, 18, \ldots, 120$ h, in order to focus on the medium-term spatiotemporal dependence structure. Before computation, streamflow forecasts and observations are standardized by the basin area such that each basin contributes equally. Multivariate calibration is assessed by the rank histogram from the forecasts of the univariate quantities sum and maximum, similarly to $\mathrm{CRPSS}_{\mathrm{sum}}$ and

$\text{CRPSS}_{\max}$. Alternative tools for multivariate calibration assessment are the multivariate-, average-, band depth- or minimum spanning tree- rank histograms (Thorarinsdottir *et al.*, 2016; Wilks, 2017), but they have not been considered here.

8.4 Methods for sampling-reordering

8.4.1 General formulation

The general formulation for sampling-reordering can be described as follows. Let M_p be the desired size of the postprocessed ensemble. From each marginal distribution $\tilde{G}_{j,k}$ within the collection $\{\tilde{G}_{1,1}, \ldots, \tilde{G}_{J,K}\}$ that results from univariate postprocessing, a sample of size M_p is drawn to construct the marginal ensemble $\tilde{\mathbf{x}}^{j,k}$:

$$\tilde{\mathbf{x}}^{j,k} = \left({\tilde{G}_{j,k}}^{-1}\left(\alpha_1^{j,k}\right), \ldots, {\tilde{G}_{j,k}}^{-1}\left(\alpha_{M_p}^{j,k}\right)\right), \tag{8.3}$$

where $\alpha_1^{j,k}, \ldots, \alpha_{M_p}^{j,k}$ are sampling probabilities between 0 and 1. These are referred hereafter to as levels. The $J \times K$ marginal ensembles are then aggregated to construct the complete multivariate ensemble

$$\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}^{1,1}, \ldots, \tilde{\mathbf{x}}^{J,K}). \tag{8.4}$$

Finally, the ensemble $\tilde{\mathbf{x}}$ is backtransformed to the original space using the inverse eNQT function t^{-1}. The resulting ensemble, denoted by $\tilde{\mathbf{f}}$, is the postprocessed forecast that will be evaluated against streamflow observations. In the above procedure, the reordering part consists in coupling the levels $\alpha_1^{j,k}, \ldots, \alpha_{M_p}^{j,k}$ between the different dimensions $\{j,k\}$ in a way that reflects a specified dependence structure.

8.4.2 Gaussian copula approach

Before turning to ECC-based methods on which this paper focuses, we present the parametric method referred to as GCA (Pinson et Girard, 2012) that will be considered as the benchmark method. Note that the technique suggested shortly afterward by Möller *et al.* (2013) is very similar. It consists first in setting a $(J \times K)$-variate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_y)$ with mean zero and covariance matrix equal to the correlation matrix $\mathbf{\Gamma}_y$, which contains the specified dependence structure. Then, for each forecast case, one must draw independent realizations $(v_m^{1,1}, \ldots, v_m^{J,K})$ for $m \in \{1, \ldots, M_p\}$. Taking for each dimension $\{j,k\}$ the probability integral transform of these realizations, which marginally follow a standard normal distribution, leads to the levels

$$\alpha_1^{j,k} = \Phi\left(v_1^{j,k}\right), \ldots, \alpha_{M_p}^{j,k} = \Phi\left(v_{M_p}^{j,k}\right), \tag{8.5}$$

which have the desired dependence structure. Finally, the general procedure described in section 8.4 applies. The parameter of GCA is the correlation matrix $\mathbf{\Gamma}_y$, which accounts for both between-basins and between-lead times dependencies. It is here empirically estimated from the observations used for training during the postprocessing step, which have been

transformed via the eNQT such that they follow a standard normal distribution. Although GCA allows one to sample as many members as desired, we have set here $M_p = M = 51$ for comparison with the ECC-based methods.

From the univariate perspective, GCA applies a random sampling scheme, since the levels in equation (8.5) follow the standard uniform distribution $\mathcal{U}(0,1)$. As a result, assuming that a postprocessed distribution $\tilde{G}_{j,k}$ is calibrated, the univariate ensemble $\tilde{\mathbf{x}}^{j,k}$ that results from GCA is also calibrated. The main limitation, though, is that the multivariate dependence structure is stationary, as it is estimated from training data that are not conditioned on the forecast case. That is, GCA does not distinguish for example between convective situations and large-scale weather fronts, which may affect spatial correlations, neither between full baseflow and mostly overland flow conditions, which may affect the temporal correlation. Figure 8.3a shows the ensemble $\tilde{\mathbf{f}}$ obtained from GCA for the forecast example. One can notice that streamflow trajectories for given member labels are completely independent from ones in the ECC-based methods, which we describe from now on.

8.4.3 Standard ECC (ECC-Q)

Non-parametric approaches for reordering aim at coupling the levels $\alpha_1^{j,k}, \ldots, \alpha_{M_p}^{j,k}$ between the different dimensions $\{j,k\}$ in a way that reproduces on $\tilde{\mathbf{x}}$ the rank dependence structure of a suitable multivariate dependence template. Such a template, which takes the form of a multivariate ensemble denoted by $\mathbf{z} = (\mathbf{z}^{1,1}, \ldots, \mathbf{z}^{J,K})$, must comprise M_p exchangeable members. In this paper, we focus on ECC-based methods (Schefzik *et al.*, 2013) whose core idea is to consider the raw ensemble forecast as the template. The approach has been primarily designed for meteorological postprocessing but can easily be applied to hydrological postprocessing (e.g., Hemri *et al.*, 2015). In our case study, the raw forecast comprises $I = 3$ groups that correspond to different hydrological models, each one comprising $M = 51$ exchangeable members. We choose here to consider as the template a subset of the raw forecast that originates from a single model, and set $M_p = M = 51$. Each streamflow trajectory in the postprocessed ensemble is thus linked to a certain meteorological scenario. It is logical to select the hydrological model that offers the best performances, which happens in our case study to be GRP ($i = 2$, see section 8.2). Besides, we consider for the template the raw forecasts after transformation on the normal space. The latter consideration is optional for the standard ECC described below, but does matter for some variants that come after. That being said, we define the ECC template $\mathbf{z} = (\mathbf{z}^{1,1}, \ldots, \mathbf{z}^{J,K})$ via

$$z_1^{j,k} = x_{2,1}^{j,k}, \ldots, z_M^{j,k} = x_{2,M}^{j,k} \tag{8.6}$$

for each dimension $\{j,k\}$. Figure 8.4a gives an illustration of this template for the forecast example.

Schefzik *et al.* (2013) have described different implementations of ECC that differ in the sampling procedure. We here consider as the standard implementation the version based on equidistant quantiles and referred to as ECC-Q, since it is the most natural way

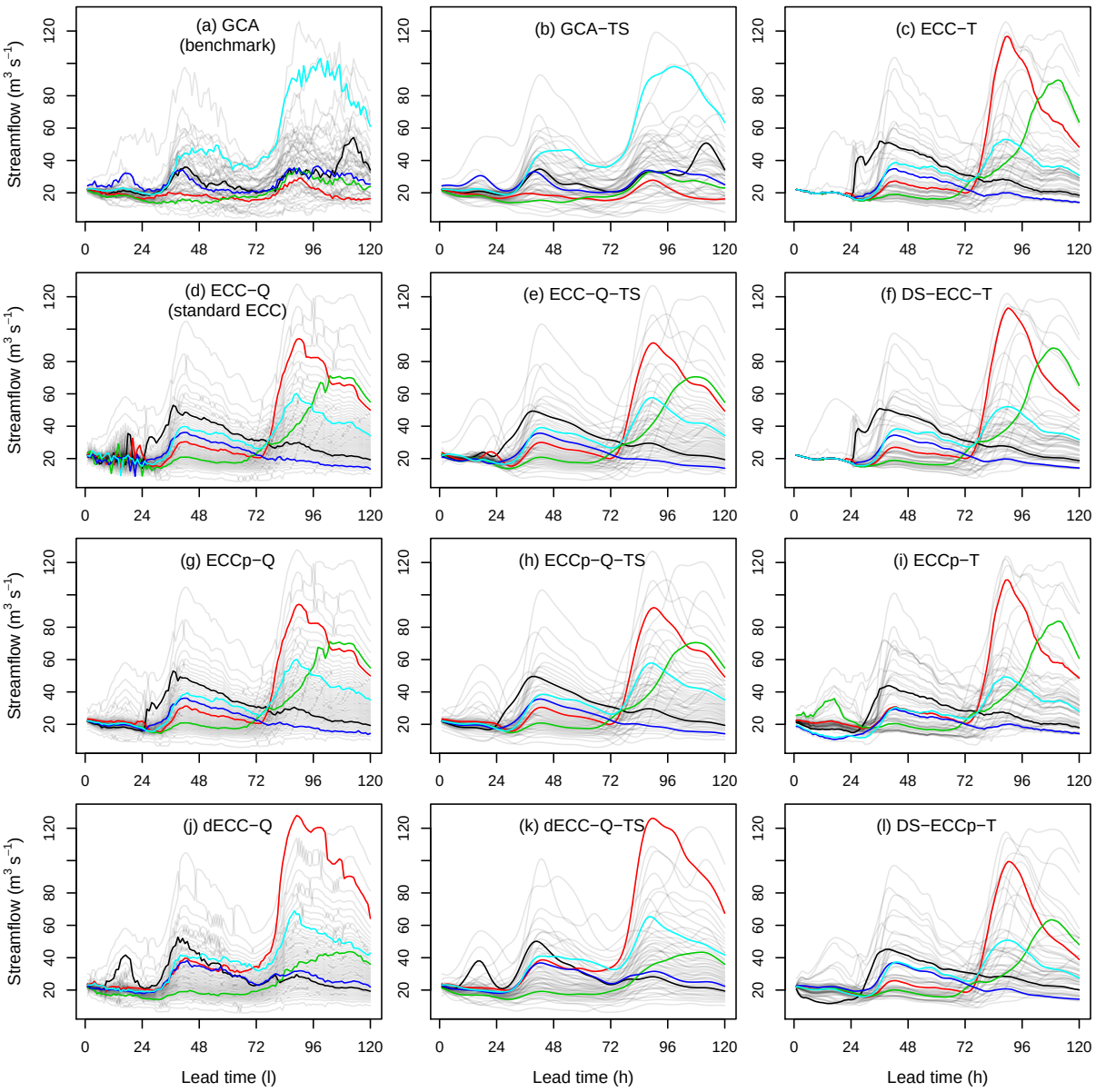

FIGURE 8.3 – Streamflow forecasts for the forecast example, obtained with each of the mixed methods. For ease of comparison, the first five members are picked to appear in colors.

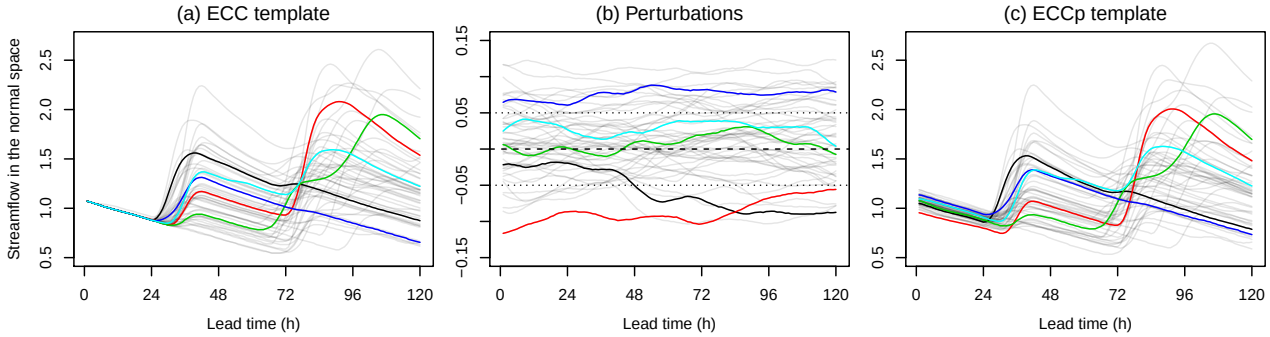

FIGURE 8.4 – Illustration of the (a) ECC and (c) ECCp reordering templates, for the forecast example. The first five members are picked to appear in colors for illustrative purpose. The ECCp template differs from that of ECC in the addition of an autocorrelated perturbations (plotted in b) on each of the $M = 51$ member. The parameter σ setting the mean amplitude of the perturbations is set to 0.05 for this figure.

of transforming continuous distributions into discrete ensembles. The ECC-Q technique consists in setting the levels as

$$\alpha_1^{j,k} = \frac{\mathrm{rk}(z_1^{j,k})}{M+1}, \ldots, \alpha_M^{j,k} = \frac{\mathrm{rk}(z_M^{j,k})}{M+1} \tag{8.7}$$

for each dimension $\{j, k\}$, where $\mathrm{rk}(z_m^{j,k})$ denotes the rank of $z_m^{j,k}$ within the vector $(z_1^{j,k}, \ldots, z_M^{j,k})$, with ties resolved as random. Finally, the general procedure described in section 8.4 applies. The equation (8.7) suggests that the levels, when sorted, take the values $\frac{1}{M+1}, \ldots, \frac{M}{M+1}$. Such a sampling scheme, hereafter referred to as the Q scheme, maintains the univariate calibration with respect to the rank histogram, since the sorted members of the ensemble $\tilde{\mathbf{x}}^{j,k}$ delineate intervals of equal probability for the observation to fall in. Note that it is not exactly the optimal scheme derived by Bröcker (2012) which minimizes the CRPS (levels then take the values $\frac{1-0.5}{M}, \ldots, \frac{M-0.5}{M}$), but is preferred since the latter optimal scheme does not display a flat rank histogram.

The ensemble $\tilde{\mathbf{f}}$ obtained from ECC-Q for the forecast example is illustrated in Figure 8.3d. Despite its simplicity, the method produces unrealistic streamflow trajectories regarding the temporal aspect, for three reasons. Let us consider a given basin $j \in \{1, \ldots, J\}$. First, the Q scheme imposes equally spaced levels $\frac{1}{M+1}, \ldots, \frac{M}{M+1}$ for sampling each of the distributions $\{\tilde{G}_{j,1}, \ldots, \tilde{G}_{j,K}\}$. Therefore, a given postprocessed trajectory must jump from levels to others as its counterpart in the template intersects other trajectories, and thus changes ranks. The distributions being strongly autocorrelated, these jumps between different levels lead to streamflow gradients that may be unrealistic, in particular in the tails of the distributions where the resulting quantiles are more widely spaced. This is visible in Figure 8.3d, around lead time 96 h. The second reason is the presence of discontinuities in the distributions $\{\tilde{G}_{j,1}, \ldots, \tilde{G}_{j,K}\}$, as it can be observed in Figure 8.2d. These derive from postprocessing being univariate and based on training samples that are of finite sizes. Thus, a streamflow trajectory with the same level across several consecutive lead times may still fluctuate, because of these discontinuities. Finally, the third reason arises when the template ensemble has no dispersion. Recall that the template in ECC-

based methods is the raw streamflow ensemble forecast, and it is frequent that, for any of the members, precipitation is not sufficient for the hydrological model to react. This yields a non-dispersive ensemble over several consecutive lead times, which leads the ranks in equation (8.7) to be allocated randomly. Meanwhile, the distributions $\{\tilde{G}_{j,1}, \ldots, \tilde{G}_{j,K}\}$ for these lead times have non-zero variances, as a result of the hydrological uncertainty being accounted for by the postprocessing. In those circumstances, the ECC-Q trajectories fluctuate strongly and to a wide extent from one lead time to the next, until the raw forecast becomes dispersive and thus the ranks are non-randomly fixed. This issue can be observed in Figure 8.3d, during the first 24 h.

To summarize, thanks to its sampling properties, the standard method ECC-Q is expected to perform well with regard to the univariate criterion, but has strong limitations with regard to the autocorrelation criterion. Temporal trajectories being not realistic, it is likely that the resulting forecasts are not satisfactory with respect to the multivariate criterion neither. Note that ECC implementations based on random sampling schemes exist (Schefzik *et al.*, 2013; Hu *et al.*, 2016), but have not been considered here since they are not expected to bring any benefit regarding the autocorrelation criterion, while their sampling properties are inferior to those of the Q scheme.

8.4.4 Trajectory smoothing (TS)

The first variant, which addresses the autocorrelation criterion only, consists in smoothing the temporal streamflow trajectories of $\tilde{\mathbf{f}}$, that is after their backtransformation to the original space, in such a manner that they are more realistic. For this purpose we use cubic smoothing splines, implemented in the R base function `smooth.spline`. This requires that a smoothing parameter be fixed. To do so, we suggest an ACF matching routine that consists in selecting the parameter that minimizes, over the training period, the mean absolute error between the ACF of the forecasts $\tilde{\mathbf{f}}$ and that of the forecasts $\mathbf{f}$, which is considered as reference regarding the autocorrelation criteria (see section 8.3). This smoothing procedure is referred hereafter to as trajectory smoothing (TS), and may be executed at the end of any sampling-reordering method as a way of making temporal trajectories more realistic. In return, it potentially affects the two other criteria, by modifying the values that result from sampling. Figure 8.3e gives an example of TS when it is executed after ECC-Q. We observe that the sharp fluctuations during the first 24 h, which are caused by the random rank structure, have been completely smoothed, but in turn the ensemble dispersion has greatly reduced.

8.4.5 Transformation sampling (ECC-T)

An alternative for avoiding jumpy trajectories is to change from the Q sampling scheme to another scheme where the levels are not equally spaced. Schefzik *et al.* (2013) have proposed a "transformation" approach where the spacing between the levels depends on the spacing between the members of the template ensemble. We refer hereafter to as the T scheme, in order to avoid any confusion with the term "transformation" that refers in this paper to the data mapping from the original to the normal space. Consider a given dimension $\{j, k\}$. The idea of this new scheme is to fit a parametric, continuous CDF denoted by $S_{j,k}$ to the template ensemble $z_1^{j,k}, \ldots, z_M^{j,k}$. In the same vein as quantile mapping, the

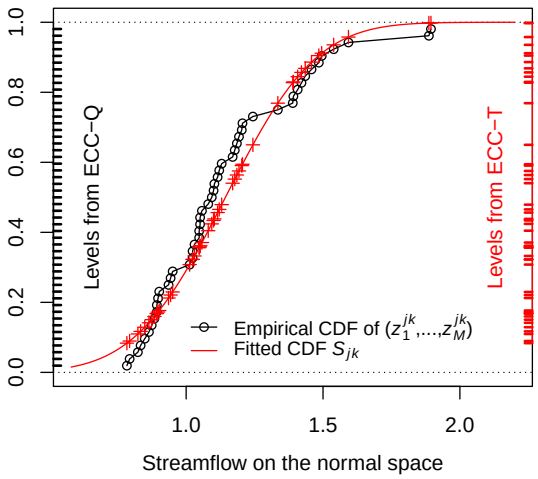

FIGURE 8.5 – Illustration of the sampling levels derived from the ECC-T approach (right axis), for a given dimension $\{j,k\}$, here the lead time 36 h of the forecast example. For comparison, the equally spaced levels of ECC-Q are plotted on the left axis.

approach aims at determining to which quantiles of $S_{j,k}$ the members $z_1^{j,k},\dots,z_M^{j,k}$ of the template ensemble correspond, before mapping them to the corresponding quantiles of the postprocessed distribution $\tilde{G}_{j,k}$. Formally, it consists in setting the sampling levels as

$$\alpha_1^{j,k} = S_{j,k}(z_1^{j,k}),\dots,\alpha_M^{j,k} = S_{j,k}(z_M^{j,k}) \tag{8.8}$$

for each dimension $\{j,k\}$. As we are in the normal space, the distributions $S_{j,k}$ are taken as normal CDF with mean $\mu_{j,k}$ and variance $\sigma_{j,k}^2$, estimated using maximum-likelihood. By construction, the T scheme automatically reproduces on $\tilde{\mathbf{x}}$ the rank structure of the template $\mathbf{z}$, hence we refer to as ECC-T. Unlike ECC-Q though, the levels may take any values within $[0,1]$, which lets the evolution across the lead times be smoother and avoid jumps. Figure 8.5 illustrates the differences in the levels obtained with the two approaches, for a given dimension $\{j,k\}$. We observe that the ECC-Q levels explore in a regular way the full range $[0,1]$, which ensures that calibration is maintained. On the other hand, the repartition of the ECC-T levels within $[0,1]$ is directly impacted by the fit of $S_{j,k}$. That is, the more normally distributed the template ensemble $z_1^{j,k},\dots,z_M^{j,k}$, the better the fit of $S_{j,k}$, and thus the better the calibration of the resulting forecast.

Some adjustments are necessary in order to prevent unrealistically low or high values in the postprocessed ensembles $\tilde{\mathbf{x}}$ when levels approach 0 or 1. First, in case of very low or null dispersion in a given template ensemble $\mathbf{z}_{j,k}$, the distribution $S_{j,k}$ cannot be fitted correctly, and some levels are likely to group around 0 or 1. We adopt here the heuristic adjustment proposed by Hemri *et al.* (2015) that consists in setting the variance $\sigma_{j,k}^2$ to

$$\max\left\{\sigma_{j,k}^2, \left[t((1+d)\bar{r}_{j,k}) - t((1-d)\bar{r}_{j,k})\right]^2\right\}, \tag{8.9}$$

where $\bar{r}_{j,k}$ is the mean of the (backtransformed) template for the dimension $\{j,k\}$ and d is a tuning parameter. After some tests, we have set $d = 0.0001$ as it gave the highest CRPSS. The above adjustment ensures that the levels group around 0.5 when the template

dispersion is very low. However, it does not prevent them, when the template dispersion is large enough, to be critically close to 0 or 1. This occurs if the fit of $S_{j,k}$ is not correct in the tails. We therefore suggest a second adjustment that constrains the levels to stay within the interval $[\gamma, 1-\gamma]$, where γ is a second tuning parameter. The procedure works as follows. Consider a given basin $j \in \{1, \ldots, J\}$. The levels are set according to equations (8.8), with the adjustment of Hemri *et al.* (2015). Then, for each trajectory $m \in \{1, \ldots, M\}$, the levels $\alpha_m^{j,1}, \ldots, \alpha_m^{j,K}$ are adjusted by the affine function

$$\mathcal{F}_{\text{sup}} : [0,1] \to [0, 1-\gamma] \tag{8.10}$$

if and only if there exists, for that trajectory, at least one lead time $k \in \{1, \ldots, K\}$ for which $\alpha_m^{j,k} > 1-\gamma$. Levels are then adjusted a second time by the affine function

$$\mathcal{F}_{\text{inf}} : [0,1] \to [\gamma, 1] \tag{8.11}$$

if and only if there exists at least one $k \in \{1, \ldots, K\}$ for which $\alpha_m^{j,k} < \gamma$. In other words, the functions $\mathcal{F}_{\text{sup}}$ and $\mathcal{F}_{\text{inf}}$ perform a linear mapping on the levels of the identified trajectories which enter the upper and lower "non-sampling" area defined via γ. Note that, for these trajectories, the adjustment of the K levels simultaneously maintains the autocorrelation, while the adjustment of only the levels falling into the "non-sampling" areas would create discontinuities. The choice of γ is a compromise between univariate calibration and sharpness. Assume that, for a given dimension $\{j, k\}$, sampling is performed within a postprocessed distributions $\tilde{G}_{j,k}$ that is calibrated. If $\gamma = 0$ (no adjustment), the frequency of outliers in the rank histogram, i.e., the frequency the observation falls outside the range of the ensemble, should be correct, that is around $\frac{1}{M+1}$ on average. In return, some extreme members may affect the forecast sharpness. If γ is too high, say 0.1, no extreme trajectories occur, but the forecast faces the risk of showing too many outliers. In this study, we have found $\gamma = \frac{1}{M+1} \simeq 0.019$ to give the highest CRPSS, at the expense of a limited increase of the number of outliers.

The ensemble $\tilde{\mathbf{f}}$ obtained with ECC-T and the two above adjustments is illustrated in Figure 8.3c for the forecast example. Visually, one can identify two limitations. First, the cases of non-dispersive templates are not handled satisfactorily, since the method produces non-dispersive trajectories although the postprocessed distributions have non-zero variance. Second, the ECC-T trajectories show unrealistic discontinuities at the lead times where the postprocessed distributions themselves have discontinuities, as discussed in section 8.4.

8.4.6 Distribution smoothing (DS)

To address the discontinuity problem of ECC-T, Hemri *et al.* (2015) suggest to perform a smoothing, across the lead times, of the postprocessed distribution parameters. We refer to this procedure as distribution smoothing (DS). In our case study, the 9 parameters of the BMA model [see equation (8.2)] need to be smoothed: a_i, ω_i and σ_i^2, for $i \in \{1, \ldots, 3\}$. As in the TS procedure in section 8.4, we use the R function `smooth.spline`, and select the smoothing parameter via ACF matching. The effect of the DS procedure is observable when comparing Figures 8.3c and 8.3f, although the discontinuity problem in this

particular example is not dramatic.

8.4.7 Template perturbation (ECCp)

The problematic case of non-dispersive template concerns both ECC-Q and ECC-T, although it is handled differently: the former method explores the full postprocessed distributions but yields noisy trajectories, while the latter produces more realistic trajectories, at the price of underdispersive forecasts. In other words, neither ECC-Q nor ECC-T are satisfying regarding both the univariate and the autocorrelation criteria. To address this, we propose a variant that consists in adding a random perturbation to each member in the ECC template in order to systematically create a minimum of dispersion for the trajectories to separate from each other. Let us denote by $\check{\mathbf{z}} = (\check{\mathbf{z}}^{1,1}, \ldots, \check{\mathbf{z}}^{J,K})$ a new template that has components $\check{\mathbf{z}}^{j,k}$ for each dimension $\{j, k\}$ defined as

$$\check{z}_1^{j,k} = z_1^{j,k} + \epsilon_1^{j,k}, \ldots, \check{z}_M^{j,k} = z_M^{j,k} + \epsilon_M^{j,k}, \tag{8.12}$$

where $\epsilon_m^{j,k}$ denotes a perturbation term. The idea is that when the original template is non-dispersive, the procedure picks the dependence structure of the perturbations only, as the original template values cancel out. Otherwise, when the template is dispersive enough, the perturbations are negligible so that the procedure reduces to the standard ECC. The construction of the perturbations is the core of this ECC variant, hence we refer hereafter to as ECCp. Its principle is illustrated in Figure 8.4.

First of all, we have found that these perturbations must be constructed for each basin independently, such that they do not assume any spatial correlation. As a matter of fact, the procedure aims at addressing the ECC limitation when the template is non-dispersive, which occurs when meteorological forcings have not been sufficient for the hydrological model to react. In those circumstances, the postprocessed streamflow uncertainty envelop accounts, in a lumped form, for non-meteorological sources such as initial catchment conditions, latest streamflow observations, and hydrological model parameters and structures. Unlike the meteorological uncertainties, the above uncertainties have no reasons to be spatially correlated, since hydrological modeling is performed separately for each basin. Note that this assumption may not hold in the case where the meteorological forcings are strongly under-dispersive, and thus the hydrological postprocessing accounts for some missing meteorological uncertainty.

Therefore, let us describe ECCp for a given basin $j \in \{1, \ldots, J\}$. The perturbation terms in equation (8.12) are the k-th components of temporal vectors $(\epsilon_m^{j,1}, \ldots, \epsilon_m^{j,K})$ for $m \in \{1, \ldots, M\}$ that are randomly drawn from a K-variate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_z)$ with mean zero and covariance matrix $\mathbf{\Sigma}_z$. This covariance matrix can be written as

$$\mathbf{\Sigma}_z = \mathrm{diag}(\mathbf{\Sigma}_z)^{1/2}\ \mathbf{\Gamma}_z\ \mathrm{diag}(\mathbf{\Sigma}_z)^{1/2}, \tag{8.13}$$

where $\mathrm{diag}(\mathbf{\Sigma}_z)$ is the matrix of the diagonal entries of $\mathbf{\Sigma}_z$, i.e., the variances, and $\mathbf{\Gamma}_z$ is the correlation matrix. Such a decomposition enables to specify the dependence structure (via $\mathbf{\Gamma}_z$) and the amplitude (via $\mathrm{diag}(\mathbf{\Sigma}_z)^{1/2}$) of the perturbations separately. The correlation matrix $\mathbf{\Gamma}_z$ is here empirically estimated from the past samples of $\mathbf{z}$ used for training during

the postprocessing step. This way, the perturbations retain the desired autocorrelation characteristics. Then, we assume a constant amplitude over the lead times, and set

$$\mathrm{diag}(\mathbf{\Sigma}_z)^{1/2} = \begin{pmatrix} \sigma & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma \end{pmatrix}, \tag{8.14}$$

where σ is a tuning parameter. The transformation step via the eNQT has not only normalized but also standardized the streamflow data, so σ is to be defined regardless of the basin or the forecast case (e.g., low, medium, or high flow).

For the approach coupling ECCp with the Q sampling scheme (i.e., ECCp-Q), taking a small albeit positive value for σ ensures that perturbations are negligible when the original template becomes dispersive, although it is sufficient for the trajectories in the perturbed template to systematically separate from each other. Hence, $\sigma = 0.01$ has been found satisfactory. To give an idea, this corresponds for the Guiers basin to a mean perturbation amplitude of 0.4 $\mathrm{m^3s^{-1}}$ for streamflow around 50 $\mathrm{m^3s^{-1}}$, which is negligible. The resulting ECCp-Q streamflow forecast for the forecast example is plotted in Figures 8.3g, where it can be noticed that the trajectories are no longer noisy during the first 24 h, while the rest of the forecast is nearly identical.

When coupling ECCp with the T sampling scheme (i.e., ECCp-T), the sensitivity of forecast skills to different values of σ is more important. Recall that, for the T scheme, the fit of the distribution $S_{j,k}$ to the template ensemble $z_1^{j,k}, \ldots, z_M^{j,k}$ has a direct impact on univariate calibration, and thus on forecast skill. A poor fit of $S_{j,k}$ typically occurs when only few trajectories in the template depart from a group of non-dispersive ones, i.e., the hydrological model has reacted for a few members only. Perturbing the template with a small σ value, say 0.01, does not modify much the configuration of the trajectories, and the fit remains poor. However, if σ is too high, say 0.5, the distribution $S_{j,k}$ fits well, since a normal distribution is being fitted to normal perturbations, but in return the perturbations are no longer negligible when the template becomes dispersive. That is, the effect of the ECC reordering reduces. After tests, we have found $\sigma = 0.1$ to give the highest CRPSS for ECCp-T. This corresponds for the Guiers basin to a mean perturbation amplitude of approximately 5 $\mathrm{m^3s^{-1}}$ for streamflow around 50 $\mathrm{m^3s^{-1}}$, which is not completely negligible, although it ensures that the fits of $S_{j,k}$ are correct for all the lead times. The resulting streamflow forecast for the forecast example is plotted in Figures 8.3i.

8.4.8 Dual-ECC (dECC)

The ECCp variant has relationship with the dual-ECC (hereafter referred to as dECC) proposed by Ben Bouallègue *et al.* (2016), which also aims at modifying the dependence template by adding a specified term. We briefly describe the procedure, but refer to the original article for further details. The motivation of dECC is that postprocessing may modify the ensemble spread to a large extent, and thus ECC magnifies raw ensemble correlation structures that would be nonrepresentative. It is therefore based on the assumption that the raw ensemble does not systematically describe correctly the spatiotemporal dependence structure, and would benefit from a second source of information (hence the

term "dual"), which is past forecast error statistics. We define the forecast error $\mathbf{e}$ as the difference $(J \times K)$-vector, here on the normal space, between the observation and the ensemble mean of the raw forecast:

$$\mathbf{e} = (y^{1,1} - \bar{z}^{1,1}, \ldots, y^{J,K} - \bar{z}^{J,K}), \tag{8.15}$$

where $y^{j,k}$ and $\bar{z}^{j,k}$ are respectively the observation and the template ensemble mean for the dimension $\{j, k\}$. We can empirically estimate the correlation matrix $\mathbf{\Gamma}_e$ of this forecast error, from the data used as training during the postprocessing step. Then, dECC aims at constructing, similarly to ECCp, a new dependence template on which reordering will be based. This involves the following steps:

1. Apply the ECC-Q procedure with the original template $\mathbf{z}$, to obtain a temporary postprocessed ensemble that we denote by $\dot{\tilde{\mathbf{x}}}$. Ben Bouallègue *et al.* (2016) do not mention alternative sampling schemes than the Q scheme, although there should not be any specific constraint regarding this point.

2. Derive the error correction $\mathbf{c}_m$ for each ensemble members $m \in \{1, \ldots, M\}$:

$$\mathbf{c}_m = (\dot{\tilde{x}}_m^{1,1} - z_m^{1,1}, \ldots, \dot{\tilde{x}}_m^{J,K} - z_m^{J,K}). \tag{8.16}$$

3. Transform the error correction $\mathbf{c}_m$ of each ensemble member $m \in \{1, \ldots, M\}$ into the adjusted correction $\check{\mathbf{c}}_m$, such that its correlation structure is pushed towards the correlation structure of the past forecast errors:

$$\check{\mathbf{c}}_m = \mathbf{\Gamma}_e^{1/2}\, \mathbf{c}_m. \tag{8.17}$$

   This transformation, which is inspired by a process known in signal processing as statistical coloring, requires the eigendecomposition of the correlation matrix $\mathbf{\Gamma}_e$.

4. Set the new template $\check{\mathbf{z}} = (\check{\mathbf{z}}^{1,1}, \ldots, \check{\mathbf{z}}^{J,K})$ such that each component $\check{\mathbf{z}}^{j,k}$ is now defined as

$$\check{z}_1^{j,k} = z_1^{j,k} + \check{c}_1^{j,k}, \ldots, \check{z}_M^{j,k} = z_M^{j,k} + \check{c}_M^{j,k}. \tag{8.18}$$

The steps 1-4 above are illustrated in Figure 8.6, while the streamflow forecast obtained by combining dECC with the Q sampling scheme (i.e., dECC-Q) is shown in Figure 8.3j for the forecast example. Besides the common idea of modifying the dependence template by adding a specified term, ECCp and dECC differ in several aspects. First, the amplitude of the perturbation vectors in ECCp is controlled by a tuning parameter, while the amplitude of $\check{\mathbf{c}}_m$ in dECC depends on both the error correction $\mathbf{c}_m$ and the correlation matrix $\mathbf{\Gamma}_e$. In particular, it is found that the correction $\mathbf{c}_m$ is greatly amplified when premultiplied by $\mathbf{\Gamma}_e^{1/2}$, because of the strong autocorrelation structure in $\mathbf{\Gamma}_e$. Streamflow data are indeed much more autocorrelated than wind speed data for which dECC has been developed in Ben Bouallègue *et al.* (2016). The second difference is that the additive terms $\check{\mathbf{c}}_m$ for $m \in \{1, \ldots, M\}$ account for both spatial and temporal correlation, while spatial correlations are ignored in the ECCp perturbations. The dECC method could be adapted

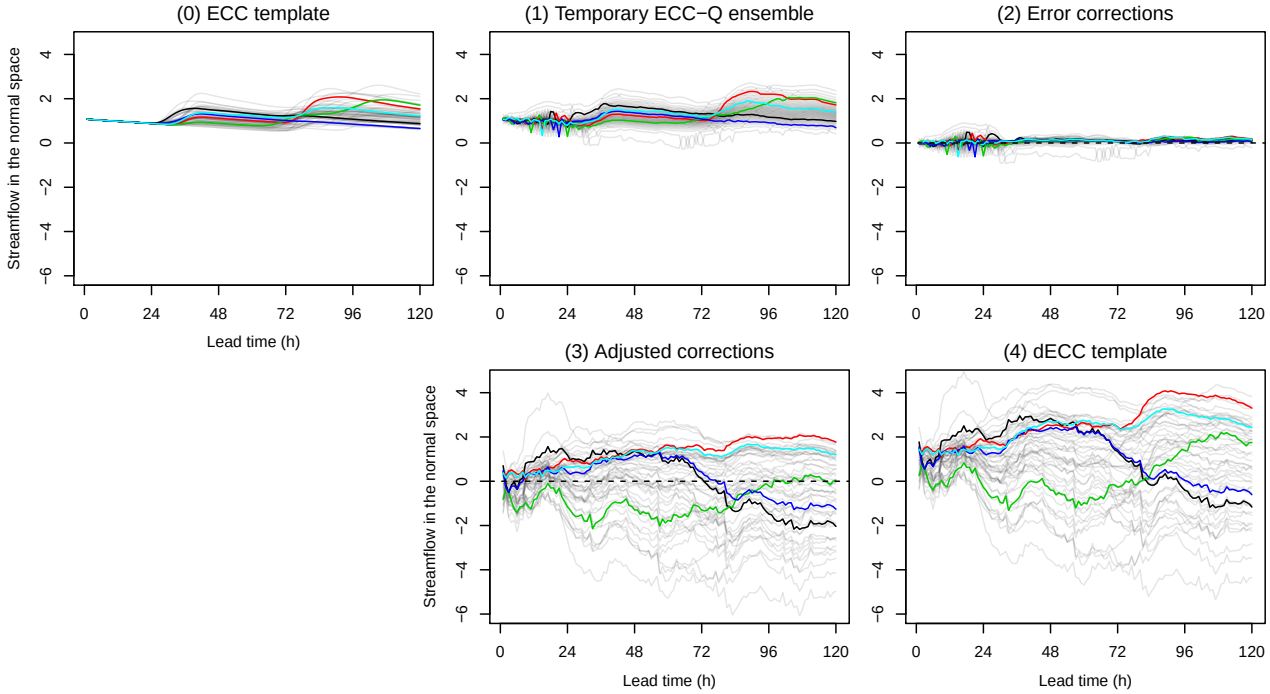

FIGURE 8.6 – Illustration of the different steps of dECC for setting a new template before reordering. The label of the plots corresponds to the steps in the text.

by redefining the correlation matrix $\mathbf{\Gamma}_e$ such that spatial correlations are ignored, but it would then deviate from the original idea of Ben Bouallègue *et al.* (2016).

8.5 Verification results and discussions

Different procedures have been described so far, which address specific shortcomings of the standard ECC-Q method. Some of these procedures may be combined, based on the following assumptions. First, TS has an interest only if combined with the Q sampling scheme, since the T scheme already aims at producing realistic streamflow trajectories. Nonetheless, DS may still be necessary for the T scheme to avoid any discontinuities in the trajectories. Besides, ECCp may be useful for both Q and T schemes, while dECC has been developed in hand with the Q scheme. All in all, we suggest a set of 10 different mixed methods, which are listed in Table 8.2, in addition to GCA (benchmark method) and ECC-Q (standard ECC). Figure 8.3 shows all the resulting streamflow forecasts for the forecast example. Supporting information Figures B.2-B.6 propose similar additional plots for randomly selected dates. We now report the results of the evaluation on the setup described in section 8.2, under the verification framework presented in section 8.3 that includes three criteria: autocorrelation, univariate skill, and multivariate skill.

8.5.1 Autocorrelation

We first evaluate the capacity of the different methods to produce forecast trajectories that are realistic regarding the strong autocorrelation of hourly streamflow, thus a pro-

TABLE 8.2 – Summary of the procedures involved in the 12 evaluated methods

Name	Distribution smoothing (DS)	Template modification	Sampling scheme	Trajectory smoothing (TS)
GCA (benchmark)	-	-	GCA	-
GCA-TS	-	-	GCA	Yes
ECC-Q (standard ECC)	-	-	Q	-
ECCp-Q	-	ECCp ($\sigma = 0.01$)	Q	-
dECC-Q	-	dECC	Q	-
ECC-Q-TS	-	-	Q	Yes
ECCp-Q-TS	-	ECCp ($\sigma = 0.01$)	Q	Yes
dECC-Q-TS	-	dECC	Q	Yes
ECC-T	-	-	T ($d = 10^{-4}$; $\gamma = 0.019$)	-
ECCp-T	-	ECCp ($\sigma = 0.1$)	T ($d = 10^{-4}$; $\gamma = 0.019$)	-
DS-ECC-T	Yes	-	T ($d = 10^{-4}$; $\gamma = 0.019$)	-
DS-ECCp-T	Yes	ECCp ($\sigma = 0.1$)	T ($d = 10^{-4}$; $\gamma = 0.019$)	-

Note: The values of the tuning parameters are indicated in brackets.

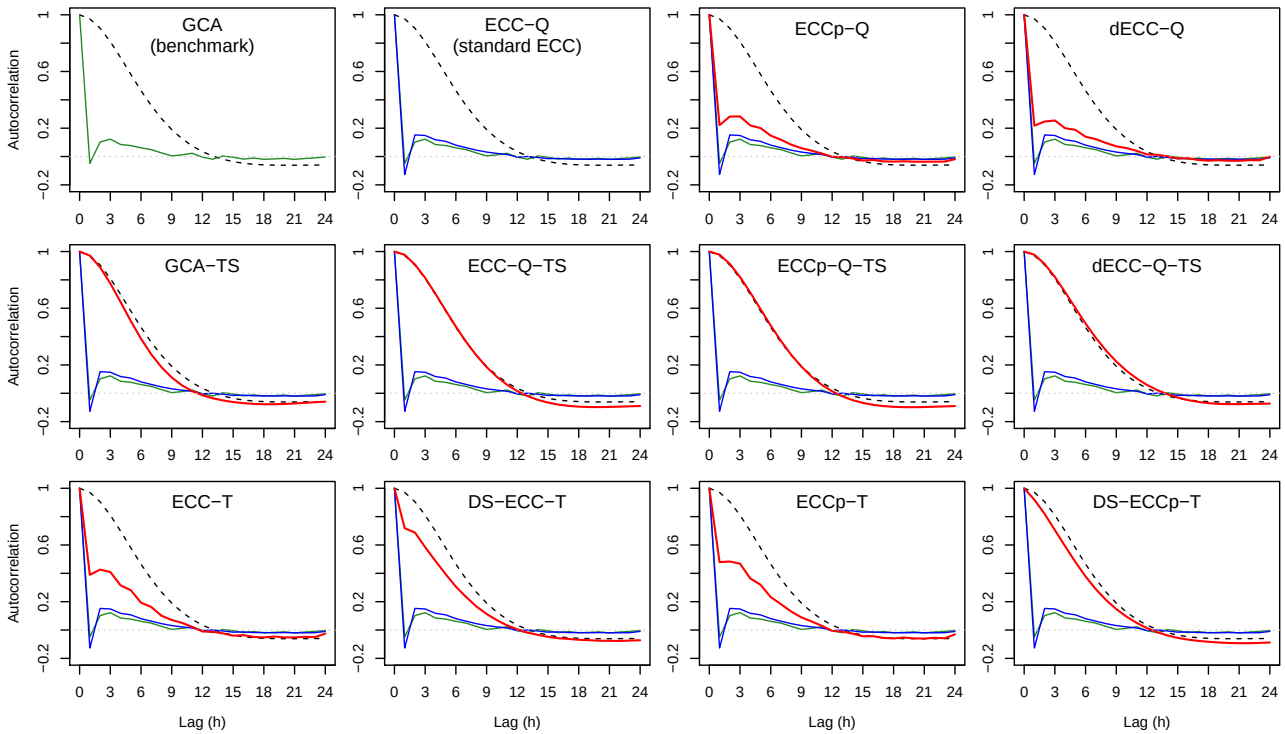

FIGURE 8.7 – ACF of lag-one differenced streamflow on the Guiers basin, for GCA (solid green line), ECC-Q (solid blue line), and each of the 10 mixed methods (solid red line). The reference ACF is in dashed black line.

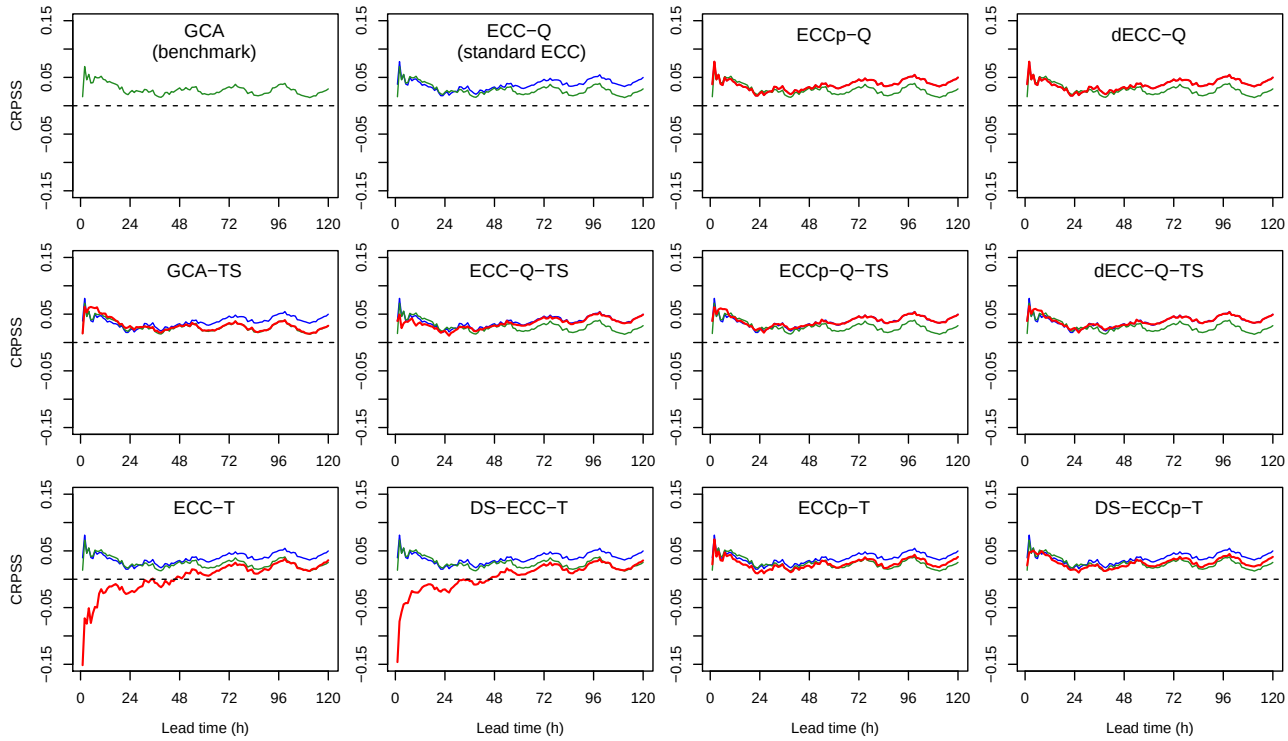

FIGURE 8.8 – CRPSS of the univariate streamflow forecasts on the Guiers basin, for GCA (solid green line), ECC-Q (solid blue line), and each of the 10 mixed methods (solid red line). Raw forecasts serve as reference.

perty of the forecasts only. Figure 8.7 shows the ACF of lag-one differenced streamflow over the Guiers basin for lags of 1 to 24 h. The reference ACF we want the methods to match is plotted in dashed black line. The rapid drop of the ACF of GCA and ECC-Q reflects the severe fluctuations of the streamflow forecasts. For ECCp-Q and dECC-Q, adapting the template reduces part of the unwanted fluctuations (those due to a non-dispersive template), but the smoothing of the trajectories remains necessary to obtain the desired ACF. Also, we notice that ECC-T improves upon ECC-Q thanks to the T sampling scheme that enables a smoother evolution of the levels, but the method as such is still not satisfying, and the DS and ECCp procedures are both needed to approach the desired ACF. The reason why ECCp is necessary is that, otherwise, the transitions between a non-dispersive and a dispersive template engender strong gradients that are not physically justified, as it can be seen in Figure 8.3f around the 24 h lead time. Other basins have led to similar results, which are provided in supporting information Figures B.7-B.11.

8.5.2 Univariate skills

The second criterion concerns the univariate forecast skill. It does not evaluate the effect of the reordering itself, but rather the skill of the different sampling schemes. Figure 8.8 depicts the CRPSS as a function of lead time for the Guiers basin. The results for the others basins are similar and provided in supporting information Figures B.12-B.16. In parallel, Figure 8.9 shows the rank histograms for the Guiers basin and a specific lead

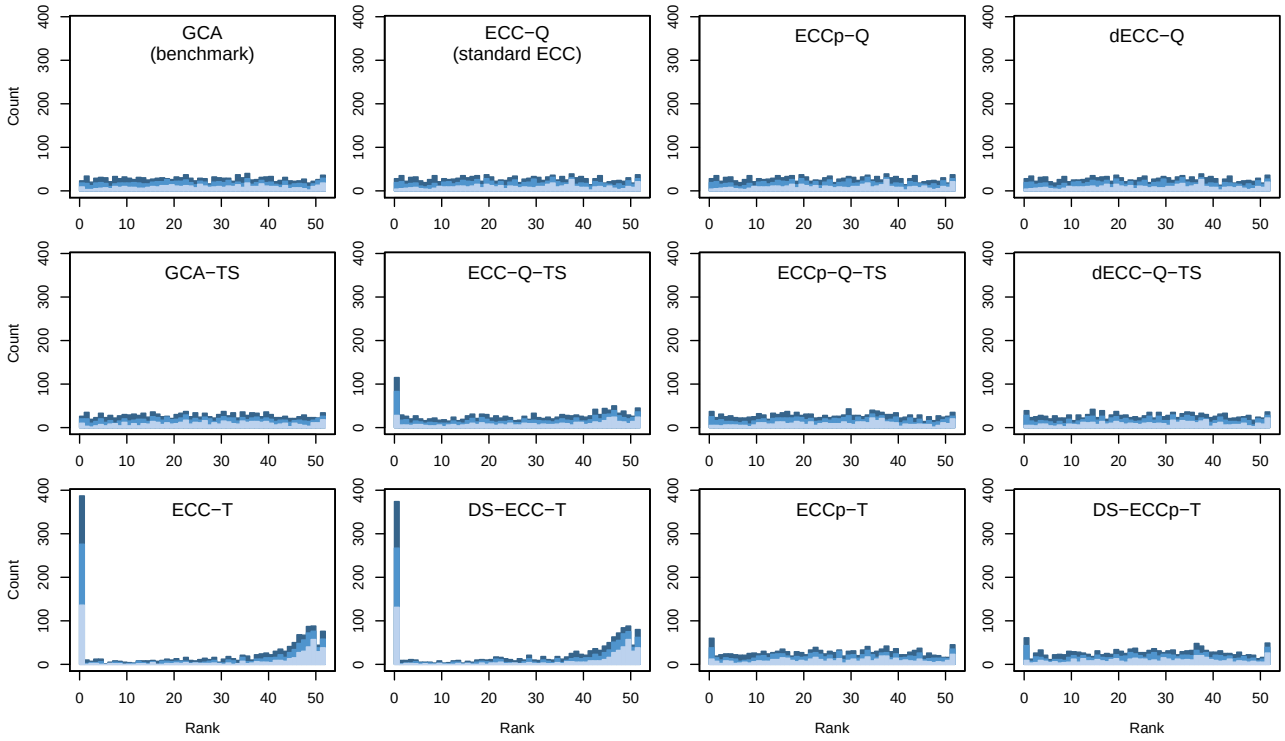

FIGURE 8.9 – Accumulated stratified rank histograms of the univariate streamflow forecasts on the Guiers basin for the 12 h lead time. The stratification criterion is the forecast ensemble mean. The numbers of forecasts within each strata are equal. Light, medium, and dark-blue strata contain lowest, intermediate, and highest streamflow forecast, respectively.

time (12 h), which help understand where the deficiencies of each method lie. Other lead times are proposed in supporting information Figures B.17-B.19.

It is found that the Q sampling scheme outperforms the random scheme implemented in GCA, to a larger extent as lead times increase and the dispersion grows. This is in accordance with Bröcker (2012) who states that sampling levels should be equally spaced for the CRPSS to be maximized. These two schemes however outperform the T scheme implemented in ECC-T, especially at first lead times. Figure 8.9 shows that, unlike the Q and the random schemes of GCA, the T scheme fails in maintaining the calibration gained through postprocessing, because of the poor fits of the distributions $S_{j,k}$ that yield sampling levels not exploring the full range $[0, 1]$. Moreover, we observe that the ECC-T rank histogram is not symmetric, which reflects an inherent limitation of the T scheme regarding its application to streamflow. Indeed, the hydrological model provides a strongly non-linear response to the meteorological forcings. To diverse precipitation, the model may or may not react, which very often produces in the ensemble forecast a group of members that share the exact same streamflow value during a series of lead times, while the other members have superior values. The eNQT transformation helps correct the non-Gaussian behavior of streamflow data, but it cannot handle such a point mass. Therefore, the distributions $S_{j,k}$, taken as normal CDF, have no chance to correctly fit in both tails. Note that this issue progressively reduces as lead times increase and the ensembles get closer to a Gaussian behavior. The ECCp procedure, by adding a perturbation, is very effective for improving the fits of the distributions $S_{j,k}$, as shown by the great improvement

TABLE 8.3 – Multivariate skill scores from the forecast vectors of lead times $1, 2, 3, \ldots, 24$ h

	$\text{CRPSS}_{\text{sum}}$	$\text{CRPSS}_{\text{max}}$	ESS	VSS
GCA (benchmark)	0.037	0.009	0.030	0.016
GCA-TS	0.038	0.001	0.036	0.017
ECC-Q (stand. ECC)	0.030	−0.091	0.027	−0.086
ECCp-Q	0.041	0.007	0.036	0.015
dECC-Q	0.042	0.004	0.034	0.003
ECC-Q-TS	0.031	−0.002	0.035	0.017
ECCp-Q-TS	0.042	0.004	0.041	0.025
dECC-Q-TS	0.043	0.000	0.039	0.012
ECC-T	−0.001	−0.009	−0.008	−0.011
ECCp-T	0.033	0.006	0.030	0.015
DS-ECC-T	−0.003	−0.013	−0.002	−0.006
DS-ECCp-T	0.029	0.002	0.034	0.020

Note: The average skill score over all basins is reported. Raw forecasts serve as reference.

toward flatness between the rank histograms of ECC-T and ECCp-T in Figure 8.9. As a consequence, the gain in CRPSS is also significant. The remaining gap between ECCp-T and ECCp-Q is imputed to sampling levels in the former method being not equally spaced. Also, note that the univariate skills of ECC-Q, ECCp-Q and dECC-Q are strictly identical, since the methods differ only in the multivariate rank structure.

Last but not least, the TS and DS procedures can theoretically affect the univariate skills, since they modify the sampled values. However, Figure 8.8 and 8.9 show that their impact is hardly detectable, with the exception of ECC-Q-TS which, during the first lead times, tends to produce underdispersive forecasts, with an emphasis on the low tie where the number of outliers soars. The reason is that the fluctuations of the ECC-Q trajectories are more important than those of the other methods, so the amplitude of the smoothing corrections is larger. Turning to ECCp-Q-TS or dECC-Q-TS corrects for this behavior.

8.5.3 Multivariate skills

The third criterion concerns the multivariate forecast skill. We first focus on the scores reported in Table 8.3 that spot serial dependence issues regardless of the spatial correlation. The standard ECC method, ECC-Q, performs fairly compared to others according to $\text{CRPSS}_{\text{sum}}$ and ESS, but its skill drops by an order of magnitude when looking at $\text{CRPSS}_{\text{max}}$ and VSS. These two scores are indeed particularly sensitive to the dependence structure, and heavily penalize the fluctuations in the ECC-Q streamflow trajectories. Besides, these fluctuations induce a miscalibration when forecasting the maximum across several lead times, as shown by the ECC-Q rank histogram in Figure 8.10, where the first rank is overpopulated. We also notice that the ECC-T forecasts are strongly miscalibrated, but this comes mainly from their limited sampling properties. Table 8.3 and Figure 8.10 indicate that modifying the template before reordering, whether using the ECCp or dECC variant, is beneficial. This is valid for both the Q and the T sampling

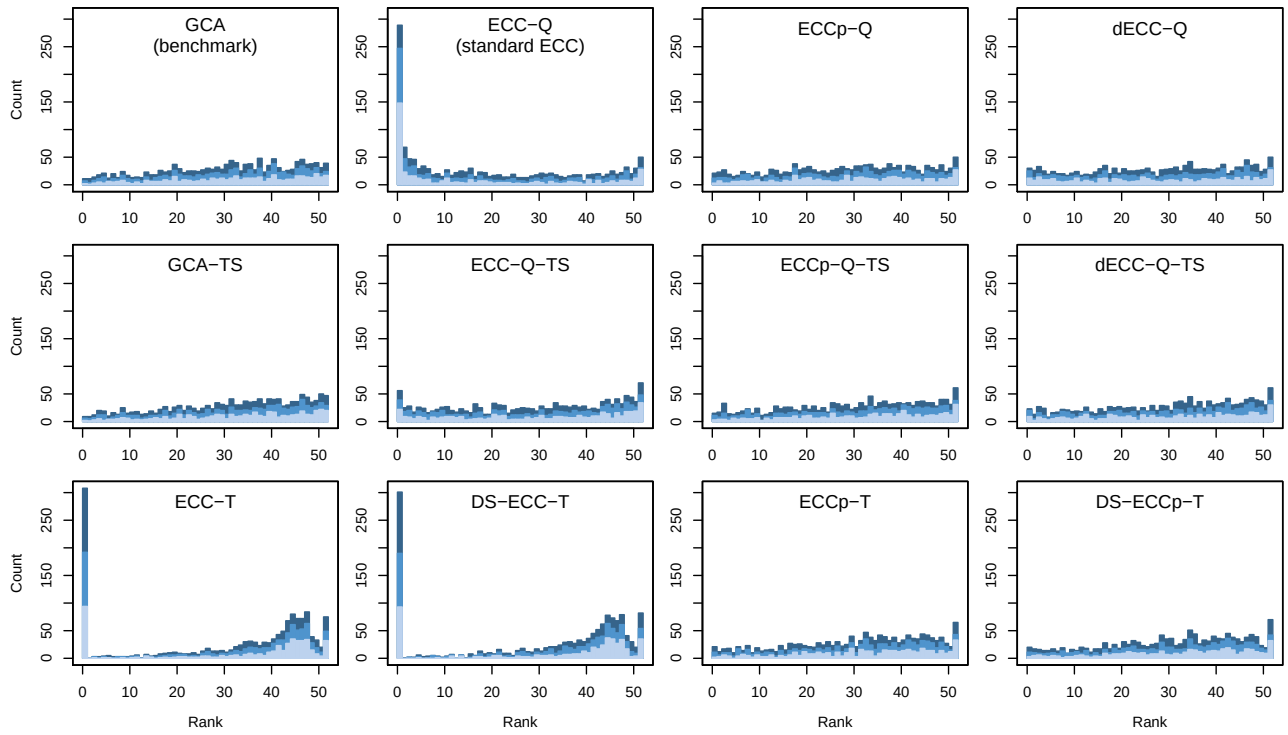

FIGURE 8.10 – Accumulated stratified rank histograms of the forecasts of the maximum across the lead times $1, 2, 3, \ldots, 24$ h, over the Guiers basin. The same stratification as in Figure 8.9 is applied.

scheme, although dECC is shown here for the Q scheme only. Furthermore, the smoothing procedures (TS and DS) are found to have a negligible impact, if not beneficial. Finally, our benchmark method, GCA, fares relatively well here.

Table 8.4 now reports the skill scores when focusing on the medium term spatiotemporal dependence structure. It highlights specific shortcomings that were invisible when looking at the temporal structure only. Indeed, we now observe a large gap in skill between GCA and the ECC-based methods. It has been shown via the univariate criterion that the sampling properties of GCA are inferior to those of ECC-Q. In addition, it appears here that the stationarity of the dependence structure is a fairly limiting factor as well. This contrasts with the results of Hemri *et al.* (2015), who compared GCA to an ECC-based method (i.e., ECC-T) and found out that GCA performs slightly better. Nonetheless, they did not account for between-basins dependencies when verifying, although the conditioning of the spatial dependence structure on the forecast case seems to be crucial. More precisely, GCA tends to overestimate in average the spatial correlations, since the spatial dependence structure is the same whether the raw forecasts over the different basins are dispersive or not, that is, whether the meteorological forcings (which are spatially correlated) have triggered a hydrological reaction or not. This behavior is confirmed by the rank histograms of GCA in Figure 8.11 for the spatiotemporal sum, which shows a marked $\cap$-shape. A similar issue is reported for dECC-Q. It appears here that the modification of the template by the dECC procedure leads to an over-estimation of the spatial correlations, unlike the ECCp variant where the template is perturbed independently for each basin. Moreover, dECC modifies the template to a large extent (not only when it is non-

Table 8.4 – Multivariate skill scores from the forecast vectors of the 6 basins and lead times $6, 12, 18, \ldots, 120$ h

	$\text{CRPSS}_{\text{sum}}$	ESS	VSS
GCA (benchmark)	0.001	0.002	−0.004
GCA-TS	0.001	0.001	−0.005
ECC-Q (stand. ECC)	0.061	0.028	0.027
ECCp-Q	0.061	0.028	0.029
dECC-Q	0.048	0.018	0.011
ECC-Q-TS	0.060	0.028	0.029
ECCp-Q-TS	0.061	0.028	0.029
dECC-Q-TS	0.047	0.018	0.011
ECC-T	0.052	0.023	0.025
ECCp-T	0.054	0.027	0.030
DS-ECC-T	0.052	0.024	0.026
DS-ECCp-T	0.055	0.027	0.031

Note: Raw forecasts serve as reference.

dispersive), which tends to reduce the conditioning of the dependence structure on the forecast case. As a consequence, Table 8.4 shows that dECC-Q performs worse than the other ECC-based methods. This also contrasts with the results of Ben Bouallègue *et al.* (2016) who found out that dECC-Q outperforms ECC-Q but, again, they verify forecasts of time series only. It is possible that dECC proves to be useful when the raw template does not have a coherent spatiotemporal structure, as the method benefits from a second source of information. This would occur for instance when the meteorological forcings are not themselves coherent. However, we believe that more gains would be expected by making these forcing coherent first.

8.6 Conclusions

This paper has focused on sampling-reordering methods that take place after univariate hydrological postprocessing, with the aim of generating coherent streamflow ensemble forecasts. The first objective was to compare a parametric method that assumes a stationary dependence structure, the Gaussian copula approach (GCA), against various methods based on the ensemble coupling copula (ECC) technique, which reproduces the dependence structure of the raw ensemble and therefore adapts to each forecast case. The evaluation was based on three criteria: autocorrelation, univariate skill, and multivariate skill. It was found that GCA suffers from a non-optimal sampling scheme and, above all, from the stationarity of the dependence structure. The ECC-based methods were found more attractive, although there are numerous limitations that must be addressed. The standard implementation, ECC-Q, shows good univariate skills thanks to its sampling properties, but generates streamflow trajectories that contain unrealistic jumps. Moreover, the resulting ensembles are miscalibrated with respect to the multivariate aspect, due to the random allocation of the ranks when the raw streamflow forecast used as template for reordering is non-dispersive.

The second objective was to test different procedures that address these limitations.

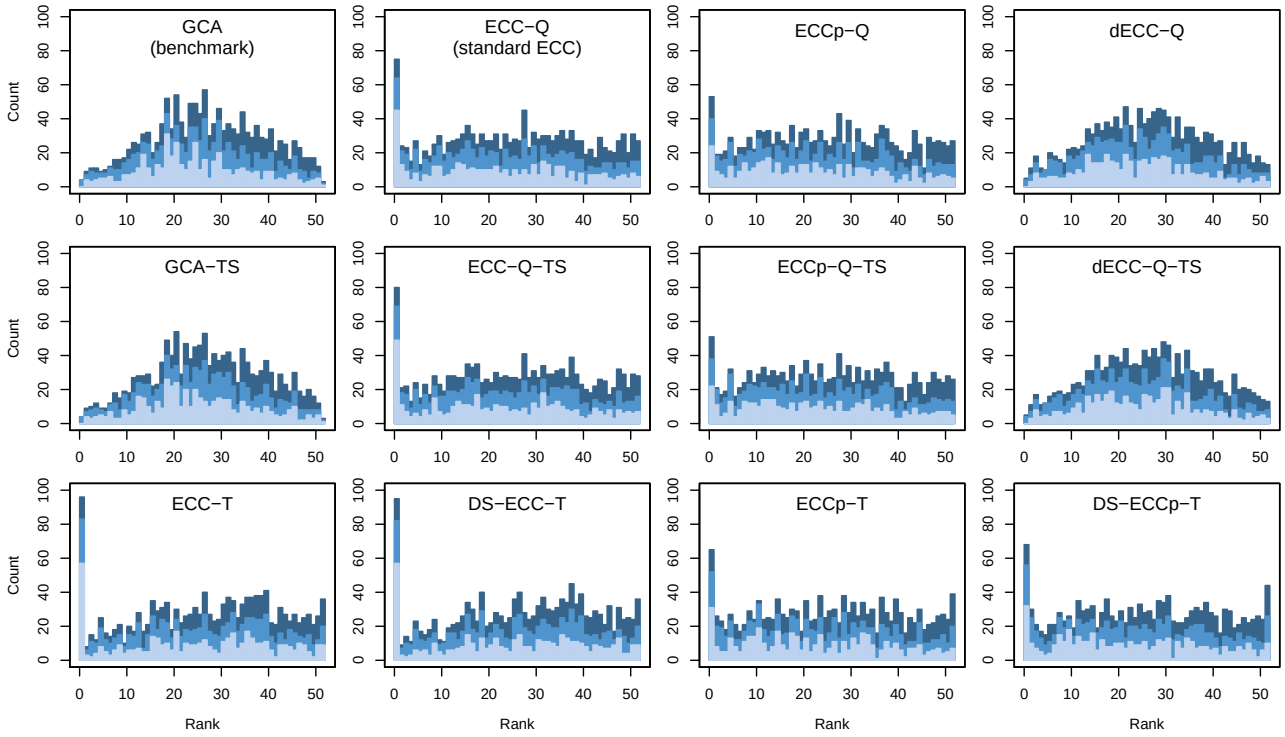

FIGURE 8.11 – Accumulated stratified rank histograms of the forecasts of the spatiotemporal sum across the 6 study basins and the lead times $6, 12, 18, \ldots, 120$ h. The same stratification as in Figure 8.9 is applied.

The first one, ECC-T, was used by Hemri *et al.* (2015) as a way of avoiding the unrealistic jumps that arise in the ECC-Q trajectories. It however requires empirical adjustments to prevent unrealistically extreme streamflow values. In addition, the postprocessing parameters need to be smoothed prior the sampling-reordering step in order to avoid any discontinuities in the streamflow trajectories. As such, the approach succeeds in producing realistic trajectories, but fails in maintaining the calibration gained through the univariate postprocessing. Two novel procedures have therefore been proposed. First, the template perturbation (referred to as ECCp) consists in adding to the raw streamflow forecast a perturbation with a suitable dependence structure, to handle the cases where it is non-dispersive. It improves the skills of both the ECC-Q and ECC-T methods, and succeeds in maintaining simultaneously the univariate and multivariate calibration. This procedure is similar in some ways to the dual-ECC variant proposed by Ben Bouallègue *et al.* (2016), but was found to perform better. The second procedure, trajectory smoothing (TS), enables to obtain realistic trajectories, while the effects on the forecast skill are negligible, if not beneficial.

Combining these procedures has conducted to a set of different mixed methods. In overall, we argue for using ECC-Q coupled with the template perturbation and the trajectory smoothing procedures, i.e., ECCp-Q-TS. It is the mixed method that shows the best forecast skills, and produces ensemble forecasts that are fairly calibrated regarding both univariate and multivariate aspects, in addition to the trajectories being realistic. Note that these two additional procedures increase the complexity and computational cost of ECC-Q only to a minor extent.

To conclude, while we have found ECC-based methods to outperform a parametric technique such as GCA, the latter has the advantage not to depend on the size of the raw ensemble. For instance, ECC-based methods do not suit the context where a deterministic streamflow forecast based on a single meteorological forcing would be dressed with an uncertainty envelop. Note that the Schaake shuffle or adaptations thereof, which are based on observation records, are also free of any constraints regarding the ensemble size, and could be employed as well in this specific context. Furthermore, in setups where hydrological modeling involves different models (as it is the case in our study), ECC-based methods can rely only the information from a single model. Further investigation on these aspects would therefore be necessary.

Acknowledgments

Joseph Bellier's research is supported by grants from Labex OSUG@2020 (Investissement d'avenir – ANR 10 LABX56) and Compagnie Nationale du Rhône. The authors gratefully acknowledge IRSTEA (Antony) for sharing the GRP hydrological model, and Michael Scheuerer for the EMOS precipitation postprocessing code. They also thank the three reviewers for their helpful comments and suggestions. Streamflow observations data were provided by Compagnie Nationale du Rhône. These data, as well as the raw streamflow forecasts and the BMA postprocessed distributions, can be obtained on request for research purpose with the signature of a data provision agreement (contact: g.bontron@cnr.tm.fr).

Chapitre 9

Quels maillons pour constituer la chaîne de prévision ?

Une chaîne de prévision hydrologique est assimilable à un « système complexe », dans le sens où elle est construite à partir d'un grand nombre d'éléments qui interagissent de manière non simple (en témoigne par exemple la non-linéarité de la modélisation hydrologique, ou encore le comportement chaotique de l'atmosphère). Cette complexité augmente encore lorsque l'on considère une chaine probabiliste. Afin de tenter de la réduire, nous avons opté jusqu'à maintenant pour une approche modulaire, en identifiant des étapes distinctes qui puissent être développées de manière indépendante : choix du forçage, prétraitement, modélisation hydrologique, post-traitement.

Arrivés au dernier chapitre de ce travail de thèse, nous adoptons une approche intégrée afin de répondre à plusieurs questions essentielles que soulève le déploiement d'une chaîne de prévision probabiliste opérationnelle. Il faut s'assurer, d'une part, que chacune des améliorations apportées sur un des maillons de la chaîne soit bénéfique quels que soient les choix fait concernant les autres maillons. Ensuite, ces améliorations ont un coût, qui peut être d'ordre financier (accès aux forçages météorologiques par exemple), humain (compétences requises pour développer les algorithmes et assurer leur support), ou encore informatique (ressources de calcul). Il semble alors important de hiérarchiser leur apport respectif dans la chaîne de prévision complète.

Pour cela, nous proposons d'évaluer les performances de chacun des maillons lorsqu'ils sont intégrés à des chaînes de prévision de complexité variable. Cela implique la construction de multiples « combinaisons », ainsi que le choix de critères d'évaluation. Ces aspects méthodologiques seront détaillés dans la section 9.1. Ensuite, nous identifions dans la section 9.2 quelques questions précises, et tentons d'y répondre en comparant les performances de certaines combinaisons choisies.

TABLE 9.1 – Maillons retenus pour chacune des étapes de la chaîne de prévision.

	Nom	Remarques	Nbr. de membres en sortie
Forçage météorologique	`ECMWF`	Correspond à ECMWF-Ens.	51
	`GEFS`		21
	`grand ens.`	Grand ensemble constitué de ECMWF-Ens et GEFS.	72
Pré-traitement	`brut`	Aucun pré-traitement.	Idem entrée
	`EMOS rdm`	Pré-traitement univarié EMOS-CSG (précipitation) et EMOS-normal (température), et structure de dépendance aléatoire.	Idem entrée (sauf combinaison avec **grand ens.** où on revient à 51)
	`EMOS ECC`	La structure de dépendance est cette fois reconstruite via l'ECC.	Idem entrée (sauf combinaison avec **grand ens.** où on revient à 51)
Modélisation hydrologique	`GRP`	Plus précisément, c'est le couplage Cemaneige-GRP.	Idem entrée
	`multi mod.`	Modélisation multi-modèle à l'aide des modèles GRP, TOPMODEL et ARX (couplés à Cemaneige).	Multiplié par 3
Post-traitement	`brut`	Aucun post-traitement.	Idem entrée
	`BMA ECC`	Post-traitement univarié BMA, et reconstruction de la structure de dépendance via ECCp-Q-TS .	Idem entrée

9.1 Présentation du plan d'expérience

9.1.1 Choix des maillons retenus

Comme rappelé en introduction de ce chapitre, quatre étapes distinctes ont été identifiées dans la chaîne de prévision : forçage météorologique, pré-traitement, modélisation hydrologique, et post-traitement. Les chapitres précédents nous ont permis de discuter, pour chacune de ces étapes, de différentes approches (maillons) permettant d'améliorer les prévisions. Désormais, nous proposons de sélectionner un nombre restreint de maillons, de manière à pouvoir tester toutes les combinaisons possibles entre ces maillons. Les choix que nous avons fait quant aux maillons sélectionnés sont résumés dans le Tableau 9.1.

Pour ce qui est des forçages météorologiques, le choix a été dicté par la disponibilité des archives. Ainsi, seuls ECMWF-Ens et GEFS, qui fournissent les prévisions de précipitations et de température sur une période suffisamment longue, sont retenus. Par ailleurs, nous constituons également un « grand ensemble » à partir de ces deux forçages. Un grand ensemble peut être défini comme la combinaison de plusieurs ensembles provenant de centres météorologiques différents. L'objectif est alors de profiter des différences entre les systèmes (résolution, schémas de paramétrisation, méthodes d'assimilation, de perturbation, etc) pour agrandir le spectre des valeurs possibles pour la prévision d'une

variable météorologique donnée. La manière de procéder la plus simple est de mélanger les membres des différents ensembles en attribuant le même poids à chacun des membres. Les bénéfices de grands ensembles ainsi construits ont été démontrés par de nombreuses études qui se sont appuyées sur la base de données TIGGE, qui archive les prévisions d'ensemble d'une dizaine de centres météorologiques. Sans être exhaustif, nous pouvons citer Park *et al.* (2008), Johnson et Swinbank (2009) et Hagedorn *et al.* (2012), qui ont démontré les bénéfices dans un contexte de prévision météorologiques, ou encore Pappenberger *et al.* (2008) et He *et al.* (2009) qui se sont intéressés aux grand ensembles dans la prévision des crues. Un récapitulatif plus complet des études menées grâces à TIGGE est disponible dans Swinbank *et al.* (2016). Notre grand ensemble, qui comprend 72 membres, a une taille très modeste comparée à celles des études citées plus haut qui comprennent parfois plus de 200 membres. Cependant, il permet d'entrevoir les bénéfice éventuels de cette stratégie, d'autant plus que les ensembles ECMWF-Ens et GEFS diffèrent sur bien des aspects, ce qui est souhaitable.

Concernant les étapes de pré-traitement et post-traitement, nous comparons les stratégies avec et sans correction, en considérant les méthodes ayant montré les meilleures performances lors des chapitres précédents [1]. Pour le pré-traitement, nous proposons de rajouter un maillon correspondant à la correction statistique avec un réarrangement aléatoire, de manière à quantifier l'apport d'un réarrangement sophistiqué des prévisions météorologiques dans une chaîne de prévision complète.

Enfin, il nous a semblé légitime, pour la modélisation hydrologique, de comparer le multi-modèle au modèle hydrologique donnant les meilleures performance en validation, à savoir GRP.

Ainsi, en choisissant un maillon par étape pour constituer les combinaisons, nous obtenons $3 \times 3 \times 2 \times 2 = 36$ combinaisons possibles. Pour classer ces combinaisons entre elles, il nous faut désormais choisir des critères de vérification.

9.1.2 Critères de vérification

Nous nous plaçons dans le même contexte que le chapitre précédent, à savoir une vérification portant sur les bassins Arve, Valserine, Usses, Fier, Séran et Guiers, sur la période 2011-2014. L'objectif de ce chapitre étant d'évaluer des chaînes de prévision complètes, nous considérons des critères de vérification qui s'appliquent sur les prévisions de débit en sortie. Pour cela, nous suggérons d'utiliser l'approche de vérification multivariée proposée dans la section 8.3, qui a permis de discriminer entre les différentes méthodes de reconstruction de prévisions hydrologiques cohérentes.

Cette approche consiste à évaluer des prévisions de vecteurs multivariés contenant les débits à différent bassins et échéances. Afin de prendre en compte la cohérence spatiale et temporelle tout en réduisant la dimension des vecteurs, nous considérons les vecteurs comprenant les 6 bassins mais les échéances $6, 12, 18, \ldots, 120$ h seulement. Ces prévisions

1. Pour mémoire, la méthode AnSS-ECC a montré des résultats légèrement meilleurs que ECC pour le réarrangement après pré-traitement, mais nous ne l'avons pas utilisé car elle ne s'applique pas trivialement au réarrangement précipitation-température (cf. 6.2).

multivariées, de dimension $6 \times 20 = 120$, sont confrontées aux observations de débit (également sous formes de vecteurs de dimension 120) via les scores de compétence $\text{CRPSS}_{\text{sum}}$, ESS et VSS. La combinaison choisie pour servir de référence dans le calcul des scores de compétence est : `ECMWF` − `brut` − `GRP` − `brut`.

Enfin, pour évaluer la fiabilité, nous traçons les histogrammes de rang des prévisions de la fonctionnelle « somme » de ces mêmes vecteurs, de manière analogue au calcul du score $\text{CRPSS}_{\text{sum}}$. Cette approche d'évaluation de la fiabilité a déjà été entreprise dans le chapitre précédent (cf Figure 8.11).

9.2 Résultats

9.2.1 Vue d'ensemble

La Figure 9.1 présente les performances de l'intégralité des combinaisons. Bien que les trois scores ne classent pas les combinaisons exactement dans le même ordre, il apparaît que, de manière générale, la sophistication de la chaîne de prévision va de pair avec une amélioration des performances. Ainsi, au regard des trois scores, les meilleures performances sont obtenues avec la combinaison la plus sophistiquée, à savoir la combinaison : `grand ens.` − `EMOS ECC` − `multi mod.` − `BMA ECC`.

En présentant ces résultats de manière différente, il est possible de montrer que chacun des 4 maillons de la combinaison ci-dessus donne de meilleures performances que ses alternatives au sein de la même étape de modélisation, et ce quel que soient les maillons retenus pour les autres étapes de modélisation. En d'autres termes, nous montrons que :

- `grand ens.` est systématiquement meilleur que `ECMWF` et `GEFS`,
- `EMOS ECC` est systématiquement meilleur que `EMOS rdm` et `brut` [2],
- `multi mod.` est systématiquement meilleur que `GRP`,
- `BMA ECC` est systématiquement meilleur que `brut`.

Ce gain peut devenir faible pour certaines combinaisons, cependant les améliorations apportées sur les différentes étapes n'annulent jamais le gain apporté par le choix du maillon le plus sophistiqué pour une étape de modélisation donnée. Cela valide donc notre stratégie « modulaire » d'amélioration de la chaîne de prévision.

Au delà de ces résultats, il est intéressant d'identifier les maillons qui permettent les gains de performance les plus importants. Ce travail s'inscrit dans une optique de maximisation du ratio coût/bénéfice de la chaîne de prévision complète. C'est pourquoi nous sélectionnons, dans les sections suivantes, quelques questions spécifiques, et tentons d'y répondre en comparant les performances de certaines combinaisons choisies. Dans les figures proposées, les maillons qui diffèrent d'une combinaison à une autre sont indiqués en gras.

2. Excepté dans certains cas lorsque l'on regarde le score VSS, mais cela est discuté dans la section 9.2.4.

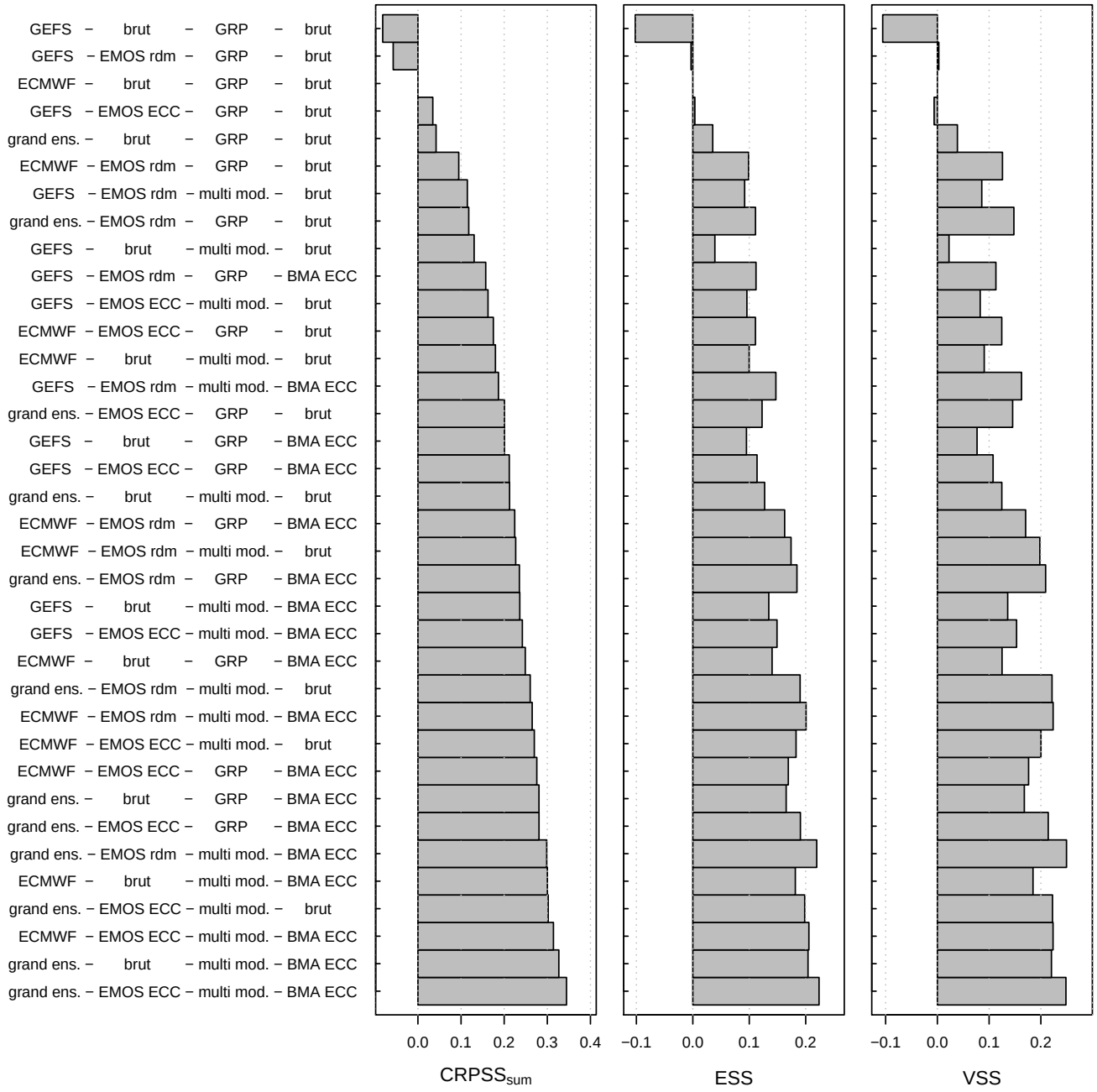

FIGURE 9.1 – CRPSS$_{\text{sum}}$, ESS et VSS de l'ensemble des combinaisons, rangés par ordre croissant de CRPSS$_{\text{sum}}$. La combinaison de référence est : `ECMWF` − `brut` − `GRP` − `brut`.

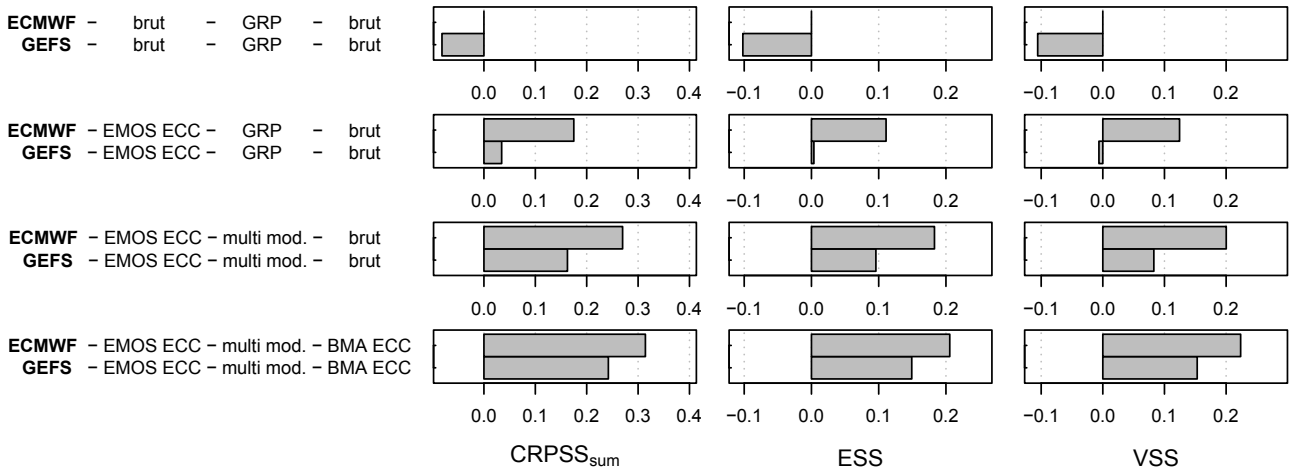

FIGURE 9.2 – Comparaison des forçages `ECMWF` et `GEFS`, pour des chaînes de prévision de complexité croissante.

9.2.2 De l'intérêt d'un forçage performant en entrée

L'accès aux forçages météorologiques, dans un contexte de prévision opérationnelle, peut représenter un coût non négligeable. Il peut alors être tentant de choisir un forçage moins performant mais « facile d'accès », et de miser sur une chaîne de prévision sophistiquée pour corriger les défauts initiaux. La question qui se pose alors est la suivante :

Les écarts de performances entre différents forçages perdurent-ils lorsque la chaîne de prévision se complexifie ?

Pour répondre à cette question, nous comparons à la Figure 9.2 les forçages `ECMWF` et `GEFS`, lorsqu'ils sont intégrés à des chaînes de prévision de complexité croissante. Il apparaît alors que la différence de performance entre les deux forçages se réduit peu lorsque la chaîne se complexifie. Notamment, le pré-traitement ne permet pas de rattraper un manque de performance initial du forçage brut. Au contraire, l'ajout du maillon `EMOS ECC` semble même accentuer les écarts entre `ECMWF` et `GEFS`. Ainsi, le choix du forçage en entrée est déterminant, et ce quelle que soit la sophistication de la chaîne de prévision.

9.2.3 Affiner l'incertitude météo : grand ensemble ou pré-traitement ?

Si le forçage est déterminant, alors il est logique de chercher à l'améliorer. La Figure 9.2 a permis d'entrevoir le gain significatif apporté par le pré-traitement. Celui-ci peut néanmoins représenter des contraintes opérationnelles majeures, comme par exemple la disponibilité d'une archive de prévisions passées suffisamment longue et homogène [3]. Une alternative possible pour s'affranchir de cette contrainte est la constitution d'un grand

3. Dans cette thèse, nous avons utilisé des archives contenant les prévisions émises au fur et à mesure par les centres de prévision, et par conséquent ces archives ne sont pas parfaitement homogènes (contrairement à des jeux de *reforecasts*). Pour réduire cette non-homogénéité, nous avons dû écarter certaines périodes et/ou modèles (cf. 2.3.1.1) pour lesquels des changements de version rendaient impossible l'application d'un pré-traitement.

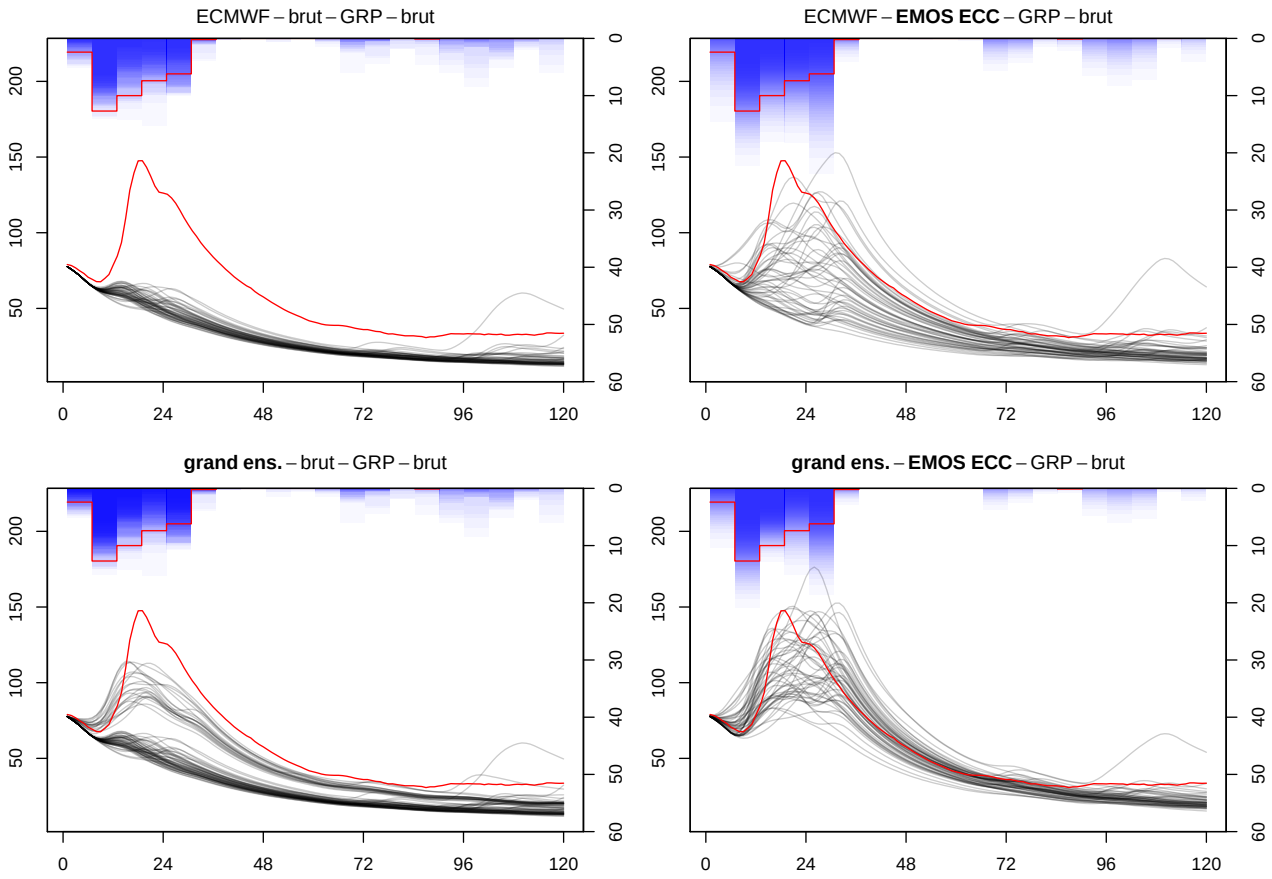

FIGURE 9.3 – Exemple de prévision du 4 janvier 2014 sur la Valserine, pour différentes stratégies d'amélioration du forçage météorologique. L'axe des débits (à gauche) est en m^3/s, tandis que l'axe des précipitations (à droite) est en mm/6h. L'observation est tracée en rouge.

ensemble à partir de plusieurs forçages, à condition d'avoir accès à ces forçages de manière opérationnelle. Nous nous posons ainsi la question suivante :

Vaut-il mieux combiner les forçages disponibles pour former un grand ensemble, ou bien pré-traiter correctement le meilleur d'entre eux ?

En se tournant vers le pré-traitement statistique, on cherche à corriger et amplifier le « signal » contenu dans le forçage brut, tandis qu'en formant un grand ensemble on cherche davantage à capter d'autres signaux qui auraient été potentiellement absents dans le premier forçage. Cette interprétation s'illustre parfaitement dans l'exemple de la Figure 9.3, où le forçage `ECMWF − brut` ne parvient pas à déclencher une réaction hydrologique. D'un côté, `ECMWF − EMOS ECC` parvient à amplifier le signal brut de manière à ce qu'un certain nombre de membres réagissent. De l'autre côté, `grand ens. − brut` apporte, grâce à GEFS, un signal que `ECMWF − brut` avait manqué, et par conséquent la prévision laisse entendre qu'une réaction est également possible. Précisons que, pour cet example hivernal, ce signal ne provient pas des précipitations mais davantage des températures, qui sont plus élevées pour GEFS que pour `ECMWF − brut`, et donc produisent davantage de précipitations liquides. Enfin, combiner les deux approches semble ici bénéfique, en témoigne la prévision `grand ens. − EMOS ECC` où la totalité des membres prévoient une réaction hydrologique (qui a été effectivement observée).

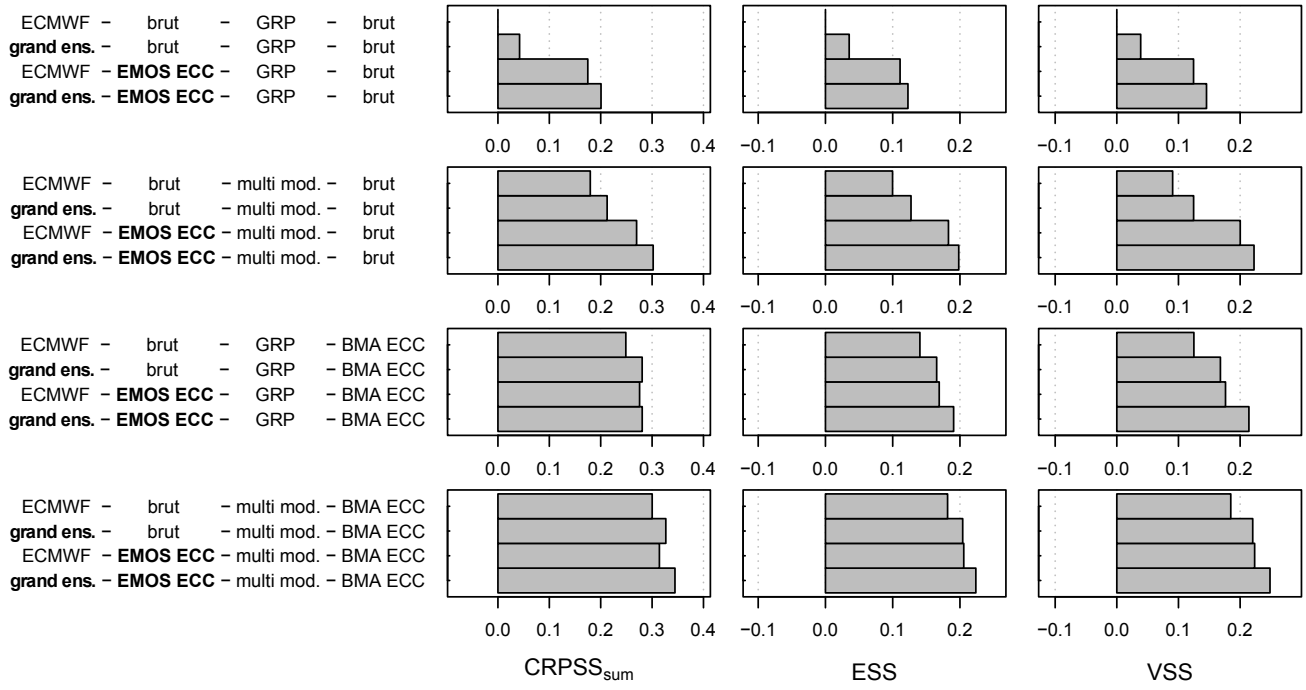

FIGURE 9.4 – Comparaison des performances obtenues avec différentes stratégies d'amélioration du forçage météorologique, pour différents niveaux de complexité de la chaîne aval de prévision.

La Figure 9.4 compare les performances de ces deux stratégies d'amélioration du forçage, pour différents niveaux de complexité de la chaîne « aval » de prévision (i.e., la chaîne à partir de la modélisation hydrologique). Lorsque la chaîne aval ne comporte pas de post-traitement, le gain apporté par le pré-traitement est supérieur à celui apporté par le grand ensemble. En effet, contrairement au grand ensemble, le pré-traitement assure une correction à minima de la fiabilité, même si cela concerne le forçage météorologique. En revanche, dès lors que la chaîne aval comporte un post-traitement, l'apport du grand ensemble est égal si ce n'est supérieur à celui du pré-traitement. Ainsi, en réponse à la question posée, notre grand ensemble est préférable au meilleur forçage pré-traité à condition seulement que la chaîne de prévision comporte un maillon de correction statistique (pré-traitement ou post-traitement).

La Figure 9.5 permet d'évaluer l'effet de ces deux stratégies d'amélioration du forçage sur la fiabilité des prévisions de débit. Les histogrammes de rang (des prévisions ramenées à un cadre univariée grâce à la fonctionnelle « somme ») qui sont tracés concernent les mêmes combinaisons qu'à la Figure 9.4, mais pour deux niveaux seulement de complexité de la chaîne aval. Il s'avère alors que les deux stratégies (grand ensemble et pré-traitement) ont toutes deux un effet sur la fiabilité des prévisions de débit relativement faible comparé à l'effet de la chaîne aval de prévision.

9.2.4 De l'importance de la cohérence des forçages

Pré-traiter « correctement » un forçage requiert la reconstruction de la structure de dépendance multivariée, étape longuement étudiée dans le chapitre 6. On peut néanmoins

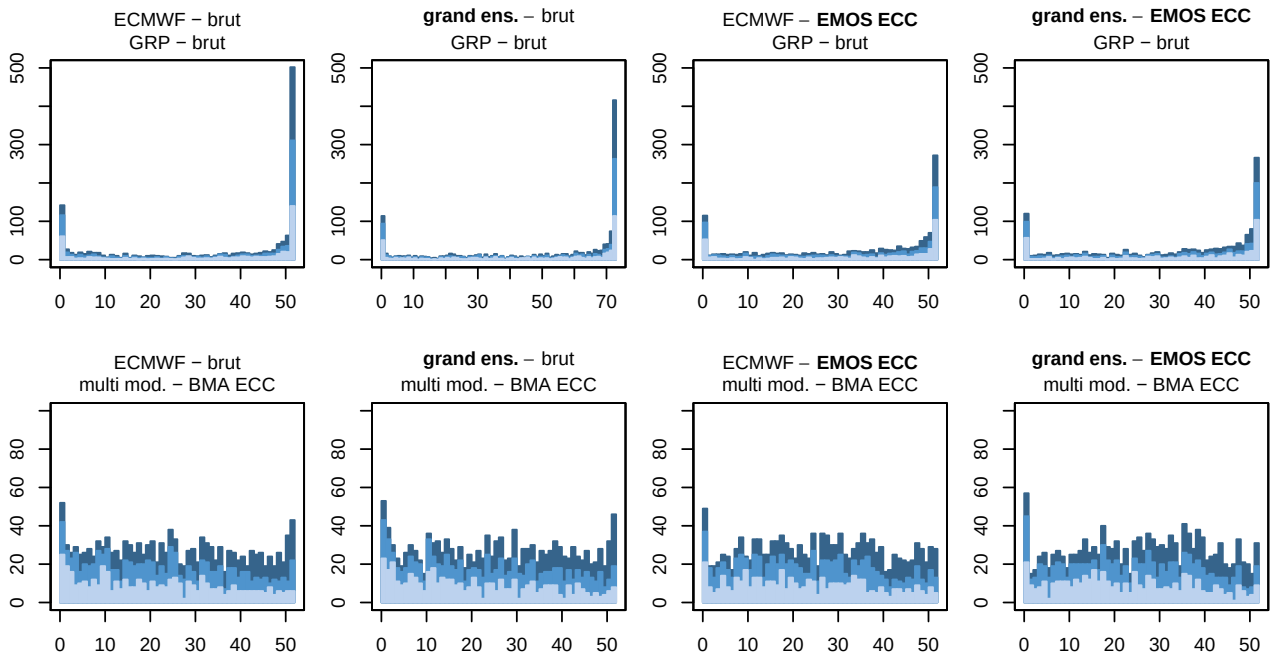

FIGURE 9.5 – Fiabilité des prévisions de débit pour différentes stratégies d'amélioration du forçage météorologique, pour la chaîne aval la moins sophistiquée (haut) et la plus sophistiquée (bas). Les histogrammes de rang sont calculés sur les prévisions multivariées ramenées à un cadre univarié via la fonctionnelle « somme ».

se poser la question de sa pertinence, le reste de la chaîne de prévision (en l'occurrence le post-traitement) étant susceptible de « casser » de nouveau la cohérence des prévisions.

Si l'étape de reconstruction de forçages cohérents est négligée, dans quelle mesure cela impacte t-il les performances de la chaîne complète de prévision ?

La Figure 9.6 permet de comparer les performances obtenues avec `EMOS ECC` et `EMOS rdm` pour différentes complexités de la chaîne aval de prévision. Les scores $\text{CRPSS}_{\text{sum}}$, ESS et VSS donnent ici des résultats parfois contradictoires. Ainsi, le VSS des combinaisons incluant `EMOS rdm` est très proche de celui des combinaisons incluant `EMOS ECC`, voir légèrement supérieur dans le cas de la chaîne aval la moins sophistiquée.

Ce résultat surprenant peut s'expliquer par le fait que le score VSS est insensible à un biais dans les prévisions, tandis qu'il est très sensible à leur structure de dépendance. Rappelons que la différence entre `EMOS ECC` et `EMOS rdm` réside dans la structure multivariée des forçages météorologiques, tandis que la vérification porte sur les prévisions multivariées de débit. Ainsi, les combinaisons incluant `EMOS rdm` s'appuient sur des forçages météorologiques non cohérents ce qui causera, sur les prévisions de débit qui en résultent, des erreurs dans la moyenne. Toutefois, ces prévisions de débit auront une structure temporelle réaliste, du fait du caractère intégrateur de la modélisation hydrologique (voir par exemple l'hydrogramme « Random » de la Figure 6.8). Appliqué aux prévisions de débit, le score VSS pénalise donc peu `EMOS rdm`. En revanche, le score $\text{CRPSS}_{\text{sum}}$ est bien plus sensible à la tendance centrale, et par conséquent il pénalise davantage la non-cohérence des forçages météorologiques.

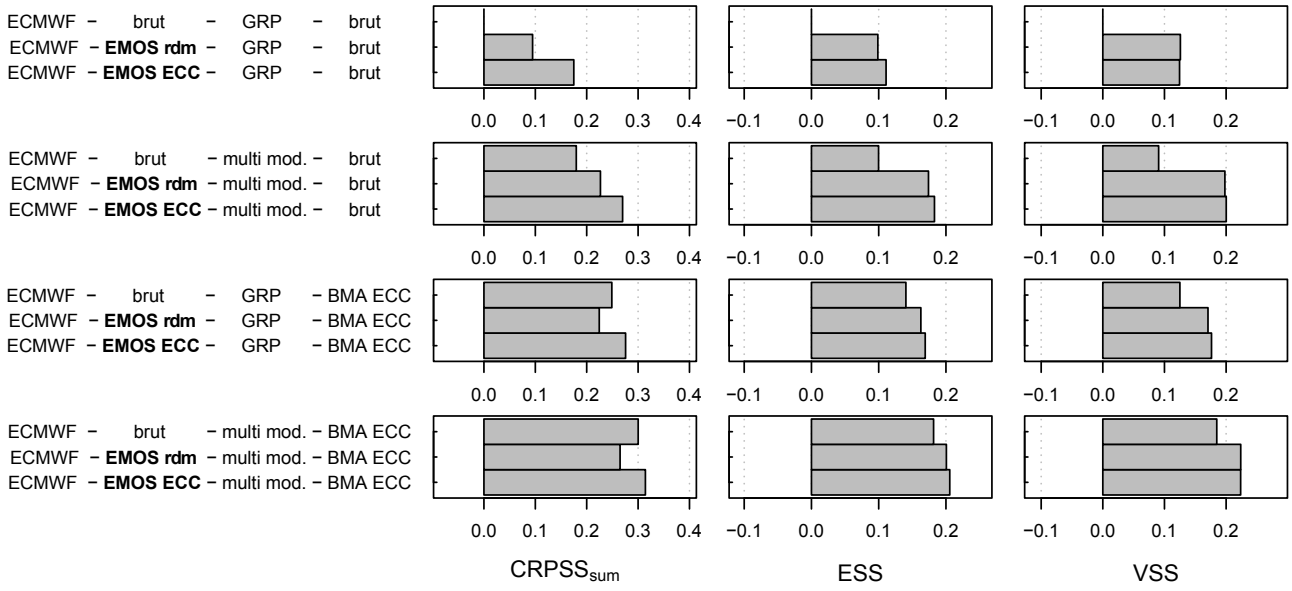

FIGURE 9.6 – Comparaison des performances de différentes chaînes de prévision incluant le pré-traitement EMOS, avec ou sans reconstruction de la structure de dépendance via l'ECC.

Selon ce score $\text{CRPSS}_{\text{sum}}$, un pré-traitement « complet » (i.e., `EMOS ECC`) apporte systématiquement un gain par rapport à un pré-traitement « incomplet » (i.e., `EMOS rdm`) ou bien par rapport à l'utilisation des forçages bruts. Cependant, plus la chaîne de prévision est sophistiquée, moins ce gain est significatif. Par ailleurs, dans le cas d'une chaîne incluant un post-traitement, le pré-traitement incomplet `EMOS rdm` est moins pertinent que l'utilisation des forçages bruts. Nous dirons alors qu'il vaut mieux « ne rien faire » plutôt que de « faire les choses à moitié » !

9.2.5 Quelle stratégie pour prendre en compte l'incertitude hydrologique ?

Intéressons-nous désormais à la partie aval de la chaîne de prévision, qui repose en premier lieu sur la modélisation hydrologique. Nous avons montré dans le chapitre 7 que la stratégie multi-modèle permettait d'améliorer les performances, grâce au fait qu'elle représente une partie de l'incertitude de modélisation. Un environnement multi-modèle peut cependant s'avérer coûteux, à la fois lors de sa mise en place (déploiement informatique des modèles, calage, etc.) et dans son fonctionnement opérationnel (il requiert davantage de ressources de calcul, le nombre de membres étant démultiplié). Choisir un modèle unique puis ajouter un post-traitement statistique est alors une alternative qui semble moins coûteuse. Cependant, le post-traitement est une correction statistique qui, comme le pré-traitement, repose sur l'homogénéité de l'archive des prévisions passées. Toute modification d'un des éléments situés à l'amont (changement des forçages, recalage du modèle hydrologique, etc) nécessite alors une mise à jour de ses paramètres. Cette mise à jour peut s'avérer chronophage, du fait du positionnement en bout de chaîne du post-traitement qui requiert, à chaque modification d'un élément, de refaire tourner toute la chaîne pour reformer une archive de prévisions passées. Ainsi, il nous semble judicieux

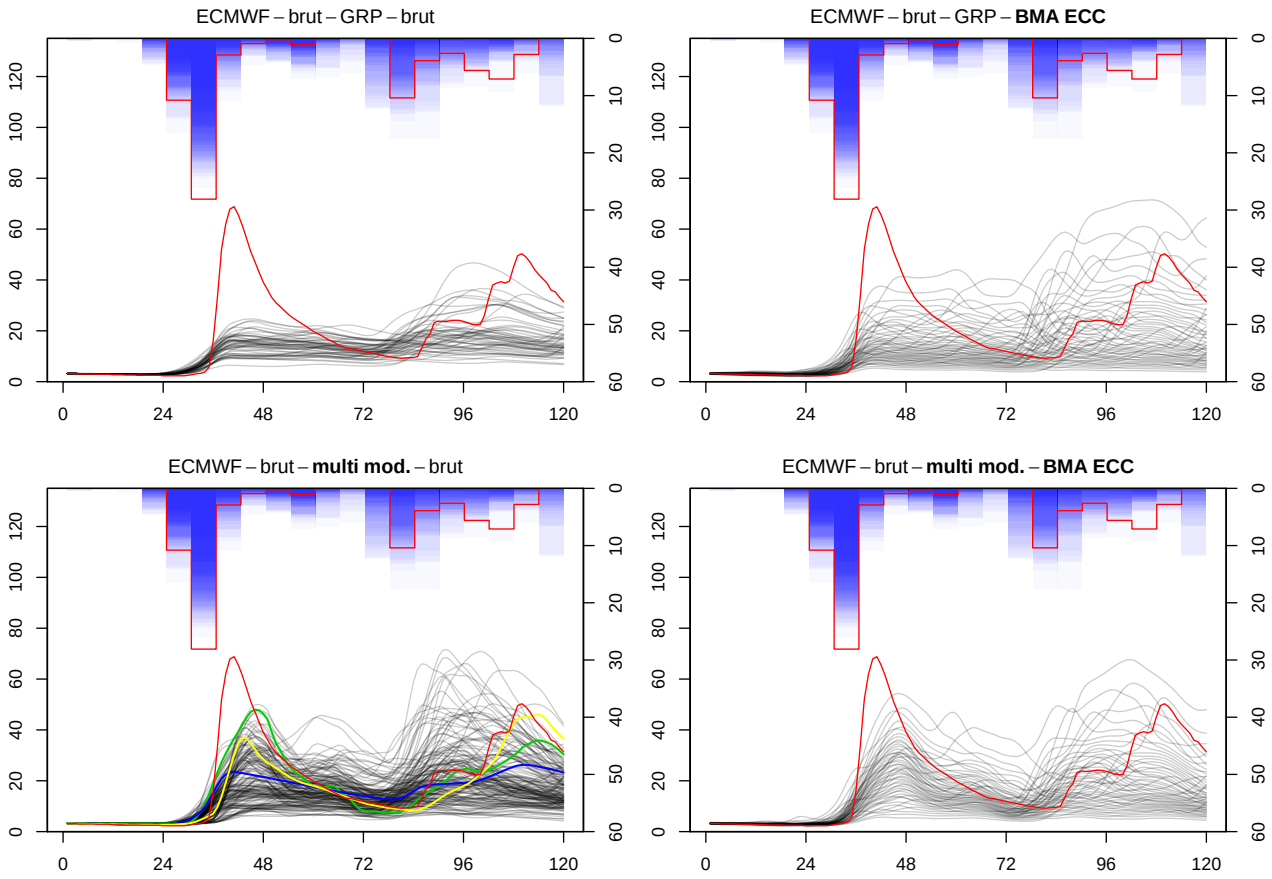

FIGURE 9.7 – Exemple de prévision du 29 avril 2014 sur le Guiers, pour différentes stratégies d'amélioration des prévisions hydrologiques. Sur le graphique `ECMWF − brut − multi mod. − brut`, les débits simulés (avec les forçages observés) par les modèles GRP, TOPMODEL et ARX sont tracés en bleu, vert et jaune, respectivement.

de se poser la question suivante :

Pour prendre en compte l'incertitude hydrologique, vaut-il mieux opter pour une approche multi-modèle, ou bien post-traiter correctement les prévisions issues du meilleur modèle ?

La question est finalement assez similaire à celle posée précédemment à propos des forçages météorologiques (en 9.2.3), du fait de l'analogie entre `grand ens.` et `multi mod.` d'une part, et `EMOS ECC` et `BMA ECC` d'autre part. Ainsi, est-il préférable de corriger et amplifier le signal contenu dans les prévisions d'un seul modèle hydrologique, ou bien de miser sur le fait qu'une variété de modèles hydrologiques simulera des comportements différents, et par conséquent élargira le spectre des valeurs possibles ? La Figure 9.7 illustre bien ces deux angles de vue, pour un exemple de prévision où la configuration standard `GRP − brut` a très fortement sous-estimé le pic de débit. Ici, `GRP − BMA ECC` parvient certes à augmenter la dispersion, mais la faiblesse du signal brut fait que le premier pic est « manqué ». La stratégie `multi mod. − brut` semble alors intéressante, grâce aux modèles ARX et TOPMODEL qui apportent un comportement différent de celui de GRP. Même si ces deux modèles ont des performances en simulation en moyenne inférieures à celles de GRP sur l'échantillon, ils peuvent se montrer davantage pertinents sur certains évènements.

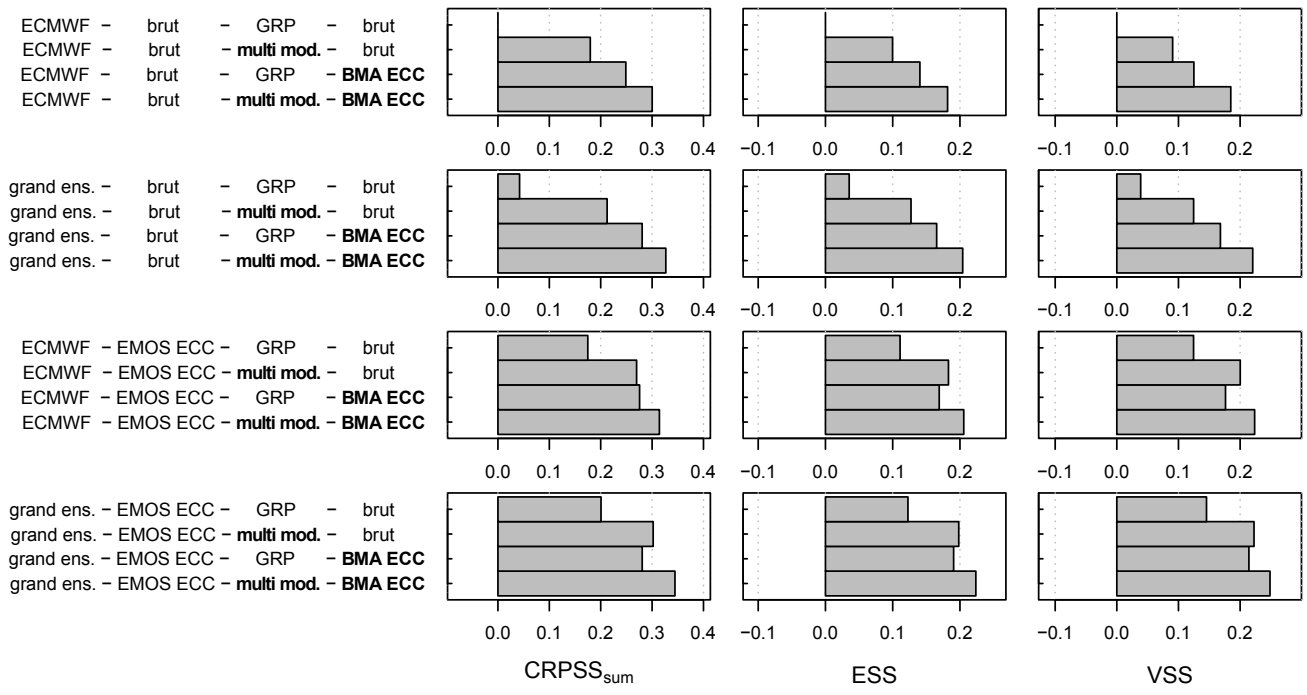

FIGURE 9.8 – Comparaison des performances obtenues avec différentes stratégies d'amélioration des prévisions hydrologiques, pour différents niveaux de complexité de la chaîne amont.

Cependant, en rajoutant à la Figure 9.7 les débits simulés (avec les forçages observés) des trois modèles hydrologiques, on se rend compte qu'il aurait été impossible, même avec une prévision météorologique parfaite, de prévoir correctement le pic de débit. En effet, les trois modèles sous-estiment (dans des proportions différentes) la réaction hydrologique. Ceci illustre le constat établi en section 7.2, à savoir que notre approche multi-modèle permet d'améliorer la qualité globale des prévisions, mais qu'elle reste insuffisante pour capturer entièrement l'incertitude hydrologique.

La Figure 9.8, qui compare les gains en performance obtenus avec ces deux stratégies pour différentes complexités de la chaîne amont, permet de répondre à la question posée. Ainsi, la stratégie `multi mod.` – `brut` est préférable à celle consistant à post-traiter les prévisions issues du meilleur modèle (i.e., `GRP` – `BMA ECC`), à condition seulement que la chaîne de prévision comporte un pré-traitement. Dans le cas contraire, c'est la seconde des stratégies qui est la plus pertinente. Enfin, combiner les deux approches (i.e., `multi mod.` – `BMA ECC`) permet d'obtenir les meilleurs résultats, et ce quel que soit la chaîne amont de prévision.

L'effet de ces deux approches sur la fiabilité des prévisions est illustrée à la Figure 9.9. Nous retrouvons le même constat que celui dressé dans le chapitre 7, à savoir que le multi-modèle seul n'est pas capable de produire des prévisions fiables, contrairement au post-traitement. Par ailleurs, ce constat est valable quel que soit le forçage météorologique, qu'il soit sophistiqué (`grand ens.` – `EMOS ECC` ; histogrammes du bas) ou non (`ECMWF` – `brut` ; haut).

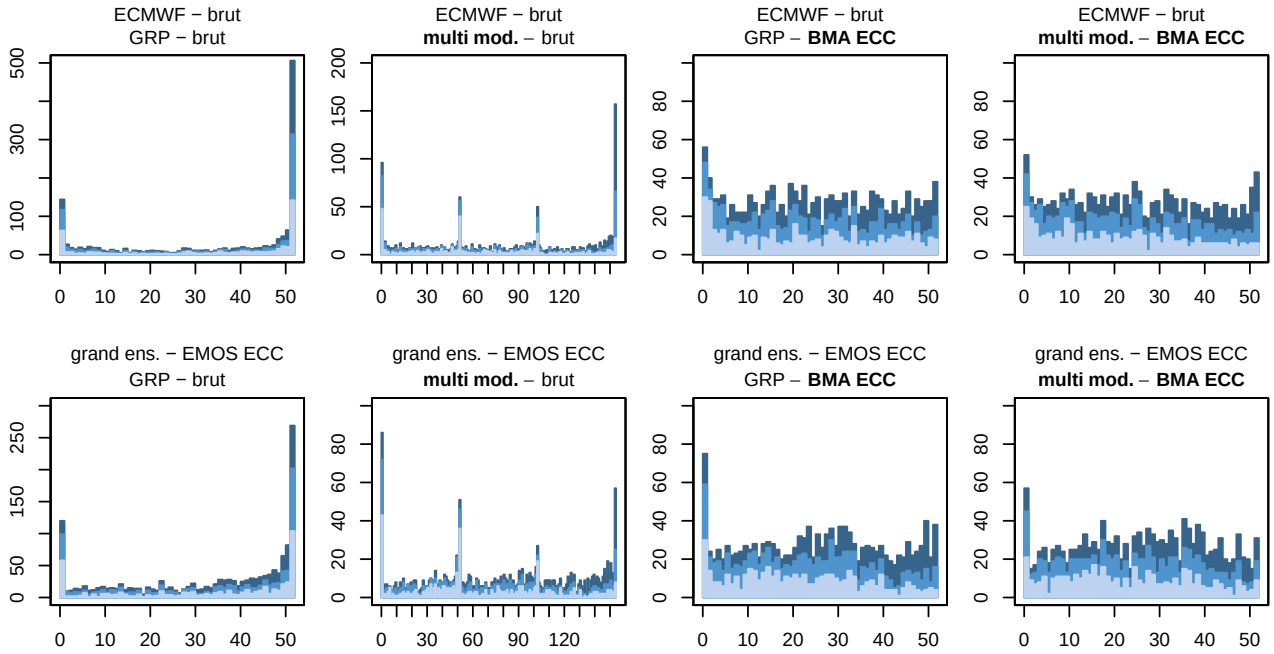

FIGURE 9.9 – Fiabilité des prévisions pour différentes stratégies d'amélioration des prévisions hydrologiques, pour la chaîne amont la moins sophistiquée (haut) et la plus sophistiquée (bas).

9.2.6 Pré-traitement ou post-traitement ?

En termes de performance globale (au travers des scores), les résultats discutés jusqu'à présent ont montré qu'il était essentiel d'inclure au moins un pré-traitement ou un post-traitement statistique dans la chaîne de prévision. Ainsi, la dernière question que nous nous posons est la suivante :

Doit-on opter en priorité pour un pré-traitement ou un post-traitement ?

Avec un pré-traitement seul, on réduit les biais des prévisions météorologiques, et par conséquent on minimise le risque de « manquer » des évènements hydrologiques pour cause de seuils critiques non franchis dans le modèle hydrologique. Rappelons que ces biais peuvent concerner les précipitations mais également la température, qui jouera un rôle crucial dans les évènements hivernaux où elle avoisine les 0°C. En revanche, cette stratégie ne traite pas de manière statistique l'incertitude hydrologique, et par conséquent les prévisions de débit risquent d'être sous-dispersives. A l'inverse, un post-traitement seul est supposé garantir une fiabilité correcte des prévisions de débit, du fait de son positionnement en bout de chaîne où il agrège l'incertitude météorologique manquante dans le forçage brut avec l'incertitude hydrologique. Cependant, il peut arriver que cette incertitude totale soit fortement décorrélée de la situation météorologique. C'est le cas notamment dans l'exemple illustré à la Figure 9.10, où les prévisions `ECMWF – brut – GRP – BMA ECC` sont certes bien davantage dispersives que celles sans post-traitement, mais cette dispersion représente peu la dynamique des évènements prévus de précipitation, contrairement aux prévisions `ECMWF – EMOS ECC – GRP – brut`.

La Figure 9.11 confirme que l'une et l'autre stratégie conduisent systématiquement à une amélioration significative des performances. Ce résultat est valable que l'on enrichisse

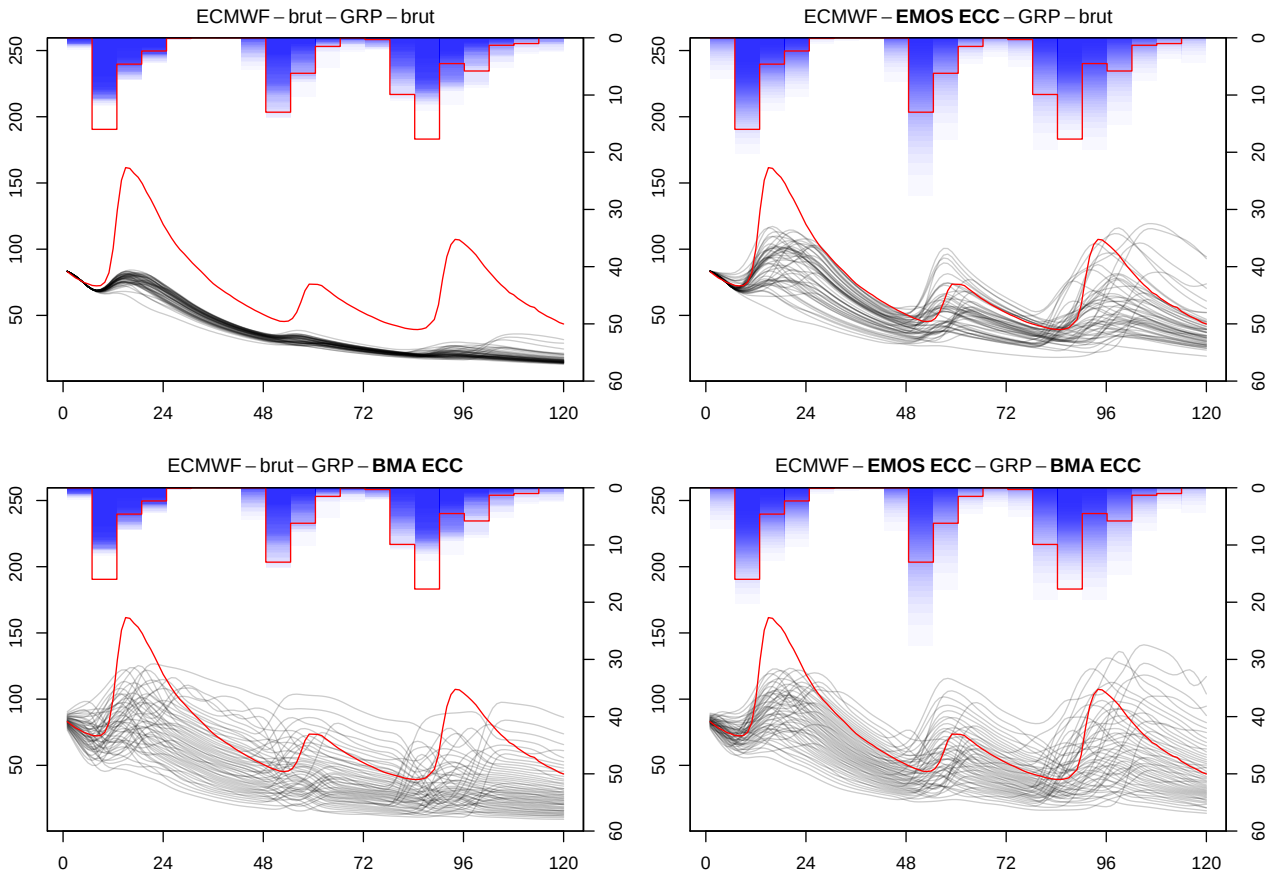

FIGURE 9.10 – Exemple de prévision du 2 janvier 2012 sur la Valserine, pour différents choix quant au positionnement de la correction statistique dans la chaîne de prévision.

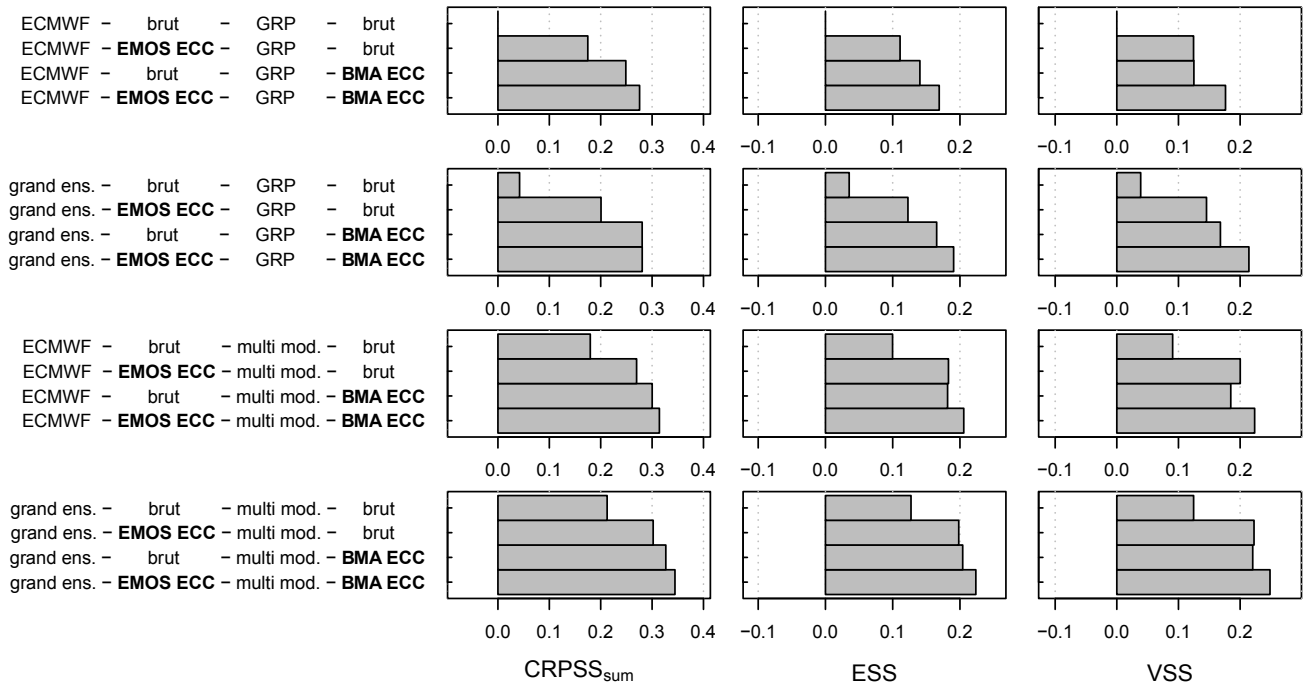

FIGURE 9.11 – Comparaison des performances obtenues avec les stratégies de pré-traitement et post-traitement, pour différents niveaux de complexité du reste de la chaîne de prévision.

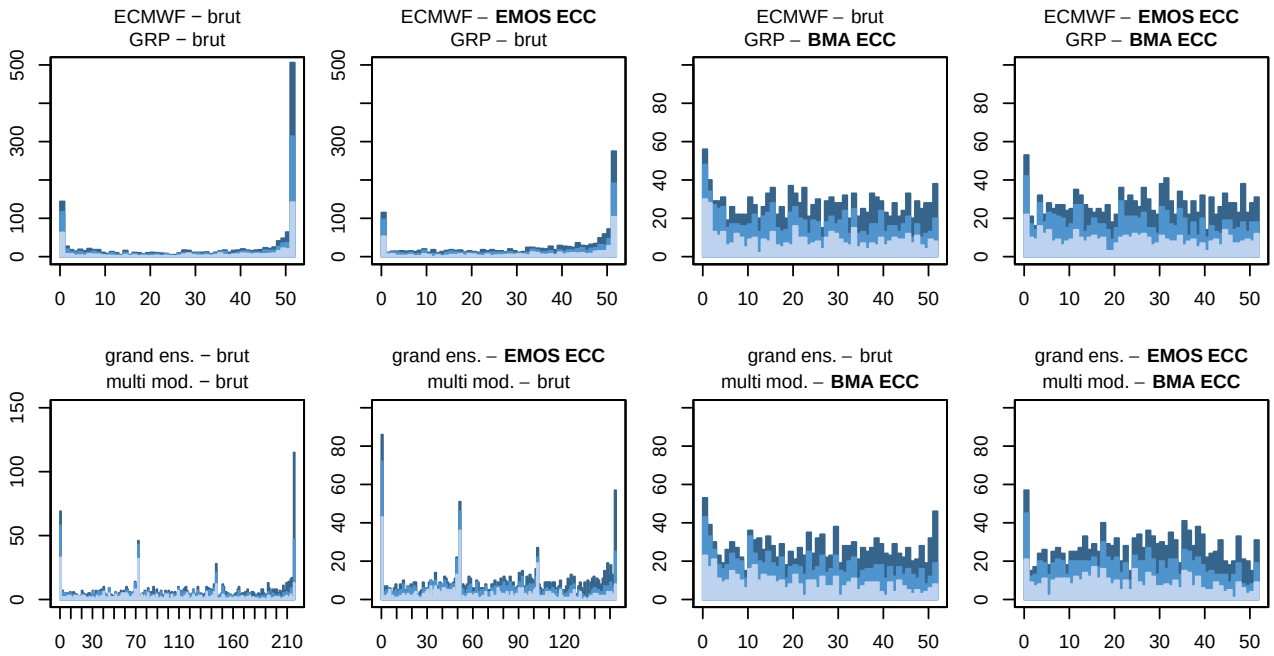

FIGURE 9.12 – Fiabilité des prévisions pour les stratégies de pré-traitement et post-traitement, pour deux niveaux (haut et bas) de complexité du reste de la chaîne.

les forçages (via le grand ensemble) ou non, et/ou que l'on fasse appel à du multi-modèle hydrologique ou non. Néanmoins, si l'on devait se limiter à une seule des corrections, les résultats montrent qu'il vaut mieux privilégier le post-traitement. Et ce d'autant plus que le post-traitement est l'unique maillon de la chaîne permettant de garantir la fiabilité des prévisions de débit (ce résultat est confirmé par la Figure 9.12). Enfin c'est bien l'utilisation conjointe du pré-traitement et du post-traitement qui conduit aux meilleurs performances.

Ces résultats diffèrent quelque peu des conclusions tirées dans de précédents travaux ayant cherché à quantifier l'apport du pré-traitement dans une chaîne de prévision hydrologique. Kang *et al.* (2010) et Zalachori *et al.* (2012) ont notamment montré que l'effet du pré-traitement s'effaçait avec la modélisation hydrologique. Verkade *et al.* (2013) ont obtenu des résultats similaires, et avancé plusieurs raisons. Tout d'abord, ils pointent la très forte non-linéarité de leur bassin d'étude (le Rhin) et par conséquent de leur modèle hydrologique, qui n'est alors pas suffisamment sensible à la correction et l'amplification du signal réalisée par le pré-traitement. Il est ainsi possible que les bassins de notre zone d'étude, qui sont par ailleurs de taille bien inférieure, présentent une non-linéarité moindre. Verkade *et al.* (2013) discutent également du fait que leurs prévisions après pré-traitement sont réarrangées via le Schaake shuffle, une méthode qui ne permet pas de conditionner la structure de dépendance à chaque situation de prévision. Cette méthode est également utilisée par Kang *et al.* (2010) [4]. Or, les résultats que nous avons obtenus dans le chapitre 6 ont montré que le Schaake shuffle réduisait très significativement le gain apporté par le pré-traitement (cf. Figure 6.11). Il est donc possible que leurs conclusions

4. Concernant Zalachori *et al.* (2012), l'article ne mentionne pas la problématique du réarrangement après pré-traitement.

aient été différentes si une version plus sophistiquée du Schaake shuffle ou bien l'approche ECC avait été utilisée. Comme l'a montré la Figure 9.6, lorsqu'un post-traitement est présent dans la chaîne, il vaut mieux « ne rien faire » en matière de pré-traitement que de « faire les choses à moitié ». Enfin, la qualité des forçages bruts ainsi que la méthode de pré-traitement utilisée jouent probablement un rôle important. En effet, de l'amplitude des corrections apportées dépend la capacité des forçages pré-traités à franchir des seuils critiques dans le modèle hydrologique qui ne l'auraient pas été sans pré-traitement.

9.3 Synthèse

L'objectif de ce chapitre était, dans un premier temps, de valider la stratégie modulaire adoptée jusqu'alors pour le développement de la chaîne de prévision. En d'autres termes, de vérifier que chacune des améliorations apportées sur le forçage brut, le pré-traitement, la modélisation hydrologique ou encore le post-traitement, n'engendrait pas de dégradations des performance lorsque combinée à d'autres améliorations. Le second objectif était de dégager certaines priorités à mettre en œuvre en vue de la mise en place d'une chaîne opérationnelle. Nous avons ainsi adopté une démarche intégrée, en évaluant les prévisions de débit obtenues avec différentes combinaisons de maillons pour former la chaîne de prévision. Les principales conclusions sont les suivantes :

- Contrairement au dicton, le « mieux n'est pas l'ennemi du bien ». En d'autres termes, les performances des prévisions de débit sont systématiquement améliorées en choisissant pour une étape donnée le maillon le plus performant.

- Le choix du forçage est déterminant, car les différences de performance obtenues avec différents forçages persistent quels que soient les autres maillons de la chaîne.

- Les stratégies visant à agrandir l'ensemble de manière ensembliste (grand ensemble et multi-modèle hydrologique) améliorent sensiblement les performances, et ce d'autant plus que la chaîne de prévision comporte au moins une correction statistique (pré-traitement ou post-traitement).

- Pré-traitement et post-traitement sont tous deux souhaitables. Néanmoins, si un choix devait vraiment être fait, ce serait en faveur du post-traitement, car son positionnement en bout de chaîne permet de garantir la fiabilité des prévisions hydrologiques.

Conclusion générale

La prévision des débits est une activité intrinsèquement entachée d'incertitudes. Être en mesure de quantifier ces incertitudes tout en les réduisant autant que possible est un défi scientifique, auquel nous espérons avoir participé grâce à ce travail de thèse. Nous nous sommes placés dans un contexte multivarié impliquant plusieurs échéances et plusieurs bassins, aux débits plus ou moins corrélés. L'objectif était de définir les méthodes à mettre en œuvre pour assurer la fiabilité et la cohérence spatio-temporelle des prévisions.

Pour clore ce mémoire, nous proposons de résumer la démarche adoptée, de synthétiser les principaux résultats obtenus, et enfin de proposer plusieurs perspectives pour des travaux futurs.

La démarche adoptée

En premier lieu, nous avons cherché à nous inscrire dans une démarche probabiliste où le but est de s'assurer de la fiabilité des prévisions avant de chercher à maximiser leur finesse. En effet, c'est seulement si les prévisions sont fiables que le caractère probabiliste peut être exploité.

La chaîne de prévision sépare deux objets d'étude que sont les prévisions météorologiques (i.e., le forçage), qui concernent dans notre cas les variables précipitation et température, et les prévisions de débit. Comme l'on dispose pour l'un et l'autre d'observations, il est possible de quantifier et de chercher à améliorer la qualité de ces deux objets de manière indépendante ; c'est la démarche qui a été adoptée. Restait encore à en vérifier la pertinence, plusieurs études ayant montré que le gain apporté par l'amélioration des forçages s'effaçait après la modélisation hydrologique.

Pour prendre en compte les incertitudes de prévision, nous avons combiné des approches ensemblistes avec des méthodes de correction statistique. L'objectif de l'approche ensembliste est de construire un ensemble de simulations s'appuyant sur des « bases » légèrement différentes (conditions initiales, paramètres et structure des modèles, etc.) mais néanmoins cohérentes au regard de l'incertitude qui règne sur notre connaissance des systèmes. Ces simulations se dispersent dans le temps, générant de manière dynamique une incertitude propre à chaque situation. Cette approche de prévision probabiliste permet en théorie de traiter adéquatement et sans redondance les différentes sources d'incertitude. Dans cette thèse, nous avons employé la prévision d'ensemble pour l'incertitude météorologique, et le multi-modèle pour l'incertitude hydrologique.

La correction statistique poursuit en revanche un objectif plus pragmatique : se servir

des erreurs passées pour corriger à posteriori la prévision, par exemple en augmentant la dispersion des membres, ou en corrigeant un biais systématique. Cette correction a été appliquée à la fois sur les prévisions météorologiques et hydrologiques. Nous avons alors adopté la convention sémantique de la communauté hydrologique, qui consiste à nommer ces deux corrections « pré-traitement » et « post-traitement », en référence à leur positionnement de part et d'autre du modèle hydrologique. La démarche adoptée pour l'une et l'autre correction a été similaire, à savoir :

1. Corriger les distributions de probabilité des prévisions dans un cadre statistique univarié. Cela permet de garantir la fiabilité, mais entraîne la perte de la structure de dépendance spatio-temporelle présente dans les prévisions ensemblistes brutes.

2. Reconstruire des prévisions ensemblistes multivariées cohérentes.

Les principaux résultats obtenus

Tout d'abord, nos résultats ont mis en lumière l'importance du forçage météorologique. Les différentes tentatives d'amélioration, que ce soit le changement du forçage brut (de GEFS à ECMWF-Ens), le passage à un grand ensemble ou encore le pré-traitement statistique (via l'approche EMOS), ont systématiquement conduit à une amélioration des prévisions de débit. Par ailleurs, nous avons montré que le choix du forçage brut en entrée était déterminant, les différences de performance persistant tout au long de la chaîne, et ce, quelle qu'en soit sa complexité.

Dès lors que ce forçage météorologique subit un pré-traitement univarié, il est nécessaire de reconstruire, via une procédure ad hoc, une structure de dépendance qui permette d'obtenir des prévisions multivariées cohérentes. Nous nous sommes intéressés tout particulièrement à la structure de dépendance spatio-temporelle des prévisions de précipitation, et avons montré qu'elle jouait un rôle crucial dans la modélisation hydrologique. Un exercice d'inter-comparaison de méthodes a permis de pointer les limites d'une méthode très couramment utilisée, le Schaake shuffle. En exploitant le principe d'analogie, nous avons proposé des variantes plus à même de conditionner la structure de dépendance à la situation de prévision. Au final, nous avons retenu pour cette étape de pré-traitement la méthode existante ECC, qui reproduit la structure de dépendance du forçage brut. Bien qu'une des méthodes développées ait été légèrement plus performante dans notre expérience sur les précipitations, le choix de retenir l'ECC a été motivé par, d'une part, l'avantage qu'elle présente de ne nécessiter aucune autre source de données que le forçage brut et, d'autre part, le fait qu'elle s'applique naturellement à la cohérence inter-variable (ici, précipitation et température).

L'incertitude qui entache la prévision des variables météorologiques ne fait guère de doute, ainsi l'utilisation de prévisions d'ensemble ou autres forçages ensemblistes s'est largement démocratisée dans la prévision hydrologique opérationnelle. En revanche, l'incertitude inhérente à la modélisation hydrologique est moins souvent prise en compte. Nous avons testé l'approche multi-modèle hydrologique, en nous appuyant sur un modèle purement statistique (ARX) et deux modèles conceptuels, l'un à l'inspiration davantage

physique (TOPMODEL) et l'autre plutôt empirique (GRP). Cette approche a permis, par sa prise en compte de manière non statistique d'une partie de l'incertitude, d'obtenir des gains significatifs de performance. Cependant, elle s'est avérée insuffisante pour prendre en compte toute l'incertitude, produisant ainsi des prévisions de débit encore fortement sous-dispersives.

Le post-traitement statistique via la méthode BMA a permis d'obtenir des prévisions de débit fiables. Cette méthode est en effet alléchante grâce à sa capacité à combiner et pondérer des prévisions provenant de plusieurs modèles, tout en les habillant de l'incertitude encore non prise en compte. S'il est possible d'appliquer la BMA à partir des prévisions de débits provenant d'un seul modèle hydrologique, c'est bien la combinaison avec le multi-modèle qui a permis d'obtenir les meilleurs résultats.

Comme le pré-traitement, le post-traitement des prévisions de débit entraine la perte de la structure de dépendance. Or, celle-ci est bien souvent cruciale pour l'utilisateur, qui a besoin de prévoir des scénarios multivariés et non pas des valeurs indépendantes pour chaque bassin et échéance. Nous avons ainsi développé différentes variantes de la méthode ECC, de manière à satisfaire les contraintes qu'impose la prévision des débits, notamment la forte autocorrélation, ainsi que la fréquente non-dispersion des prévisions brutes lors des phases de récession.

En construisant pas à pas notre chaîne de prévision probabiliste, nous avons constaté que les méthodes ensemblistes utilisées (la prévision d'ensemble, le grand ensemble météorologique ou le multi-modèle hydrologique) étaient indispensables pour générer dynamiquement une dispersion propre à chaque prévision, mais s'avéraient encore insuffisantes pour rendre compte de l'incertitude totale. Une correction statistique, en un endroit au moins de la chaîne de prévision, est donc indispensable. Nous avons cherché à savoir quelle stratégie devait être prioritaire entre le pré-traitement et le post-traitement. Il s'est avéré que le post-traitement était indispensable pour produire des prévisions hydrologiques fiables, du fait de son positionnement en bout de chaîne où il « reprend » l'incertitude manquante. Cependant, contrairement à d'autres études, nous avons observé que l'effet du pré-traitement ne s'effaçait pas complètement lors de la modélisation hydrologique, et qu'il permettait d'augmenter les performances des prévisions de débits. En d'autres termes, corriger les prévisions météorologiques permet parfois d'amplifier suffisamment un signal pour qu'il franchisse un seuil critique dans le modèle hydrologique, et donc que le gain se retrouve sur les prévisions de débit. Enfin, c'est bien la combinaison des deux corrections, pré-traitement et post-traitement, qui permet d'obtenir les meilleurs résultats.

Quelques perspectives

Mener ce travail jusqu'à son terme aura exigé de faire des choix ici et là, mais en notant malgré tout précieusement les nombreuses pistes que le temps imparti ne nous permettait pas d'explorer. Nous proposons désormais de restituer celles qui semblent à nos yeux les plus importantes.

Mieux prendre en compte les incertitudes Une prévision (probabiliste) fiable c'est bien, mais une prévision fiable et fine, c'est mieux ! Hormis l'amélioration intrinsèque des outils de simulation, une meilleure finesse des prévisions peut être obtenue en identifiant et traitant des sources d'incertitude mieux définies, l'idée étant d'obtenir des prévisions dont la dispersion est la plus spécifique possible à chaque situation.

Cette démarche peut être entreprise d'une part en poursuivant les travaux démarrés sur le forçage météorologique. Nous avons notamment montré l'intérêt d'enrichir le forçage via un grand ensemble, sans toutefois explorer suffisamment le lien avec le prétraitement. S'il semble en effet prometteur de combiner différentes sources de prévision météorologique, la combinaison de ces sources est un aspect non moins essentiel. Par exemple, les modèles météorologiques déterministes à haute résolution fournissent des prévisions performantes sur les courtes échéances mais, noyées sans pondération au milieu d'un ensemble ou d'un grand ensemble, elles auront un poids négligeable. Il sera par ailleurs nécessaire, en parallèle, de développer la procédure ECC de manière à ce qu'elle puisse exploiter ces différentes sources d'information, en s'affranchissant notamment de l'hypothèse d'échangeabilité des membres.

Concernant la modélisation hydrologique d'autre part, il existe plusieurs pistes potentielles qui permettraient une meilleure prise en compte des incertitudes. Tout d'abord, la stratégie multi-modèle que nous avons entreprise mériterait être approfondie. Par exemple, dans quelle mesure le rajout ou le retrait d'un modèle impacte t-il les performances ? Par ailleurs, nous avons mentionné diverses stratégies ensemblistes alternatives, comme par exemple le multi-jeux de paramètres, dont il serait intéressant de tester la pertinence. Enfin, l'assimilation de données dans le modèle hydrologique est une piste qui nous semble particulièrement importante. Les méthodes ensemblistes telles que le filtre particulaire ou le filtre de Kalman d'ensemble sont attrayantes, car permettant de quantifier l'incertitude (vue par le modèle) de l'état initial du bassin. L'objectif ainsi visé est de se rapprocher de l'incertitude intrinsèque à des processus hydrologiques mieux définis, et par conséquent de générer une dispersion qui soit davantage spécifique à chaque situation. Bien qu'attractives, ces perspectives sont cependant susceptibles d'entraîner une inflation du nombre de membres des prévisions ensemblistes. Est-ce alors compatible avec les contraintes de temps de calcul d'une chaîne opérationnelle ?

Comment intégrer nos outils dans une chaîne opérationnelle ? Outre la réflexion nécessaire sur le temps de calcul, l'intégration des outils dans un environnement opérationnel de prévision hydrologique requiert de se placer dans un nouveau contexte. Les objectifs restent les mêmes, à savoir garantir la fiabilité et la cohérence des prévisions. En revanche de nouvelles problématiques apparaissent.

La première concerne l'intégration de l'expertise. Dans une chaîne de prévision déterministe, un prévisionniste peut considérablement améliorer les prévisions en s'appuyant sur son expertise. En revanche, les contours de son action au sein d'une chaîne probabiliste sont complexes à définir, car la chaîne inclue des méthodes de correction statistique qui reposent sur l'homogénéité des prévisions passées. Dès lors que les prévisions sont modifiées « à la main », on réduit cet homogénéité, et par conséquent il devient plus difficile

de garantir la fiabilité des prévisions. S'il est fort dommage de se passer de l'expérience des prévisionnistes, des travaux futurs sont nécessaires pour comprendre comment mieux combiner expertise et fiabilité des prévisions.

Par ailleurs, aboutir à des prévisions de débit en tout point du linéaire d'un fleuve comme le Rhône nécessite la modélisation d'affluents ayant un fonctionnement hydrologique non naturel, du fait, entre autres, d'acteurs opérant des aménagements hydroélectriques avec capacité de stockage. La prévision sur de tels bassins ne relève alors guère plus de la modélisation pluie-débit, mais fait davantage intervenir la modélisation de systèmes autres tels que les acteurs économiques. La fiabilité des prévisions de débits sur ces affluents sera une tâche délicate, de même que leur cohérence avec celles des affluents aux régimes hydrologiques plus naturels.

Étendre le domaine d'application de notre travail La démarche de développement poursuivie dans cette thèse, autour notamment de l'architecture entre (i) les approches ensemblistes, (ii) la correction statistique univariée, et (iii) la reconstruction de prévisions multivariées cohérentes, pourrait s'étendre à d'autres domaines d'application que la prévision météorologique et hydrologique.

Nous pensons en premier lieu à la problématique de la propagation hydraulique, qui est la dernière étape avant la production de prévisions probabilistes des débits en tout point du linéaire d'un fleuve collecteur d'affluent. On peut alors se poser les questions suivantes, qui doivent désormais avoir pour le lecteur un air de déjà-vu : la fiabilité résiste t-elle à la modélisation hydraulique, ou bien est-il nécessaire de tenir compte de sources d'incertitude supplémentaires ? Une nouvelle correction statistique est-elle nécessaire, et si oui, se substitue t-elle aux pré-traitement et post-traitement précédents ? Comment reconstruire ensuite des scénarios de débits cohérents à l'échelle de plusieurs échéances et stations sur le fleuve ?

La problématique de la cohérence de prévisions probabilistes multivariées concerne de manière plus large le champ de la prévision de production d'énergie renouvelable. De nombreux producteurs d'énergie tentent aujourd'hui de diversifier leur porte-feuille d'actifs de production. Ce travail de thèse est ainsi une étape vers la prévision de scénarios météorologiques cohérents à l'échelle de plusieurs variables (débit, vent, rayonnement solaire) et sur un large domaine spatial.

Bibliographie

Abramowitz, G., et Gupta, H. (2008). Toward a model space and model independence metric. *Geophysical Research Letters*, 35(5).

Addor, N., Jaun, S., Fundel, F., et Zappa, M. (2011). An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrology and Earth System Sciences*, 15(7), 2327–2347.

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9(7), 1518–1530.

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., et Lettenmaier, D. P. (2016). Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. *Water Resources Research*, 52(6), 4209–4225.

Baran, S., Horanyi, A., et Nemoda, D. (2014). Comparison of the BMA and EMOS statistical methods in calibrating temperature and wind speed forecast ensembles. *Quarterly Journal of the Hungarian Meteorological Service*, 118(3), 217–241.

Baran, S., et Möller, A. (2015). Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics*, 26(2), 120–132.

Baran, S., et Möller, A. (2017). Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature. *Meteorology and Atmospheric Physics*, 129(1), 99–112.

Baran, S., et Nemoda, D. (2016). Censored and shifted gamma distribution based emos model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27(5), 280–292.

Bauer, P., Thorpe, A., et Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567), 47-55.

Bellier, J., Bontron, G., et Zin, I. (2017b). Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. *Water Resources Research*, 53(12), 10085–10107.

Bellier, J., Zin, I., et Bontron, G. (2017a). Sample stratification in verification of ensemble forecasts of continuous scalar variables: potential benefits and pitfalls. *Monthly Weather Review*, 145(9), 3529–3544.

Bellier, J., Zin, I., Siblot, S., et Bontron, G. (2016). Probabilistic flood forecasting on the Rhone River: Evaluation with ensemble and analogue-based precipitation forecasts. *E3S Web of Conferences (FLOODrisk 2016)*, 7.

Ben Bouallègue, Z., Heppelmann, T., Theis, S. E., et Pinson, P. (2016). Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Monthly Weather Review*, 144(12), 4737–4750.

Ben Daoud, A., Sauquet, E., Bontron, G., Obled, C., et Lang, M. (2016). Daily quantitative precipitation forecasts based on the analogue method: Improvements and application to a French large river basin. *Atmospheric Research*, 169, 147–159.

Ben Daoud, A., Sauquet, E., Lang, M., Bontron, G., et Obled, C. (2011). Precipitation forecasting through an analog sorting technique: a comparative study. *Advances in Geosciences*, 29, 103–107.

Bennett, J. C., Robertson, D. E., Shrestha, D. L., Wang, Q., Enever, D., Hapuarachchi, P., et Tuteja, N. K. (2014). A system for continuous hydrological ensemble forecasting (SCHEF) to lead times of 9 days. *Journal of Hydrology*, 519, 2832–2846.

Berner, J., Shutts, G., Leutbecher, M., et Palmer, T. (2009). A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, 66(3), 603–626.

Berrocal, V. J., Raftery, A. E., et Gneiting, T. (2007). Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135(4), 1386–1402.

Berthet, L. (2010). *Prévision des crues au pas de temps horaire : pour une meilleure assimilation de l'information de débit dans un modèle hydrologique* (Thèse de doctorat). AgroParisTech & Cemagref.

Berthet, L., Andréassian, V., Perrin, C., et Javelle, P. (2009). How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments. *Hydrology and Earth System Sciences*, 13(6), 819–831.

Beven, K. J. (2012). *Rainfall-runoff modelling*: *The primer* (2e éd.). Chichester, UK : Wiley.

Beven, K. J., et Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, 249(1-4), 11–29.

Beven, K. J., et Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Journal*, 24(1), 43–69.

Beven, K. J., Lamb, R., Quinn, P., Romanowicz, R., et Freer, J. (1995). TOPMODEL. In V. P. Singh (Ed.), *Computer models of watershed hydrology* (pp. 627–668). Colorado, USA : Water Resources Publications.

Blasone, R.-S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A., et Zyvoloski, G. A. (2008). Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. *Advances in Water Resources*, 31(4), 630–648.

Bogner, K., et Pappenberger, F. (2011). Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resources Research*, 47(7).

Bogner, K., Pappenberger, F., et Cloke, H. (2012). Technical note: The normal quantile transformation and its application in a flood forecasting system. *Hydrology and Earth System Sciences*, 16(4), 1085–1094.

Bompart, P., Bontron, G., Celie, S., et Haond, M. (2009). Une chaîne opérationnelle de prévision hydrométéorologique pour les besoins de la production hydroélectrique de la cnr. *La Houille Blanche*(5), 54–60.

Bontron, G. (2004). *Prévision quantitative des précipitations*: *adaptation probabiliste par recherche d'analogues. Utilisation des réanalyses NCEP/NCAR et application aux précipitations du sud-est de la France* (Thèse de doctorat). Institut National Polytechnique de Grenoble.

Boucher, M.-A., Tremblay, D., Delorme, L., Perreault, L., et Anctil, F. (2012). Hydro–economic assessment of hydrological forecasting systems. *Journal of Hydrology*, 416, 133–144.

Bourgin, F. (2014). *Comment quantifier l'incertitude prédictive en modélisation hydrologique ? Travail exploratoire sur un grand échantillon de bassins versants* (Thèse de doctorat). AgroParisTech & Irstea.

Box, G. E. P., et Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, 26(2), 211–252.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., et Ljung, G. M. (2015). *Time series analysis*: *Forecasting and control* (5^{e} éd.). Hoboken, NJ, USA : Wiley.

Bröcker, J. (2008). On reliability analysis of multi-categorical forecasts. *Nonlinear Processes in Geophysics*, 15(4), 661–673.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643), 1512–1519.

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, 138(667), 1611–1617.

Buizza, R. (2008). The value of probabilistic prediction. *Atmospheric Science Letters*, 9(2), 36–42.

Buizza, R., Leutbecher, M., et Isaksen, L. (2008). Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, 134(637), 2051–2066.

Buizza, R., Milleer, M., et Palmer, T. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560), 2887–2908.

Caillouet, L. (2016). *Reconstruction hydrométéorologique des étiages historiques en france entre 1871 et 2012* (Thèse de doctorat). Université Grenoble Alpes & Irstea.

Candille, G., Côté, C., Houtekamer, P., et Pellerin, G. (2007). Verification of an ensemble prediction system against observations. *Monthly Weather Review*, 135(7), 2688–2699.

Candille, G., et Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131(609), 2131–2150.

Casati, B., Wilson, L., Stephenson, D., Nurmi, P., Ghelli, A., Pocernich, M., ... Mason, S. (2008). Forecast verification: Current status and future directions. *Meteorological Applications*, 15(1), 3–18.

Clark, M. P., Gangopadhyay, S., Hay, L., Rajagopalan, B., et Wilby, R. (2004). The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1), 243–262.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., ... others (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, 51(4), 2498–2514.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., ... Hay, L. E. (2008). Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44(12).

Cloke, H. L., et Pappenberger, F. (2009). Ensemble flood forecasting: a review. *Journal of Hydrology*, 375(3-4), 613–626.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., et Andréassian, V. (2017). The suite of lumped GR hydrological models in an R package. *Environmental Modelling and Software*, 94, 166–171.

DeChant, C. M., et Moradkhani, H. (2011). Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. *Hydrology and Earth System Sciences*, 15(11), 3399–3410.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., ... others (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), 553–597.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., et Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141(10), 3498–3516.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., ... others (2014). The science of NOAA's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society*, 95(1), 79–98.

Demeritt, D., Nobert, S., Cloke, H., et Pappenberger, F. (2010). Challenges in communicating and using ensembles in operational flood forecasting. *Meteorological applications*, 17(2), 209–222.

Demirel, M. C., Booij, M. J., et Hoekstra, A. Y. (2013). Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. *Water Resources Research*, 49(7), 4035–4053.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., et Cébron, P. (2014). PEARP, the Météo-France short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 141(690), 1671–1685.

Diaconescu, E. P., et Laprise, R. (2012). Singular vectors in atmospheric sciences: A review. *Earth-Science Reviews*, 113(3-4), 161–175.

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

Duan, Q., Sorooshian, S., et Gupta, V. K. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. *Journal of Hydrology*, 158(3-4), 265–284.

Efron, B., et Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL, USA : CRC press.

Elmore, K. L. (2005). Alternatives to the Chi-square test for evaluating rank histograms from ensemble forecasts. *Weather and Forecasting*, 20(5), 789–795.

Engeland, K., et Steinsland, I. (2014). Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times. *Water Resources Research*, 50(1), 182–197.

Ferro, C. A., Richardson, D. S., et Weigel, A. P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15(1), 19–24.

Fortin, V., Favre, A.-C., et Said, M. (2006). Probabilistic forecasting from ensemble prediction systems: improving upon the best-member method by using a different weight and dressing kernel for each member. *Quarterly Journal of the Royal Meteorological Society*, 132(617), 1349–1369.

Fraley, C., Raftery, A. E., et Gneiting, T. (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using bayesian model averaging. *Monthly Weather Review*, 138(1), 190–202.

Friederichs, P., et Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7), 579–594.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., et Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24(5), 1416–1430.

Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B., et Jackson, B. (2009). MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review*, 137(1), 246–268.

Gneiting, T., Balabdaoui, F., et Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, 69B(2), 243–268.

Gneiting, T., et Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., et Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5), 1098–1118.

Gneiting, T., et Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3), 411–422.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., et Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2), 211-264.

Gupta, H. V., Kling, H., Yilmaz, K. K., et Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2), 80–91.

Hagedorn, R. (2008). Using the ECMWF reforecast dataset to calibrate EPS forecasts. *ECMWF Newsletter*, 117, 8–13.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., et Palmer, T. (2012). Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138(668), 1814–1827.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3), 550–560.

Hamill, T. M. (2012). Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Monthly Weather Review*, 140(7), 2232–2252.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau Jr, T. J., ... Lapenta, W. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, 94(10), 1553–1565.

Hamill, T. M., et Colucci, S. J. (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, 125(6), 1312–1327.

Hamill, T. M., et Colucci, S. J. (1998). Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, 126(3), 711–724.

Hamill, T. M., Hagedorn, R., et Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly weather review*, 136(7), 2620–2632.

Hamill, T. M., et Juras, J. (2006). Measuring forecast skill: is it real skill or is it the varying climatology?. *Quarterly Journal of the Royal Meteorological Society*, 132(621), 2905–2924.

Hamill, T. M., et Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Monthly Weather Review*, 134(11), 3209–3229.

Hastie, T., et Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398), 371–386.

He, Y., Wetterhall, F., Cloke, H., Pappenberger, F., Wilson, M., Freer, J., et McGregor, G. (2009). Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Meteorological Applications*, 16(1), 91–101.

Hemri, S., Fundel, F., et Zappa, M. (2013). Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. *Water Resources Research*, 49(10), 6744–6755.

Hemri, S., Lisniak, D., et Klein, B. (2015). Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, 51(9), 7436–7451.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570.

Hingray, B., Picouet, C., et Musy, A. (2009). *Hydrologie. 2. Une science pour l'ingénieur*. Lausanne, Suisse : Presses Polytechniques Universitaires Romandes.

Hou, D., Toth, Z., Zhu, Y., Yang, W., et Wobus, R. (2010). A stochastic total tendency perturbation scheme representing model-related uncertainties in the NCEP global ensemble forecast system. *Submitted to Tellus*.

Hu, Y., Schmeits, M. J., Jan van Andel, S., Verkade, J. S., Xu, M., Solomatine, D. P., et Liang, Z. (2016). A stratified sampling approach for improved sampling from a calibrated ensemble forecast distribution. *Journal of Hydrometeorology*, 17(9), 2405–2417.

Hyndman, R. J., et Athanasopoulos, J. (2012). *Forecasting*: *principles and practice*. Melbourne, Australie : Otexts.

Johnson, C., et Swinbank, R. (2009). Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, 135(640), 777–794.

Jolliffe, I. T., et Primo, C. (2008). Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136(6), 2133–2139.

Jolliffe, I. T., et Stephenson, D. B. (2003). *Forecast verification*: *A practitioner's guide in atmospheric science*. Chichester, UK : Wiley.

Kang, T.-H., Kim, Y.-O., et Hong, I.-P. (2010). Comparison of pre- and post-processors for ensemble streamflow prediction. *Atmospheric Science Letters*, 11(2), 153–159.

Keune, J., Ohlwein, C., et Hense, A. (2014). Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts. *Monthly Weather Review*, 142(11), 4074–4090.

Krzysztofowicz, R. (1997). Transformation and normalization of variates with specified distributions. *Journal of Hydrology*, 197(1-4), 286–292.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249(1-4), 2–9.

Krzysztofowicz, R. (2002). Bayesian system for probabilistic river stage forecasting. *Journal of Hydrology*, 268(1-4), 16–40.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., et Gneiting, T. (2017). Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science*, 32(1), 106–127.

Leutbecher, M., et Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, 227(7), 3515–3539.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., et Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews*: *Water*, 4(6).

Liu, Y., Weerts, A., Clark, M. P., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., ... others (2012). Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences*, 16(10), 3863–3887.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2), 130–141.

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences*, 26(4), 636–646.

Madadgar, S., et Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research*, 50(12), 9586–9603.

Marty, R., Zin, I., Obled, C., Bontron, G., et Djerboua, A. (2012). Toward real-time daily PQPF by an analog sorting approach: application to flash-flood catchments. *Journal of Applied Meteorology and Climatology*, 51(3), 505–520.

Matheson, J. E., et Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10), 1087–1096.

Mathevet, T. (2005). *Quels modèles pluie-débit globaux au pas de temps horaire ? Développements empiriques et comparaison de modèles sur un large échantillon de bassins versants* (Thèse de doctorat). Ecole Nationale du Génie Rural, des Eaux et Forêts & Cemagref.

Michel, C. (1989). *Hydrologie appliquée aux petits bassins versants ruraux* (Thèse de doctorat). Cemagref.

Michelangeli, P.-A., Vautard, R., et Legras, B. (1995). Weather regimes: recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, 52(8), 1237–1256.

Möller, A., Lenkoski, A., et Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139(673), 982–991.

Morton, F. I. (1983). Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology. *Journal of Hydrology*, 66(1-4), 1–76.

Mullen, S. L., et Buizza, R. (2002). The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Weather and Forecasting*, 17(2), 173–191.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595–600.

Murphy, A. H. (1995). A coherent method of stratification within a general framework for forecast verification. *Monthly Weather Review*, 123(5), 1582–1588.

Murphy, A. H., et Epstein, E. S. (1967). Verification of probabilistic predictions: a brief review. *Journal of Applied Meteorology*, 6(5), 748–755.

Murphy, A. H., et Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7), 1330–1338.

Murtagh, F., et Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?. *Journal of Classification*, 31(3), 274–295.

Nash, J. E., et Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I–A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.

Nelder, J. A., et Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4), 308–313.

Obled, C., Bontron, G., et Garçon, R. (2002). Quantitative precipitation forecasts: A statistical adaptation of model outputs through an analogues sorting approach. *Atmospheric Research*, 63(3), 303–324.

Obled, C., et Zin, I. (2004). TOPMODEL : principes de fonctionnement et application. *La Houille Blanche*(1), 65–77.

Oudin, L., Andreassian, V., Mathevet, T., Perrin, C., et Michel, C. (2006). Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. *Water Resources Research*, 42(7).

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., et Loumagne, C. (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. *Journal of Hydrology*, 303(1-4), 290–306.

Pagano, T. C., Wood, A. W., Ramos, M.-H., Cloke, H. L., Pappenberger, F., Clark, M. P., ... others (2014). Challenges of operational river forecasting. *Journal of Hydrometeorology*, 15(4), 1692–1707.

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., ... Weisheimer, A. (2009). Stochastic parametrization and model uncertainty. *ECMWF Technical Memorandum*, 598, 1–42.

Pappenberger, F., Bartholmes, J., Thielen, J., Cloke, H. L., Buizza, R., et de Roo, A. (2008). New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophysical Research Letters*, 35(10).

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., ... Salamon, P. (2015). How do i know if my forecasts are better? using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, 522, 697–713.

Pappenberger, F., Scipal, K., et Buizza, R. (2008b). Hydrological aspects of meteorological verification. *Atmospheric Science Letters*, 9(2), 43–52.

Park, Y.-Y., Buizza, R., et Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, 134(637), 2029–2050.

Perrin, C. (2000). *Vers une amélioration d'un modèle global pluie-débit au travers d'une approche comparative* (Thèse de doctorat). Institut National Polytechnique de Grenoble & Cemagref.

Perrin, C., Michel, C., et Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of hydrology*, 279(1-4), 275–289.

Perrin, C., Michel, C., et Andréassian, V. (2007). *Modèles hydrologiques du génie rural (GR)* (Rapport technique). Cemagref, UR Hydrosystèmes et Bioprocédés.

Pinson, P., et Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96, 12–20.

Pinson, P., et Tastu, J. (2013). *Discrimination ability of the energy score* (Rapport technique). Technical University of Denmark.

Qu, B., Zhang, X., Pappenberger, F., Zhang, T., et Fang, Y. (2017). Multi-model grand ensemble hydrologic forecasting in the fu river basin using Bayesian model averaging. *Water*, 9(2), 74.

R Core Team. (2014). R: A language and environment for statistical computing [Manuel de logiciel]. Vienna, Austria. Consulté sur `http://www.R-project.org/`

Raftery, A. E., Gneiting, T., Balabdaoui, F., et Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5), 1155–1174.

Ramos, M. H., Van Andel, S. J., et Pappenberger, F. (2013a). Do probabilistic forecasts lead to better decisions?. *Hydrology and Earth System Sciences*, 17(6), 2219–2232.

Ramos, M. H., Voisin, N., et Verkade, J. S. (2013b). *Hepex-sip topic*: *post-processing (2/3)* [Blog post]. `https://hepex.irstea.fr/hepex-sip-topic-post-processing-23/`.

Reggiani, P., Renner, M., Weerts, A., et Van Gelder, P. (2009). Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research*, 45(2).

Remesan, R., et Mathew, J. (2015). *Hydrological data driven modelling*. Berlin, Germany : Springer.

Riboust, P., Thirel, G., Le Moine, N., et Ribstein, P. (2018). Revisiting a simple degree-day model for integrating satellite data: implementation of SWE–SCA hystereses. *Journal of Hydrology and Hydromechanics*, accepted.

Roulin, E. (2007). Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrology and Earth System Sciences*, 11, 725–737.

Roulston, M., et Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus A*, 55(1), 16–30.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., ... others (2010). The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, 91(8), 1015–1058.

Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., ... Seo, D. (2007). Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrology and Earth System Sciences Discussions*, 4(2), 655–717.

Schefzik, R. (2016a). A similarity-based implementation of the Schaake shuffle. *Monthly Weather Review*, 144(5), 1909–1921.

Schefzik, R. (2016b). Combining parametric low-dimensional ensemble postprocessing with reordering methods. *Quarterly Journal of the Royal Meteorological Society*, 142(699), 2463–2477.

Schefzik, R., Thorarinsdottir, T. L., Gneiting, T., *et al.* (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical science*, 28(4), 616–640.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140(680), 1086–1096.

Scheuerer, M., et Hamill, T. M. (2015a). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4), 1321–1334.

Scheuerer, M., et Hamill, T. M. (2015b). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions. *Monthly Weather Review*, 143(11), 4578–4596.

Scheuerer, M., Hamill, T. M., Whitin, B., He, M., et Henkel, A. (2017). A method for preferential selection of dates in the Schaake shuffle approach to constructing spatio-temporal forecast fields of temperature and precipitation. *Water Resources Research*, 53(4), 3029–3046.

Seiller, G., Anctil, F., et Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, 16(4), 1171–1189.

Seiller, G., Anctil, F., et Roy, R. (2017). Design and experimentation of an empirical multistructure framework for accurate, sharp and reliable hydrological ensembles. *Journal of Hydrology*, 552, 313–340.

Serinaldi, F. (2008). Analysis of inter-gauge dependence by kendall's τk, upper tail dependence coefficient, and 2-copulas with application to rainfall fields. *Stochastic Environmental Research and Risk Assessment*, 22(6), 671–688.

Siegert, S., Bröcker, J., et Kantz, H. (2012). Rank histograms of stratified Monte Carlo ensembles. *Monthly Weather Review*, 140(5), 1558–1571.

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., et Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135(9), 3209–3220.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., ... others (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, 97(1), 49–67.

Tabios, G. Q., et Salas, J. D. (1985). A comparative analysis of techniques for spatial interpolation of precipitation. *Water Resources Bulletin*, 21(3), 365–380.

Taillardat, M., Mestre, O., Zamo, M., et Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6), 2375–2393.

Talagrand, O., Vautard, R., et Strauss, B. (1997). Evaluation of probabilistic prediction systems. In *Proceedings of the ECMWF Workshop on predictability, 20-22 October 1997, ECMWF, Reading, UK.*

Tangara, M. (2005). *Nouvelle méthode de prévision de crue utilisant un modèle pluie-débit global* (Thèse de doctorat). École pratique des hautes Études de Paris & Cemagref.

Teweles, S., et Wobus, H. (1954). Verification of prognostic charts. *Bulletin of the American Meteorological Society*, 35, 455–463.

Thévenot, N. (2004). *Provision quantitative des précipitations par une méthode d'analogie. Utilisation de la prévision d'ensemble du CEPMMT* (Mémoire de DEA). Institut National Polytechnique de Grenoble.

Thiboult, A., et Anctil, F. (2015). On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments. *Journal of Hydrology*, 529, 1147–1160.

Thiboult, A., Anctil, F., et Boucher, M.-A. (2016). Accounting for three sources of uncertainty in ensemble hydrological forecasting. *Hydrology and Earth System Sciences*, 20(5), 1809–1825.

Thielen, J., Bartholmes, J., Ramos, M. H., et Roo, A. d. (2009). The European flood alert system–Part 1: concept and development. *Hydrology and Earth System Sciences*, 13(2), 125–140.

Thorarinsdottir, T. L., Gneiting, T., et Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal of Uncertainty Quantification*, 1(1), 522–534.

Thorarinsdottir, T. L., Scheuerer, M., et Heinz, C. (2016). Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, 25(1), 105–122.

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, 6(2), 123–137.

Trinh, B., Thielen-del Pozo, J., et Thirel, G. (2013). The reduction continuous rank probability score for evaluating discharge forecasts from hydrological ensemble prediction systems. *Atmospheric Science Letters*, 14(2), 61–65.

Valéry, A., Andréassian, V., et Perrin, C. (2014). 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2–Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments. *Journal of Hydrology*, 517, 1176–1187.

Valéry, A. (2010). *Modélisation précipitations - débit sous influence nivale : Elaboration d'un module neige et évaluation sur 380 bassins versants* (Thèse de doctorat). AgroParisTech & Cemagref.

Van Schaeybroeck, B., et Vannitsem, S. (2014). Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141(688), 807–818.

Velázquez, J., Anctil, F., Ramos, M., et Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 french catchments using 16 hydrological model structures. *Advances in Geosciences*, 29, 33–42.

Verkade, J. S., Brown, J. D., Reggiani, P., et Weerts, A. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, 501, 73–91.

Verkade, J. S., et Werner, M. G. F. (2011). Estimating the benefits of single value and probability forecasting for flood warning. *Hydrology and Earth System Sciences*, 15(12), 3751–3765.

Voisin, N., Schaake, J. C., et Lettenmaier, D. P. (2010). Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Weather and Forecasting*, 25(6), 1603–1627.

Vrac, M., et Friederichs, P. (2015). Multivariate–intervariable, spatial, and temporal—bias correction. *Journal of Climate*, 28(1), 218–237.

Vrac, M., et Yiou, P. (2010). Weather regimes designed for local precipitation modeling: application to the mediterranean basin. *Journal of Geophysical Research*, 115(D12).

Vrugt, J. A., Gupta, H. V., Bouten, W., et Sorooshian, S. (2003). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39(8).

Wang, X., et Bishop, C. H. (2005). Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, 131(607), 965–986.

Wei, M., Toth, Z., Wobus, R., et Zhu, Y. (2008). Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A*, 60(1), 62–79.

Wilks, D. S. (2004). The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, 132(6), 1329–1340.

Wilks, D. S. (2015). Multivariate ensemble model output statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, 141(688), 945–952.

Wilks, D. S. (2017). On assessing calibration of multivariate ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143(702), 164–172.

Wilks, D. S. (2018a). Univariate ensemble postprocessing. In S. Vannitsem, D. S. Wilks, et J. Messner (Eds.), *Statistical postprocessing of ensemble forecasts* (pp. 49–89). Amsterdam, Netherlands : Elsevier.

Wilks, D. S. (2018b). Enforcing calibration in ensemble postprocessing. *Quarterly Journal of the Royal Meteorological Society*, 144(710), 76–84.

Wu, L., Seo, D.-J., Demargne, J., Brown, J. D., Cong, S., et Schaake, J. (2011). Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *Journal of Hydrology*, 399(3-4), 281–298.

Wu, L., Zhang, Y., Adams, T., Lee, H., Liu, Y., et Schaake, J. (2018). Comparative evaluation of three Schaake shuffle schemes in postprocessing GEFS precipitation ensemble forecasts. *Journal of Hydrometeorology*, 19(3), 575–598.

Yates, J. F. (1982). External correspondence: decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30(1), 132–156.

Yoo, C., et Ha, E. (2007). Effect of zero measurements on the spatial correlation structure of rainfall. *Stochastic Environmental Research and Risk Assessment*, 21(3), 287–297.

Zalachori, I. (2013). *Prévisions hydrologiques d'ensemble : développements pour améliorer la qualité des prévisions et estimer leur utilité* (Thèse de doctorat). AgroParisTech & Irstea.

Zalachori, I., Ramos, M. H., Garçon, R., Mathevet, T., et Gailhard, J. (2012). Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Advances in Science and Research*, 8(1), 135–141.

Zamo, M., et Naveau, P. (2018). Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, 50(2), 209–234.

Zappa, M., Jaun, S., Germann, U., Walser, A., et Fundel, F. (2011). Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research*, 100(2-3), 246–262.

Annexe A

Figures supplémentaires du chapitre 6

Cette annexe contient les figures additionnelles (*supporting information Figures*) de l'article inclus dans le chapitre 6, et publié dans la revue *Water Resources Research* :

Using Meteorological Analogues for Reordering Postprocessed Precipitation Ensembles in Hydrological Forecasting

J. Bellier[1], G. Bontron[2], and I. Zin[1]

[1] Université Grenoble Alpes, Grenoble INP, CNRS, IGE, Grenoble, France
[2] Compagnie Nationale du Rhône, Lyon, France

(Manuscript received 1 Jun. 2017, accepted 31 Oct 2017, published online 1 Dec. 2017)

DOI: 10.1002/2017WR021245

Contents

Introduction

Contents of Figures A.1 to A.3 aim at showing the effect of univariate postprocessing on mean areal precipitation (MAP) forecasts, regardless of multivariate dependence structures. Figure A.1 displays the CRPS improvement of postprocessed MAP forecasts over raw forecasts for the 5 study basins. Figure A.2 and A.3 show stratified rank histograms of MAP forecasts before and after postprocessing, respectively, for the 5 study basins and lead times of 24, 72 and 120 h. Each histogram is decomposed into different strata (whose sum forms the overall rank histogram) representing subsets of the verification sample. The subdivision is made here according to the forecast ensemble mean.

Figures A.4 to A.7 display examples of streamflow forecasts in addition to the one given in the paper.

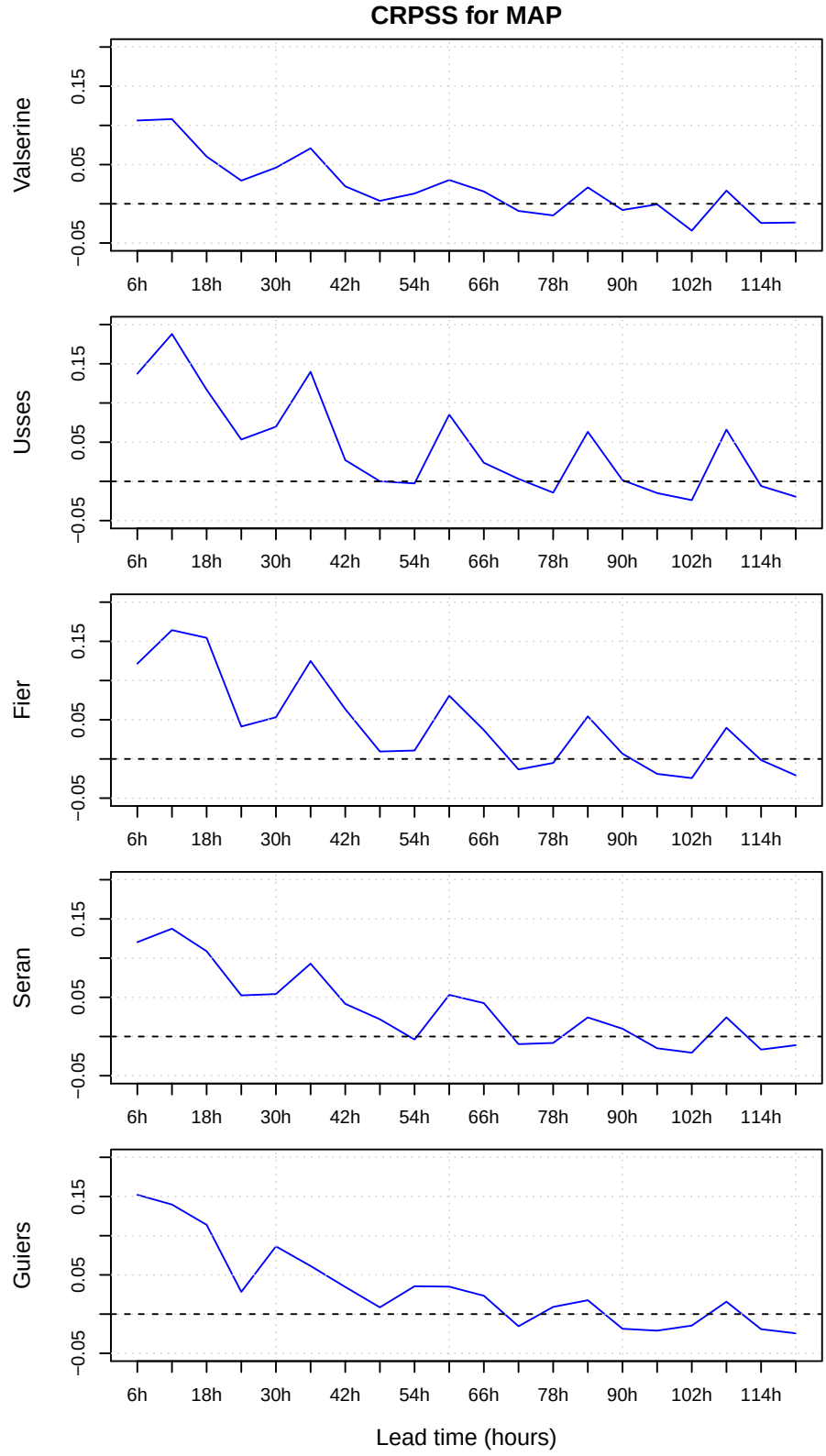

FIGURE A.1 – CRPSS of postprocessed MAP forecasts, as a function of lead time, for the 5 study basins. Raw MAP forecasts serve as reference.

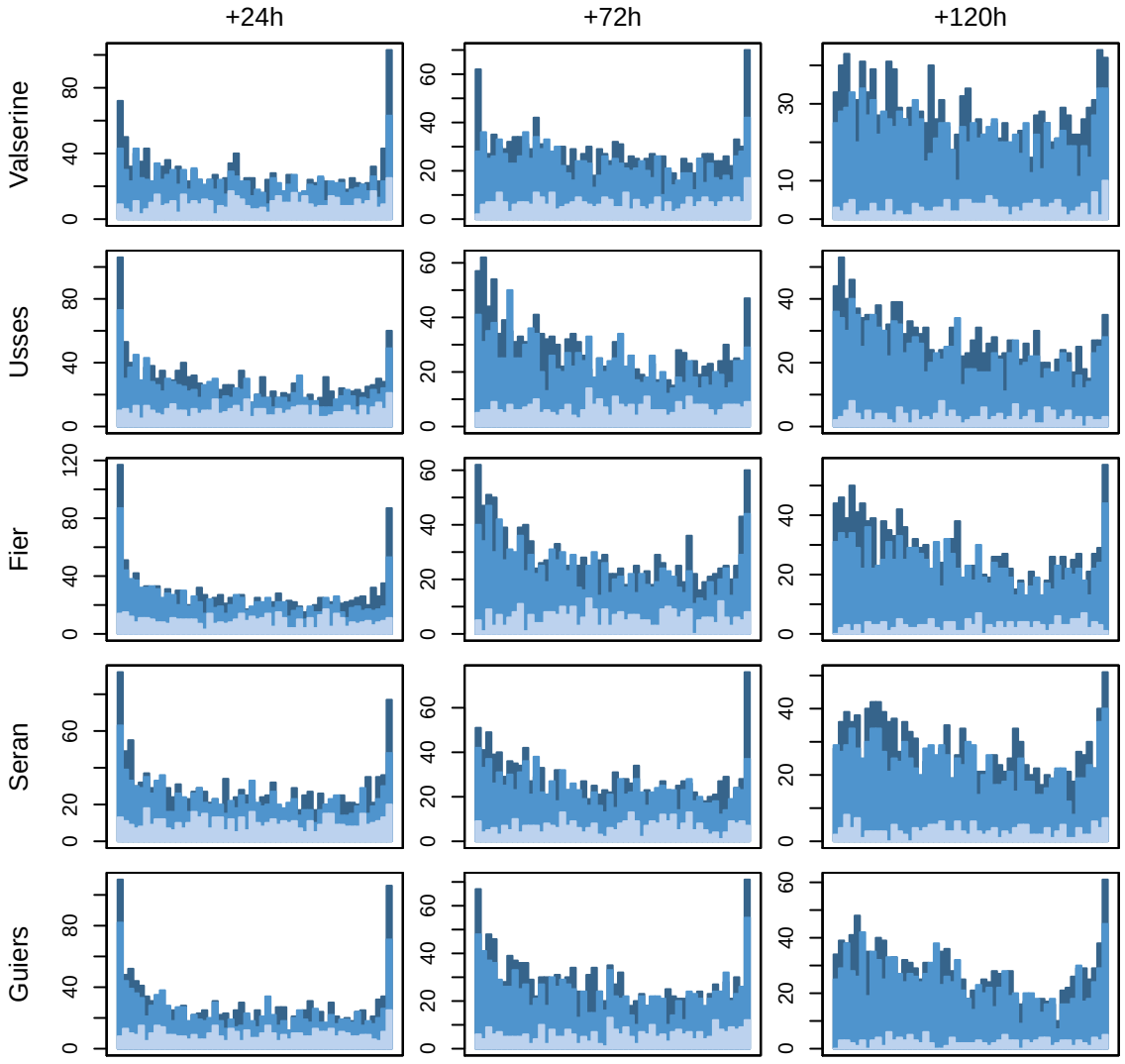

FIGURE A.2 – Accumulated stratified rank histograms of raw MAP forecasts for lead times 24, 72 and 120 h and the 5 study basins. The stratification criterion is the forecast ensemble mean denoted by $\bar{x}^{j,k}$. The light-blue stratum contains forecasts where $\bar{x}^{j,k} = 0$, the medium-blue stratum contains forecasts where $\bar{x}^{j,k} \in [0, 2[$ and the dark-blue stratum contains forecasts where $\bar{x}^{j,k} \geq 2$ (unit : mm $6h^{-1}$).

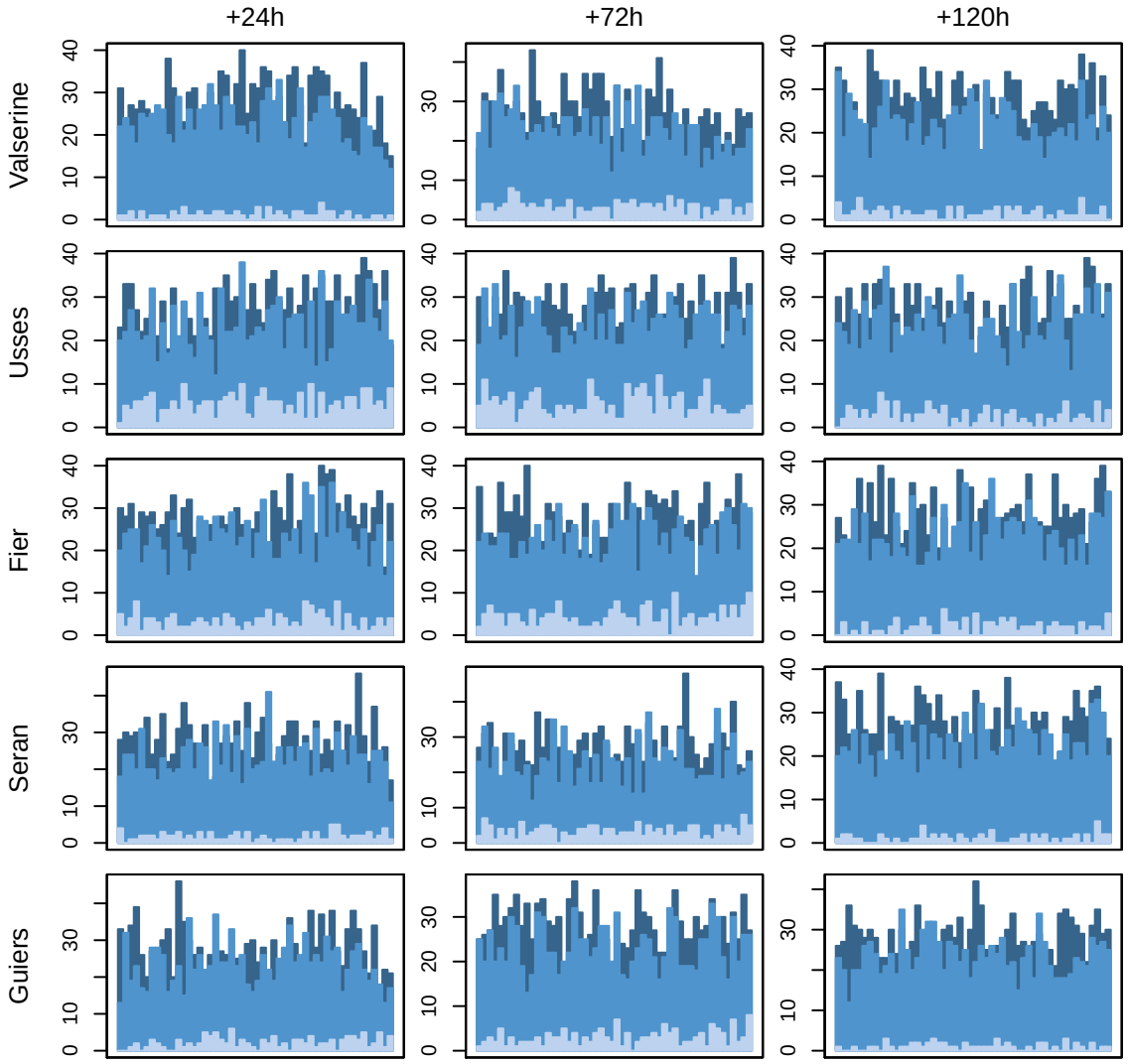

FIGURE A.3 – Same as Figure A.2, after postprocessing.

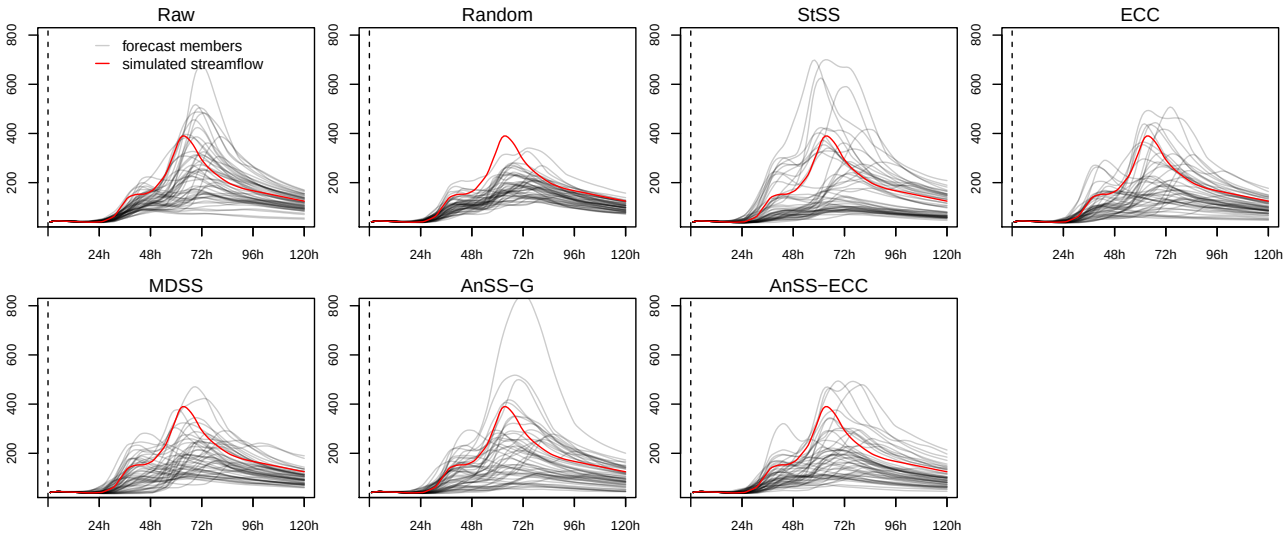

FIGURE A.4 – Examples of streamflow forecasts issued at Pont d'Evieu on 20 May 2012 at 00 UTC. The y-axis unit is m^3s^{-1}.

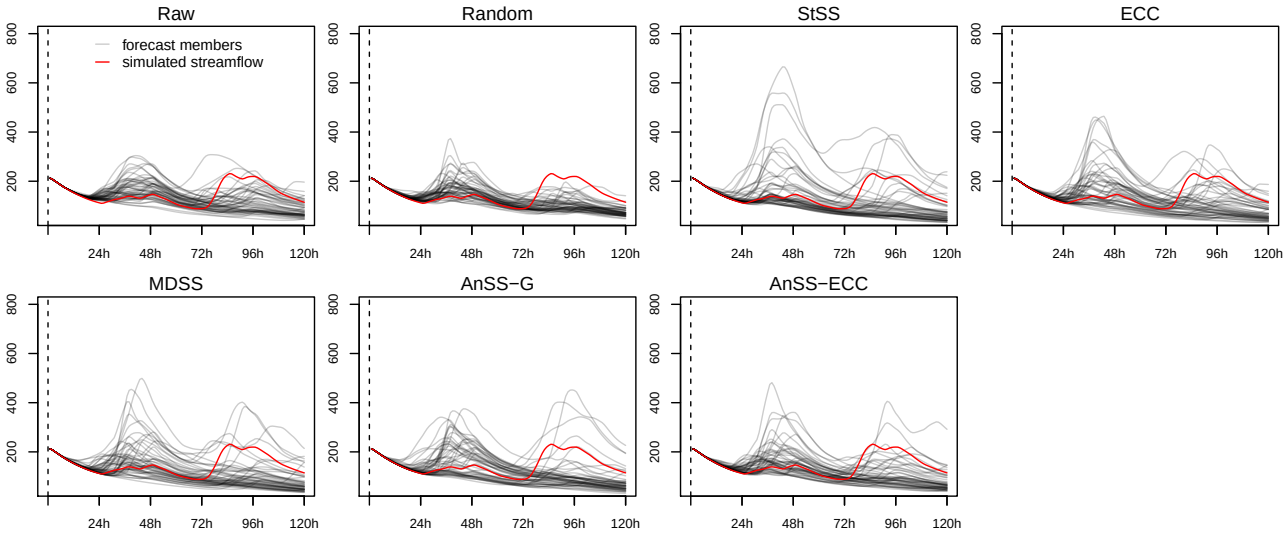

FIGURE A.5 – Same as Figure A.4 for a forecast issued on 26 July 2014 at 00 UTC.

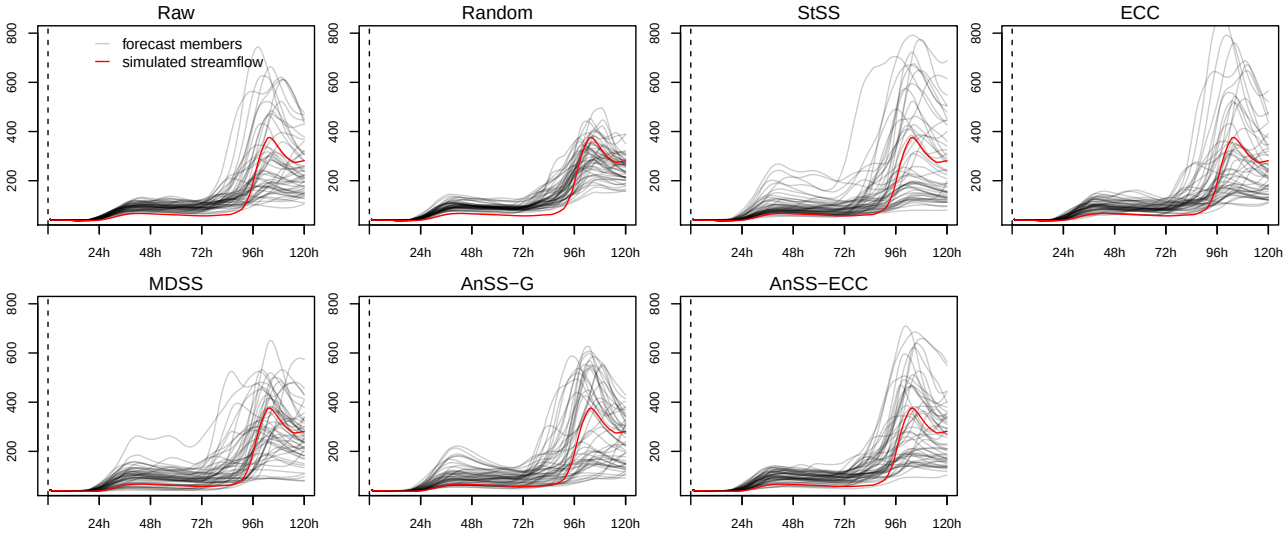

FIGURE A.6 – Same as Figure A.4 for a forecast issued on 01 November 2012 at 00 UTC.

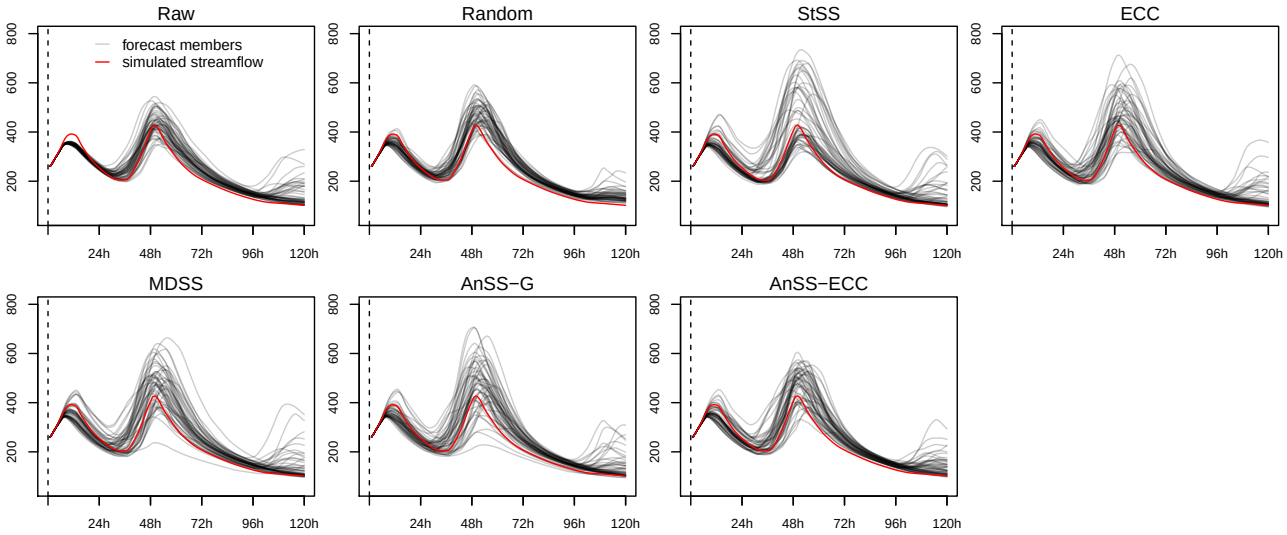

FIGURE A.7 – Same as Figure A.4 for a forecast issued on 26 December 2012 at 00 UTC.

Annexe B

Figures supplémentaires du chapitre 8

Cette annexe contient les figures additionnelles (*supporting information Figures*) de l'article inclus dans le chapitre 8, et publié dans la revue *Water Resources Research*, Vol. 54, no 8, p. 5741-5762. :

Generating coherent ensemble forecasts after hydrological postprocessing: Adaptations of ECC-based methods

J. Bellier[1], I. Zin[1], and G. Bontron[2]

[1] Université Grenoble Alpes, Grenoble INP, CNRS, IGE, Grenoble, France
[2] Compagnie Nationale du Rhône, Lyon, France

(Manuscript received 19 Jan. 2018, accepted 4 May 2018, published online 28 Aug. 2018)

DOI: 10.1029/2018WR022601

Contents

Introduction

Figure B.1 shows the performances in validation of each of the three hydrological models of the multi-model system used in the study. Figures B.2 to B.6 provide additional forecast examples (similarly to the Figure 8.3 in the article), but for randomly selected dates that cover different flow regimes. Figures B.7 to B.11 show the results of the evaluation of the autocorrelation criterion (similarly to the Figure 8.7 in the article), but for

the five other basins. Figures B.12 to B.16 show the CRPSS of the univariate forecasts (similarly to the Figure 8.8 in the article), but for the five other basins. Figures B.17 to B.19 show the rank histograms of the univariate forecasts for the Guiers basin (similarly to the Figure 8.9 in the article), but for other lead times : 6, 48 and 120 h.

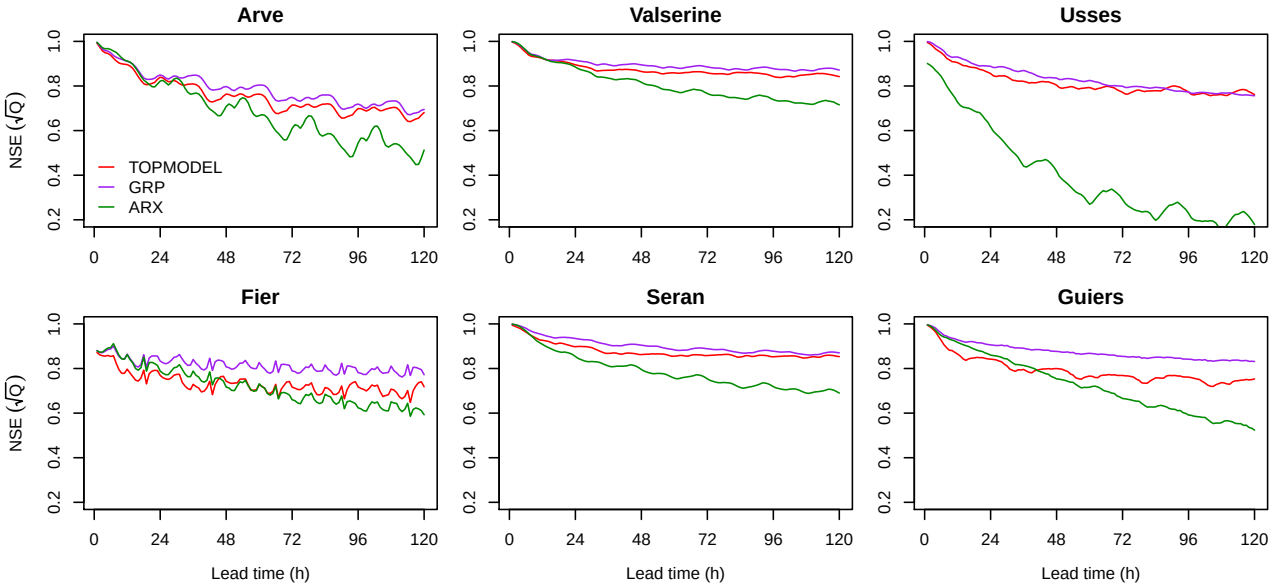

FIGURE B.1 – Performances of the three hydrological models under the perfect forecast mode (i.e., with observed precipitation and temperature forcings, and with the streamflow assimilation). The criterion is the Nash-Sutcliffe efficiency (NSE) coefficient, computed on the square roots of streamflow, over the 2007-2010 validation period.

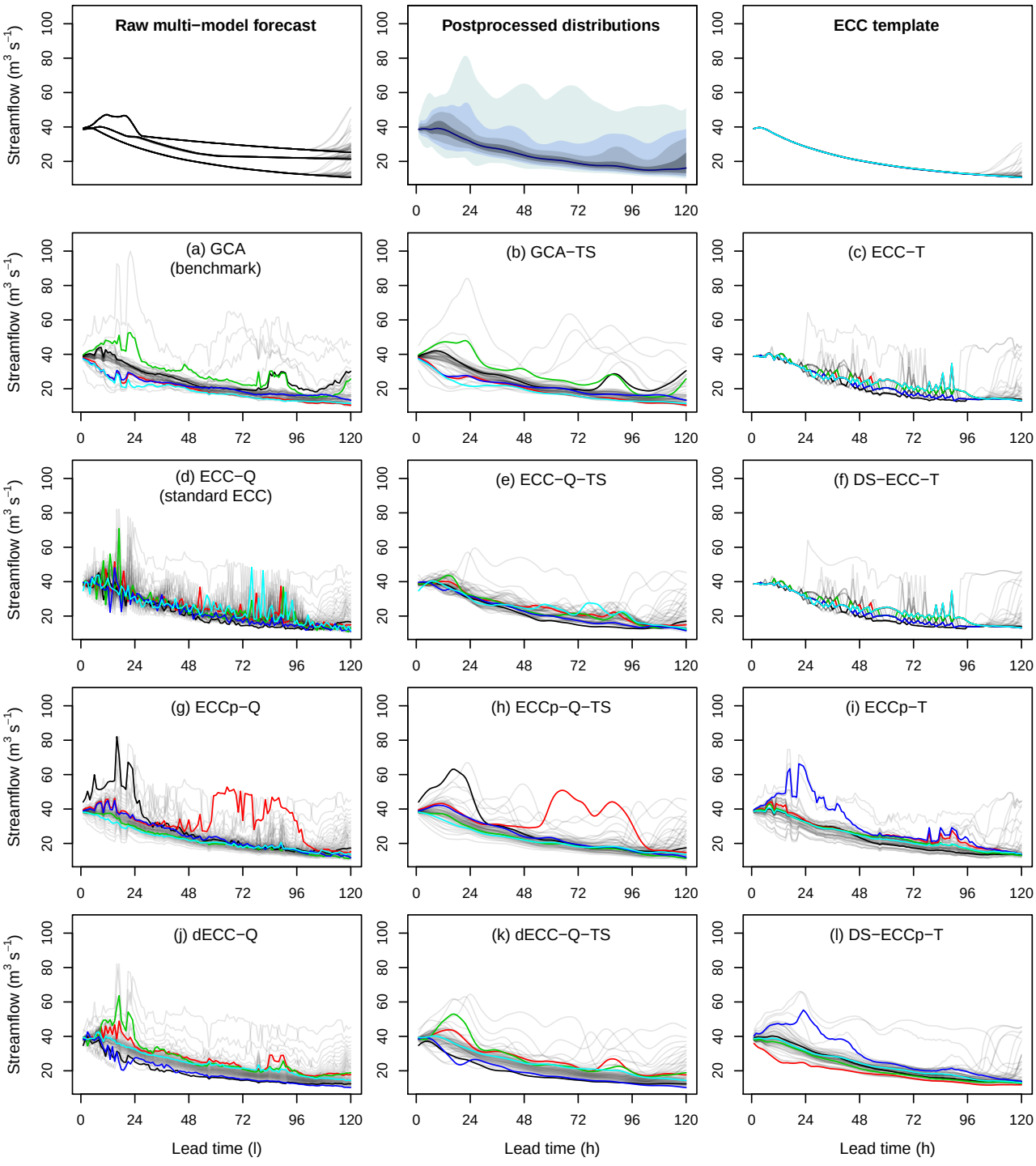

FIGURE B.2 – Additional forecast example for a randomly selected date (19 November 2014, Guiers basin).

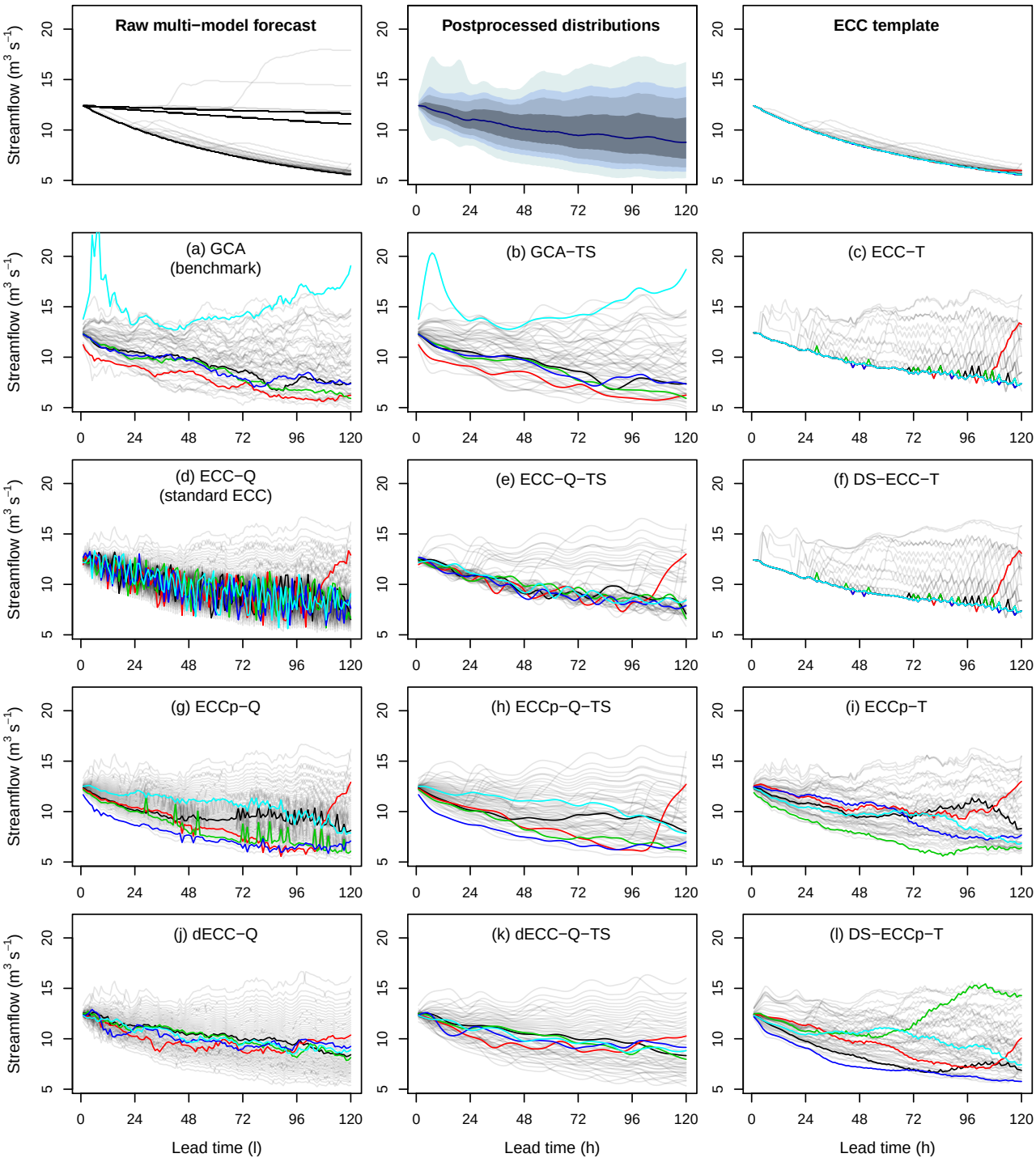

FIGURE B.3 – Additional forecast example for a randomly selected date (13 April 2011, Guiers basin).

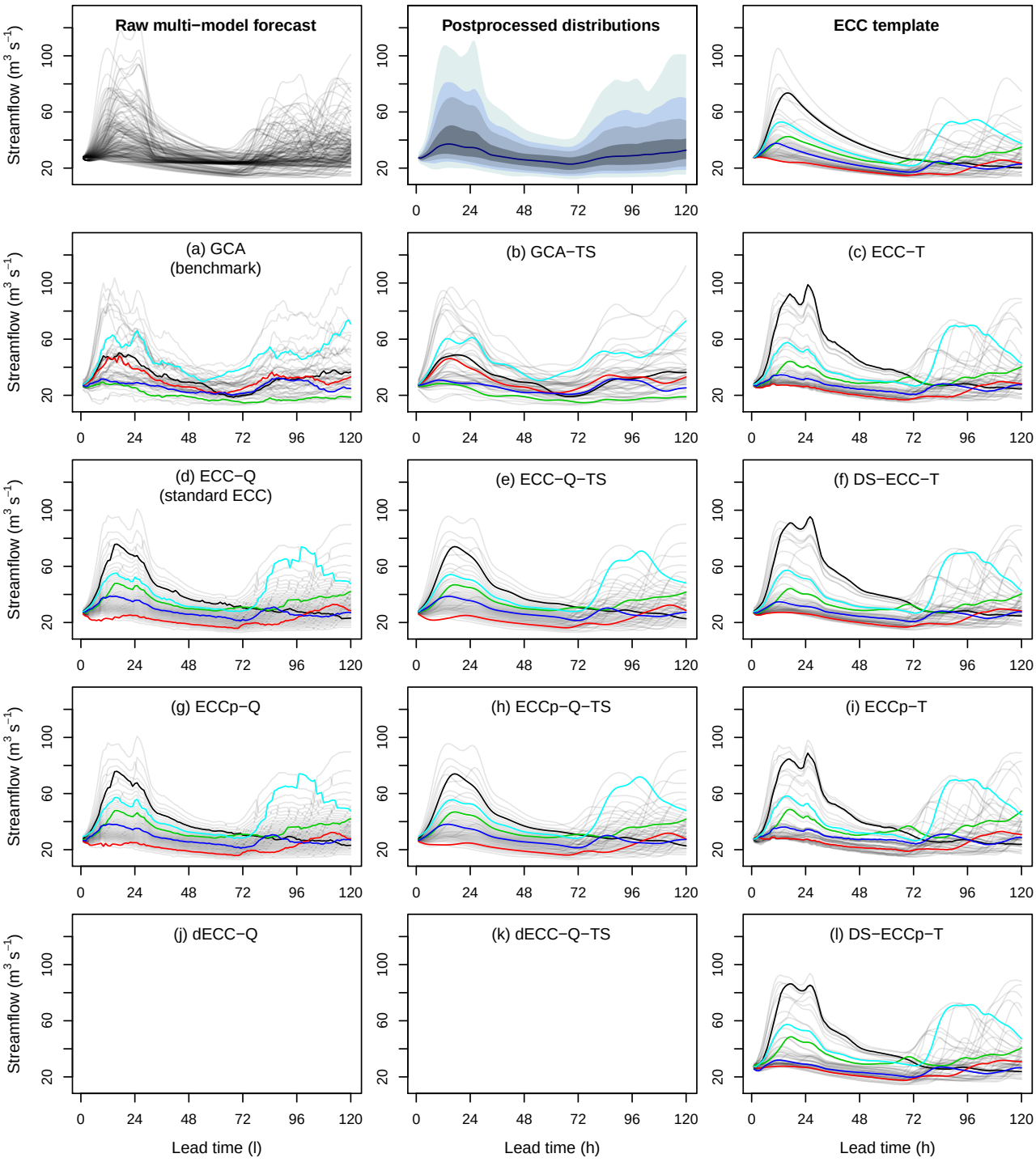

FIGURE B.4 – Additional forecast example for a randomly selected date (2 May 2012, Guiers basin).

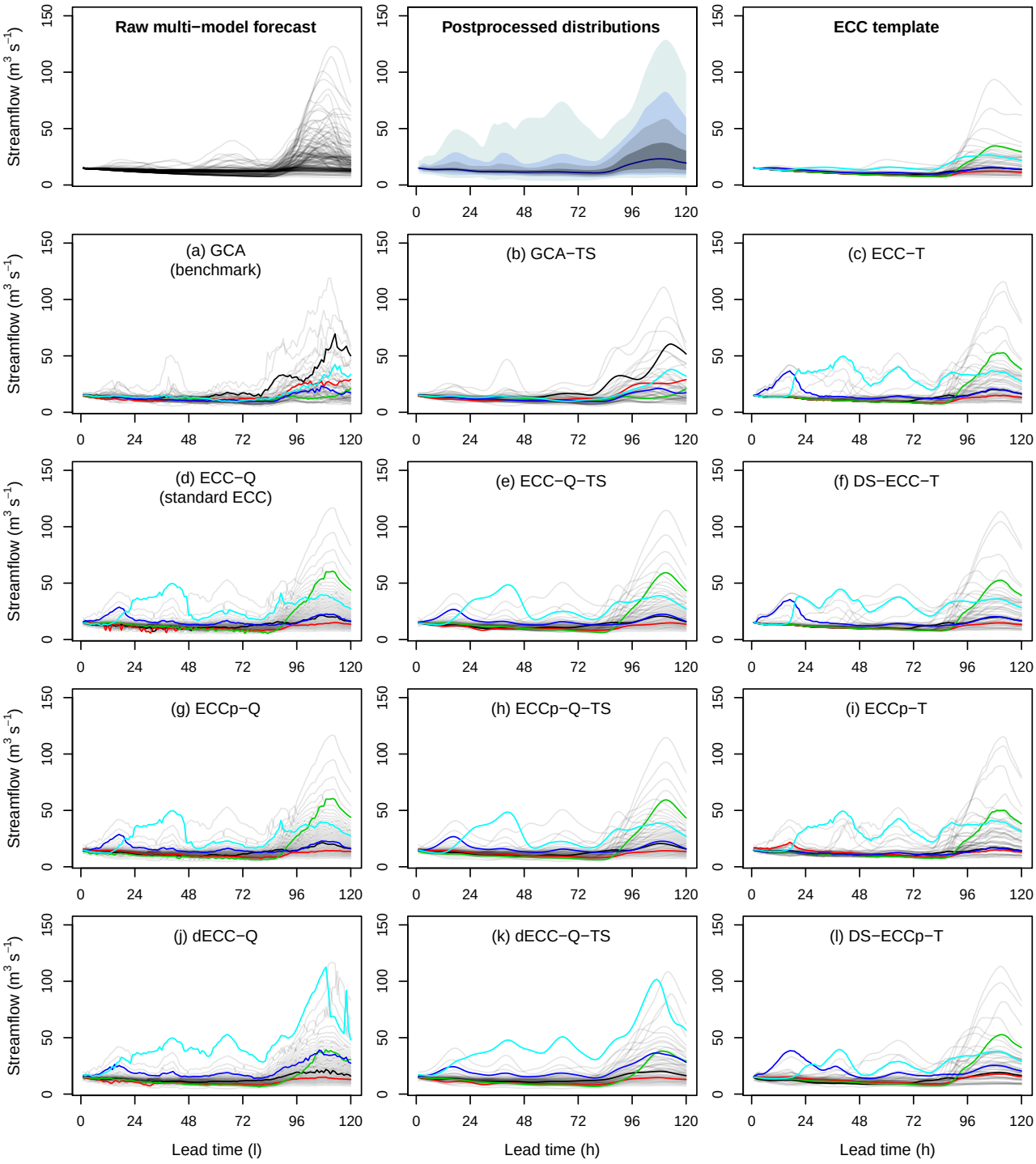

FIGURE B.5 – Additional forecast example for a randomly selected date (7 October 2013, Guiers basin).

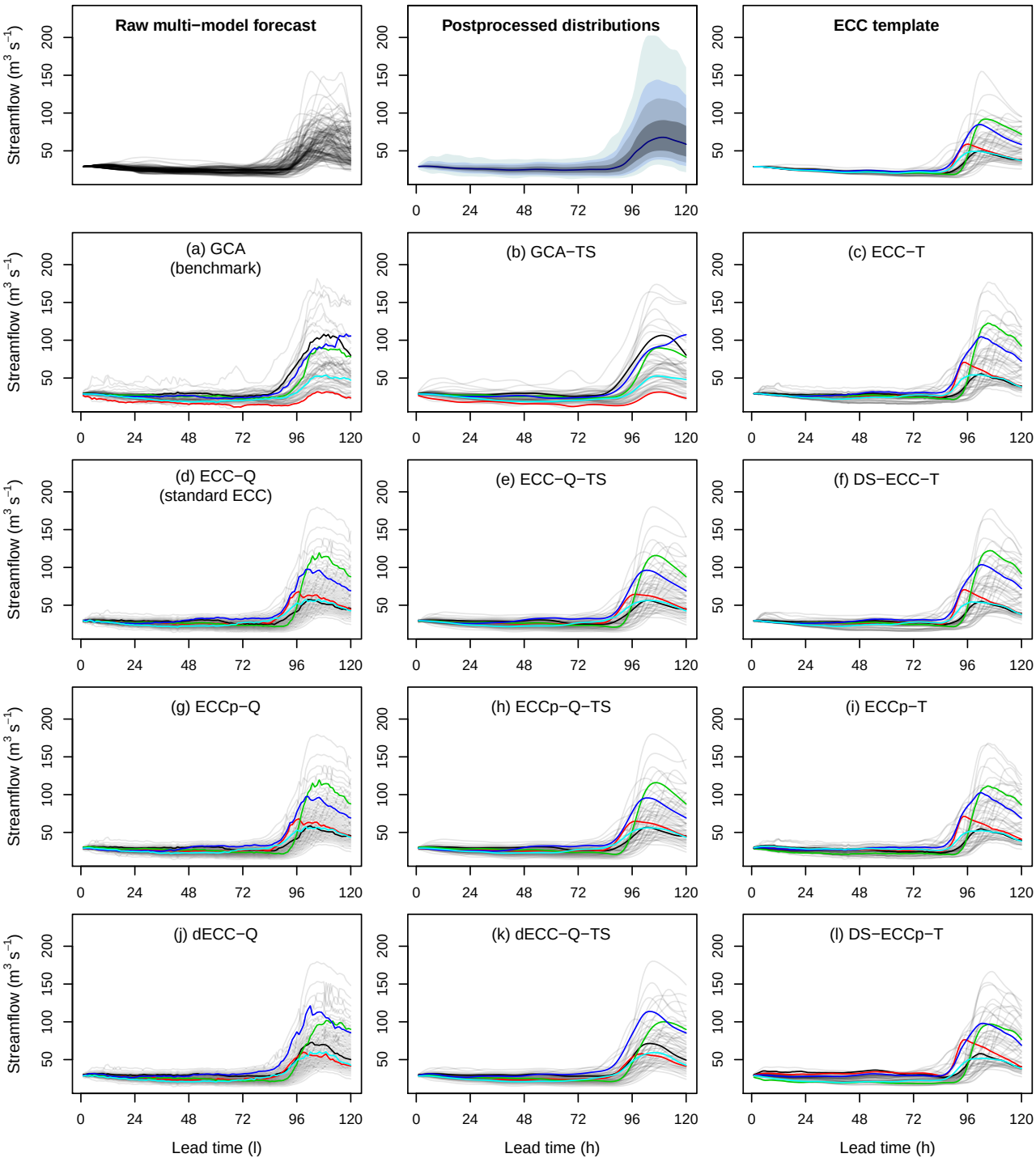

FIGURE B.6 – Additional forecast example for a randomly selected date (19 Mars 2014, Guiers basin).

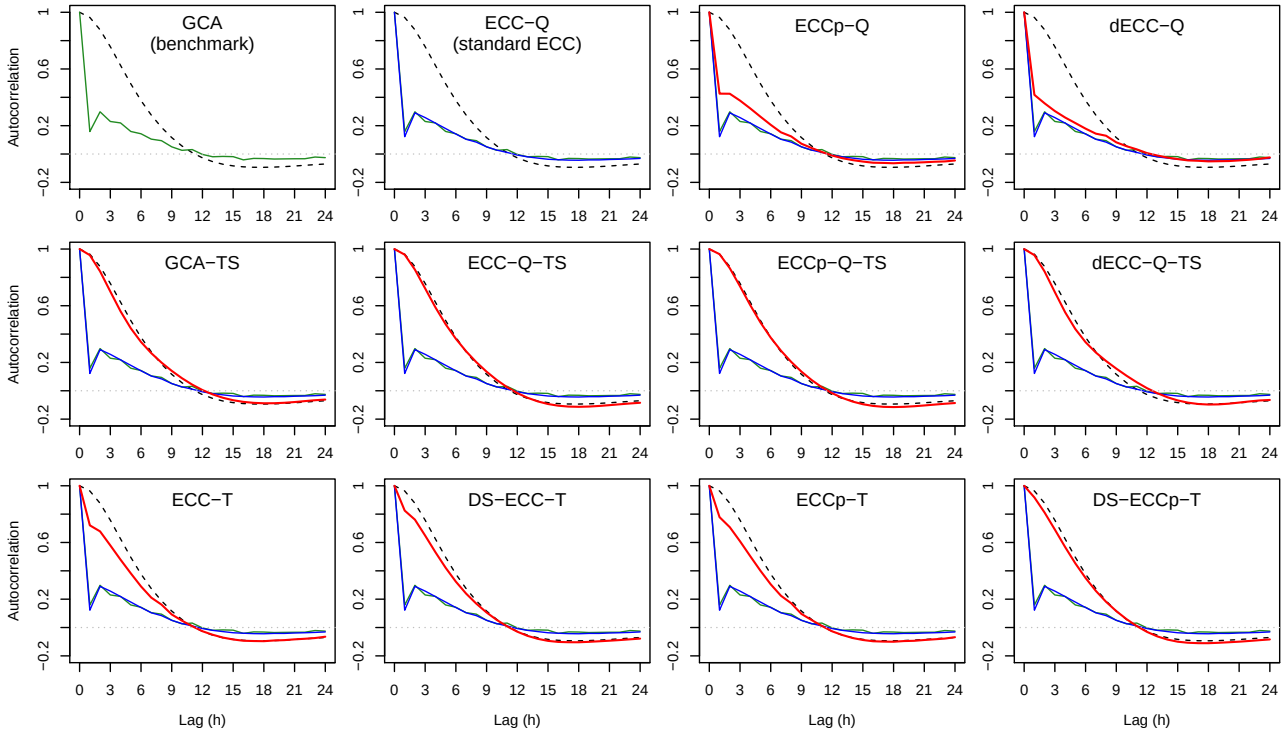

FIGURE B.7 – Results of the autocorrelation evalution for the Valserine basin.

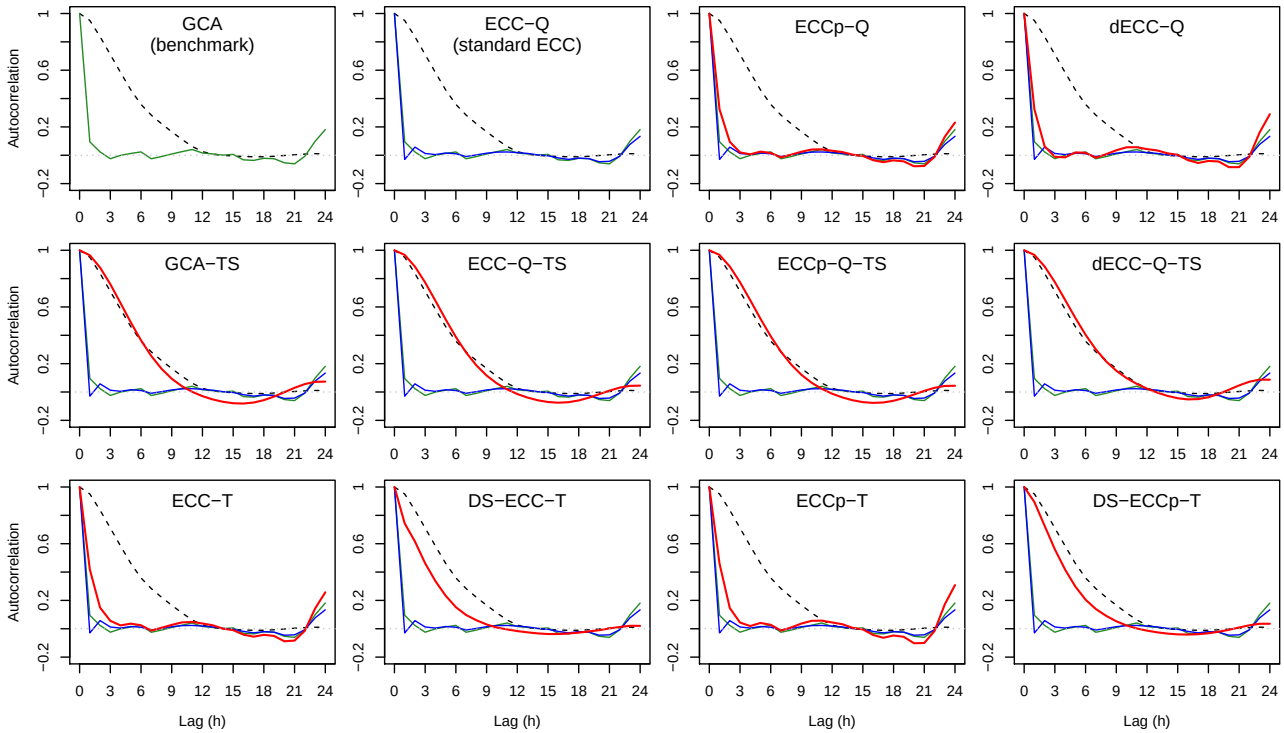

FIGURE B.8 – Results of the autocorrelation evalution for the Arve basin.

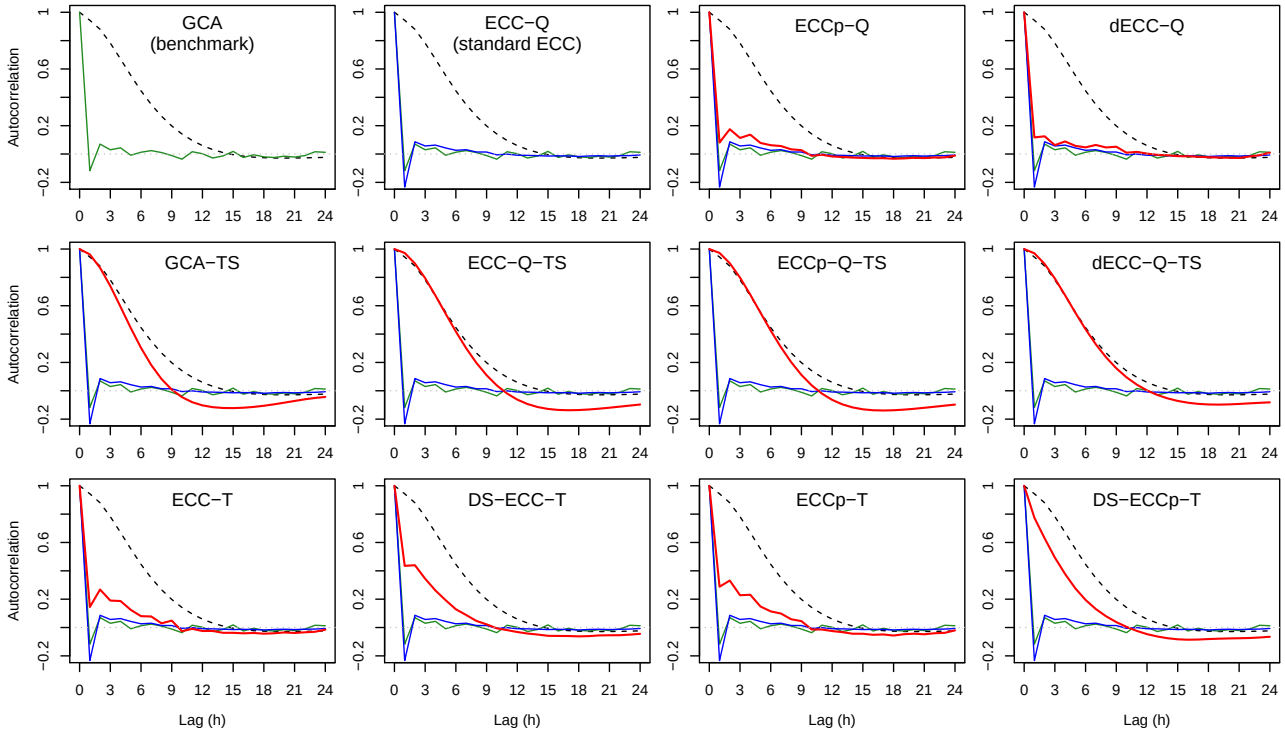

FIGURE B.9 – Results of the autocorrelation evalution for the Usses basin.

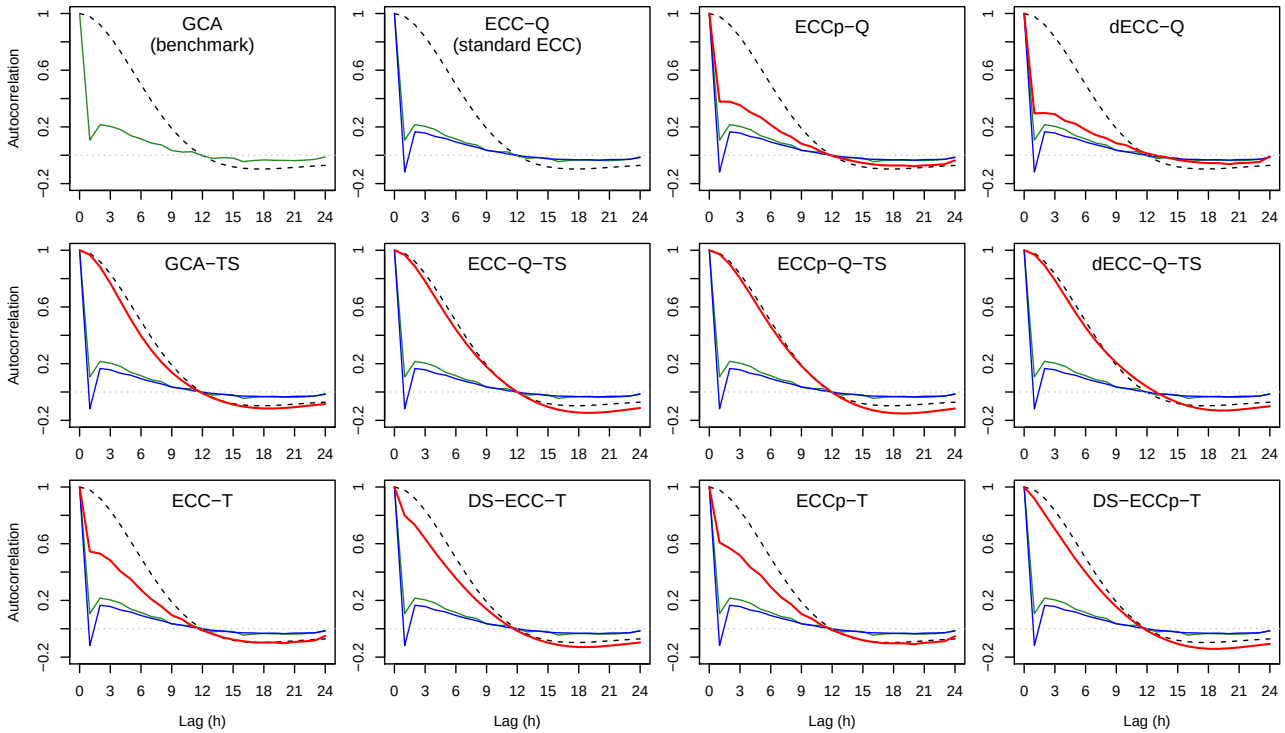

FIGURE B.10 – Results of the autocorrelation evalution for the Fier basin.

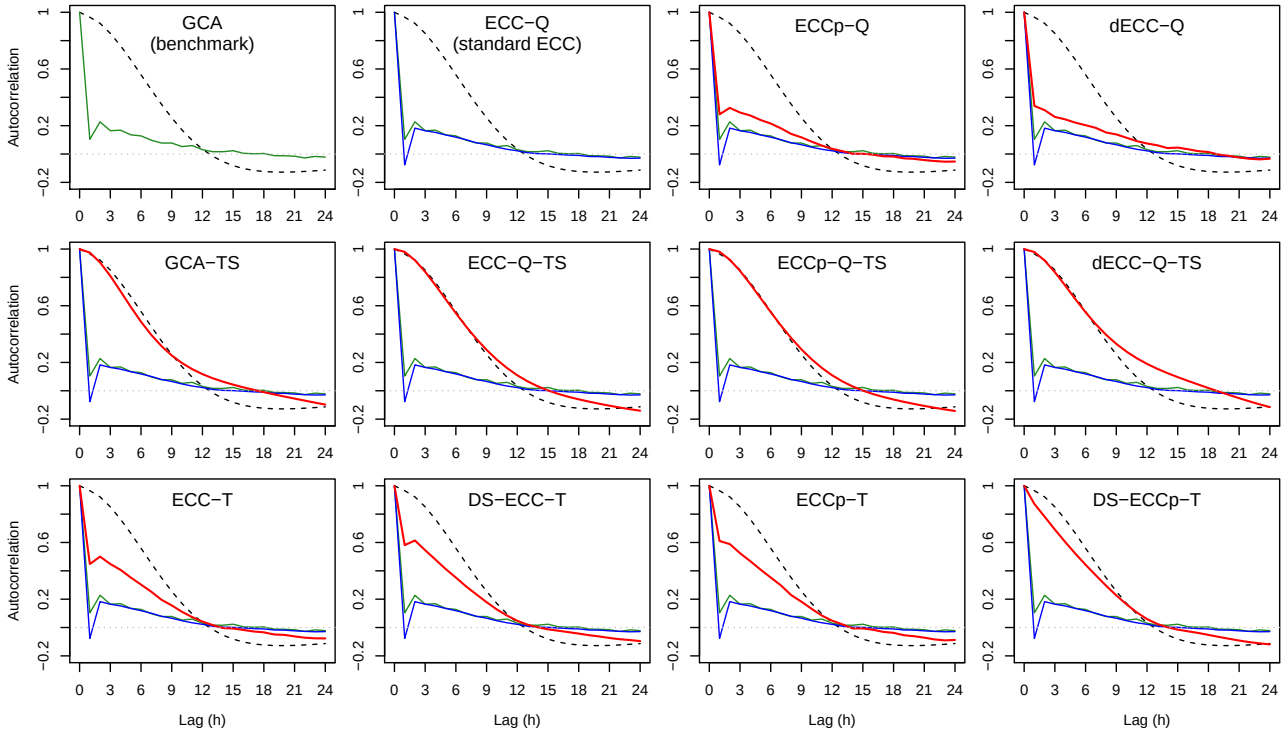

FIGURE B.11 – Results of the autocorrelation evalution for the Seran basin.

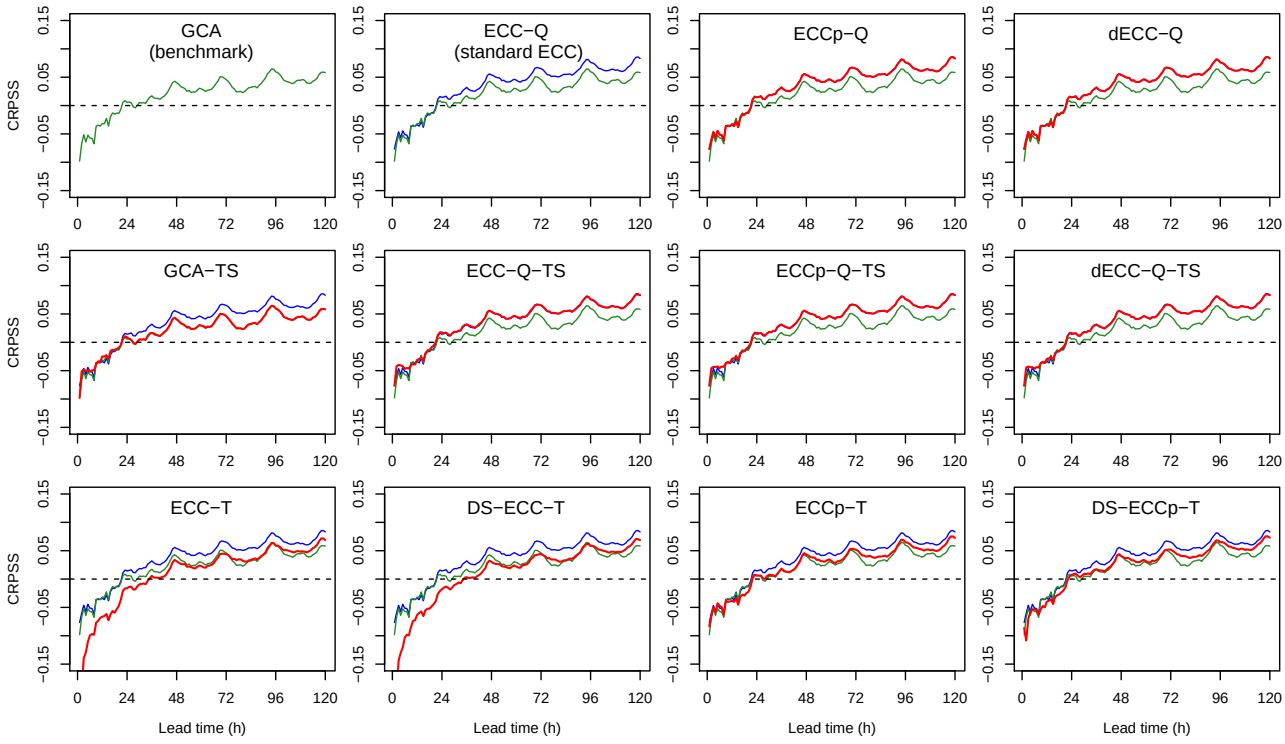

FIGURE B.12 – CRPSS of the univariate streamflow forecasts for the Valserine basin.

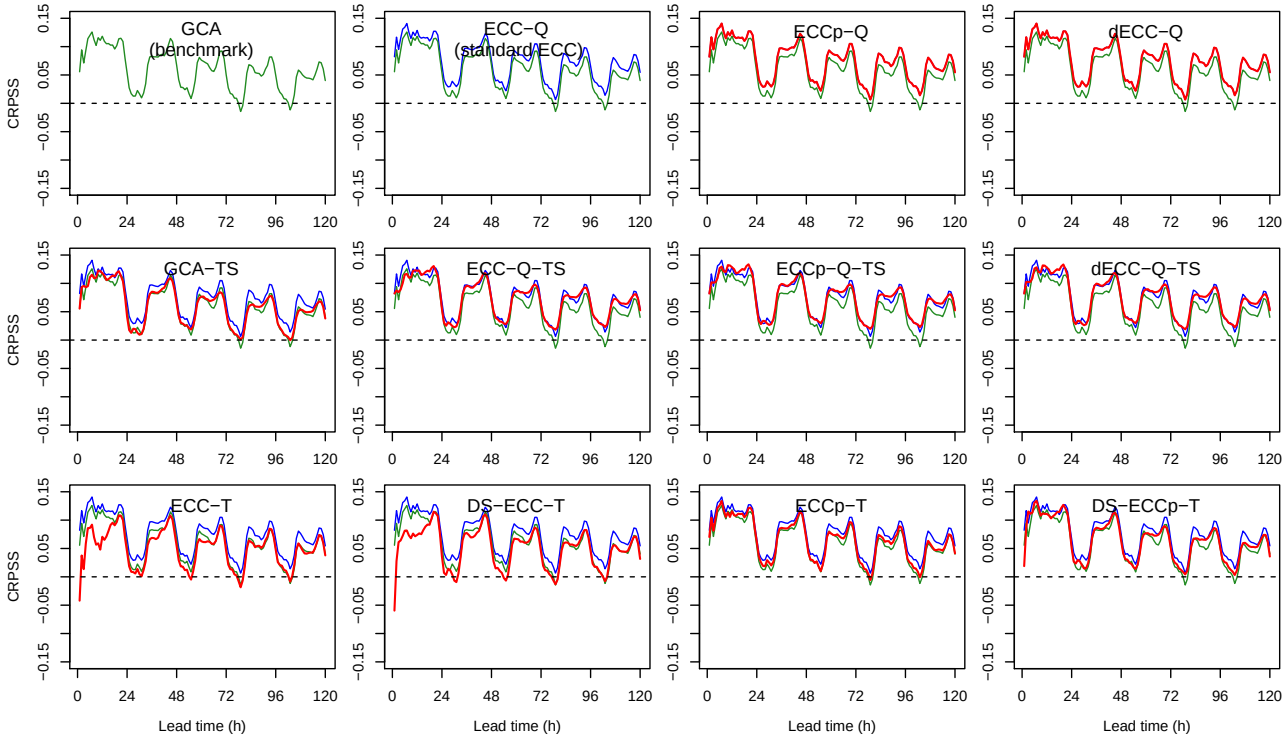

FIGURE B.13 – CRPSS of the univariate streamflow forecasts for the Arve basin.

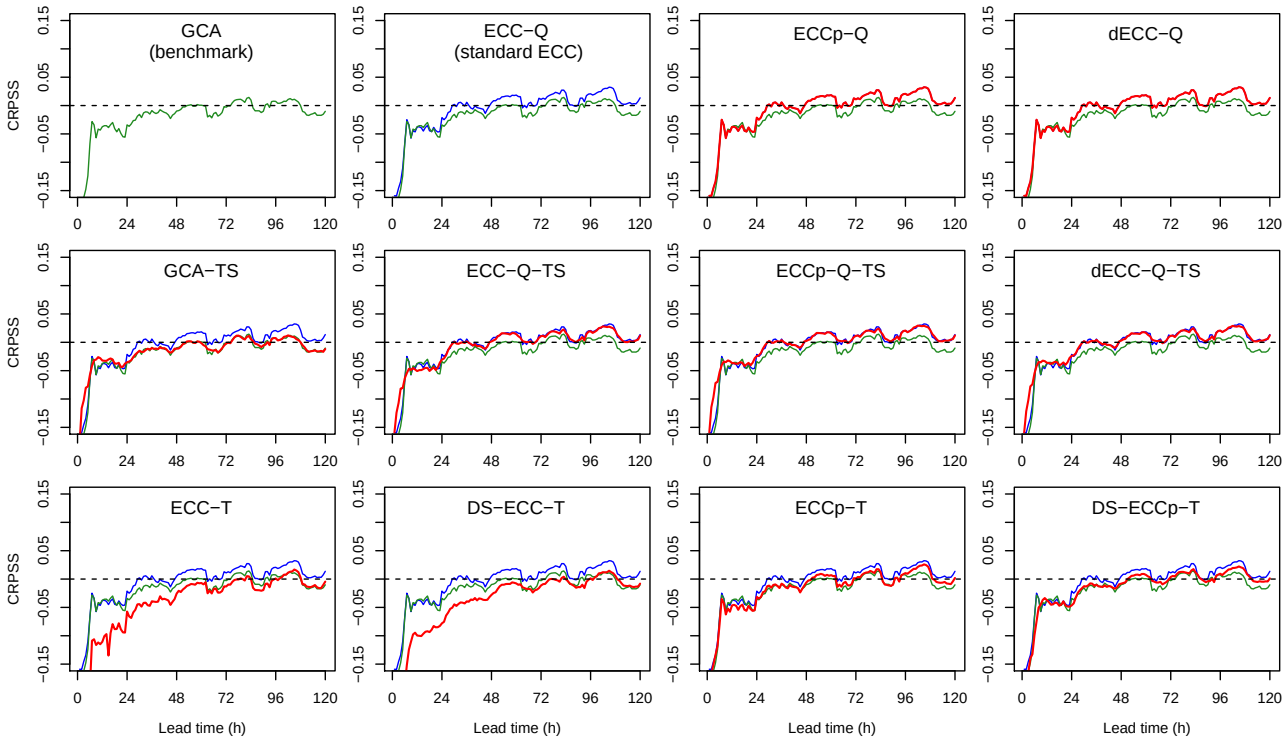

FIGURE B.14 – CRPSS of the univariate streamflow forecasts for the Usses basin.

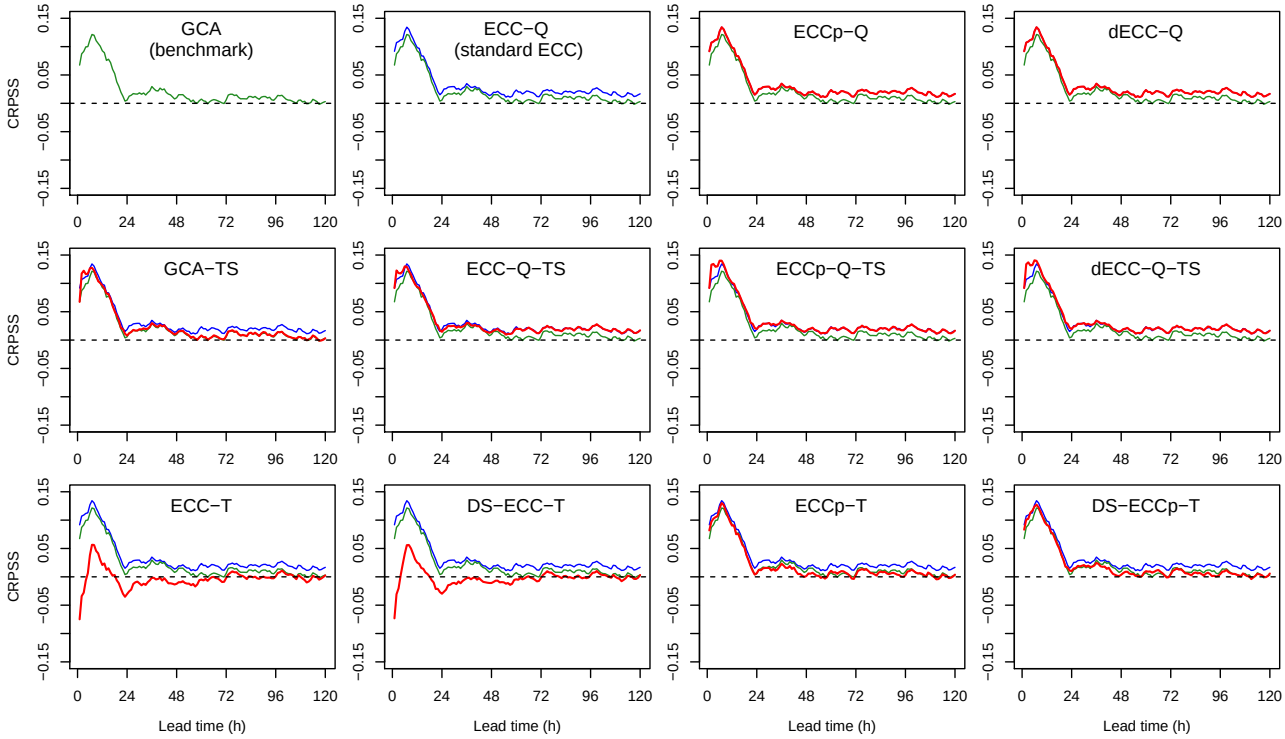

FIGURE B.15 – CRPSS of the univariate streamflow forecasts for the Fier basin.

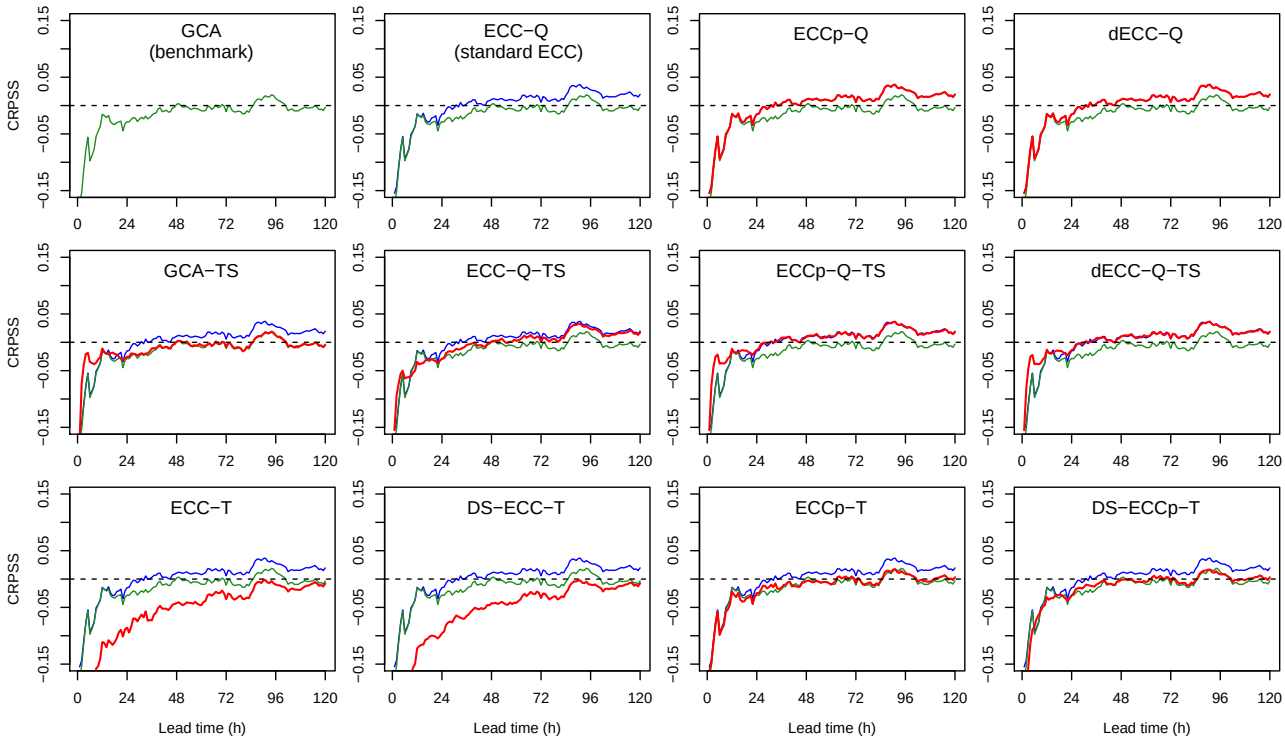

FIGURE B.16 – CRPSS of the univariate streamflow forecasts for the Seran basin.

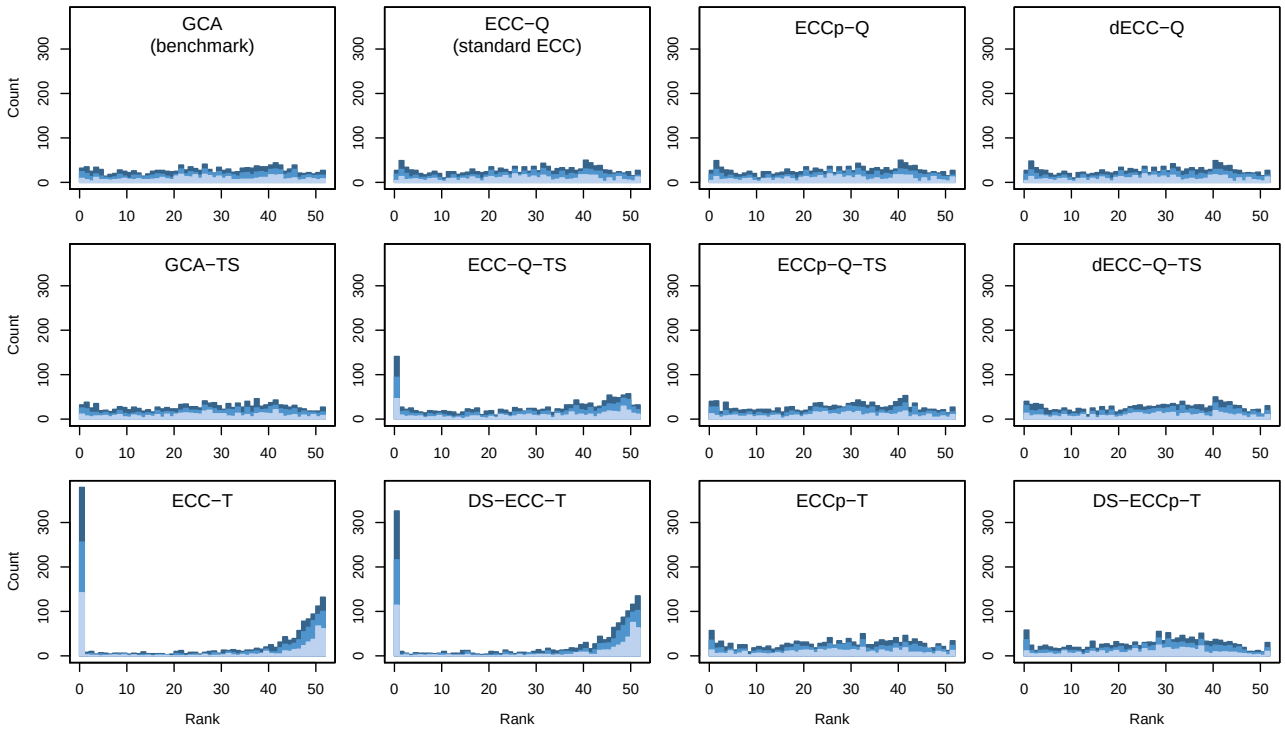

FIGURE B.17 – Rank histograms of the univariate streamflow forecasts for the Guiers basin and the 6 h lead time.

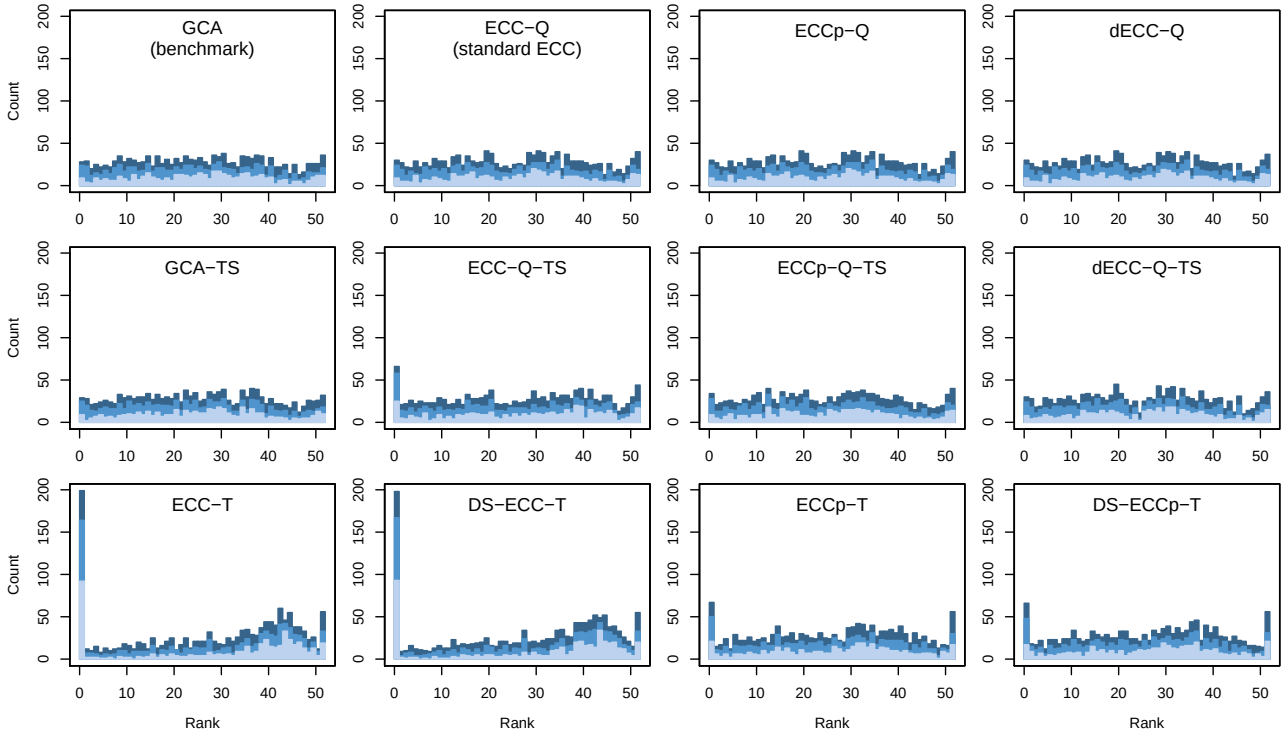

FIGURE B.18 – Rank histograms of the univariate streamflow forecasts for the Guiers basin and the 48 h lead time.

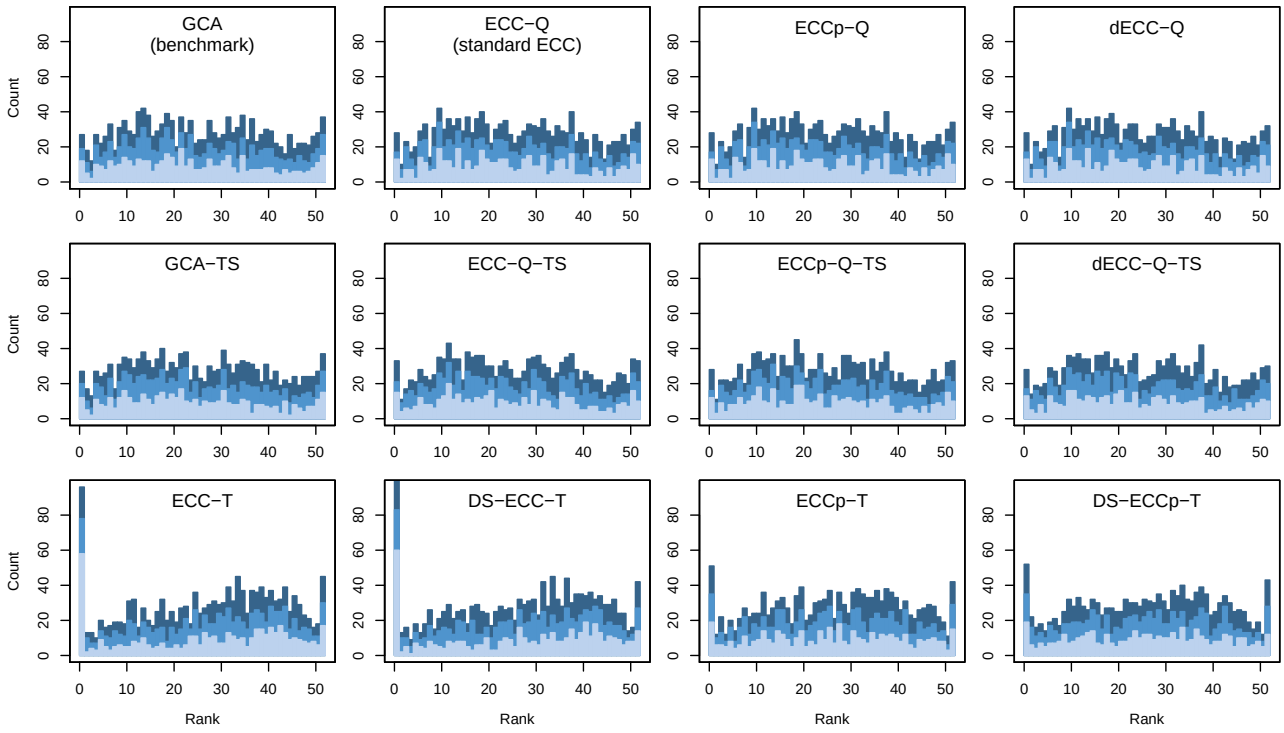

FIGURE B.19 – Rank histograms of the univariate streamflow forecasts for the Guiers basin and the 120 h lead time.