Spatially coherent probabilistic precipitation downscaling with meteorological analogues

Sabine Radanovics

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Océan, Atmosphère, Hydrologie**

Arrêté ministériel : 7 août 2006

Présentée par

Sabine Radanovics

Thèse dirigée par **Eric Sauquet**
et codirigée par **Jean-Philippe Vidal**

préparée au sein **d'Irstea, Centre de Lyon-Villeurbanne, France**
et de **l'École Doctorale Terre, Univers et Environnement**

Descente d'échelle probabiliste par analogues météorologiques. Étude de la cohérence spatiale.

Thèse soutenue publiquement le **3 Novembre 2014**,
devant le jury composé de :

Mme Anne-Catherine FAVRE
Professeur, INP Grenoble, France, Présidente
M. Bruce HEWITSON
Professeur, University of Cape Town, Rapporteur
M. Mathieu VRAC
Chargé de recherche CNRS, LSCE, Rapporteur
M. András BÁRDOSSY
Professeur, University of Stuttgart, Examinateur
M. Eric SAUQUET
Chargé de recherche Irstea, Irstea, Directeur de thèse
M. Jean-Philippe VIDAL
Chargé de recherche Irstea, Irstea, Co-Directeur de thèse
M. Guillaume BONTRON
Docteur, Ingenieur, Compagnie Nationale du Rhône, Invité

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Océan, Atmosphère, Hydrologie**

Arrêté ministériel : 7 août 2006

Présentée par

Sabine Radanovics

Thèse dirigée par **Eric Sauquet**
et codirigée par **Jean-Philippe Vidal**

préparée au sein **d'Irstea, Centre de Lyon-Villeurbanne, France**
et de **l'École Doctorale Terre, Univers et Environnement**

Spatially coherent probabilistic precipitation downscaling with meteorological analogues

Thèse soutenue publiquement le **3 Novembre 2014**,
devant le jury composé de :

Ms Anne-Catherine Favre
Professor, INP Grenoble, France, Présidente
Mr Bruce Hewitson
Professor, University of Cape Town, Rapporteur
Mr Mathieu Vrac
Chargé de recherche CNRS, LSCE, Rapporteur
Mr András Bárdossy
Professor, University of Stuttgart, Examinateur
Mr Eric Sauquet
Chargé de recherche Irstea, Irstea, Directeur de thèse
Mr Jean-Philippe Vidal
Chargé de recherche Irstea, Irstea, Co-Directeur de thèse
Mr Guillaume Bontron
Docteur, Ingenieur, Compagnie Nationale du Rhône, Invité

Remerciements

Au début de ce manuscript je voudrais exprimer ma gratitude pour tout les découvertes et rencontres que je pouvais faire pendant les dernières trois ans et remercier tout ceux et celles qui ont rendu cette très bonne expérience possible.

Tout d'abord je voudrais remercier Jean-Philippe Vidal, qui était toujours disponible pour tout les genres de questions scientifiques, techniques, linguistique, administratives et même personnelles, qui m'a appris pleine des choses, qui était toujours ouvert pour mes idées un peu fous, qui me faisait partir de son bureau presque toujours avec un sourire, qui a supporté mon esprit de la contradiction et bien géré mes émotions parfois un peu trop fortes même quand je faisais la crise sur les histoires de la cantine ou du CRPS. Merci pour les sessions des casse-têtes au tableau blanc, les astuces du ggplot, les blagues, les traductions, les explications, les discussions et la bonne énérgie. Merci également à Eric Sauquet qui comme directeur de thèse m'a laissé beaucoup de liberté dans mes démarches, apporté un regard complémentaire dans nos discussions et qui apprécie toujours une citation du hitchhikers guide bien placé.

Pour faire une thèse il faut un financement et du support scientifique et technique. Merci à la CNR pour le financement et leurs représentants Guillaume Bontron et Aurélien Ben Daoud pour le suivi. Merci aux membres externes du comité de thèse Julien Boé, Benoît Hingray et Maria-Helena Ramos pour la discussion enrichissante. Merci à Etienne Leblois pour les échanges philosophiques et la relecture d'une première version de ce manuscript. Merci aux rapporteurs Bruce Hewitson et Mathieu Vrac qui ont accepté d'évaluer ce manuscrit et aux autres membres du jury Anne-Catherine Favre, András Bárdossy et Guillaume Bontron d'avoir pris le temps de lire le manuscrit et d'en discuter pendant la soutenance. Merci à Anne Eicholz et Hélène Faurant-Philippe pour leur aide dans les démarches administratives. Merci à Antoine, Jean-Pierre et Fréderic du service informatique pour le bon fonctionnement des systèmes informatiques même pendant le déménagement du centre et Vincent Chevallereau pour le support compétent et rapide concernant le cluster de calcul. Un grand merci aussi à Judith Eeckman qui à fait un gros travail sur la modélisation hydrologique pendant son stage sans lequelle le chapitre 11 de cette thèse n'existais pas.

J'ai beaucoup apprecié la bonne ambiance humaine dans l'Unité de Recherche Hydrologie-Hydraulique et dans l'équipe Hydrologie qui permettait que tout le monde pouvait s'exprimer de sa façon pendant les réunions et dans les discussions entre collègues. Je voudrais remercier Isabelle Braud qui gère l'équipe sans beaucoup de bruit mais avec sagesse et efficacité et qui est devenue un modèle professionnel pour moi. Merci à Sonja Jankowfsky, Pierre-Henri Bazin, Victor Dupuis, Laurent Bonnifait, Clotaire Catalogne, Marko Adamovic et Mohammed Jaballah pour la bonne compagnie du bureau. Merci á François pour les crèpes, merci à Céline, Martin, Delphine, Laurie, Carla, Anne-Laure, Lucie et tout les autres collègues pour les bons moments au coin café et aux soirées.

De temps en temps il faut se défouler ou se détendre. Merci à tout les organisateurs et organisatrices d'escapade pour les sorties au théâtre et au ski et pour le yoga. Merci à Caroline Jean pour les précieux moments de la détente et les cours de confiance en soi à travers des postures improbables et à Claire Beraud qui m'a fait découvrir la biennale de la danse.

Un défi supplémentaire pendant les dernières trois ans était d'apprendre une troisième langue. Je voudrais donc remercier l'Unité de Recherche qui m'a paié des cours de français pendant deux ans et les professeurs de l'Alliance Française notamment Isabelle, Alice, Fréderic, Damian et Rémy, qui m'ont aidé à passer d'un niveau débutant en français à un niveau B2 en deux ans et demi. Merci aussi à Benjamin Renard dont les blagues m'ont motivés à améliorer mon français pour mieux les comprendre. C'est déjà beaucoup mieux, mais j'avoue que je ne comprends pas encore tout.

Partir à l'étranger c'est aussi une aventure personelle. Je voudrais donc remercier Georg Kotschy qui m'a encouragé de partir à l'étranger, Stanislas Siblot qui m'a herbergé pendant mon premier mois en France, Günther Vogelsam für die unterhaltsamen chats et Stephane, Marion et Mimi pour une expérience de la vie en collocation inoubliable. Un très grand merci aussi à Christian Cauchois pour le soutien pendant la dernière année, la consolation, les bisous et les repas, pour la motivation de faire du sport et pour avoir tenté l'aventure de démenager avec une doctorante deux mois avant le rendu du manuscrit (tu te ne rendes pas compte à quel point c'est courageux ;-)) et ma famille qui m'a soutenue de loin.

Contents overview

Contents

Résumé étendu

1 Contexte et état de l'art

Contexte

L'anticipation des ressources en eau futures est un enjeu de société important car elles concernent des besoins humains en termes d'eau potable, de production agricole et d'énergie. Anticiper la production hydroélectrique potentielle est notamment un enjeu économique pour les producteurs d'électricité comme la Compagnie Nationale du Rhône (CNR). Aux échéances courtes, l'enjeu porte sur la prévision de la production tandis qu'aux échéances longues c'est l'anticipation du rendement des investissements. Par ailleurs, étudier les débits passés sur de longues périodes permet d'améliorer notre connaissance des régimes et des extrêmes, ce qui est essentiel par exemple pour le dimensionnement d'ouvrages hydrauliques. Cependant, la longueur limitée des séries de débits mesurées implique parfois le besoin de reconstruire les événements passés anciens.

Les modèles de simulation du climat et du système terrestre permettent d'étudier ce type de problème, en fournissant par exemple des projections climatiques futures. Des réanalyses atmosphériques globales d'un siècle ou plus ont par ailleurs été récemment mises à disposition de la communauté et peuvent être utilisées pour des reconstructions historiques (Compo et al., 2011; Dee et al., 2014). Dans les deux cas, les processus météorologiques locaux ne sont pas bien représentés et rendent nécessaire une étape de régionalisation – ou descente d'échelle – à l'amont de toute étude d'impact.

Si la descente d'échelle dynamique utilise des modèles climatiques régionaux à résolution plus fine, la descente d'échelle statistique exploite les liens statistiques entre des variables de large échelle comme la circulation atmosphérique (les prédicteurs) et les variables locales comme les précipitations (les prédictants).

Il existe une grande diversité des méthodes de descente d'échelle statistique (Maraun et al., 2010; Fowler et al., 2007) : méthodes de régression linéaire ou non-linéaire, générateurs de temps, méthodes de correction de biais, ou encore méthodes basées sur des types de temps. La méthode utilisée ici est une méthode

par analogie, qui est un cas particulier des méthodes par types de temps. Les méthodes par analogie sont également utilisées en prévision opérationnelle, par example par la CNR pour la prévision des précipitations sur le bassin versant du Rhône (Bompart et al., 2009). Les méthodes par analogie parviennent en général à reproduire de manière satisfaisante la variance observée du prédictant peu importe la distribution de celui-ci. Une limitation intrinsèque est leur incapacité à produire des événements extrêmes qui n'ont pas été observés, quand elles reposent sur un simple ré-échantillonnage du passé.

L'utilisation de méthodes de descente d'échelle statistiques pour des applications hydrologiques entraîne d'importantes contraintes. Tout d'abord, et par opposition aux méthodes déterministes, les méthodes probabilistes peuvent apporter des informations sur leurs incertitudes dans une situation donnée. De plus, dans le contexte des simulations pluie-débit, la structure spatiale des précipitations est aussi importante que les quantités locales. En effet, notamment dans le contexte de la prévision opérationnelle, il est important de savoir dans quel affluent une augmentation de débit est générée. Les caractéristiques spatiales des sorties de méthodes de descente d'échelle statistiques sont cependant rarement étudiées. Un défi supplémentaire est la cohérence inter-variables indispensable pour toute modélisation hydrologique. Parmi les méthodes capable de répondre à ces exigences, les méthodes statistiques basées sur du rééchantillonnage – comme l'approche par analogie – côtoient les méthodes dynamiques bien plus coûteuses en termes de temps de calcul.

Objectifs

Les objectifs de cette thèse sont donc :

- de mettre en place une méthode de descente d'échelle probabiliste de type analogue sur l'ensemble de la France métropolitaine ;
- de mesurer et améliorer la cohérence spatiale des précipitations régionalisées par cette méthode ;
- de développer un outil de descente d'échelle applicable dans des contextes variés, de la reconstitution de précipitations passées au changement climatique en passant par la prévision à court terme, pour des applications sur les ressources en eau et la production hydroélectrique.

Les données et scores utilisés sont présentés dans la partie I. La partie II regroupe une présentation de la méthode SANDHY (*Stepwise ANalogue Downscaling method for HYdrology*) (Chapitre 4), son optimisation sur la France entière dans le Chapitre 5, et sa validation dans le Chapitre 6. Dans la partie III, trois pistes

sont explorées pour réduire l'espace des paramètres : utiliser ceux optimisés pour d'autres zones (Chapitre 7), regrouper les zones utilisant les mêmes paramètres (Chapitre 8), et utiliser un prédictant moins asymétrique pendant l'optimisation (Chapitre 9). Dans la partie IV, le Chapitre 10 s'intéresse à mesurer la cohérence spatiale en utilisant une méthode de vérification spatiale (SAL) sur les bassins du Rhône et de la Durance. La méthode SAL est ici adaptée à des simulations probabilistes et des stratégies diverses concernant la cohérence spatiale et la performance locale sont évaluées. Enfin, l'impact des différentes stratégies de descente d'échelle sur la performance des simulations des débits de la Durance est discutée dans le Chapitre 11.

Données

Les prédicteurs utilisés dans cette étude proviennent de la réanalyse ERA40 (Uppala et al., 2005) mise à disposition par le Centre Européen de Prévision Météorologique à Moyen Terme. Les prédictants utilisés sont les précipitations issues de la réanalyse Safran (Vidal et al., 2010), basée sur un découpage de la France en 608 zones climatiquement homogènes d'une taille moyenne de 900km^2. L'optimisation et la validation de la méthode utilisent les précipitations moyennes par zones, alors que la partie IV tire profit des sorties de la réanalyse Safran sur une grille de résolution 8km. La période commune aux deux réanalyses court du 1er août 1958 au 31 juillet 2002.

Critères de performance

Le principal critère de performance utilisé ici est le CRPS (*Continous Ranked Probability Score*) (Brown, 1974), un critère de vérification pour des prévisions probabilistes d'une variable continue, largement adopté pour la vérification des prévisions d'ensemble. Le CRPS mesure la différence entre l'observation et une fonction de répartition simulée, sans nécessiter le recours à une répartition des simulations en différentes classes. Dans le cas d'une simulation déterministe, il est égal à l'erreur absolue moyenne. Le CRPSS (*Continous Ranked Probability Skill Score*) est de plus utilisé pour comparer la performance des simulations par rapport à une simulation de référence. Celle-ci correspond tout au long du document à la distribution climatique de la saison – 121 jours autour de la date cible extraits de l'archive des dates analogues.

Par ailleurs, des méthodes de vérification spatiale ont été développées pour la vérification des champs de précipitations déterministes issues des modèles atmosphérique de haute résolution et non-hydrostatiques qui permettent de simuler explicitement la convection profonde. La méthode SAL (*Structure, Amplitude, Location*) (Wernli et al., 2008) regroupe trois caractéristiques : l'amplitude (A) –

la quantité totale de précipitation tombée sur le domaine, la localisation (L) – le centre de masse des précipitations et la distance entre les objets de précipitation et leur centre de masse, et la structure spatiale (S) – la taille et la forme des objets de précipitation. Cette méthode est ici adaptée pour des simulations probabilistes.

2 Descente d'échelle sur la France

SANDHY

La méthode de descente d'échelle SANDHY (Ben Daoud et al., 2011a) utilisée ici est une méthode de sélection par analogie à 4 niveaux successifs. Le premier niveau est une sélection sur la température à 600 hPa et 925 hPa, basé sur une distance euclidienne. Le deuxième niveau est une sélection sur le géopotentiel à 500 hPa et 1000 hPa avec le critère de Teweles-Wobus (TWS) (Teweles and Wobus, 1954). Le TWS mesure la similarité des gradients des champs et est donc un critère de forme. Le troisième niveau est une sélection sur la vitesse verticale à 850 hPa et le quatrième niveau est une sélection sur l'humidité relative à 850 hPa multiplié par l'eau précipitable. Ces deux niveaux utilisent la distance euclidienne comme critère de similarité. 25 dates analogues sont retenues à la fin du processus.

Le logiciel implémentant la méthode SANDHY rassemble des fonctionnalités pour l'optimisation, la simulation et la validation dans un seul programme. Il est écrit en Fortran 2003, parallélisé avec OpenMP, et utilise un fichier de configuration sous forme de *namelist*. Les entrées et sorties sont au format NetCDF. Les tâches implémentées sont :

- L'optimisation des domaines du prédicteur géopotentiel avec un algorithme de domaines rectangulaires croissants multiples, qui utilise le CRPS comme fonction objectif. Le nombre de domaines à agrandir et retenir peut être choisi par l'utilisateur (Chapitre 5) ;
- Le calcul de cartes de pertinence (Chapitre 5) ;
- La validation avec un calcul du CRPS et du biais (Chapitre 6) ;
- L'agrégation des zones utilisant les mêmes domaines de prédicteur avec trois méthodes différentes (Chapitre 8) ;
- Le calcul du CRPS climatologique ;
- Le calcul du pourcentage de dates analogues communes entre deux zones ou stations.

Optimisation locale des domaines de prédicteur

Une optimisation des domaines spatiaux du prédicteur géopotentiel est ici réalisée. Les domaines pour les autres prédicteurs correspondent au point de grille ERA-40 le plus proche de la zone cible.

Les cartes de pertinence représentent la performance en termes de CRPSS pour des domaines élémentaires à quatre points de grille pour le prédicteur géopotentiel. Celles-ci sont établies pour quelques zones d'étude et montrent que la région de pertinence maximale est en général au sud-ouest de la zone d'étude pour les zones sous influence Atlantique, et au sud-est de la zone d'étude pour les zones sous influence Méditerranéenne. La région de forte pertinence forme un cercle relié à la région présentant les plus forts gradients des anomalies de géopoteniel sur les jours pluvieux. La région de pertinence maximale est au sud de ce cercle pour les zones sous influence Atlantique et à l'est du cercle pour les zones sous influence Méditerranéenne.

L'optimisation des domaines est réalisée avec un algorithme de domaines rectangulaires croissants qui est ici adapté pour faire croître plusieurs domaines en parallèle afin de tester plus de domaines et de proposer un ensemble de domaines optimisés à la sortie du processus. L'optimisation est réalisée individuellement pour chaque zone Safran en partant des points de grilles autour de la zone cible et 5 domaines sont finalement retenus pour chaque zone. On constate une grande diversité de domaines pour le prédicteur géopotentiel sur l'ensemble de la France. Les centres des domaines sont distribués suivant l'emplacement géographique des zones, mais avec des différences marquées entre les versants au vent et sous le vent du Massif Central. La variabilité du centre entre les cinq domaines pour chaque zone est en général faible, sauf pour quelques zones montagneuses dans la partie sud-est de la France. La taille des domaines trouvés est elle plus variable : de petits domaines sont trouvés à l'ouest et au nord du pays, et de grands domaines sont trouvés dans les régions montagneuses du sud-est. Dans les régions de plaine du sud-est les domaines sont allongés en direction nord-sud et dans une bande orientée est-ouest au nord du Massif Central – juste sous le parallèle situé à 47,5° nord – les domaines sont allongés en direction est-ouest. La variabilité des tailles entre les cinq domaines pour chaque zone est très grande pour certaines zones.

Des tests complémentaires montrent une sensibilité au point du départ de l'optimisation pour une zone située au nord de la bande identifiée ci-dessus, et qui pour un point de départ au sud de celle-ci, trouve des domaines allongés en direction est-ouest qui remportent une meilleur performance. Une sensibilité à la longueur de l'archive est trouvée pour une zone en Ardèche avec une forte saisonnalité. Pour cette dernière zone, une expérience où 99 domaines sont retenus lors de l'optimisation est menée dans l'idée de mieux explorer l'espace des domaines possibles. Parmi les 5 meilleurs domaines, deux sont communs aux expériences avec 99 et 5

domaines. Ceux de l'expérience à 99 domaines sont en moyenne plus grands.

Validation de SANDHY

Une étape de validation est nécessaire pour vérifier le comportement de la méthode dans des conditions d'expérience diverses, qui ont en commun d'utiliser la période Août 1982 – Juillet 2002 (nommée *late*) pour optimiser les domaines pour le prédicteur géopotentiel :

La simulation de référence Elle simule les précipitations de la période *late* en utilisant cette même période comme archive pour chercher des dates analogues ;

La validation hors échantillon Elle simule la période Août 1958 – Juillet 1978 (nommée *early*) – période indépendante de la période d'optimisation – en utilisant la période *late* comme archive ;

L'expérience d'archive alternative Elle simule la période *late* en utilisant la période *early* comme archive ;

L'expérience de domaines de prédicteurs suboptimaux Cette dernière expérience simule la période *early* en utilisant cette même période comme archive. Elle utilise de fait la meilleure archive possible et teste l'influence de domaines de prédicteurs optimisés sur une autre période.

La distribution spatiale des valeurs de CRPSS pour la simulation de référence est directement liée à celles des précipitations moyennes. Les plus fortes valeurs correspondent aux régions les plus arrosées, comme les Alpes, les Cévennes, la façade ouest du Massif Central, les Vosges, ainsi que la côte Atlantique. Les performances les plus basses se situent sur la côte méditerranéenne, la côte est de la Corse ainsi que l'est du Massif Central. Une diminution du CRPSS de l'ordre de 0,03 est observée pour la plupart des zones dans l'expérience de validation hors échantillon. On constate de plus une légère augmentation de la performance pour quelques zones dans le sud-est du pays, ainsi qu'une très forte diminution (-0,17) pour une zone située dans le Massif Central. Cette zone présentait une large différence de précipitation moyenne identifiée entre les périodes *late* et *early*, en raison de l'ajout d'une station d'altitude entre ces deux périodes. Cette inhomogénéité dans la chronique de précipitations sur la zone conduit à une non-stationnarité de la relation prédicteur-prédictant puisque une situation atmosphérique donnée ne se traduit plus par les mêmes cumuls de précipitations. C'est une hypothèse centrale de la descente d'échelle statistique qui n'est pas respectée ici. Dans l'expérience d'archive alternative, la perte de performance est faible et très uniforme spatialement.

Enfin, dans l'expérience de domaines de prédicteurs suboptimaux, les pertes de performance sont similaires à celles de l'expérience de validation hors échantillon.

Un critère de biais est de plus calculé entre les valeurs observées et la moyenne des 25 valeurs issues des dates analogues. Cette moyenne est en général biaisée positivement pour la simulation de référence. Les régions les plus arrosées, notamment les façades au vent des chaînes de montagnes, tendent à présenter de plus forts biais positifs que les régions plus sèches. La distribution spatiale des biais dans l'expérience des domaines de prédicteurs suboptimaux est similaire à celle de la simulation de référence, avec une tendance à de plus forts biais positifs dans le sud du pays. Pour l'expérience de validation hors échantillon, les changements ne sont pas homogènes spatialement. Pour l'expérience d'archive alternative, les biais changent substantiellement par rapport à la simulation de référence. Ces différences semblent être reliées à la différence de précipitation moyenne entre les deux périodes.

Les pertes de performance de la méthode apparaissent lorsque la période simulée est différente de la période d'optimisation, tandis que les changements sur le biais sont plus marqués lorsqu'elle est différente de la période d'archive. Ces changements sur le biais, qui dépendent de la différence de climatologie entre les périodes de simulation et d'archive, ont de sérieuses implications sur l'application de la méthode. En effet, toute correction de biais simple basée sur l'hypothèse d'une homogénéité temporelle pour un endroit donné ne serait pas valide.

3 Réduction de l'espace de paramètres

Exploitation des domaines des zones voisines

Chacun des 847 domaines de prédicteurs trouvés localement sur la France est appliqué à l'ensemble des 608 zones pour la période *late*. Cela permet de rechercher quel domaine donne la meilleure performance pour chacune des zones. Idéalement, celui-ci devrait avoir été trouvé localement par l'algorithme d'optimisation, mais ce n'est en fait pas le cas pour la majorité des zones.

Il y a deux raisons possibles à cela. La première concerne le choix du domaine élémentaire de départ de l'algorithme. Ce domaine élémentaire a été choisi comme le plus près de la zone cible, mais ce n'est pas toujours le meilleur choix. Au nord d'un parallèle situé à 47,5° de latitude nord se trouve une région présentant de larges différences en termes de CRPSS qui semble lié à cette limitation de l'algorithme. Une seconde raison concerne le choix de la maille ERA-40 la plus proche prise comme domaine d'analogie pour les prédicteurs locaux. Ce choix apparaît comme non optimal pour un nombre important de zones situées principalement sur la côte méditerranéenne.

Les meilleurs domaines pour les prédicteurs issus de cette analyse (appelés *best*) peuvent être comparés à ceux optimisés localement (appelés *optim*). Les centres des domaines pour le prédicteur géopotentiel sont modifiés pour les zones situées au nord du parallèle à 47,5° de latitude nord et sont moins variables dans l'espace sur les massifs montagneux. La variabilité au sein des 5 domaines est plus forte dans certaines régions et moins importantes dans d'autres. La taille moyenne des domaines varie généralement de manière plus lisse sur le territoire et leur variabilité au sein des 5 domaines apparaît plus petit. En résumé, les limitations de l'algorithme imposées par les choix évoqués ci-dessus peuvent être nettement atténuées en considérant des domaines optimisés pour d'autres zones.

Agrégation des zones de prédictants

Les différentes configurations de paramètres obtenues sur la France peuvent être réduites en regroupant les zones partageant les mêmes paramètres. Ceci devrait permettre de renforcer la cohérence spatiale à l'intérieur de chaque groupe et aider à définir une échelle spatiale type en dessous de laquelle des paramètres identiques ne conduisent qu'à une perte de performance limitée.

Les zones sont tout d'abord regroupées selon la corrélation des rangs des précipitations journalières observées, en utilisant l'algorithme *affinity propagation* (Frey and Dueck, 2007), qui sélectionne un nombre de groupes optimal ainsi qu'un membre représentatif de chaque groupe. Cet algorithme conduit à 52 groupes de taille similaire. Cette approche qui agrège les zones selon le comportement de leurs prédictants n'est toutefois pas nécessairement la plus adaptée dans le cadre de l'approche par analogie, et il est plus pertinent de réaliser l'agrégation selon le comportement des prédicteurs associés à ces zones.

La similarité entre zones est alors exprimée en termes de fraction de dates analogues communes afin d'inclure des informations sur les domaines des prédicteurs et de bénéficier de distances continues. Des groupes plus petits sont trouvés ainsi, notamment dans la partie sud du pays, avec un nombre total de 87.

La fraction de dates analogues communes contient de l'information sur la similarité des prédicteurs, mais elle n'assure pas que la performance locale est entièrement maintenue. Les zones sont finalement regroupées selon la définition même des domaines des prédicteurs. L'inconvénient est que la distance associée – le nombre de domaines *optim* ou *best* communs entre deux zones – peut seulement prendre des valeurs discrètes (0 à 5), ce qui pose problème aux algorithmes d'agrégation. En effet, des égalités de distance apparaissent très souvent et conduisent à des solutions de regroupement non-uniques.

Trois méthodes d'agrégation – simple, maximum d'occurrence, et *variable group agglomerative hierarchical clustering* (Fernández and Gómez, 2008) – sont appliquées pour constituer des groupes utilisant les mêmes domaines de prédicteurs.

Le plus petit nombre de groupes est obtenu par l'algorithme le plus simple, même si ce nombre exact dépend de l'ordre dans lequel les zones sont traitées. Regrouper les domaines *best* permet de réduire fortement le nombre de groupes par rapport aux domaines *optim*, et les barrières d'agrégabilité ont un sens physique beaucoup plus marqué que pour *optim* où elle sont plus ou moins imposées par la grille des prédicteurs. Dans tous les cas, la réduction du nombre de groupes total sur la France provient des zones de plaine du pays. Le sud-est montagneux du pays ne voit pas la taille de ses regroupements évoluer considérablement.

Les bassins versants de la Durance et du Rhône, utilisés plus tard comme bassins d'étude, présentent une faible agrégabilité par domaines de prédicteurs communs ou dates analogues communes. Assurer une cohérence spatiale au sens de ces agrégations implique de fait une perte de performance sur ces bassins.

Utilisation de précipitations transformées comme prédictant

Jusqu'ici les données de précipitation ont été considérées telles quelles comme prédictant, au contraire des études précédentes (Bontron, 2004; Ben Daoud, 2010) utilisant des versions antérieures de SANDHY. L'effet d'une transformation des précipitations sur la performance de la méthode et sur l'aggrégabilité est ici étudiée.

Les précipitations transformées sont obtenues en les divisant par le quantile de pluie journalière maximal annuel de période de retour 10 ans, et en prenant la racine carrée de ce rapport. Cette transformation est ainsi censée éliminer les écarts systématiques entre différentes zones et réduire l'asymétrie de la variable.

L'optimisation des domaines du prédicteur géopotentiel est effectuée à nouveau en utilisant les précipitations transformées comme prédictants. Le centre des domaines trouvés dans cette expérience – appelée *transformed* – est très similaire à celui de l'expérience *optim*. En revanche, la variabilité au sein des 5 meilleurs domaines est plus restreinte et aussi plus homogène dans l'espace. La taille moyenne tend elle à augmenter et à s'homogénéiser spatialement. La variabilité de cette taille au sein des 5 meilleurs domaines tend elle aussi à devenir plus homogène sur la France.

La performance des simulations de la variable de précipitations transformées est généralement plus élevée que celle des simulations de pluies brutes, quelle que soit l'expérience. En revanche, la différence de performance entre les expériences *optim* et *transformed* est en faveur des domaines de l'expérience *optim* lorsqu'elle est calculée sur les précipitations brutes. La structure spatiale du CRPSS reste similaire quelle que soit l'expérience ou le prédictant considéré. Les faibles performances rencontrées dans les vallées soumises à l'effet de foehn peuvent ainsi être reliées aux erreurs déjà faibles de la simulation de référence (climatologique) sur ces zones.

En résumé, utiliser des précipitations transformées pour l'optimisation des domaines de prédicteurs ne conduit pas à de meilleures performances, mais à une meilleure aggrégabilité selon les domaines de prédicteurs communs.

4 Vérification spatiale à l'échelle du bassin versant

Cohérence spatiale – Cas d'étude sur la Durance et le Rhône

Simuler des champs de précipitations aux propriétés spatiales réalistes est crucial pour la modélisation hydrologique distribuée. La méthode de vérification spatiale SAL (Wernli et al., 2008), qui identifie des propriétés spatiales des champs pertinentes pour l'hydrologie, est ici adaptée pour l'évaluation de champs de précipitations régionalisés probabilistes. Cette adaptation se base ainsi sur des champs de probabilité de dépassement d'un seuil de précipitations. L'utilisation de seuils différents pour les observations et les simulations permet d'évaluer les caractéristiques spatiales des événements pluvieux intenses. Des scores de performance sont construits à partir des composantes structure et localisation de la version probabiliste du SAL afin d'évaluer la cohérence spatiale des simulations probabilistes.

Un ensemble d'expériences est mené pour identifier la meilleure stratégie en termes de configuration de domaines de prédicteur pour garantir à la fois une performance locale élevée et une cohérence spatiale sur le bassin considéré. Ces expériences vont ainsi d'une configuration uniforme des domaines sur le bassin à une configuration hétérogène utilisant des domaines spécifiques pour chaque zone du bassin. La configuration uniforme conduit à des champs de précipitations réguliers et une structure des objets de précipitations plus réaliste. En revanche, la localisation est moins précise que pour la configuration hétérogène pour les deux bassins d'étude. Les expériences mettant en œuvre 5 domaines de prédicteurs plutôt qu'un seul conduisent à une meilleure performance locale et une meilleure représentation de la structure des objets de précipitation.

L'augmentation de la performance locale due à l'utilisation de domaines de prédicteurs multiples s'avère être une propriété générale pour toutes les zones en France, et trouve sa source dans une meilleure résolution des distributions simulées.

Sur le bassin de la Durance, la médiane des simulations sous-estime les précipitations pour toutes les expériences, alors que la moyenne présente un biais positif dans le nord du bassin et négatif dans le sud, avec une zone de transition variable selon l'expérience. Les histogrammes des rangs montrent que les simulations sont trop dispersives pour toutes les expériences.

Impacts hydrologiques de la cohérence spatiale dans la descente d'échelle

Les scores de performance pour la structure et la localisation des objets de précipitation montrent des résultats différents selon les expériences et le bassin considéré. On essaye ici d'identifier quelle expérience conduit aux meilleures simulations hydrologiques sur le bassin de la Durance et si les scores en précipitation peuvent être reliés aux scores sur les débits.

Les simulations hydrologiques et le calcul des performances associées ont été réalisés dans le cadre du stage de Master de Judith Eeckman (Eeckman, 2014) supervisé par J.-P. Vidal, F. Tilmant et moi-même. Le modèle distribué à base physique J2000 (Krause, 2002) est utilisé pour simuler les débits à partir de données météorologiques issues des expériences décrites ci-dessus. Les températures minimum et maximum ainsi que l'évapotranspiration potentielle sont échantillonnées depuis le jeu de données DuO (Magand et al., 2014) aux mêmes dates analogues que les précipitations échantillonnées dans la réanalyse Safran. Le modèle hydrologique ne peut pas être forcé directement avec des entrées probabilistes à chaque point de grille telles que produites par SANDHY. Un ensemble de scénarios déterministes et équiprobables est construit pour chaque expérience en utilisant des permutations aléatoires des dates analogues pour chaque jour et chaque zone. Les mêmes permutations sont utilisées pour les zones utilisant les mêmes domaines de prédicteur pour conserver cette élément de cohérence spatiale. Cet ensemble de scénarios permet ainsi de produire un ensemble probabiliste de simulations de débit.

La performance des simulations à la station de Cadarache (aval du bassin) est étudiée sur la période *late*. Les simulations utilisant une configuration uniforme des domaines présentent une plus large dispersion que celles utilisant une configuration hétérogène. Les efficiences de Nash-Sutcliffe probabiliste (Bulygina et al., 2009) et de Kling-Gupta (Kling et al., 2012) montrent des résultats similaires avec de meilleurs scores pour les expériences utilisant une configuration hétérogène. Les expériences mettant en œuvre des domaines de prédicteurs multiples conduisent à des résultats très proches de celles utilisant un seul domaine, malgré leur meilleure performance sur les précipitations. Parmi les trois scores de précipitation, celui sur la localisation des objets de précipitation présente la plus grande corrélation avec les performances sur les débits. En conclusion, il apparaît intéressant d'optimiser localement les domaines de prédicteurs, puisque cela conduit à de meilleures simulations de débit, notamment en termes de corrélation temporelle.

5 Conclusions et perspectives

Conclusions

La méthode de descente d'échelle SANDHY (*Stepwise ANalog Downscaling method for HYdrology*) a été étendue à l'ensemble de la France continentale et à la Corse en optimisant les domaines spatiaux du prédicteur géopotentiel localement pour chacune des 608 zones climatiquement homogènes couvrant le territoire. Une variabilité spatiale importante de ces domaines de prédicteurs a ainsi été trouvée. La non-optimisation des domaines pour les trois autres prédicteurs (température, humidité et vitesse verticale) s'est révélée être une faiblesse de la stratégie d'optimisation. Celle-ci a pu en revanche être surmontée en considérant localement des domaines de prédicteurs identifiées pour d'autres zones en France.

La performance de la méthode sur les précipitations apparaît plus importante pour les climats humides. Les expériences de validation ont montré une perte de performance non uniforme lorsque la période de simulation est différente de la période d'optimisation, et un changement non-uniforme sur le biais lorsque la période de simulation est différente de la période d'archive. L'optimisation de la méthode sur une variable de précipitation transformée n'induit pas de changements majeurs sur ces résultats.

Deux grandes approches peuvent être identifiées pour assurer la cohérence spatiale dans une descente d'échelle par analogie. La première consiste à sélectionner les mêmes dates analogues pour une région la plus grande possible, a priori au détriment de la performance locale. L'alternative consiste à utiliser les dates analogues de domaines de prédicteurs optimisés localement, mais sous l'hypothèse que des zones proches présentent des paramètres suffisamment similaires pour assurer des transitions spatiales douces. Pour la première approche, il est possible de définir un seuil de perte de performance pour définir les limites de la région où une configuration uniforme est pertinente. L'étude d'aggrégabilité des zones de prédictants selon les domaines de prédicteurs associés et une étude récente sur la transférabilité spatiale des dates analogues (Chardon et al., 2014) montrent que la transférabilité spatiale de SANDHY est peu importante et que la première approche n'est pas très adaptée pour une méthode d'analogie à plusieurs niveaux.

Dans le cadre de la première approche, d'importantes discontinuités des champs de précipitation régionalisés peuvent apparaître à la frontière de deux régions considérées comme homogènes, surtout quand les régions sont grandes et le seuil de perte de performance lâche. Lors de l'application de la méthode de descente d'échelle sur des bassins versants emboîtés ces frontières devraient ainsi être redéfinies quand un bassin est ajouté. La deuxième approche évite cet écueil en rendant la région cible indépendante de l'application et plus aisément extensible spatialement.

L'optimisation locale sur des petites unités spatiales favorise des transitions lisses entre des zones voisines dans le cadre de la deuxième approche. Elle évite des discontinuités non-physiques importantes, néanmoins des discontinuités mineures un peu partout doivent être acceptées. Pour obtenir un résultat encore plus lisse, des précipitations aux dates analogues trouvées chez les voisins pourraient être intégrées dans l'estimation locale probabiliste des précipitations. L'utilisation des précipitations analogues issues de domaines de prédicteurs multiples augmente la performance locale des précipitations régionalisées.

Une méthode de vérification spatiale a été adaptée pour des simulations probabilistes et a été mise en œuvre pour évaluer les propriétés spatiales des champs de précipitations régionalisés par SANDHY. Un ensemble d'expériences a été mis en place pour comparer les deux grandes approches pour la cohérence spatiale, en termes de performance locale des précipitation simulées, et du réalisme de la structure spatiale et de la localisation des objets de précipitation simulés. Une configuration homogène des domaines de prédicteurs conduit à une structure plus réaliste des objets de précipitation, alors qu'une configuration hétérogène capture mieux la localisation de ces objets.

Des simulations hydrologiques avec un modèle distribué sur le bassin de la Durance réalisées durant un stage ont montré que les différentes approches pour la cohérence spatiale induisent des différences sur la performance des débits simulés associés. En revanche, l'utilisation de domaines de prédicteurs multiples localement a comparativement peu d'influence. Une configuration hétérogène avec des domaines de prédicteurs variables spatialement conduit à une meilleure performance des simulations de débit.

Le logiciel SANDHY developpé durant cette thèse a été mise en œuvre avec succès par plusieurs collègues. L. Caillouet l'utilise pour reconstruire des sécheresses historiques sur la France et J.-P. Vidal l'a appliqué pour faire de la descente d'échelle de précipitations sur 81 stations en Argentine dans le cadre du deuxième atelier CORDEX-ESD. Ces deux applications ont nécessité l'utilisation de prédicteurs et prédictants différents du contexte de la thèse (réanalyses différentes, données station ponctuelles).

Perspectives

Plusieurs questions restent ouvertes à l'issue de cette thèse, par exemple comment les domaines *best* ou ceux optimisés sur la précipitation de bassin se comportent en validation hors échantillon, quelle est la significativité statistique des différences de CRPSS, ou encore quelle est la sensibilité de la version probabiliste de SAL à certains choix effectués durant son développement.

Des développements additionnels sur SANDHY pourraient porter sur l'utilisation de la température de surface de la mer comme prédicteur additionnel, du flux

d'humidité comme prédicteur alternatif à la variable composite d'humidité utilisée actuellement, des prédicteurs spécifiques à certaines régions comme les Tropiques, ou encore sur la prise en compte de données manquantes, essentielle lorsque la méthode utilise des données stations comme prédictants.

Pour une application à la descente d'échelle de projections climatiques futures, plusieurs questions supplémentaires se posent. Le développement de méthodes de sous-échantillonage des projections désagrégées qui puissent concilier d'une part l'intégration des différentes types d'incertitudes liées à la chaîne de modélisation hydroclimatique et d'autre part les contraintes des modèles d'impact, s'avère par exemple indispensable pour fournir une estimation plausible des incertitudes sur les impacts locaux, notamment hydrologiques.

La descente d'échelle par analogie et les questions afférentes liées à la cohérence spatiale connaissent un réel intérêt en France. Mis à part les présents travaux, deux autres thèses par J. Chardon et G. Dayon portent actuellement sur des thèmes similaires avec des questions complémentaires. Les producteurs d'électricité comme la CNR utilisent depuis plusieurs années la descente d'échelle par analogie pour la prévision opérationnelle, et le Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations (SCHAPI) cherche à présent à implémenter un modèle d'analogues sur la France pour la prévision opérationnelle de crues.

Part I

Context and state of the art

1 Context and objectives

1.1 Context

Anticipating future freshwater ressources on short and long timescales is an important societal issue, because human needs in terms of drinking water-, food- and energy supply depend on it. Specifically the availability of freshwater for human consumption, irrigation and hydropower production are concerned. Future drought characteristics are for example studied to potentially develop adaptation strategies, for example planting crops that need less irrigation. Anticipating the potential hydropower production is an economic matter for hydropower companies on the short timescale to forecast the production and get the best possible price and on the long timescales to calculate the profitability of investments. Furthermore, studying past and present day streamflow regimes, droughts and floods helps to increase our understanding of their characteristics. However, the directly measured records are still of very limited length in most places. To extend the record backwards in time, reconstructions are needed.

Global climate models or earth system models (GCMs) are valuable tools to study global climate change and processes. 110 to 150 years long reanalysis datasets have recently become available (Compo et al., 2011; Dee et al., 2014) and can be used for reconstructions of the less recent past. Both the GCMs and the extended reanalyses have a rather coarse skillful resolution and use parametrisations of subgridscale processes. This leads to strongly smoothed local surface variables like precipitation which is why they are not directly suited for studying local impacts. A possible solution to this is downscaling.

1.2 Statistical downscaling methods

Downscaling seeks to deduce local scale climate elements over a limited area from larger scale information at coarser resolution. There are two main types of downscaling: dynamical downscaling using regional models (e.g. Philandras et al., 2011;

Christensen and Christensen, 2003, 2007; Hohenegger et al., 2008; Pfeiffer and Zängl, 2011) and statistical downscaling. In the case of dynamical downscaling regional models are provided with boundary values from GCMs and perform simulations of the atmospheric state over their limited domain with higher resolution. This allows for more processes being explicitly simulated than in the GCMs, but subgridscale processes like for example cloud microphysics are still parametrised and the model topography is smoothed compared to the real world one. The idea of statistical downscaling is to use observed relationships between large scale predictors and local climate variables to build statistical models that can translate variations of large scale variables to variations of local scale ones. Downscaling models should be able to reproduce the historical evolution of local variables when they are driven with observed large scale predictors (Zorita and von Storch, 1999).

There is a large variety of statistical downscaling methods. Following Schoof (2013) they evolved from synoptic climatology, a discipline that describes surface climate as a function of large scale atmospheric circulation and local conditions. In review papers downscaling methods have been classified into families by type of statistical technique used like regression methods, weather pattern based approaches and stochastic weather generators (Wilby and Wigley, 1997; Fowler et al., 2007) or by the way predictors are considered (Maraun et al., 2010). In a perfect prognosis (PP) setting predictors and predictands are related event by event and the predictor-predictand relationship is established using observed predictors and predictand. Model output statistic (MOS) relates simulated predictors and observed predictands during the calibration period and weather generators generate timeseries with prescribed statistical properties. MOS corrections are specific to the numerical model used, while PP relationships are independent of the model, but assume that the models correctly simulate the predictors (Maraun et al., 2010). Nowadays most downscaling methods are "hybrid" methods, that is methods that combine techniques from different families to benefit from the advantages of different methods and to capture deterministic variance, that can be explained by large scale variations and stochastic variance due to local phenomena and conditions. This makes them more difficult to classify. The following overview of statistical downscaling methods describes the main ingredients used in precipitation downscaling. The techniques are grouped into regression methods, weather generators, bias correction methods (MOS), weather pattern based approaches and examples of hybrid applications. The overview focus on daily precipitation, probabilistic methods and multisite applications since these aspects are relevant for this study.

1.2.1 Regression methods

Linear regression or mulitple linear regression is one of the most widely used downscaling methods (e.g. Nicholas and Battisti, 2012; Ning et al., 2012; Hanssen-Bauer

et al., 2003; Wilby et al., 1998a), but it assumes a gaussian distribution of the predictand (Maraun et al., 2010; Zorita and von Storch, 1999), or at least predictors and predictand having similar shaped cumulative distribution functions (CDFs) (Ribalaygua et al., 2013). This works well for predictands like temperature, but is not suited for precipitation or wind speed for example. Themeßl et al. (2010) found multiple linear regression to be defective in estimating non-normally distributed daily precipitation and concluded that this technique can not be recommended for regional climate model (RCM) precipitation correction.

Generalised linear models (e.g. Chandler and Weather, 2002; Chandler, 2005; Frost et al., 2011) are better suited for precipitation downscaling since they allow for non-gaussian predictands (Maraun et al., 2010). Constructed analogues (Hidalgo et al., 2008) search linear combinations of past large scale variables that match the target situation and use the same linear combination for obtaining the local scale counter part. It has already been mentioned in Van den Dool (1994) as an idea to overcome the problem of too short archives for finding good analogues for the analogue method. Canonical correlation analysis (CCA) is a linear method that identifies pairs of patterns whose time evolution is optimally correlated (Zorita and von Storch, 1999) (e.g. Fernández and Sáenz, 2003; Hertig et al., 2012). Typically CCA performs comparatively well when correlation between observed and downscaled variables is measured, which is not very surprising given that it is optimised to reproduce correlations. Other linear regression methods are censored quantile regression (Friederichs and Hense, 2007) and vector generalised linear models (Maraun et al., 2011). Regression models allow to extrapolate out of the observed range of variability and are therefore used as components to account for trends and long term deterministic variance. Care has to be taken that the relationships established are actually valid within and out of the observed range. It is therefore important to not solely rely on statistical correlations but to keep in mind the physical relationships that may have led to the observed correlations. This kind of considerations are important as well to increase the confidence that the predictor-predictand relationship will be stationary.

Since simple linear methods are not that successful for precipitation downscaling and arguing that the processes leading to precipitation are non-linear, artificial neural networks (ANN) were developed for downscaling (e.g. Fernández-Ferrero et al., 2009; Trigo and Palutikof, 2001; Haylock et al., 2006). ANNs tend to overestimate trace precipitation and thus wet-day occurrence (Wilby and Wigley, 1997; Wilby et al., 1998b; Fowler et al., 2007), but this problem can be overcome by ANNs that treat the occurrence and the amount of precipitation separately (Harpham and Wilby, 2005). Non-linear censored quantile regression or quantile regression neural network (QRNN) is proposed by Cannon (2011). A major issue with ANNs and other regression methods is that they tend to underestimate the

variance of precipitation and the frequency of heavy rainfall because they are fitted to give the best estimation of the mean of the observed distribution (Zorita and von Storch, 1999).

1.2.2 Weather- or rainfall generators

Weather generators generate random sequences of weather variables, that is synthetic time series, with statistical properties resembling those of observed weather (Maraun et al., 2010).

Rainfall occurrence is often modelled using a first-order Markov chain (e.g. Cordano and Eccel, 2012), but this leads to an underestimation of the interannual variability (Maraun et al., 2010). Chen et al. (2010) solved this problem by modelling the low frequency variability using observed power spectra. An alternative to Markov chains are Nyman-Scott rectangular pulses (Kilsby et al., 2007; Burton et al., 2008, 2010; Blenkinsop et al., 2010; van Vliet et al., 2012).

Several choices are possible to model the skewed distribution of rainfall intensities: the gamma distribution (e.g. Bellone et al., 2000; Hingray and Mezghani, 2007), mixtures of exponential distributions (Wilks, 1998), mixtures of gamma-, normal- or log-normal distributions with extreme value distributions (Carreau and Vrac, 2011) or distributions from the Tweedie family (Dunn, 2004). To model extreme precipitation the Generalized Pareto Distribution (e.g. Cooley et al., 2007) or the Generalized Extreme Value distribution (e.g. Maraun et al., 2011) can be used. There are fewer models for multisite precipitation simulations, most of them using a transformed Gaussian distribution (e.g. Leblois and Creutin, 2012; Kioutsioukis et al., 2008; Brussolo et al., 2008; Rebora et al., 2006).

In a downscaling context modern weather generators are often conditioned on the daily evolution of large-scale weather states and thus coupling a weather pattern based component with the random sampling (Ailliot et al., 2014). The nonhomogeneous hidden Markov models (NHMM) belong to this category (Hughes and Guttorp, 1999; Bellone et al., 2000; Charles et al., 2004; Greene et al., 2011). Mehrotra and Sharma (2005) extended the usual NHMM using continuous weather states instead of a few discrete ones.

Issues with weather generators are for example underestimated spell lengths (e.g. Vrac et al., 2007) and how to adjust the parameters for future projections in a physically realistic and consistent way (Wilby, 1997). The conditional random sampling techniques are used in downscaling as components representing the stochastic variance.

1.2.3 MOS - Bias correction methods

While it is debateable if bias correction alone is already downscaling, many downscaling methods include a bias correction component, either for the predictand variable in case of MOS approaches or for the predictor variables for PP approaches. Therefore bias correction methods are presented here.

The simplest bias correction approaches correct only the bias of the mean (e.g. Gutmann et al., 2012), but there are more sophisticated approaches. Quantile mapping (e.g. Amengual et al., 2012; Piazza et al., 2011; Vidal and Wade, 2009; Quintana Seguí et al., 2011) considers different intensities individually and cumulative distribution function transform (CDF-t) (Michelangeli et al., 2009; Piani et al., 2010; Vrac et al., 2012; Lavaysse et al., 2012) seeks at finding a transfer function for the whole distribution. XCDF-t (Kallache et al., 2011) takes extremes into account when fitting the transfer function and is therefore better suited when studying extremes.

Bias correction and spatial disaggregation (BCSD) (Wood et al., 2004) consists of two steps: first a CDF transform type bias correction and then an interpolation to finer scales. BCSD is frequently used in the United States (e.g. Voisin et al., 2010). A similar method is applied by Früh et al. (2006) in the Alps with the difference that only the bias in the mean is corrected in the bias correction step and errors in the variability are accounted for in the disaggregation procedure. Joint variable spatial downscaling (Zhang and Georgakakos, 2012) maps multiple variable CDFs and uses analogues instead of interpolation in the spatial disaggregation step along with a technique to expand the range of historical analogues.

Maraun (2013) warns that quantile mapping should not be used to change scales because this introduces artefacts in the spatial structure of precipitation. Similarly Vrac et al. (2012) notes that CDF-t is not designed to correct the spatial correlation of the simulations. A further challenge is to bias correct the rain day frequency that is scale-dependent as well.

Bias correction methods assume that the (distributional) bias is stationary. This assumption is not valid where model biases are related to physical processes for example the parametrisation of melting. Such a bias will always occur around 0° C, but the quantile of 0° C in the temperature distribution changes under global warming (Formayer et al., 2008).

1.2.4 Weather pattern based approaches

Weather pattern based approaches aim at classifying days into weather patterns, states or classes and then resample the local variables from the days belonging to the same weather state as the target day. Weather pattern based approaches and specifically analogue resampling methods are discussed in a bit more detail below

since the downscaling method used in this work (presented in chapter 4) belongs to this group of methods.

Wilby and Wigley (1997) states that weather pattern approaches are appealing, because they are based on sensible physical linkages between the scales. On the downside they noted, that the relationships between weather patterns and precipitation might not be stable enough, such that the precipitation characteristics of a specific weather type may vary from year to year. Another drawback could be that GCMs are not able to resolve the circulation patterns that lead to hydrological extreme events (Maraun et al., 2010) or have biases concerning the position of the storm tracks that can impact the detection of weather patterns depending on the way they are defined. An advantage is that weather pattern techniques allow for nonlinear relationships in a straightforward way (Maraun et al., 2010).

Weather pattern components are used in hybrid approaches either to describe deterministic or stochastic variance. In the first case they are used in a first step to define to which circulation type a day belongs and then using for example a separate model or different model parameters for each circulation type. In the second case a resampling from similar past situations (analogues) is used to account for the stochastic variance.

1.2.4.1 Weather typing schemes

There are different ways to define weather patterns or weather classes. Empirical orthogonal functions (EOFs) from pressure data (Goodess and Palutikof, 1998), indices from sea level pressure (SLP) data (Conway et al., 1996; Goodess and Jones, 2002), cluster analysis (Fowler et al., 2005; Gutiérrez et al., 2004; Rust et al., 2010; Boé et al., 2006; Plaut et al., 2001), classification and regression trees (CART) analysis (Zorita and von Storch, 1999; Zorita et al., 1995), self organising maps (Hewitson and Crane, 2006; Ning et al., 2012), correlation clustering (Vrac and Yiou, 2010) or fuzzy rules (Bárdossy et al., 2002) are applied. Philipp et al. (2010) provides a collection of 17 automated and 5 subjective weather- and circulation type classifications for Europe.

Vrac and Naveau (2007) integrates extreme value models within a weather typing approach in order to improve the simulation of extremes that are in general not well simulated using weather type or analogue approaches. Weather pattern based approaches reproduce well observed means and variability, but underestimate persistence (Zorita and von Storch, 1999). Weather type approaches can account for changes in precipitation due to changes in the occurrence of certain weather types, but have difficulties to account for changes of the precipitation characteristics inside a given weather type (Boé and Terray, 2008).

1.2.4.2 Analogue methods

The analogue method can be seen as a limit case of weather typing where each day defines a weather type (Boé et al., 2006). Benestad (2007) described the analogue method as a search in a historical archive describing all large scale weather events in the past together with the local measurements of the quantity of interest. Before its use in downscaling since the 1990s (e.g. Zorita et al., 1995; Timbal et al., 2003), the analogue method has been used in forecasting (e.g. Lorenz, 1969; Gutiérrez et al., 2004; Bannayan and Hoogenboom, 2008a,b; Guilbaud and Obled, 1998b) or to predict short term climate fluctuations (e.g. Barnett and Preisendorfer, 1978) – nowadays it would be rather called seasonal forecasting. An example of an analogue method used in operational precipitation forecasting for the Rhône catchment in France is OPALE developed at Compagnie Nationale du Rhône (CNR) (Bompart et al., 2009). Most often analogue methods are employed in a perfect prognosis setting, but model output statistics (e.g. Turco et al., 2011) and weather generator applications (e.g. Yiou, 2014) exist as well.

Important choices in the application of the analogue method include the choice of the distance measure to define similarities (Martin et al., 1997; Matulla et al., 2008) and the choice of the predictor domains (Gutiérrez et al., 2004; Wetterhall et al., 2005). Euclidean distance is probably the most common distance measure. Hamill and Whitaker (2006) found that for a skewed predictor measuring the distance in terms of ranks rather than absolute values leads to less biased results. Matulla et al. (2008) tested different distance measures, Euclidean distance, absolute distance, cosine distance and Mahalanobis distance to measure distances in the EOF space and found that the Euclidean distance performs well in most cases while the Mahalanobis distance is less advisable. Ribalaygua et al. (2013) found Euclidean distance to perform better than correlation. Toth (1991) compared similarity measures to define the most similar days for an analogue method, however the purpose was not downscaling but forecasting geopotential fields of the following days in this case. Gradients were the best measure for 1 day leadtime, while for longer leadtimes Euclidean distance was better. This is in line with the better performance of the Teweles and Wobus score (TWS) that is based on gradients in a PP setting (Guilbaud and Obled, 1998b).

A drawback of the standard analogue method is that it can not produce precipitation amounts that were not observed in the past (Maraun et al., 2010; Benestad, 2007). This makes it not suited for studying extremes, however it is not the only method that tend to produce too moderate extremes (Fowler et al., 2007). Van den Dool (1994) states that it is easier to find good analogues closer to the mean, because the probability density is higher there compared to the extremes. Imbert and Benestad (2005) suggested an extension of the analogue method to allow extrapolation outside the range of the observed values. The linear trend of the analogue

simulation is subtracted and the linear trend simulated using a linear regression method is added instead. Benestad (2010) used quantile mapping in addition to the imposed linear trend to correct the extremes, however the high percentiles and their trend turned out to be difficult to estimate from the data record.

The analogue method is often used in the form of a k-nearest neighbour approach (knn) where not the most similar historical situation is selected but it is randomly chosen between the k most similar situations as proposed by Lall and Sharma (1996) and Gangopadhyay et al. (2005). Boé et al. (2006) found that averaging several nearest neighbours doubles the correlation for daily precipitation but the variance becomes greatly underestimated. Buishand and Brandsma (2001) found a tendency to underestimate the autocorrelation and the variance when averaging.

The analogue method is generally good at reproducing the observed variance of precipitation which is in general not the case for pure regression methods (Benestad, 2007). Zorita and von Storch (1999) pointed out that the analogue method can be used for normally and nonnormally distributed local variables, produces the right level of variability and preserves the spatial covariance of local variables.

The standard procedure for multisite simulations with the analogue method is that the same analogues are used for the whole area and thus maintaining the spatial covariance structure, but this is not always the case. Themeßl et al. (2010) for example uses individual analogue models for each station with separately selected predictors. To apply the analogue method to larger areas, Hamill and Whitaker (2006) used a moving spatial window, tiling together local analogues.

Hwang and Graham (2013) combines bias correction and analogues in a MOS setting and concludes that this method is superior to BCSD in reproducing the observed spatial correlation and variance and performs equally well for other statistics. This highlights the capacity of analogue methods to reproduce spatial characteristics. Similarly Voisin et al. (2010) uses analogues to calibrate and downscale medium range ensemble weather forecasts and found that analogue methods improve the reliability of the forecast while BCSD does not. This is at the cost of lower temporal correlation.

1.2.5 Hybrid methods

In this section some examples are presented that combine at least two of the techniques described in the previous sections.

Hybrids of weather generators and analogues provide an easy way to multisite downscaling (Mehrotra and Sharma, 2005). For example Buishand and Brandsma (2001) proposes a nearest neighbour resampling scheme conditioned on summary statistics of the weather variables and large scale circulation. Orlowsky et al. (2008, 2010) resample time blocks instead of single events, respecting a trend derived

using linear regression to improve the temporal consistency of the simulated time series and this method indeed improves the dry- and wet spell statistics.

The widely used Statistical DownScaling Model (SDSM) (Wilby et al., 2002; Souvignet and Heinrich, 2011; Wetterhall et al., 2006, 2007; Wilby and Dawson, 2013) is a hybrid of a weather generator and a regression method. SDSM underestimates precipitation and its variance (Souvignet and Heinrich, 2011). Zorita and von Storch (1999) used a simple version of the analogue method as benchmark and found that more complicated methods not necessarily outperform the analogue method. However, Wetterhall et al. (2007, 2006) found that SDSM outperforms the analogue benchmark method.

Yang et al. (2010) proposed a method that combines three techniques, regression, circulation type classification and stochastic rainfall generation. The drawback of this method is the high number of parameters to estimate and a possible lack of data for these estimates especially in the case of less frequent circulation types. Hybrids of weather pattern or analogues and regression methods following a similar principle are widely used (Boé et al., 2006, 2007, 2009; Cannon, 2007; Hingray and Mezghani, 2007; Vrac et al., 2007; Mezghani and Hingray, 2009; Fernández-Ferrero et al., 2009, 2010; Piazza et al., 2011).

1.2.6 General remarks

Common to all statistical downscaling methods is that they assume stationarity of the predictor-predictand relationships (Schmith, 2008). Fowler et al. (2007) concludes that in general, temperature can be downscaled with more skill than precipitation, winter climate can be downscaled with more skill than summer due to stronger relationships with large-scale circulation, and wetter climates can be downscaled with more skill than drier climates. Regression methods perform best in terms of correlation, but analogue methods and weather generators better reproduce observed distributions (Gutiérrez et al., 2013). In a deterministic setting most downscaling methods underestimate the spatial variability, while adding uncorrelated noise risks to destroy the spatial structure (Maraun et al., 2010).

1.3 Issues relevant for hydrology

Torres et al. (2008) suggests that downscaling methods should provide probabilistic output reflecting their uncertainty for a given situation and Krzysztofowicz (1983, 1998) notes that probabilistic forecasts are beneficial in decision making. While Maraun et al. (2014) states that stochastic methods allow for random sampling of time series that can be treated as an ensemble of deterministic simulations, it could be argued that this approach is likely to underestimate the value of prob-

abilistic simulations and that probabilistic measures are preferable for assessing probabilistically downscaled fields.

Maraun et al. (2010) notes that in the context of rainfall-runoff modeling there is evidence that spatial structure is important for small, rapidly responding catchments and for catchments that are larger than the scale of typical precipitation events. Given the importance of the spatial structure of precipitation, an ideal downscaling method for hydrological modelling delivers probabilistic results with realistic spatial patterns. The spatial properties of statistically downscaled precipitation are rarely studied. To diagnose if a (downscaling) model simulates precipitation fields with a realistic spatial structure, measures for the spatial properties are needed. One possibility is to use spatial correlation (Voisin et al., 2010; Moron et al., 2006) or anomaly correlation (Jolliffe et al., 2012, section 6.3.2 pp.98). Another possibility is to employ spatial verification methods (Gilleland et al., 2009). Neither of them have been applied to probabilistically downscaled fields yet.

In addition inter-variable coherence of the downscaled fields is important for hydrological modelling, because streamflow depends not only on precipitation but also on other atmospheric variables like temperature and evapotranspiration as well as vegetation and soil properties. Storage of water in solid state and releases from this storage through melting depend on temperature and atmospheric humidity. The evapotranspiration depends on temperature, humidity, wind, radiation and the vegetation.

Among the downscaling methods that are in principle capable to respond to these requirements are dynamical downscaling using regional high resolution models and resampling methods like analogues. Dynamical high resolution modelling is rather costly, especially if ensemble simulations are run in order to provide probabilistic output and they are often subject to large biases. Resampling methods can provide probabilistic output in a straightforward way, but are limited by the observed record and tend to underestimate extremes.

1.4 Objectives

The objectives of this thesis are (1) to extend the stepwise analogue downscaling method for hydrology, SANDHY hereafter, optimised for the Seine and the Saône catchment by Ben Daoud (2010), to the whole of France and to analyse its performance on precipitation, (2) to study and improve the spatial coherence of the downscaled precipitation preserving the local performance and (3) to build a downscaling tool (software) that can be used in future research projects.

After the presentation of the data and diagnostics used (part I), the approach pursued is to first build and validate local downscaling relationships over France(part II), then explore ways to reduce the parameter space maintaining the

local downscaling skill (part III) and finally to define measures and experiments for assessing the spatial coherence of downscaled precipitation at the catchment scale and its hydrological consequences (part IV).

More specifically, the SANDHY method is presented in chapter 4 and the predictor domains for the geopotential predictor are optimised locally over France in chapter 5. The the SANDHY method and its local downscaling relationships are validated on an independent period in chapter 6. Three ways are explored to reduce the parameter space maintaining the local downscaling skill: (1) Considering the relationships found for zones elsewhere in France (Chapter 7), (2) building groups of zones that are supposed to use the same relationships using cluster algorithms (Chapter 8) and (3) considering a less skewed predictand variable in the optimisation procedure (Chapter 9). Measures for spatial coherence based on the spatial verification method SAL are developed and applied to several downscaling experiments over the Durance and the Rhône catchments (Chapter 10). Finally the impact of these experiment designs on the performance of streamflow simulations with a distributed hydrological model for the Durance catchment is discussed (Chapter 11).

2 Data, study area and precipitation characteristics

In this chapter the study region and the datasets for the predictors and the predictands for SANDHY, that aims at downscaling precipitation are introduced. Some characteristics of the precipitation in the study region are presented below.

2.1 Study area and predictand data set – Safran

The study area, that is the target area of the downscaling, comprises the mainland of France and Corsica. Figure 2.1 shows an annotated map of France with the mountain ranges, catchments and regions that are mentioned later in the text.

The predictand dataset is the precipitation from the Safran reanalysis. Safran is a meteorological analysis system that performs an optimum interpolation of station measurements using vertical gradients from a background forecast taken from the ERA-40 reanalysis until the year 2002 and the operational analysis from the European Center of Medium-range Weather Forecast (ECMWF) weather forecast model for the later years (Quintana-Seguí et al., 2008; Vidal et al., 2010). The analysis is performed for 608 climatologically homogeneous zones covering France, that are shown in Figure 2.2. The size of the zones varies, but is about 1000 km^2 on average. The Safran near surface reanalysis starts in August 1958 and is continued in near-realtime. The data are available as gridded data on an 8km grid, that is used in part IV. For the development of the downscaling relationships (part II) zone average time series are used. Vidal et al. (2010) and Quintana-Seguí et al. (2008) found the precipitation analysis to be robust and not biased but with an average root-mean-square error (RMSE) of 2.4mm day^{-1} . Quintana-Seguí et al. (2008) attributes this to the spatial heterogeneity within the zones. The Safran analysis assumes that differences inside a zone depend only on elevation while the occurrence of small storms contributes to the spatial heterogeneity of precipitation as well. The RMSE is proportional to the precipitation. There are not enough dry

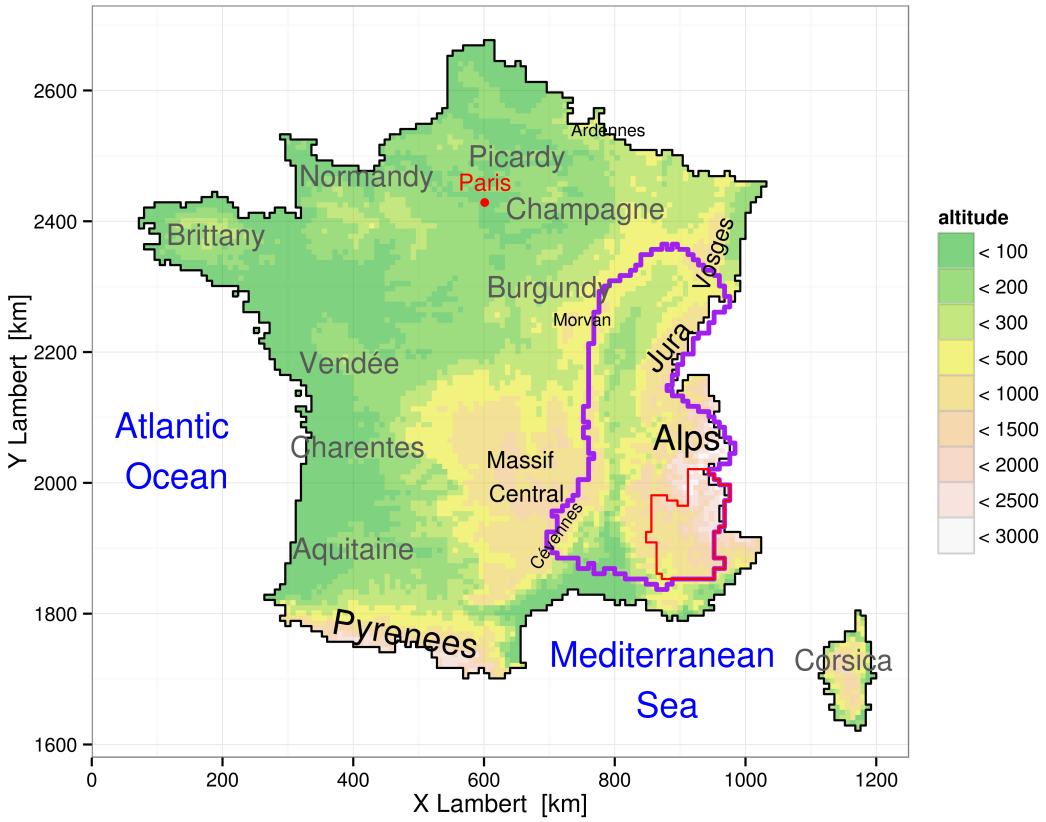

Figure 2.1: Map of France with the contours of the Rhône catchment between Pougny and Beaucaire in purple and the Durance catchment at Cadarache in red.

days, too much trace precipitation and underestimated strong precipitation due to the aggregation on zone level (Quintana-Seguí et al., 2008). Safran is more precise over France than other gridded observation datasets like E-OBS (Haylock et al., 2008) or the global scale precipitation datasets GPCC or GPCP (Szczypta et al., 2011). Safran data are available at an hourly time step, but the hourly values of precipitation are interpolated from a daily analysis. Daily data are used in this work.

2.2 Predictor data set – ERA40

ERA-40 (Uppala et al., 2005) is a 44-year long (1958 – 2002) global reanalysis dataset from the ECMWF. The three dimensional variational data assimilation method (3DVAR) was used to create the dataset. A characteristic of global re-

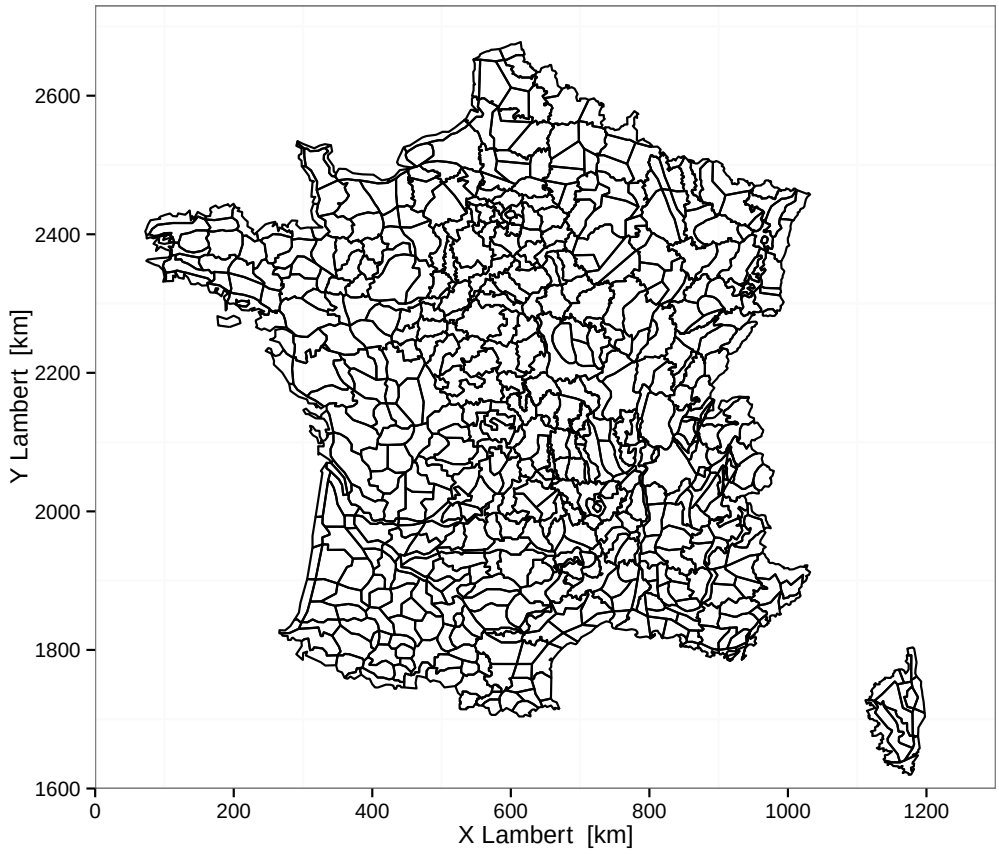

Figure 2.2: Climatologically homogeneous zones.

analyses in general is that the model used is the same for the whole time period to maximise the temporal consistency. The model used for ERA-40 is the forecast model that was in use at the ECMWF when the reanalysis project started in the late 1990s but with a lower horizontal spectral resolution of T159. Here ERA-40 data are used on a 2.5° grid. Ben Daoud (2010) tested ERA-40 data on a 1.25° grid but didn't find any improvement of the downscaled precipitation using the higher resolution predictor dataset.

ERA-40 incorporates many different data sources including weather stations, radiosoundings, drifting buoys and radiance measurements from satellites. The changes in the observing system over time, especially the beginning of the satellite era (1979) and the assimilation of data from infrared sounders in the 1990s, affect the temporal consistency of the reanalysis but improves the quality. This is the case for any reanalysis that seeks to incorporate as many data as possible. In the more recent ERA-Interim reanalysis homogenised radiosonde data and adaptive bias correction of satellite data is used to mitigate this issue (Dee et al., 2011).

ERA-40 is however used (1) to ensure consistency with Safran where ERA-40 has been used as a first guess for computing vertical profiles of near-surface variables, (2) to ensure consistency with previous development work by Ben Daoud et al. (2011a) and exclude that differences in the results are due to the use of a different predictor dataset and (3) to have a long archive that allows for validation with a rigorous split-sample approach (see also section 2.1.1 in section 5.1).

2.3 Precipitation characteristics

In this section some characteristics of the zone average Safran precipitation are shown. The graphical representation shows 8-km gridded data, but with identical values for all grid cells belonging to a given zone.

Figure 2.3 shows the average daily precipitation over the 1 Aug. 1958 – 31 July 2002 common archive period between ERA-40 and Safran. Unsurprisingly the highest precipitation amounts are found at the windward side of the mountains. It is quite wet along the Atlantic coast and rather dry along the Mediterranean coast, the flat areas in the northern part of the country and in the foehn-prone valleys of the Allier and the Loire in the Massif Central and the Rhine valley in the Northeast.

Apart from the daily average sums of precipitation it is also interesting to know if the precipitation is distributed over a large number of days or if it falls on a few days per year only. Therefore figure 2.4 shows the percentage of dry days for each zone. A dry day is defined as a day with less than 0.1mm day^{-1} of zone average precipitation. The zones with a high percentage of dry days are mainly located in the southeastern part of the country between the Massif Central and the southern Alps and along the Mediterranean coast. The Mediterranean climate in the southeastern part of France is characterised by dry summers. The lowest percentage of dry days are found in Brittany.

High precipitation percentiles are of interest because they are related to the precipitation distributions studied later in this section and they show what high daily precipitation amounts that occur a few times per year mean in different places. The 99th percentile of precipitation at each zone is shown in figure 2.5. The values of the 99th percentile range from 16mm day^{-1} to nearly 90mm day^{-1} in the Cévennes. In general the 99th percentile of precipitation has high values in the mountains and the south-eastern part of the country and lower values around 20-30mm day^{-1} elsewhere.

Another way to characterise the precipitation is by its distribution. If extremes are not of interest, wet day precipitation amounts can be approximated using gamma distributions (e.g. Yang et al., 2010; Bellone et al., 2000; Hingray and Mezghani, 2007). Since at this stage the interest is rather to explore the main

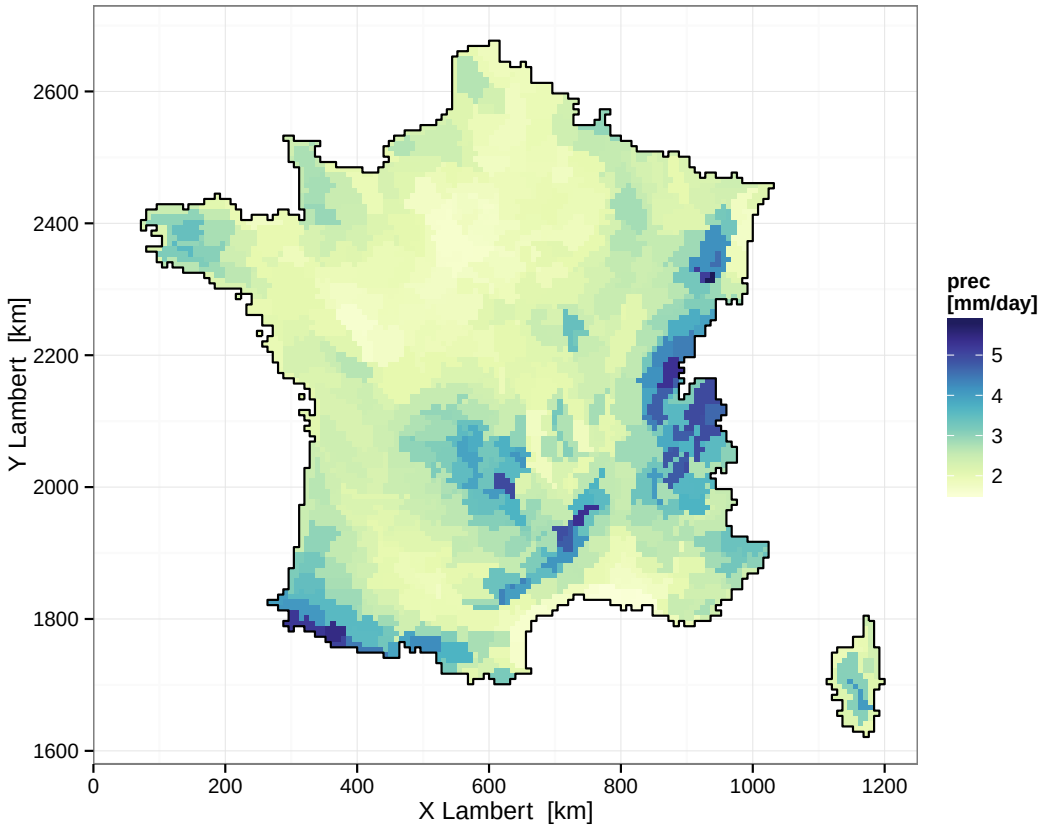

Figure 2.3: Mean precipitation over the 1 Aug. 1958 – 31 July 2002 period.

characteristics of the dataset than finding the best possible fit, this approximation is considered to be sufficient.

The probability density function of the gamma distribution can be written as

$$f(x;k,\theta) = \frac{1}{\theta^k}\frac{1}{\Gamma(k)}x^{k-1}e^{-\frac{x}{\theta}} \tag{2.1}$$

for x, k and $\theta > 0$. k is the shape parameter, θ the scale parameter and Γ the Gamma function. If k is a positive integer $\Gamma(k) = (k-1)!$.

Gamma distributions are here fit to the wet day precipitation amounts for the 608 zones using maximum likelihood fitting. Figures 2.6 and 2.7 show the spatial distribution of the scale and the shape parameters. The scale parameter is large in the Southeast and at the windward side of the mountain ranges and shows a very similar spatial pattern as the 99th percentile daily amounts in figure

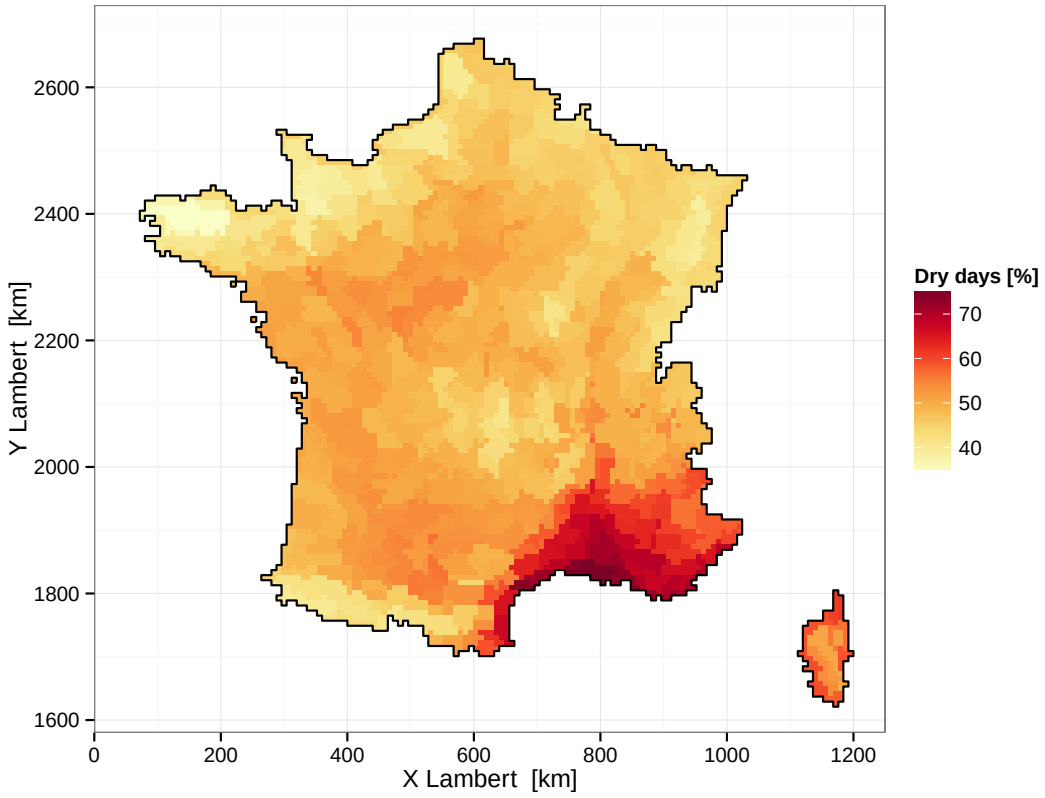

Figure 2.4: Percentage of dry days.

2.5. The shape parameter is small in the Southeast, medium-sized in the Alps, the Pyrenees and Brittany and large elsewhere. Figure 2.8 shows the observed frequency distributions of wet day amounts for three selected zones and the fitted gamma distributions. The three selected zones will be used as case study zones later on in the chapters 5, 7 and 9. The zone named *Saône* is located in the eastern part of the Burgundy region and is an example for a zone where the fitted gamma distribution has a comparatively small scale- and a large shape parameter. The second zone *Arve* is located in the Alps and has medium values of the two parameters. The last zone *Ardèche* is located in the Cévennes and has large scale and small shape parameter. The parameters values for these three zones are tabulated in table 2.1. The fitted distributions appear to fit well for the *Saône*, quite well for the *Arve* and less well for the *Ardèche*.

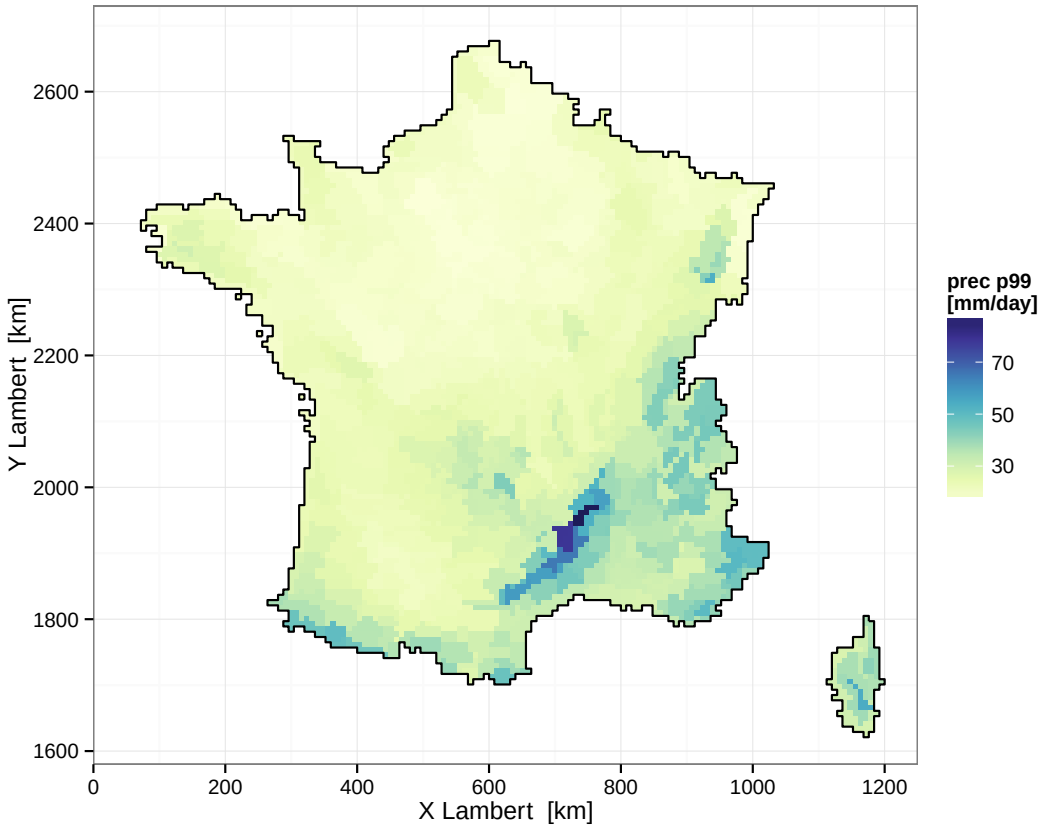

Figure 2.5: 99th percentile of precipitation in the 1 Aug. 1958 – 31 July 2002 period.

zone	Saône	Arve	Ardèche
shape	0.80	0.57	0.42
scale	6.07	15.13	19.89

Table 2.1: Scale- and shape parameters of gamma distributions fitted to wet day precipitation amounts of selected zones.

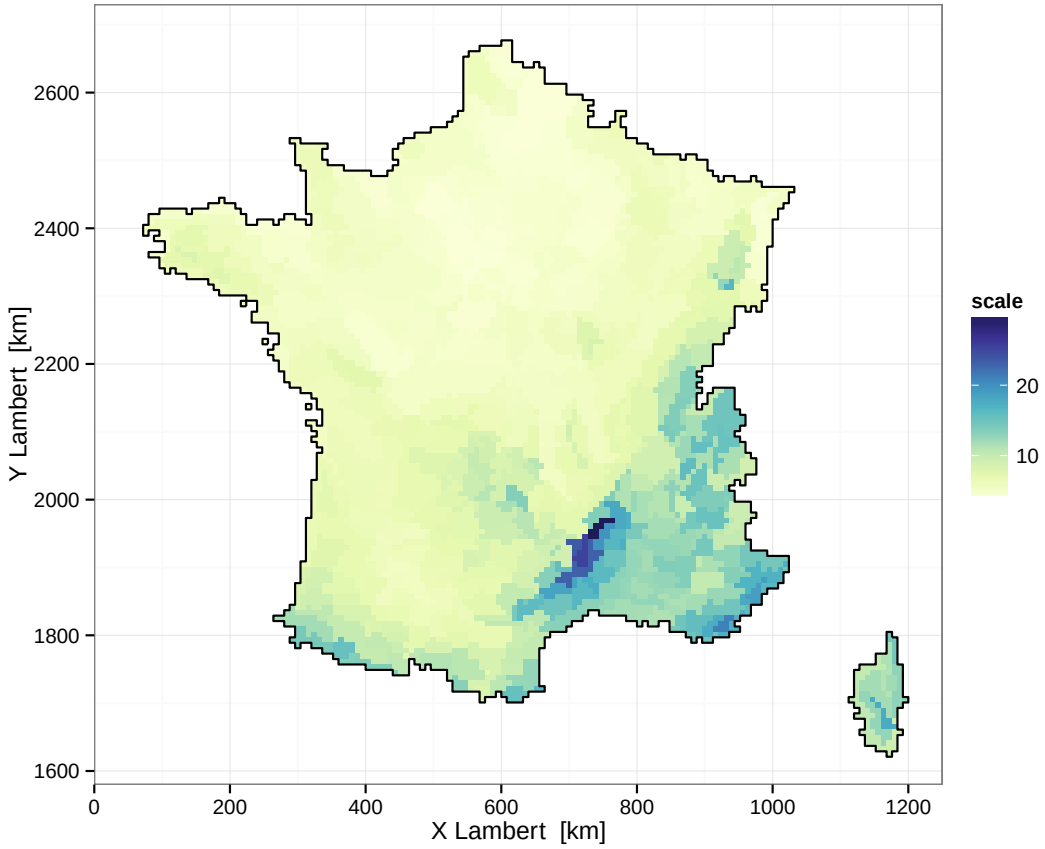

Figure 2.6: Scale parameter of gamma distributions fitted to wet day amounts.

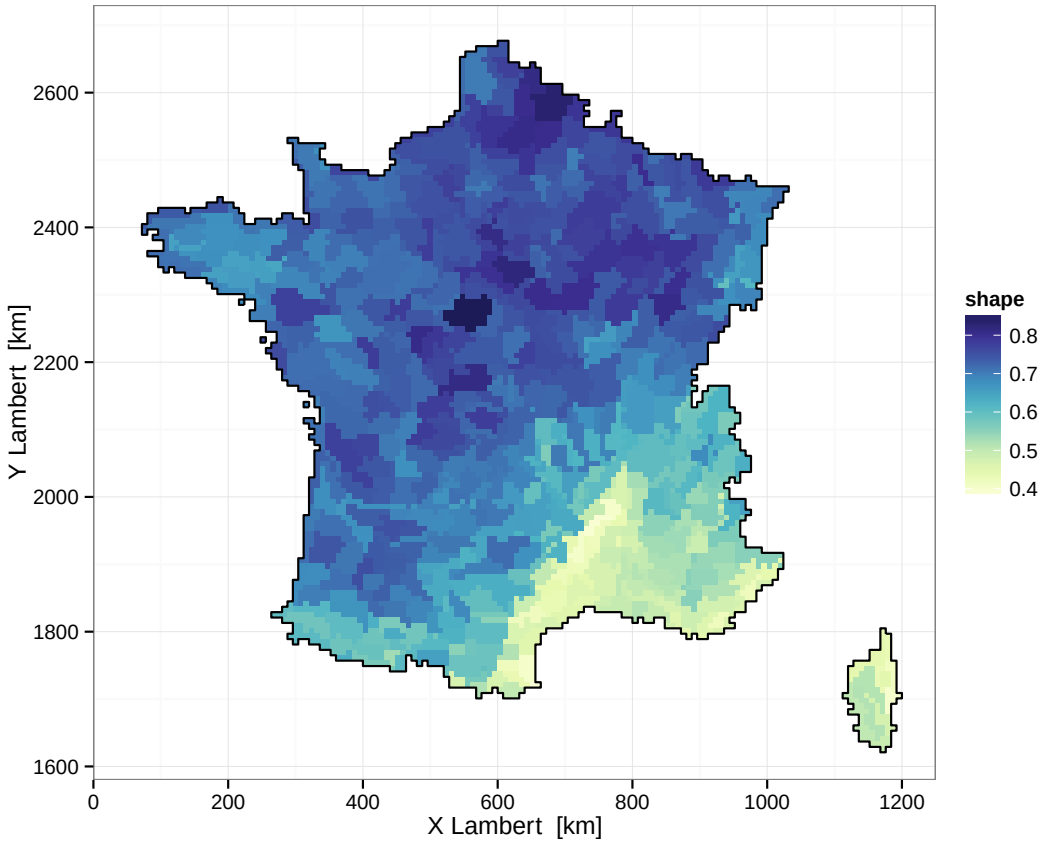

Figure 2.7: Shape parameter of gamma distributions fitted to wet day amounts.

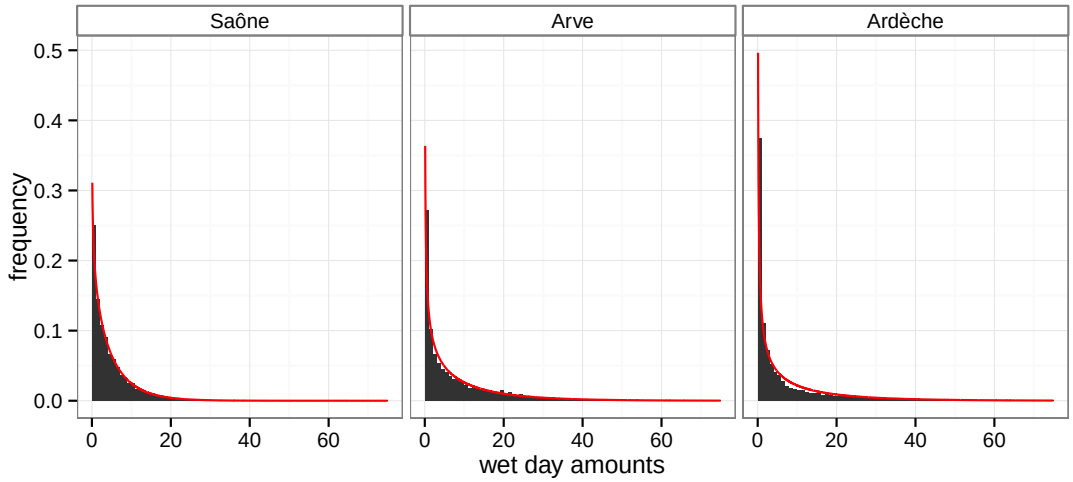

Figure 2.8: Observed frequency distribution of wet day amounts at three selected zones (black) and fitted gamma distributions (red).

2.4 Temporal differences

In chapter 5 predictor domains will be optimised using the last 20 years of the common archive period between ERA-40 and Safran from August 1982 to July 2002 called *late* in the following. Then validation will be carried out on the first 20 years of the common dataset from August 1958 to July 1978 called *early* in the following. In this section the differences in the precipitation characteristics between these two periods are presented.

Figure 2.9 that shows the difference between the average precipitation in the *late* period and the *early* period. During the *early* period there was more precipitation in the southern part of the country and less in the northern part. This is in line with Vidal et al. (2010) who found non-significant positive linear precipitation trends in the North and negative ones in the South over the 1958–2008 period. Zone 367 in the Massif Central, or more precisely the Sancy Massif, stands out with a high positive difference, while the zones around show rather small and mostly negative differences. During most of the *early* period there was only one station reporting in this zone, while later on a second station at higher altitude has been added, such that the mean altitude of the precipitation records in this zone increased by 150m from 1050m to 1200m. Since in general higher precipitation amounts are observed at higher altitudes this leads to higher average precipitation in the *late* period and to an inhomogeneity in the zone average precipitation.

Figure 2.10 shows the difference in the number of dry days between the *late* and the *early* period. In the northern part of the country there are less dry days

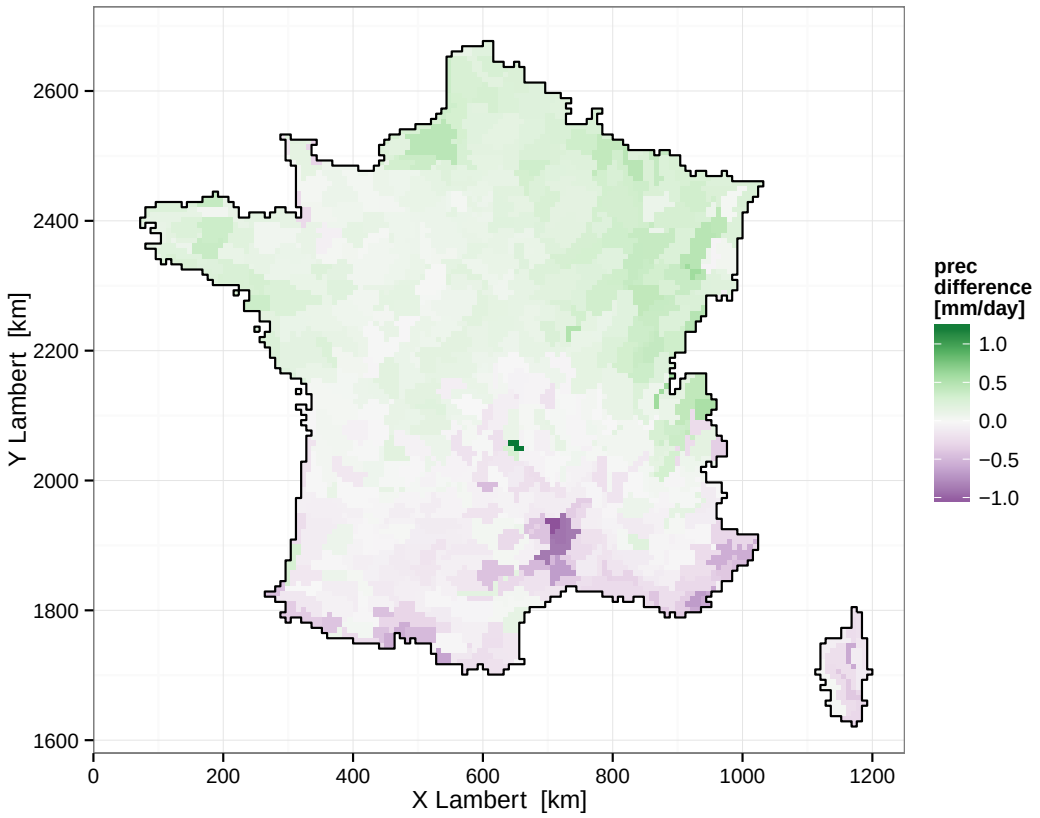

Figure 2.9: Difference in mean precipitation *late* minus *early*.

in the *late* period than in the *early* period. Especially Brittany is affected by more rainy days in the *late* period. In the Pyrenees and the Massif Central there are more dry days in the *late* period than in the *early* period. In the southeastern part of the country the difference between the two periods is small.

Figure 2.11 shows the difference between the 99th percentiles of the *late* and the *early* period precipitation. In the northern part of the country most of the zones have higher values in the *late* period than in the *early* period. In the southern part of the country the picture is more patchy. Along the Mediterranean coast and the southern Cévennes the 99th percentiles are higher in the *early* period. The largest absolute difference can be found for a zone at Corsica where the picture is extremely patchy.

The precipitation characteristics presented in this chapter, notably their spatial distributions, are compared to the spatial distribution of performance measures in

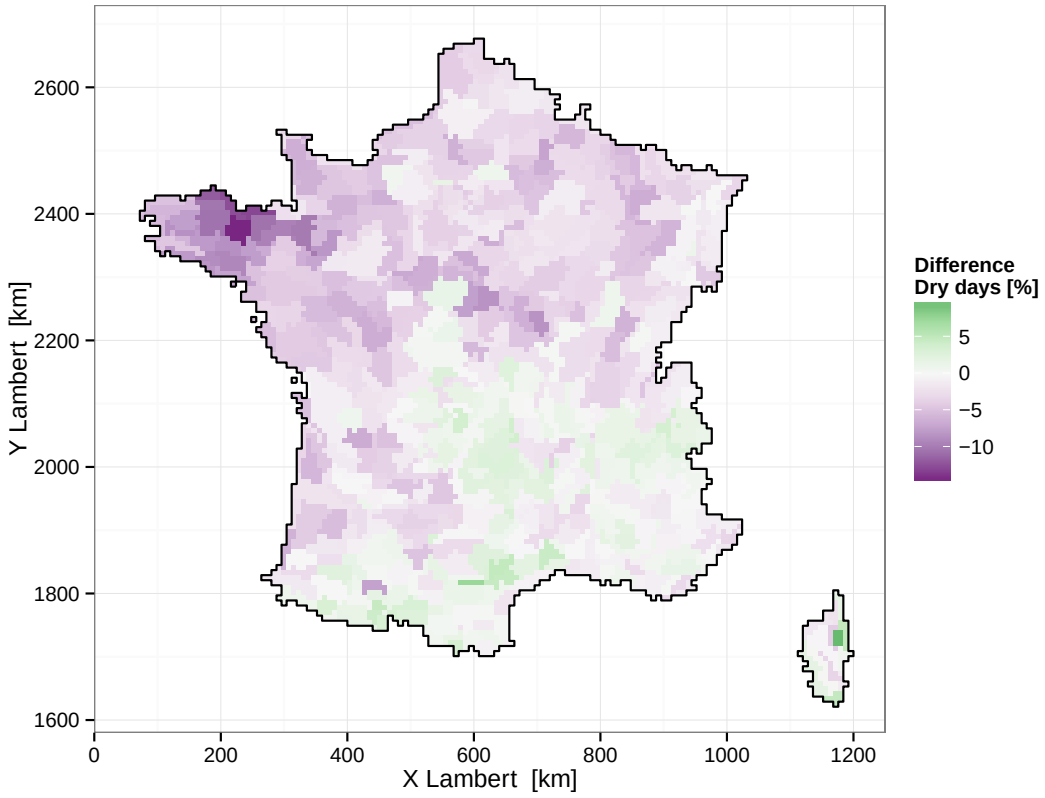

Figure 2.10: Difference in the percentage of dry days *late* minus *early*.

the chapters 3 and 9. Now that we are familiar with the data we proceed to the main diagnostics for simulated precipitation used in this study.

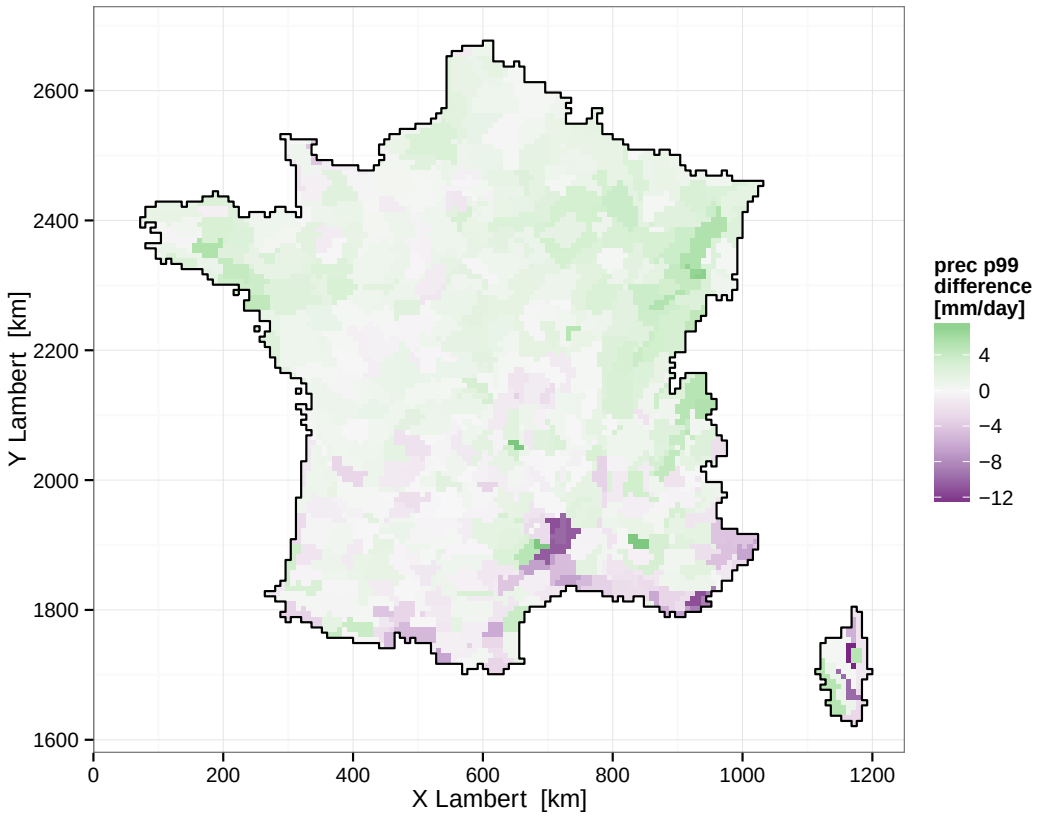

Figure 2.11: Difference of 99th percentiles *late* minus *early*.

3 Performance diagnostics

To diagnose the performance of SANDHY in simulating precipitation, measures for the performance are needed. Verification indices developed to diagnose the performance of weather- or streamflow forecasts are used to measures some kind of error or performance over a given time period. For different kinds of forecasts like binary events (e.g., rain or no rain, exceedance of a threshold...), multi-category events, continuous variables or spatial fields, deterministic or probabilistic forecasts, different verification methods or scores are available.

For deterministic forecasts of categoric events the most common methods are based on contingency tables, for example hit rate, false alarm ratio, probability of detection, critical success index, equitable thread score or the relative operating characteristic (ROC) (Mason, 1982). Well known measures for deterministic forecasts of continuous variables are bias, mean absolute error (MAE), root mean squared error (RMSE) and correlation. The Taylor diagram (Taylor, 2001) summarises correlation, variance ratio and RMSE. For probabilistic forecasts of categoric events, that is the probability that a given threshold will be exceeded, the probability score or brier score (Brier, 1950) is commonly used in the case of binary events and the ranked probability score (RPS) (Epstein, 1969) in the multi-threshold case. Alternatively the Ignorance score (IGN) and its generalisations for the multi-threshold (ranked IGN) and continuous case (continuous ranked IGN) (Tödter and Ahrens, 2012) are sometimes used. Jolliffe et al. (2012) provides a detailed guide for verification in atmospheric science.

In the following selected methods for probabilistic forecasts of continuous variables and spatial verification methods are presented as these are the ones that are relevant for our purpose.

3.1 Continuous Ranked Probability Score (CRPS)

3.1.1 The CRPS as an error measure

The continuous ranked probability score (Brown, 1974) is a score for probabilistic forecasts of continuous variables. Unlike most of the other scores it does not require discrete probability categories which makes it particularly useful for this study where meaningful categories are not at hand. The CRPS measures the area between a simulated cumulative distribution function (cdf) and an observed one. Note that for a typical observation without uncertainty assigned the observed cdf has the form of a unit step function. The CRPS for a given timestep t is thus:

$$CRPS_t = \int_{-\infty}^{\infty} \left[F(x) - H_{x^0_{obs}}(x) \right]^2 dx \tag{3.1}$$

with F(x) being the simulated cdf and $H_{x^0_{obs}}(x)$ the Heaviside function with its step at the observed value x^0_{obs}. Matheson and Winkler (1976) proposed a discretised version of the CRPS for cases where the cdf is defined by discrete values as it is the case for ensemble predictions and for simulations from SANDHY.

As the quality of a probabilistic forecast can not be judged on a single event, the CRPS is the time average of the $CRPS_t$s over a number of time steps N:

$$CRPS = \frac{1}{N} \sum_{t=1}^{N} CRPS_t \tag{3.2}$$

The CRPS of a perfect forecast is 0. The units of the CRPS are the same as the ones of the predicted variable, in the case of precipitation for example [mm day^{-1}]. The CRPS can be interpreted as an integral over all possible Brier score values. It is identical to the mean absolue error (MAE) in the case of a deterministic forecast and thus has a clear interpretation in the deterministic case. Gneiting and Raftery (2007) proved that the CRPS is a proper score and a special case of the energy score, a very general type of scoring rule. A proper score is one that encourages the forecaster to issue forecasts corresponding to his best judgement, because the score take its ideal value only for ideal forecasts (Murphy, 1969).

3.1.2 Continuous Ranked Probability Skill Score (CRPSS)

To test the performance of a method compared to the performance of a reference method, skill scores are commonly used. Skill scores report the performance of a simulation with respect to a reference simulation thus answering the question if

and how much the simulation is better than the reference. The skill score of the CRPS is the continuous ranked probability skill score (CRPSS) and is defined as

$$CRPSS = 1 - \frac{CRPS}{CRPS_{ref}} \tag{3.3}$$

$CRPS_{ref}$ being the CRPS of a reference forecast. The CRPSS is 1 for a perfect forecast, 0 for a forecast that is equally skillful as the reference forecast and negative for a forecast less skillful than the reference.

The CRPSS depends on the predictive capability of the reference forecast. One can choose a simple (low effort) reference forecast like random forecast, persistence forecast or climatological forecast. The climatological forecast is an estimate of the marginal distribution of the predictand (Gneiting and Raftery, 2007). Alternatively a concurrent method may be used if one wants to stress the improvement over a particular concurrent method as has been done by Ben Daoud (2010) to measure the improvement achieved over the reference method from Bontron (2004). For this study the climatological forecast is used as reference forecast.

Gneiting and Raftery (2007) claims that skill scores of the type of the CRPSS are generally improper even if the underlying score is proper, but might be asymptotically proper for a high number of independent forecasts.

To rate the utility of a simulation, user specific costs and potential benefits from the forecasts have to be known. Assumptions about these highly user dependent numbers are often necessary, since they are frequently not even known by the user himself. Murphy (1966) proposed a simple cost-loss model based on cost/loss ratios and Murphy (1977) showed that the brier skill score is related to the utility of a simulation under the assumption that the cost/loss ratios of the users are uniformly distributed over the interval [0,1]. Since the CRPS can be interpreted as an integral over the brier scores for all possible thresholds Epstein (1969), Bontron (2004) and Laio and Tamea (2007) showed that the CRPSS can be interpreted as a measure of utility under the assumption of uniformly distributed cost/loss ratios as well.

3.1.3 CRPS of the climatological reference forecast

A climatological forecast means that on each day the predicted cdf is the cdf defined by the observations in the observation data base. The CRPS of the climatological forecast, called $CRPS_{clim}$ in the following, is equal to the uncertainty component in the decomposition of the CRPS by Hersbach (2000) in reliability, resolution and uncertainty. The decomposition will be explained and used in section 10.4.

Since there might be some sensitivity of the CRPS on the discretisation of the cdf, the climatological cdf is approximated with the same number of values as

provided by the simulation. Here, it is considered that if the simulation consists of N values the climatological cdf is approximated using N values as well.

There are different ways to define which observations to take into account when defining the climatological cdf. Essentially it comes down to the question of how to take seasonality into account. ±60 days around the target day are chosen as season for our CRPSS calculations in line with Bontron (2004) and Ben Daoud (2010) to render the CRPSS values comparable with theirs. Here it is explored what differences arise in the CRPS_{clim} if a shorter season (±15 days) or the whole year is used instead.

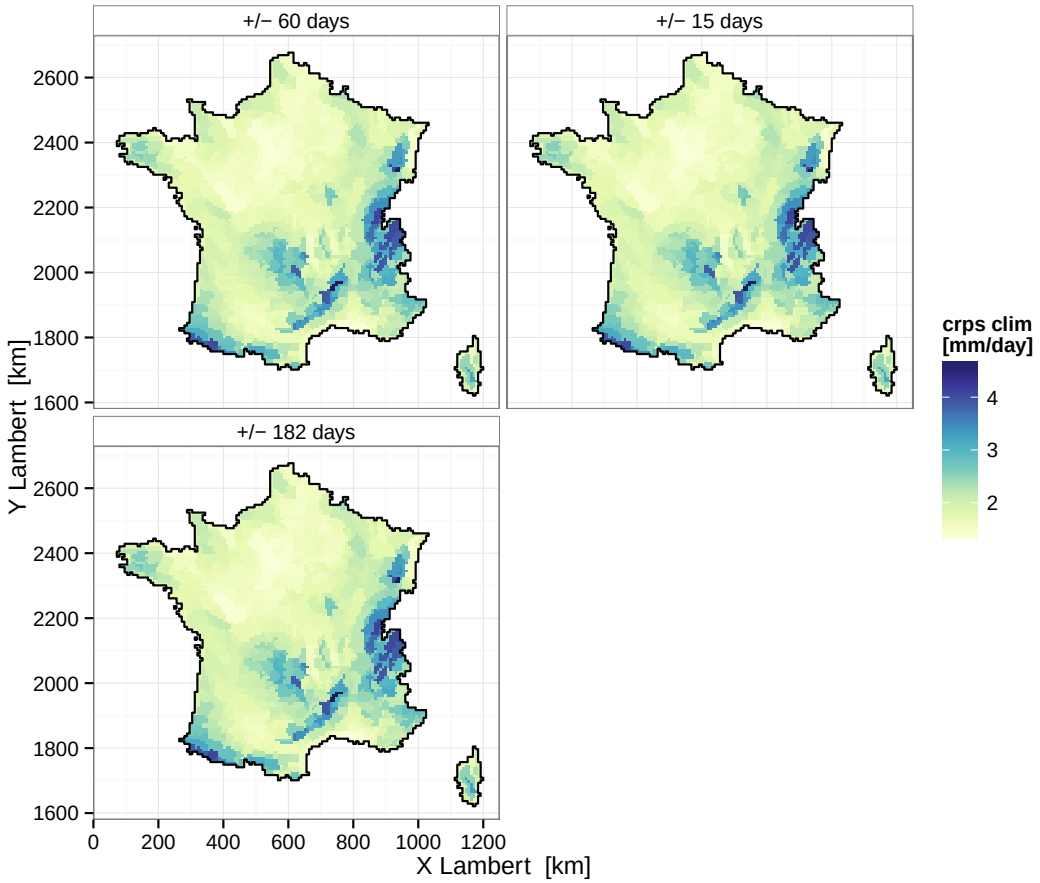

Figure 3.1: CRPS_{clim} calculated with top left: ± 60 days around the target day, top right: ± 15 days around the target day and bottom: ± 182 days around the target day.

Figure 3.1 shows the climatological CRPS calculated using ± 60 days around

the target day (from different years), ± 15 days or ± 182 days over the *late* period. The results are very similar. The difference is not visible here. The spatial pattern is very similar to what we saw in terms of mean precipitation in figure 2.3. Since we are not able to distinguish differences from the maps with absolute CRPS values, difference maps are shown in the following.

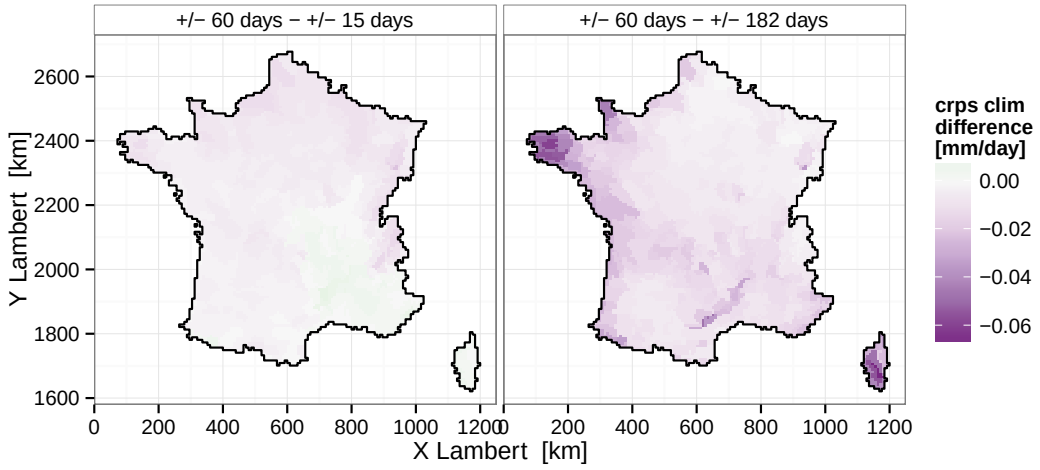

Figure 3.2: Difference between the $CRPS_{clim}$ calculated for left: ± 60 minus ± 15 days around the target day, right: ± 60 days around the target day minus ± 182 days around the target day.

Figure 3.2 shows the difference between the $CRPS_{clim}$ calculated using different lengths of season. The left map shows the difference between the ± 60 and the ± 15 days version. Here regional differences can be seen. The highest differences are found where the highest precipitation and climatological CRPS values can be found. The longer season (± 60 days) leads to smaller errors of the climatological forecast in the North and the shorter ± 15 day season is better in the southeastern part of the country. The right map shows the difference of the climatological CRPS calculated with ± 60 days around the target day and ± 182 days around the target day (corresponding to the whole year). The values obtained with ± 60 days around the target day are smaller than the ones for the whole year, the absolute values of the differences correspond to around 1% of the CRPS_{clim}. The highest differences are found in Brittany and Corsica.

Overall the ± 60 days around the target day lead to the smallest CRPS_{clim} of

the three tested versions and therefore represents the strongest reference forecast, that is the one that is the hardest to beat. In the southeastern part of the country the ± 15 days version would be even more challenging.

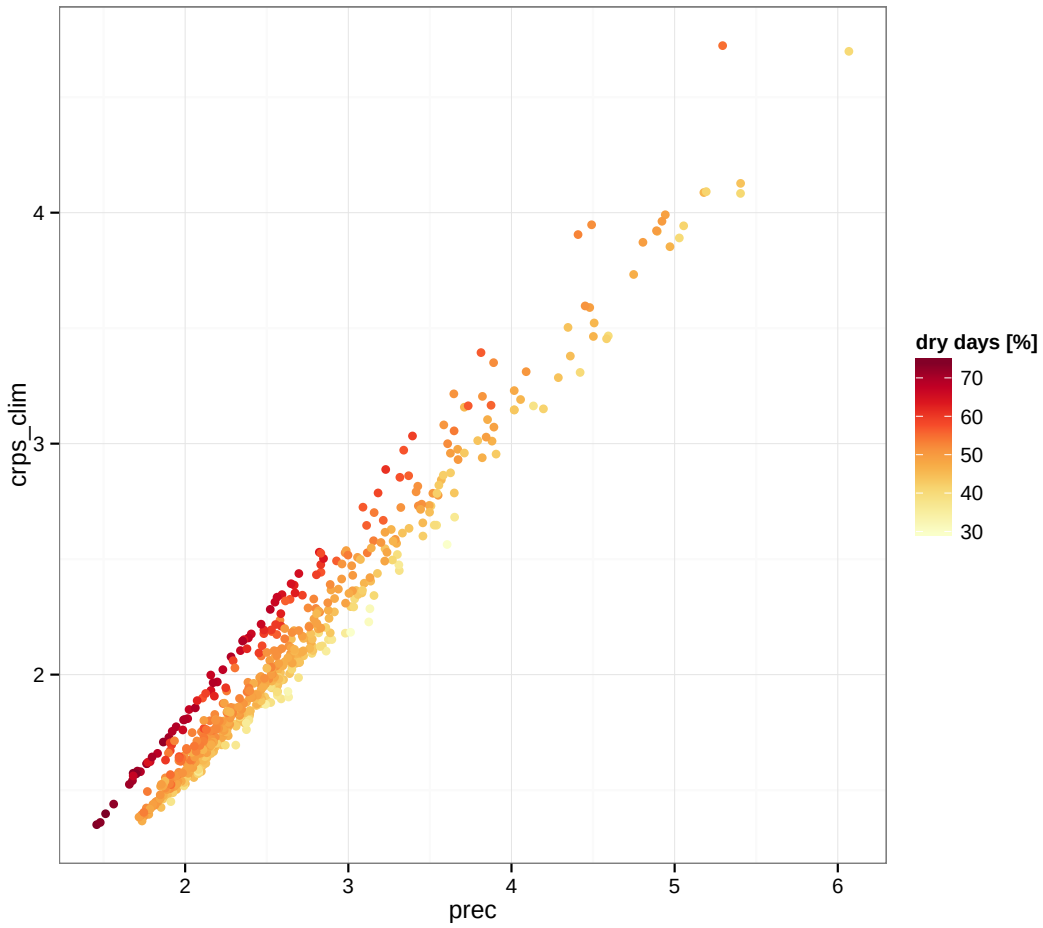

Figure 3.3: Mean daily precipitation [mm day^{-1}] vs. climatological CRPS [mm day^{-1}] calculated over the *late* period with 60 days around the target day for each symposium zone. Colours correspond to the percentage of dry days.

In figure 3.1 we saw that the spatial pattern of the CRPS$_{clim}$ is similar to the one of average precipitation. Figure 3.3 shows a scatter plot between these two variables and there is indeed a strong linear relationship between the mean daily precipitation and the climatological CRPS. In regions with large mean precipitation amounts the climatological CRPS is higher then for the others. Where the climatological CRPS is high there is more "room for improvement" which in general leads to higher CRPSS. Furthermore figure 3.3 shows that for zones with a

high percentage of dry days the climatological CRPS is higher than for zones with a low percentage of dry days. In particular for low precipitation amounts the dots representing zones with a large number of dry days cluster around a different line than the ones with a smaller number of dry days. The zones with a high percentage of dry days can be found in the Southeast of the country (see figure 2.4) and have a Mediterranean climate.

3.2 Rank histograms

The rank histogram or Talagrand diagram (Talagrand et al., 1997) tests the reliability of an ensemble-type forecast. It counts where the observation ranks with respect to the ensemble members. For a perfectly reliable ensemble each of the m+1 ranks, m being the number of ensemble members, has the same frequency of occurrence and thus the resulting histogram is flat. Biased forecasts lead to overpopulation of either low ranks for a high bias or high ranks for a low bias. Underdispersive ensembles, that is forecasts with too low spread, lead to overpopulation of the extreme ranks while overdispersive ones lead to peaked rank histograms. In the case of precipitation, ties often occur for the ranks of zero precipitation. Hamill and Colucci (1998) uses small random perturbations of the zero precipitation forecasts and observation to determine the ranks. This is necessary to have equal probability of the observation falling in each of the tied ranks. This correction is implemented in the R package ensembleBMA (Fraley et al., 2013) that will be used for obtaining the rank histograms in chapter 10.

3.3 Spatial verification methods

Precipitation behaves quite differently on different spatial and temporal scales and the smaller the scales the more complicated it is to predict. Numerical models typically parametrise subgridscale processes, among them convective processes that are important for local scale precipitation. With grid scales of a few kilometers and non-hydrostatic models with more explicit treatment of deep convection that became available during the last decade it became possible to explicitly simulate storms (Nuissier et al., 2008). These more realistic precipitation fields produced by the new generation of forecast models presented a challenge for verification, because they are less smooth than before and a gridpointwise verification thus leads to poorer scores for small displacement errors due to the double penalty problem. To make results from models with different resolutions more comparable, a number of spatial verification techniques have been developed. Gilleland et al. (2009) performed a comparison of spatial verification techniques using synthetic and real

cases to evaluate which kind of errors are detected by the different methods. Most of the tested methods are for example sensitive to displacement errors but not every method actually quantifies them. There are not so many methods that account for structure errors. Apart from intensity errors that nearly all methods including traditional ones can measure, displacement- and structure errors in precipitation are relevant for hydrological modelling. In chapter 10 a spatial verification technique will be used to assess displacement and structure errors in downscaled precipitation data. In the following an overview of spatial verification methods is given, looking for methods that (1) can be applied to downscaled precipitation, (2) can be adapted for probabilistic simulations and (3) quantify types of errors that are of interest for hydrological modelling.

3.3.1 Scale decomposition, neighbourhood and fuzzy methods

Scale decomposition methods aim at separating different spatial scales of the forecast and its errors and thus determine the skillful scales of the forecast. Examples are the intensity-scale verification approach (Casati et al., 2004), wavelet decomposition (Briggs and Levine, 1997) and discrete cosine transform (Denis et al., 2002).

Neighbourhood or fuzzy methods allow for some uncertainty of the forecast in space and time by evaluating the forecasts within a spatial (or temporal) window around the observation. Similar to scale decomposition methods a skillful scale can be determined by using several window sizes. Forecasts for grid points inside the window may be interpreted as a probabilistic forecast for an observation point. In contrast to scale decomposition methods neighbourhood methods successively smooth the fields. Examples for neighbourhood or fuzzy methods are multi-scale statistical organization (Zepeda-Arce et al., 2000) and fraction skill score (Roberts and Lean, 2008; Duc et al., 2011). An overview of fuzzy methods can be found in Ebert (2008).

Scale decomposition methods and neighbourhood methods are powerful in determining skillful scales for example, but they do not treat displacement and structure errors explicitly. In a downscaling study where we already know the scale at which we want to assess the skill and the spatial features of the downscaled fields, these methods will not tell us what we want to know.

3.3.2 Object based methods

The second group of spatial verification methods are object based methods. Common to these methods is that precipitation objects are defined in the observed- and

the simulated precipitation fields typically using a threshold, and then properties of these objects are compared. Some, but not all methods try to match objects, that is they explicitly try to find corresponding objects in the observation and the simulation. This is definitly useful for verifying high resolution numerical weather predictions but may complicate the adaptation to probabilistic simulations.

The oldest object based method is the continous rain areas (CRA) method (Ebert and McBride, 2000). CRA decomposes the mean squared error (MSE) into a displacement, a volume and a pattern error while the pattern error is the residual of the other two. The method is based on the total MSE over the verification domain, that is the area average of the MSE values at each grid point. CRA could be adapted for probabilistic forecasts by replacing the MSE with a probabilistic score, for example the CRPS or the Brier score. A drawback of CRA in the downscaling context is the requirement of matching individual precipitation objects.

MODE (Davis et al., 2006, 2009) describes intensity, area, centroid, axis angle, aspect ratio and curvature of precipitation objects. With MODE one can chose to match objects or to compare the climatologies of their attributes. The calculation of the attributes could be adapted for probabilistic forecasts. The large number of attributes that can – but not have to – be calculated and that are combined in a total interest value makes the method very flexible. However, the final score strongly depends on the weighting of the components which complicates its interpretation.

Morphing methods measure the modification that are needed to match the observation with the forecast or the forecast with the observation. Examples are the procrust shape analysis method (Micheas et al., 2007), and the optical flow method used for the displacement and amplitude score DAS (Keil and Craig, 2007, 2009). Morphing methods seem tricky to adapt for probabilistic simulations.

3.3.3 Structure, Amplitude, Location (SAL)

SAL is an object based spatial verification technique and will be described in more detail because it is used in this study and adapted for probabilistic simulations in chapter 10.

Wernli et al. (2008, 2009) introduced SAL (Structure, Amplitude, Location) that is designed to evaluate these three properties of the precipitation forecast over a given domain, for example a specific catchment. SAL is implemented in the R package spatialVx (Gilleland, 2013). SAL has been used since in high resolution numerical weather forecast model development (e.g. Termonia et al., 2011; Haiden et al., 2011; Zappa et al., 2010).

3.3.3.1 The SAL components

The amplitude component is the normalised difference of the domain averaged precipitation values:

$$A = \frac{\widehat{R_{fc}} - \widehat{R_{obs}}}{0.5[\widehat{R_{fc}} + \widehat{R_{obs}}]} \tag{3.4}$$

where R denotes the precipitation value and $\widehat{R}$ the spatial average of precipitation over the verification domain. A ranges from -2 to +2, taking negative values for too little precipitation in the forecast and positive values for too much precipitation in the forecast. The perfect forecast has an amplitude error of 0.

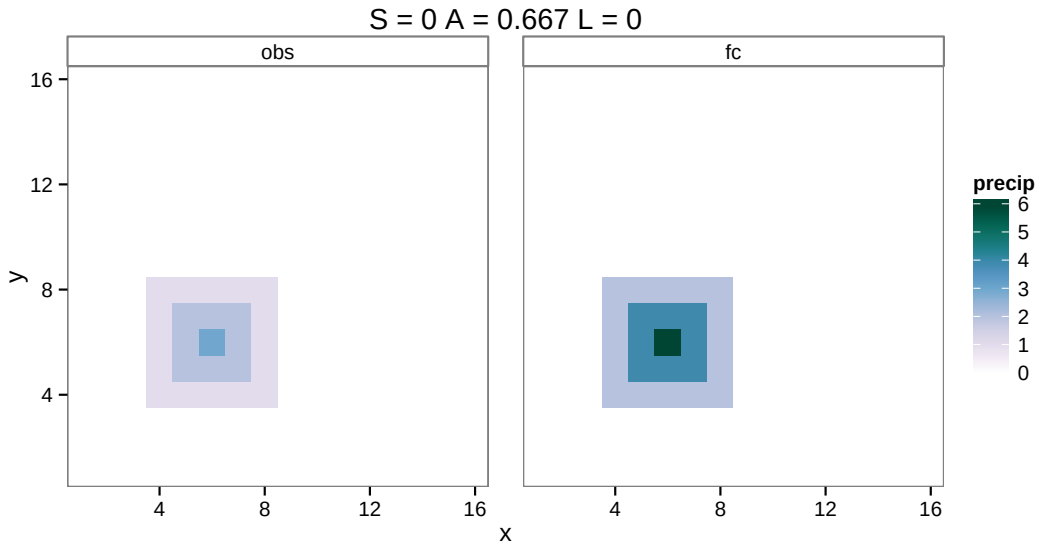

Figure 3.4: Synthetic example of a pure amplitude error. Left: observation, right: simulation.

Figure 3.4 shows a synthetic example of a pure amplitude error. The precipitation values in the forecast (fc) (right hand side) are twice as high as the ones observed (obs) (left hand side). This leads to a positive amplitude error of 0.667 and no error in the structure and location components.

For the S and the L components precipitation objects have to be defined in the domain. Objects are defined as areas where a given threshold is exceeded. Wernli et al. (2009) defined this threshold as:

$$f = \frac{1}{15} R^{95} \tag{3.5}$$

R^{95} being the 95th percentile value of all nonzero rainfall in the domain. With this threshold it is attempted to approximate what a human forecaster would define as precipitation object. A limitation of the SAL method is that S and L can be calculated only if there is at least 1 object in the observation and in the simulated field.

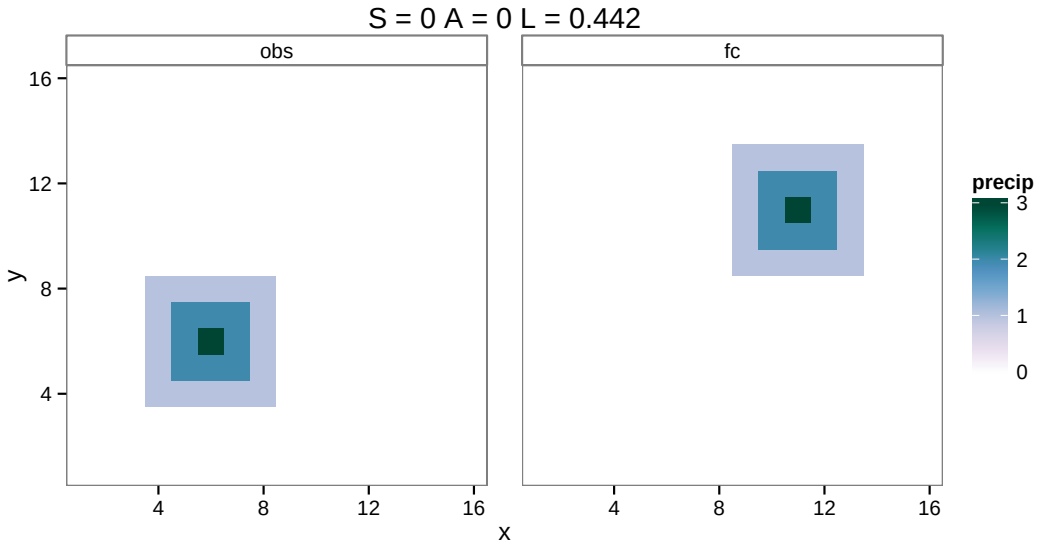

Figure 3.5: Synthetic example of a pure location error.

The location component L consists of two parts: $L = L_1 + L_2$. L_1 is the normalised distance between the centers of mass of obs and fc.

$$L_1 = \frac{|\overrightarrow{x}(fc) - \overrightarrow{x}(obs)|}{d} \tag{3.6}$$

where $\overrightarrow{x}$ is the location vector of the center of mass of the objects and d is the largest distance between two boundary points of the domain. L_2 measures the weighted average distance of the precipitation objects to the center of mass and gives the difference of this distance between observation and forecast. L_2 can only differ from 0 if there are at least two precipitation objects in one of the fields (obs or fc).

$$L_2 = 2\left[\frac{|r(fc) - r(obs)|}{d}\right] \tag{3.7}$$

where r is the weigthed average distance of the precipitation objects to the center of mass. The perfect value for L_1 and L_2 is zero, the maximum value is 1 for each of the components adding up to a maximum of 2 for the entire location component L.

In figure 3.5 the forecasted precipitation object is shifted compared to the observed one. This leads to a location error, more precisely an error of type L_1. The other two components are zero.

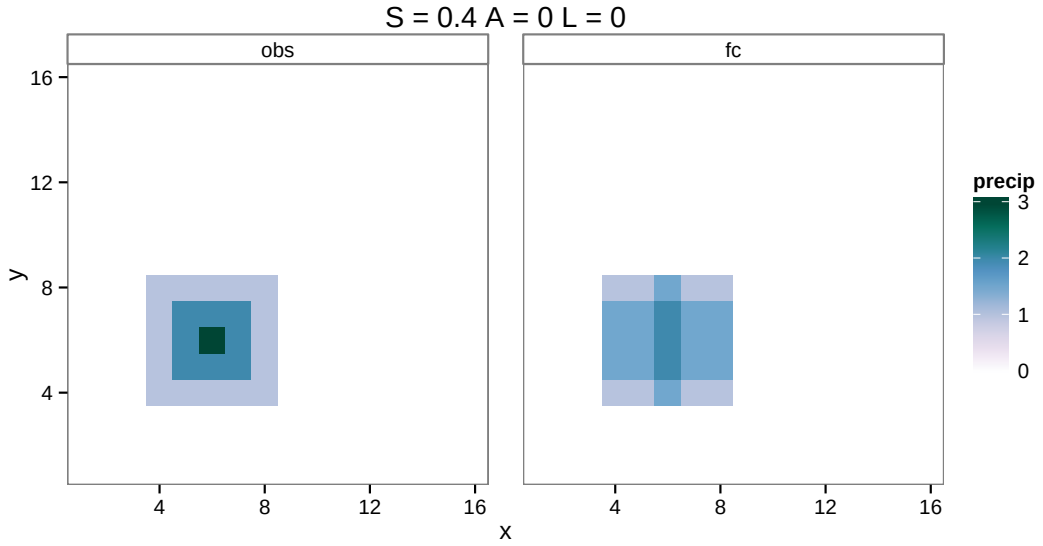

Figure 3.6: Synthetic example of a pure structure error.

The structure component S compares the volume of normalised precipitation objects. It contains information about the size and the shape of precipitation objects. S is the normalised difference between the mean scaled volumes of the precipitation objects in the forecast and the observation:

$$S = \frac{V(fc) - V(obs)}{0.5[V(fc) + V(obs)]} \tag{3.8}$$

where

$$V = \frac{\sum_{n=1}^{M} R_n V_n}{\sum_{n=1}^{M} R_n} \tag{3.9}$$

and

$$V_n = \sum_{i,j} \frac{R_{ij}}{R_n^{max}} \tag{3.10}$$

where n is an individual precipitation object, R_n its precipitation, V_n its scaled volume and M the total number of objects. S ranges between -2 and 2 where too

small or too peaked objects in the forecast lead to negative structure errors and too large or too flat objects to positive structure error. Forecasts with a perfect structure have a structure error of 0. In figure 3.6 the forecasted object is less peaked than the observed one, thus leading to a posititive structure error. The other two components are zero in this example.

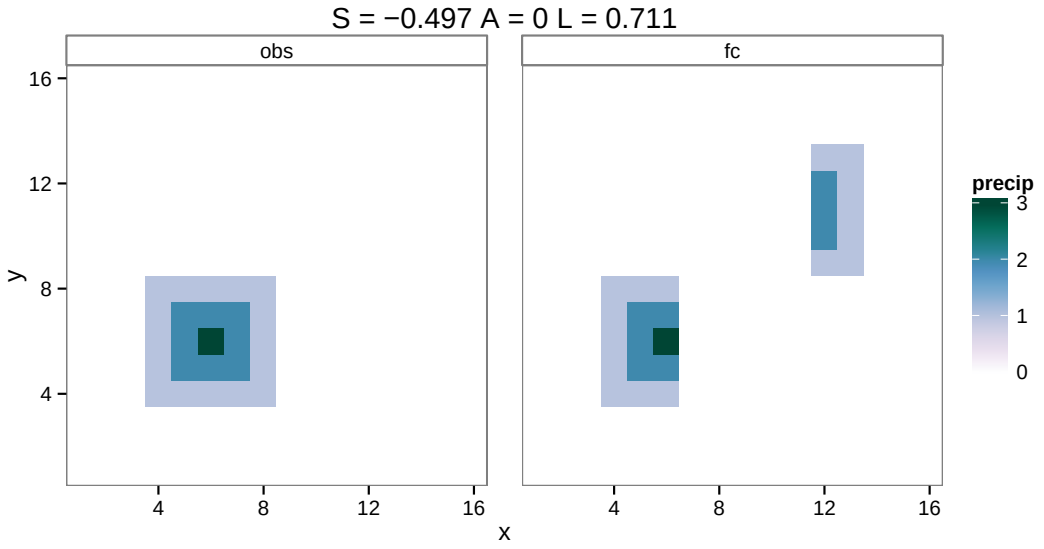

Figure 3.7: Synthetic example with two small objects instead of one larger object observed.

Figure 3.7 shows a synthetic example of a combined structure and location error. The forecast shows two smaller objects compared to one larger object observed. This leads to a negative structure error of -0.497 and even if one of the two forecasted objects being located very close to the observed one there is a location error of 0.711 stemming from a L_1 and a L_2 error. The amplitude error A is 0 because the domain average precipitation is identical.

3.3.3.2 Limitations of SAL

SAL is insensitive to rotation errors as shown in figure 3.8. All SAL components are 0 in the case of a simulated precipitation object that is rotated with respect to the observed object. Furthermore SAL does not assess missed events and false alarms, they have to be considered separately when comparing model performances.

Concerning the relevance of the SAL for hydrological modelling which is the purpose of the precipitation downscaling conducted in this study, Wernli et al. (2008) notes that errors in the A component describe the overall bias of the precip-

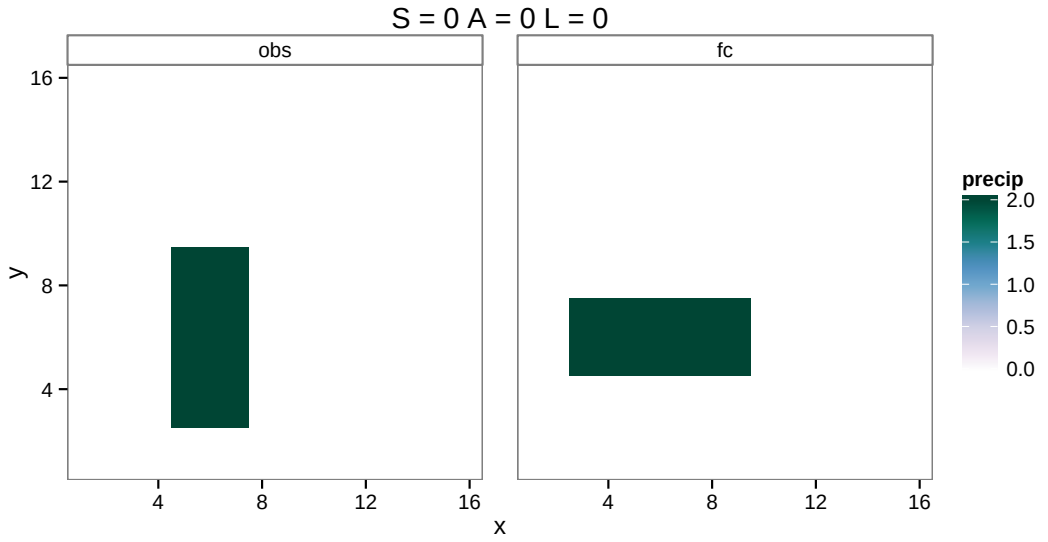

Figure 3.8: Synthetic example with a rotated precipitation object.

itation input and therefore can result in runoff biases. The L component describes how accurately the precipitation is distributed over subcatchments, so errors in the L component are supposed to lead to random errors in the runoff simulation. Errors in the spatial structure of precipitation as expressed through the S component are supposed to affect the repartition between surface runoff and infiltration and therefore affect the soil water balance. This together with the explicit calculation of structure and location errors makes SAL interesting in the context of downscaling for hydrological modelling.

3.3.3.3 SAL diagram

Wernli et al. (2008) proposes a SAL diagram for a climatological analysis and comparison of different models. For every wet day that is not a missed event, a point is drawn in the SAL diagram where the S component is mapped on the x-axis, the A component on the y-axis and the L component is represented by the colour of the dots. So a good model will have many light colored points near the center of the diagram. An example of a SAL diagram is shown in figure 3.9. A potential weakness of this diagram is that it does not distinguish between very light precipitation events and intense precipitation events. For very light precipitation events points are likely to appear near the upper right corner of the diagram, in the case somewhat more precipitation has been forecasted. Due to the small observed value the relative difference becomes large and the forecasted objects are

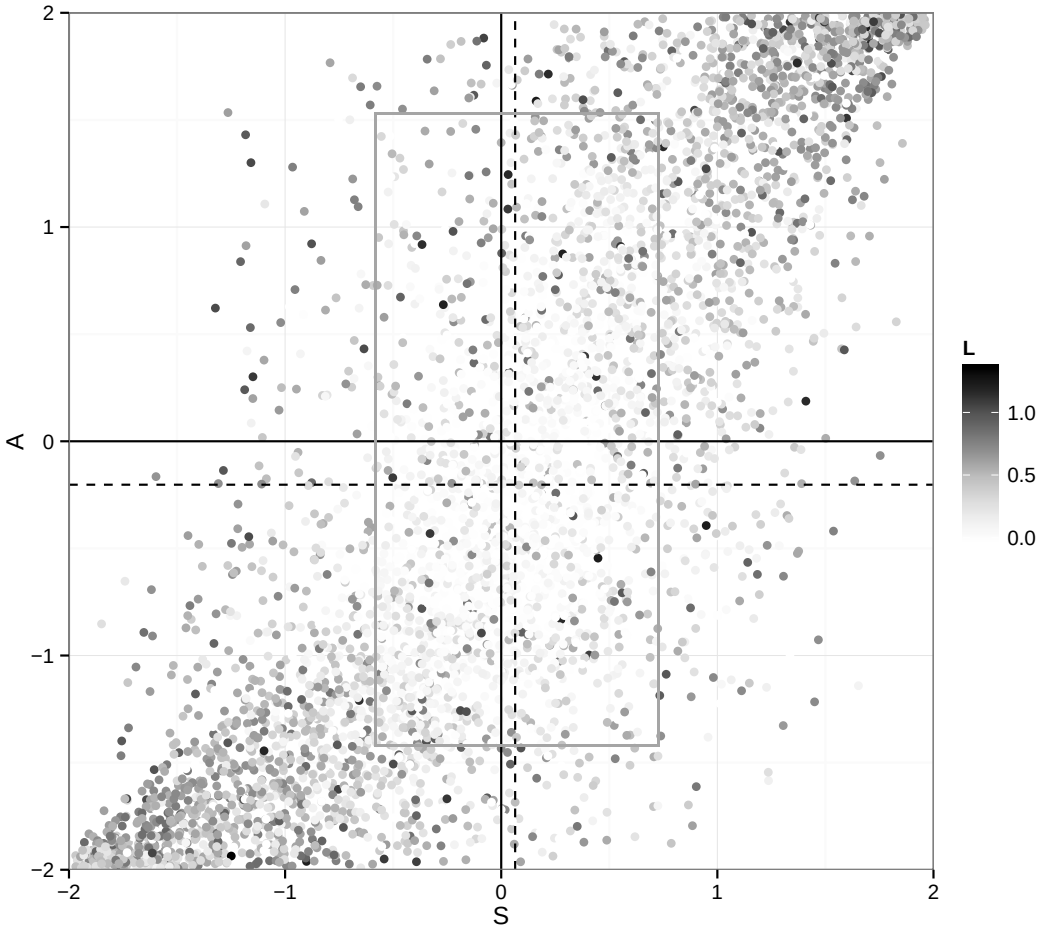

Figure 3.9: SAL diagram example

larger leading to large S components. Cases where light precipitation events are underestimated typically don't appear on the diagram because this often means that no rain has been forecasted and no dot is drawn on the diagram. Points in the lower left corner are typically events where an important precipitation event is underestimated. This means that "good" models don't produce SAL diagrams that appear symetric around the center point while random forecasts and persistence forecasts do, as shown in Wernli et al. (2008). Wernli et al. (2008) noted further that it seems intrinsic to the SAL method that points tend to appear in the first and third quadrants of the SAL diagram, since it is difficult to overestimate the precipitation amount with too small objects and underestimate it with too large ones.

To compare models Wernli et al. (2008) suggested to calculate the radius r of

a sphere in the three dimensional SAL space that contains a fixed percentage of best forecasts:

$$r = \sqrt{S^2 + A^2 + L^2} \tag{3.11}$$

3.4 Performance criteria for streamflow

Two widely used citeria are employed to assess the performance of streamflow simulations in Chapter 11.

Probabilistic Nash

A probabilistic version of the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) has been proposed by Bulygina et al. (2009). It is defined as:

$$Nash^{prob} = 1 - \frac{\sum_{t=0}^{T} Var[\xi_t] - (E[\xi_t] - q_t^0)^2}{\sum_{t=0}^{T} (E[q_t^0] - q_t^0)^2} \tag{3.12}$$

where q_t^0 is the observed streamflow at time t, ξ_t the modelled streamflow at time t, T the number of timesteps, $E[\bullet]$ the expectation and $Var[\bullet]$ the variance of the modelled streamflow ensemble.

Kling-Gupta Efficiency (KGE)

The Kling-Gupta Efficiency (Gupta et al., 2009) with corrected standard deviation ratio (Kling et al., 2012) is used. It is defined as:

$$KGE = 1 - \sqrt{(r-1)^2 + (\gamma - 1)^2 + (\beta - 1)^2} \tag{3.13}$$

Where r is the correlation coefficient between the simulation and the reference, γ is the bias corrected standard deviation ratio

$$\gamma = \frac{\frac{\sigma^{sim}}{\mu^{sim}}}{\frac{\sigma^{obs}}{\mu^{obs}}} \tag{3.14}$$

where μ is the mean and σ the standard deviation, and β the ratio of the means.

$$\beta = \frac{\mu^{sim}}{\mu^{obs}} \tag{3.15}$$

Critically, the components of the KGE, β, γ and r can be used separately to diagnose different model behaviours.

With the diagnostic tools discussed in this chapter at hand we start downscaling over France in the next part.

Part II

Downscaling over France

4 Stepwise Analogue Downscaling method for Hydrology (SANDHY)

In this chapter the downscaling method used throughout this work, SANDHY, is presented. First we will have a look at the development history of SANDHY, then the chosen predictors and possible treatment of predictors are discussed and finally the implementation and the technical details are presented.

4.1 History

Analogue methods were originally developed for short term weather forecasting and so was SANDHY. Guilbaud and Obled (1998a,b) described an analogue method for daily quantitative precipitation forecast and tested principal components from the gridded 700 hPa geopotential field, raw radiosonde data and gridded radiosonde data at two pressure levels as predictors. Euclidean distance is used as similarity criterion in the first two cases and the Teweles and Wobus shape criteria (TWS) (Teweles and Wobus, 1954), which measures the similarities between the zonal- and the meridional gradients, in the third case. It was found that gridded data with the TWS criterion perform best. Temperature and humidity fields did not lead to improvements at this stage but later Gibergans-Báguena and Llasat (2007) found improvements adding predictors related to thermodynamics.

Obled et al. (2002) tested different numbers of analogues retained and found a flat optimum around the value of 50. Bontron (2004) and Ben Daoud (2010) studied how many analogue dates should be retained in the end as well and found that 25-35 analogues lead to the best results. For low precipitation values a larger number of analogues leads to better defined distributions but for high precipitation events taking less analogue dates should be preferred, because for more rare events not enough good analogues are available (Hamill and Whitaker, 2006). Furthermore Obled et al. (2002) noted that a single predictor domain for all French catchments is probably not optimal and some splitting should be considered.

Bontron et al. (2002) introduced the use of reanalysis data (NCEP/NCAR reanalysis (Kalnay et al., 1996)) instead of interpolated radiosonde data. This version has since been used in various studies (e.g. Auffray et al., 2011; Obled et al., 2009).

Bontron (2004) and Bontron and Obled (2005) used geopotential heights at two pressure levels, 1000 hPa and 500 hPa, and introduced an additional humidity predictor. This is in line with Martin et al. (1997) who found humidity and variables related to the general circulation dynamics to perform best as predictors. Fowler et al. (2007) adds that circulation predictors alone are unlikely to be sufficient, because they fail to capture important processes related to thermodynamics and moisture content and recommends geopotential heights and specific humidity as predictors. The version from Bontron and Obled (2005) is used for example in Horton et al. (2012); Marty et al. (2012, 2013); Chardon et al. (2014).

Ben Daoud et al. (2009a,b) compared the use of ERA-40 and NCEP/NCAR reanalysis data in the analogue downscaling and found slightly better results with ERA-40.

Ben Daoud (2010) and Ben Daoud et al. (2011a,b) replaced the restriction of the analogue search to the current four month moving season with a selection on temperature. The four month season is a rather large window compared to for example Martin et al. (1997) who used only one month, but has the advantage of providing a larger pool of potential analogues. In addition vertical velocity is added as an additional predictor that leads to lower false alarm rates for the first forecast days. This method with its four predictor variables is used in this work and is called SANDHY since 2013.

SANDHY performs a four step analogue selection using temperature, geopotential heights, vertical velocity and humidity as predictor variables to identify analogue dates in the archive as illustrated in Figure 4.1. The predictor variables, levels, times and similarity criterion for each step are as identified by Ben Daoud (2010) for the Seine and Saône river basins and are summarised in Table 4.1. The precipitation value for day D corresponds to the precipitation accumulated between 06:00 UTC day D and 06:00 UTC day D+1. The first step is a selection on temperature at 925 hPa at 12:00 UTC day D+1 and 600 hPa at 12:00 day D. The similarity criterion is the Euclidean distance with equal weights for the two pressure levels.

The second step is a selection on geopotential at 1000 hPa at 12:00 UTC day D and 500 hPa at 00:00 UTC day D+1. The similarity criterion used is the TWS. Again equal weights are given for the two pressure levels. The same predictor domain is used for the two pressure levels and is optimised in Chapter 5.

The third step is a selection on vertical velocity at 850 hPa at 06:00, 12:00 and 18:00 day D and 00:00 day D+1. The similarity criterion is the Euclidean distance

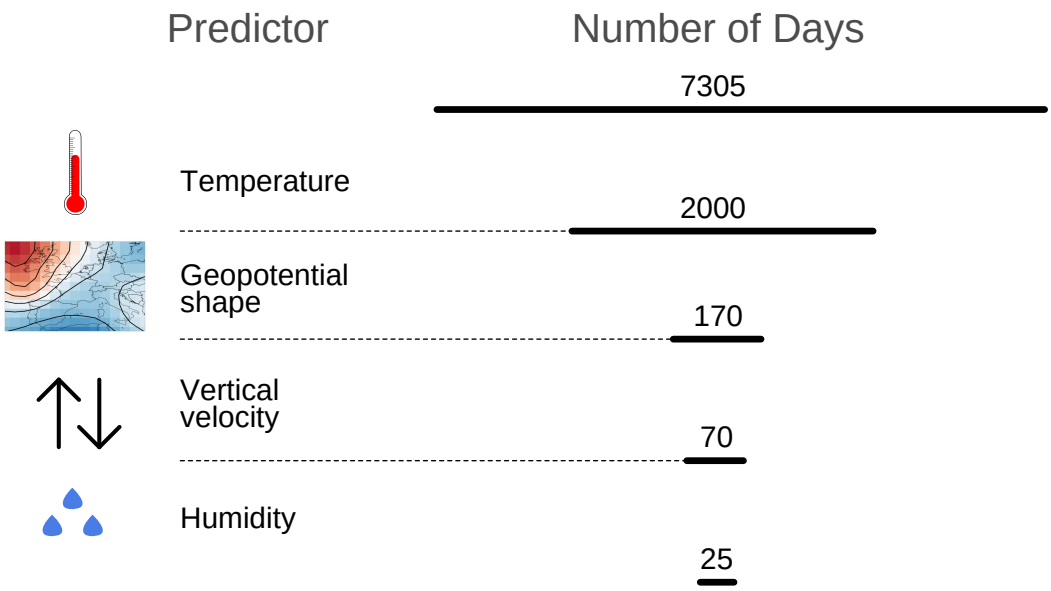

Figure 4.1: Analogue selection steps.

and equal weights are given to the different times.

The fourth step is a selection on humidity, more precisely the product of the total column water (TCW) and relative humidity at 850 hPa (RH) at 12:00 day D and 00:00 day D+1. The similarity criterion is the Euclidean distance.

Identical combinations of variables, pressure levels and hours in steps 2 and 4 are already used in Bontron and Obled (2005) for an application at various locations in south-eastern France.

4.2 Predictors and Predictands

Predictors should be informative, that is having predictive power, physically meaningful, reasonably well simulated by the climate models and capture the large scale climate change signal and climatic variability on a range of timescales (Maraun et al., 2010; Fowler et al., 2007; Schoof, 2013). The requirement that predictor variables should be well simulated by the GCMs is especially important for perfect prognosis (PP) approaches because well simulated predictors is one of their assumptions. Common to most, if not all, downscaling methods is the assumption that the predictor-predictand relationship is stationary (Zorita and von Storch,

Predictor	**Pressure level [hPa] and Time [UTC]**	**Similarity citerion**
Temperature	925 at 12:00 D+1, 600 at 12:00 D	E.d.
Geopotential	1000 at 12:00 D, 500 at 00:00 D+1	TWS
Vertical velocity	850 at 06:00, 12:00, 18:00 D and 00:00 D+1	E.d.
Humidity (TCW×rh)	850 at 12:00 D and 00:00 D+1	E.d.

Table 4.1: Predictors and their corresponding pressure levels, times and similarity criterion for each of the four steps of SANDHY. E.d.: Euclidean distance, TWS: The Teweles and Wobus shape criteria (Teweles and Wobus, 1954) (Radanovics et al., 2013a)

1999), which is not guaranteed but more likely for physically meaningful predictors (Ribalaygua et al., 2013). Fowler et al. (2007) notes that the choice of the predictor variables for precipitation is as important as the choice of the downscaling method.

4.2.1 Predictor selection

The choice of the predictor variables should therefore be given high consideration (Fowler et al., 2007). Wilby and Wigley (2000) examined predictors and predictor domains for precipitation downscaling analysing correlations and concluded that the choice of the predictor and the predictor domain are critical factors affecting the realism and stability of downscaled precipitation scenarios. Bontron (2004) and Fernández-Ferrero et al. (2009) found that taking dynamical predictors over a wide region around the studied area but thermal and moisture variables only in a small neighbourhood gives the best result.

The most common methods for selecting predictor variables are correlation analysis (e.g. Frost et al., 2011), expert judgement or a combination of both (e.g. Khan et al., 2006). Ribalaygua et al. (2013) recommends to select the predictors based on theoretical considerations rather than emprirical analysis to avoid non-physical relationships that may not be stationary. Khan et al. (2006) use a sensitivity analysis to select the predictors for an ANN downscaling model. A less common selection method are relevance maps (e.g. Bontron, 2004), that are described in Section 5.1.

4.2.2 Predictors of the SANDHY method

As we have seen in section 4.1, SANDHY predictors are the result of a long development history. The predictor variables and their particular sequence have been selected in Ben Daoud (2010) and Bontron (2004) who tested many different variables, different similarity measures and predictor domains sizes. The following paragraphs highlight some reasons for using temperature, geopotential, vertical velocity and humidity as predictor variables as well as alternative possible predictors.

Using a temperature predictor rather than a restriction to seasons has the advantage that a situation that is extreme in one season may be rather common in another season and therefore it is possible to find better analogues in such situations. Furthermore, if we think of applications in a climate change context predictors representing the climate change signal are needed (Schoof, 2013). A method that uses a temperature predictor may account for a shift in seasons (except for extreme seasons) which is not possible using fixed seasons (Ribalaygua et al., 2013). The drawback of not using seasons is that it can not be distinguished between spring and autumn precipitation events that occur at the same air temperature but do not at all lead to similar amounts of precipitation in certain regions. Large climatological differences between spring and autumn precipitation are due to differences in sea-surface temperature that play a key role for moisture supply in these regions. An alternative would be to use sea surface temperature as an additional predictor.

Circulation related predictors are commonly used in statistical downscaling of precipitation but various ones may be used. Mean sea level pressure (mslp) (e.g. Wetterhall et al., 2005; Timbal, 2004), geopotential heights, geostrophic wind and vorticity are the most common ones. Ribalaygua et al. (2013) recommends to use upper air fields rather than boundary layer variables and point values arguing that they are more reliably simulated by GCMs. Lavers et al. (2013) argues that mslp is related to the vertical velocity in the mid-troposphere and therefore related to cyclonic development and precipitation and should therefore be preferred. In SANDHY geopotential height fields are used, but with a similarity criterion that is sensitive to gradients of these fields. The gradients of the geopotential fields are related to the geostrophic wind (compare Section 5.3).

Though closely related to precipitation generation, vertical velocity is not a very common predictor in precipitation downscaling. Perhaps because it is not as reliably simulated by GCMs as geopotential fields for example. Nevertheless it proved to be useful in a forecasting context (Ben Daoud et al., 2011a) and is therefore included in SANDHY.

It is common sense that circulation and temperature predictors alone are not enough to capture future precipitation changes and that a humidity predictor

should be included (Wilby and Wigley, 1997; Nicholas and Battisti, 2012). However, there is some debate about which humidity predictor this should be. Some argue that relative humidity is not suited, because it does not carry the climate change signal and specific humidity should be used, while others say that specific humidity is likely to exaggerate the climate signal due to the non-linear relation to temperature in the clausius-clapeyron equation and relative humidity should be used (Timbal et al., 2008; Ning et al., 2012). It can be argued that if a temperature predictor is already present, the humidity predictor does not need to carry a climate change signal. Relative humidity better describes the occurrence of precipitation, while amounts are more related to specific humidity (Yang et al., 2010). Some go even further, saying the moisture flux or moisture flux convergence should be used as predictor (Schoof, 2013; Teutschbein et al., 2011). Yang et al. (2010) for example uses specific humidity and moisture flux. Hamill and Whitaker (2006); Timbal (2004) use column precipitable water for downscaling with analogues. In SANDHY the product of relative humidity at 850 hPa and the total column water is used, which has been found to be the best performing humidity variable by Bontron (2004) and combines a relative and an absolute measure of humidity.

The chosen predictor variables are similar to the ones a human forecaster tends to look at to forecast non-convective precipitation in the mid-latitudes. The predictor variables are therefore considered to be quite robust for the mid-latitudes and the winter season and less adapted for the summer season and tropical and sub-tropical climates (Vidal and Radanovics, 2014).

4.2.3 Treatment of Predictor data

In SANDHY the predictor fields are taken as they are, but this is not always the case in statistical downscaling. Some use principal component analysis (PCA) of single predictors, for atmospheric fields usually termed empirical orthogonal functions (EOF) (e.g. Zorita and von Storch, 1999; Timbal and McAvaney, 2001; Trigo and Palutikof, 2001; Benestad, 2004; Matulla et al., 2008; Themeßl et al., 2010; Nicholas and Battisti, 2012; Schmith, 2008; Friederichs and Hense, 2007), or all predictor variables together (e.g. Gangopadhyay et al., 2005; Rust et al., 2010) to reduce the number of dimensions and/or potential redundancy between predictor variables. Timbal and McAvaney (2001) found principal components to be useful when large predictor domains are used, but for smaller predictor domains raw fields lead to better results. Van den Dool (1994) argued that the restriction to smaller spatial domains or a few principal components allow to find reasonably similar analogues within a historical record of limited length. Ribalaygua et al. (2013) advised against the use of principal components in order to make use of the full range of data variability. Cannon (2007) criticises that principal components are calculated without reference to the predictands. Fernández and Sáenz (2003) uses

canonical correlation analysis (CCA) to reduce the dimensionality of the predictor space.

Sometimes predictors are standardised (e.g. Martin et al., 1997; Teng et al., 2012) in order to give the same weight to predictor variables with different magnitudes and variations. For example Carreau and Vrac (2011) performs a PCA on all predictor variables together and then standardises the retained principal components. Gutiérrez et al. (2004) does it the other way round: the predictor variables are standardised prior to the PCA.

Biases of the GCM with respect to the reanalysis the PP downscaling relationships have been built on are an issue when GCMs are downscaled. One option is to bias-correct the GCM simulated predictors prior to downscaling (e.g. Timbal and McAvaney, 2001; Chiew et al., 2010; Frost et al., 2011). Martin et al. (1997) found that bias correcting the predictor dataset prior to downscaling with the analogue method improved the results. A second issue is that GCM fields are available on different grids than the reanalysis and thus are often regridded (e.g. Frost et al., 2011).

4.3 Implementation

This section concerns the third objective of this thesis which is to build a downscaling software that can be used in future research projects. The aim was to integrate functionalities for predictor domain optimisation, simulation and validation in a single program.

The SANDHY software is written in Fortran 2003 and the main loops are parallelised using openMP. Input and output files are in NetCDF format except the configuration file which is a namelist file. SANDHY works for a spatial set of predictand series.

The main implemented tasks are:

- Optimisation of geopotential predictor domains with the multiple growing rectangular domains algorithm and the CRPS as objective function. The number of domains can be selected by the user. Results are presented in Chapter 5.
- Calculation of relevance maps (Chapter 5).
- Analogue downscaling simulation.
- Validation calculating CRPS and Bias (Chapter 6).
- Aggregation of predictand areas that can use the same predictor domains with three different methods that will be detailed in Chapter 8.

- Calculation of the CRPS_{clim} for given observations.
- Calculation of the percentage of common analogue dates between zones or stations.

The configuration file allows to select which tasks should be performed and to adjust some parameters related to these tasks. Information on the spatial- and temporal dimensions of the predictor- and predictand input files, their names and paths to read and write files have to be specified as well. The time periods to be actually used as archive or for simulation can be selected.

To extend the SANDHY method to the entire mainland of France, predictor domains are optimised in the next chapter for each of the 608 climatologically homogeneous target zones in the Safran dataset (Vidal et al., 2010).

5 Local optimisation of spatial predictor domains

In this chapter we will have a look at the predictor domains and their spatial optimisation, especially the ones for the geopotential shape predictor. We want to optimise them locally, that is for a high number of locations. Which algorithm is fast enough to allow an optimisation for over 600 locations in reasonable computation time? And how can equifinality, that is the fact that different parameter combinations may lead to very similar performance – which is likely to occur if the performance is summarised over a long period – be taken into account? The first section of this chapter consists of an article published in Hydrology and Earth System Sciences (Radanovics et al., 2013a). As a reader of this thesis you may want to skip the sections 1.1, 2.1.1, 2.2 and 2.3 of the article since they mainly contain information already presented in the previous chapters. The predictor domain optimisation was subject of a presentation at the EGU General Assembly 2012 in Vienna (Radanovics et al., 2012) and a poster at the next generation climate data workshop 2013 at National Center for Atmospheric Research (NCAR) in Boulder, Colorado (Radanovics et al., 2013d). Equifinality is explored in more detail in the second section and possible reasons for the shape of the relevance maps in section one are explored in the third section.

5.1 Optimising predictor domains for spatially coherent precipitation downscaling (Article HESS)

Radanovics, S., Vidal, J.-P., Sauquet, E., Ben Daoud, A., and Bontron, G. (2013). Optimising predictor domains for spatially coherent precipitation downscaling. Hydrology and Earth System Sciences, 17(10):4189–4208.

Hydrol. Earth Syst. Sci., 17, 4189–4208, 2013
www.hydrol-earth-syst-sci.net/17/4189/2013/
doi:10.5194/hess-17-4189-2013

Hydrology and Earth System Sciences
Open Access

Optimising predictor domains for spatially coherent precipitation downscaling

S. Radanovics[1], J.-P. Vidal[1], E. Sauquet[1], A. Ben Daoud[2], and G. Bontron[2]

[1]Irstea, UR HHLY, 5 rue de La Doua, 69100 Villeurbanne, France
[2]Compagnie Nationale du Rhône (CNR), 2 rue André Bonin, 69316 Lyon Cedex 04, France

Correspondence to: S. Radanovics (sabine.radanovics@irstea.fr)

Received: 18 March 2013 – Published in Hydrol. Earth Syst. Sci. Discuss.: 2 April 2013
Revised: 29 August 2013 – Accepted: 9 September 2013 – Published: 28 October 2013

Abstract. Statistical downscaling is widely used to overcome the scale gap between predictors from numerical weather prediction models or global circulation models and predictands like local precipitation, required for example for medium-term operational forecasts or climate change impact studies. The predictors are considered over a given spatial domain which is rarely optimised with respect to the target predictand location. In this study, an extended version of the growing rectangular domain algorithm is proposed to provide an ensemble of near-optimum predictor domains for a statistical downscaling method. This algorithm is applied to find five-member ensembles of near-optimum geopotential predictor domains for an analogue downscaling method for 608 individual target zones covering France. Results first show that very similar downscaling performances based on the continuous ranked probability score (CRPS) can be achieved by different predictor domains for any specific target zone, demonstrating the need for considering alternative domains in this context of high equifinality. A second result is the large diversity of optimised predictor domains over the country that questions the commonly made hypothesis of a common predictor domain for large areas. The domain centres are mainly distributed following the geographical location of the target location, but there are apparent differences between the windward and the lee side of mountain ridges. Moreover, domains for target zones located in southeastern France are centred more east and south than the ones for target locations on the same longitude. The size of the optimised domains tends to be larger in the southeastern part of the country, while domains with a very small meridional extent can be found in an east–west band around 47° N. Sensitivity experiments finally show that results are rather insensitive to the starting point of the optimisation algorithm except for zones located in the transition area north of this east–west band. Results also appear generally robust with respect to the archive length considered for the analogue method, except for zones with high interannual variability like in the Cévennes area. This study paves the way for defining regions with homogeneous geopotential predictor domains for precipitation downscaling over France, and therefore de facto ensuring the spatial coherence required for hydrological applications.

1 Introduction

For both, climate change impact studies and operational hydrological forecasts, precipitation information on the scale of small subcatchments is needed. Numerical weather prediction (NWP) models and general circulation models (GCMs) provide relevant information about the atmospheric large-scale circulation but have too coarse a resolution to be directly used in impact models like hydrological models or for precipitation forecasts on the scale of small subcatchments. A downscaling step is therefore required, and this can be done dynamically using regional climate models and limited-area models or using statistical methods that make use of statistical relationships between large-scale predictors and local-scale predictands.

Requirements for hydrological use of predictands specifically include the spatial coherence of precipitation fields – i.e. a realistic spatial distribution of precipitation at any time step – over potentially large basins. Indeed, the generation of floods is, for example, particularly sensitive to the spatial distribution of precipitation over the catchment

considered. While dynamical downscaling methods naturally provide such a sought-after spatial coherence, this is not necessarily the case for statistical methods.

This paper proposes some development on how to ensure such a spatial coherence in precipitation by optimising the predictor domains of an analogue downscaling method for different individual target locations over France and analysing the spatial variability of results. This work will help in identifying regions with homogeneous geopotential predictor domains for precipitation, over which the spatial coherence would be de facto ensured by the selection of common analogue dates.

1.1 Statistical downscaling methods

In statistical downscaling, a relationship between large-scale predictors provided by GCMs and local-scale predictands is established. There are three major groups of statistical downscaling methods used in a climate change context: model output statistics (MOS) (e.g. Chandler, 2002; Friederichs and Hense, 2007; Vidal and Wade, 2008; Lavaysse et al., 2012), perfect prognosis (PP) (e.g. Timbal et al., 2003; Boé et al., 2006; Hertig et al., 2012) and weather generators (e.g. Vrac et al., 2007; Chen et al., 2010; Bellone et al., 2000). A review of methods and their strengths and weaknesses to produce relevant input for impact models can be found in Maraun et al. (2010). Both PP and MOS methods are also applied for operational precipitation forecast (e.g. Marty et al., 2008, 2012, 2013; Voisin et al., 2010; Nam et al., 2011; Liu and Coulibaly, 2011; Muluye, 2011).

A number of statistical downscaling studies with various methods have been performed over France over the last few years, but mainly for specific regions like the French Mediterranean (e.g. Quintana Seguí et al., 2010, 2011; Lavaysse et al., 2012; Kallache et al., 2011; Carreau and Vrac, 2011; Nuissier et al., 2011; Beaulant et al., 2011), western France (e.g. Timbal et al., 2003), the French Alps (e.g. Martin et al., 1997) or the Seine Basin (e.g. Boé et al., 2006). Until the present study, only a few of them had been performed at the country scale (Boé and Terray, 2008; Boé et al., 2009).

The downscaling method used in this work follows an analogue approach. It belongs to the PP methods, and is based on the idea introduced by Lorenz (1969) in weather forecasting that similar causes have similar effects; that is, similar predictor fields lead to similar predictand values. Nowadays numerous variants using different types of predictor fields and distance measures are in use. They range from weather-typing-based methods based on principal components of mean sea level pressure fields (Boé et al., 2006) to MOS-like techniques based on precipitation field analogues (Hamill and Whitaker, 2006; Turco et al., 2011). A description of the theory of probabilistic forecasts with analogues can be found in Hamill and Whitaker (2006). Analogue methods have been applied in different regions of the world with very diverse climates, e.g. Switzerland (Horton et al., 2012), Australia (Timbal and McAvaney, 2001), central Sweden (Wetterhall et al., 2005), Punjab (India) (Raje and Mujumdar, 2011), southeast USA (Zhang and Georgakakos, 2012), the Alpine region (Themeßl et al., 2011), and northeast Spain (Ibarra-Berastegi et al., 2011).

1.2 Predictor domains: optimisation

The predictor variables used for statistical downscaling and the predictor domains have to be chosen carefully. The predictor variables should have predictive skill for the quantity to predict – in this case, precipitation. These predictors should be quantities that are reliably simulated by NWPs and GCMs, and ideally they should be related to the processes leading to precipitation, and for climate change applications this relationship should persist in a changing climate (Wilby et al., 1998).

In most downscaling studies, no optimisation of the predictor domains has been performed, and only a few of them have tested even a handful of different domains (Timbal and McAvaney, 2001; Timbal et al., 2003; Gutiérrez et al., 2013). Timbal and McAvaney (2001) especially found that choosing an informative predictor domain is an important issue for the analogue selection. Ben Daoud (2010) found that some predictors like temperature or moisture variables have their main influence close to the target location, and therefore a small predictor domain close to the target location is likely to be sufficient. The predictor domains for the shape of the geopotential field are usually larger, and their optimum location depends on the meteorological situations that lead to precipitation at the target location.

Various algorithms may be used to optimise predictor domains. Ideally all predictor variables, predictor domains and other parameters should be optimised together and predictor domains of any size and shape should be possible. This was done by Sauter and Venema (2011) for an artificial neural network downscaling method and one target location in the Rhineland (Germany). Large computer resources were needed to do so because the search space is huge. A global optimisation of the analogue method was done by Horton (2012) for some stations in the Swiss Alps using genetic algorithms, with substantial computational costs as well.

This work focuses on optimising the predictor domain of one variable: the geopotential height. Restricting the parameters to be optimised allows for optimising domains for a large number of target zones separately and for exploration of the near-optimum domains for each target zone rather than searching for a unique optimum following the equifinality thesis (Beven, 2006).

1.3 Predictor domains: spatial variability

When analogue methods are applied, only one predictor domain for all target locations is generally used (see, e.g.

Timbal et al., 2003), because this ensures that the same analogue date will be selected for the whole region, and this naturally leads to spatial coherence of the precipitation field as long as individual fields and no summary measures are used. But for large target regions like France or large catchments with diverse precipitation climates like the Rhône Basin, this will likely result in a lower skill for smaller subcatchments. Bontron (2004) optimised the geopotential predictor domains for individual groups of precipitation stations located in France and northern Italy and compared the performance with those optimised for all groups together. For the groups near the barycentre of all groups, the difference in skill was small, but for groups far from the barycentre the skill was clearly better using the individually optimised predictor domains. Furthermore, he suggested to use the same predictor domain for groups situated not more than 250 km apart from each other and not separated by a major "climatological barrier".

This work considers a large number of individual target locations over France in order to assess the spatial variability in locally optimised domains, in terms of both location and shape. The use of several near-optimum domains furthermore allows for assessment of the diversity of domains associated with very similar performance for single target locations.

1.4 Objectives and outline of the paper

The first objective of this paper is to present an extended version of the growing rectangular domain algorithm for optimising the predictor domains used by a statistical downscaling method. Such an algorithm may be used to find an ensemble of near-optimum predictor domains for any statistical downscaling method and consequently to address the issue of equifinality raised when trying to perform a numerical optimisation based on some summary skill score.

The second objective of the paper is to answer the following question: is the assumption made for example by Timbal et al. (2003) and Boé and Terray (2008) of a common predictor domain for large regions in France actually valid? To this aim, the extended version of the growing rectangular domain algorithm is applied to derive an ensemble of near-optimum geopotential predictor domains for 608 target zones covering the whole of France. The downscaling method considered for this application is an analogue method that has a long history of development with various applications, like hydrological forecasts (Ben Daoud et al., 2011b) or historical flood reconstruction (Auffray et al., 2011). However, it is applied here for the first time to the whole of France.

The methods used for downscaling and optimisation are described in Sect. 2, and results are described in Sect. 3. Section 4 presents some sensitivity tests that have been performed to check the robustness of findings with respect to (1) the archive length, (2) the version of the optimisation algorithm and (3) the optimisation starting point. Critical methodological choices are discussed in Sect. 5, and conclusions are given in Sect. 6.

2 Data and methods

2.1 Data

2.1.1 Reanalyses

The predictor domain optimisation is done using two archives of reanalysis data. ERA-40 data (Uppala et al., 2005) at 2.5° resolution were selected as the large-scale archive against other global reanalyses because of the trade-off between archive length and data assimilation technique, following Ben Daoud et al. (2011b, a). The archive length is critical (1) for including as many diverse analogue situations as possible, (2) for studying the sensitivity on the archive length (see Sect. 4.3) and (3) for having a completely independent (i.e. not used either for optimising domains or as an archive) time period left for validation using a rigorous split-sample approach as defined by Klemeš (1986). The NCEP/NCAR reanalysis (Kalnay et al., 1996) has a longer archive, but ERA-40 made use of the more advanced three-dimensional variational data assimilation. Ben Daoud et al. (2009) compared ERA-40 and NCEP/NCAR reanalysis as sources for large-scale predictors for the downscaling method used here and found a slightly higher skill using ERA-40. ERA-Interim (Dee et al., 2011) uses an even more advanced data assimilation technique and has a higher spatial resolution, leading to a better temporal consistency and a better representation of the hydrological cycle (see Dee et al., 2011), but has still a shorter archive than ERA-40. Preliminary tests by Ben Daoud (2010) with a 1.125° version of ERA-40 and a simpler variant of the downscaling method (see Bontron and Obled, 2005) showed only very small improvements in skill and quite similar optimised domains with the higher resolution archive. Moreover, a higher resolution large-scale archive would increase both the equifinality issue and the computation time. Furthermore, some hypotheses, notably on the predictor domains for temperature, vertical velocity and humidity (see Sect. 2.2), may not be appropriate anymore. Lastly, using ERA-40 data ensures consistency with the local-scale archive as this global reanalysis has been used as a first guess by the Safran system for computing vertical profiles of near-surface variables.

Safran (French near-surface reanalysis) data (Vidal et al., 2010) are used as predictands for the local daily precipitation, which is the target variable addressed by the downscaling. The Safran reanalysis data are defined on 608 climatologically homogeneous zones covering France. Inside these zones the meteorological variables are supposed to depend only on altitude. These zones are used as elementary units in this work and are shown in Fig. 1. The algorithm used for the Safran analysis as well as its validation and application over

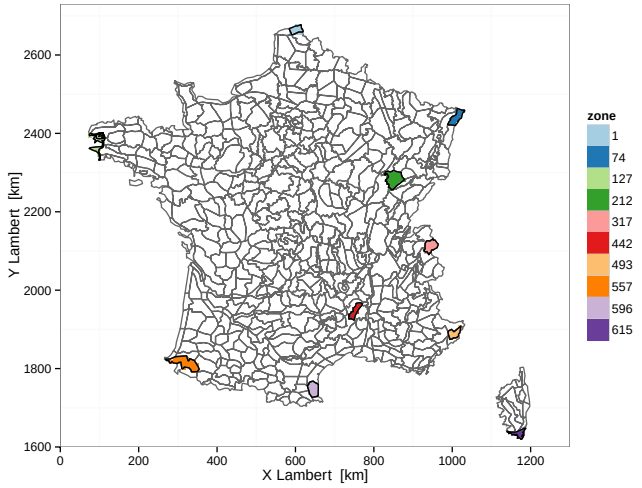

Fig. 1. The 608 climatologically homogeneous zones defined in the Safran data. The case study zones used in this study are coloured. The three zones situated in the Rhône catchment are Saône (212, dark green), Arve (317, light red) and Ardèche (442, dark red). The seven zones located at the geographical limits of the country are shown with other colours.

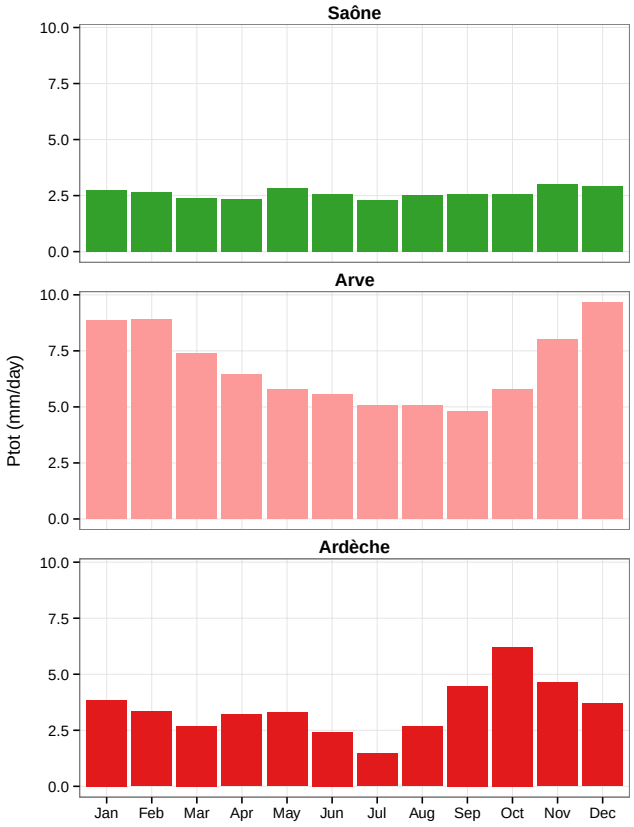

Fig. 2. Monthly mean precipitation, 1 August 1958–31 July 2002, from the Safran data for the three case study zones Saône, Arve and Ardèche.

France are described by Quintana-Seguí et al. (2008). A detailed validation of this 50 yr atmospheric reanalysis over France has been carried out by Vidal et al. (2010). Of particular interest to this study, they found that the reanalysis uncertainty on precipitation is both very low and relatively constant over the 1958–2008 period when considering both dependent and independent validation data. The bias calculated with 83 high-quality independent validation stations is smaller than 0.1 mm day^{-1}, and the root-mean-square error is around 2.5 mm day^{-1} (Vidal et al., 2010).

The common archive period for the two reanalysis data archives is from 1 August 1958 to 31 July 2002. The period 1 August 1982–31 July 2002 is used to optimise the geopotential predictor domains except for in the sensitivity test on archive length, where the whole common archive is used. This is discussed later in Sect. 5.

2.1.2 Case study zones

The domain optimisation was performed for all 608 zones in the Safran data set, but detailed sensitivity tests focused on three case study zones. All three selected zones are part of the Rhône catchment, but have different precipitation climates as shown in Fig. 2. This has implications on the spatial coherence, since different parts of the catchment receive precipitation in different meteorological situations, and this may lead to different informative spatial predictor domains and therefore different analogue dates. Furthermore, in Sect. 3.1 results are shown for zones located at the geographical limits of the country. Maps showing the skill of the downscaling method using a unitary-sized domain at all possible locations, so-called relevance maps, for these zones were used in a preliminary analysis to define the edges of the search domain. These zones at the geographical limits of the country as well as the case study zones are coloured in Fig. 1.

The case study zone called *Saône* (212) is located in the Burgundy region, in the Saône River valley. The terrain is rather flat and the zone is mainly influenced by the westerlies. The precipitation is uniformly distributed over the whole year. The second zone, named *Arve* (317), is located in the upper Arve catchment near Mont Blanc. The precipitation has a yearly cycle with a maximum in winter and a minimum in summer and early autumn. The third case study zone, named *Ardèche* (442), is located in the upper Ardèche catchment in the Cévennes area, and has a precipitation maximum in October with a high inter-annual variability (see Fig. 4 in Vidal et al., 2010). The precipitation maximum in autumn results from heavy precipitation events that are frequently observed in the Cévennes region during this season (see, e.g. Ricard et al., 2012).

2.2 Downscaling method

The downscaling method used here is an analogue approach that has already a long history of development in weather forecasting context, and some developments are underway to

Table 1. Predictors and their corresponding pressure levels, times, similarity criterion and number of situations selected in the given step used by the downscaling method.

Predictor	Pressure level [hPa] and time [UTC]	Similarity criterion	Number of analogues
Temperature	925 at 12:00 $D+1$, 600 at 12:00 D	E.d.[1]	2000*
Geopotential	1000 at 12:00 D, 500 at 00:00 $D+1$	TWS[2]	170
Vertical velocity	850 at 06:00, 12:00, 18:00 D and 00:00 $D+1$	E.d.[1]	70
Humidity (TCW × Rh)	850 at 12:00 D and 00:00 $D+1$	E.d.[1]	25

* depending on the length of the archive: 100 × number of years in the archive (44 yr → 4400, 20 yr → 2000, ...). [1] E.d., Euclidean distance; [2] TWS, the Teweles and Wobus shape criteria (Teweles and Wobus, 1954).

apply it in a climate change context. Duband (1981) was the first who applied the analogue method in France. Guilbaud and Obled (1998) introduced the analogue selection on gridded geopotential fields with the Teweles and Wobus shape criteria (Teweles and Wobus, 1954). Obled et al. (2002) calibrated the method for 50 French, Spanish and Italian catchments. Bontron and Obled (2005) introduced the use of reanalysis data as a historical archive – the NCEP/NCAR reanalysis (Kalnay et al., 1996) – instead of interpolated radiosonde data, and added local humidity to the predictor variables. Ben Daoud et al. (2011a, b) then introduced the temperature and the vertical velocity predictor variables.

This downscaling algorithm developed by Ben Daoud et al. (2011a) and applied here performs a four-step selection on temperature, geopotential heights, vertical velocity and humidity, respectively, to identify analogue dates in the archive. The predictor variables, similarity criterion and the number of analogue situations selected after each step are summarised in Table 1. The main characteristics of the method are summarised below; for details, see Ben Daoud (2010), who identified the optimum combinations of variables and times for the Seine and Saône river basins. The number of analogue dates retained after step 2, 3 and 4 were again taken from Ben Daoud (2010). The precipitation value for day D corresponds to the precipitation accumulated between 06:00 UTC day D and 06:00 UTC day $D+1$.

The first step is a selection on temperature at 925 hPa at 12:00 UTC day $D+1$ and 600 hPa at 12:00 UTC day D. The pressure levels and corresponding times were optimised by Ben Daoud et al. (2011a). The predictor domain is the ERA-40 grid point closest to the target location, which is reasonable as temperature can be seen as a proxy for the thermodynamical properties of the air on the local scale. A similar choice was made by Hanssen-Bauer et al. (2003). The similarity criteria is the Euclidean distance with equal weights for the two pressure levels. As shown by Timbal et al. (2008) and Hanssen-Bauer et al. (2003), including a temperature variable as a predictor is especially important in a climate change context since different temperatures may occur in a given season and the amount of water the atmosphere can hold depends on temperature. The number of analogue situations selected in step 1 depends on the length of the archive; it is 100× number of years in the archive – for example 2000 analogue situations for a 20 yr archive, as is used for the optimisation of the predictor domains for geopotential. This approximates the 4-month season length used by Bontron (2004) and the 2900 days Ben Daoud (2010) used with a 30 yr archive. The four days before and after the target date are excluded to avoid the selection of days within possibly the same low-pressure system.

The second step is a selection on geopotential at 1000 hPa at 12:00 UTC day D and 500 hPa at 00:00 UTC day $D+1$. The similarity criteria used is the Teweles and Wobus criteria S1 (Teweles and Wobus, 1954), called TWS in the following, which measures the similarity between the zonal- and meridional gradients expressed as the difference between each point of the predictor domain and all other points with the same longitude or latitude. Therefore the TWS measures the similarity of the shape of the fields. Guilbaud and Obled (1998) found that the TWS leads to better downscaling performance than the Euclidean distance for the geopotential predictor. This criterion has been widely used in various analogue methods (e.g. Wetterhall et al., 2005; Wetterhall et al., 2007; Teutschbein et al., 2011; Horton et al., 2012; Brigode et al., 2012) and weather-type classification (e.g. Garavaglia et al., 2010). Again equal weights are given for the two pressure levels. The same predictor domain is used for the two pressure levels. For this step the predictor domains are optimised using the method described later in Sect. 2.5. The 170 most similar days regarding geopotential shape out of the 2000 with the most similar temperature are selected. Geopotential- or pressure fields are often used as predictors because they are well simulated by the GCMs and contain information about the atmospheric dynamics like flow strength and direction or divergence (Wilby and Wigley, 2000).

The third step is a selection on vertical velocity at 850 hPa at 06:00, 12:00 and 18:00 UTC day D and 00:00 UTC day $D+1$. The similarity criterion is Euclidean distance, and the predictor domain is the nearest ERA-40 grid cell. Equal weights are given to the different times. Upward motion is necessary for the formation of clouds and precipitation. With a model resolution of 2.5° this predictor can only account

for large-scale upward motion due to dynamical reasons, and not for upward motion due to local convection or orography. Ben Daoud et al. (2011a) found some additional skill for the first two forecast days and fewer false alarms with the vertical velocity added as a predictor. The most similar 70 days out of the 170 days remaining after step 2 are selected.

The fourth step is a selection on humidity, more precisely the product of the total column water (TCW) and relative humidity at 850 hPa (RH) at 12:00 UTC day D and 00:00 UTC day $D+1$. This compound variable was found to be more informative than other simple indicators by Bontron (2004). The similarity criteria is Euclidean distance and the predictor domain is the nearest ERA-40 grid cell. The most similar 25 days out of the 70 days remaining after step three are selected.

The predictor variables, their pressure levels and hours, and the number of analogues to select after each step were taken from Ben Daoud (2010), where they were selected for the Seine and Saône basins. It has also to be noted that identical combinations of variables, pressure levels and hours in steps 2 and 4 had also been selected by Bontron and Obled (2005) for application at various locations in southeastern France.

2.3 Performance criterion

The skill of the downscaling method is assessed with the continuous ranked probability score (CRPS) (Brown, 1974; Matheson and Winkler, 1976). The CRPS is widely used for the verification of probabilistic atmospheric or hydrological forecasts (see, e.g. Hagedorn et al., 2008; Demargne et al., 2010; Aspelien et al., 2011). It is defined as follows:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left[F(x) - H^0_{x_{\mathrm{obs}}}(x) \right]^2 \mathrm{d}x, \quad (1)$$

where $F(x)$ is the forecasted cumulative distribution function of the variable x, x^0_{obs} the observed value and $H^0_{x_{\mathrm{obs}}}(x)$ the Heaviside function of $x - x^0_{\mathrm{obs}}$. The properties of the CRPS are as described in Hersbach (2000). The CRPS is sensitive to the entire range of the parameter, and no predefined classes are required; it is equal to the mean absolute error (MAE) in the case of a deterministic forecast, and it can be interpreted as an integral over all possible Brier scores. In order to compare results from different zones, the continuous ranked probability skill score (CRPSS) with the climatology as a reference forecast is used:

$$\mathrm{CRPSS} = 1 - \frac{\langle \mathrm{CRPS} \rangle}{\langle \mathrm{CRPS}_{\mathrm{clim}} \rangle}, \quad (2)$$

where $\langle\ \rangle$ denotes the time average and the $\mathrm{CRPS}_{\mathrm{clim}}$ is calculated over the 1 August 1982–31 July 2002 period – except for the 44 yr experiments, where the whole archive period is considered – using precipitation data from ± 60 days around the target day from different years in order to take seasonality into account.

2.4 Relevance maps

A relevance map represents the forecast skill for each grid cell of the predictor data set. Relevance maps were used, for example, by Bontron (2004) and Horton et al. (2012) to select the most predictive pressure level and time step for the geopotential predictor. Relevance maps are obtained by fixing every parameter except the location of a unitary-sized spatial domain (2×2 ERA-40 grid points, 2.5° resolution) that moves across the whole map (Horton et al., 2012). Using the TWS 2×2 grid points is the smallest possible domain since the TWS is based on the calculation of gradients, i.e. differences between two grid points. By iterating the position of this small domain, the CRPS score corresponding to every location is obtained. Relevance maps thus allow for one to see where the synoptic circulation information is relevant to explain observed or analysed precipitation time series. It is expected that the best predictor locations are consistent with the meteorological characteristics that are responsible for the region's weather. The predictor's best locations are therefore expected to be different for sub-catchments or stations influenced by different meteorological phenomena (Horton et al., 2012).

Relevance maps are used in this study (1) to illustrate the different atmospheric influences for zones in different parts of the country (Sect. 3.1), (2) to compare the optimised domains with the regions of high skill in the relevance maps (Sect. 3.2.2) and (3) to have an additional starting point for experiments on the sensitivity of the optimisation algorithm (Sect. 4.1).

2.5 Optimisation method

Geopotential was chosen for optimisation because it is the most important predictor in the downscaling method used, and the size and location of the predictor domain is supposed to depend more strongly on the typical weather pattern causing precipitation in the target area than for the other predictor variables. The predictor domain optimised here is a domain common to both geopotential levels described in Table 1.

The selected optimisation method is based on the idea of growing rectangular domains as applied by Bontron (2004) and Ben Daoud (2010). The basic version starts from a given 2×2 grid point domain (here the nearest one to the target zone), calculates a score (here the CRPS) and then expands the domain in four directions by adding one grid point east, west, north or south. For these four resulting domains the CRPS is calculated and the domain with the smallest CRPS is selected. This selected domain is then used as a starting domain in the next step. This is done until the score is not improved during four consecutive steps or the edge of the search domain is reached. This method is very fast, but explores

only a very small subset of the space of predictor domains, and therefore it is likely that some relevant domains are not tested.

For this work an extended version of this method was developed: five domains, instead of a single one, are retained and expanded in each step. With this method a larger number of domains are explored, and the five best domains found in this procedure are returned, thus providing an indication of variability around optimal domains. The equifinality thesis (Beven, 2006) postulates that very similar skill score values can be obtained with different parameter sets, and that it is therefore beneficial to search for a number of very good performing parameter sets rather than for the best one.

Like in the basic version, the search is started from a given 2×2 grid point domain, here the nearest one to the target zone where not indicated differently. After calculating the CRPS this domain is expanded in four directions by adding one grid point east, west, north or south, and calculating the CRPS for each of them. For the second step all four domains from the previous step are expanded. This gives 16 domains, but only 10 actually different ones, so 10 new domains are explored. From these 10 domains in the second step, the 5 best are selected to be expanded in the next step. Theoretically there are up to 20 domains (5×4) to explore from step 3 on, but there is some redundancy or some domains have already been explored in a previous step, and as such between 13 and 18 actually new ones were found. In the end the five best domains found during the whole procedure, in general stemming from different steps, are returned.

Brigode et al. (2012) optimised the predictor domains for a rainfall-based weather pattern classification and therefore tested domains of three different sizes for every possible location similar to the relevance map calculation. This is a complementary approach to the one used in this work since they assumed a domain size and shape and then tested where to centre it. In the present work, a starting point is fixed, which defines a location that has to be included in the final domain, and then domains of different size and shapes but all containing the starting point are tested. An approach similar to the one adopted by Brigode et al. (2012) was chosen by Obled et al. (2002) for a previous version of the analogue method used here. They first tested domains of six different sizes centred over the target location and then shifted the best one to find the best location. Sauter and Venema (2011) optimised the predictor domains for an artificial neural network for all predictor variables together, allowing for three-dimensional and disjointed predictor domains. In contrast to our study, they did the optimisation for only one target location, and even for this one location they stated that the computational costs were very high. It needs to be noted that the approach selected here does not allow for exploration of non-rectangular domains suggested by the examination of relevance maps (see Sect. 3.1) and that may lead to better skill values. Allowing for this type of domains would, however, involve much higher computational costs.

3 Results

In the following sections, results are shown on the different regions of influence as mapped with relevance maps in Sect. 3.1, the size and location of the optimised predictor domains (Sect. 3.2) and the downscaling skill of the method using the optimised domains during the optimisation period (Sect. 3.2.1). In Sect. 4 sensitivity experiments on the choice of the starting point of the optimisation, the choice of the basic or extended optimisation method and the archive length are shown.

The downscaling and optimisation methods are implemented in Fortran 2003 using the NetCDF Fortran90 library (Pincus and Rew, 2011) for data input and output. The subsequent analysis and figures are done using the R software environment (R Development Core Team, 2012) with packages ncdf (Pierce, 2011), ggplot2 (Wickham, 2009), reshape2 (Wickham, 2007), RColorBrewer (Neuwirth, 2011; Harrower and Brewer, 2003), sp (Pebesma and Bivand, 2005; Bivand et al., 2008), zoo (Zeileis and Grothendieck, 2005) and gridExtra (Auguie, 2012).

3.1 Different regions of influence

The relevance maps for different Safran zones located at the geographical limits of France and in the Rhône Basin (cf. Fig. 1) and calculated from the 20 yr archive are compared in Fig. 3. This figure also contains the corresponding optimised domains that will be discussed later in Sect. 3.2.2. First the magnitude of the skill differs between the relevance maps for different zones. The highest skill is found for zones that are mainly exposed to the westerlies (127, 557, 317). Furthermore, there is a clear difference in the spatial pattern between different zones. The zones in western, northern and northeastern France (001, 074, 127, 557, 317) have their region of maximum skill located west or southwest of the zone. Their regions of high skill are larger in zonal direction than in meridional direction and are cyclonically curved. They are exposed to the westerlies and receive precipitation mainly from frontal systems. A similar shape was found by Horton et al. (2012) for the Marécottes station in Switzerland, located close to the Arve zone (317). The zones in southeastern France have their region of maximum skill located south or southeast of the zone (493, 596, 442, 615). Indeed, the heavy precipitation events in this region are associated with southerly or southeasterly flow (e.g. Ricard et al., 2012). Their regions of highest skill are more north–south oriented, with high-skill regions extending westward at the southern end and eastward and northwestward at the northern end. What all relevance maps have in common is a local minimum of skill surrounded by regions with higher skill. This is due to the use of the TWS criterion that is sensitive to the gradients of the geopotential fields and their anomalies on days with precipitation. The region of low skill corresponds to the location of a minimum in the mean geopotential

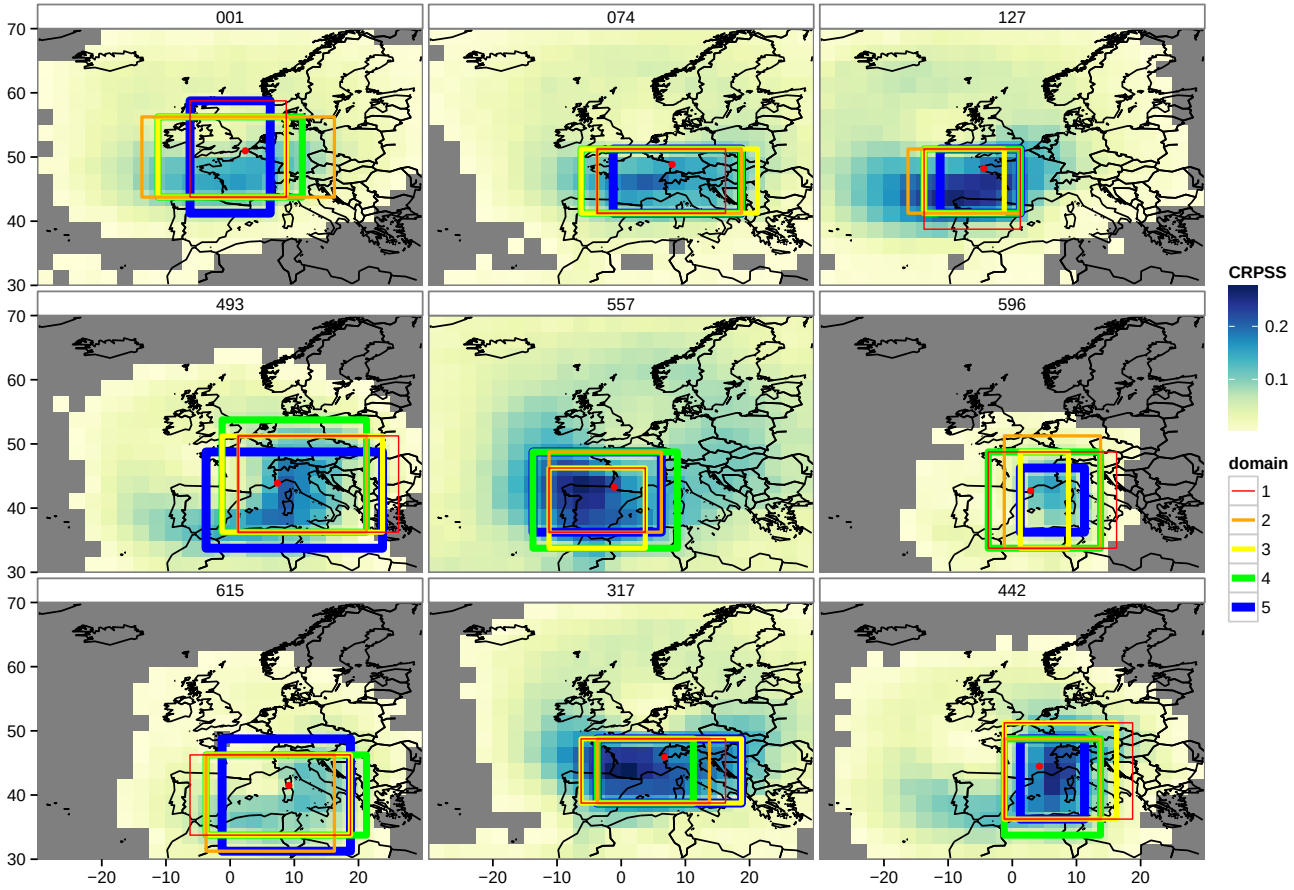

Fig. 3. Relevance maps truncated at CRPSS = 0 (areas with higher skill than the climatology) and optimised domains for nine zones identified in Fig. 1 using the 20 yr archive. The best domain found is drawn in red, followed then by those in orange, yellow, green and blue. The zone location is indicated by a red dot.

anomaly fields for rainy days (not shown). The largest gradients are situated around this minimum, which makes these regions more relevant using a similarity measure based on gradients.

Given the high seasonality with the precipitation maximum in autumn for the Ardèche case study zone (442) (cf. Fig. 2) due to specific atmospheric flow conditions, we investigated the score variations for different seasons. Seasonal relevance maps were obtained by averaging the CRPSS over different seasons instead of the whole year. In order to have enough data for each season, the score was calculated for the whole 44 yr archive. Relevance maps for the Ardèche case study zone for different seasons are shown in Fig. 4. The highest skill can be found for the winter season, followed by autumn. The location of the maximum of skill southeast to south-southeast from the Ardèche target zone corresponds well with the south-southeasterly flow found by Duffourg and Ducrocq (2011) for heavy precipitation events in the Cévennes region. In spring and summer the skill is lower due to convective precipitation, which is more difficult to predict based only on large-scale fields. This is a common feature for the three case study zones (not shown). Interestingly the shape of the region with high skill is very similar between the seasons, which was not expected for the Ardèche zone due to the specific flow condition that leads to the autumn precipitation maximum in this zone (cf. Fig. 2). Further investigation could look at relevance maps for days with different precipitation thresholds, but this is beyond the scope of this paper.

3.2 Optimised predictor domains

In this section we will show results on the optimised domains. The downscaling skill measured with the CRPSS for all 608 zones in France is shown first. The near-optimum domains found for the case study zones are then presented, before looking at summary characteristics for all 608 zones in France.

3.2.1 Downscaling skill

Figure 5 (left panel) shows the CRPSS calculated over the 20 yr optimisation period (1 August 1982–31 July 2002) for the best domain found for each of the 608 climatologically

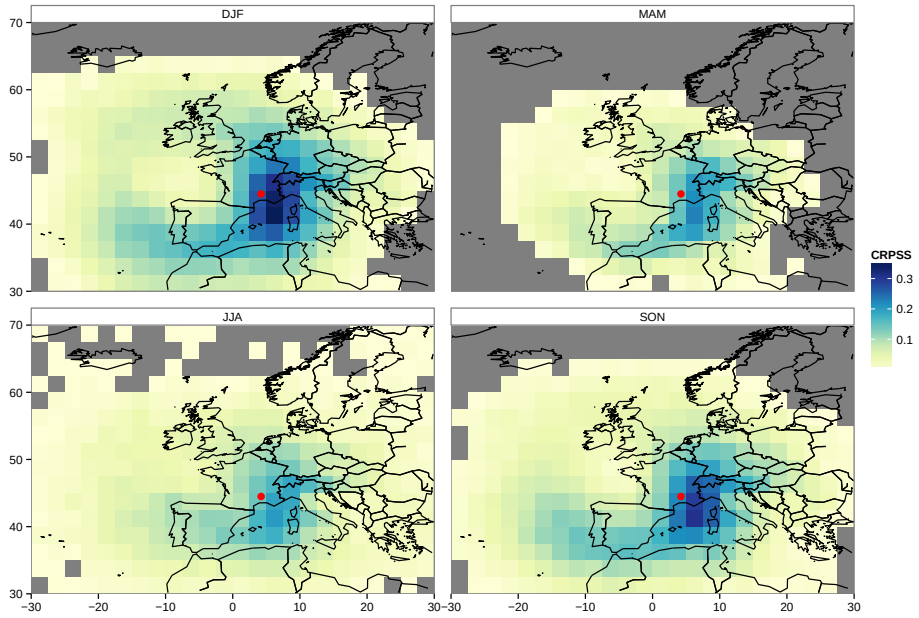

Fig. 4. Seasonal relevance maps for the Ardèche case study zone.

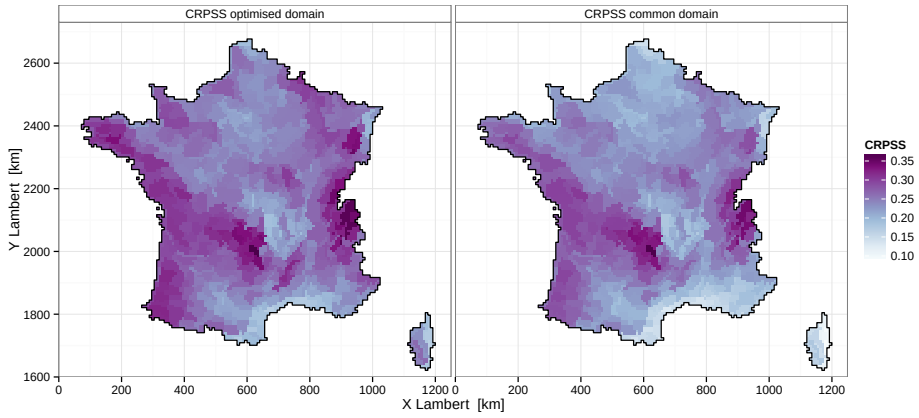

Fig. 5. Left panel: CRPSS obtained with the best domain found during optimisation. Right panel: CRPSS obtained with a domain optimised for average precipitation over France. Dark purple corresponds to a higher (that is, better) skill score, while light blue corresponds to lower skill.

homogeneous zones in France. The zones are coloured according to the CRPSS value obtained. Unsurprisingly, the CRPSS shows a spatial distribution similar to the one of the mean precipitation (Vidal et al., 2010). The more precipitation a region receives, the higher the CRPSS. The highest skill, between 0.30 and 0.35, is found on the windward side (west side in this case) of the Alps, the Massif Central and the Vosges, and along the Atlantic coast. Poorer skill, around 0.2, can be found on the lee side of mountains and around the Mediterranean coast. Quite interestingly, the difference in skill measured with the CRPSS between the best and the fifth-best domain found is never larger than 0.01. So the difference in skill between different optimised domains for the same zone are about one order of magnitude smaller than the differences in skill between different zones, which makes all five domains equally plausible. Additionally, the skill difference does not show any apparent spatial structure.

In order to compare these CRPSS values with some reference values, a set of common geopotential predictor domains were optimised using the average precipitation time series over France. The starting point as well as the predictor domains for the other predictor variables were chosen to be close to the centroid of the country. The right-hand side of Fig. 5 shows the CRPSS obtained for each zone with the best of the common predictor domains found. The mean CRPSS over the whole country is 0.24, compared to 0.26 for the individually optimised domains. Optimising the domains locally corresponds to improvements ranging from 0.45 to 77 % for

specific locations. The largest differences can be seen close to the country borders, in southeastern France and especially on Corsica, i.e. in areas with very specific regions of influence (see Fig. 3).

3.2.2 Case study zones

As shown in Fig. 3, the optimised domains tend to include the most relevant area depicted by the relevance maps for the case study zones. They differ reasonably between different locations and inside the ensemble for a given zone. For the majority of the zones (074, 127, 493, 557, 615, 317) the aspect ratio of the optimised domains varies little inside the five-domain ensemble, while there are larger variations of this property for the zones 001, 596 and 442. These large differences in aspect ratio do not lead to larger differences in skill as mentioned above. This exemplifies the equifinality issue mentioned in Sect. 2.5.

3.2.3 Domain characteristics at the scale of France

Figure 6a shows the mean location of the centre of the optimised domains for each of the 608 zones using a 2-D colour scheme for bivariate maps introduced by Teuling et al. (2011). Here the two variables that are combined are the longitude and the latitude of the domain centre. Thus the colours correspond to the mean location of the domain centres of the five best domains for each target zone. The domain centres for the best domains are mainly distributed following the location of the target zone, but in general the mean domain centre is situated south of the target zone. Nevertheless there are some deviations from this general pattern. For zones on the east side of the Massif Central, the centres of the optimised domains are located clearly more east than the ones for zones on the west side of the massif. The same feature can be seen at other mountain ridges, for example the Vosges mountain range. Furthermore the domain centres for the zones in southeastern France are located more east than north of this area. In some regions such as, for example, Champagne in the northeast of the country (approximate Lambert coordinates $X = 700$, $Y = 2400$), we can see that many zones have their average optimised domain centre at approximately the same location. In contrast, for the Cévennes and the southern Alps regions, the average domain centres differ more often between neighbouring zones.

Figure 6b shows the maximum difference in domain centre location between two domains in the five best domain ensemble in degrees longitude and latitude. For the majority of the zones the domain centre location is a very stable property (green colour), especially in latitude direction, where differences of more than 2° are rare. So, in general, the centre points of the five near-optimum domains for a zone are close to each other. Zones with larger differences, up to 8° in longitude, are located in the southeastern part of the country at the slopes of the Alpes and the Massif Central.

Figure 7a shows the mean size in degrees longitude and latitude of the optimised domains for each zone. Again the 2-D colour scheme is used, with the mean domain length in zonal direction and the mean domain length in meridional direction being the two variables. Small optimised domains (green) can be found in Brittany (150, 2400), Champagne (750, 2500), Lorraine (850, 2450), Poitou–Charentes (300, 2200) and in some parts of Normandy (300, 2500). Optimised domains with small extent in longitude direction but somewhat larger extent in latitude direction (blue) can be found along the Mediterranean coast and in the northernmost part of the country. Domains with small extent in the latitude direction and larger extent in the longitude direction (yellow) form an east–west-oriented band in the middle of the country (around 2250 km Y Lambert). Medium-sized domains (grey, brown) are found north of this band (500–900, 2350), in the southwest of the country and on the west side of Corsica (1150, 1700). The largest domains (purple, red, dark blue) tend to be situated in the southeastern part of France, except near the coast. The most prominent feature in this map is the area in the middle, where the optimised domains are very small in the meridional direction, while being reasonably stretched zonally.

The domain sizes used in other downscaling studies were compared to the domain sizes found in this study. Bontron (2004), Ben Daoud (2010), Timbal and McAvaney (2001), Boé et al. (2006) and Guilbaud and Obled (1998) used predictor domains with sizes of 20–25° longitude and 10–15° latitude, which corresponds to upper-medium-sized domains found in this study. Timbal and McAvaney (2001) (for daily minimum and maximum temperatures) tested somewhat smaller and much larger domains as well, but found the one of $20 \times 12°$ to perform best. The domains tested by Brigode et al. (2012) correspond to small- to medium-sized ones found in this study. Timbal et al. (2003) (for daily minimum and maximum temperatures) used a domain somewhat larger in north–south direction. Larger domains were used by Boé and Terray (2008), Hanssen-Bauer et al. (2003), Matulla et al. (2008) and Obled et al. (2002).

Figure 7b shows the ratio of domain size range in the five-domain ensemble, defined as follows:

$$\text{ratio} = \frac{\max(X) - \min(X)}{\text{mean}(X)}, \tag{3}$$

where X is the extent of the domains in degrees longitude or latitude. A size ratio of 0 means that all five domains have equal extent. A size ratio of 1 means that the difference in extent between the largest and the smallest domain is equal to the mean extent. On average the size ratio is larger in the longitude direction than in the latitude direction. The figure is quite patchy, with individual zones showing large ratios in one or both dimensions. In the north of the country and along the Mediterranean coast, these individual zones have large ratios in longitude or both dimensions. The zones 001 and 596 shown in Fig. 3 are examples of such zones. The zones

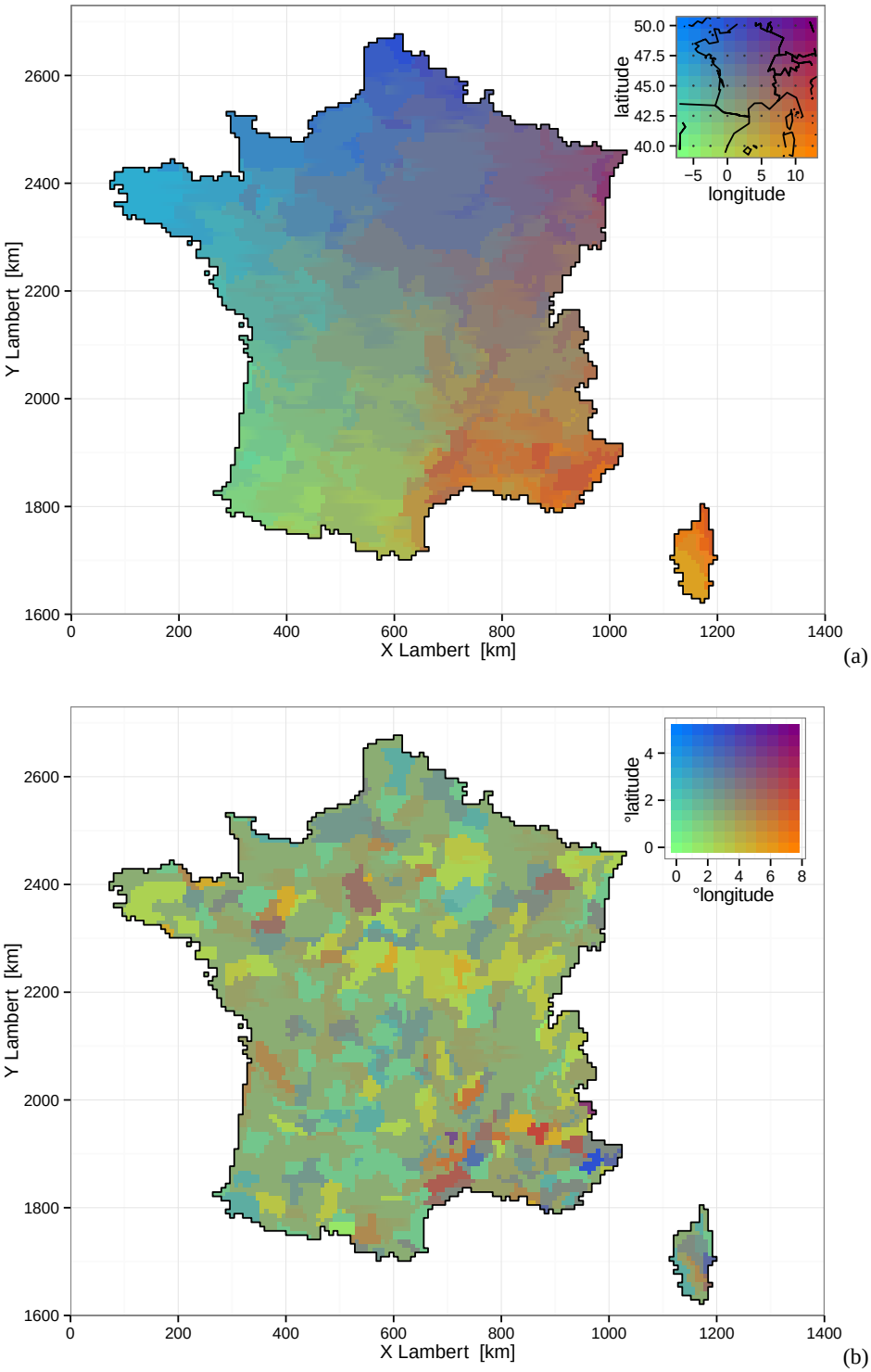

Fig. 6. (a) Mean domain centre of the five best domains found during the domain optimisation. The colours correspond to the mean location of the domain centres of the five best domains for each target zone. **(b)** Maximum difference in domain centre location between two domains in the five-best-domain ensemble. The colours correspond to the maximum distance of centre points in degrees longitude and latitude. Green, the centres are very close to each other; purple, the centres are very far from each other; blue, the centres are close in longitude but comparatively far in latitude direction; and orange, far in longitude and close in latitude direction.

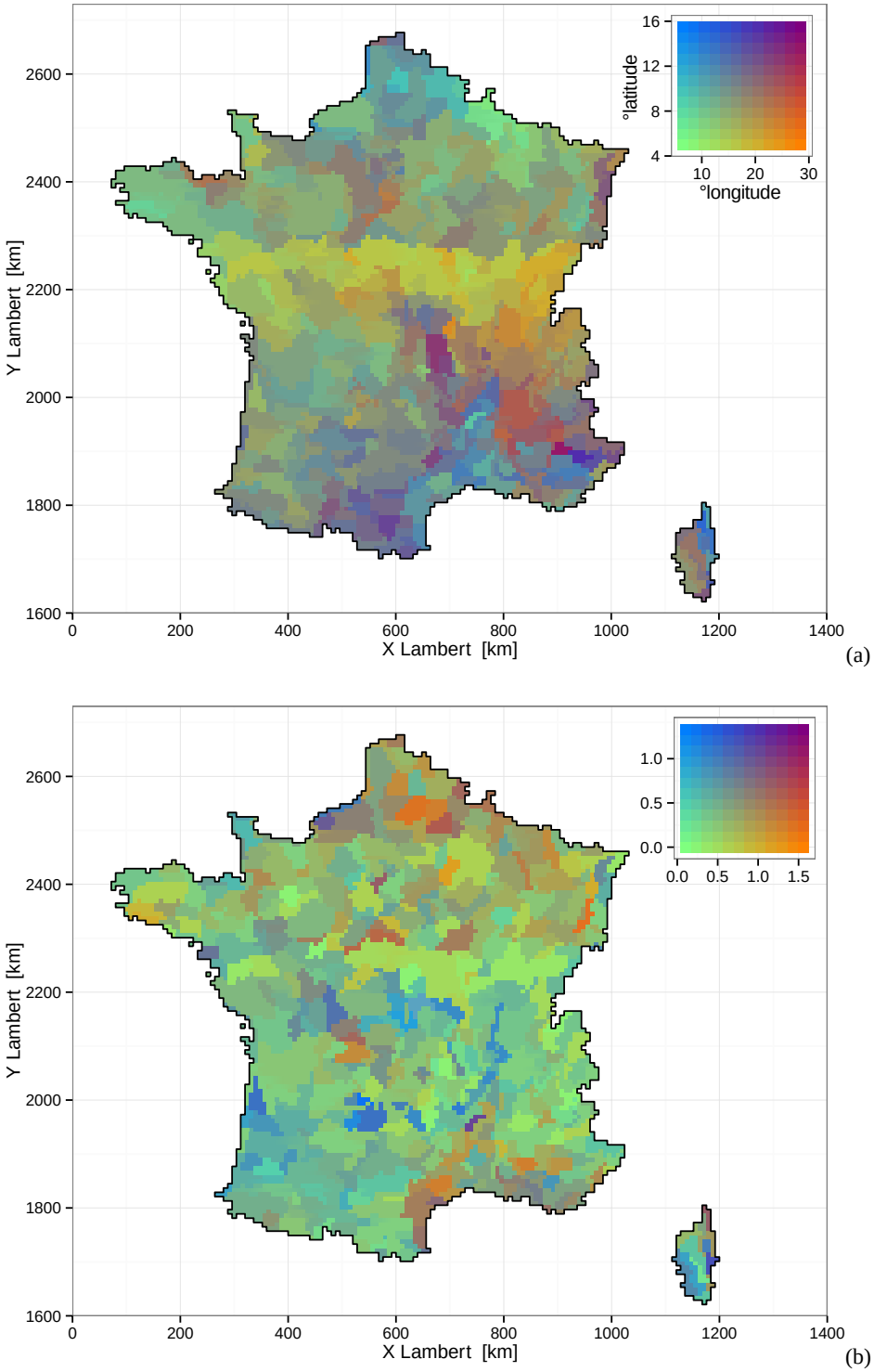

Fig. 7. (a) Mean domain size of the five best domains found during the optimisation procedure for each zone. The size is given in degrees longitude and latitude. Green, small domains; purple, large domains; blue, small in longitude direction and comparatively large in latitude direction; orange, large in longitude direction and small in latitude direction.**(b)** Ratio of domain size range in the five-domain ensemble (range/mean size). 0, no difference in domain extent; 1, the difference in extent between the largest and the smallest equals the medium extent.

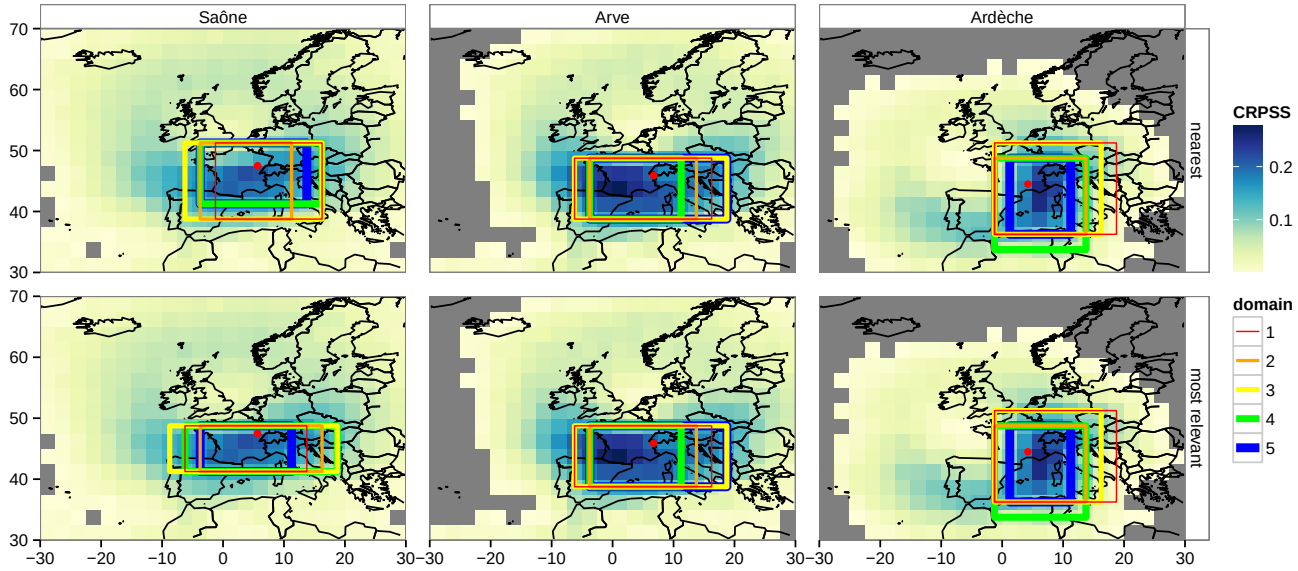

Fig. 8. Optimised predictor domains for three case study zones and different starting domains for the optimisation. First row, start at the nearest elementary domain; second row, start from the most relevant elementary domain from the relevance map. The relevance maps for each zone are shown with a colour scale underneath the predictor domains.

with the largest domain size range in the latitude direction are situated in the southern half of the country, except near the Mediterranean coast.

4 Sensitivity experiments

For the optimisation study, options were selected concerning the choice of the algorithm, the starting point and the archive length. In this section we take a detailed look at the impact of these choices on the optimised domains for the three case study zones by comparing with results for alternative choices.

4.1 Starting domain for optimisation

The growing rectangular domain algorithm requires the definition of a starting domain for the optimisation. This is true for other algorithms as well, but since the growing rectangular domain algorithm only adds grid cell rows or columns in each step and never subtracts any, the starting domain will automatically be included in the final domain (see Sect. 2.5). Therefore the choice of the starting domain can influence the predictor domains found and a poorly chosen starting point may lead to less skillful predictor domains.

One reasonable assumption is that the best predictor domain will comprise the large-scale grid cell closest to the target location, as is done here or in Obled et al. (2002). Another possibility is to start at the most relevant elementary domain, as obtained through a relevance map as done by Bontron (2004) and Ben Daoud (2010), to make sure that the most relevant location is included in the final predictor domain. The drawback of the second approach is that the computational costs for the relevance maps are high if performed for over 600 target locations. Roughly 2.8 million CRPS calculations per zone are, for example, needed for a $40° \times 60°$ sized relevance map with a 20 yr archive. Therefore the relevance maps were computed only for the case study zones, and for these zones the optimised domains obtained with the two different starting domains are compared.

The first line of Fig. 8 shows the five best domains found with the optimisation procedure starting at the nearest elementary domain, with a 20 yr archive. In the second line the same procedure is used, but the optimisation was started from the most relevant elementary domain as found with the relevance map. Comparing them we can see for the Arve zone and the Ardèche zone that exactly the same five domains are found even if the two starting domains are different. For the Saône zone, five different domains are found, with lower meridional extent and systematically higher zonal extent when starting from the most relevant elementary domain. The domains found starting from the most relevant elementary domain have higher CRPSS.

4.2 Optimisation method

Results obtained with the basic growing rectangular domain algorithm and the extended one developed here are compared for the case study zones. Figure 9 shows the optimised domain found with the extended algorithm (first row) and the ones found with the basic algorithm. For the Arve zone the best domain is the same for the two algorithms. For the Saône zone and the Ardèche zone the domain found with the basic algorithm is the second-best found with the extended version.

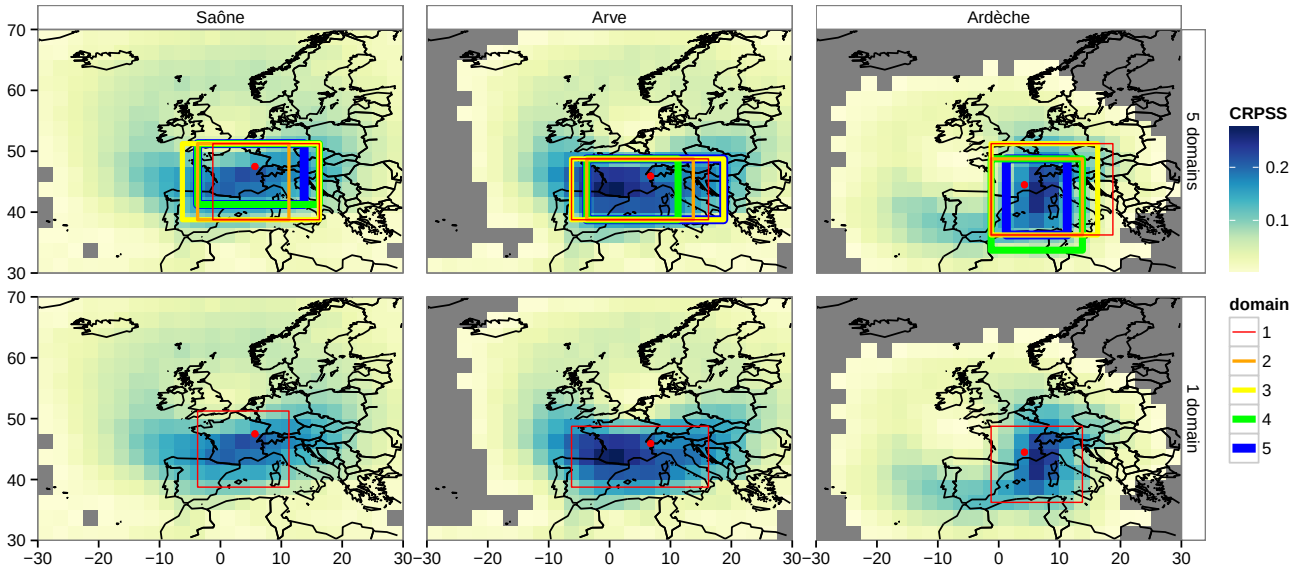

Fig. 9. Optimised predictor domains for three case study zones using the extended optimisation method with five domains (first row) or the basic optimisation method with one domain (second row). The relevance maps for each zone are shown with a colour scale underneath the predictor domains.

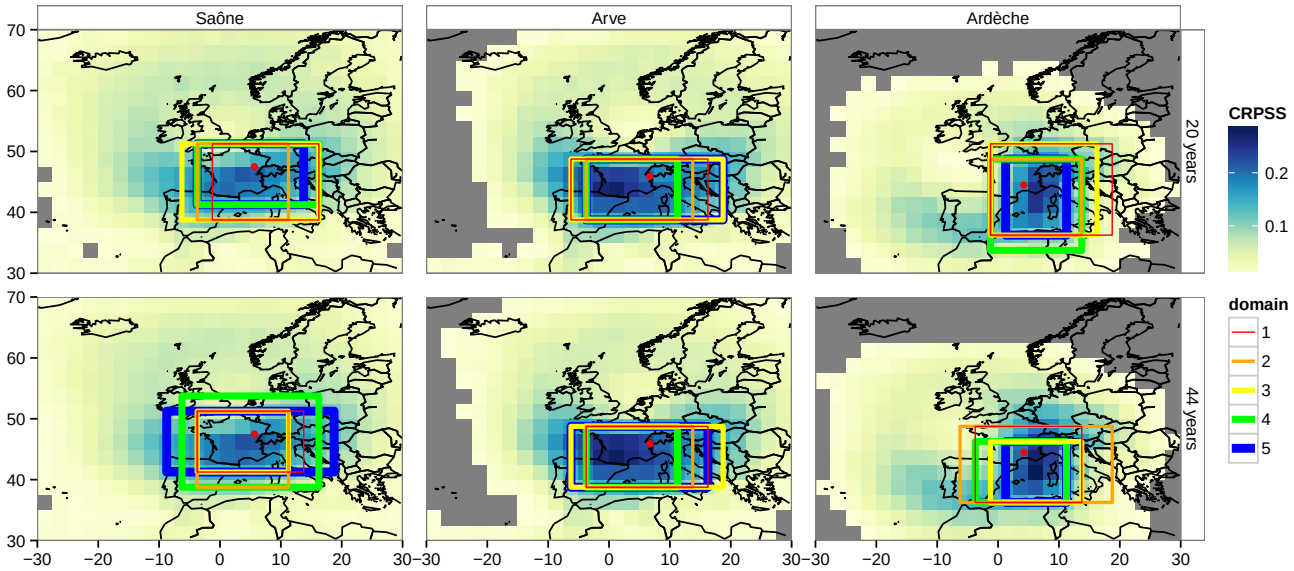

Fig. 10. Optimised predictor domains for three case study zones using 20 yr (first row) and 44 yr (second row) archives for optimisation. The relevance maps for each zone and each archive length are shown with a colour scale underneath the predictor domains.

This shows that for some zones the extended algorithm finds domains with slightly better CRPSS, together with an indication of variability between near-optimum ones.

4.3 Archive length

Figure 10 shows the optimised domains found with the 20 yr archive (1 August 1982–31 July 2002, first row) and 44 yr archive (1 August 1958–31 July 2002, second row). For the Saône zone (first column) the domains found with different archive lengths differ, but the second-best domains found are the same, and the best domain found with the 44 yr archive is the same as the fifth best found with the 20 yr archive. For the Arve zone (second column) the differences between the best domains found with the two archive lengths are small. The best domain found with the 20 yr archive is one grid cell larger in the west than the one found with the 44 yr archive, and was found to be fifth best with the 44 yr archive. All top-five domains found have the same extent in meridional

direction. For the Ardèche zone, five completely different domains are found with the different archive lengths. This is probably related to the high inter-annual variability in the target region (see Fig. 4 in Vidal et al., 2010). A slight reduction of skill over these 20 yr (0.282 to 0.275) can be observed for the Ardèche zone when considering the domains optimised over 44 yr. Additionally, the skill computed over the 44 yr is slightly lower (0.305 to 0.310) for the domains optimised over 20 yr compared to the ones optimised over 44 yr.

Additionally, the relevance maps obtained with different archive lengths show the same structure and the same location of maximum values, but the absolute values are slightly higher with the longer archive. On relevance maps obtained with a 10 yr archive (not shown) the same overall structure is still visible, but with a decrease in CRPSS of approximately one-third.

5 Discussion

5.1 Choice of the archive period

For successful statistical downscaling it is necessary to have long data sets of predictors and predictands for building and validating the model (Timbal and McAvaney, 2001). The archive length and optimisation period chosen for statistical downscaling development depend strongly on the data that are available and the validation strategy. Bontron (2004) and Ben Daoud (2010) left only five years of their archive for validation. Ben Daoud (2010) for example excluded the years 1978, 1983, 1988, 1993 and 1998 from the 1972–2002 optimisation period. The specific years were chosen to resemble the 1972–2002 climate as closely as possible in order to validate the method for forecast purposes, i.e. in the same climate. Timbal et al. (2003) found that using more than 20 yr of the reanalysis archive does not further reduce the error in the reconstructed time series of minimum and maximum temperature as long as the more recent part of the data is used, indicating that the quality of the observation data in terms of homogeneity and the reliability of the reanalysis plays an important role too. A 20 yr recent period has been considered here for optimising the predictor domains in order to (1) leave out enough data for future validation and (2) retain a period with the highest number of observations entering the ERA-40 reanalysis system (Uppala et al., 2005). Section 4.3 above provides some preliminary analysis of the sensitivity to the archive length.

5.2 Optimisation starting point

The starting point for optimisation was chosen to be the nearest elementary domain to the target zone. The optimisation method used requires that this elementary domain is included in the final domains. As seen above, using an alternative starting point at the most relevant elementary domain instead of the nearest one results in the same domains for two of the case study zones but in different ones for the third one, where more skillful domains, 3 % higher CRPSS, were found starting at the most relevant elementary domain. In Fig. 7a a sudden change in domain size can be seen around 47.5° N, with the domains north of this line having slightly larger domains in meridional direction. The Saône case study zone happens to be situated north of this line, and the experiment with the most relevant elementary domain as a starting point, more southwest in this case, showed that the optimised domains differ for this case study zone. The domains found starting the optimisation from the most relevant elementary domain are indeed very similar to those found for zones south of the Saône case study zone (not shown). Thus the sudden change in the domain sizes in Fig. 7a is likely to be a result of the starting point choice.

6 Conclusions

6.1 An algorithm to provide near-optimum predictor domains

An extended version of the growing rectangular domain algorithm has been described and applied for deriving ensembles of five near-optimum geopotential predictor domains for 608 individual target zones covering France. This algorithm allowed for us to find that different predictor domains may lead to very similar performances for the analogue downscaling method considered here. It exemplifies the equifinality issue in statistical downscaling that has been recognised in many other research domains (Beven, 2006). The equifinality is a consequence of a single-objective optimisation approach; that is, the use of a single-valued objective function. Indeed, the CRPSS used in this study as the objective function gives only an overall skill of the method. Consequently, for a given target location, some near-optimum domains may perform better than others – for example, for days with specific circulation patterns. This algorithm is potentially applicable in other contexts. This study has already shown that it could be applied at different target locations, but one may also think of considering another predictand, such as minimum or maximum temperature (Gutiérrez et al., 2013), or optimising the spatial domain of other predictors. As a result, this algorithm could be perfectly applied to another type of statistical downscaling method.

This first application of the downscaling procedure by Ben Daoud (2010) to the whole of France together with the use of the optimisation algorithm led here to a country-wide assessment of predictor domains. The domains resulting from an optimisation with the presented algorithm include the most relevant area depicted by the relevance maps for all three case study zones. The domains differ moderately between different locations and inside the ensemble for a given zone. In some regions, such as Brittany, we found a larger region with the same optimised domain, while

especially in the Rhône catchment we found high variability in the location and even more in the size of the optimised predictor domains. For the majority of the zones the aspect ratio of their five domains is rather similar, but for some zones, equally skillful domains with very different aspect ratios are found. The centres of the optimised domains are mainly distributed following the geographical location of the target zone but with clear differences between eastern and western slopes of mountain ridges. The domain centres for zones in southeastern France are located more east than north of this area. The domain centre location is a stable property in the five-domain ensemble except for isolated zones at the slopes of the Alps and the Massif Central. The domain sizes vary considerably between the zones with ensemble mean zonal extents between 6.5 and 28.5° and meridional extents between 5 and 15.5°.

6.2 On the assumption of a common predictor domain

This work addressed the hypothesis of a common predictor domain for different regions of France for statistical downscaling of precipitation. This assumption has been indeed made implicitly by all previous studies over France, e.g. Timbal et al. (2003) for western and southern France separately, and Boé and Terray (2008) for the whole of France. Results from the optimisation of geopotential predictor domains showed a large diversity of near-optimum domains for the set of 608 climatically homogeneous zones covering France, and therefore suggest that this assumption is questionable, at least when one seeks to obtain the most skillful method for each individual zone. However, relatively large zones have been found to share similar near-optimum predictor domains, and making this assumption within each of them may lead to limited loss of skill compared to domains optimised for individual locations. This is seemingly the case for the Seine Basin, where only minor variations in the optimised domains can be found (see Figs. 6 and 7), supporting the hypothesis made by Boé et al. (2006, 2007) for a common predictor domain over this basin.

Conversely, large river basins like the Rhône Basin include zones with very diverse influence as exemplified by the three case study zones located in the Saône, upper Arve and upper Ardèche catchments (see Fig. 3). The present work suggests that the performance of any perfect prognosis downscaling method using a common predictor domain is far from optimal for individual locations in France as a consequence of the assumption of a common predictor domain, as shown in Fig. 5 for the analogue downscaling method used here. This may be specifically the case for the method developed by Boé et al. (2006), which was later extended to the whole of France by Boé and Terray (2008), Pagé et al. (2008) and Pagé and Terray (2010). This method has been used in many subsequent national-scale climate change impact studies on hydrology (see, e.g. Boé et al., 2009; Vidal et al., 2012), and downscaled products are now disseminated through a national climate service platform built in the DRIAS project (Lémond et al., 2011). This issue of a common predictor domain thus provides some explanation for the identified biases (Boé, 2007) and weak correlations (Boé and Terray, 2008) in downscaled precipitation outputs for regions around the Mediterranean coast. Indeed, as shown in Figs. 6 and 7, the optimum geopotential predictor domains for these regions are quite different from the rest of the country.

6.3 Towards predictand areas with homogeneous predictors

The spatial coherence of the downscaled precipitation is often taken as a given when using the analogue method, but this is only true if the same analogue dates are found for the whole target region, which is not guaranteed if different subtarget regions are using different predictor domains. On the other hand, if the target region is large, like a large river basin, a common predictor domain is likely to be suboptimal on the local scale as the best domains differ for the subcatchments, as shown in this study, for example, for the Rhône catchment.

Despite the simplicity of the concept, the analogue method has a large number of parameters: the predictor variables and their spatial and temporal domains, the similarity criteria and the number of analogues. A global optimisation of all these parameters together is desirable but involves high computational costs. In this work the optimisation was restricted to the horizontal domains of the geopotential predictor but was performed for a large number of predictand zones.

Using individual predictor domains for each zone will in general result in different analogue dates, thereby not ensuring systematically the spatial coherence it has if a common domain is used for all predictand locations. Therefore it will be beneficial to group zones together that can use the same parameters, i.e. the same geopotential predictor domain. The presented analysis will help to this end by building on the idea of grouping zones by equal domains in the five-domain ensemble, as equal optimised predictor domains reflect proximity and similar flow exposure. To this end, for each zone, one predictor domain from the five-domain ensemble has to be selected such that contiguous areas with the same predictor domain are formed. The smooth distribution of the domain centre locations together with their rather small range should facilitate this. The domain size has a higher spatial variability that could hamper the attempt of aggregating zones by same domain, but this goes along with a larger range that may compensate it to a certain degree.

Acknowledgements. The authors would like to thank Météo-France for providing access to the Safran database and ECMWF for providing access to the ERA-40 data. We would like to thank Pascal Horton, Tobias Sauter and an anonymous reviewer for their constructive comments on the discussion paper that helped to significantly improve the manuscript.

Edited by: F. Pappenberger

References

Aspelien, T., Iversen, T., Bremnes, J. B., and Frogner, I.-L.: Short-range probabilistic forecasts from the Norwegian limited-area EPS: long-term validation and a polar low study, Tellus A, 63, 564–584, doi:10.1111/j.1600-0870.2010.00502.x, 2011.

Auffray, A., Clavel, A., Jourdain, S., Ben Daoud, A., Sauquet, E., Lang, M., Obled, C., Panthou, G., Gautheron, A., Gottardi, F., and Garçon, R.: Reconstitution hydrométéorologique de la crue de l'Isère de 1859 (Reconstructing the hydrometeorological scenario of the 1859 flood of the Isère river), La Houille Blanche, 44–50, doi:10.1051/lhb/2011005, 2011.

Auguie, B.: gridExtra: functions in Grid graphics, r package version 0.9.1, http://CRAN.R-project.org/package=gridExtra, last access: 24 October 2013, 2012.

Beaulant, A.-L., Joly, B., Nuissier, O., Somot, S., Ducrocq, V., Joly, A., Sevault, F., Deque, M., and Ricard, D.: Statistico-dynamical downscaling for Mediterranean heavy precipitation, Q. J. Roy. Meteorol. Soc., 137, 736–748, doi:10.1002/qj.796, 2011.

Bellone, E., Hughes, J. P., and Guttorp, P.: A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts, Clim. Res., 15, 1–12, 2000.

Ben Daoud, A.: Améliorations et développements d'une méthode de prévision probabiliste des pluies par analogie, Application à la prévision hydrologique sur les grands bassins fluviaux de la Saône et de la Seine, Thèse de doctorat, Université Joseph Fourier, Grenoble, 2010.

Ben Daoud, A., Sauquet, E., Lang, M., Obled, C., and Bontron, G.: La prévision des précipitations par recherche d'analogues: etat de l'art et perspectives (Precipitation forecasting through an analog sorting technique: state of the art and further investigations), La Houille Blanche, 6, 60–65, doi:10.1051/lhb/2009079, 2009.

Ben Daoud, A., Sauquet, E., Lang, M., Bontron, G., and Obled, C.: Precipitation forecasting through an analog sorting technique: a comparative study, Adv. Geosci., 29, 103–107, doi:10.5194/adgeo-29-103-2011, 2011a.

Ben Daoud, A., Sauquet, E., Lang, M., and Ramos, M.-H.: Peut-on étendre l'échéance de prévision des crues en optimisant la prévision de pluies par recherche d'analogues? Application au bassin de la Seine à Paris (Can we extend flood forecasting lead-time by optimising precipitation forecasting based on analogs? Application to the Seine river basin), La Houille Blanche, 37–43, doi:10.1051/lhb/2011004, 2011b.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Bivand, R. S., Pebesma, E. J., and Gomez-Rubio, V.: Applied spatial data analysis with R, http://www.asdar-book.org/, Springer, New York, NY, 2008.

Boé, J.: Changement global et cycle hydrologique : Une étude de régionalisation sur la France, Thèse de doctorat, 3, http://thesesups.ups-tlse.fr/227/, Université Toulouse, 2007.

Boé, J. and Terray, L.: A weather-type approach to analyzing winter precipitation in France: Twentieth-century trends and the role of anthropogenic forcing, J. Climate, 21, 3118–3133, doi:10.1175/2007JCLI1796.1, 2008.

Boé, J., Terray, L., Habets, F., and Martin, E.: A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling, J. Geophys. Res., 111, D23106, doi:10.1029/2005JD006889, 2006.

Boé, J., Terray, L., Habets, F., and Martin, E.: Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies, Int. J. Climatol., 27, 1643–1655, doi:10.1002/joc.1602, 2007.

Boé, J., Terray, L., Martin, E., and Habets, F.: Projected changes in components of the hydrological cycle in French river basins during the 21st century, Water Resour. Res., 45, W08426, doi:10.1029/2008WR007437, 2009.

Bontron, G.: 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, 2004.

Bontron, G. and Obled, C.: L'adaptation probabiliste des prévisions météorologiques pour la prévision hydrologique (A probabilistic adaptation of meteorological model outputs to hydrological forecasting), La Houille Blanche, 23–28, doi:10.1051/lhb:200501002, 2005.

Brigode, P., Bernardara, P., Gailhard, J., and Ribstein, P.: Optimization of the geopotential heights informatio used in a rainfall based weather patterns classification over Austria, Int. J. Climatol., 33, 1563–1573, doi:10.1002/joc.3535, 2013.

Brown, T. A.: Admissible scoring systems for continuous distributions, The Rand Paper Series P-5235, The Rand Corporation, Santa Monica, California, 1974.

Carreau, J. and Vrac, M.: Stochastic downscaling of precipitation with neural network conditional mixture models, Water Resour. Res., 47, W10502, doi:10.1029/2010WR010128, 2011.

Chandler, R. E.: On the use of generalized linear models for interpreting climate variability, Environmetrics, 16, 699–715, doi:10.1002/env.731, 2002.

Chen, J., Brissette, F. P., and Leconte, R.: A daily stochastic weather generator for preserving low-frequency of climate variability, J. Hydrol., 388, 480–490, doi:10.1016/j.jhydrol.2010.05.032, 2010.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteorol. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

Demargne, J., Brown, J., Liu, Y., Seo, D.-J., Wu, L., Toth, Z., and Zhu, Y.: Diagnostic verification of hydrometeorological and hydrologic ensembles, Atmos. Sci. Lett., 11, 114–122, doi:10.1002/asl.261, 2010.

Duband, D.: Prévision spatiale des hauteurs de précipitations journalières (A spatial forecast of daily precipitation heights), La Houille Blanche, 497–512, doi:10.1051/lhb/1981046, 1981.

Duffourg, F. and Ducrocq, V.: Origin of the moisture feeding the Heavy Precipitating Systems over Southeastern France, Nat. Hazards Earth Syst. Sci., 11, 1163–1178, doi:10.5194/nhess-11-1163-2011, 2011.

Friederichs, P. and Hense, A.: Statistical downscaling of extreme precipitation events using censored quantile regression, Mon. Weather Rev., 135, 2365–2378, doi:10.1175/MWR3403.1, 2007.

Garavaglia, F., Gailhard, J., Paquet, E., Lang, M., Garçon, R., and Bernardara, P.: Introducing a rainfall compound distribution model based on weather patterns sub-sampling, Hydrol. Earth Syst. Sci., 14, 951–964, doi:10.5194/hess-14-951-2010, 2010.

Guilbaud, S. and Obled, C.: Prévision quantitative des précipitations journalièes par une technique de recherche de journées antérieures analogues: optimisation du critère d'analogie (Daily quantitative precipitation forecast by an analogue technique: optimisation of the analogy criterion), Comptes Rendus de l'Académie des Sciences – Series IIA, Earth Planet. Sci., 327, 181–188, doi:10.1016/S1251-8050(98)80006-2, 1998.

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., and Herrera, S.: Reassessing statistical downscaling techniques for their robust application under climate change conditions, J. Climate, 26, 171–188, doi:10.1175/JCLI-D-11-00687.1, 2013.

Hagedorn, R., Hamill, T. M., and Whitaker, J. S.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts, Part I: Two-meter temperatures, Mon. Weather Rev., 136, 2608–2619, doi:10.1175/2007MWR2410.1, 2008.

Hamill, T. M. and Whitaker, J. S.: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application, Mon. Weather Rev., 134, 3209–3229, doi:10.1175/MWR3237.1, 2006.

Hanssen-Bauer, I., Førland, E. J., Haugen, J. E., and Tveito, O. E.: Temperature and precipitation scenarios for Norway: comparison of results from dynamical and empirical downscaling, Clim. Res., 25, 15–27, doi:10.3354/cr025015, 2003.

Harrower, M. and Brewer, C. A.: ColorBrewer.org: An online tool for selecting colour schemes for maps, Cartogr. J., 40, 27–37, doi:10.1179/000870403235002042, 2003.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems, Weather Forecast., 15, 559–570, 2000.

Hertig, E., Paxian, A., Vogt, G., Seubert, S., Paeth, H., and Jacobeit, J.: Statistical and dynamical downscaling assessments of precipitation extremes in the Mediterranean area, Meteorol. Z., 21, 61–77, doi:10.1127/0941-2948/2012/0271, 2012.

Horton, P.: Améliorations et optimisation globale de la méthode des analogues pour la prévision statistique des précipitations, Développement d'un outil de prévision et application opérationnelle au bassin du Rhône à l'amont du Léman, Thèse de doctorat, Université de Lausanne, Switzerland, 2012.

Horton, P., Jaboyedoff, M., Metzger, R., Obled, C., and Marty, R.: Spatial relationship between the atmospheric circulation and the precipitation measured in the western Swiss Alps by means of the analogue method, Nat. Hazards Earth Syst. Sci., 12, 777–784, doi:10.5194/nhess-12-777-2012, 2012.

Ibarra-Berastegi, G., Saénz, J., Ezcurra, A., Elías, A., Diaz Argandoña, J., and Errasti, I.: Downscaling of surface moisture flux and precipitation in the Ebro Valley (Spain) using analogues and analogues followed by random forests and multiple linear regression, Hydrol. Earth Syst. Sci., 15, 1895–1907, doi:10.5194/hess-15-1895-2011, 2011.

Kallache, M., Vrac, M., Naveau, P., and Michelangeli, P.-A.: Non-stationary probabilistic downscaling of extreme precipitation, J. Geophys. Res., 116, D05113, doi:10.1029/2010JD014892, 2011.

Kalnay, E., Kanamitsu, M., Kistler, R. W. C., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janoviak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-year Reanalysis Project, B. Am. Meteorol. Soc., 77, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrolog. Sci. J., 31, 13–24, 1986.

Lavaysse, C., Vrac, M., Drobinski, P., Lengaigne, M., and Vischel, T.: Statistical downscaling of the French Mediterranean climate: assessment for present and projection in an anthropogenic scenario, Nat. Hazards Earth Syst. Sci., 12, 651–670, doi:10.5194/nhess-12-651-2012, 2012.

Lémond, J., Dandin, P., Planton, S., Vautard, R., Pagé, C., Déqué, M., Franchistéguy, L., Geindre, S., Kerdoncuff, M., Li, L., Moisselin, J.-M., Noël, T., and Tourre, Y. M.: DRIAS: a step toward Climate Services in France, Adv. Sci. Res., 6, 179–186, doi:10.5194/asr-6-179-2011, 2011.

Liu, X. and Coulibaly, P.: Downscaling ensemble weather predictions for improved week-2 hydrologic forecasting, J. Hydrometeorol., 12, 1564–1580, doi:10.1175/2011JHM1366.1, 2011.

Lorenz, E. N.: Atmospheric predictability as revealed by naturally occurring analogues, J Atmos Sci, 26, 636–646, doi:10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2, 1969.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I.: Precipitation downscaling under climate change. Recent developments to bridge the gap between dynamical models and the end user, Rev. Geophys., 48, RG3003, doi:10.1029/2009RG000314, 2010.

Martin, E., Timbal, B., and Brun, E.: Downscaling of general circulation model outputs: simulation of the snow climatology of the French Alps and sensitivity to climate change, Clim. Dynam., 13, 45–56, doi:10.1007/s003820050152, 1997.

Marty, R., Zin, I., and Obled, C.: On adapting PQPFs to fit hydrological needs: the case of flash flood forecasting, Atmos. Sci. Lett., 9, 73–79, 2008.

Marty, R., Zin, I., Obled, C., Bontron, G., and Djerboua, A.: Toward real-time daily PQPF by an analog sorting approach: Application to flash-flood catchments, J. Appl. Meteorol. Clim., 51, 505–520, doi:10.1175/JAMC-D-11-011.1, 2012.

Marty, R., Zin, I., and Obled, C.: Sensitivity of hydrological ensemble forecasts to different sources and temporal resolutions of probabilistic quantitative precipitation forecasts: flash flood case studies in the Cévennes-Vivarais region (Southern France), Hydrol. Process., 27, 33–44, doi:10.1002/hyp.9543, 2013.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Manage Sci., 22, 1087–1095, 1976.

Matulla, C., Zhang, X., Wang, X. L., Wang, J., Zorita, E., Wagner, S., and von Storch, H.: Influence of similarity measures on the performance of the analog method for downscaling daily precipitation, Clim. Dynam., 30, 133–144, doi:10.1007/s00382-007-0277-2, 2008.

Muluye, G. Y.: Deriving meteorological variables from numerical weather prediction model output: A nearest neighbor approach, Water Resour. Res., 47, W07509, doi:10.1029/2010WR009750, 2011.

Nam, D. H., Udo, K., and Mano, A.: Downscaling global weather forecast outputs using ANN for flood prediction, J. Appl. Math., 2011, 246–286, doi:10.1155/2011/246286, 2011.

Neuwirth, E.: RColorBrewer: ColorBrewer palettes, http://CRAN.R-project.org/package=RColorBrewer (last access: 24 October 2013), 2011.

Nuissier, O., Joly, B., Joly, A., Ducrocq, V., and Arbogast, P.: A statistical downscaling to identify the large-scale circulation patterns associated with heavy precipitation events over southern France, Q. J. Roy. Meteorol. Soc., 137, 1812–1827, doi:10.1002/qj.866, 2011.

Obled, C., Bontron, G., and Garçon, R.: Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach, Atmos. Res., 63, 303–324, doi:10.1016/S0169-8095(02)00038-8, 2002.

Pagé, C. and Terray, L.: Nouvelles projections climatiques à échelle fine sur la France pour le 21ème siècle : les scénarii SCRATCH2010, Technical Report TR/CMGC/10/58, CERFACS, Toulouse Cedex, France, 2010.

Pagé, C., Terray, L., and Boé, J.: Projections climatiques à échelle fine sur la France pour le 21ème siècle: les scénarii SCRATCH08, Technical note, Climate Modelling and Global Change TR/CMGC/08/64, CERFACS, Toulouse, France, 2008.

Pebesma, E. and Bivand, R.: Classes and methods for spatial data in R, r news 5th Edn., http://cran.r-project.org/doc/Rnews/ (last access: 24 October 2013), 2005.

Pierce, D.: ncdf: Interface to Unidata netCDF data files, r package version 1.6.6, http://CRAN.R-project.org/package=ncdf (last access: 24 October 2013), 2011.

Pincus, R. and Rew, R.: NetCDF Fortran 90 Guide, netcdf 4.1.3 Edn., UCAR Unidata, Boulder, Colorado, 2011.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchistéguy, L., and Morel, S.: Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France, J. Appl. Meteorol. Clim., 47, 92–107, doi:10.1175/2007JAMC1636.1, 2008.

Quintana Seguí, P., Ribes, A., Martin, E., Habets, F., and Boé, J.: Comparison of three downscaling methods in simulating the impact of climate change on the hydrology of Mediterranean basins, J. Hydrol., 383, 111–124, doi:10.1016/j.jhydrol.2009.09.050, 2010.

Quintana-Seguí, P., Habets, F., and Martin, E.: Comparison of past and future Mediterranean high and low extremes of precipitation and river flow projected using different statistical downscaling methods, Nat. Hazards Earth Syst. Sci., 11, 1411–1432, doi:10.5194/nhess-11-1411-2011, 2011.

Raje, D. and Mujumdar, P. P.: A comparison of three methods for downscaling daily precipitation in the Punjab region, Hydrol. Process., 25, 3575–3589, doi:10.1002/hyp.8083, 2011.

R Development Core Team: R: A Language and Environment for Statistical Computing, http://www.R-project.org/, R Foundation for Statistical Computing, Vienna, Austria, 2012.

Ricard, D., Ducrocq, V., and Auger, L.: A climatology of the mesoscale environment associated with Heavily Precipitating Events over a Northwestern Mediterranean Area, J. Appl. Meteorol. Clim., 51, 468–488, doi:10.1175/JAMC-D-11-017.1, 2012.

Sauter, T. and Venema, V.: Natural three-dimensional predictor domains for statistical precipitation downscaling, J. Climate, 24, 6132–6145, doi:10.1175/2011JCLI4155.1, 2011.

Teuling, A. J., Stöckli, R., and Seneviratne, S. I.: Bivariate colour maps for visualizing climate data, Int. J. Climatol., 31, 1408–1412, doi:10.1002/joc.2153, 2011.

Teutschbein, C., Wetterhall, F., and Seibert, J.: Evaluation of different downscaling techniques for hydrological climate-change impact studies at the catchment scale, Clim. Dynam., 37, 2087–2105, doi:10.1007/s00382-010-0979-8, 2011.

Teweles, S. and Wobus, H. B.: Verification of prognostic charts, B. Am. Meteorol. Soc., 35, 455–463, 1954.

Themeßl, M. J., Gobiet, A., and Leuprecht, A.: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models, Int. J. Climatol., 31, 1530–1544, doi:10.1002/joc.2168, 2011.

Timbal, B. and McAvaney, B. J.: An analogue-based method to downscale surface air temperature: application for Australia, Clim. Dynam., 17, 947–963, doi:10.1007/s003820100156, 2001.

Timbal, B., Dufour, A., and McAvaney, B.: An estimate of future climate change for western France using a statistical downscaling technique, Clim. Dynam., 20, 807–823, doi:10.1007/s00382-002-0298-9, 2003.

Timbal, B., Hope, P., and Charles, S.: Evaluating the consistency between statistically downscaled and global dynamical model climate change projections, J. Climate, 21, 6052–6059, doi:10.1175/2008JCLI2379.1, 2008.

Turco, M., Quintana Seguí, P., Llasat, M. C., Herrera, S., and Gutiérrez, J. M.: Testing MOS precipitation downscaling for ENSEMBLES regional climate models over Spain, J. Geophys. Res., 116, D18109, doi:10.1029/2011JD016166, 2011.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Q. J. Roy. Meteorol. Soc., 131, 2961–3012, doi:10.1256/qj.04.176, 2005.

Vidal, J.-P. and Wade, S. D.: A framework for developing high-resolution multi-model climate projections: 21st century scenarios for the UK, Int. J. Climatol., 28, 843–858, doi:10.1002/joc.1593, 2008.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, Int. J. Climatol., 30, 1627–1644, doi:10.1002/joc.2003, 2010.

Vidal, J.-P., Martin, E., Kitova, N., Najac, J., and Soubeyroux, J.-M.: Evolution of spatio-temporal drought characteristics: validation, projections and effect of adaptation scenarios, Hydrol. Earth Syst. Sci., 16, 2935–2955, doi:10.5194/hess-16-2935-2012, 2012.

Voisin, N., Schaake, J. C., and Lettenmaier, D. P.: Calibration and downscaling methods for quantitative ensemble precipitation forecasts, Weather Forecast., 25, 1603–1627, doi:10.1175/2010WAF2222367.1, 2010.

Vrac, M., Stein, M., and Hayhoe, K.: Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing, Clim. Res., 34, 169–184, doi:10.3354/cr00696, 2007.

Wetterhall, F., Halldin, S., and Xu, C.-y.: Statistical precipitation downscaling in central Sweden with the analogue method, J. Hydrol., 306, 174–190, doi:10.1016/j.jhydrol.2004.09.008, 2005.

Wetterhall, F., Halldin, S., and Xu, C.-Y.: Seasonality properties of four statistical-downscaling methods in central Sweden, Theor. Appl. Climatol., 87, 123–137, doi:10.1007/s00704-005-0223-3, 2007.

Wickham, H.: Reshaping data with the reshape package, J. Stat. Softw., 21, 1–20, 2007.

Wickham, H.: ggplot2: elegant graphics for data analysis, Use R!, Springer, New York, doi:10.1007/978-0-387-98141-3, 2009.

Wilby, R. L. and Wigley, T. M. L.: Precipitation predictors for downscaling: observed and general circulation model relationships, Int. J. Climatol., 20, 641–661, 2000.

Wilby, R. L., Hassan, H., and Hanaki, K.: Statistical downscaling of hydrometeorological variables using general circulation model output, J. Hydrol., 205, 1–19, doi:10.1016/S0022-1694(97)00130-3, 1998.

Zeileis, A. and Grothendieck, G.: zoo: S3 infrastructure for regular and irregular time series, J. Stat. Softw., 14, 1–27, 2005.

Zhang, F. and Georgakakos, A. P.: Joint variable spatial downscaling, Climatic Change, 111, 945–972, doi:10.1007/s10584-011-0167-9, 2012.

5.2 A closer look at equifinality or the 99 domains experiment

In the last section the geopotential predictor domains have been optimised using five growing rectangular domains. But the number five is an arbitrary choice. So what happens if we choose a much higher number instead? With this experiment we want to assess how the relation between CRPS, domain size and aspect ratio behaves for many different domains for a zone where the near optimum domains show considerable differences in size and aspect ratio. Furthermore we want to see if the error for using too small domains and the one for too large domains behaves similarly or not.

For one zone, the Ardèche case study zone, the optimisation with the multiple growing rectangular domain algorithm is repeated with 99 domains instead of five. The number of 99 domains has been chosen because this is the maximum number that is possible in the current implementation of SANDHY. (This limitation is related to output filename format rather than calculation.) The Ardèche case study zone has been chosen, because it is a zone that is subject to high seasonality of precipitation and its near optimum domains show comparatively large differences in size and aspect ratio (see Section 5.1).

The resulting domains are characterised by their diagonal length and aspect ratio in the following. Both characteristics are defined in degrees longitude and latitude.

$$diagonal = \sqrt{(lon_{max} - lon_{min})^2 + (lat_{max} - lat_{min})^2} \tag{5.1}$$

$$log(aspectratio) = log\left(\frac{lon_{max} - lon_{min}}{lat_{max} - lat_{min}}\right) \tag{5.2}$$

The logarithm is taken in order to have equal distances between the ratios 1:2 - 1:1 and 1:1 - 2:1 on the plots.

Figure 5.1 shows the length of the diagonal, as an indication of domain size, in degrees on the x-axis and the CRPS on the y-axis. The smaller the CRPS the better, so the points corresponding to the best domains can be found near the bottom of the diagram. The best domain has a length of diagonal of about 25 degrees and the other points are rather symmetrically distributed around this value. The differences in the CRPS values are often very small with the total range of CRPS values smaller than 0.05mm day^{-1} . From this figure it does not seem that a larger than optimal domain leads to less increase of the CRPS than a smaller than optimal one. The five domains found using five growing rectangular domains are marked in red. The best two domains are the same as with the 99 domains but for the domains 3, 4 and 5 better ones have been found using 99 domains and they are all larger than the ones found using five growing domains.

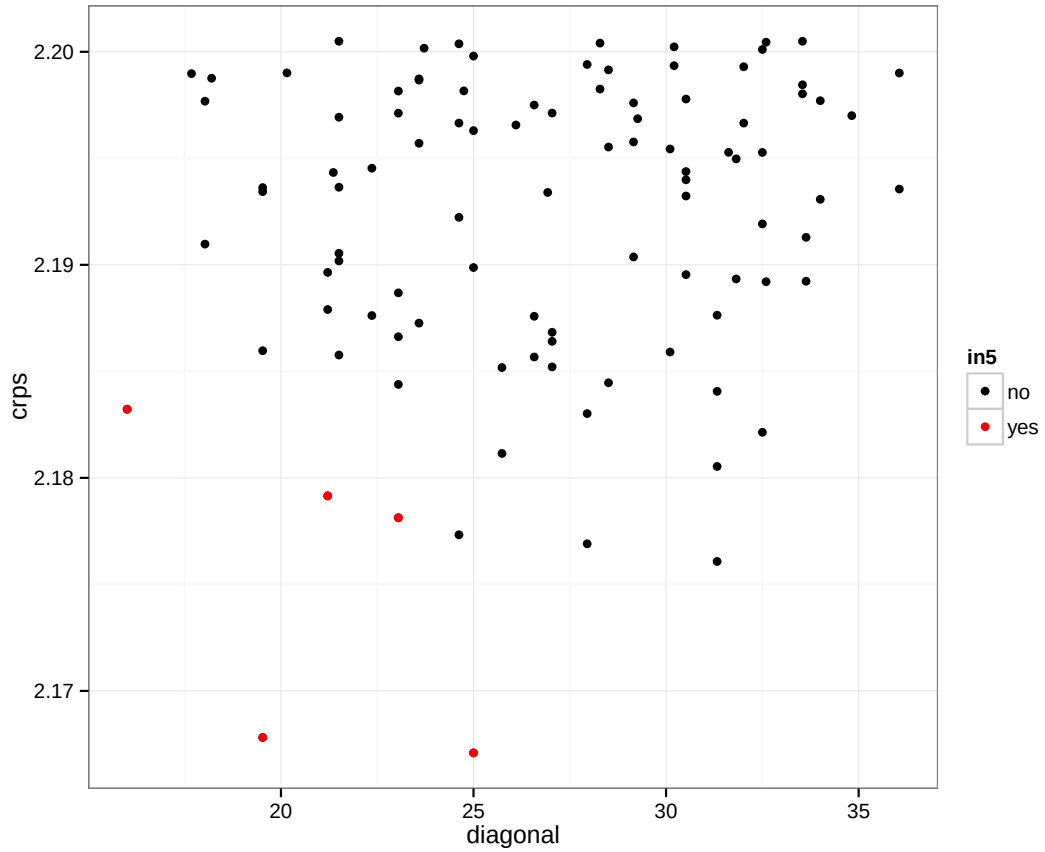

Figure 5.1: The length of the diagonal vs. the CRPS for the 99 best domains found in the optimisation procedure for the Ardèche zone. The archive length is 20 years and the starting point the nearest grid cell. In red the points found with N=5.

Figure 5.2 shows the logarithm of the aspect ratio of the 99 domains on the x-axis and the CRPS on the y-axis. The domains that had been found using five growing domains are marked in red while the others are shown in black. The point size corresponds to the length of the diagonal of the domain. For the aspect ratio a clear optimum and a symmetric distribution around this optimum can be observed. The two best domains have been found using five growing domains as well, but for the domains 3-5 smaller domains with smaller aspect ratios have been found. There is a gap in aspect ratio between the third best domain found with five growing domains and the one found with 99 growing domains. Obviously a precursor domain of the third best in the 99 domains experiment was not under the five best of the corresponding step in the five domains optimisation experiment

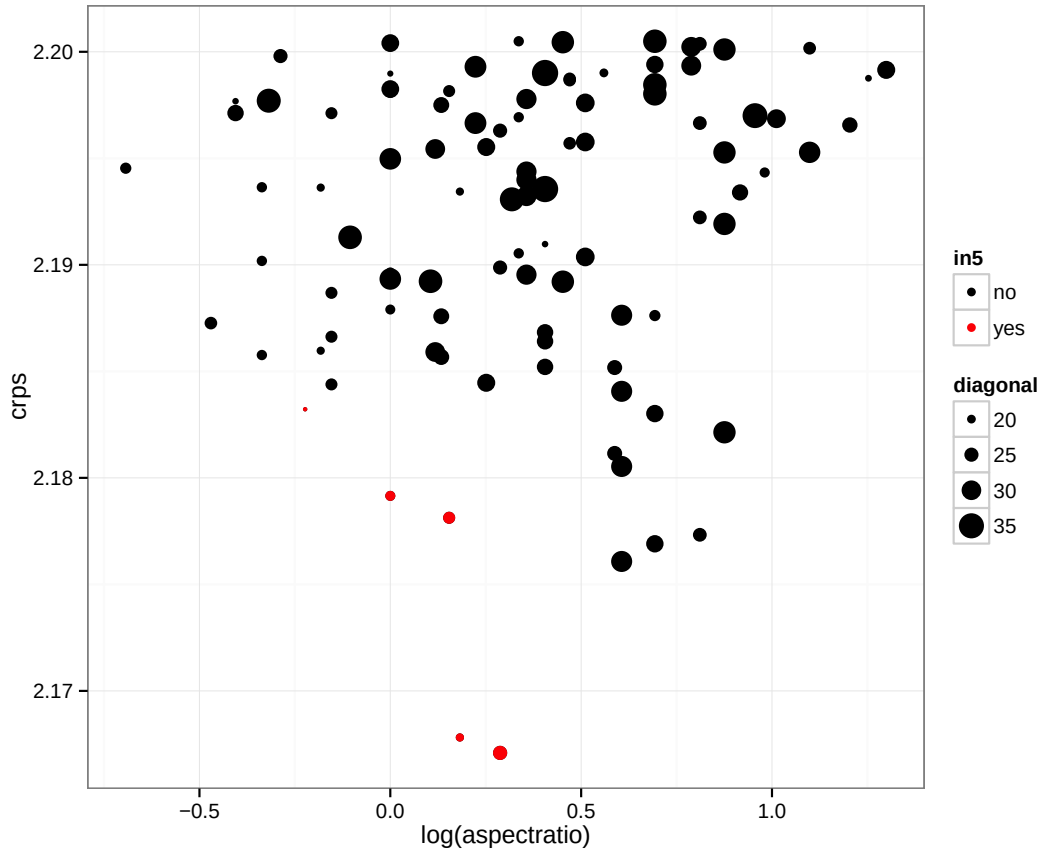

Figure 5.2: The logarithm of the aspect ratio vs. the CRPS for the 99 best domains found in the optimisation procedure for the Ardèche zone. The archive length is 20 years and the starting point the nearest grid cell. In red the points found with N=5. The size of the points corresponds to the length of the diagonal.

and has therefore been abandoned. Since the optimisation method can only add rows or columns of grid cells and not subtract some, these domain could not be found in the five domains experiment.

Figures 5.3, 5.4 and 5.5 show the location of the domain centers for the 99 domains for the Ardèche zone. The majority of the domain centers are located south-east of the zone. Domains further away from the target zone tend to have longer diagonals, specifically they are larger in east-west direction. Since the optimisation is started from the nearest 2x2 grid point domain from the target zone, a domain has to grow for example at least 10 degrees in one direction to move its center 5 degrees away from the target zone. The domain centers are situated on a 1.25 degree grid because the domain size is discrete with 2.5 degree steps due to

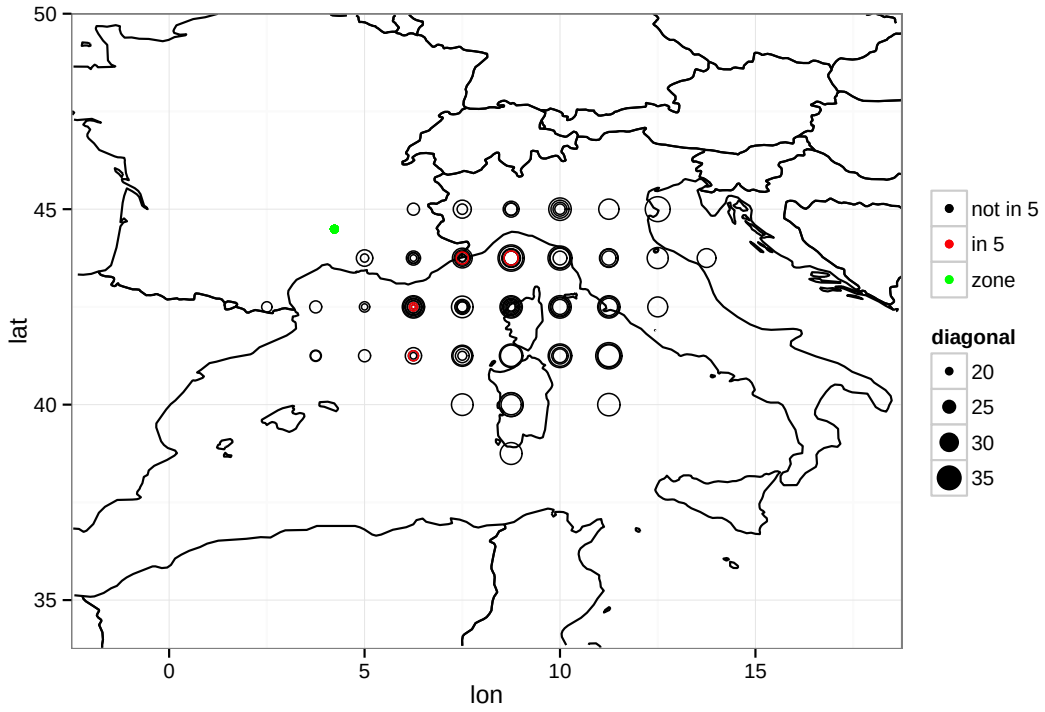

Figure 5.3: Domain center locations for the 99 domains optimised for the Ardèche zone. The zone is marked in green, the centers of the domains found with N=5 are drawn in red. The circle size corresponds to the length of the diagonal of the domain.

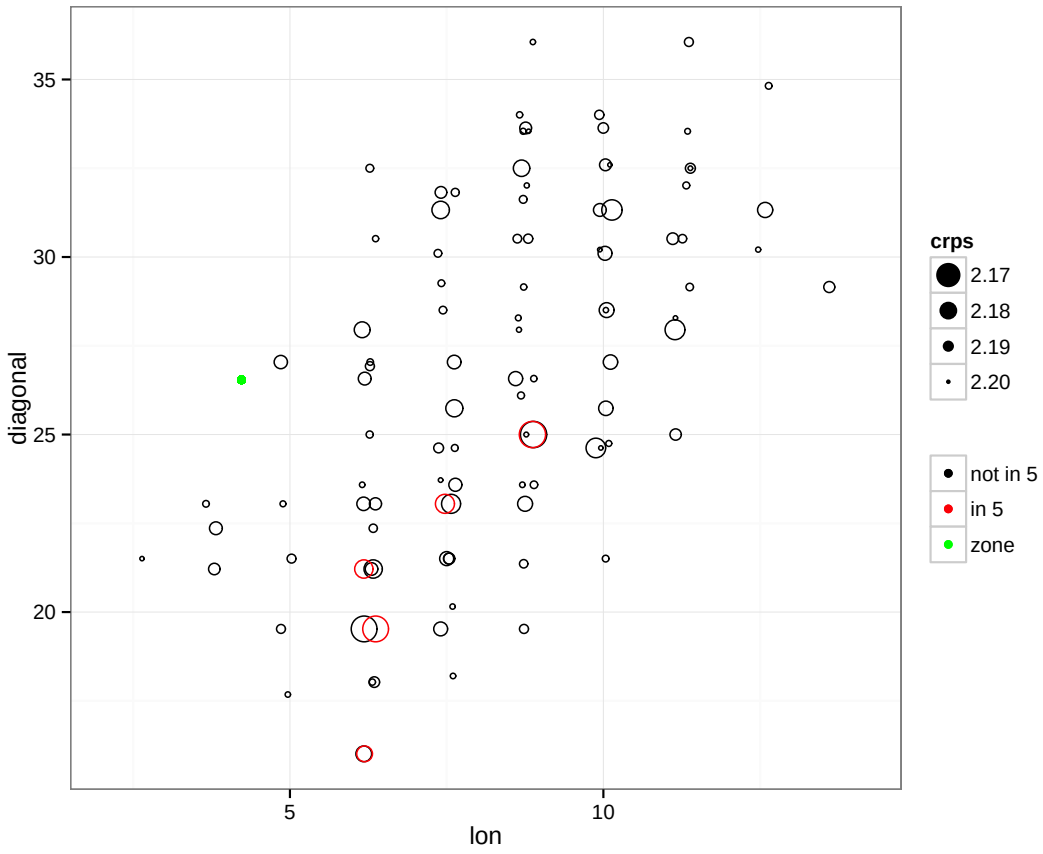

Figure 5.4: Domain center longitude vs. diagonal. Larger cercles correspond to smaller (that is better) CRPS values.

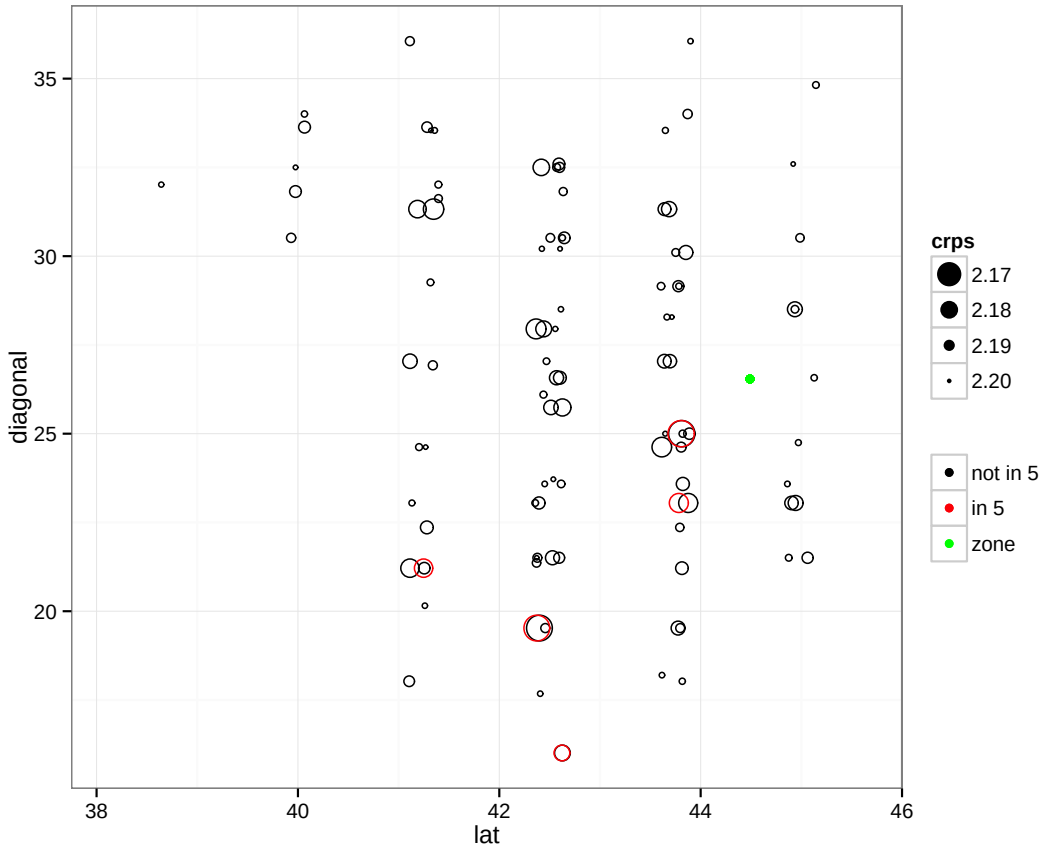

Figure 5.5: Domain center latitude vs. diagonal. Larger cercles correspond to smaller (that is better) CRPS values.

the resolution of the reanalysis data. On the scatter plots (figures 5.4 and 5.5) the points for the different domains are dispersed a bit to reduce overplotting. The centers of the five domains found in the original optimisation are marked in red. The location of their centers are quite representative but they have below average diagonal length.

In the case of the Ardèche case study zone the growing rectangular domain algorithm missed out some near optimum domains that are not explored using 5 domains but are found using a much higher number of domains that explores a larger part of the search space. This is related to the fact that the algorithm always adds rows or columns of grid points but can not subtract some. The hypothesis that taking too large domains has less effect on the skill than taking too small ones has not been confirmed.

5.3 Analysis of geopotential predictor fields or can we reconstruct a relevance map?

Looking at the relevance maps in Figures 3, 4, 8, 9 and 10 of the article in section 5.1, the question arises where their shape comes from. Some features like the region of highest skill that can be found a bit upstream with respect to the main flow direction in situations that lead to high precipitation at the zone in question seem rather natural, but why does the most relevant region tend to form a circle with lower skill in the center? Why is this lower skill region found in the middle of a higher skill region and does this shape correspond to something? Is there a way to obtain a relevance map in a "cheap" way?

5.3.1 Average geopotential fields

Let us have a look at average geopotential fields from ERA40 over the same time period (1 August 1982 – 31 July 2002) and the same time steps that were used to create the relevance maps. The first row of Figure 5.6 shows the mean geopotential height at 500 hPa for the 20-year period 1 August 1982 - 31 July 2002 at 00:00 UTC day D+1 (=+24h). Since there is no threshold that depends on the zone, the picture is the same for all three zones. The geopotential is higher in the south and lower in the north. The gradients in zonal direction are small. The average only over the days with precipitation at a given zone (second row, threshold 0.1mm day^{-1}) shows a very weak trough over western europe. For the average over days with high precipitation, that is above the 80% percentile, this trough becomes a little bit more intense. The orientation of the trough axis is slightly different for the Ardèche zone than for the other two where the pictures are very similar. There is nothing that corresponds to the circle or the center of the circle on the relevance maps.

Figure 5.7 shows the mean geopotential height at 1000 hPa over 20 years (1 August 1982 – 31 July 2002) at 12:00 UTC day D (=+12h). The first row is again the average over all days. Again the higher geopotential is in the south with the maximum in the south-west of the domain the lower geopotential is in the north with a minimum over Iceland. Here the geopotential gradient is stronger in the west of the domain than in the east. For the rainy days (second row) we can see a low over Italy for the Saône and the Arve zone and west of Italy for the Ardèche zone and maybe a second low over the north-sea. Again the structure gets more intense for the higher precipitation threshold (third row). Again there is nothing corresponding to the center of the cercle on the relevance maps.

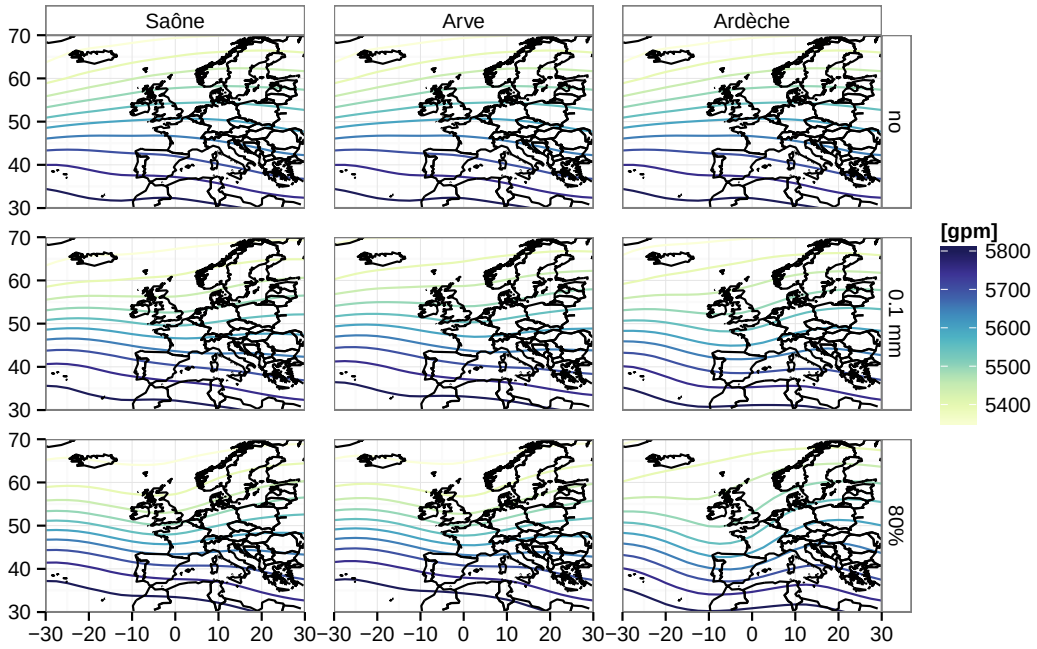

Figure 5.6: Mean geopotential height at 500 hPa over 20 years (1 August 1982 – 31 July 2002) at 00:00 UTC day D+1 (=+24h) for three zones and for all days (first row), for rainy days, threshold is 0.1mm day^{-1} (second row) and high precipitation days, threshold is the 80% percentile (third row).

5.3.2 Geopotential anomalies

The mean field over all days is subtracted from the fields with thresholds to look at the anomalies. Figures 5.8 and 5.9 show the mean geopotential height anomalies for rainy days and high precipitation days (same thresholds as above) for the 500 hPa and the 1000 hPa pressure level respectively. For example the first row in figure 5.8 is the difference between the second and the first row in Figure 5.6, that is the mean over all days is subtracted from the mean over the rainy days. Here more (500 hPa) or less (1000 hPa) circular shaped negative geopotential height anomalies can be seen, that are more intense for the higher precipitation threshold. Sanchez-Gomez and Terray (2005) found similar geopotential anomalies for days with intense precipitation events in western France. The centers, the place with the most negative anomaly, is more south-west for the Ardèche zone than for the two others and is not exactly at the same place for the two pressure levels. The center of the circle on the relevance maps is actually situated between the two,

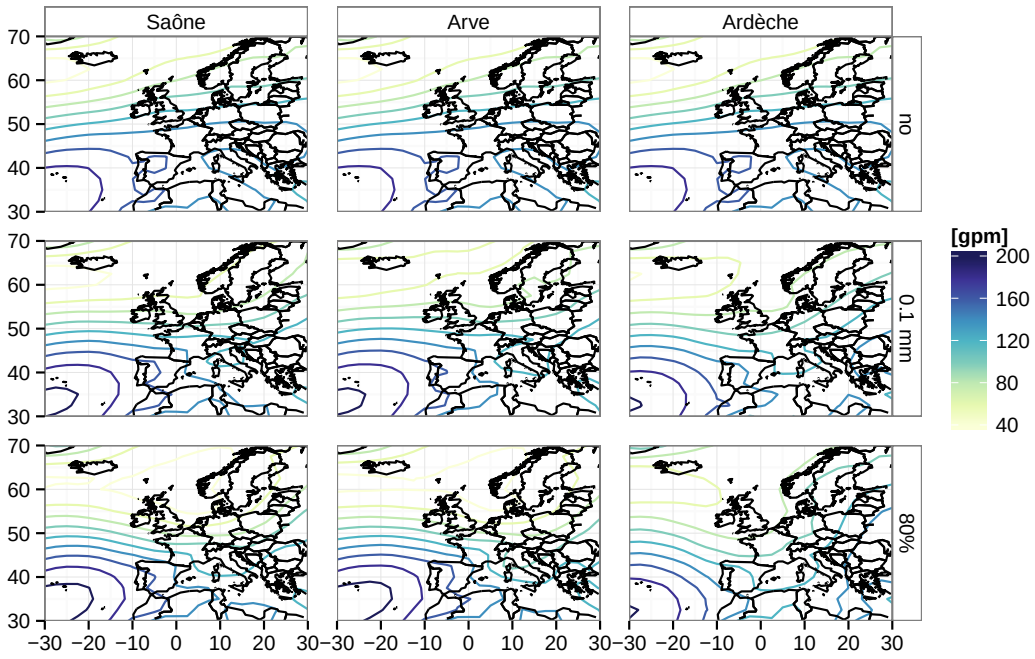

Figure 5.7: Mean geopotential height at 1000 hPa over 20 years (1 August 1982 – 31 July 2002) at 12:00 UTC day D (=+12h) for three zones and for all days (first row), for rainy days, threshold is 0.1mm day^{-1} (second row) and high precipitation days, threshold is the 80% percentile (third row).

which corresponds well to the fact that the relevance maps include both fields with equal weight.

So the center corresponds to the strongest negative geopotential anomaly, but why do the relevance maps show a local minimum of skill there?

5.3.3 Geopotential gradients

The TWS (Teweles and Wobus, 1954) that is used as distance measure for the geopotential predictor fields measures the gradients of the geopotential fields. From a meteorological point of view these gradients are strongly related to the geostrophic wind. The geopotential field is considered as a field of the stream function Ψ and the gradients of this stream function correspond to the geostrophic wind.

$$\begin{aligned} v_x &= \frac{\partial \Psi}{\partial y} \\ v_y &= -\frac{\partial \Psi}{\partial x} \end{aligned} \tag{5.3}$$

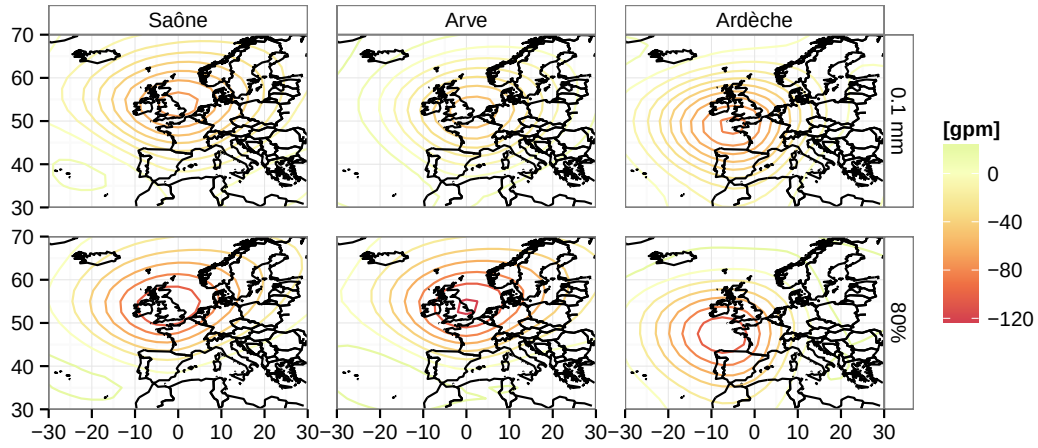

Figure 5.8: Mean geopotential height anomalies at 500 hPa over 20 years (1 August 1982 - 31 July 2002) at 00:00 UTC day D+1 (=+24h) for three zones at rainy days, threshold is 0.1mm day^{-1} (first row) and high precipitation days, threshold is the 80% percentile (second row).

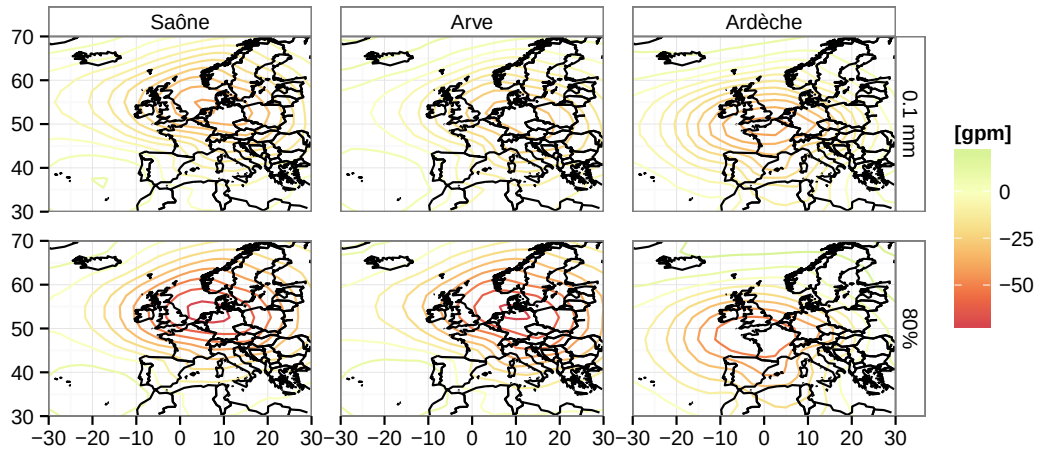

Figure 5.9: Mean geopotential height anomalies at 1000 hPa over 20 years (1 August 1982 - 31 July 2002) at 12:00 UTC day D (=+12h) for three zones at rainy days, threshold is 0.1mm day^{-1} (first row) and high precipitation days, threshold is the 80% percentile (second row).

where v_x and v_y are the components of the geostrophic wind field. To compute this from a gridded field a simple discretisation is applied:

$$\begin{aligned} v_x &= \frac{\Delta\Psi}{\Delta y} \\ &= \frac{\Psi(x_i,y_{j+1})-\Psi(x_i,y_{j-1})}{2\Delta y} \\ v_y &= -\frac{\Delta\Psi}{\Delta x} \\ &= -\frac{\Psi(x_{i+1},y_j)-\Psi(x_{i-1},y_j)}{2\Delta x} \end{aligned} \tag{5.4}$$

The gradients can be calculated for the mean field or for the anomaly field. Figure 5.10 shows the magnitude (M) of the gradient of the geopotential height anomalies at 500 hPa for rainy days and high precipitation days.

$$M_{anomaly} = \sqrt{v^2_{x,anomaly} + v^2_{y,anomaly}} \tag{5.5}$$

The arrows correspond to v_x and v_y interpreted as "wind components". The magnitude of the gradient shows a circle that is similar to the circular shape in the relevance maps but with higher values in the north compared to the relevance maps. This indicates that the circular shape of the high relevance region in the relevance maps is related to the gradients of the geopotential anomalies on rainy days, which is reasonable given the use of the TWS criteria. The relevance maps show a minimum of skill in the center of the cercle, because the gradients are less strong there.

We have found the reason for the circular shape, but can we find the maxima of the relevance maps by analysing the geopotential fields as well? The magnitude of the gradient M derived from the anomalies as shown before is multiplied with the sum of v_x and v_y derived from the mean geopotential field at rainy days and high precipitation days.

$$A = M_{anomaly} * (v_x + v_y) \tag{5.6}$$

Figure 5.11 shows this quantity A for the 500 hPa level and it can be seen that the shape for the high precipitation days resembles the one on the relevance maps. Compared to the real relevance maps there is a minor shift in the location of the maxima and too high values in the west while the eastward extension of the high skill region for the Ardèche zone is missing.

5.3.4 Vorticity

To find the missing east branch of the high skill region for the Ardèche zone the relative vorticity ζ, another quantity that is of interest in meteorology, is calculated.

$$\zeta = \frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \tag{5.7}$$

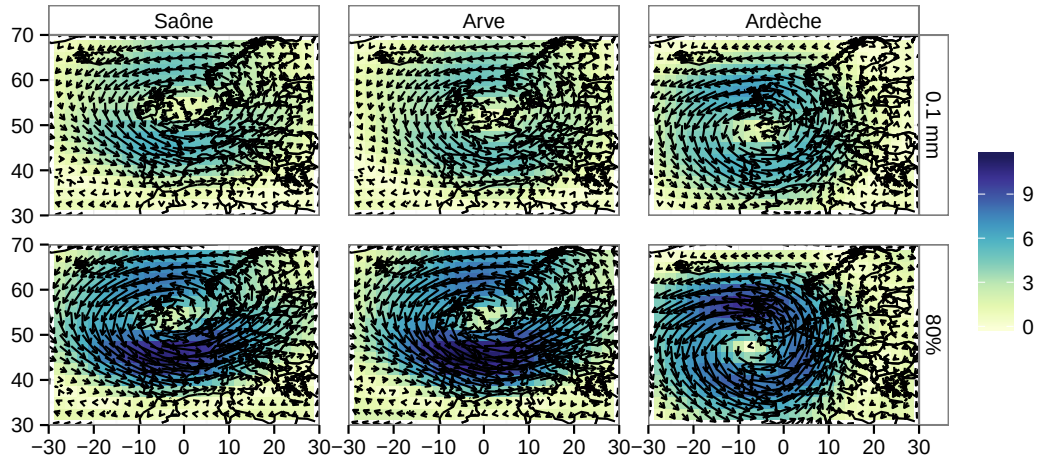

Figure 5.10: Gradient of mean geopotential height anomalies at 500 hPa over 20 years (1 August 1982 – 31 July 2002) at 00:00 UTC day D+1 (=+24h) for three zones at rainy days, threshold is 0.1mm day^{-1} (first row) and high precipitation days, theshold is the 80% percentile (second row). The colour correspond to the magnitude, the arrows the resulting anomaly of the geostrophic wind.

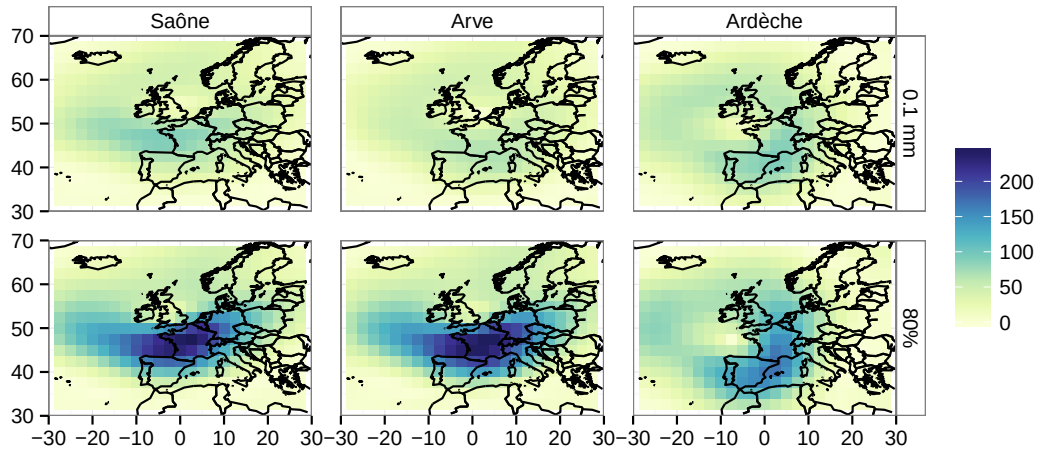

Figure 5.11: Magnitude of the gradient from the geopotential anomalies multiplied with the sum of the "wind components" derived from the mean fields on rainy days and high precipitation days at 500 hPa. (Times and thresholds are the same as for the previous figures.)

With discretisation we obtain:

$$\begin{aligned} \zeta &= \frac{\Delta v_y}{\Delta x} - \frac{\Delta v_x}{\Delta y} \\ &= \frac{v_y(x_{i+1},y_j) - v_y(x_{i-1},y_j)}{2\Delta x} - \frac{v_x(x_i,y_{j+1}) - v_x(x_i,y_{j-1})}{2\Delta y}. \end{aligned} \tag{5.8}$$

Figure 5.12 shows the relative vorticity calculated from the anomalies at 1000 hPa for rainy days and high precipitation days. An interesting detail on the map for high precipitation and the Ardèche zone is the large negative vorticity anomaly north of the Alps that corresponds to the "high skill" region on the relevance map that is missing in Figure 5.11.

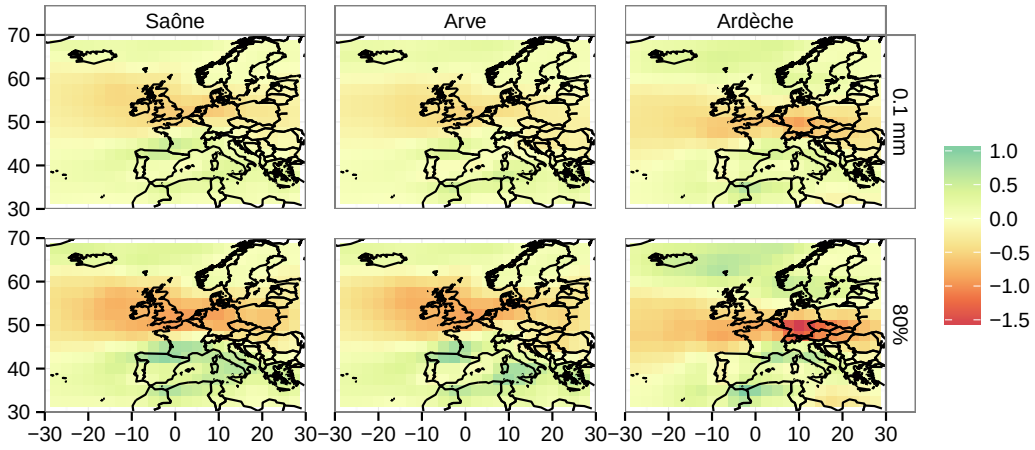

Figure 5.12: Vorticity ζ of the geopotential anomalies at 1000 hPa for rainy days and high precipitation days. (Times and thresholds like in figure 5.9.)

5.4 Conclusions

The domains resulting from the optimisation with an extended version of the growing rectangular algorithm differ moderately between neighbouring locations and inside the ensemble for a given zone. In some regions like Brittany larger regions were found with the same optimised domains while especially the Rhône catchment is characterised by high variability in the location and even more in the size of the optimised predictor domains. This high variability suggests that the assumption of a unique predictor domain for large parts of France is questionable, at least when one seeks to obtain the most skillful method for each individual zone. The optimum geopotential predictor domains for regions around the Mediterranean coast are for example quite different from the rest of the country.

The differences in skill measured with the CRPSS between the best and the fifth best domain found is one order of magnitude smaller than the differences in skill between different zones which makes all five domains equally plausible.

Even using multiple growing rectangular domains instead of a single one for the optimisation of the predictor domains of the geopotential shape predictor, the growing rectangular domain algorithm rests a simple and therefore rather fast optimisation algorithm. This allows the optimisation to be performed for a high number of different locations, but it comes at the cost of some limitations. One limitation is related to the insufficient exploration of the search space as has been demonstrated in the 99 domain experiment in Section 5.2. A second limitation is the choice of the starting point for the optimisation as has been mentioned for the Saône case study zone. Better downscaling skill in terms of CRPSS would be possible for this zone with domains similar to those found for the zones further south. This would require to use a different starting domain or to be able to subtract rows of grid points of the predictor domain during the optimisation process instead of always adding some. The larger than optimal domains found for a couple of zones concerned with this issue lead to a break-line in the mean domain size graph (Figure 7a of the article in Section 5.1). A third limitation arises from the decision not to optimise the predictor domains for the other three predictor variables at all. The last two limitations are further discussed in Chapter 7.

The center of the circle shape on the relevance maps corresponds to the center of the negative mean geopotential anomaly on rainy days or heavy precipitation days. The circular shape can be recovered by the magnitude of the gradient of the mean geopotential anomaly field. There are apparent similarities between the magnitude of the gradient of geopotential anomalies multiplied with the sum of the geostrophic wind components from the mean geopotential on high precipitation days at 500 hPa and the relevance maps. This is due to the use of the TWS, that uses essentially the same information: the magnitude and the direction of the geopotential gradients. The north-eastward extension of the high skill region on the relevance map for the Ardèche zone is not recovered by this gradient and wind map at 500 hPa. This seems to be rather due to features in the 1000 hPa field. Notably this region is characterised by a strong negative anomaly of relative vorticity at 1000 hPa. If only the maximum value is of interest, the map of the product of the gradients of the anomaly field for high precipitation days with the geostrophic wind sum from the mean field can serve as an approximation of a relevance map. The actual relevance maps combine information from both pressure levels and from all days without a precipitation threshold. It is therefore probably possible to get something even closer to a relevance map by combining quantities from both pressure levels and maybe different precipitation thresholds. In practice this

is not that easy, because for example the different quantities have very different magnitudes and even the same quantity can have different magnitudes on different pressure levels. Therefore some scaling would be required.

6 Validation of SANDHY

Having optimised the predictor domains of our downscaling method, it is important to see how the method with its parameters – the domain definition – behaves for independent data, that is data that has not been used during the optimisation process. The question is: are the parameters that work well during optimisation actually valid? Do they represent the general characteristics we want them to represent or are they too strongly fitted to the optimisation period? We can think of validation as a kind of crash test for models: we take a model (SANDHY), put it in some situations it was not constructed for (independent data), note its degradation (in terms of performance) and report the related implications for applications. This gives valuable information to later users of the model about what they can expect it to do well or not so well. Parts of this chapter have been presented as a poster at the EGU General Assembly 2014 (Radanovics et al., 2014).

6.1 Validation periods

Out of sample validation tests the stationarity assumption inherent to all downscaling methods (Frost et al., 2011) and is needed to ensure robust results as pointed out by Maraun et al. (2010). The choice of the length of calibration and validation periods is often restricted by the length of the available datasets. For example Hughes and Guttorp (1999); Bellone et al. (2000); Timbal and McAvaney (2001); Piani et al. (2010); Lavaysse et al. (2012); Souvignet and Heinrich (2011) use periods of not more than 12 years.

The division of the available data into calibration and validation periods can be a difficult task because on one hand sufficient data for calibration is needed to get adequate parameters and robust results. Specific to the analogue method the calibration period has to provide a reasonably large pool of analogue dates. On the other hand the validation period has to be long enough to ensure a stable calculation of the validation scores. If authors do not choose periods of equal

length, usually the calibration period is chosen to be longer than the validation period, for example Vrac and Yiou (2010) and Khan et al. (2006) use a 30-year learning period and a 15-year and 10-year validation period respectively, Bontron (2004) 44 years for calibration and 5 years for validation, Kallache et al. (2011) 35 years of calibration and 14 years for validation, Carreau and Vrac (2011) 25 years for training and 21 years for validation. An exception is Hanssen-Bauer et al. (2003) who used 61 years for training and 38 years for validation of precipitation downscaling models but 30 years calibration and 61 years validation periods for temperature downscaling, and Frost et al. (2011) who used 20 years for calibration and 25 years for validation.

Finally one has to decide if one wants to validate the performance in a similar climate or in a slightly different climate. For example Ben Daoud (2010) chose to calibrate on the period 1972–2002 excluding the years 1978, 1983, 1988, 1993, 1998 for validation, where the excluded years are chosen to represent as well as possible the 1972-2002 climate. Haylock et al. (2006) and Yang et al. (2010) placed their 15-year validation period in the middle of a 28 years or 26 years calibration period respectively. If one wants to test the ability of the method to adapt at a climate that is slightly different from the calibration climate the more common separated periods set-up is used. This is especially important when the downscaling method is to be applied in a climate change context. The SANDHY method is here validated on a completely independent period, that is a time period that has neither been used for optimising the predictor domains nor as an archive.

Here it is important to have two periods of the same length, because the skill of the analogue method strongly depends on the archive length (Radanovics et al., 2013a) and we want to test the influence of an alternative archive as compared to suboptimal predictor domains separately as will be described in more detail later in section 6.2. Periods of the same length are used for example in Vrac et al. (2007) and Orlowsky et al. (2010) and in Beuchat et al. (2012). The predictor domain optimisation has been performed for the 1 Aug. 1982 – 31 July 2002 period that is called *late* period in the following, the independent period from 1 Aug. 1958 – 31 July 1978 is referred to as *early*. These are the same periods as used for the difference maps in section 2.4 where it has been shown that there are indeed some slight differences between the two periods in terms of mean daily precipitation, high percentiles and percentage of dry days.

6.2 Validation experiments

A specificity of the analogue method is that it has an archive period, that is the pool from which analogue days can be selected. During the optimisation the archive period is equal to the simulation period with some days around the currently

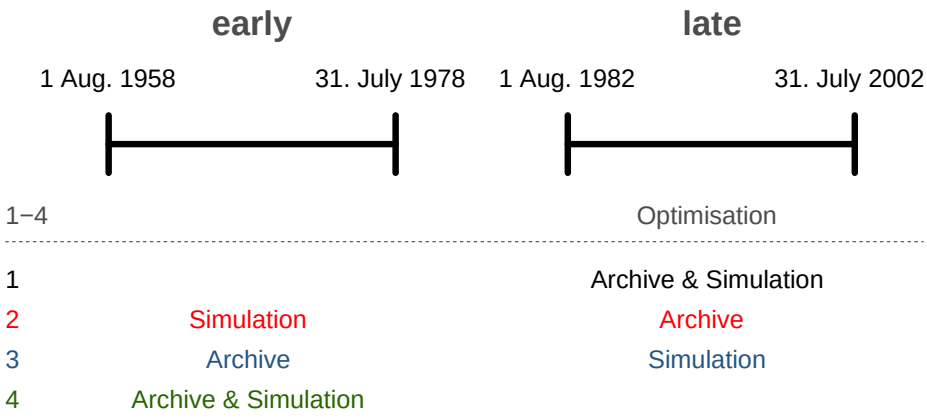

Figure 6.1: Time periods used in the validation experiments. 1 is the reference experiment, 2 the out of sample validation, 3 the alternative archive experiment and 4 the suboptimal predictor domains experiment.

simulated day excluded. It should be noted that for the validation experiments only the first of the optimised domains for each zone is used. Four experiments are conducted:

Reference experiment (simulation 1) against which the performance of the downscaling in the validation experiments is compared is a simulation under optimisation conditions. This means that the archive period and the simulation period are equal to the optimisation period, that is the *late* period.

Out of sample validation (simulation 2) is a simulation of the *early* period with the *late* period as archive. This set-up simulates realistic circumstances in reconstruction applications.

Alternative archive experiment (simulation 3) is a simulation of the *late* period but with the *early* period as archive, thus testing the influence of an alternative archive.

Suboptimal predictor domain experiment (simulation 4) is a simulation of the *early* period with the *early* period as archive, thus using the perfect archive for the simulation and testing the influence of suboptimal predictor domains that were optimised for the *late* period.

Figure 6.1 summarises the validation setup. The experiments 2, 3 and 4 are compared below to the reference simulation 1 in terms of CRPSS and bias. The

CRPS_{clim} for each of the experiments is calculated using the climatology from the archive period to predict the simulation period. The CRPSS always express the skill with respect to the climatological "model" reference.

6.3 Validation Results

6.3.1 CRPS and CRPSS

6.3.1.1 Reference simulation

First we are interested in the spatial distribution of the CRPS and the CRPSS for the reference simulation shown in Figures 6.2 and 6.3. The CRPS and CRPSS show a distribution related to the precipitation regime. The spatial distribution of the CRPS is very similar to the one of the climatological CRPS (compare Figure 3.1) with larger errors where the precipitation amounts are higher (compare Figure 2.3). At the same time these are the regions where we find the largest improvement over the climatological CRPS as can be seen on the map for the CRPSS (Figure 6.3 and Figure 5 in the article of Section 5.1). As has already been discussed in the article, the highest skill in terms of CRPSS can be found in regions with high mean precipitation for example in the Alps, the Cévennes, the western side of the Massif Central and the Vosges as well as at the Atlantic coast. The lowest skill is found along the Mediterranean coast, the eastern side of Corsica and the eastern side of the Massif Central.

6.3.1.2 Validation experiments

Next we want to know how much skill is lost if an independent period is simulated. Therefore the differences in CRPSS between the experiments described in Section 6.2 are shown. Figure 6.4 shows the CRPSS of the reference experiment on the x-axis and the CRPSS of the other experiments on the y-axis for all 608 zones. The upper left subfigure with the red dots correspond to the out-of-sample validation experiment. For most of the zones some skill loss is observed compared to the reference simulation (points below the first diagonal), about 0.03 on average. This skill loss is of the same order of magnitude as the difference between using (1) locally optimised domains or (2) one domain optimised for the France average precipitation, as shown in Section 3.2.1 of the article in Section 5.1. For a few zones we see a very small increase of skill in the out-of-sample validation experiment compared to the reference experiment (points above the first diagonal) and for one zone a dramatic skill loss of 0.17 can be seen. The zone in question is zone 367 with a large difference in average daily precipitation between the *late* and the *early* period as shown in Figure 2.9. The zone is under the most skillful ones in

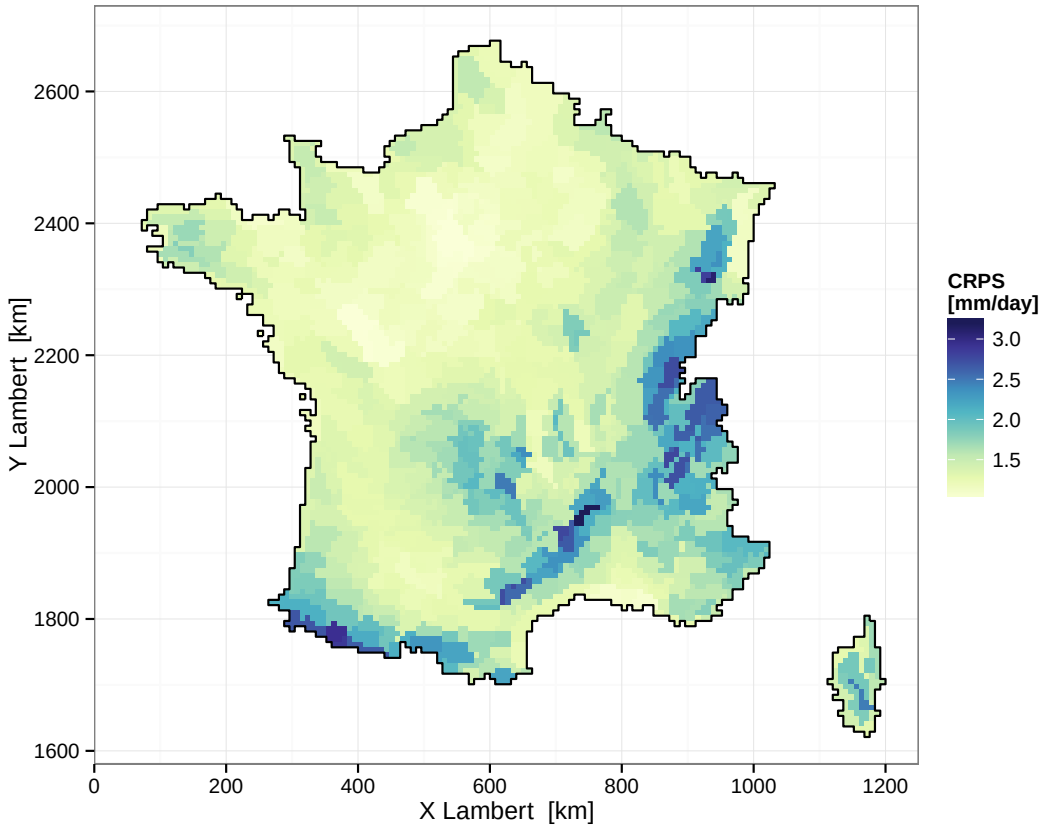

Figure 6.2: CRPS for the reference simulation.

the reference experiment and among the ones with the poorest skill in the out of sample validation experiment. This is further examined later in Section 6.4.

The upper right figure with the blue dots in Figure 6.4 corresponds to the alternative archive experiment. All the dots are below the first diagonal, but in general rather close to it, which means that we observe a rather small skill loss everywhere.

For the suboptimal predictor domains experiment in the bottom subfigure with the green dots in Figure 6.4 a similar behaviour as for the out of sample validation can be seen. Note that for zone 367 the skill loss in the suboptimal predictor domains experiment is about half the one observed in the out of sample validation. If the suboptimal predictor domain explains only half of the skill loss for zone 367, what is responsible for the other half? If it is the archive, why don't we see anything extreme in the blue dots? In fact our four experiments are only one half of the

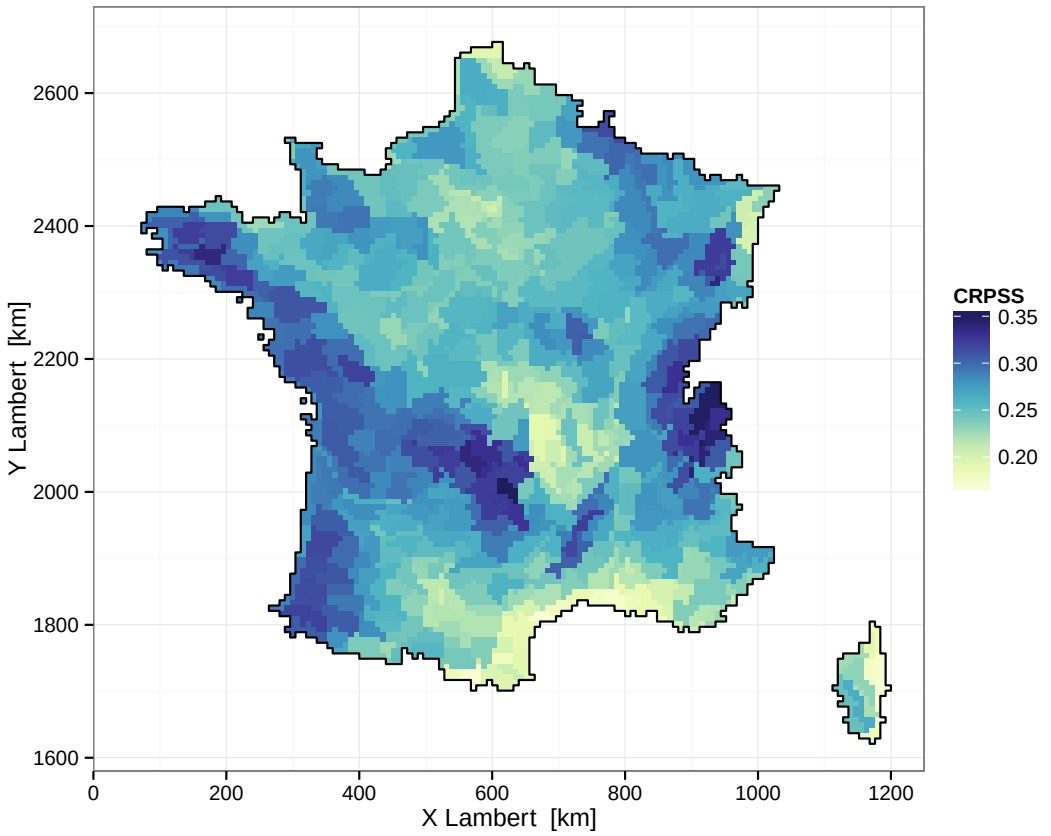

Figure 6.3: CRPSS for the reference simulation.

experiments one could do, because the optimisation period is here never changed. The experiment that is supposed to explain the other half of the skill loss would be a simulation of the *early* period with domains optimised for the *early* period and the *late* period as archive. Such an experiment has not been done, because it implies to redo the optimisation of predictor domains for the *early* period and the optimisation is comparatively costly in terms of computer resources. It was hoped that the two periods would be more or less interchangeable in terms of behaviour in the validation set-up, but obviously it can not be concluded from the *early* period being a suitable archive for a simulation of the *late* period that the *late* period is a suitable archive for the simulation of the *early* period.

Figure 6.5 shows maps of the difference in CRPSS between each of the four experiments and the reference experiment. Obviously for the reference experiment the difference is zero everywhere. Purple colour indicates skill loss and green colour

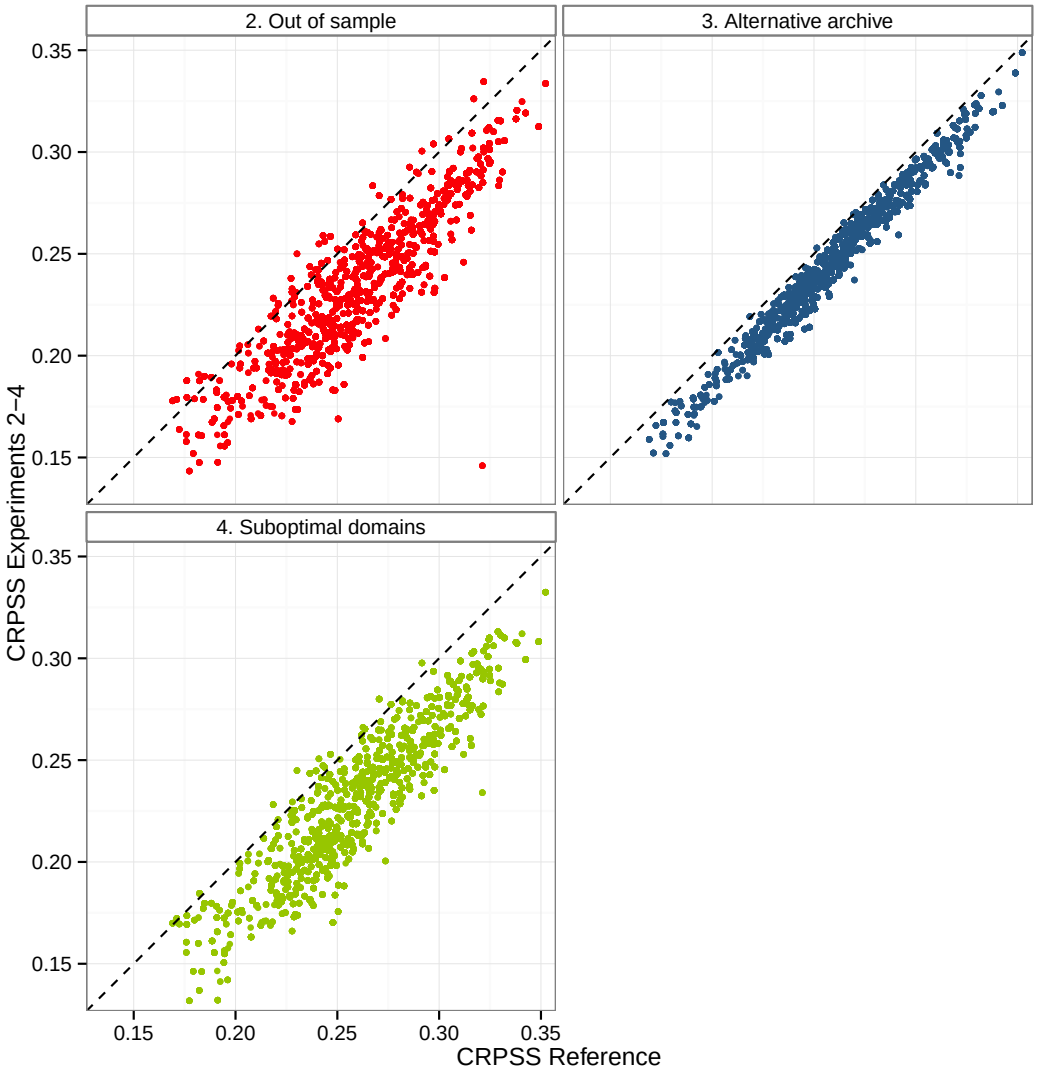

Figure 6.4: CRPSS of the reference experiment vs. CRPSS of the other experiments for all 608 zones.

skill gain with respect to the reference simulation. For the out of sample validation (upper right panel in Figure 6.5) it can be seen that the zones exhibiting a skill gain are situated in the south-eastern part of the country. Zone 367 is even off the scale and therefore in grey colour and can be found in the Massif Central, more precisely in the Sancy Massif that is the highest peak in the Massif Central. The spatial structure of the skill differences of the suboptimal predictor domain experiment (bottom right panel) is similar to the one of the out of sample validation except that there are less zones with a skill gain. For the alternative archive a nearly spatially

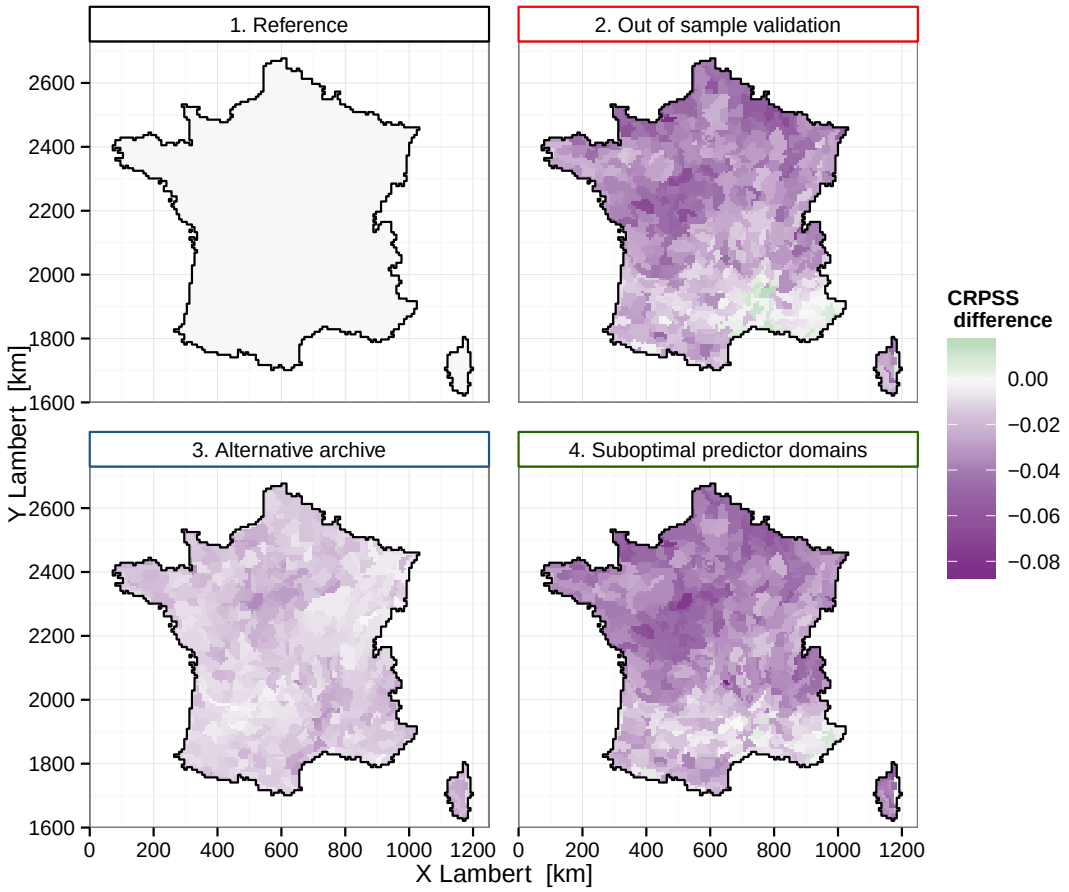

Figure 6.5: CRPSS difference between the validation experiments and the reference experiment.

uniform skill loss can be seen with somewhat more skill loss in the Normandy, the Paris region and the Cévennes and less in Aquitaine and Burgundy. Note that the spatial structure of the CRPSS differences are similar between experiments simulating the same period.

The predictor domains for SANDHY have been optimised using the CRPSS as an objective function, that means that SANDHY is supposed to do rather well in terms of CRPSS since it has been optimised for it. In the next subsection a second verification measure that is widely used to rate and compare models, the bias, is employed.

6.3.2 Bias

The bias is a basic deterministic verification measure. In contrast to the CRPSS it has not been used during optimisation and is thus an independent criterion. Since the bias is a deterministic criterion it has to be decided with respect to which deterministic quantity that can be derived from the probabilistic simulations it should be calculated. It is chosen to calculate the bias with respect to the mean of the 25 analogue precipitation values following Chardon et al. (2014). Other reasonable choices would be the median or an other fixed quantile of the empirical distribution. For asymmetric distributions like the ones typical for daily precipitation, the resulting biases will strongly depend on this choice.

In Figure 6.6 the spatial distribution of the bias with respect to the mean of the 25 analogue precipitation values is shown for the four experiments. For the reference simulation, upper left panel, the mean is generally positively biased. The strongest positive biases are found in the Alps, in the Cévennes mountains and along the western slopes of the Massif Central. Negative biases can be found in the eastern Pyrenees, along the Mediterranean coast, in northern Corsica and some zones in northern France. In general wet areas, for example at the windward side of mountain ridges, tend to have stronger positive biases than drier areas that have rather small biases. This shows that the bias is location specific and it can not be concluded from seeing a positive bias in one region that the method leads to a positive bias everywhere.

The spatial distribution of the biases in the suboptimal predictor domain experiment, lower right panel, is rather similar to the one for the reference experiment, with some tendency to stronger positive biases in the southern part of the country for example in the southern Alps, the Cévennes and the western Pyrenees.

For the out of sample validation, upper right panel, the changes in bias differ from zone to zone. The positive bias in the Alps and the northern part of the country increases, while it stays about the same in the northern Cévennes. There are more zones with positive bias in the southern Cévennes, where the sign of the bias changes, and along the Mediterranean coast. The zone 367, that suffers from substantial skill loss in this experiment, has a strong positive bias, while in the reference simulation the positive bias for this zone is much smaller and quite similar to the one for the zones around.

For the alternative archive simulation, lower left panel, the biases change substantially compared to the reference simulation. There are many zones in the northern part of the country, that have a negative bias in the alternative archive simulation and a positive one in the reference simulation. In the southern part of the country the positive biases tend to be more intense in the Cévennes and less intense in the Alps compared to the reference simulation. Along the Mediterranean coast the sign of the bias changes: there are positive biases in the alternative

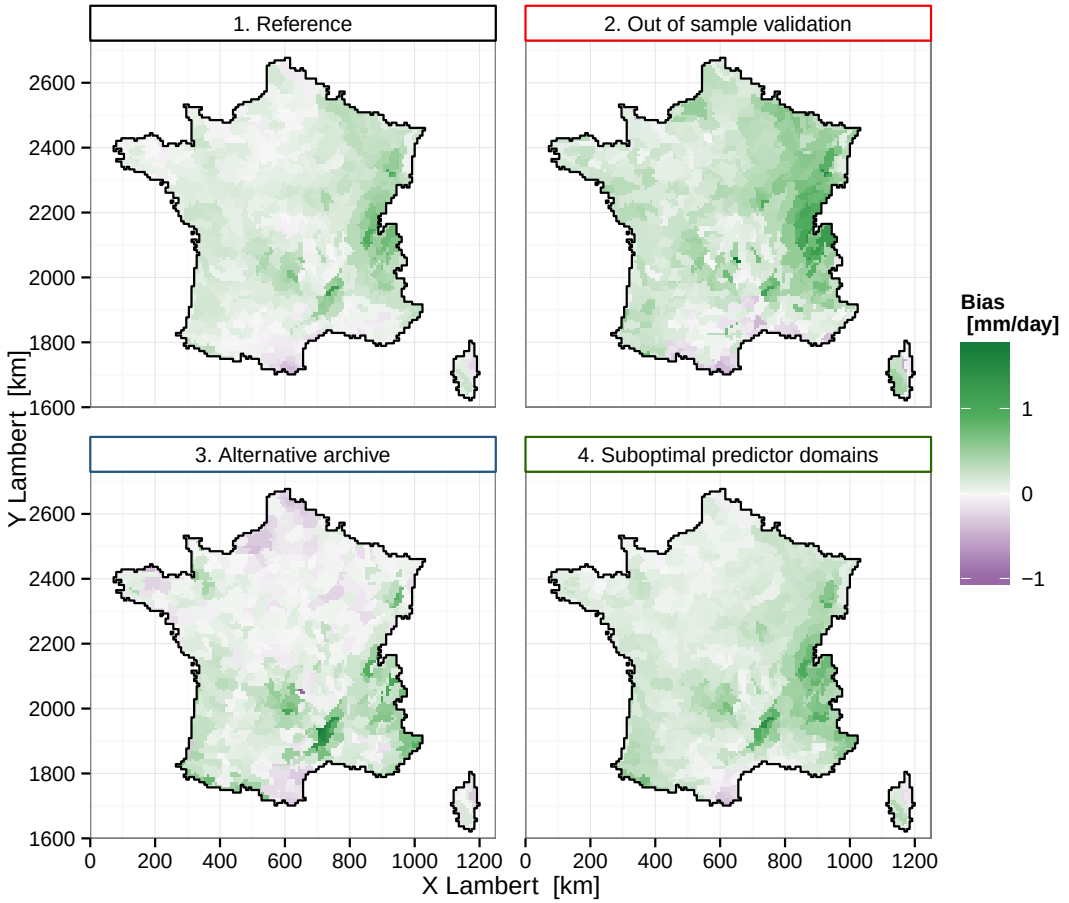

Figure 6.6: The spatial distribution of the bias for the four simulations.

archive simulation and negative ones in the reference simulation. The negative biases in the eastern Pyrenees are still present. The zone 367 has a strong negative bias in this experiment. The difference in bias for the alternative archive simulation seems to be related to the difference in mean precipitation during the two different periods, since the pattern shows some similarity to the mean precipitation difference between the two periods as shown in Figure 2.9. The analogue dates are sampled from a period with more precipitation in the southern part and less precipitation in the northern part compared to the reference, which explains stronger positive biases in the south and negative biases in the north. In summary changes in the bias that are spatially non-uniform occur if the archive period is different from the simulation period.

The odd behaviour of zone 367 leads to the question why this zone behaves differently from the others. What causes the downscaling to fail the validation in

this case?

6.4 Some details about zone 367

The most important reason is probably the difference in the average precipitation between the *early* and the *late* period seen in Figure 2.9 that stems from additional measurements at higher altitude that are included in the *late* period but not in the *early* one. This leads to an inhomogeneity in the precipitation time series and therefore an instationarity of the predictor-predictand relationship since the same large scale situation does not lead to the same precipitation any more. In this case a key assumption of the statistical downscaling is violated.

Nevertheless there might be other influences which is why the time series of zone 367 is inspected in a bit more detail.

6.4.1 Precipitation time series

Figure 6.7 shows the precipitation time series of zone 367. Due to the large variance of daily precipitation the difference in the mean is hardly visible. The most noticable feature is that in the *late* period two events with very high precipitation amounts occur that are higher than anything observed during the *early* period. Notably on the 13 February 1990 190 mm of precipitation in one day is given for zone 367. Regarding the station records used in Safran, there is actually the station Besse – Saint Anastasie at 1340 m, where 216 mm of precipitation were observed on the day in question so this is not an error in the Safran reanalysis. The zone contains a second station at 1000 m where 86 mm were observed. So the values in the Safran data are not inconsistent with the observations used and it can be concluded that these high precipitation values are not due to an error in the reanalysis.

6.4.2 Case study day 13 February 1990

So if it was not a reanalysis error, what was the large-scale situation leading to the very high precipitation amount? Figures 6.8, 6.9 and 6.10 show maps of the predictor variables used in SANDHY for every 12 hours from 13 February 1990 00:00 UTC to 14 February 1990 12:00 UTC. On the 13 February at 00:00 UTC there was a trough over central Europe and a ridge over the Atlantic Ocean moving towards Europe during the two days. At the 1000 hPa level there is a low pressure center over the Adriatic Sea on the 13 Feb. at 00:00 moving eastward. A second one is located over the North Sea, that belongs to a low pressure system whose center becomes visible at 12:00 UTC the same day north-west of the British

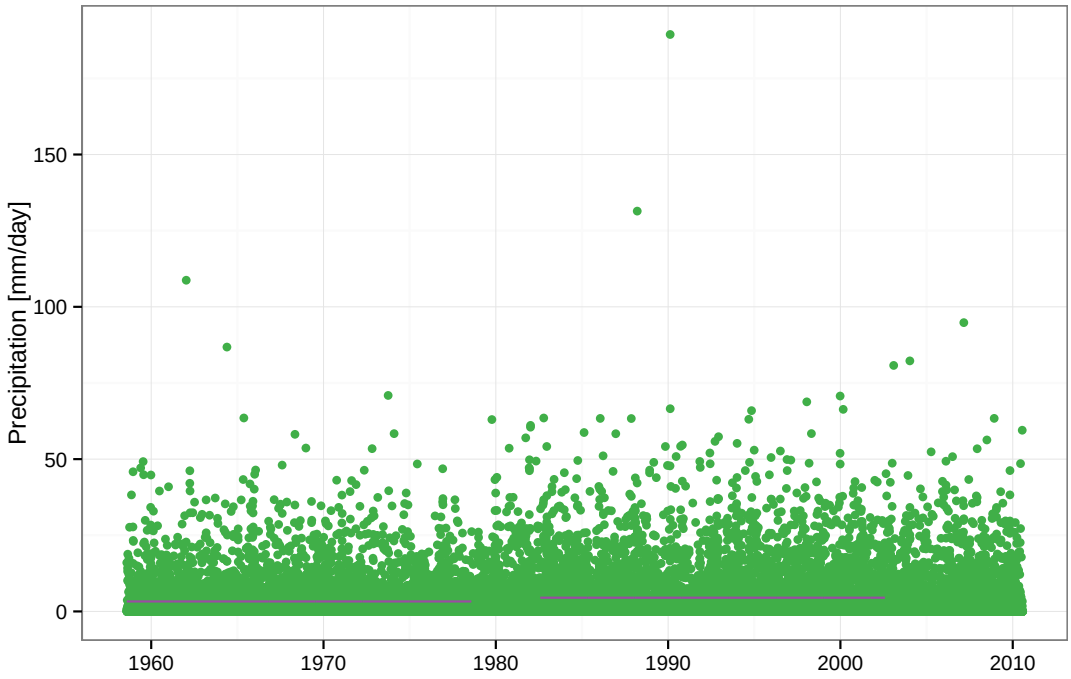

Figure 6.7: Precipitation time series of zone 367 for the period 1 Aug. 1958 - 31 July 2002. The purple lines represent the mean over the *early* and the *late* period.

Islands. This low pressure system is moving south-eastward such that it is over the North sea the 14 Feb. at 00:00 UTC and over central Europe twelve hours later. If we compare the low-level geopotential maps with the (low level) temperature maps and the humidity maps we can see that this low-pressure system is associated with a frontal system. On the 13 Feb. 12:00 the warm front is located over the British Islands and continues over France until the Mediterranean Sea. The cold front is north of Ireland. On the 14 Feb. at 00:00 UTC the fronts have already started to occlude over northwestern Germany and the Benelux, the warm front passes over eastern France, along the Alps to the Mediterranean and the cold front from Belgium westward. During the next twelve hours the occlusion gets deformed along the Alps and the cold front passes over northern France. At 12:00 UTC we can already see the next low pressure center with a new frontal system approaching from the north-west. The precipitation observed over France on the 13 Feb. stems mainly from the warm front and probably the point of occlusion passing over on the 14 Feb. in the early morning.

Figure 6.11 shows the corresponding precipitation over France accumulated from the 13 Feb. 1990 6:00 UTC to the 14 Feb. 1990 6:00 UTC. Heavy precipitation

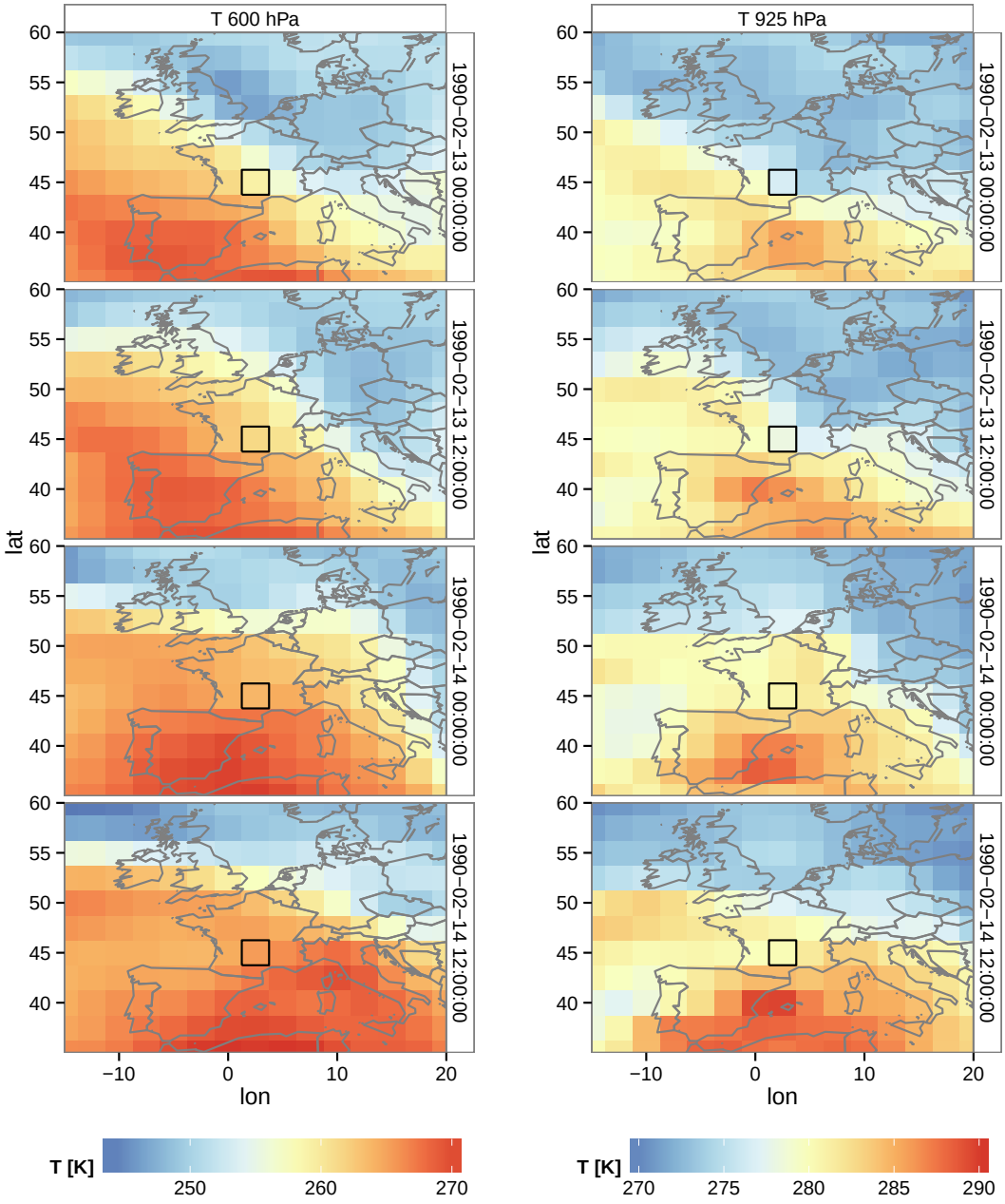

Figure 6.8: Temperature at left: 600 hPa, right: 925hPa, first row: 13 Feb. 1990 00:00 UTC, second row: 13 Feb. 1990 12:00, third row 14 Feb. 1990 00:00, fourth row 14 Feb. 1990 12:00. Black rectangle: Predictor domain for zone 367.

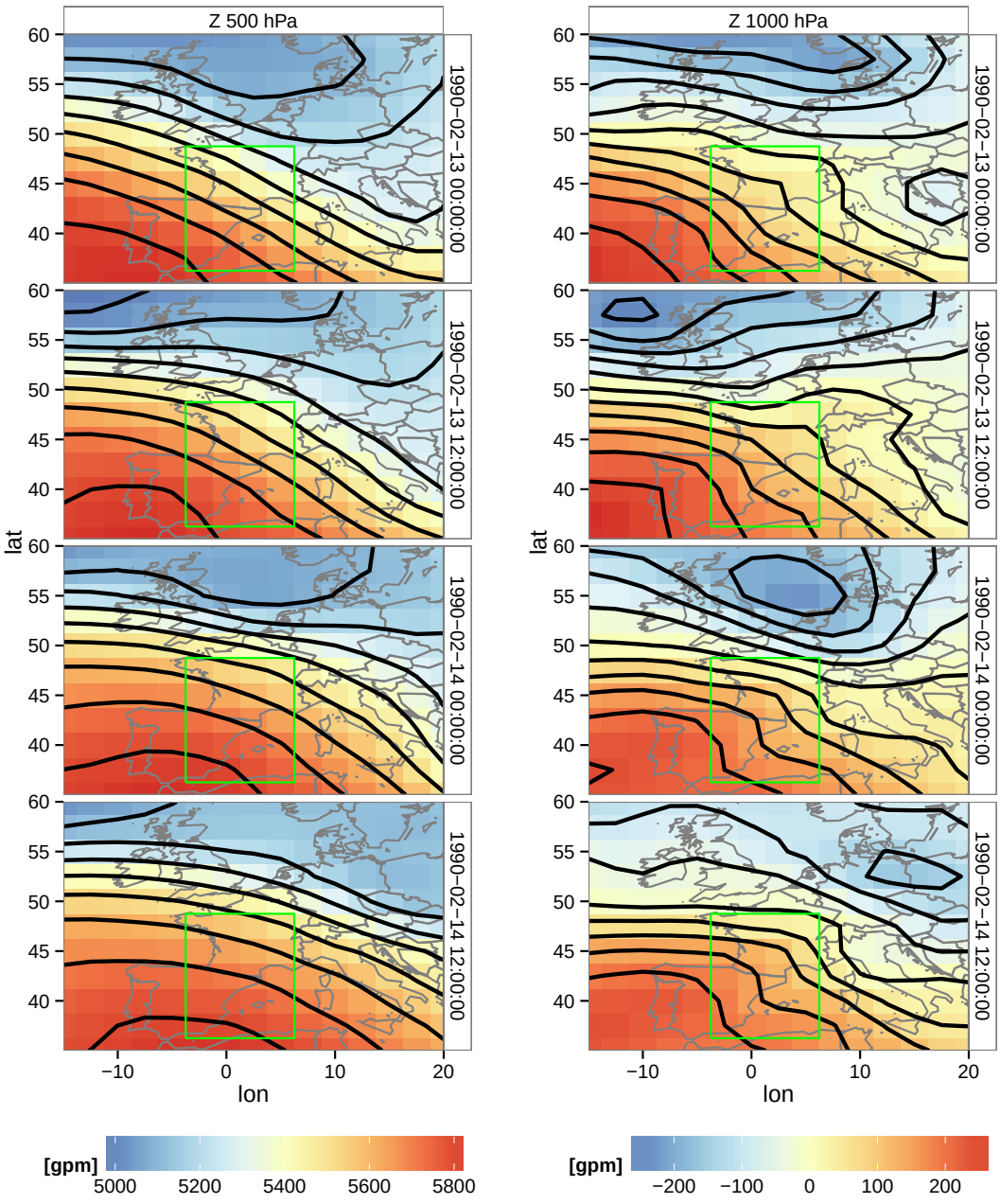

Figure 6.9: As Figure 6.8 but for geopotential height at left: 500 hPa, right: 1000 hPa. Green rectangle: the first optimised predictor domain for geopotential.

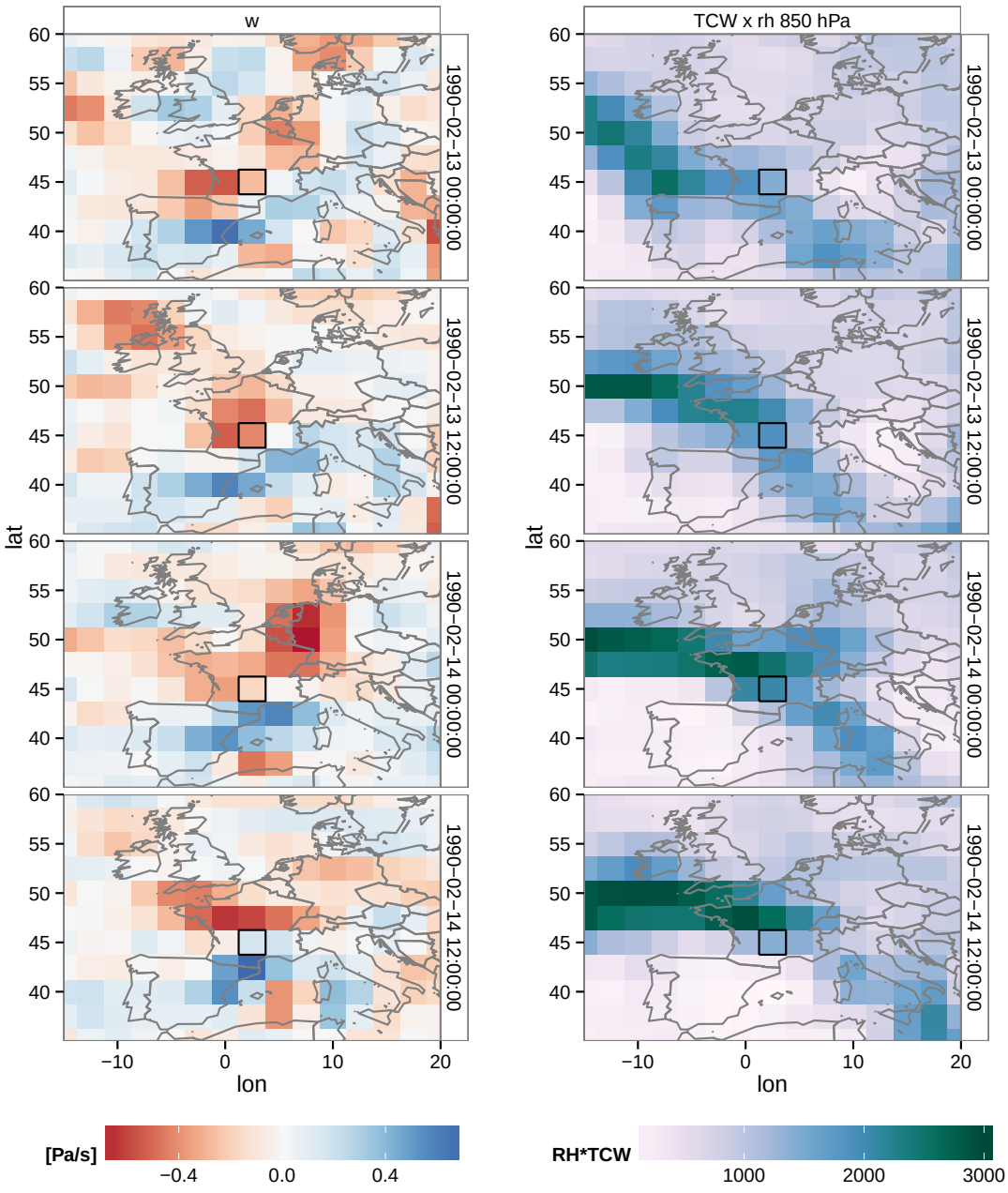

Figure 6.10: As Figure 6.8 but for left: vertical velocity at 850 hPa, right: relative humidity at 850 hPa times total column water

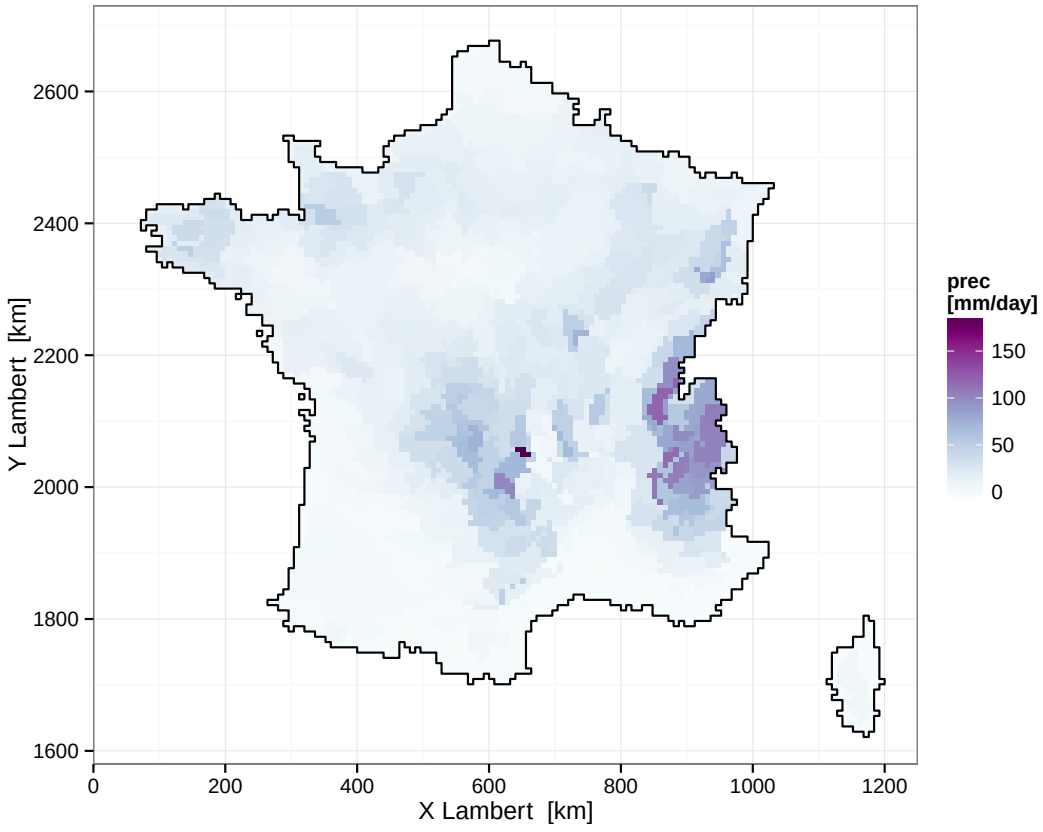

Figure 6.11: Precipitation from zone averaged Safran data on the 13 February 1990.

has been observed over most of northern and eastern France with orographically enhancement due to the Alps, the Jura, the Massif Central, the Vosges, the Morvan and the Ardennes. The highest precipitation values are observed in the Alps and at the zone 367 in the Massif Central.

6.4.3 SANDHY predictor fields for the 13 February 1990

Now that we have an overview of the situation on the day in question we can look more specifically at the predictor fields at the times actually used in SANDHY. Figure 6.12 shows these predictor fields. The top right panel shows the isohypses of the 500 hPa pressure level on the 14 Feb. 00:00 UTC and the top left panel the temperature at the 600 hPa pressure level on the 13 Feb. at 12:00 UTC.

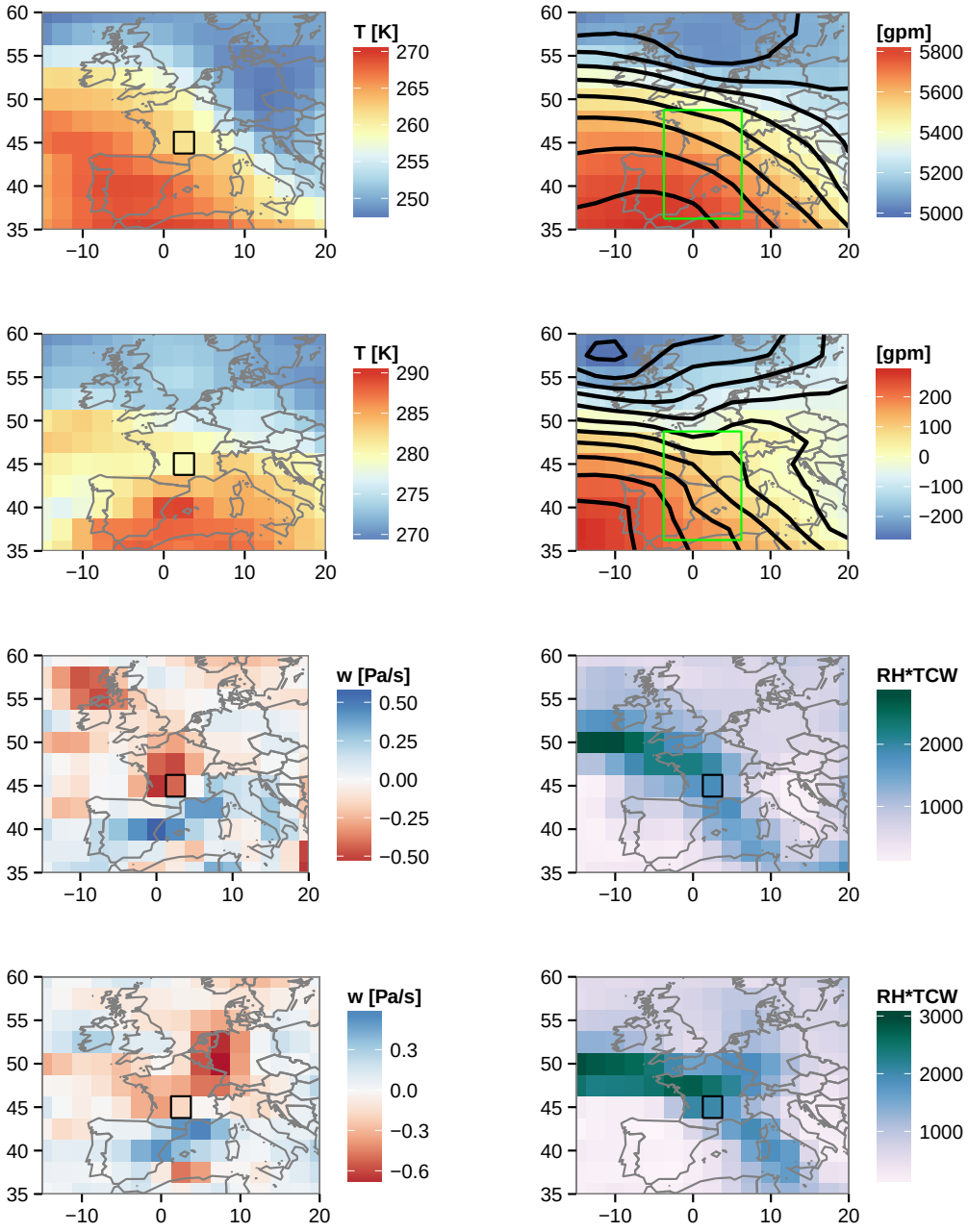

Figure 6.12: Predictor fields for the 13 February 1990. Top left: temperature at 600 hPa 13 Feb. 12h, top right: geopotential at 500 hPa 14 Feb. 0h, second row left: temperature at 925 hPa 14 Feb. 12h, second row right: geopotential at 1000 hPa 13 Feb. 12h third row and bottom row left: vertical velocity at 850 hPa at two of the four timesteps used 13 Feb 12h and 14 Feb 0h, middle and bottom right: relative humidity at 850 hPa times total column water at same timesteps as vertical velocity. Predictor domains for zone 367 are depicted in green for geopotential and black for the other variables.

We can see a trough over the North sea and northwesterly upper-air flow over France with slightly anticyclonically curved isohypses in the predictor domain for the geopotential predictor (green rectangle). Warm air masses are situated over southwestern Europe and colder ones over central Europe. Note that the isohypses in the predictor domain are rather anticyclonic, which is not a typical situation for heavy precipitation.

In the second row the left panel shows the isohypses of the 1000 hPa pressure level on the 13 Feb. 12:00 and the right panel the temperature at the 925 hPa pressure level on the 14 Feb. 12:00. We see a low pressure center north-west of the British Islands and a short wave trough over the British Islands and northern France in about the same place were the warm front is located the 13th at 12:00. These cyclonically curved isohypses together with the north-westerly flow lead to positive vorticity advection and to upward motion as can be seen in the vertical velocity field at the same time (third row left panel in Figure 6.12). This upward motion together with the available humidity are the ingredients for precipitation. The low level temperature is taken the day after where not only the warm front has already passed but the cold front as well and thus the temperature in the predictor domain (small black rectangle) is not as high as the one that was actually associated with the warm front. There is more or less strong upward motion at all four time steps considered for vertical velocity in the predictor domain (only two timesteps are shown) and high humidity values at both time steps.

So we can say that the event is well captured by the humidity and vertical velocity predictors, but the low level temperature has been taken too late to capture the event and the predictor domain for the geopotential predictors is more south than a human forecaster would tend to look. As a consequence SANDHY considers the situation to be more anticyclonic than it actually is.

6.4.4 Influence of an heavy precipitation event

What are the consequences of having an event like that in the archive? Does it influence the predictor domain optimisation such that it leads to unreasonable predictor domains? How often is the 13. Feb. 1990 selected as an analogue date when simulating the *early* period?

To answer the first question the predictor domain optimisation for zone 367 has been repeated using a data set where the precipitation on the 13 Feb. 1990 has been replaced by a value linearly interpolated between the precipitation values of the 12 Feb. 1990 and the 14 Feb. 1990. The resulting predictor domains, shown in Figure 6.13, are identical to the ones found for the original data, so it can be concluded that the predictor domain optimisation does not depend on one extreme event.

To answer the second question it is counted how often each day in the archive

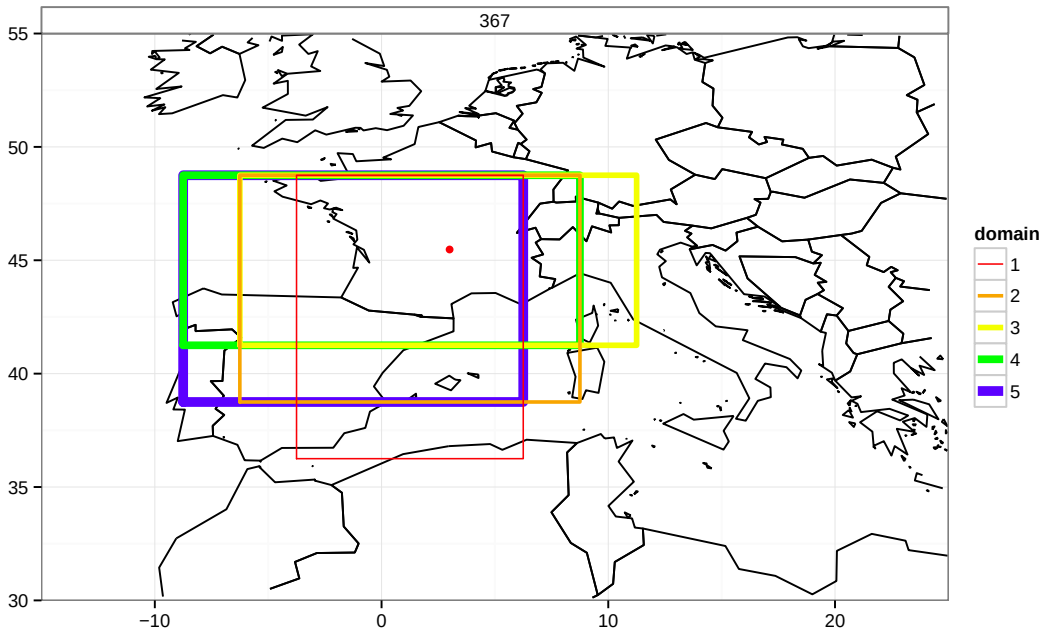

Figure 6.13: Optimised predictor domains for zone 367.

has been selected as an analogue date for the simulation period. Figure 6.14 shows the distribution of the selection frequencies for all archive days in the late period when simulating the early period for zone 367. The average selection frequency is 25 times because 25 analogue days are chosen for each simulated day and the archive and the simulation periods are of the same length. The distribution of the selection frequencies is skewed with a median selection frequency of 22, a minimum of 0 and a maximum of 132. The 13 February 1990 has been selected 52 times and is therefore under the 8% of most selected days.

6.5 Conclusions

The skill loss in the out of sample validation is reasonable in most cases, but we have to be aware that substantial skill loss may occur for some zones. Skill loss mainly occurs when the simulation period is different from the optimisation period which questions the robustness of the predictor domain optimisation. In the presented validation setting it can not be determined weather the skill loss occurs

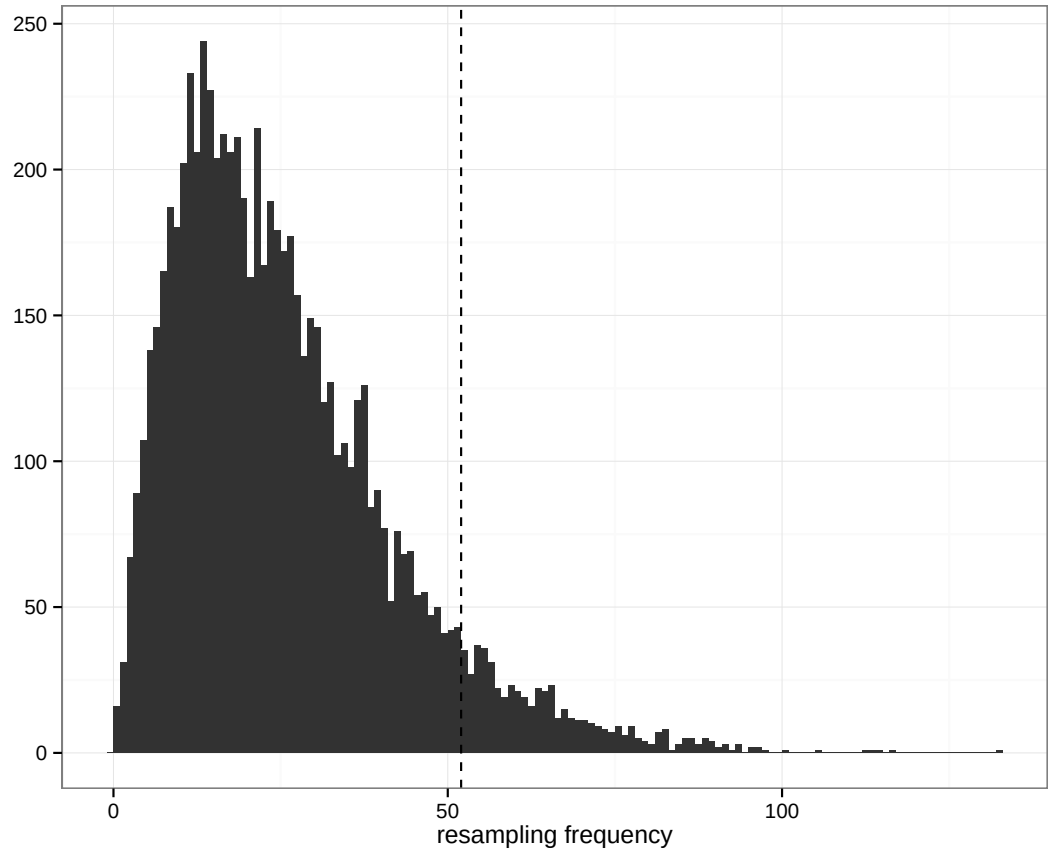

Figure 6.14: Distribution of selection frequency for the days in the archive (late period) for the simulation of the early period for zone 367. The dashed line indicates the selection frequency of the 13 February 1990.

because the *early* period is more difficult to simulate or because the optimisation is not robust. To conclude on this point experiments with predictor domains optimised for the *early* period would be required which implies important additional optimisation costs. Changes in the bias occur mainly when the simulation period is different from the archive period. These changes in bias, that depend on the difference in climatology between the archive and the simulation period, have serious implications for the application of the method: we can not assume that the bias at a given place will be constant over time and thus any attempt to do a simple bias correction that relies on this assumption will not be valid. Furthermore it shows that the downscaling method has difficulties to adapt to a slightly different climate and that in the context of a changing climate biases have to be expected such that long-term trends and variability are underestimated. More precisely, if there is for

example a positive long-term trend the bias is expected to become more negative than for recent climate conditions.

An extreme event that occurs in the archive period under large-scale conditions that are not rare can affect the downscaling results, because the extreme precipitation is resampled far too often. On the other hand if there aren't any extreme events in the archive, they can not be simulated at all with a pure resampling method like SANDHY.

Part III

Reducing the parameter space

7 Using better domains from neighbouring locations

During the preparation of the validation of the downscaling method that was the subject of Chapter 6, a CRPSS over the 20-year long optimisation period was calculated for every zone and every predictor domain found for any zone. In total, optimising predictor domains for 608 zones using five growing rectangular domains, 847 different domains, or more precisely combinations of geopotential predictor domain and nearest ERA-40 grid cell, were found. 847 times 608 CRPSS values are calculated which allows to search for every zone which of the 847 predictor domains gives the largest, that is best, CRPSS. Ideally, if the optimisation algorithm works well, this *best* predictor domain should be the same as the one that has been found during optimisation, but looking at Figure 7.1, that shows the difference between the largest CRPSS and the CRPSS of the first of the locally optimised domains, it can be seen that this is actually not the case for the majority of the zones. The largest differences are found along the Mediterranean coast and the spatial pattern of the differences is governed rather by the ERA-40 grid cell limits than climatologically meaningful features. The predictor domains that result directly from the optimisation are called *optim* in the following and the ones with the largest CRPSS are called *best*.

7.1 Why have the best domains not been found?

There are two reasons why the optimisation algorithm was not able to find the *best* domains. The first one is the choice of the starting domain. The nearest 2x2 grid point domain from the target zone has been chosen, but this may not always be the best choice as has been already shown in section 4.1 of the article in Section 5.1. The red grid in Figure 7.1 indicates where the starting domain changes. North of the 47.5 deg. line we can actually see a zone with considerable differences in CRPSS that seems to be related to this limitation. The other regions with

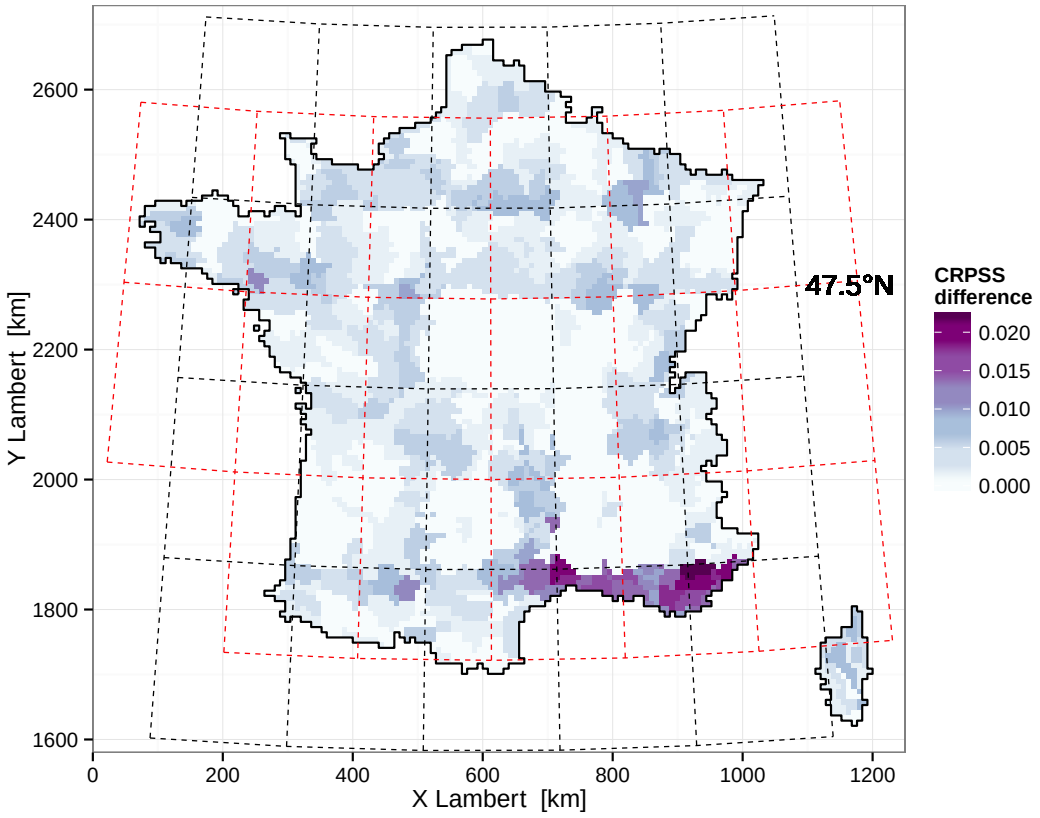

Figure 7.1: Difference between the CRPSS for the best domain found and the one for the first optimised domain. The black grid indicates where the nearest ERA-40 grid cell changes, the red grid indicates where the starting domain of the optimisation changes.

comparatively large differences align rather along the black grid which indicates where the nearest ERA-40 grid cell – and therefore the predictor domain of the temperature, vertical velocity and humidity predictors – changes. This brings us to the second important choice, which is the fact that the predictor domains for the other three predictor variables were not optimised. The nearest ERA-40 grid cell to the target location has been used. This seems not the best choice for many zones and especially for the zones along the Mediterranean coast. But where exactly has been found the *best* domain of a given zone?

7.2 Where have the best domains been found originally?

In Figure 7.2 some arrows have been added that start from the zones for whom the difference in CRPSS between the first *optim* domain and the first *best* domain is larger than 0.01. The arrows point to the zones where the first *best* domain in question has been found as one of the five optimised domains. If this *best* domain for zone A has been among the five *optim* ones for several zones (let's say B, C and D), there are several arrows starting from the zone A and pointing to the zones B, C and D respectively. Where arrows cross the black grid, the nearest ERA-40 grid cell and therefore the predictor domain for the temperature, vertical velocity and humidity predictors changes. We can not tell from this graph if the geopotential predictor domain changes as well or not, but what we can say, is that if only the black grid is crossed and not the red one, the search space for the geopotential predictor domains has had the same a priori restrictions during the optimisation. Where arrows cross the red grid only, only the geopotential predictor domain has changed, but this does not occur for the largest differences.

To see arrows that cross red lines, we have to look at Figure 7.3. Here there is an arrow for every zone where the first *best* predictor domain is not equal to the first *optim* predictor domain, but it points only to one of the zones for whom their *best* domain has been found during optimisation. To keep the example from above, the arrow from zone A points to zone B only now if B is the first encountered zone. Many arrows point towards the same zones, which is related to the fact that they always point towards the first encountered zone for whom the *best* domain has been found during the optimisation. Therefore this does not mean that these zones are more representative than others.

The five *best* domains out of the pool of 847 domain are compared with the five optimised ones for the three main case study zones that are described in section 2.1.2 of the article in Section 5.1. Figure 7.4 shows the five optimised domains and the five *best* domains for the three case study zones. For the Saône case study zone the five *best* domains are smaller in north-south direction than the *optim* ones. The five *best* domains are similar to the domains found in the sensitivity study using an alternative starting point, that is the most relevant elementary domain instead of the nearest grid cell, in Figure 8 of the article in section 5.1. As discussed in the article, the Saône case study zone is one of the zones where the nearest elementary domain is clearly not the optimal starting point for the predictor domain optimisation. The predictor domains for the other variables, that is the ERA-40 grid cell used, are shown in Figure 7.5. The first, third and fourth *best* domains have the same predictor domain for the local predictor variables as in the *optim* experiment. Figure 7.6 shows where the *best* predictor domains for the

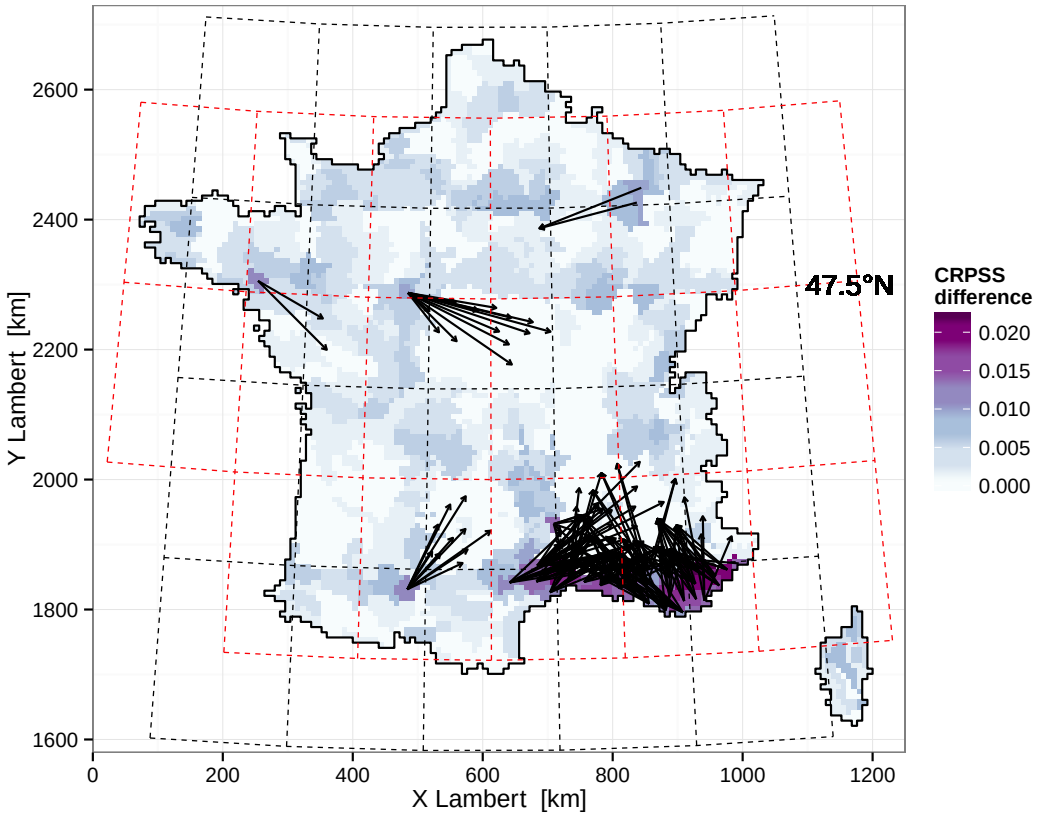

Figure 7.2: As Figure 7.1 but with arrows from the zones where the difference is larger than 0.01 to all the zones for whom their first best domain has been found during optimisation.

case study zones have been found during the optimisation and indeed the arrows corresponding to the first, third and fourth domain do not cross a black line. For the domains 2 and 5 the local predictor domain differs. Note that the geopotential predictor domain is three times the same for domains 2, 4 and 5, which means that the only difference between them is the local predictor domain. This is in contrast to the zones along the Mediterranean coast, where important skill differences arise due to a suboptimal local predictor domain used in the *optim* case.

For the Arve case study zone the first best geopotential domain has been found fourth best during optimisation, but with a different nearest grid cell further north. The zones for whom the best and the third best domains have been found are the same which makes the corresponding arrows in Figure 7.6 hard to see because they are the same. The same applies to the second and the fifth domain, such that the

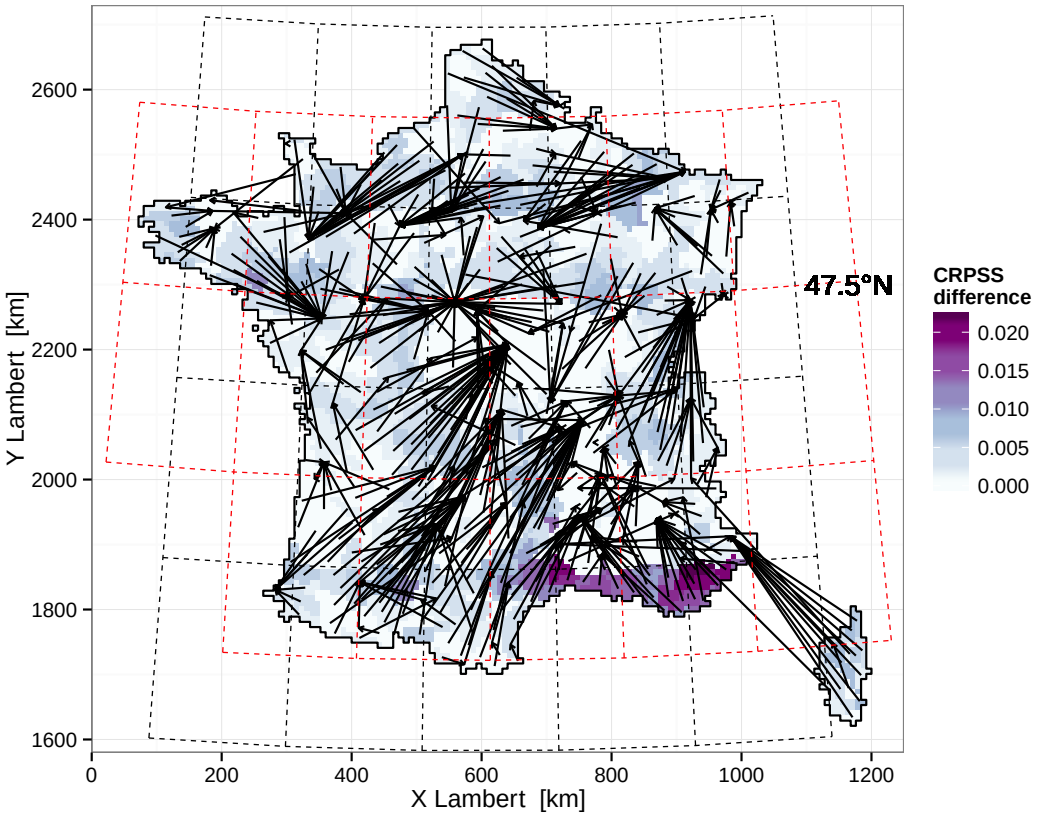

Figure 7.3: As figure 7.1 but with arrows from the zones with a difference in CRPSS to one of the zones for whom their best domain has been found during optimisation.

blue arrows are below the orange ones. For the Arve zone it can be seen that no arrow crosses a red line, but the ones for the first, third and fourth best domain cross a black line, leading to a different local predictor domain as can be seen in Figure 7.5. The second and the fifth *best* domains correspond to the first and the second *optim* ones. The third and the fourth best domains have the same predictor grid cell for the temperature, vertical velocity and humidity predictors as the first *best* and geopotential predictor domains similar but not equal to the optimised ones. In summary, for the Arve catchment there is a small but not crucial benefit from using different predictor domains for the temperature, vertical velocity and humidity predictors.

For the Ardèche zone the first, second, fourth and fifth *best* domains correspond to the first four *optim* ones. The third best domain has been found for a zone next

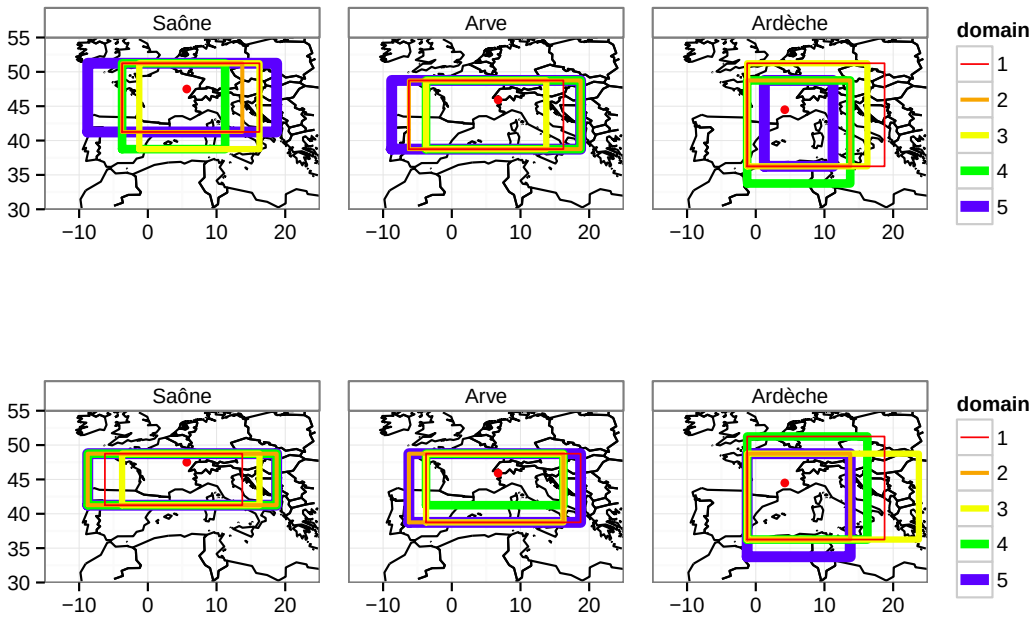

Figure 7.4: top: five *optim* geopotential predictor domains, bottom: five *best* geopotential predictor domains for three case study zones.

to the Ardèche zone and has not been found because the search algorithm did not search far enough to the east. This is in line with what has been found in the 99 domains experiment in Section 5.2, where we saw a "gap" in the aspect ratio (Figure 5.2) that the search algorithm could not overcome using the five domain configuration. Obviously only one of the three domains, that have been found to perform better than the third and fourth *optim* domain in the 99 domain experiment has been found for other zones. There is no change in the local domains for the Ardèche zone.

Following these details about the *best* domains compared to the *optim* domains for the three case study zones, where we saw very little changes for the Ardèche zone and much more important changes for the Saône zone, we will see how the domain characteristics at the scale of France differ between the optimised and the best domains.

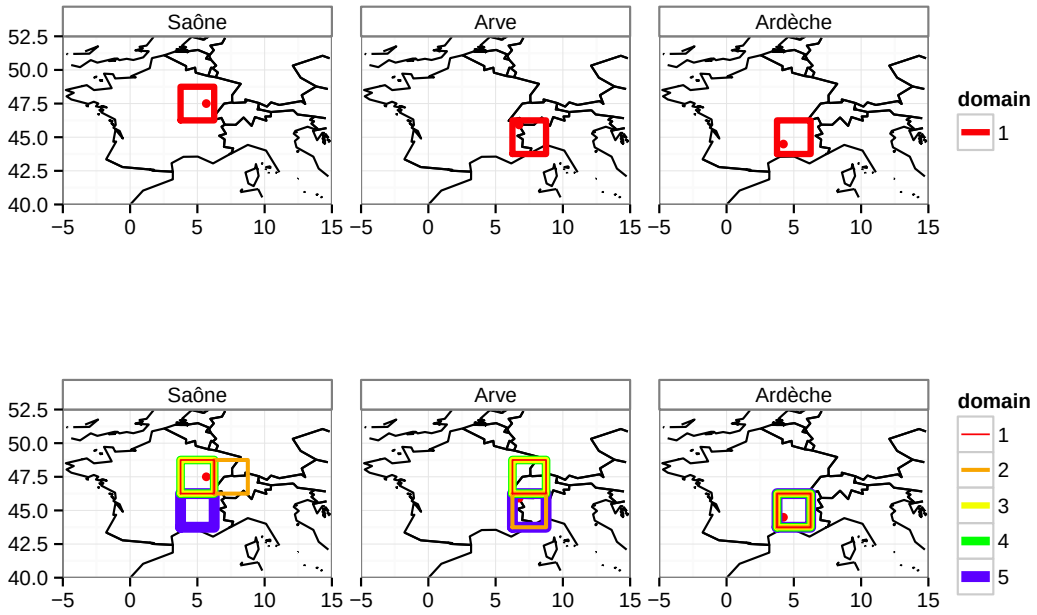

Figure 7.5: top: predictor domain for temperature, humidity and vertical velocity in the *optim* experiment, bottom: five predictor domains for temperature, humidity and vertical velocity in the *best* experiment for three case study zones.

7.3 Characteristics of the *best* domains at the scale of France

Figure 7.7 shows the mean domain center for the geopotential predictor for the optimised and the best five domains. The overall pattern of the mean domain center remains the same, but for individual zones changes can be noted. North of the 47.5 deg line the domain centers tend to be more south, in the northern part of the Massif Central and in the Alps the pattern is less patchy.

The next interesting question is how the domain center variability behaves, because on one hand there is more variability for zones like the Ardèche case study zone, where we saw that one of the best domains is reaching some degrees further to the east than the optimised ones, on the other hand there is less variability for zones like the Saône case study zone where three times the same geopotential predictor domain with different domains for the local predictors is under the best five. Figure 7.8 shows the domain center range for the five optimised domains and the five best domains. With the *best* domains there are larger areas with small

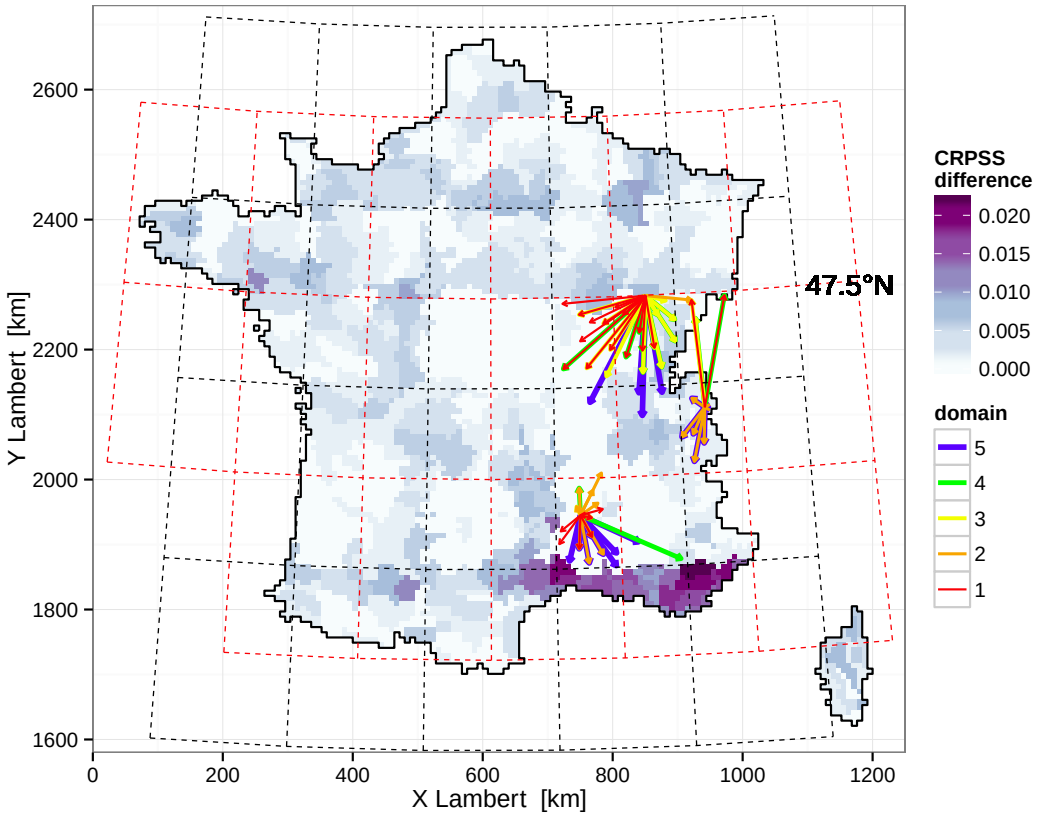

Figure 7.6: As figure 7.2 but with arrows from the three case study zones for all 5 best domains.

domain center ranges. However, there are some more zones in the Southeast that have large domain center ranges. It can therefore be concluded that the case study zones seem to be quite representative for their respective regions.

The picture for the mean domain size of the five *optim* domains was much more patchy than the one for the mean domain center. In addition a sudden change in domain size was observed along the 47.5 degree latitude line (compare Section 5.1). In Figure 7.9 the mean domain size of the five optimised domains and the five best domains is shown. For the best domains the domain size shows a much smoother overall picture, especially along the Alps. The jump along the 47.5 degree latitude line is not there any more and there are less zones with extreme domain sizes in the Massif Central. Taking the *best* domains reduces the number of different domains from 847 to 456, which may explain the smoother picture of the mean domain size. Concerning the domain size range ratio in Figure 7.10, an

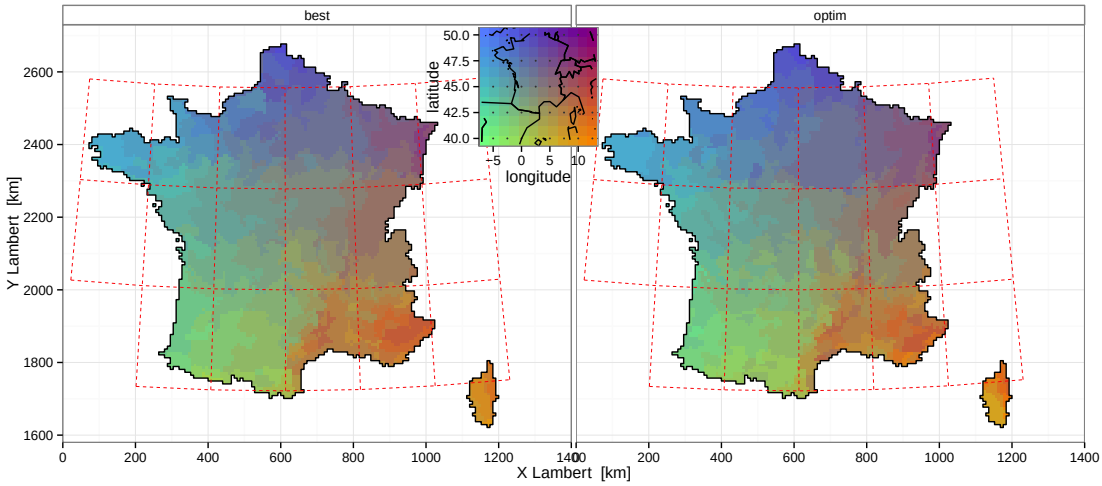

Figure 7.7: Mean domain center for left: *best* domains, right: *optim* domains.

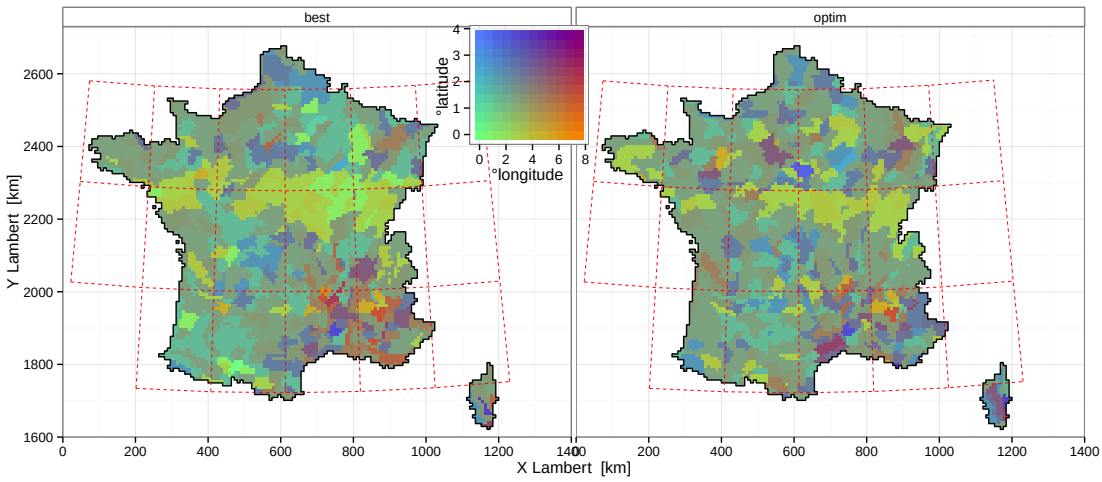

Figure 7.8: Domain center range for left: *best* domains, right: *optim* domains.

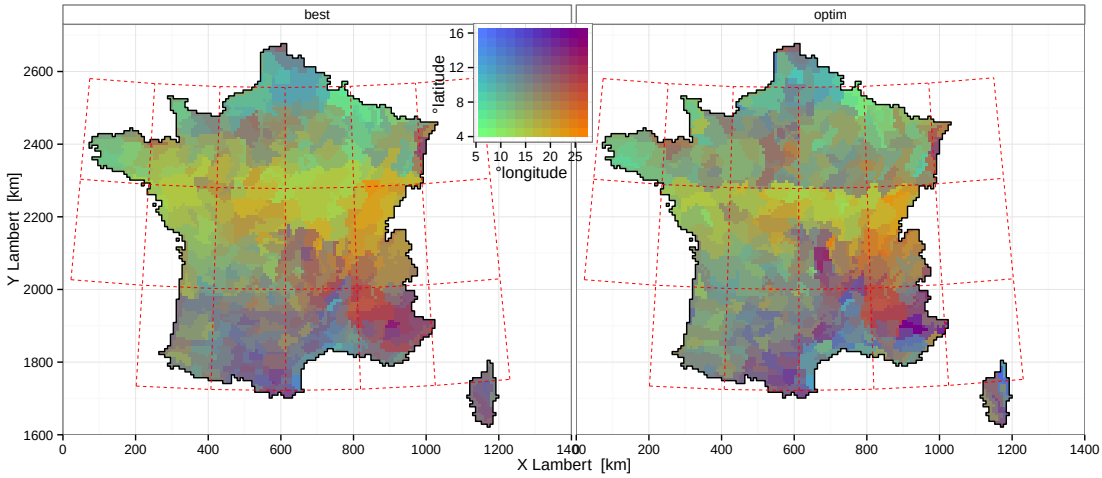

Figure 7.9: Mean domain size for left: *best* domains, right: *optim* domains.

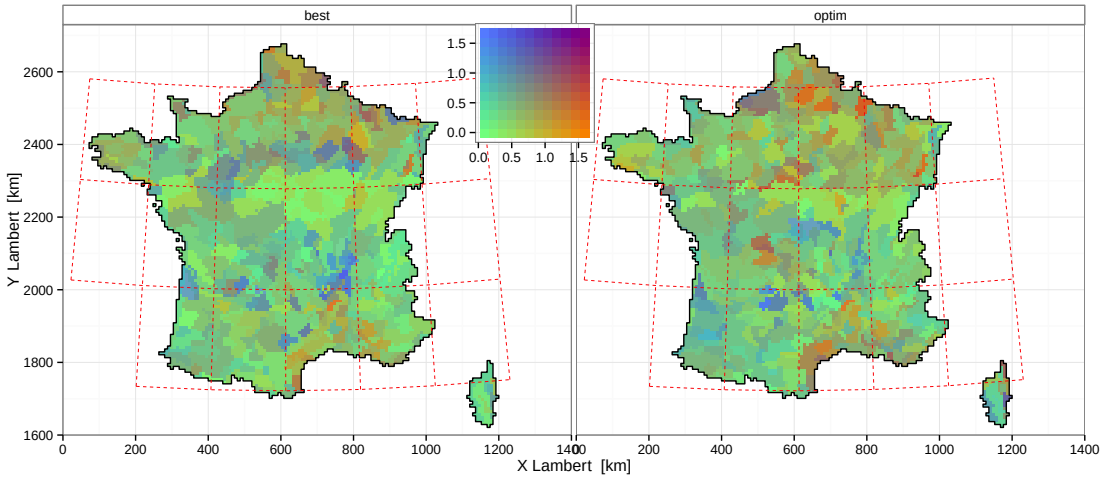

Figure 7.10: Domain size range ratio for left: *best* domains, right: *optim* domains.

overall smaller domain size range can be seen for the *best* domains, especially in the northern part of the country, while there are some exceptions like the Rhône valley where the domain size range ratio is larger for the *best* domains than for the *optim* ones.

Figure 7.11 shows the mean domain center of the predictor domain for the local predictor variables temperature, vertical velocity and humidity. For the *optim* domains these domains were imposed to be the nearest ERA-40 grid cell while this is not the case any more for the *best* domains which leads to a spatially smoother distribution of the domain centers.

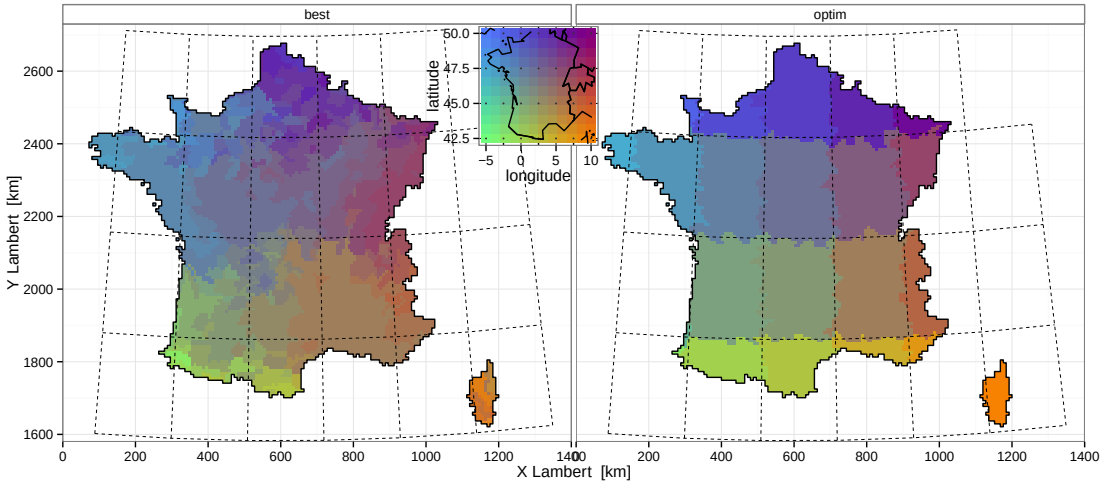

Figure 7.11: Mean domain size for left: *best* domains, right: *optim* domains.

7.4 Conclusions

A limitation of the optimisation algorithm used in Chapter 5 is the obligation to choose an elementary domain to start the optimisation from. This limitation can be mitigated by looking at the skill of the domains found for other zones but this might be less efficient for zones close to the country borders where other zones are not available in every direction. The limiting effect of the imposed and not optimised predictor domains for the local predictor variables temperature, vertical velocity and humidity is mitigated as well. The limitation related to the insufficient exploration of the search space as has been demonstrated in the

99 domain experiment in section 5.2 has been only partially mitigated by using domains found for other zones, but if we look at the CRPS differences this is only a minor limitation. The number of different predictor domains is substantially reduced from 847 to 456, suggesting an increase in the spatial coherence over the whole country.

8 Predictand area aggregation

In this chapter groups of zones are formed that are expected to use the same parameters for the downscaling and therefore reducing the different parameter configurations used. This is expected to lead to stronger spatial coherence inside the groups but introduces increased spatial decoherence at the boundaries. The process of building groups is supposed to help defining a typical spatial scale up to which not much skill loss is expected when taking the same parameters. This means that in addition it has to be decided which parameters a specific group is supposed to use. Cluster algorithms form groups of similar items. To this end the similarity between these items, in this case the zones, has to be defined. The similarity can be defined in terms of predictands or predictors. Results from this chapter have been presented at the EGU General Assembly 2013 (Radanovics et al., 2013b) and the IMSC 2013 (Radanovics et al., 2013c).

8.1 Cluster zones by correlated precipitation

This section deals with clustering in terms of predictands. This means that only precipitation data is used and it is independent of the predictors, the predictor domains and their optimisation. Seibert et al. (2006) uses a mixture of hierarchical clustering and k-means clustering of backward trajectories to define regions with similar behaviour in terms of precipitation. Matulla et al. (2003) compares three methods to define homogeneous precipitation regions: rotated principal component analysis, self organising networks and hierarchical clustering with correlation as similarity measure and concludes that the differences between the methods in terms of the resulting groups is small.

Here the zones are clustered using the affinity propagation algorithm (Frey and Dueck, 2007). Affinity propagation builds clusters by passing messages between data points and returns a representative cluster member for each cluster in addition to the definition of clusters. This has the advantage that a representative zone for each cluster is obtained that could be used to define common parameters for all

zones in the cluster and thus reducing the parameter space. A second advantage of the algorithm is that on the contrary to many other clustering methods it is not necessary to specify the number of clusters beforehand. Frey and Dueck (2007) found affinity propagation to be faster and giving more accurate results than the widely used k-means algorithm (Diday et al., 1974). Affinity propagation is implemented in the R package apcluster by Bodenhofer et al. (2011).

The spearman rank correlation coefficient is calculated between daily precipitation for each of the 608 zones over a 44 years period from 1 August 1958 to 31 July 2002 and is used as similarity matrix. Clustering the zones by correlated precipitation results in 52 clusters that are shown in Figure 8.1. Champeaux and Tamburini (1996) found 51 clusters using an agglomerative hierarchical classification method with the euclidean distance between 20 years (1971-1990) of daily precipitation at 1976 stations as distance matrix. They found larger clusters in the north of France and smaller ones in the mountainous regions, particularly in the Cévennes. With further aggregation they reduced the number of clusters to 12. Compared with their results the clusters obtained with affinity propagation are rather homogeneous in size.

Aggregating zones by correlated precipitation is not necessarily the most adapted approach if we think of the analogue method, because using the same analogue dates does not automatically mean that the precipitation is correlated and using different analogue dates does not automatically imply uncorrelated precipitation. Aggregating by correlated precipitation means aggregation in terms of similar behaviour of the predictand, but actually we are more interested in similar behaviour of parameters concerning the predictors.

8.2 Cluster zones in terms of common analogue dates

To include some information on the predictor domains in the clustering but maintaining the advantages of continuous distances the similarity between zones is expressed in terms of the fraction of common analogue dates between two zones. The fraction of common analogue dates is calculated over the 20-year *late* period for the 5 *optim* domains found in Section 5.1 for each pair of zones. The analogue dates from the five domains were pooled together. If an analogue date for a given day has been found with several domains (out of the five) it is counted as often as it appears for both zones that are compared. For example for zone A the day d has been selected with 3 of its domains and for zone B the day d has been selected with 2 domains, it is counted two times for the fraction of common dates. This ensures that the fraction of common dates of A with B is the same as the one for

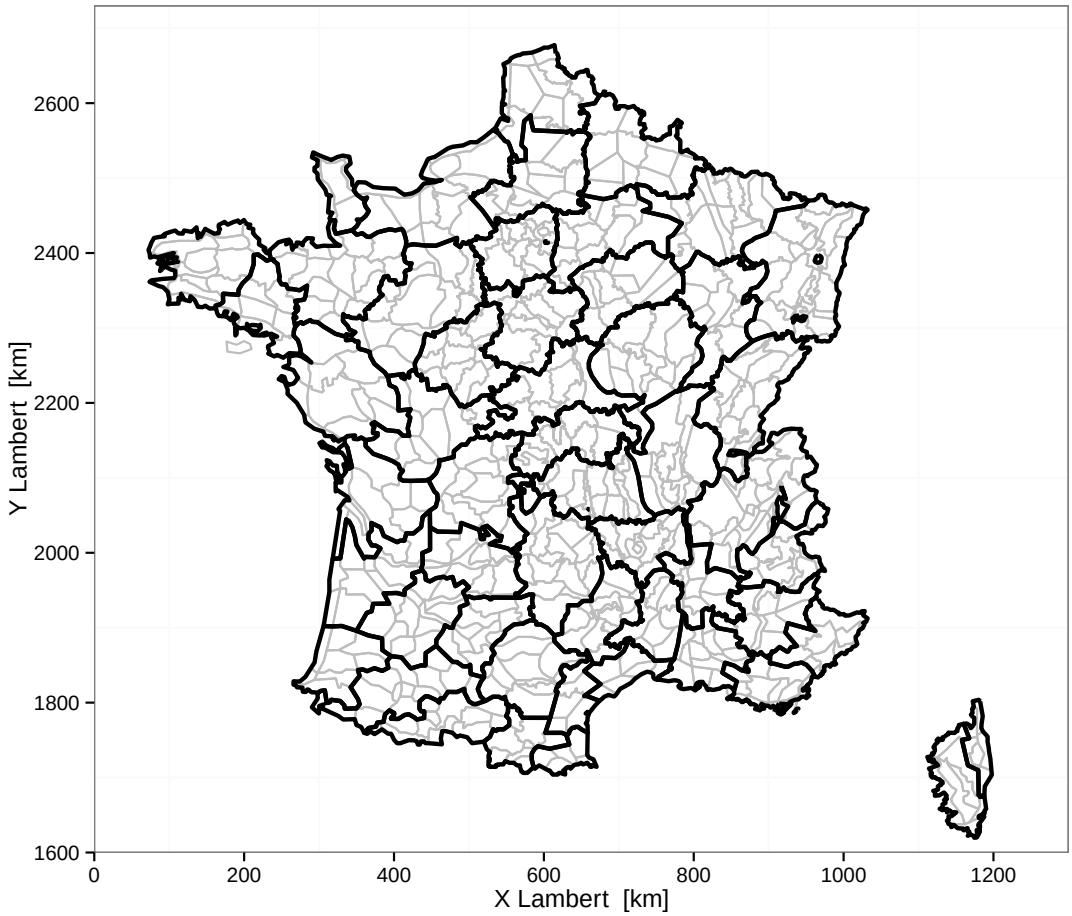

Figure 8.1: Zones clustered to 52 clusters with the affinity propagation algorithm by rank correlation coefficient of precipitation.

B with A, so that the resulting similarity matrix is symmetric.

The 608 zones are clustered again using the affinity propagation algorithm, but this time with the fraction of common analogue dates as similarity matrix. The resulting clusters can be seen in Figure 8.2. Here a much larger number of clusters, 87, is obtained but they are now smaller in the southern part of the country and larger in the northern part, which is in line with the results of Champeaux and Tamburini (1996) and reflects the less similar predictor domains found in the southern part of the country in Section 5.1.

The fraction of common analogue dates contains some information about the similarity in terms of predictors, however, it does not ensure that the local skill is maintained. In order to aggregate zones without loosing skill they are now aggregated by the predictor domains itself.

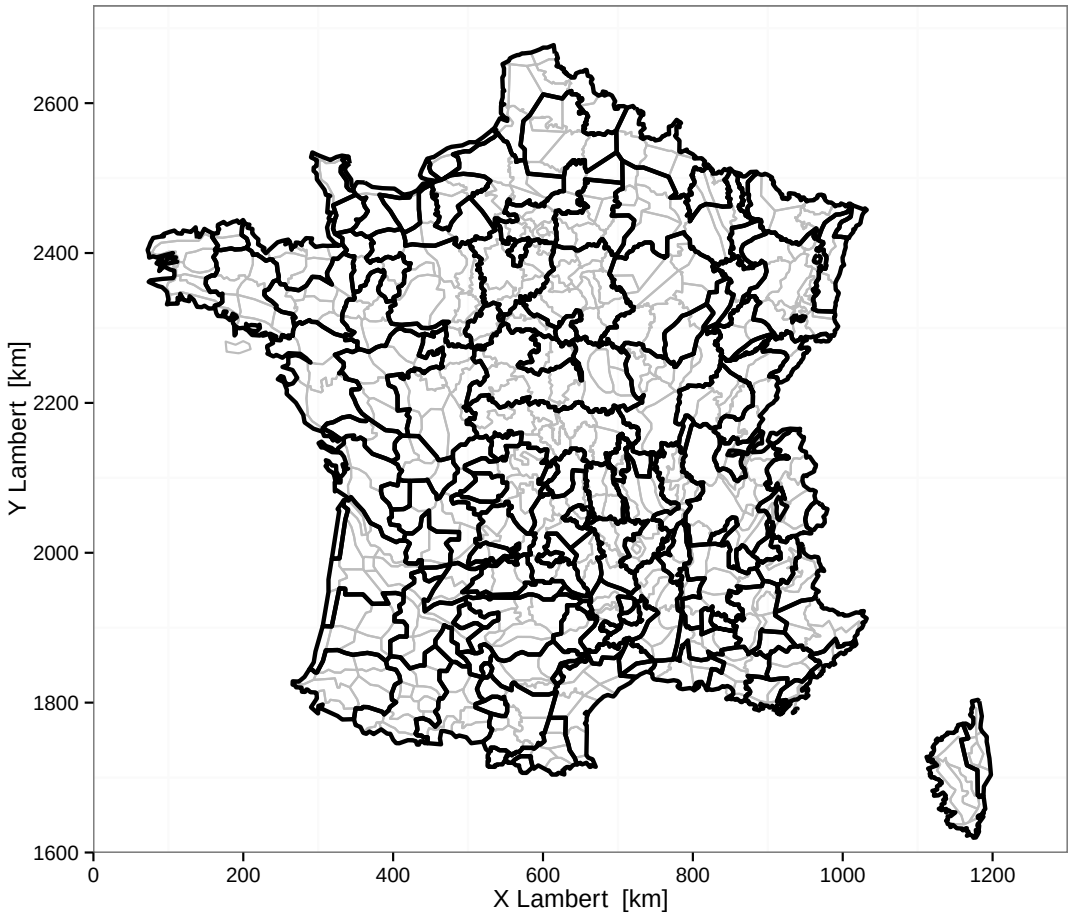

Figure 8.2: Zones clustered with the affinity propagation algorithm by fraction of common analogue dates.

8.3 Aggregate zones by same predictor domains

Aggregating zones by the same predictor domain imposes homogeneity in terms of predictor domains inside the groups and thus ensures that using the same parameters for the whole group does not lead to a loss in skill. The drawback is that the resulting distances – the number of common *optim* predictor domains between two zones – take only a few discrete values (0 to 5) which is an issue for most clustering algorithms. Therefore affinity propagation is not used here. Equal distances occur frequently if the distances between elements comprise just a few distinct values. Those so called ties in proximity lead to non-unique cluster solutions.

8.3.1 Aggregation methods

For each domain that is near-optimum for at least one symposium zone, an identifier is assigned. The predictor domains for all predictor variables are considered, which means to get the same identifier the geopotential predictor domain and the nearest ERA40 grid cell must be equal. This ensures the same analogue dates. Three aggregation methods are considered and are described in the following.

8.3.1.1 Simple Aggregation (SG)

This algorithm is designed to yield a small number of groups and to minimise the number of very small groups. The algorithm works as follows:

1. Find the zones that do not share any of their near-optimum domains with any other zone and assign them their first domain.

2. Find the zones that share one out of their five near-optimum domains with any other zone and assign it to them. Assign this domain to all other zones sharing this domain too.

3. Recalculate the number of shared domains for the remaining zones and repeat steps 1 and 2.

4. If all the remaining zones share at least two of their near-optimum domains take one of the zones that share two domains and assign the one that occurs more often. Assign this domain to all remaining zones sharing this domain too.

5. Repeat steps 3 and 4 until all zones have a domain assigned.

This algorithm, like most clustering algorithms in the presence of ties in proximity, depends on the order of the zones in the input file and is run forward (SGF) and backward (SGB) through the input file.

8.3.1.2 Maximum occurrence (MOC)

For every predictor domain the number of zones for whom they are near-optimum is calculated. For every zone the near-optimum domain out of the five *optim* domains with the highest occurrence count is chosen. This leads to some large groups and many very small groups.

This algorithm is extremely fast and it does not depend on the order of the zones in the input file in case of equal distances, which is the exception rather than the rule for aggregation algorithms. However, it depends on the order of the near-optimum domains in case of equal occurrence counts.

8.3.1.3 Variable group agglomerative hierarchical clustering (VGH)

In agglomerative hierarchical clustering items are grouped by distances between them, starting with the smallest distances. The result is often presented as a dendrogram, but since a dendrogram with over 600 roots is not intuitively interpretable, it will not be shown here. Fernández and Gómez (2008) extended the widely used pair-group algorithm that aggregates the two closest items in each step in order to explicitly account for ties in proximity, that is equal distances between clusters. To explicitly account for the resulting non-uniqueness Fernández and Gómez (2008) proposed an algorithm that merges all clusters in the same supercluster that fall into a tie, while assigning a lower and an upper bound of the amalgamation interval, that is the minimum and the maximum distance between the clusters that are merged. The algorithm has been used for example in Gomez et al. (2013); van Dijk et al. (2013); Arslan et al. (2012).

The algorithm has the following steps:

1. Initialize a single item cluster for each item and initialize the distances between them.

2. Find the shortest distance separating two clusters

3. Select all groups of clusters separated by this shortest distance and merge them into superclusters.

4. Compute the distances between all superclusters

5. For superclusters containing more than one item calculate the common amalgamation interval, that is the minimum and the maximum distance between two items in the supercluster.

6. Repeat steps 2-5.

It turned out that standard methods to calculate the distance between superclusters are not restrictive enough to ensure a common predictor domain for all supercluster elements. Therefore the distance calculation has been adapted such that distances are calculated considering only domains that are shared by all elements of a cluster.

For clusters having an upper bound of the amalgamation interval that corresponds to zero common predictor domains, the maximum occurrence method is used to split these clusters.

8.3.2 Aggregateability

To aggregate zones we have two possible sets of predictor domains: the ones optimised locally *optim* and the *best* ones as described in Chapter 7. Both of them are used and compared. The first question to answer when aggregating zones by equal predictor domains is where are they actually aggregateable? That is where neighbour zones have predictor domains in common. In the following the term aggregateability describes the potential of two zones for being aggregated by common predictor domains. Figure 8.3 shows the aggregateability for the *optim* set of domains in a graphical representation proposed by Bertin (1967, p. 339) where the dissimilarity of predictor domains corresponds to the line width. The main feature are the black lines along the ERA-40 grid cells, indicating that all five predictor domains differ. Two zones with a different ERA-40 nearest grid cell have always different predictor domains for the local predictors and can never be aggregated. An additional black line is present around the 47.5° N line discussed in Chapter 5 and 7 (compare Figure 7.9) that results from a change in the elementary domain to start the optimisation from. Apart from that, neighbour zones have less predictor domains in common in regions with complex terrain, notably the Massif Central and the southern Alps, but also around Paris, and many common predictor domains in the Aquitaine, Burgundy and Champagne regions.

Figure 8.4 shows that taking the *best* instead of the *optim* domains, the ERA-40 grid does not play its restricting role anymore. There is still low aggregateability around the Massif Central, but high aggregateability in the northern French Alps and the Jura, Burgundy, north of the Massif Central, the Vendée and in Brittany. Interestingly not only the mountains of the Massif Central and the Vosges but also minor hills in Normandy induce low aggregateability. It is expected that using the *best* domains leads to a smaller number of groups, since the number of different domains in the data set has already been reduced to 456 compared to 847 in the *optim* case.

8.3.3 Aggregation results

Figure 8.5 shows the groups obtained with the different aggregation methods described above using the *optim* and the *best* domains. Overall, the groups tend to be smaller in complex terrain, which is in line with Champeaux and Tamburini (1996). The number of groups obtained for each method are shown in Table 8.1. The algorithm designed to minimise the number of groups (SGF and SGB) leads to a smaller number of groups than the other two algorithms. Using the *best* domain set strongly reduces the number of groups for all algorithms. MOC and to some extent VGH lead to a large number of small groups around the lines of low aggregateability shown in Figures 8.3 and 8.4 and large groups in between.

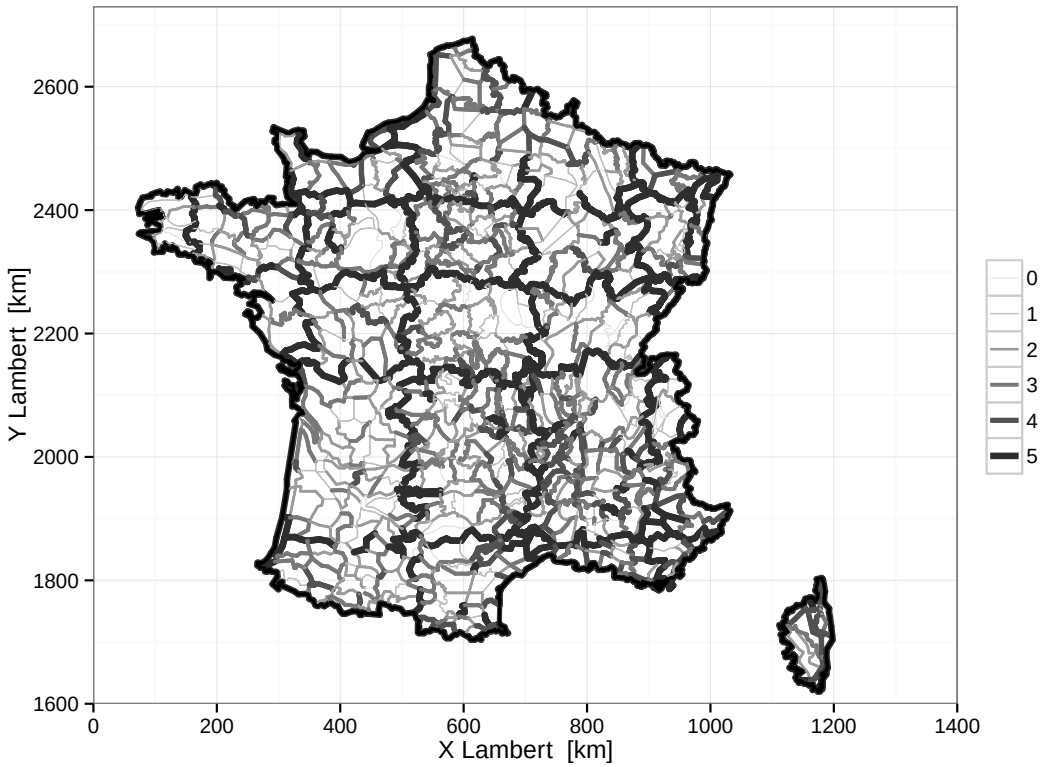

Figure 8.3: Aggregateability of locally optimised domains. The thicker and darker the line between two zones the less predictor domains they have in common.

In the case of *optim* the differences in group size distribution between the methods is not so easily visible on the map representation, therefore the cumulative size distribution of the groups obtained with the different aggregation methods are shown in Figure 8.6. Differences in the size distribution between the SG and the MOC method are expected due to the way they are constructed. As expected, for the MOC method there are more small groups with a size between 15 and 50 grid cells and less between 100 and 200 grid cells. The VGH method has more small groups than the SG method and less groups with a size between 200 and 300 grid cells.

Figure 8.5 shows that groups tend to be larger using the *best* domains than using the *optim* domains. Figure 8.7 shows how the choice of the domain dataset influences the group size distribution for the different methods. The maximum

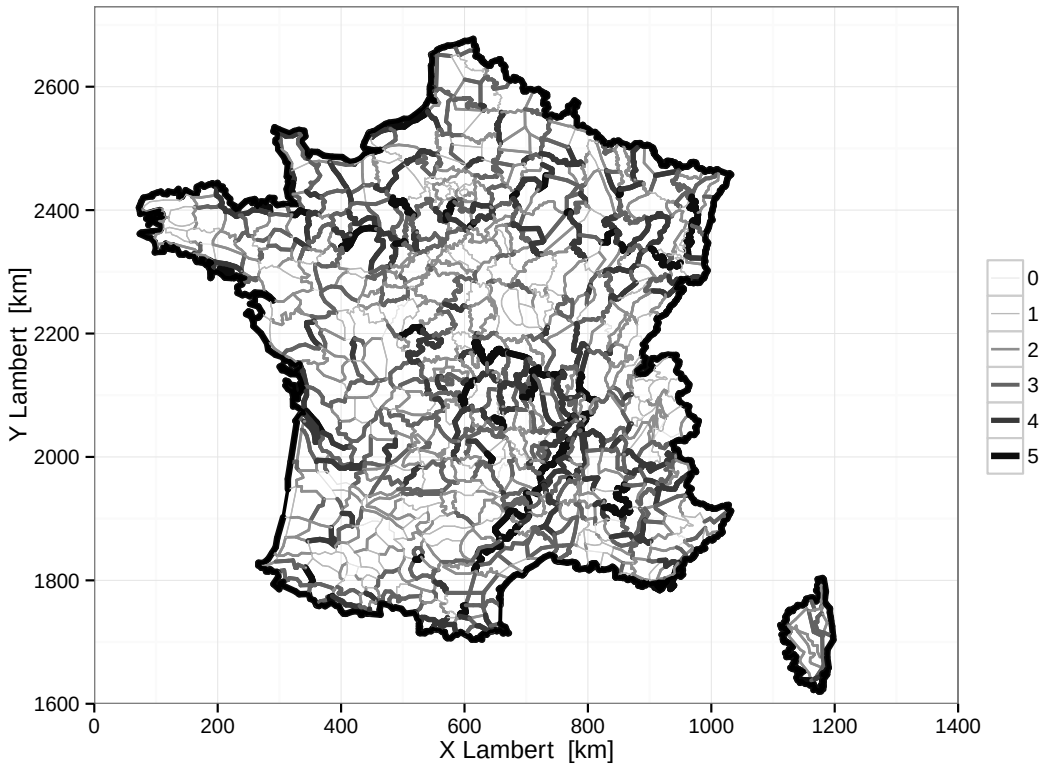

Figure 8.4: As figure 8.3 but for the best domains.

group size increases for all methods, but there are differences between the methods concerning the small groups. The number of groups with a size smaller than 50 grid cells is about the same for the MOC and the VGH methods, but decreases for the SG method. The number of groups with a size between 50 and 250 grid cells decreases more strongly for the MOC and VGH methods than for the SG method.

8.4 Conclusions

Grouping the zones by fraction of common analogue dates with the affinity propation algorithm or by common *best* predictor domain with the variable group hierarchical clustering algorithm leads to nearly the same number of groups (87 vs. 88), but the size distribution of the groups is quite different.

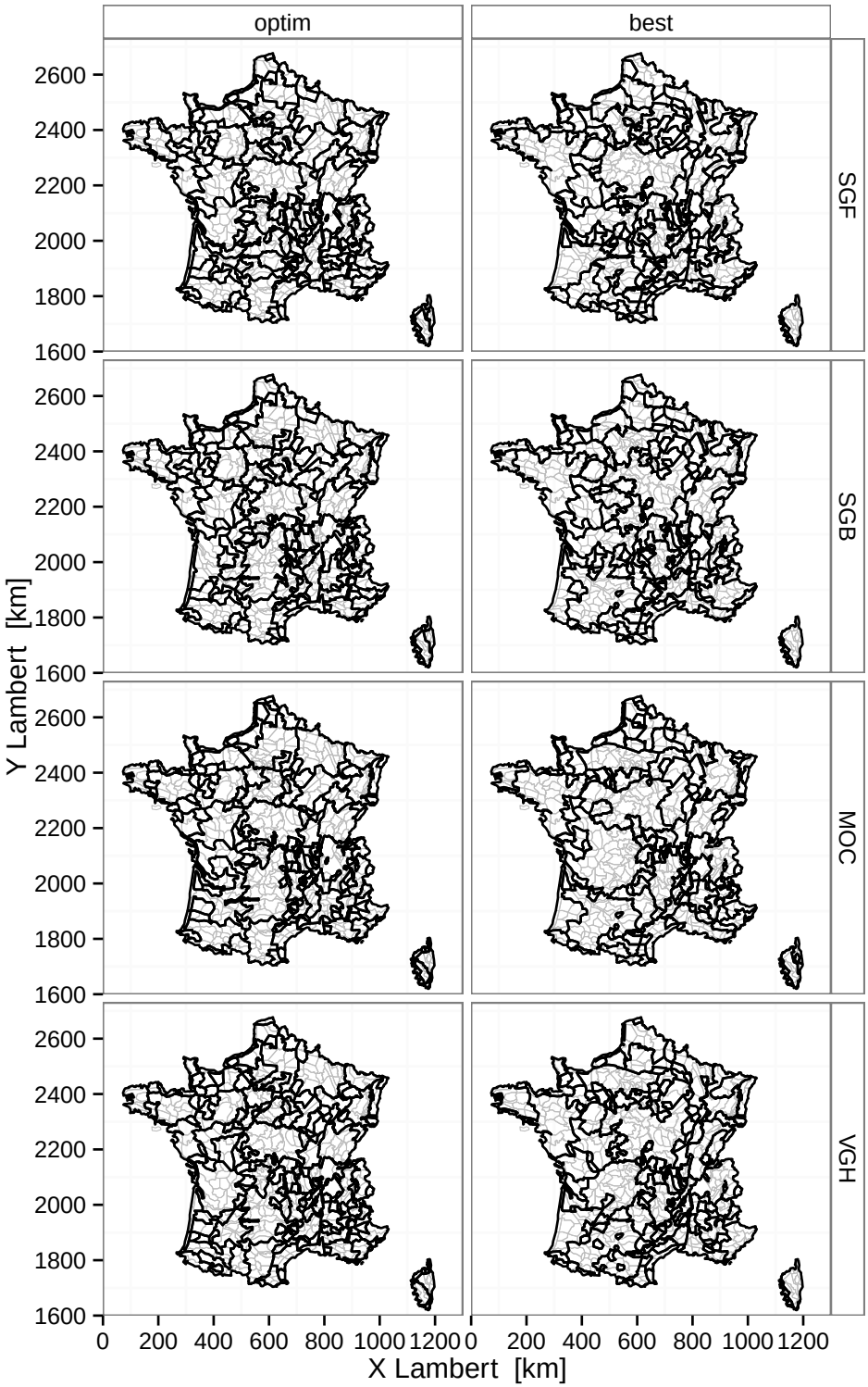

Figure 8.5: Groups obtained with different methods for left: *optim* and right: *best* domains.

	SGF	SGB	MOC	VGH
optim	100	99	131	132
best	61	60	83	88

Table 8.1: Number of groups obtained for each method with *optim* and *best* domains.

Aggregating zones by same predictor domains, the smallest number of groups is obtained using the simple aggregation algorithm even if the exact number depends on the order in which the zones are considered. To test all possible order of input data is not feasible for the whole country, but could be done for some catchments consisting of a few zones.

Aggregating by *best* domains rather than *optim* domains strongly reduces the number of groups obtained and the barriers of aggregateability are physically more meaningful than the ones for *optim* where barriers are somewhat imposed by the predictor grid.

The reduction of the number of groups is in any case mainly due to fewer groups in the flat areas of the country, while the number of groups in the southeastern part of the country with its complex terrain is not substantially reduced. This does not mean that the same groups are obtained there.

With the VGH method superclusters with an upper bound of the amalgamation interval that corresponds to zero common domains have to be splitted. Here this is done using the MOC method for each of the superclusters. This has the advantage that MOC does not depend on the order of the zones and thus having a unique solution, but one could think of other possibilities, for example SG, too.

Which method or which set of groups to prefer will depend on the application and basin to study, but concerning the domain set, *best* is clearly preferred.

Using continuous distance measures like for example the fraction of common analogue dates allows to choose an aggregation level by defining a maximum acceptable distance but the effect on the skill loss can not be controlled. Aggregating large areas without loosing skill is not possible due to the large diversity of parameters. Especially the Durance and the Rhône catchments that are used as case study catchments in Chapter 10 show low aggregateability by common predictor domains. Therefore spatial coherence in the sense of common predictor domains and common analogue dates implies de facto a performance loss in these catchments.

In Part IV a different pathway is explored. Instead of seeking to aggregate, the spatial properties of simulated precipitation fields using common predictor domains or locally optimised ones are measured and the effects on streamflow simulations are explored. Before that the effect of using transformed predictand data in the

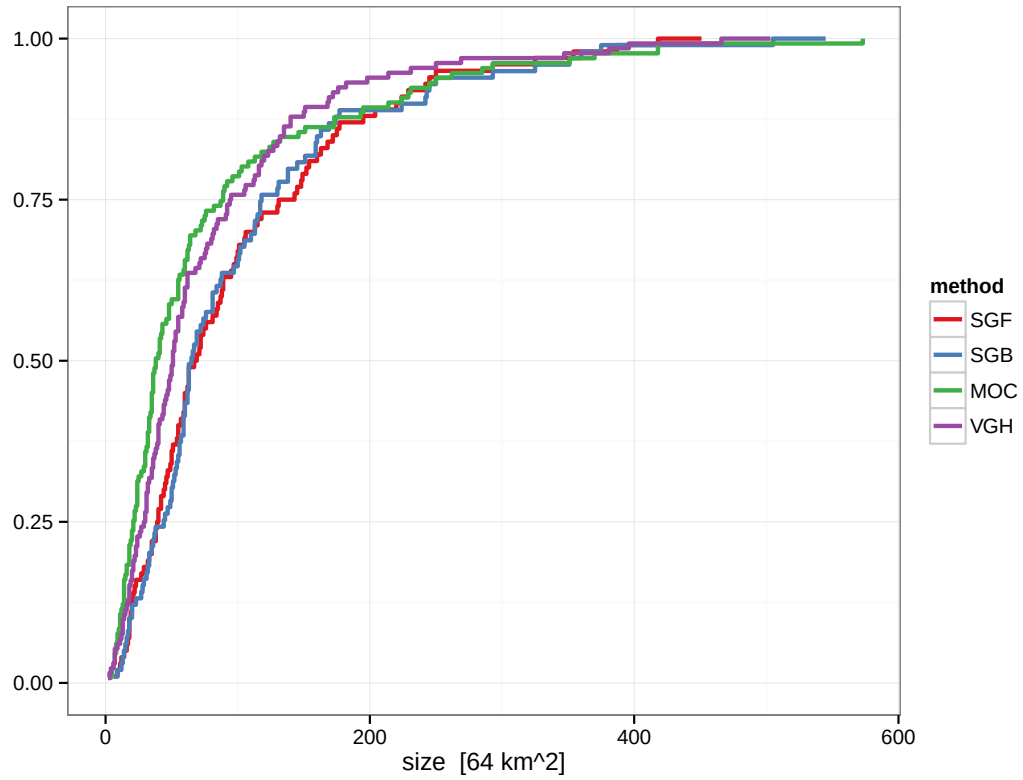

Figure 8.6: Cumulative size distribution of aggregated predictand areas using the *optim* domain set.

optimisation is discussed in Chapter 9.

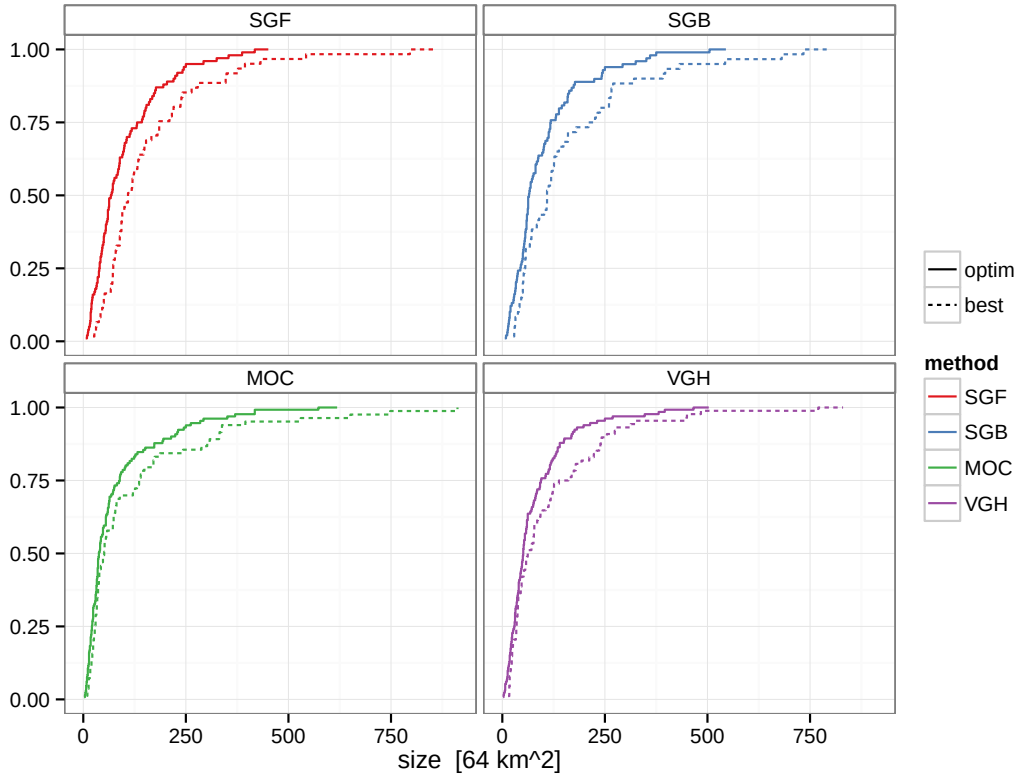

Figure 8.7: Cumulative size distribution of aggregated predictand areas comparing the *optim* and the *best* domain set for each method.

9 Should we use transformed precipitation?

In previous studies on the stepwise analogue downscaling method (Bontron, 2004; Ben Daoud, 2010) predictor domains have been optimised using transformed precipitation as the predictand variable instead of the predictand of interest, raw precipitation. So far the precipitation data has not been transformed in this study, but when comparing the performance scores obtained with the ones from other studies the question arises what influence this choice has on the results. Do the resulting predictor domains differ, and how? How do the CRPSS values differ? Are the validation results better? Are the predictor domains more similar between zones, such that the zones can be more easily aggregated? To answer these questions some of the analysis presented in the previous chapters are repeated using transformed precipitation as a predictand during optimisation and the results are compared.

9.1 Why to transform and how?

In some studies the precipitation predictand variable is transformed, often in order to reduce its skewness. For example Yang et al. (2010) uses the Box-Cox power transformation to be able to take advantage of methods requiring Gaussian data. Khan et al. (2006) and Tryhorn and DeGaetano (2011) use a fourth root transformation prior to a regression analysis and Nicholas and Battisti (2012), Boé et al. (2006) and Rakovec et al. (2012) a square root transformation. Hwang and Graham (2013) uses the normal score transformation to calculate spatial correlations on the normalised variable. Themeßl et al. (2010) tests different predictand transformations for a multiple linear regression downscaling method and found the cube root transformation to be the best performing. The square root transformation is used in hydrology as well to give more weight to low flows in the streamflow simulation performance evaluation (Pushpalatha et al., 2012). On the other hand

Zorita and von Storch (1999) notes that variable transformations can introduce biases in the back-transformed means and variances.

In studies developing or using stepwise analogue downscaling (Bontron, 2004; Ben Daoud, 2010; Horton et al., 2012) predictor domains are typically optimised for groups of stations or catchments and their respective catchment average precipitation. That means that they are looking for domains that have good performance for all the catchments in the group. Using raw precipitation in this case would give more weight to catchments receiving larger precipitation amounts, typically the headwater catchments in the mountains. This was not desired, therefore the precipitation values were divided by the maximum daily precipitation within one year that has a return period of ten years, called adxT10 here. In addition the square root was taken, which changes the distribution of the data. It renders the distribution less skewed and therefore downweights the heavy precipitation events in the optimisation process. Here the same transformation is used such that the transformed precipitation is defined as:

$$Prec_{tf} = \sqrt{\frac{Prec_{raw}}{adxT10}} \tag{9.1}$$

The adxT10 is obtained by fitting a Gumbel distribution to the annual maximum daily precipitation values of the 1959–2009 period. It is calculated with all available data to make its calculation more stable. The adxT10 is calculated for each zone and the values for each zone are divided by their own adxT10. This leads to values that lie nearly always between 0 and 1, except for extremes with a return period of at least 10 years. This is a desirable property for analysis where equal weight should be given to each zone regardless of mean precipitation amounts. For the domain optimisation there is no comparison or interaction between zones, such that this division alone should not effect the result except that the dimension (and the values) of the CRPS are changed. Note that $Prec_{raw}$, and therefore the CRPS calculated with $Prec_{raw}$, has the dimension mm day^{-1} while $Prec_{tf}$ is dimensionless. The square-root transformation changes the distribution of the data, which has the potential to change the local optimisation results. Predictor domains obtained using transformed precipitation as target variable in the optimisation are therefore presented in Section 9.2. The advantage may be that the performance calculation could be more stable with less skewed data. On the other hand optimising for a different predictand variable does not necessarily give the best performance for the actual target variable, which is discussed in Section 9.3.

9.2 Effect of transformed precipitation on domain optimisation

The optimisation described in Section 5.1 has been repeated with transformed precipitation as predictand variable. As for the *optim* experiment 5 near optimum domains for the geopotential predictor are obtained using multiple growing rectangular domain algorithm. The predictor domains for the other predictor variables is the nearest grid cell to the target location and the starting domain for the optimisation is the nearest 2x2 grid point domain. The experiment using transformed precipitation as predictand variable is named *transformed* in the following. The resulting predictor domains are compared with those found using raw precipitation as predictand variable (*raw*) during optimisation which corresponds to the *optim* experiment.

9.2.1 Case study zones

Figure 9.1 shows the optimised domains for the same three case study zones as in Chapter 5 when optimised for raw and transformed precipitation. In general a substantial proportion of the domains in the five domain ensembles are the same for the two optimisations, but they are found in a different order. For the Saône zone 3 of 5 domains are the same, for the Arve zone 4 of 5 and for the Ardèche zone 3 of 5. The domains that differ tend to be larger for the *transformed* experiment than for the *raw* experiment. A possible reason is that heavy precipitation events are related to smaller scale disturbances in the geopotential field and therefore downweighting heavy precipitation makes smaller domains less attractive. The domain characteristics for all zones in France are shown next to see if the tendency to have larger domains for the *transformed* experiment is specific to the selected zones or a more general feature.

9.2.2 France

The first domain characteristic is the mean domain center location of the five optimised domains shown in Figure 9.2. No systematic difference can be seen in the mean domain center location, which is in line with the case studies which don't show any systematic shift of domains. There are changes in the domain center range shown in Figure 9.3. The domain center ranges are smaller, especially in north-south direction, and show less spatial variability in the *transformed* case.

The mean domain size tends to be larger for the domains from the *transformed* experiment compared to those for *raw* as can be seen in Figure 9.4. The overall spatial structure is similar, but the one for the *transformed* experiment has larger

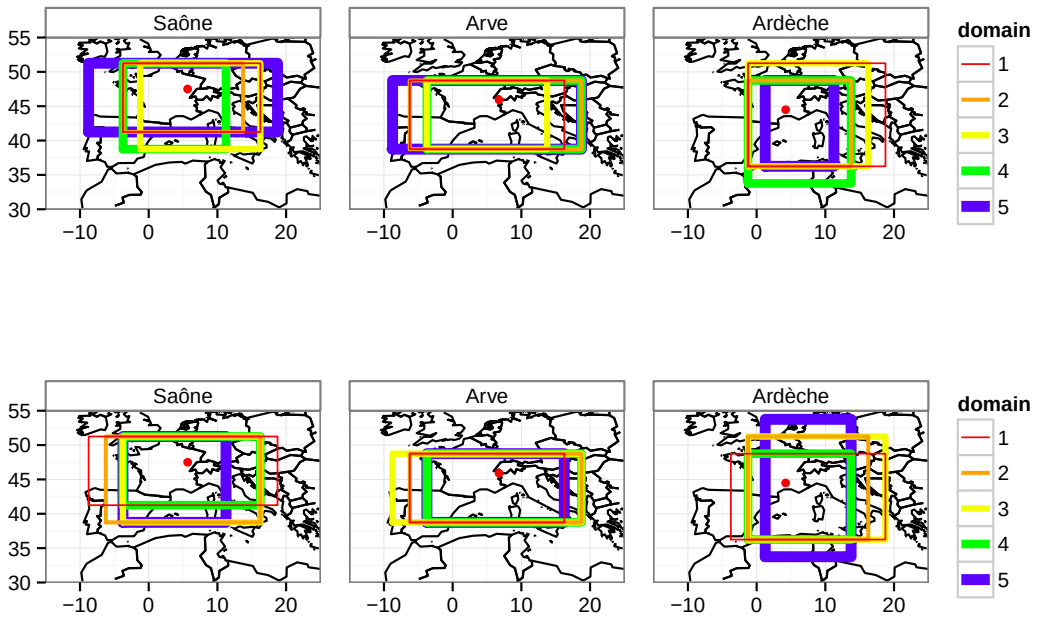

Figure 9.1: Optimised domains for three case study zones from the top row: *raw*, bottom row: *transformed* experiment.

contiguous areas with the same mean domain size than the one for *raw*. The decreased patchiness tends to be in favor of the larger domains around which is in line with what has been seen for the case study zones.

Figure 9.5 shows the domain size range ratio. Remember that the size range ratio is the difference between the largest and the smallest domain extent devided by the mean domain size for a given zone. A ratio of 0 means that all 5 domains have the same length in the given direction, a ratio of 1 means that the range is equal to the mean domain size. The domain size range ratios for the *transformed* experiment are overall smaller than the ones for *raw*. This is partly due to their larger mean extent. The total number of different domains found for the 608 zones is 694 for *transformed* compared to 847 for *raw*.

The results shown in this section indicate that the five domains found for one zone tend to be more similar to each other in the *transformed* case than the ones optimised for raw precipitation. This is in line with Horton (2012) who found only minor differences in the optimised predictor domains for 36 stations in Switzerland using raw and transformed precipitation as predictand. The spatial variability of the mean characteristics is only slightly reduced.

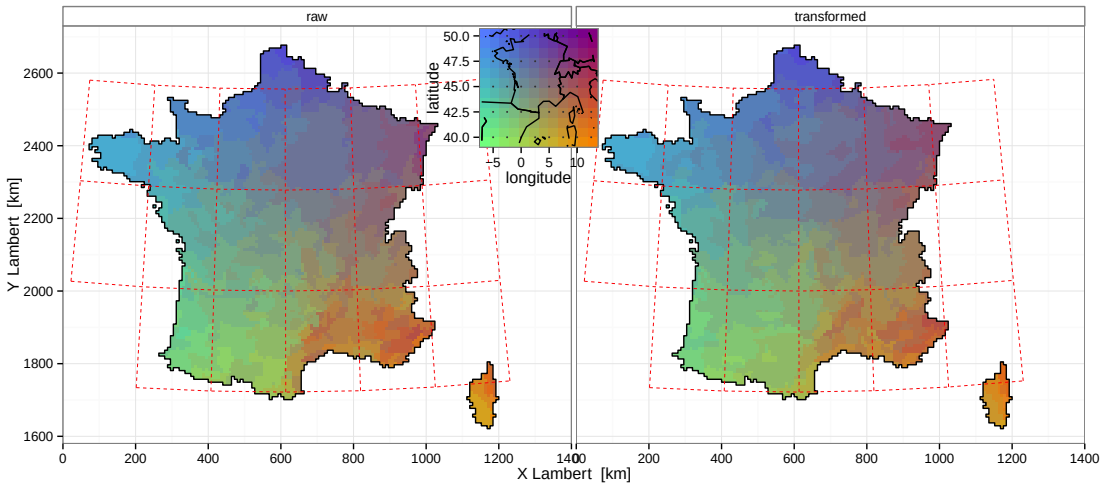

Figure 9.2: Mean domain center location for domains optimised using left: raw, right: transformed precipitation as predictand. The red grid corresponds to the places where the starting domain of the optimisation changes.

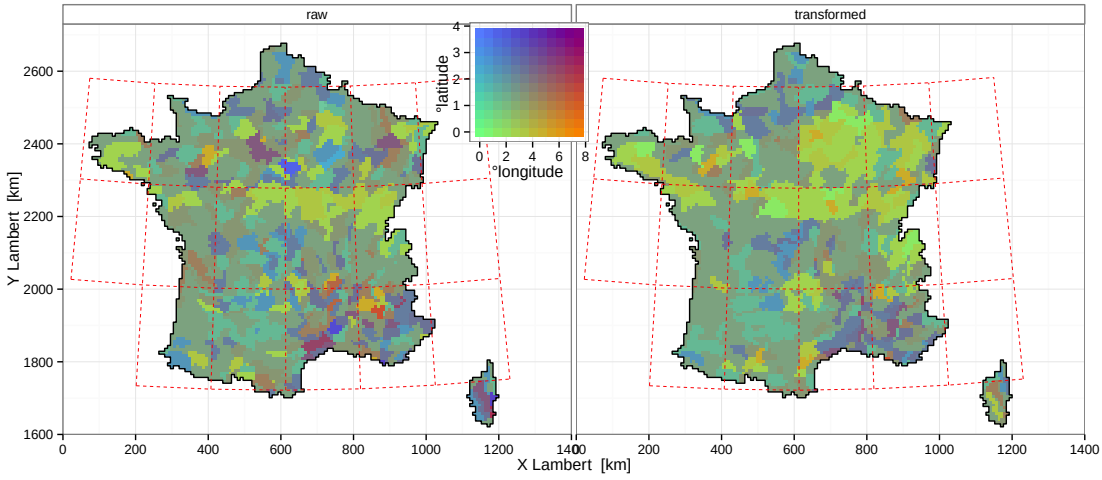

Figure 9.3: Domain center range for domains optimised using left: raw, right: transformed precipitation as predictand.

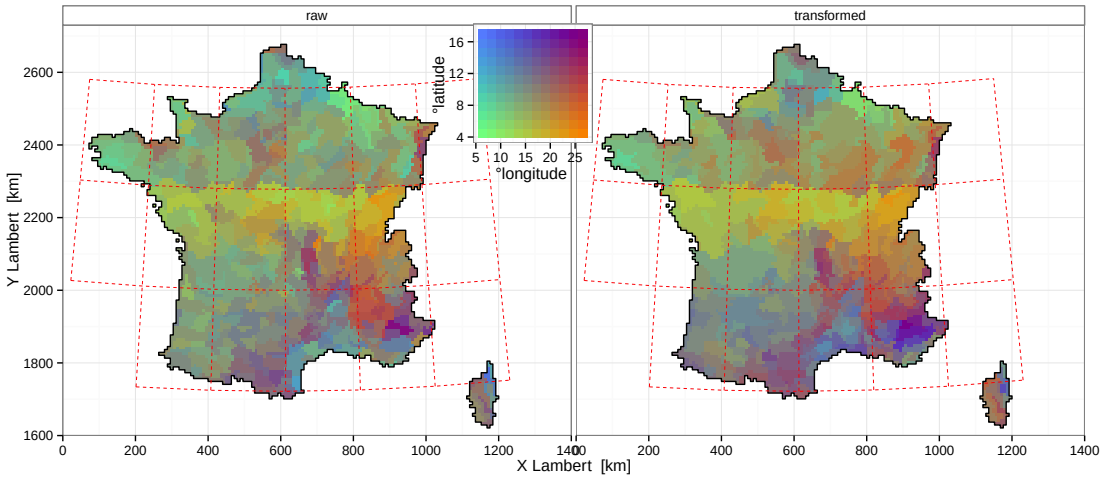

Figure 9.4: Mean domain size for domains optimised using left: raw, right: transformed precipitation as predictand.

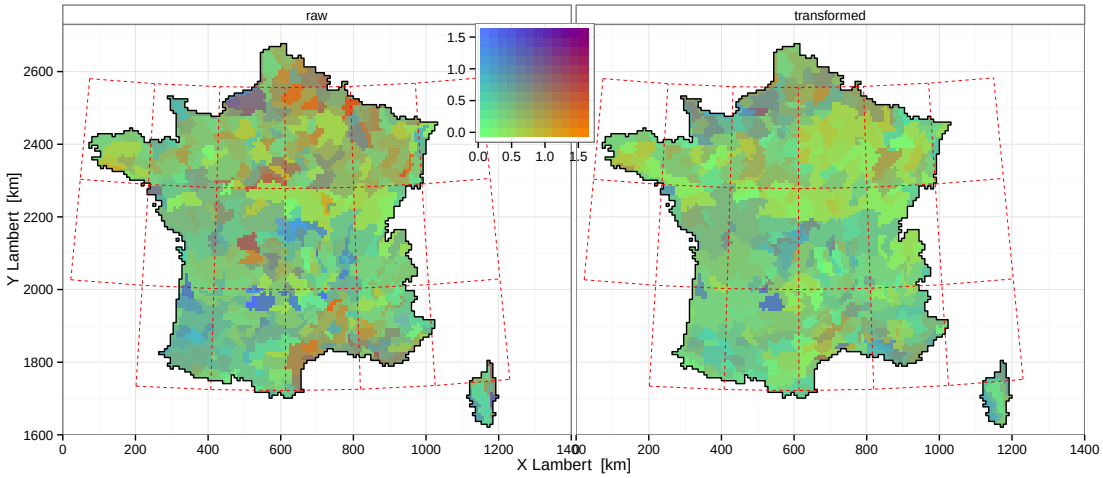

Figure 9.5: Domain size range ratio for domains optimised using left: raw, right: transformed precipitation as predictand.

Let's now look at the impact of the transformation on the skill.

9.3 Effect of transformed precipitation on CRPSS values

9.3.1 Skill evaluated under optimisation conditions

The CRPSS is calculated for the optimisation period using the first optimised domain from the *raw* and the *transformed* experiment. Figure 9.6 shows the resulting CRPSS values when evaluated for the variable the domains have been optimised for. The CRPSS values using transformed precipitation are systematically higher than those where raw precipitation is used, while the spatial structure does not change substatially. This shows that the absolute values of the CRPSS depend on the distribution of the predictand variable. The CRPSS values calculated for transformed precipitation are of the same order of magnitude as in Bontron (2004) and Horton (2012).

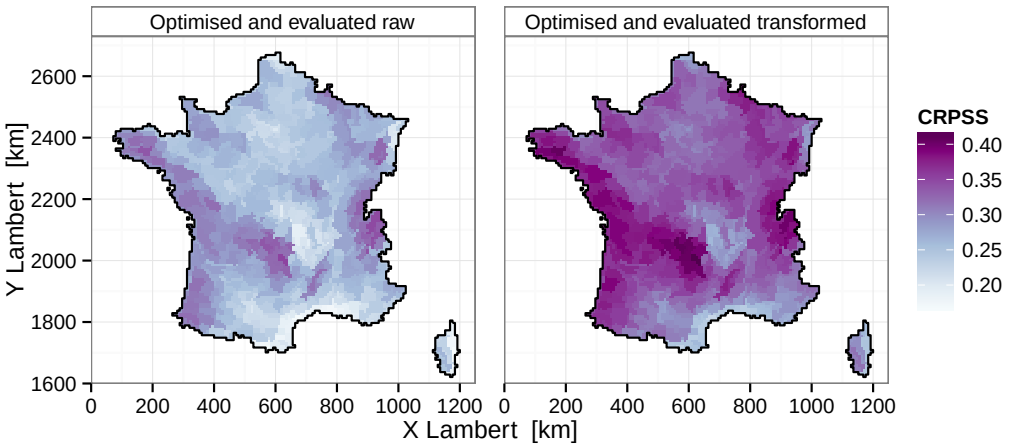

Figure 9.6: CRPSS on the optimisation period for the best domain found. Left: domains optimised and CRPSS calculated using raw precipitation (same as figure 6.3), right: domains optimised and CRPSS calculated using transformed precipitation

But how does the skill behave if the CRPSS is calculated using raw precipitation but with the domains optimised using transformed precipitation? This is an important question because the target variable is raw precipitation so this is

where we would like to have some skill. On the other hand evaluating the skill using transformed precipitation with the domains optimised using raw precipitation might indicate how the domains optimised for raw precipitation would behave for a less skewed target variable. This leads to Figure 9.7 that shows the four combinations described above. On the left hand side the first optimised domains from the *raw* experiment are used, on the right hand side the first optimised domains from the *transformed* experiment are used. In the first row the skill expressed as the CRPSS in simulation mode is calculated using raw precipitation and in the second row the skill is calculated using transformed precipitation. Using transformed precipitation in the evaluation leads to higher CRPSS values, so the skill is higher using a less skewed target variable regardless for which variable the predictor domains are actually optimised.

The difference in skill for different optimised domains are very small as can be seen in Figure 9.8. The figure shows the difference between left: the upper left and the upper right map in Figure 9.7 and right the lower left and the lower right map in Figure 9.7. Where there are differences, they are mostly in favor of the domains optimised for the variable that is simulated, that is domains optimised for raw precipitation are better for simulating raw precipitation and domains optimised for transformed precipitation are better for simulating transformed precipitation. This means that if we seek to have the highest possible skill for raw precipitation, the predictor domains should be optimised for raw precipitation.

9.3.2 Spatial pattern of CRPS and CRPSS

It has been shown that the spatial pattern of the CRPSS is remarkably similar regardless of the variable it is calculated from and regardless of the variable the predictor domains are optimised for. This leads to the question where this spatial structure comes from.

There are two factors that play a role for the CRPSS: the simulation itself and the reference simulation. In all this work the climatological distribution is used as a reference simulation. Figure 9.9 shows CRPS maps for simulating in the top row raw precipitation, and in the bottom row transformed precipitation. In the left column the CRPS for the climatological reference simulation is shown. For raw precipitation the spatial pattern of the CRPS is dominated by the distribution of average precipitation (compare Figure 2.3). The CRPS is higher where there is higher precipitation. This feature is more pronounced for the reference simulation than for the analogue simulation, which leads to somewhat higher skill in terms of CRPSS in the high precipitation regions where it is easier to “beat” the reference simulation. For the transformed precipitation the effect of precipitation amount is strongly reduced due to the division by the adxT10 and what is left is the effect of the proportion of dry days. The CRPS values are lower where there are

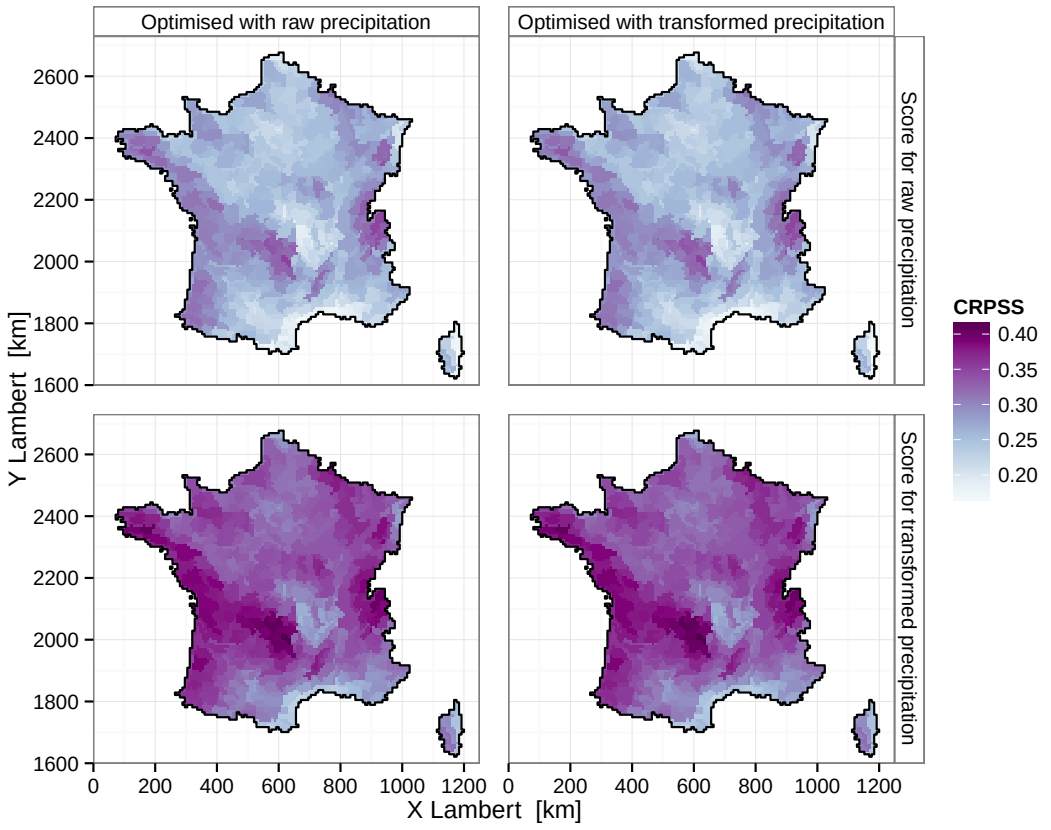

Figure 9.7: As figure 9.6 but completing the matrix. Left: Domains optimised using raw precipitation, right: domains optimised using transformed precipitation. Top: CRPSS calculated using raw precipitation, bottom: CRPSS calculated using transformed precipitation.

more dry days and again the feature is more pronounced for the climatological reference forecast (compare Figure 2.4). This leads to higher CRPSS in places with a smaller number of dry days. Again the differences between the simulations using the domains optimised for raw precipitation or transformed precipitation are hardly visible.

The lower CRPSS values in figure 9.6 along the Mediterranean coast and in foehn prone valleys are due to a high number of dry days and low average daily precipitation, which leads to comparatively small errors of the reference. They are not due to larger errors of the analogue simulation in these places.

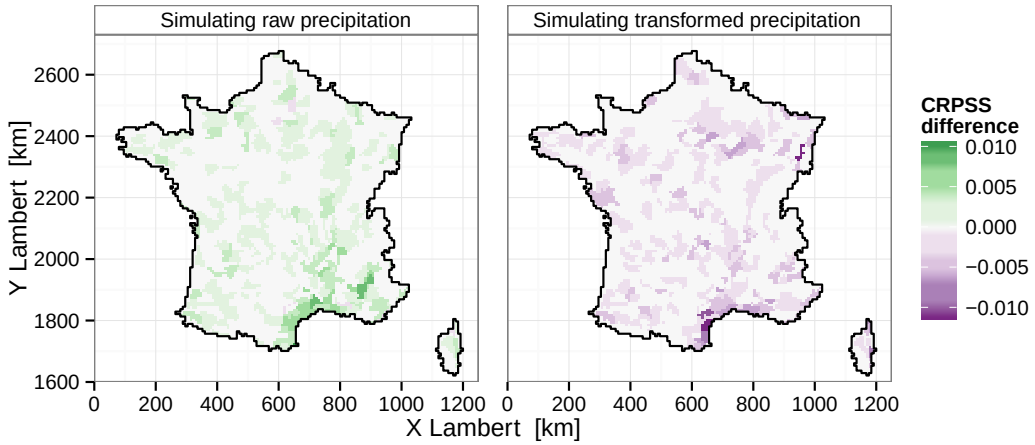

Figure 9.8: CRPSS difference when simulating left: raw precipitation, right: transformed precipitation, using the predictor domains optimised for raw precipitation minus using the domains optimised for transformed precipitation. Positive values: domains optimised for raw precipitation are better, negative values: domains optimised for transformed precipitation are better.

9.4 Do domains optimised for transformed precipitation lead to better validation results?

In the last section it has been shown that the skill of simulating raw precipitation during the optimisation period is slightly lower in the *transformed* experiment. However, this may be compensated by a smaller skill loss in validation. Therefore the validation exercise from Chapter 6 is repeated for the *transformed* experiment. The results are shown in Figure 9.10. On average the validation performance using the predictor domains from the *transformed* experiment is nearly equal to the *raw* one shown in Chapter 6. The smaller skill differences between the reference and the other validation experiments compensate the smaller skill of the reference.

Downweighting high precipitation amounts in the optimisation does not improve the validation results for zone 367 that has been discussed in detail in Section 6.4. Comparing Figures 9.10 and 6.4 it can be seen that the skill loss in the out of sample validation is even more severe using the predictor domain from the *transformed* experiment.

The spatial structure of the bias is very similar between the *raw* and the *transformed* experiment as can be seen comparing Figures 6.6 and 9.11. The area average biases are larger for the *transformed* experiment.

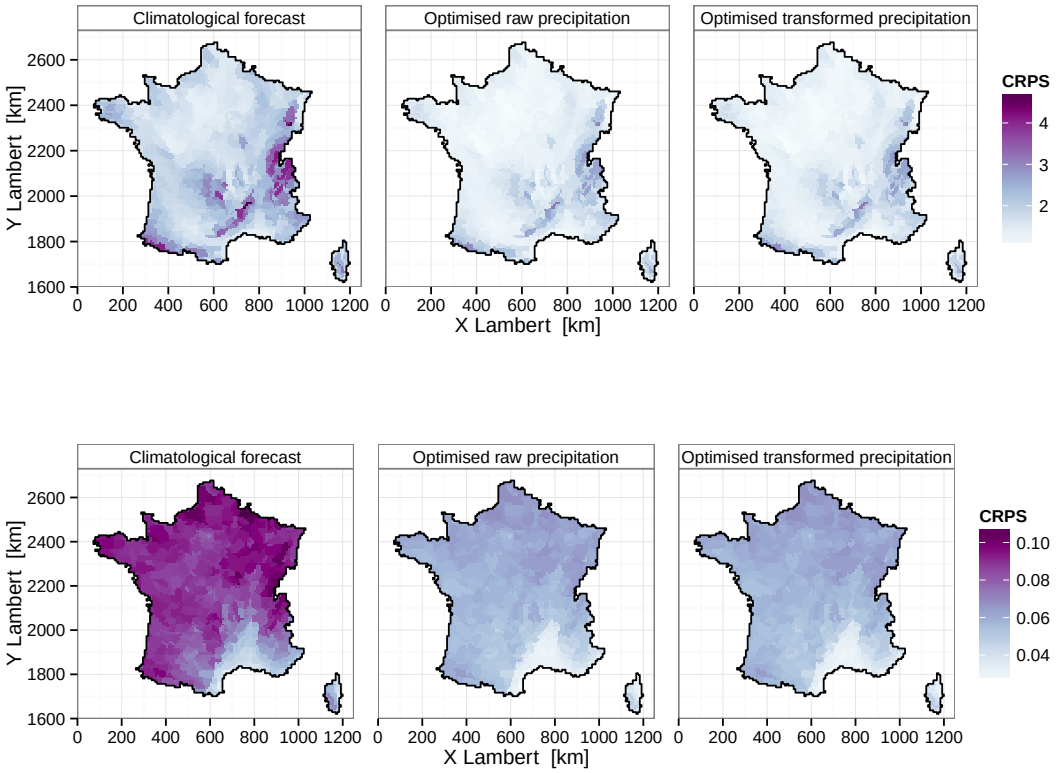

Figure 9.9: CRPS for simulating top: raw precipitation, bottom: transformed precipitation. Left: climatological forecast, center: using domains optimised for raw precipitation, right: using domains optimised for transformed precipitation.

Figure 9.12 shows the biases and the CRPSS obtained in the out of sample validation using predictor domains optimised for raw and transformed precipitation. For both diagnostics the points are in general quite close to the first diagonal and scattered to both sides which indicates equal overall performance. Three zones stand out with larger differences in the bias between the *raw* and the *transformed* experiment. In all three cases the biases are larger in the *transformed* experiment.

Overall, using transformed precipitation for the predictor domain optimisation does not lead to better validation performance.

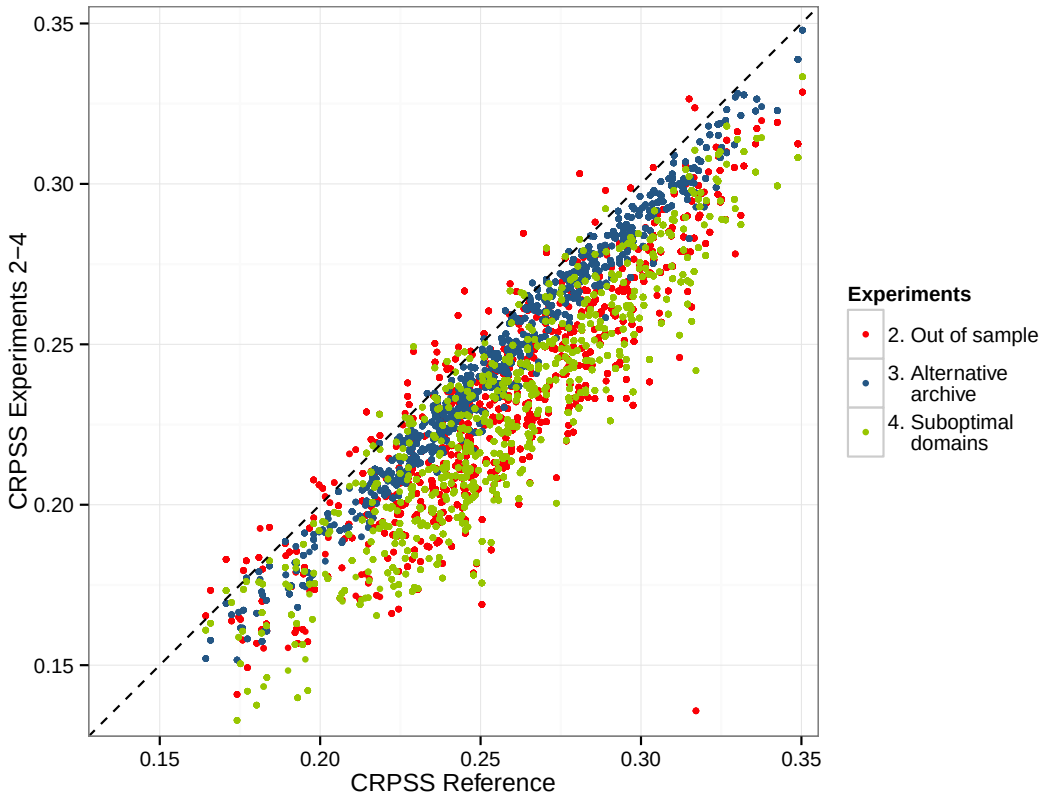

Figure 9.10: The CRPSS of the reference experiment on the x-axis vs. the CRPSS of the validation experiments for all 608 zones. (Same as Figure 6.4 but for the *transformed* experiment.)

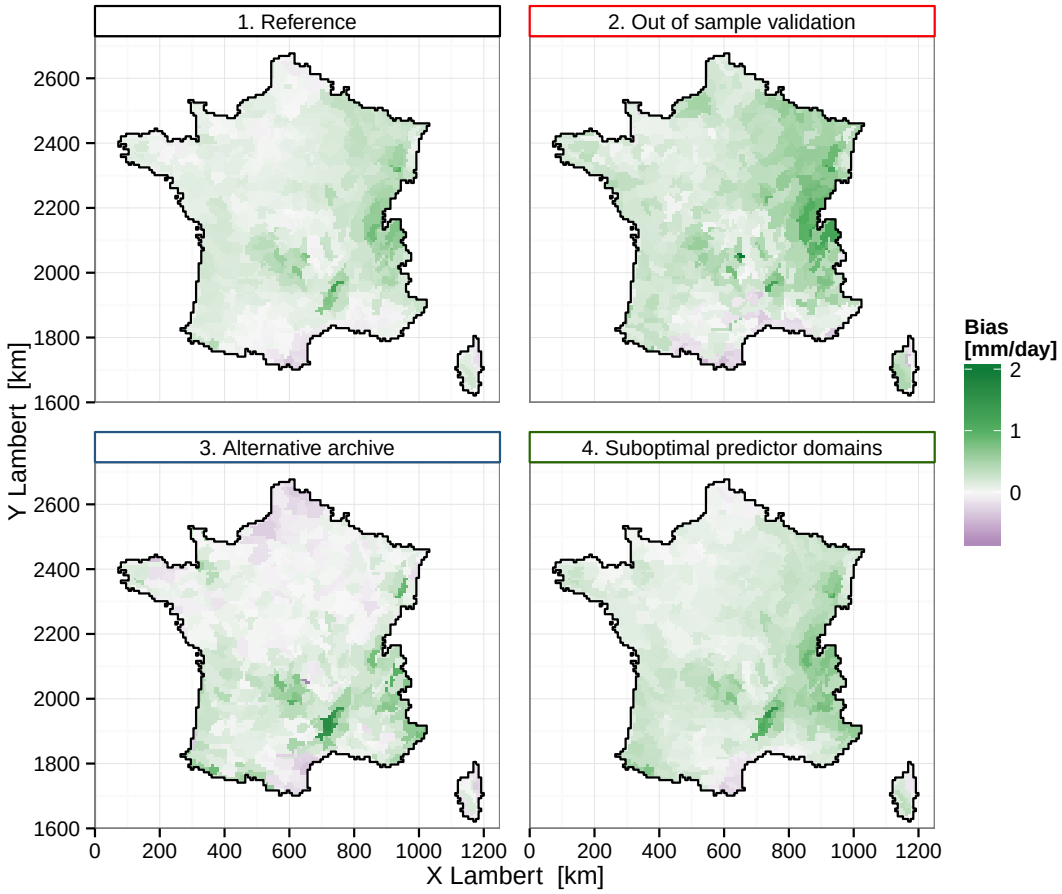

Figure 9.11: The spatial distribution of the bias for the four simulations using the predictor domains from the *transformed* experiment.

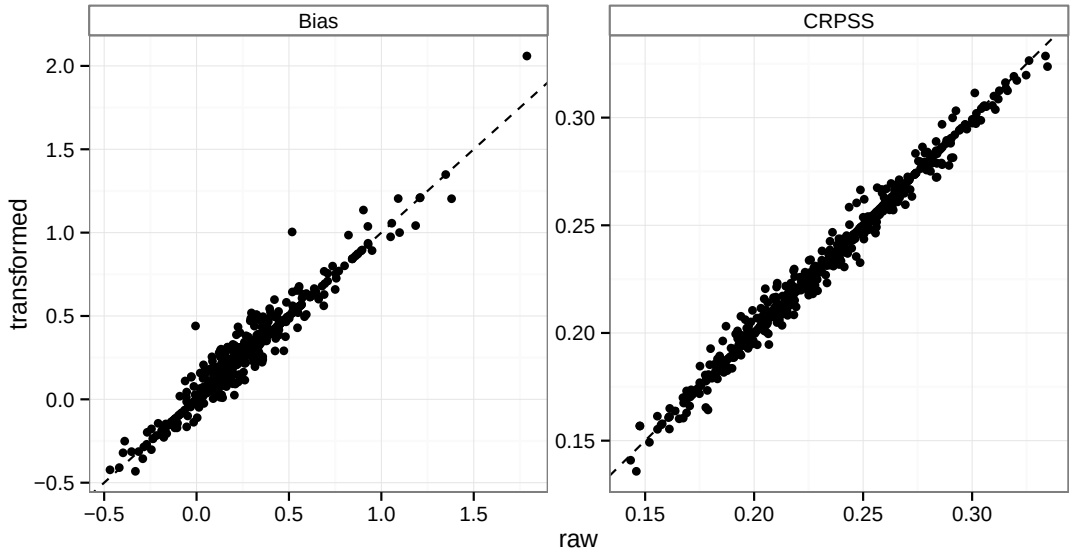

Figure 9.12: Bias and CRPSS in the out of sample validation using predictor domains optimised for raw precipitation (x-axis) and transformed precipitation (y-axis) for the 608 zones.

9.5 Effect on aggregation of predictand areas

In this section the effects of the fewer different domains found optimising for transformed precipitation on the aggregateability is examined. Figure 9.13 shows how many predictor domains are shared between neighbour zones. The same barriers of aggregateability imposed by the predictor grid as for the raw precipitation case shown in Figure 8.3 are visible. The aggregateability inside these grid cells is higher, that is there are more common optimised predictor domains between neighbour zones.

	SGF	**SGB**	**MOC**	**VGH**
raw	100	99	131	132
transformed	85	87	113	102
best	61	60	83	88

Table 9.1: Number of groups obtained for each method with domains optimised using raw and transformed precipitation. Groups from the best experiment discussed in Chapter 8 are added for comparison.

Figure 9.14 shows the groups found with the different aggregation methods pre-

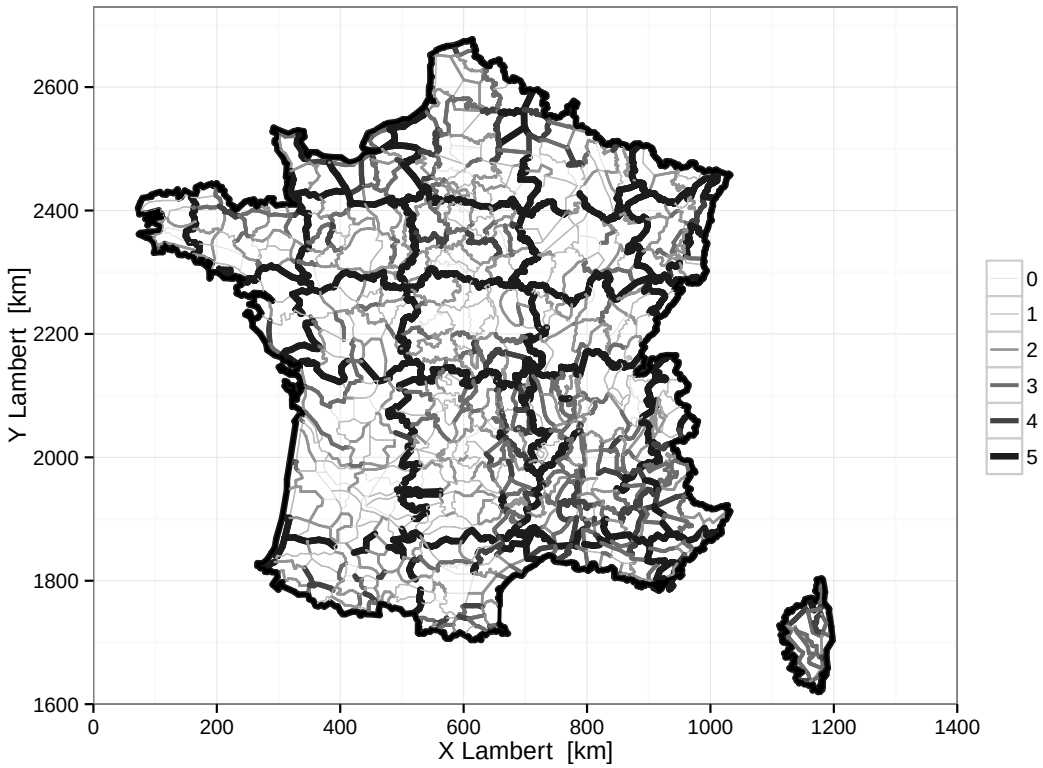

Figure 9.13: Aggregateability of predictand areas using predictor domains optimised for transformed precipitation.

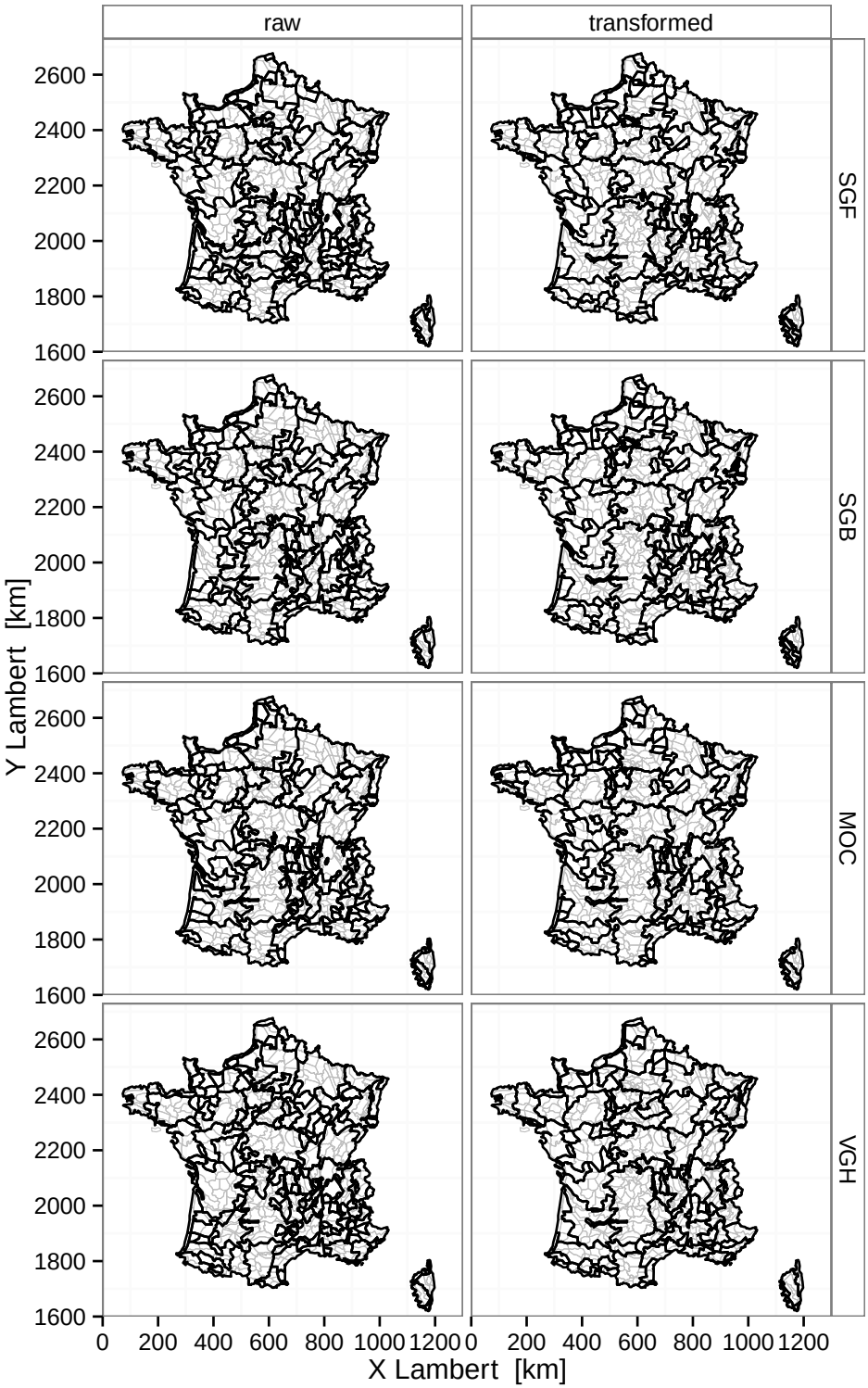

Figure 9.14: Aggregated predictand areas using four different methods for predictor domains optimised for left: raw precipitation, right: transformed precipitation

sented in Section 8.3.1 using the domains optimised for raw and transformed precipitation. Table 9.1 summarises the corresponding number of groups and Figure 9.15 shows the domain size distributions. Using the domains from the *transformed* experiment leads to fewer groups with any of the aggregation methods, with the most important reduction for the VGH method. The reduction of the number of groups obtained with the *transformed* experiment compared to the *raw* one is less than half the reduction obtained with the *best* experiment, except for the VGH method. Remember that the *best* experiment is based on the optimisation using raw data. This is in line with the total number of different domains in the domain data set, that is 847 for the *raw*, 694 for the *transformed* and 456 for the *best* experiment. The differences in terms of size distribution are as follows: For the SG and the VGH algorithm large groups get larger. For the VGH algorithm very small groups get larger as well. For the MOC method the difference in the size distribution is very small.

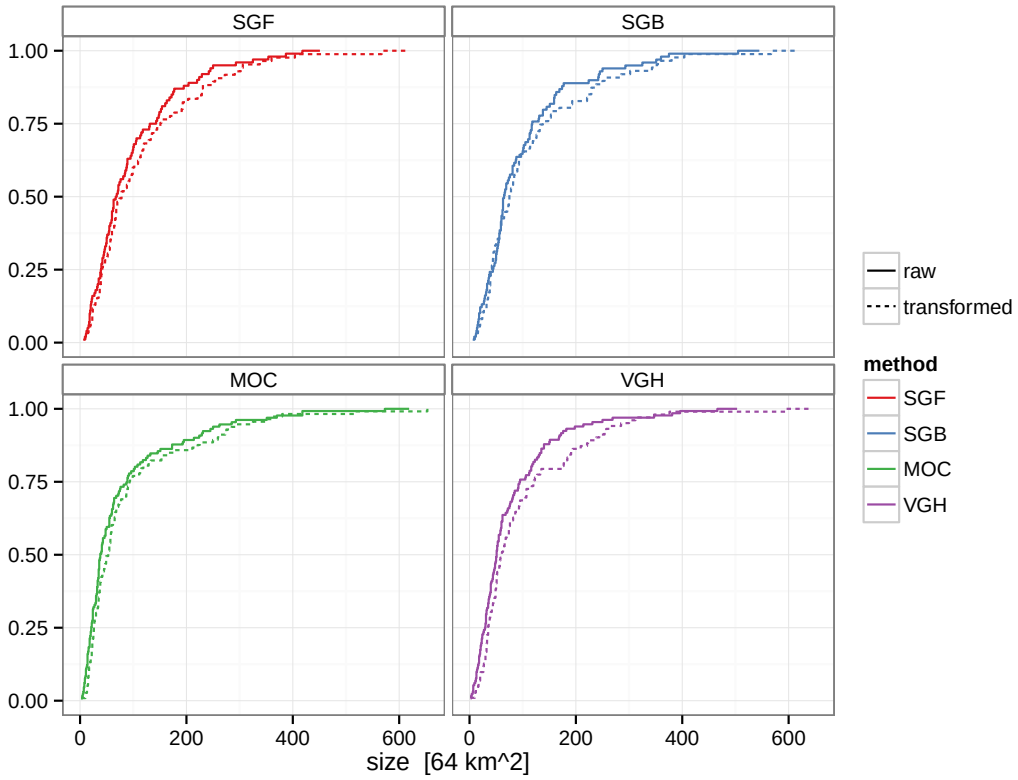

Figure 9.15: Distribution of aggregated area size for four aggregation method and domains optimised for raw and transformed precipitation.

Using transformed precipitation for the predictor domain optimisation im-

proves the aggregateability of predictand areas, but is, as *raw*, limited by the predictor domains for the temperature, vertical velocity and humidity predictor variables that are imposed to be the nearest predictor grid cell to the target location. The predictor grid thus forces a change of the predictor domains where the nearest grid cell changes.

9.6 Conclusions

Calculating the CRPSS for transformed precipitation increases the CRPSS compared to raw precipitation. The spatial structure of the CRPSS does not change regardless which domains are used or for which precipitation variable it is calculated. The low skill found in valleys subject to foehn effects is mainly due to the small errors of the reference simulation in these places.

In the validation experiments using domains from the *transformed* experiment lead to lower average skill loss, but larger biases and larger skill loss for zone 367.

Both the *transformed* and the *best* experiment decreases the number of groups obtained when aggregating predictand areas by common predictor domains. However, *best* leads to stronger reduction and has the advantage to allow for an a posteriori adaptation of the predictor domain for the predictor variables other than geopotential, which is not the case in the *transformed* experiment. A *best transformed* experiment would be possible, but in the light of the increased bias for some zones in the out of sample validation for the *transformed* experiment the *best transformed* experiment has not been conducted in this thesis.

So should the precipitation data be transformed prior to optimisation? If we are interested in spatially smooth results and better aggregateabiliy, yes, if we want to maximise the skill and minimise the bias for precipitation, no, but if we do so the skill decrease will not be dramatic. What we should definitly not do, is to confound the skill calculated for transformed precipitation with the actual skill of precipitation downscaling.

Part IV

Catchment scale spatial verification

10 Spatial Coherence - case studies Durance and Rhône

In this chapter we will assess the spatial coherence of the downscaled precipitation fields. In contrast to other studies, for example Vautard et al. (2013); Maraun et al. (2010); Moron et al. (2006); Robertson et al. (2009); Hwang and Graham (2013), where spatial coherence is used as a synoyme for "spatial smoothness" or "spatial correlation structure", spatial coherence is defined here as realistic spatial properties of the simulated fields. Voisin et al. (2010) used both definitions, defining realistic as realistically looking. Simulating precipitation fields with realistic spatial properties is important for distributed hydrological modelling. Instead of studying spatial correlation an approach based on spatial verification is presented. The spatial verification method SAL (Wernli et al., 2008) characterises features of spatial fields that are relevant for hydrological modelling, but how can SAL, that has been developed and used so far for the spatial verification of deterministic high resolution forecasts, be used to assess probabilistically downscaled fields, here from SANDHY?

The first section of the chapter consists of an manuscript article on the development of a probabilistic version of SAL and its application on downscaled precipitation over the Durance and the Rhône catchment. As a reader of this thesis you may want to skip sections 2.1, 3.1 and 3.2.1 of the article since they mainly contain information already presented in the previous chapters. Furthermore a set of experiments are presented, that aim at figuring out which strategy in terms of predictor domain configuration gives the best local skill and spatial coherence, the more common uniform configuration or fields that are tiled together using locally optimised predictor domains as has been done for example in Hamill and Whitaker (2006) and Voisin et al. (2010) who showed that spatially varying but overlapping predictor domains maintain the spatial consistency of the downscaled fields.

10.1 Measuring spatial coherence (article)

Article to be submitted to the Journal of Climate.

Spatial verification of downscaled ensemble precipitation: a probabilistic version of SAL

Sabine Radanovics, Jean-Philippe Vidal and Eric Sauquet

January 30, 2015

Abstract

With the increasing development of downscaling methods with both multi-site and probabilistic characteristics, spatial verification methods able to handle probabilistic precipitation simulations are required for informing a range of applications like distributed hydrological modelling. Indeed, assessing the spatial coherence, in the sense of realistic spatial properties, of an ensemble of downscaled precipitation fields remains an open question. A probabilistic version of the spatial verification method SAL (Structure Amplitude Location) is here proposed for assessing the spatial properties of precipitation fields, notably the structure and the position of precipitation objects inside a given area. Skill scores are developed based on the structure and location components of the probabilistic SAL, allowing comparison between different ensembles of downscaled fields. The probabilistic SAL is then applied to compare several 20-year precipitation downscaling experiments from the ERA-40 reanalysis to a 8 km resolution over two catchments in France using the Stepwise ANalogue Downscaling method for HYdrology (SANDHY). The experiments correspond to different predictor domain configurations of the SANDHY method and are compared in terms of local performance and proposed structure and location skill scores. Configurations using locally optimised predictor domains lead to higher local performance, higher location skill score but lower structure skill score than the ones using the same predictor domains–and therefore the same analogue dates–over the whole catchment. Both skill scores furthermore depend on the catchment size, as do the standard SAL components. These scores appear as valuable tools for assessing the spatial properties of probabilistic downscaled fields relevant for hydrological impacts.

1 Introduction

Climate change impact studies require precipitation information at the most relevant scale for impact models. However, general circulation models (GCMs) and reanalyses still have a too coarse resolution to resolve such scales, and appropriate downscaling strategies have therefore to be devised. Empirical-statistical methods are often developed and used for downscaling precipitation for local to regional impact studies (Maraun et al., 2010). Several large scale intercomparison experiments involving such methods are currently under way, like the COST Action VALUE (Maraun et al., 2014) or the CORDEX-ESD initiative[1].

Statistical downscaling is typically done pointwise to individual stations/areas, but in order to provide useful input to a range of applications like distributed hydrological modelling, a downscaling method has to provide spatial fields of precipitation (and other variables) that have not only realistic local amounts but also realistic spatial properties. To this aim, more and more methods aim at providing spatially realistic precipitation fields, (either multi-site or gridded, see, e.g., Jeong et al., 2013; Radanovics et al., 2013). At the same time, methods that can provide probabilistic information in form of an ensemble of scenarios or samples from a probability distribution are preferred, in order to take into account the uncertainty due to the downscaling step, identified as the small scale component of internal variability by Hingray and Saïd (2014). Once downscaling models provide local probabilistic output over spatial fields, how can the skill of such models or model set-ups to produce coherent–in the sense of realistic–spatial fields be assessed?

In the context of quantitative precipitation forecasting with high resolution numerical weather forecast models, spatial verification techniques have been developed (see Gilleland et al. (2009) for a comparison). Scale separation (Briggs and Levine, 1997; Denis et al., 2002; Casati et al., 2004) and neighbourhood methods (Zepeda-Arce et al., 2000; Roberts and Lean, 2008; Ebert, 2008) are powerful in determining skillful scales for example. However, the target scale is already determined in the downscaling context, and such methods do not treat displacement and structure errors explicitly. Object-based methods (Ebert and McBride, 2000; Davis et al., 2006; Keil and Craig, 2007; Wernli et al., 2008) can take into account displacement and spatial structure errors at a given scale and are therefore more suited in the context of downscaled fields. One of these methods, SAL (Structure Amplitude Location) (Wernli et al., 2008), has been proposed as part of the verification framework of the COST Action VALUE (ES1102) (Maraun et al., 2014) for the evaluation of downscaling methods, but to

[1] http://wcrp-cordex.ipsl.jussieu.fr/index.php/community/cordex-esd

the authors knowledge it has not been applied to statistically downscaled precipitation yet.

Additionally, whereas more and more downscaling methods are of probabilistic nature (Maraun et al., 2010), very few spatial verification measures have been developed for assessing probabilistic simulations (Jolliffe et al., 2012). Keil and Craig (2007) used for example the Displacement and Amplitude Score (DAS) to rank the members of a short-range ensemble forecast. Such measures are moreover usually applied to the ensemble forecasts regarded as a collection of deterministic forecasts.

The objective of this paper is twofold: first, a probabilistic version of the SAL method initially proposed by Wernli et al. (2008) is introduced along with skill scores corresponding to the location and structure errors. The probabilistic version of SAL is then applied to characterize probabilistic outputs from the Stepwise ANalogue Downscaling method for HYdrology (SANDHY) (Radanovics et al., 2013) over two catchments of different sizes, the Rhône and the Durance catchments in France. Several downscaling model configurations are compared using skill scores derived from structure and location components of the proposed probabilistic version of SAL.

The predictor and predictand datasets, the studied catchments, the SANDHY method and the local performance measure used are presented in section 2. The SAL and the developed extensions of the method are detailed in section 3. An application of the developed scores on downscaling experiments over two French catchments is presented in section 4. Choices, advantages and limitations are discussed in section 5 and conclusions are given in section 6.

2 Downscaling data and methods

2.1 Predictor and predictands

Predictor and predictands for downscaling experiments were similar to the ones used by Radanovics et al. (2013). ERA-40 data (Uppala et al., 2005) at 2.5 deg. resolution were used as large-scale predictors for the statistical downscaling method. Daily precipitation from the French near-surface reanalysis Safran (Vidal et al., 2010) was used as predictand. The downscaling was performed over a 20-year period from 1 August 1982 to 31 July 2002. This is the same time period as has been used for the optimisation of the downscaling method (Radanovics et al., 2013). This avoids influences from the use of different datasets or different time periods on the skill of the method that are beyond the scope of this study.

2.2 Study area

The Rhône river is one of the largest European rivers. The French part of its catchment between Pougny and Beaucaire, as it was considered for this study, has an area of $86000 km^2$ and is indicated by the purple contour in figure 1. The northern part of the catchment experiences oceanic conditions, with heavy rainfall events during winter, whereas the southern part is under a Mediterranean influence, characterized by high temperatures in summer and strong rainfall events in autumn (Etchevers et al., 2002). Orographical enhancement of precipitation occurs due to the Alps and the Jura in the eastern part of the catchment, the Vosges in the northern part and the Massif Central in the southwestern part. Drier regions can be found in the south-eastern and northwestern part of the catchment (Ottlé et al., 2001). The Cévennes mountains at the south-eastern edge of the Massif Central are particularly prone to heavy rainfall events in autumn and resulting flash floods. These heavy rainfall events are due to southerly flow, moisture supply from the Mediterranean sea that is relatively warm in autumn and a conditionally unstable air mass that is lifted due to orography (Duffourg and Ducrocq, 2011).

The Durance catchment is a subcatchment of the Rhône catchment situated in south-eastern France in the Southern Alps. In this study the Durance catchment at Cadarache was considered (red contour in figure 1), which comprises an area of $11700 km^2$. The Durance catchment has large altitudinal and climatic gradients (Magand et al., 2014). Precipitation varies from less than 600 mm $year^{-1}$ in the southwest to over 3800 mm $year^{-1}$ in the north. The spatial variability of precipitation inside the catchment is high due to the complex topography and the different climatic influences, from Alpine in the northern part to Mediterranean in the eastern and southern parts (Lafaysse et al., 2014).

2.3 Stepwise Analogue Downscaling Method for Hydrology (SANDHY)

SANDHY is a downscaling method based on the analogue principle introduced by Lorenz (1969) and performs a stepwise selection of analogue dates using a different predictor variable at each step. The first step is a selection on temperature at two pressure levels. The second step is a selection on geopotential shape at two pressure levels. The third step is a selection on vertical velocity at one pressure level but at four different timesteps and the last step is a selection on humidity. Details on the method can be found elsewhere (Ben Daoud et al., 2011a,b; Radanovics et al., 2013).

Radanovics et al. (2013) optimised the predictor domains for the geopotential predictor for 608 climatologically homogeneous zones in France using a multiple growing rectangular domain algorithm that takes into account the equifinality in parameter optimisation. In this work the resulting predictor domains were used in different combinations as detailed in the experimental setup described in section 4.1.

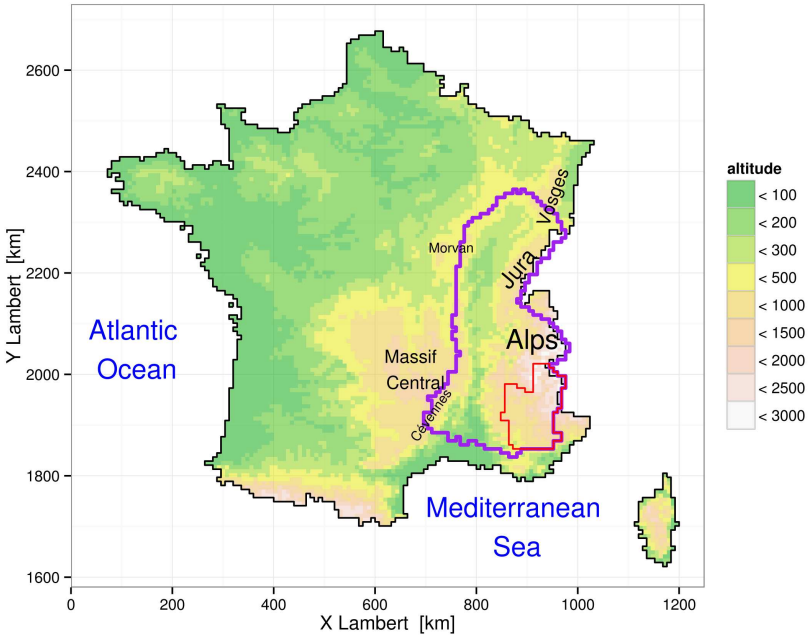

Figure 1: Study region. purple: Rhône catchment, red: Durance catchment

2.4 Local performance measure

The continuous ranked probability score CRPS (Brown, 1974; Matheson and Winkler, 1976) is a probabilistic verification score that is widely used for the verification of ensemble forecasts (see, e.g Hagedorn et al., 2008; Demargne et al., 2010; Aspelien et al., 2011).

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left[F\left(x\right) - H\left(x - x_o\right)\right]^2 dx \qquad (1)$$

where $F(x)$ is the simulated cumulative distribution function of the variable x, x_o the observed value and H the Heaviside function. The properties of the CRPS are as described in Hersbach (2000): The CRPS is sensitive to the entire range of the simulated variable, i.e. it evaluates the entire probabilistic prediction and no predefined classes are required, and it is equal to the mean absolute error (MAE) in the case of a deterministic simulation. It can be interpreted as an integral over all possible Brier scores and it is a proper score (Gneiting and Raftery, 2007). A proper score is one that encourages the forecaster to issue forecasts corresponding to his best judgement, because the score take its ideal value only for ideal forecasts (Murphy, 1969).

The continuous ranked probability skill score (CRPSS) is here used as a measure of the local performance:

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}}{\mathrm{CRPS}_{clim}} \qquad (2)$$

where CRPS_{clim} is the CRPS of the climatological distribution for a 121 days moving window centered on the target day of the year in order to take seasonality into account.

In the context of comparing downscaling model setups in terms of spatial coherence and local performance, the CRPSS was calculated for empirical precipitation distributions that consisted of precipitation from 25 or 125 analogue days depending on the experiment (see section 4.1), the same way as it is typically done for verifying ensemble simulations. A CRPSS was calculated for each grid cell in the catchment and then averaged over the catchment area, thus providing an integrated measure of local performance.

3 Structure, Amplitude, Location - SAL

3.1 Standard SAL

Structure-Amplitude-Location (SAL) is a spatial verification method developed by Wernli et al. (2008) for the verification of high-resolution NWP precipitation forecasts. The method compares three characteristics of observed and simulated precipitation: (1) the amplitude A, that is the total precipitation over the domain, (2) the location L, that is where precipitation objects are located and (3) the structure S, that is the size and the shape of precipitation objects. Precipitation objects are contiguous locations where precipitation is above a given threshold. Unlike methods like the displacement and amplitude score (Keil and Craig, 2009), SAL does not attempt to match individual objects. The difference between the A component of the SAL and error measures like the CRPS is that A is a relative measure and the CRPS an absolute one. In practice this means

that if the simulated precipitation over the verification domain is half the observed one, A will always have a value of -0.66 regardless the absolute amounts, while the CRPS will be higher for cases with higher precipitation amounts because the absolute difference is higher. A unique feature of SAL is the structure component that rates the ability of the simulation to produce precipitation objects with similar size and shape characteristics (scaled volumes). This aims at rewarding simulations that show for example small and peaked precipitation objects, typical of convective precipitation, in case such precipitation is observed.

SAL has so far mainly been used in high-resolution NWP model development (Ahijevych et al., 2009; Zappa et al., 2010; Termonia et al., 2011; Haiden et al., 2011) for the verification of deterministic forecasts but has recently been proposed as part of the verification framework of COST Action VALUE[2] for the verification of spatial properties of downscaled fields.

3.2 A proposed version of SAL for probabilistic simulations

Statistical donwscaling methods designed for probabilistic simulations like SANDHY require probabilistic verification measures. The SAL approach was therefore adapted to fields reflecting the probability that a spatially uniform precipitation threshold u is exceeded locally $P(prec > u)$. For the calculation of the S and the L component, objects need to be defined in the observed and in the simulated fields. They were defined using a threshold t, such that there is an object where $P(prec > u) > t$. t was supposed to be uniform over space and constant in time. If there are no objects in the observed or simulated field, no S and L components can be calculated, as in the standard method. The probability fields were defined independently for simulations and observations:

$$P_s = P\left(prec_s > u_s\right) \tag{3}$$

where P_s is the probability P that the simulated ensemble precipitation $prec_s$ exceeds the threshold u_s, and:

$$P_o = P\left(prec_o > u_o\right) \tag{4}$$

where P_o is the probability that the observed precipitation $prec_o$ exceeds the threshold u_o. If no uncertainty information is available for the observation data used, as it is the case here, the observed probability fields defined by P_o are binary fields of 0 where $prec_o$ is below u_o and 1 where it is above the threshold. The following paragraphs display the SAL equations adapted from those provided by Wernli et al. (2008) for an application in a probabilistic context.

When using directly probabilities into the original definition of the A component, it becomes the normalised difference between the domain average probabilities that u is exceeded:

$$A = \frac{D(P_s) - D(P_o)}{0.5\left[D(P_s) + D(P_o)\right]} \tag{5}$$

where $D(P)$ is the domain average of the exceedance probabilities. The above definition makes the interpretation of A values somewhat difficult in specific contexts, especially when different thresholds are considered for simulations and observations ($u_s \neq u_o$). When simulations suffer from daily biases with respect to the observations, which is most often the case in practice, one may want to remove them before doing the spatial verification, and using different thresholds precisely allows to do that. However in this case quantifying the biases in absolute terms rather than through A values as calculated from the above equation would be also informative. Following this line, a novel and simpler definition of A is therefore proposed, as the relative difference between the thresholds for the observation and the simulation:

$$A = \frac{u_s - u_o}{0.5(u_s + u_o)} \tag{6}$$

Although this novel definition is not based on probabilistic features, it nicely matches the idea of relative bias implemented in the original definition by Wernli et al. (2008) and makes the link with L and S components that rely on the two precipitation thresholds, as described below. It has to be noted that this definition is not informative (A = constant) when the difference between the two thresholds does not vary with time. A typical case is the spatial verification of high precipitation objects defined by an absolute threshold of, say, 10 mm. In such cases, the initial definition would be more suitable as a indicator of the daily amplitude component.

L is the sum of L_1, the normalised distance between the centers of mass of modelled and observed probabilities that are part of objects, and L_2, the difference in the average distance between individual objects and their common center of mass:

$$L = L_1 + L_2 \tag{7}$$

with:

$$L_1 = \frac{|\mathbf{x}(P_s > t) - \mathbf{x}(P_o > t)|}{d} \tag{8}$$

where d is the largest distance between two boundary points of the domain and $\mathbf{x}(P)$ the center of mass, and:

$$L_2 = 2\left[\frac{|r(P_s > t) - r(P_o > t)|}{d}\right] \tag{9}$$

where:

$$r(P) = \frac{\sum_{i=1}^{N} P_i \left|\mathbf{x} - \mathbf{x_i}\right|}{\sum_{i=1}^{N} P_i} \tag{10}$$

and r is the weighted average distance between the centers of mass of individual objects denoted $\mathbf{x_i}$ and the

[2] http://www.value-cost.eu/reports

center of mass of all objects. N is the number of objects in a given field and P_i is the sum of all probabilities belonging to the object i. The definition of the L component thus matches the standard one, simply replacing precipitation objects by exceedance probability objects. L ranges from 0 to 2, where it is 0 for a perfect location.

S is the difference of the weighted average of the scaled volumes of the objects in the verification domain.

$$S = \frac{V(P_s) - V(P_o)}{0.5\,[V(P_s) + V(P_o)]} \tag{11}$$

where $V(P)$ is the weighted average of the scaled volumes of the objects

$$V(P) = \frac{\sum_{i=1}^{N} P_i \sum_{j=1}^{M_i} P_{ij}/P_i^{max}}{\sum_{i=1}^{N} P_i} \tag{12}$$

where P_{ij} is the probability at grid cell j belonging to object i, M_i the number of grid cells in object and P_i^{max} is the maximum probability value in object i. Again, this definition of S matches the standard one. S ranges from -2 to 2, where negative values indicate that the simulated objects are too small or too peaked, while positive values indicate too large or too flat objects, for a perfect structure S equals zero.

3.3 Parameter choices

When applying the probabilistic version of SAL defined by Equations 3 to 12, except equation 5, several choices–further discussed in section 5.2–had to be made. First, different thresholds for simulated and observed precipitation have been considered: u_o was chosen to be the domain average observed precipitation and u_s the domain average of local median of simulated values. This critically allowed to evaluate the structure and location components even in the case of biased simulations.

The threshold for defining objects had been subjectively defined in the early applications of the standard SAL method, as $R_{max}/15$ (Wernli et al., 2008) where R_{max} is the maximum precipitation value over the domain, or $R^{95}/15$ where R^{95} is the "95th percentile of all gridpoint values in the domain larger than 0.1 mm" (Wernli et al., 2009). In the probabilistic version of SAL, a threshold t of 0.5 was chosen, corresponding to situations where the probability of being above the threshold u should be higher than 0.5.

Additionally, in order to penalise experiments that produce lots of missed events or false alarms, a value of -2 was assigned to S and a value of 2 to L if objects were found in the observation but not in the simulation. Similarly, a value of 2 was assigned to both components if objects were found in the simulation but in fact it was a dry day. Otherwise S and L would not be able to distinguish between a dry day and a missed event since in both cases they have no value.

3.4 Spatial coherence measures

This work aims at evaluating whether the spatial features of precipitation were simulated by the downscaling method used. Both the structure and the location components of the probabilistic SAL were considered to carry important information to this end. Therefore scores were developed based on these two components of the probabilistic SAL.

A location score was thus defined as:

$$LS = \frac{\sum_{k=1}^{T} w_k L_k}{2\sum_{k=1}^{T} w_k} \tag{13}$$

and a structure score as:

$$SS = \frac{\sum_{k=1}^{T} w_k \,|S_k|}{2\sum_{k=1}^{T} w_k} \tag{14}$$

where T is the number of time steps in the verification period, w_k the domain average observed precipitation at time step k, S_k the structure component S for time step k and L_k the location component L for time step k. LS and SS range from 0 to 1, 0 meaning perfect location or structure respectively as measured by the proposed scores. The weights w_k were used here with a hydrological objective in mind. Indeed, the SAL components are relative measures and often have large values for very light precipitation events that are of little importance for runoff generation. This approach thus allows to give more weight to important precipitation events.

To assess the performance of the downscaling method in simulating the location and structure of the observed precipitation compared to a naive prediction derived from climatology, a location skill score LSS and a structure skill score SSS were introduced:

$$LSS = 1 - \frac{LS_{fc}}{LS_{clim}} \tag{15}$$

$$SSS = 1 - \frac{SS_{fc}}{SS_{clim}} \tag{16}$$

where LS_{clim} and SS_{clim} are the LS and the SS values for a reference simulation that was constructed using at each point the local climatological distribution for a 121 days moving temporal window centered on the target day of year. As for the CRPSS, perfect simulations have a LSS and SSS of 1, simulations that are just as skillful as the climatology exhibit scores of 0 and less skillful ones have a negative skill score.

4 Application on precipitation downscaling

The daily ERA-40 data from 1 August 1982 to 31 July 2002 were downscaled to a 8 km resolution grid using the SANDHY method. The predictor domains for each climatologically homogeneous zone that contributes to the studied catchments were determined as described in the

Table 1: Experiments. N = number of predictor domains per zone, miss D = Percentage of missed events in 20 years for the Durance catchment, miss R = Percentage of missed events in 20 years for the Rhône catchment. There are no false alarms for any of the experiments. For the Durance catchment there are 24% dry days and for the Rhône catchment 6%.

Experiment	N	miss D	miss R	Symbol
optim	1	14.0%	6.9%	
optim 5	5	18.8%	12.1%	
best	1	14.8%	7.3%	
best 5	5	19.7%	12.6%	
catchment	1	19.7%	12.7%	
catchment 5	5	21.2%	14.7%	
consensus	5	22.7%	-	
5 friends	5	17.9%	-	

following experimental framework in section 4.1. Precipitation at the analogue dates was sampled for each 8 km grid cell using the same analogue dates for all grid cells in a particular zone. The probabilistic SAL was calculated for each day and the location skill score (LSS) and the structure skill score (SSS) for the whole 20yr period. Examples of the probabilistic SAL for some selected days are shown in section 4.2. The LSS, SSS and CRPSS for the Rhône and the Durance catchment for the experiments described in section 4.1 are compared in section 4.3.

4.1 Precipitation downscaling experiments

The downscaling experiments presented here aimed at figuring out which strategies in terms of predictor domain optimisation and accounting or not for equifinality lead to the better trade-off between local performance and spatial coherence. The experiments are described in the following and summarised in Table 1.

4.1.1 Optimised locally

The predictor domains for the geopotential predictor in SANDHY had been optimised for 608 climatologically homogeneous zones in France. Five near-optimum domains showing near-optimal performance during the calibration had been retained for each zone (Radanovics et al., 2013). Analogue dates were therefore found for each zone, and observed gridded fields within each zone for each analogue date were used. Locally optimised parameters are supposed to lead to the best possible local skill at the zone scale but with the risk to loose spatial coherence since precipitation from different analogue dates may be used for neighbouring zones. Two experiments used the locally optimised domains:

optim one locally optimised domain was used for each zone.

optim 5 the precipitation on all the analogue dates from all five domains formed the empirical predictive distribution. Precipitation from a specific date might be considered several times if it was selected with several different predictor domains.

4.1.2 Best in region

Due to some limitations of the domain optimisation algorithm, it can happen that the downscaling skill for a given zone would be higher if domains found for other zones during the optimisation were used. Therefore the skill of any predictor domain found for any of the 608 zones in France had been evaluated for every zone and the best 5 domains had been kept. Two experiments used these best domains:

best the best domain was used for each zone.

best 5 the precipitation from the 5 best domains found for each zone contributed to the local precipitation estimate.

4.1.3 Optimised for catchment precipitation

The predictor optimisation procedure was repeated for catchment average precipitation as predictand variable. Five near-optimum domains were retained. Using the same parameter configuration for the whole catchment assured that precipitation from the same analogue dates was used for the precipitation estimates in the whole catchment. Using the same analogue dates for the whole catchment meant that the complete precipitation fields over the catchment were resampled in contrast to the previously described experiments where the resampling was done independently for the zones in the catchment and the precipitation fields were glued together from different analogue dates. Using the complete precipitation fields was supposed to lead to the best possible spatial coherence. Two experiments used the domains optimised for catchment average precipitation:

catchment analogue precipitation was obtained using one of the five near optimum domains for catchment precipitation. Only the first one was used for the Rône catchment due to computational constraints.

catchment 5 analogue precipitation from all five domains formed the local precipitation estimate.

The *catchment* experiment served as a benchmark since it corresponds to the least costly procedure in terms of optimisation effort.

Due to the smaller size of the Durance catchment it was computationally feasible to perform two additional experiments for this catchment. They are described in the following two subsections.

4.1.4 Consensus zone

The fraction of common analogue dates between each of the 608 climatologically homogeneous zones in France was calculated using the analogue dates from the 5 locally optimised domains for each zone. The fraction was calculated for each day and then averaged over the 20 years simulation period. This was slightly different from the calculation in Chardon et al. (2014), who calculated the percentage of days on which the daily fraction exceeds 0.8.

consensus The zone in the catchment that had the highest fraction of common analogue dates with all the other zones in the catchment was identified. The five domains that had been optimised for this consensus zone were used for the whole catchment in the *consensus* experiment.

As in the *catchment 5* experiment, the same five domains and therefore the same analogue dates were used for the whole catchment.

4.1.5 5 friends

Like in the *optim 5* and *best 5* experiments the predictor domain configuration in the *5 friends* experiment was different for each zone but it was tried to introduce some kind of "spatial smoothing" at the predictor domain level by including for each zone some predictor domains found for other zones.

5 friends The first optimised domain of the target zone and the first optimised domains of the 4 other zones that had the largest fraction of common analogue dates with the target zone were taken.

This experiment was supposed to lead to a higher number of analogue dates shared by the zones in the catchment but had the disadvantage that the analogue dates from one domain might be used several times in the same precipitation estimate, leading to overconfident precipitation estimates. On the other hand the high spatial variability of the predictor domains in the Durance catchment found in Radanovics et al. (2013) should limit this risk.

4.2 Probabilistic SAL examples

Three contrasted situations in terms of probabilistic S and L values in the *catchment* experiment were first selected. Figure 2 shows the observed values and the mean of the 25 simulated precipitation values at each grid cell for the Rhône catchment for the three selected days and the *catchment*, *best* and *optim* experiments. Figure 3 shows the probability objects used for the calculation of the S and L components of the probabilistic SAL. The S and L values for the three experiments and the three selected dates are summarised in Table 2.

Table 2: Probabilistic S and L scores for the three example dates.

	1995-01-26	1995-06-30	1999-05-19
S catchment	-0.00	-1.31	-0.04
S best	-0.52	-1.32	0.02
S optim	-0.33	-1.39	0.01
L catchment	0.09	0.09	0.76
L best	0.36	0.12	0.71
L optim	0.12	0.25	0.77

The left panel in figures 2 and 3 is for the 26 January 1995 where heavy precipitation was observed in the northern part of the catchment and specially high amounts in the Vosges and the Jura mountains. In figure 2, the maximum values are with 105 mm off the scale and appear as grey pixels on the graph (figure 2). The basin average precipitation was 7.3 mm.day^{-1}. The simulations showed precipitation over the whole catchment with the highest amounts in the northern and eastern parts. Due to the averaging over the 25 analogue situations constituting the simulations, the simulated precipitation fields are smoother than the observed ones. The simulations from the different experiments showed similar mean precipitation fields but in the *optim* and *best* experiments less precipitation than in the *catchment* experiment was simulated in the south-western part of the catchment.

For the 26 January 1995 the simulated probability field from the *catchment* experiment in the left panel of figure 3 shows a structure quite similar to the observed precipitation in figure 2, while the observation is essentially a large, of course flat, object in the north of the catchment. The resulting values for S and L are very small, that is very good, with $S = -0.0001$ and $L = 0.09$.

At the first glance the simulated probability fields from the *best* and *optim* experiments were similar to the *catchment* one except that they were more patchy. This was expected, because their precipitation fields were glued together spatially from different analogue dates, while in the *catchment* experiment the same days were used for the whole catchment. The S scores were much worse than for the *catchment* experiment for both other experiments and the L score was much worse for the *best* and only slightly worse for the *optim* experiment. For the *optim* experiment the probability object was more peaked but at approximately the same place than for the *catchment* experiment, which led to a negative S of -0.33. For the *best* experiment the object was split up into several objects in the south which led to too small objects (negative S values) and large distances between the objects and the common center of gravity that explain the L value of 0.36.

The second situation shown in the middle panels of figures 2 and 3 is the 30 June 1995. The basin average precipitation was 1.4 mm.day^{-1}. Precipitation was mainly

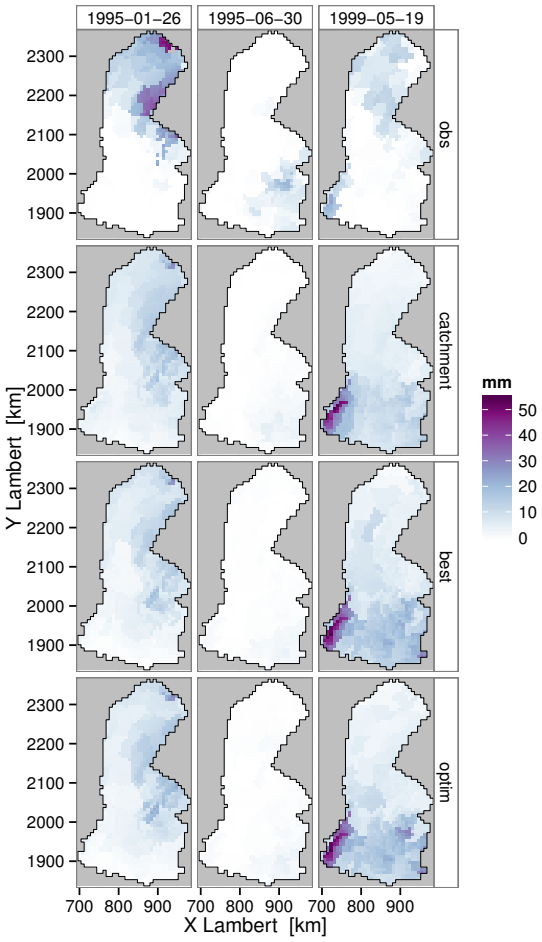

Figure 2: Precipitation over the Rhône catchment top: observed, second row: mean over 25 analogue simulations from the *catchment* experiment, third row: from the *best* and bottom row from the *optim* experiment. Left: 26 January 1995, middle: 30 June 1995, right: 19 May 1999.

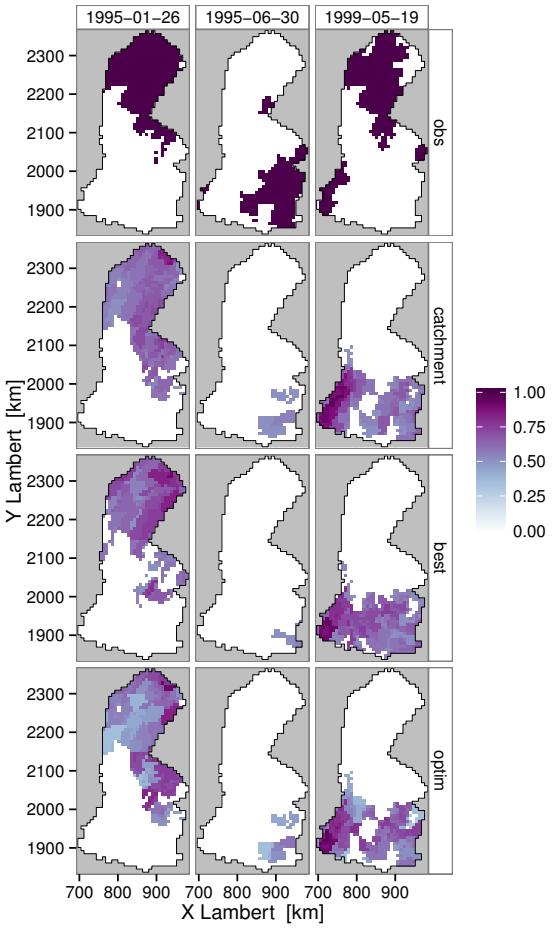

Figure 3: As figure 2 but for the probability objects used to calculate the S and L components of the probabilistic SAL.

observed in the south-eastern part of the catchment, but with some precipitation in the Jura and Cévennes mountains as well (see figure 2). The simulations showed rain mainly in the south-eastern part of the catchment and only very little in the Jura and Cévennes mountains, such that the probability objects were situated only in the south-eastern part and the objects were small compared to the observed ones (figure 3). Due to the correct center of mass of the simulated probability objects, the resulting values for L were small with $L = 0.09$, 0.12 and 0.25 for the *catchment*, *best* and *optim* experiments respectively. The too small size of the simulated precipitation objects led to a large negative error in the S component with $S = -1.31$ for the *catchment* experiment and even -1.39 for the *optim* experiment.

The third situation shown in the right panels of figures 2 and 3 is the 19 May 1999. On this day with a basin average precipitation of 3.8 mm.day^{-1} heavy precipitation was observed in the Cévennes mountains in the south-western part of the catchment, in the northern part of the catchment and in the Alps (figure 2). In the south-eastern part of the catchment precipitation was observed, but the amounts were below the catchment average and thus no "object" was defined there as can be seen in figure 3. Precipitation was simulated for the whole catchment with the highest amounts in the Cévennes mountains and in the the south-east and less precipitation in the north. This led to simulated precipitation objects in the southern part of the catchment and no simulated object in the northern part. The error in S was small in this case for all three experiments with S values between -0.04 and 0.02, since the size of the objects was similar. The location errors were large with L between 0.71 and 0.77 for the three experiments due to a large object simulated in the south-eastern part while there was one observed in the northern part of the catchment.

These three examples confirm that both the S and the L components carry information relevant for distributed hydrological modelling and therefore it is worth looking at both corresponding scores.

4.3 Results from the downscaling experiments for the Rhône and the Durance catchment

The downscaling experiments *optim*, *optim 5*, *best*, *best 5*, *catchment* and *catchment 5* described in section 4.1 were performed over a 20 yr period for the Durance catchment and the Rhône catchment. The additional experiments *consensus* and *5 friends* were computationally possible for the Durance catchment due to the smaller size of the catchment.

For each experiment the catchment averaged local CRPSS, the SSS and the LSS were calculated. What was looked for are experiments that had a comparatively high CRPSS and SSS or LSS respectively, showing a good trade-off between local performance and spatial coherence. Higher CRPSS would be expected for the experiments using location specific domains (*local*, *local 5*, *best*, *best 5*, *5 friends*) and higher SSS and LSS for experiments using the same predictor domain configuration for the whole catchment (*catchment*, *catchment 5*, *consensus*). Figure 4 shows the CRPSS on the x-axis and the SSS (left column) or LSS (right column) on the y-axis for the different experiments. In the first row results are shown for the Durance- and in the second row for the Rhône catchment. The most skillful experiments are those closest to the upper right corner of each panel.

Let's first look at the experiments using only one predictor domain per zone (*optim*, *best* and *catchment*), represented by small symbols in figure 4. The small green squares correspond to the *optim* experiment and the small red circles to the *best* experiment. By construction *best* had somewhat higher CRPSS than *optim* for both catchments. The local skill for these experiments was the same for the two catchments. This means that the Durance catchment was representative for the Rhône catchment in terms of possible local skill (cf. Radanovics et al., 2013, fig. 5). On the other hand *best* had equal or smaller SSS and LSS than *optim*, so spatial properties for the *optim* experiment tended to be somewhat better. This result may be explained by the fact that some of the *best* domains were found outside the catchment, bringing some more variability of domains – and therefore less spatial coherence – in the *best* experiments.

The small blue symbols correspond to the *catchment* experiment. Domains optimised for the Durance average precipitation were used for the Durance catchment (blue triangles in the first row) and domains optimised for the Rhône catchment average precipitation were used for the Rhône catchment (blue triangles in the second row). For the Durance catchment the five domains optimised for catchment average precipitation were used individually leading to the five small blue triangles while for the Rhône catchment only the first domain was considered due to computational constraints. As expected, the CRPSS was smaller than for the *optim* experiment, the skill difference being larger for the larger Rhône catchment than for the smaller Durance catchment, because for the larger catchment the domains optimised for the catchment average precipitation are locally less representative. The *catchment* experiment led to larger SSS than the *optim* experiment, but to smaller LSS. Note that depending on which of the *catchment* domains were used for the Durance catchment, the SSS and LSS varied considerably, but in the same direction for the SSS and the LSS, that is domains that led to better SSS led to better LSS as well.

Now let's look at the experiments using analogue dates from 5 predictor domains per zone (*optim 5*, *best 5*, *catchement 5*, *consensus* and *5 friends*), represented by the large symbols in figure 4. The CRPSS and the SSS increased for the *optim 5*, *best 5* and *catchment 5* experiments compared to their single domain counterparts for both catchments, while the LSS increased only slightly for the Rhône catchment and decreased for the Durance catchment. The relation between the three type of experiments did not change for the LSS if five domains were used compared to one single domain. In terms of CRPSS they were slightly closer for the 5 domain version than for the 1 domain version and they were much closer in terms of SSS, that is the differences between the 5 domains and the 1 domain version of the experiment was larger for the *optim* and *best* experiment than for the *catchment* experiment. The *consensus* experiment led to slightly worse results than the *catchment 5* experiment for all three skill scores. The *5 friends* experiment had similar skill as the *catchment 5* experiment, but lower SSS, the SSS was only slightly higher than for the *optim 5* experiment. On the other hand *5 friends* was the experiment with the highest LSS. Note that the differences between the experiments were smaller in terms of LSS than in terms of SSS.

Overall, the LSS and SSS for the Durance catchment were higher than for the Rhône catchment, with the SSS for the *optim* and *best* experiments being even negative for the Rhône catchment. This was due to the fact that the probabilistic S and L did not measure the same scale of phenomena because of the different catchment size. The phenomena at the scale of the Rhône catchment, notably the orographic enhancement along the main mountain ridges, are well captured by the climatology. The typical size of the objects was of the order of a whole mountain ridge and the variations inside an observed object were neglected due to the way the observed probability fields were built. For the Durance catchment, which is entirely located in the Alps, the question was not any more if it rained in the Alps, but if it rained more on the southern-, southwestern- or northwestern slopes or if the precipitation was rather of convective or of stratiform nature. These details were averaged out in the climatology and made it less skillful.

As expected the use of locally optimised domains led to higher skill, while the use of uniform predictor domains over the whole basin led to higher structure skill scores. Interestingly location skill scores tended to be higher for the locally optimised domains. Using ana-

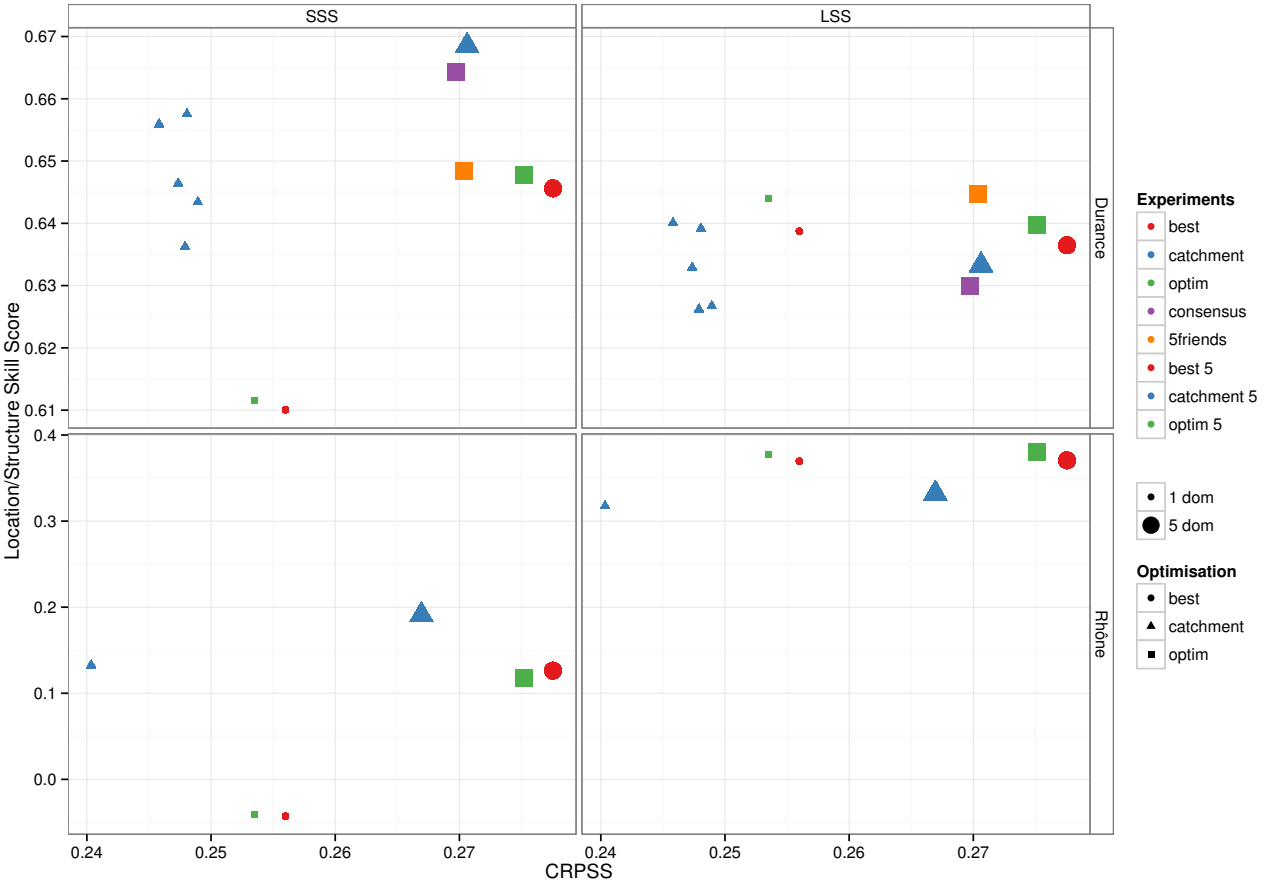

Figure 4: Left: *SSS* vs. CRPSS and right: *LSS* vs. CRPSS for the experiments defined in section 4.1 for top: the Durance catchment, bottom: the Rhône catchment. Small symbols correspond to experiments with one domain per zone, large symbols to experiments with 5 domains per zone. Squares: locally optimised domains, cercles: the best domains found searching through the domains optimised locally for any zone in the whole region, triangles: domains optimised for catchment average precipitation. Shapes, sizes and colours colours correspond to the experiment type as defined in table 1.

logue precipitation from multiple predictor domains together increased the local skill and the SSS for all experiments, the *LSS* decreased for the Durance catchment but increased for the Rhône catchment. An exception was the *5 friends* experiment, where multiple predictor domains were used, but it had the highest *LSS* of all experiments for the Durance.

The number of missed events was higher for the experiments that have a high *SSS*, that is for experiments using analogue precipitation from multiple predictor domains or the same predictor domain for the whole catchment. For the Durance the numbers range from 14% for the *optim* experiment to 22.7% for the *consensus* experiment. The percentages for the Rhône catchment are smaller and range from 6.9% for the *optim* experiment to 14.7% for the *catchment 5* experiment. This is due to the larger size of the catchment in the sense that there are fewer days where no precipitation at all is observed in the whole catchment and there are fewer days as well where no precipitation is simulated. The percentage of missed events is summarised for all experiments in Table 1.

5 Discussion

5.1 Probabilistic versus deterministic SAL

As most spatial verification methods, SAL has so far been applied only for deterministic forecasts and simulations. Figure 5 illustrates the issue considering the precipitation on the analogue dates as a collection of deterministic simulations or summarized as a fixed quantile. It shows the S and the L values for the three case study days for the Durance catchment calculated using single realisations, fixed quantiles and the final probabilistic version. The 25 deterministic realisations were widely dispersed for all three days and the two components, sometimes spanning nearly the whole range which precluded conclusive results. Note that the S and the L values are not necessarily similar for the Durance and the Rhône catchment (the latter being given in Table 2,

because the SAL components depend on the verification domain as found by Wernli et al. (2009).

In the case of the fixed quantiles both S and L values strongly depend on the specific quantile taken, as can be seen from the example days in figure 5. Figure 6 shows SAL diagrams (Wernli et al., 2008) for daily simulations over 20 years for the Durance with the *best* experiment and taking the 0.5 or 0.6 quantile of the empirical distribution of the analogue precipitation respectively as simulation. The median of A was strongly negative for the 0.5 quantile and much less negative for the 0.6 quantile. For the S component the median was slightly negative for the 0.5 quantile and slightly positive for the 0.6 quantile. The same was true for other experiments (not shown). The choice of the quantile was an arbitrary one and may favor some of the experiments over others.

The above attempt to consider the precipitation on the analogue dates from the SANDHY method for the Durance catchment as a collection of deterministic simulations did not lead to conclusive results. Furthermore this approach rises the question of how to combine the different realisations to form a score. Besides, transforming the output from the SANDHY method into a quasi-deterministic simulation by taking a fixed quantile from the output distribution leads to strong smoothing and loss of information which is not desirable either. This led to the choice of applying SAL on exceedance probability fields rather than precipitation fields. This choice respects the nature of SANDHY output, but adds the question of the choice of the threshold to be exceeded.

5.2 Probabilistic SAL choices

In the case of probability fields the observation became a binary field and thus the observed objects were completely flat. This arises questions about the S component. It would be desirable to have some uncertainty information for the observation in order to create non-binary probability fields for the observation. One could for example follow Theis et al. (2005), who transformed deterministic simulations into probabilistic ones using neighbourhood information, but this could strongly smooth the field depending on the size of the neighbourhood.

Some applications focusing on specific absolute precipitation threshold may want to use a fixed precipitation threshold u. As the study here aimed at looking at the overall behaviour of the downscaling method in terms of spatial properties of downscaled fields, a time-varying threshold depending on the catchment average precipitation of the day was chosen. Using a fixed threshold would lead to a strong dependency on the total precipitation of both the size of the objects and the number of cases for which the probabilistic S and L can be calculated. Using a threshold depending on precipitation allowed here to include a higher number of cases for calculating the scores.

The second threshold to be chosen was the exceedance probability to define the objects. Again the size and the number of cases that can be studied depend on it. Results, and especially the S component, could strongly depend on the choice of the two thresholds. Sensitivity tests on theses thresholds could shed light on these dependencies, but are beyond the scope of this paper.

5.3 Time average of SAL components

For the structure score and location score developed in this paper a weighted time average of SAL components was used. But is this a valid approach? Wernli et al. (2008) proposed a SAL diagram that summarises the three SAL components (see an example in Figure 6). The SAL diagram is a valuable tool for comparing fundamentally different models, but has its limits for comparing different configurations of the same model type that have similarly looking SAL diagrams. Furthermore SAL diagrams for good models are not necessarily symmetric, because for example light rain events that are overestimated typically have large values of S and A, leading to an accumulation of dots near the upper right corner while underestimated light rain events often fall in the missed event category and are not represented on the graph. The time average does not entirely solve the problem, but the introduced weights downweight the light precipitation events. This became particularly important as missed events and false alarms were taken into account by assigning extreme values of S and L to them in order to penalise configurations that tend to miss larger precipitation events.

5.4 Structure score or location score?

The structure score used in this work was based on the structure component of SAL, thus comparing the size and the shape of the observed and the downscaled precipitation probability objects but not their location which is an important information for distributed hydrological modelling as well and was accounted for by the location score. Concerning the relevance of the SAL components for hydrological modelling, Wernli et al. (2008) noted that the L component describes how accurately the precipitation is distributed over the subcatchments, so that errors in the L component are supposed to lead to random errors in the runoff simulation while errors in the spatial structure of precipitation, as expressed through the S component, are supposed to affect the repartition between surface runoff and infiltration and therefore affect the soil water balance. For the experiments in this study the SSS and the LSS showed opposed results. Experiments using the same predictor domain configuration for the whole catchment tended to have higher SSS and those using locally optimised domains higher LSS. For the Rhône catchment the LSS and SSS tendencies were similar to the ones for the Durance catchment, but the differences between the experiments were about three times larger than for the Durance catchment.

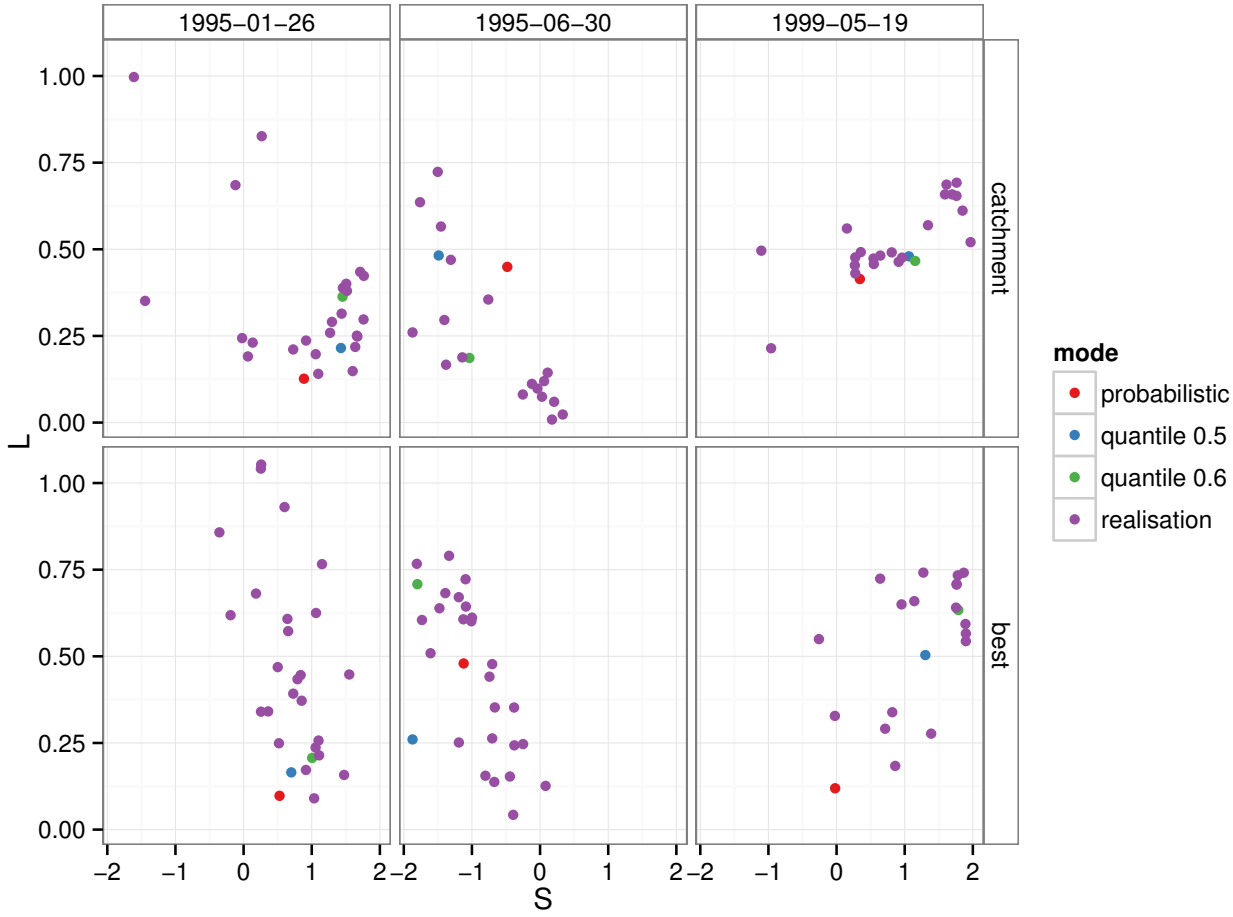

Figure 5: S and L components for three selected days for the Durance catchment from red: probabilistic SAL, purple: for single realisations interpreted as deterministic simulations, blue: the 0.5 quantile and green the 0.6 quantile of the empirical distribution for the respective day. Top row: for the *catchment* experiment, bottom row: for the *best* experiment.

The location skill score varied only little between the experiments, especially for the Durance catchment. A possible reason is the size of the catchment compared to the spatial resolution of the precipitation data. The analysis on the Durance catchment was done using 255 grid cells which might be not enough to distinguish differences in location errors. On the other hand the structure skill score might be more sensitive to the choice of the thresholds and more affected by the completely flat observed objects of the probabilistic SAL.

5.5 Definition of the Region for the *best* experiments

For the example case presented in this study we had at hand not only predictor domains for the catchment area itself but for 608 zones covering France, which allowed for the *best* experiment described in section 4.1 that led to a modest skill improvement compared to the locally optimised domains. A limitation of this approach is that the study catchments are located at the eastern border of France, and no predictor domain information was available from outside France. This could possibly penalise locations close to the border. If no predictor domain information from outside the catchment is available a possible alternative approach is to evaluate the best domains inside the catchment only. In this case locations close to the catchment borders could be penalised. Furthermore this approach will only lead to improvements for catchments that are sufficiently large compared to the predictor field resolution.

5.6 Range of applications

The structure- and location skill scores developed in this work require the calculation of SAL components on a day by day basis, which means that they can be applied to reanalysis driven experiments, reanalysis nudged GCM runs and pseudo-reality experiments (Maraun, 2012). They can therefore be a valuable tool to assess the spatial properties of downscaled ensemble fields, independently of the downscaling method considered. It can be applied for any probabilistic downscaling method that provides spatial fields, including future large reanalysis-

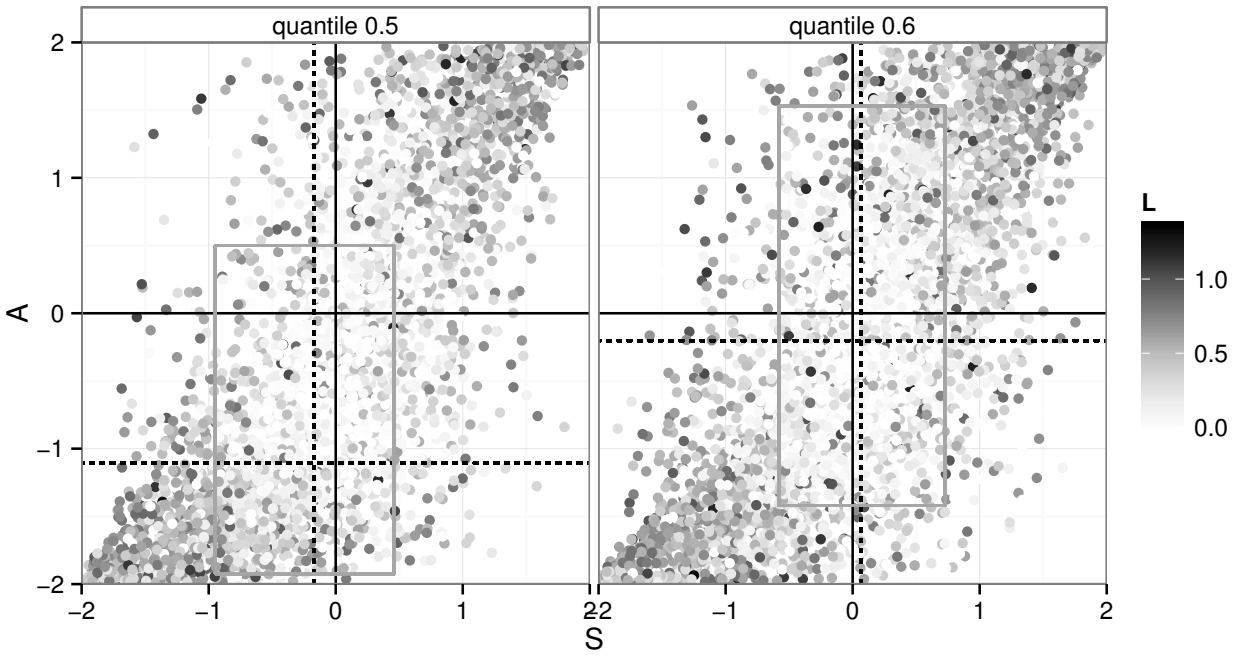

Figure 6: SAL diagrams for the Durance catchment, the *best* experiment and a 20 year simulation period for taking left: the 0.5 quantile and right: the 0.6 quantile of the 25 analogue precipitations at each gridpoint as simulation. Dashed lines: median of the S and A components, grey rectangle: inner quartile range of S and A.

driven RCM ensembles sampling RCM internal variability (Braun et al., 2012).

6 Conclusions

This paper proposed a probabilistic version of the Structure-Amplitude-Location (SAL) method for the spatial verification of ensemble downscaled precipitation fields. Two scores were derived from the structure- and the location component of the probabilistic SAL for assessing the spatial properties of probabilistic precipitation simulations.

The probabilistic SAL method and the associated spatial coherence scores were then applied to several statistical downscaling experiments. The experiments were performed with several configurations of the Stepwise ANalogue method for HYdrology (SANDHY) over a 20-year period over two nested catchments in France. The downscaling experiments considered here aimed at figuring out which strategies in terms of predictor domain optimisation lead to the better trade-off between local performance and spatial coherence.

The probabilistic SAL result depend on the catchment considered, as do the one for the standard SAL. The proposed structure- and location skill scores indeed depend on the catchment size, because for larger catchments the phenomena represented by the S and the L components are better represented by the climatological distribution than for smaller ones.

The experiments using analogue precipitation from multiple predictor domains outperformed the ones using only one predictor domain in terms of local performance. The two spatial coherence scores showed contrasting results over the range of downscaling experiments. If a good structure skill score is most important, a configuration using homogeneous predictor domains over the catchment (*catchment 5*) is preferred. On the other hand, configurations with predictor domains optimised locally (*optim 5* or *5 friends*) are recommended if location skill score is given priority. Runoff simulations using a distributed hydrological model are currently under way in order to decide which of the two skill scores proposed (LSS and SSS) is actually more relevant for runoff generation in a given catchment.

Acknowledgements

The authors would like to thank Météo-France for providing access to the Safran database and ECMWF for providing access to the ERA-40 data. The presented work was part of a PhD project co-funded by Irstea and the Compagnie Nationale du Rhône (CNR). SANDHY has been implemented in Fortran 2003 with Fortran 90 api for NetCDF (Pincus and Rew, 2011). Analysis and graphics are done using R software environement (R Development Core Team, 2012) with the packages ncdf4 (Pierce, 2013), SpatialVx (Gilleland, 2013), ggplot2 (Wickham, 2009), RColorBrewer (Neuwirth, 2011), doMC (Revolution Analytics, 2011a), foreach (Revolution Analytics, 2011b), ff (Adler et al.,

2013), reshape2 (Wickham, 2007), sp (Pebesma and Bivand, 2005), Hmisc (Harrell et al., 2011), plyr (Wickham, 2011), maptools (Lewin-Koh et al., 2010), gpclib (Peng et al., 2012), timeDate (Wuertz et al., 2011) and xtable (Dahl, 2009).

References

Adler, D., Gläser, C., Nenadic, O., Oehlschlägel, J., and Zucchini, W. (2013). *ff: memory-efficient storage of large data on disk and fast access functions*. R package version 2.2-11.

Ahijevych, D., Gilleland, E., Brown, B. G., and Ebert, E. E. (2009). Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Weather and Forecasting*, 24:1485–1497.

Aspelien, T., Iversen, T., Bremnes, J. B., and Frogner, I.-L. (2011). Short-range probabilistic forecasts from the norwegian limited-area EPS: long-term validation and a polar low study. *Tellus A*, 63:564–584.

Ben Daoud, A., Sauquet, E., Lang, M., Bontron, G., and Obled, C. (2011a). Precipitation forecasting through an analog sorting technique: a comparative study. *Advances in Geosciences*, 29:103–107.

Ben Daoud, A., Sauquet, E., Lang, M., and Ramos, M.-H. (2011b). Can we extend flood forecasting leadtime by optimising precipitation forecasting based on analogs? application to the Seine river basin. *La Houille Blanche*, (1):37–43. (in French).

Braun, M., Caya, D., Frigon, A., and Slivitzky, M. (2012). Internal variability of the Canadian RCMs hydrological variables at the basin scale in Quebec and Labrador. *Journal of Hydrometeorology*, 13(2):443–462.

Briggs, W. M. and Levine, R. A. (1997). Wavelets and field forecast verification. *Monthly Weather Review*, 125:1329–1341.

Brown, T. A. (1974). Admissible scoring systems for continuous distributions. The Rand Paper Series, P-5235, The Rand Corporation, Santa Monica, California.

Casati, B., Ross, G., and Stephenson, D. B. (2004). A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, 11:141–154.

Chardon, J., Hingray, B., Favre, A.-C., Autin, P., Gailhard, J., Zin, I., and Obled, C. (2014). Spatial similarity and transferability of analog dates for precipitation downscaling over France. *Journal of Climate*, 27(13):5056–5074.

Dahl, D. B. (2009). *xtable: Export tables to LaTeX or HTML*. R package version 1.5-6.

Davis, C., Brown, B., and Bullock, R. (2006). Object-based verification of precipitation forecasts. part I: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, 134:1772–1784.

Demargne, J., Brown, J., Liu, Y., Seo, D.-J., Wu, L., Toth, Z., and Zhu, Y. (2010). Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmospheric Science Letters*, 11:114–122.

Denis, B., Coté, J., and Laprise, R. (2002). Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Monthly Weather Review*, 130:1812–1829.

Duffourg, F. and Ducrocq, V. (2011). Origin of the moisture feeding the heavy precipitating systems over southeastern France. *Natural Hazards and Earth System Sciences*, 11:1163–1178.

Ebert, E. and McBride, J. (2000). Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology*, 239(1-4):179–202.

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, 15:51–64.

Etchevers, P., Golaz, C., Habets, F., and Noilhan, J. (2002). Impact of a climate change on the Rhône river catchment hydrology. *Journal of Geophysical Research*, 107:D16.

Gilleland, E. (2013). *SpatialVx: Spatial Forecast Verification*. R package version 0.1-5.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24:1416–1430.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hagedorn, R., Hamill, T. M., and Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part I: Two-meter temperatures. *Monthly Weather Review*, 136:2608–2619.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C. (2011). The integrated nowcasting through comprehensive analysis (INCA) system and its validation over the eastern Alpine region. *Weather and Forecasting*, 26:166–183.

Harrell, F. E. J. et al. (2011). *Hmisc: Harrell Miscellaneous*. R package version 3.9-0.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15:559–570.

Hingray, B. and Saïd, M. (2014). Partitioning internal variability and model uncertainty components in a multimodel multireplicate ensemble of climate projections. *Journal of Climate*, 27(17):6779–6798.

Jeong, D. I., St-Hilaire, A., Ouarda, T. B. M. J., and Gachon, P. (2013). A multi-site statistical downscaling model for daily precipitation using global scale GCM precipitation outputs. *International Journal of Climatology*, 33(10):2431–2447.

Jolliffe, I. T., Stephenson, D. B., Broecker, J., Brown, B. G., Déqué, M., Ebert, E. E., Ferro, C. A. T., Gilleland, E., Hogan, R., Livezey, R. E., Mason, I. B., Mason, S. J., Pocernich, M., Potts, J. M., Richardson, D. S., and Weigel, A. P. (2012). *Forecast Verification: A Practioner's Guide in Atmospheric Science*. Wiley-Blackwell.

Keil, C. and Craig, G. C. (2007). A displacement-based error measure applied in a regional ensemble forecasting system. *Monthly Weather Review*, 135:3248–3259.

Keil, C. and Craig, G. C. (2009). A displacement and amplitude score employing an optical flow technique. *Weather and Forecasting*, 24:1297–1308.

Lafaysse, M., Hingray, B., Terray, L., Mezghani, A., and Gailhard, J. (2014). Internal variability and model uncertainty components in future hydrometeorological projections: The Alpine Durance basin. *Water Resources Research*.

Lewin-Koh, N. J., Bivand, R., contributions by Edzer J. Pebesma, Archer, E., Baddeley, A., Bibiko, H.-J., Dray, S., Forrest, D., Friendly, M., Giraudoux, P., Golicher, D., Rubio, V. G., Hausmann, P., Jagger, T., Luque, S. P., MacQueen, D., Niccolai, A., Short, T., and Stabler, B. (2010). *maptools: Tools for reading and handling spatial objects*. R package version 0.7-38.

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646.

Magand, C., Ducharne, A., Le Moine, N., and Gascoin, S. (2014). Introducing hysteresis in snow depletion curves to improve the water budget of a land surface model in an Alpine catchment. *Journal of Hydrometeorology*, 15(2):631–649.

Maraun, D. (2012). Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophysical Research Letters*, 39:L06706.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I. (2010). Precipitation downscaling under climate change. recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48:RG3003.

Maraun, D., Widmann, M., Benestad, R. E., Kotlarski, S., Hertig, E., Wibig, J., Gutierrez, J. M., Huth, R., Chandler, R. E., and Wilcke, R. A. I. (2014). VALUE - a framework to validate downscaling approaches for climate change studies. *Earth's Future*.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1095.

Murphy, Allan, H. (1969). On the "Ranked Probability Score". *Journal of Applied Meteorology*, 8:988–989.

Neuwirth, E. (2011). *RColorBrewer: ColorBrewer palettes*.

Ottlé, C., Etchevers, P., Golaz, C., Habets, F., Noilhan, J., Martin, E., Ledoux, E., Leblois, E., Sauquet, E., Amraoui, N., Artinian, E., Champeaux, J. L., Guérin, C., Lacarrère, P., le Moigne, P., Saulnier, G. M., Thiéry, D., Vidal-Madjar, D., and Voirin, S. (2001). Hydro-meteorological modelling of the Rhône basin: General presentation and objectives. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, 26(5-6):443–453.

Pebesma, E. and Bivand, R. (2005). *Classes and methods for spatial data in R*, r news 5 (2) edition. sp.

Peng, R. D., Murdoch, D., Rowlingson, B., and Murta, A. (2012). *gpclib: General Polygon Clipping Library for R*. R package version 1.5-3.

Pierce, D. (2013). *ncdf4: Interface to Unidata netCDF (version 4 or earlier) format data files*. R package version 1.10.

Pincus, R. and Rew, R. (2011). *NetCDF Fortran 90 Interface Guide*. Unidata, University Corporation for Atmospheric Research, Boulder, Colorado 80307, netcdf 4.1.3 edition.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radanovics, S., Vidal, J.-P., Sauquet, E., Ben Daoud, A., and Bontron, G. (2013). Optimising predictor domains for spatially coherent precipitation downscaling. *Hydrology and Earth System Sciences*, 17(10):4189–4208.

Revolution Analytics (2011a). *doMC: Foreach parallel adaptor for the multicore package*. R package version 1.2.3.

Revolution Analytics (2011b). *foreach: Foreach looping construct for R*. R package version 1.3.2.

Roberts, N. and Lean, H. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136:78–97.

Termonia, P., Degrauwe, D., and Hamdi, R. (2011). Improving the temporal resolution problem by localized gridpoint nudging in regional weather and climate models. *Monthly Weather Review*, 139:1292–1304.

Theis, S. E., Hense, A., and Damrath, U. (2005). Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications*, 12:257–268.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. (2005). The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M. (2010). A 50-year high-resolution atmospheric reanalysis over France with the Safran system. *International Journal of Climatology*, 30(11):1627–1644.

Wernli, H., Hofmann, C., and Zimmer, M. (2009). Spatial forecast verification methods intercomparison project: Application of the SAL technique. *Weather and Forecasting*, 24:1472–1484.

Wernli, H., Paulat, M., Hagen, M., and Frei, C. (2008). SAL–a novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, 136(11):4470–4487.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Use R! Springer, New York.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Wuertz, D., Chalabi, Y., with contributions from Joe W. Byers, M. M., and others (2011). *timeDate: Rmetrics - Chronological and Calendarical Objects*. R package version 2131.00.

Zappa, M., Beven, K. J., Bruen, M., Cofiño, A. S., Kok, K., Martin, E., Nurmi, P., Orfila, B., Roulin, E., Schröter, K., Seed, A., Szturc, J., Vehviläinen, B., Germann, U., and Rossa, A. (2010). Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 working group 2. *Atmospheric Science Letters*, 11:83–91.

Zepeda-Arce, J., Foufoula-Georgiou, E., and Droegemeier, K. K. (2000). Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *Journal of Geophysical Research*, 105:10129–10146.

10.2 Different thresholds for observation and simulation

In the final version of the SS and the LS two different thresholds are used to define the probability fields for the observation and the simulation, but earlier in the development phase a common threshold has been used.

Daily probabilistic SAL scores have been calculated over 20 years (1 Aug. 1982 – 31 July 2002) for the *best* experiment using the observed catchment average precipitation of the day as a common threshold for defining the observed and simulated probability fields. The top left panel of Figure 10.1 shows the S component as a function of the observed catchment average precipitation. The high precipitation days are often missing, because no objects have been found in the simulation, specifically all events with more than 31mm day^{-1} observed catchment average precipitation are missing. At least 13 out of the 25 analogue precipitation values at a grid cell have to be above the threshold to have a simulated object but since these events are quite rare and not always at the same place, there are not enough such events in the archive to exceed the threshold. These misses thus don't mean that no precipitation is simulated. There might even be some quite high precipitation events in the sample.

To be able to evaluate the location and the structure of the simulated precipitation in such situations, two different thresholds are used. The top right panel of Figure 10.1 shows the S component with the spatial mean of the local median values of the simulation is used as threshold for the simulation while the threshold for the observation is the same as before. The high precipitation events are no missed events any more. The remaining missed events are low and very low precipitation events and have a value of -2 assigned here.

The bottom panel of Figure 10.1 shows the difference between the thresholds used for the observation and the simulation. Positive values mean that the threshold for the simulation is larger than the one for the observation. This occurs mainly for small observed precipitation. For high observed precipitation the threshold difference is systematically negative, which means that the threshold for the simulation is smaller than the threshold for the observation. The dotted line marks the threshold difference that would be obtained with a threshold of 0 for the simulation. The dots veer only slowly away from this line, which highlights that the median of the simulations strongly underestimates the high precipitation events even if there are some high values in the sample.

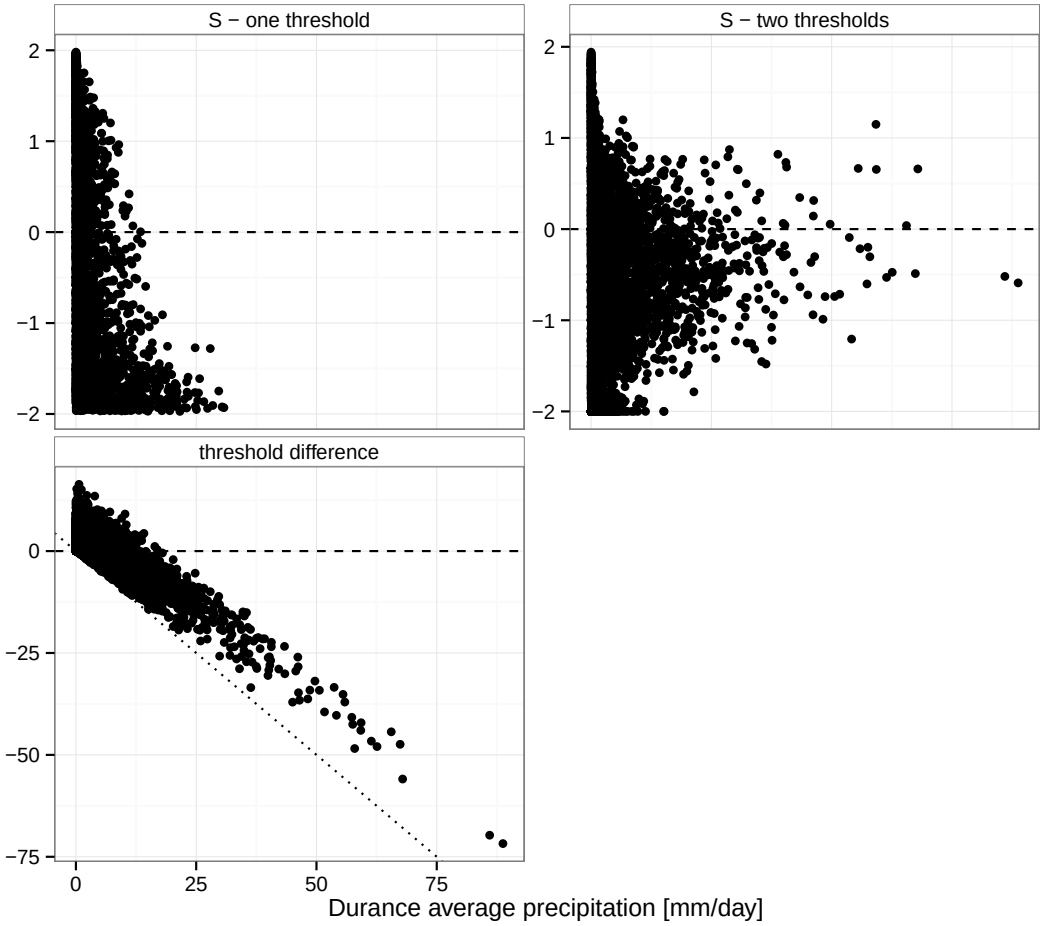

Figure 10.1: top left: S component of probabilistic SAL using the same threshold for observation and simulation - the catchment average observed precipitation, top right: the S component using two thresholds - the catchment average observed precipitation for the observation and the spatial mean of the local median of the simulated precipitation for the simulation, bottom: the difference between the two threshold - positive values mean that the observation threshold is higher. All variables are shown as a function of the catchment average precipitation of the day.

10.3 Some performance diagnostics of the experiments for the Durance

In this section some more traditional performance measures are presented for the grid cells in the Durance catchment and the six main experiments *catchment*, *best*, *optim*, *catchment 5*, *best 5* and *optim 5*.

In Figure 10.2 CRPS maps for the Durance catchment and the different experiments are shown for the *late* period. The overall spatial structure is very similar between the experiments with high CRPS values, that is large errors, at the northwestern- and southeastern edges of the catchment and the smallest errors in the southwestern part near the outlet of the catchment. This pattern corresponds to the average precipitation regime as has been discussed in Section 6.3.1.

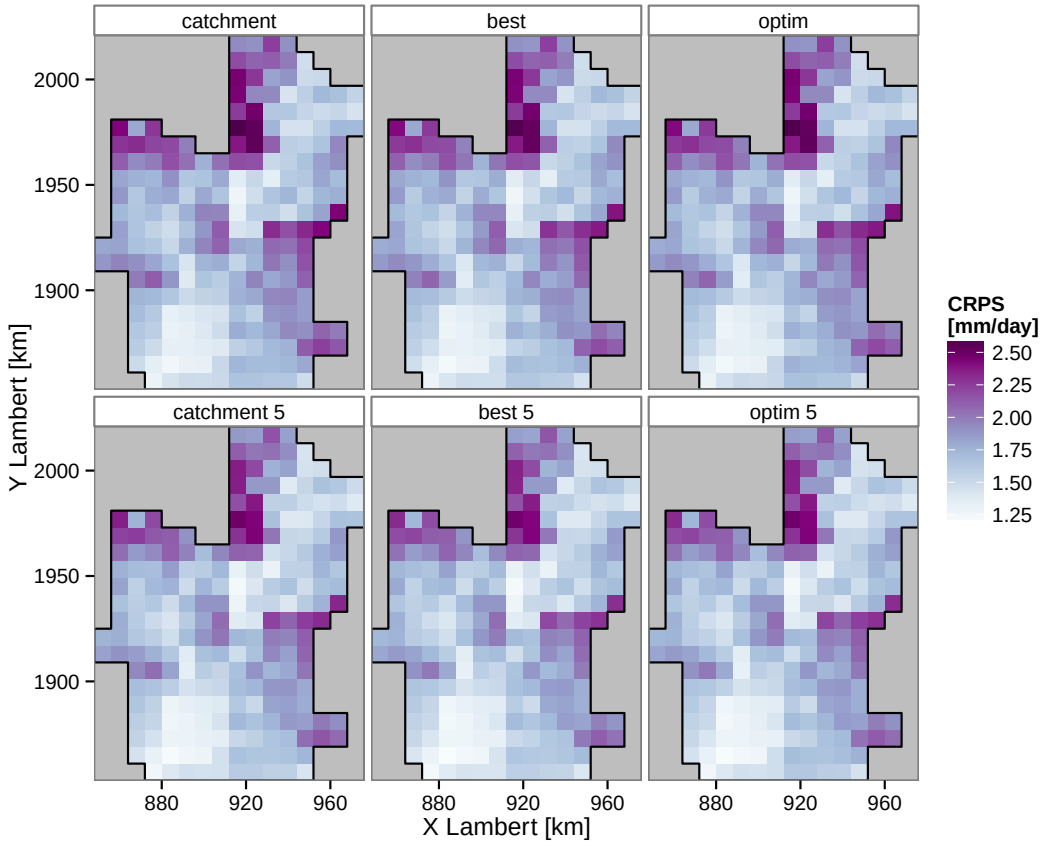

Figure 10.2: CRPS for the grid cells in the Durance catchment at Cadarache for the six main experiments.

To highlight the difference between the experiments the CRPS of the *catchment* experiment has been subtracted in Figure 10.3. The *best* experiment has slightly smaller CRPS values than the *catchment* experiment, especially in the southern part of the catchment and in areas with large CRPS. The *optim* experiment is very similar to the *best* experiment in the northern part of the catchment, but has higher CRPS values than the *catchment* experiment along the southern edge of the catchment and smaller differences in the southeastern part. The higher CRPS values of the *optim* experiment along the southern edge is due to the predictor domains used for the local predictors (temperature, vertical velocity and humidity), that are not ideal for these zones as has been discussed in chapter 7. They belong to the zones with large differences in CRPSS between the *optim* and the *best* domains as shown in Figure 7.1. The multidomain experiments (*catchment 5*, *best 5* and *optim 5*) have smaller CRPS values than their single domain counterparts. The differences are larger where the CRPS values are higher. Again, the *best 5* experiment has smaller CRPS in the southern part of the catchment whereas there is less difference for the *optim 5* experiment along the southern edge.

The bias of the median calculated over the *late* period is negative for all grid cells and all experiments as shown in Figure 10.4. The biases are more negative where the average precipitation is higher and for the multi domain experiments.

The picture is quite different for the bias of the mean shown in Figure 10.5. The bias of the mean is much smaller than the bias of the median. It is positive in the northern and eastern part of the catchment and slightly negative in the southwestern part, near the outlet of the catchment. The strongest positive biases are found along the northwestern edge of the catchment. The *best* experiment has smaller biases in the northeastern part than the other experiments. The *optim 5* experiment has the least negative biases.

Figure 10.6 shows rank histograms for the six main experiments accumulated for all grid cells in the Durance catchment and over the *late* period. Ideally the rank histograms should be flat which indicates reliable and unbiased simulations. The shape of the rank histograms obtained with the downscaling experiments using SANDHY is nearly the same for all experiments. It is a convex shape, which means that the SANDHY ensemble tends to be overdispersive, and it is asymmetric, with a maximum due to overpopulation of moderately high ranks and underpopulation of very high ranks. At low ranks the shape of the rank histogram depends on how zero precipitation is treated when the ranks are determined. Here zero precipitation is ranked using random perturbations as explained in Section 3.2.

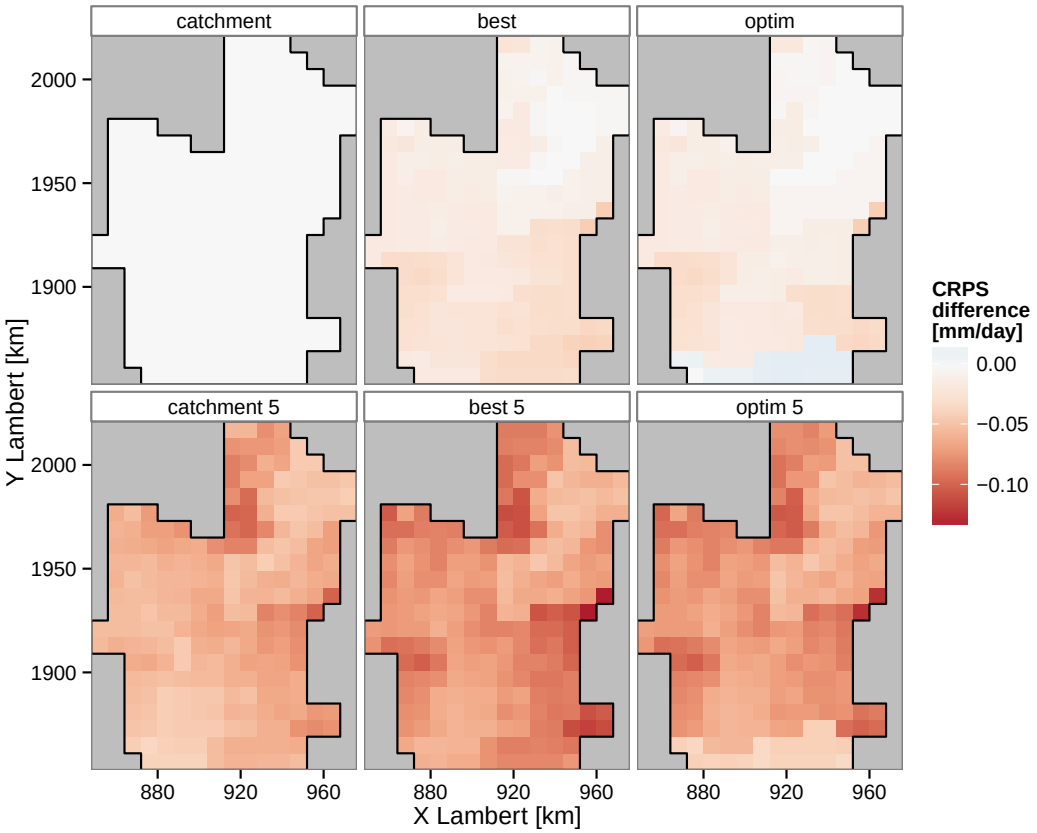

Figure 10.3: CRPS difference with respect to the *catchment* experiment.

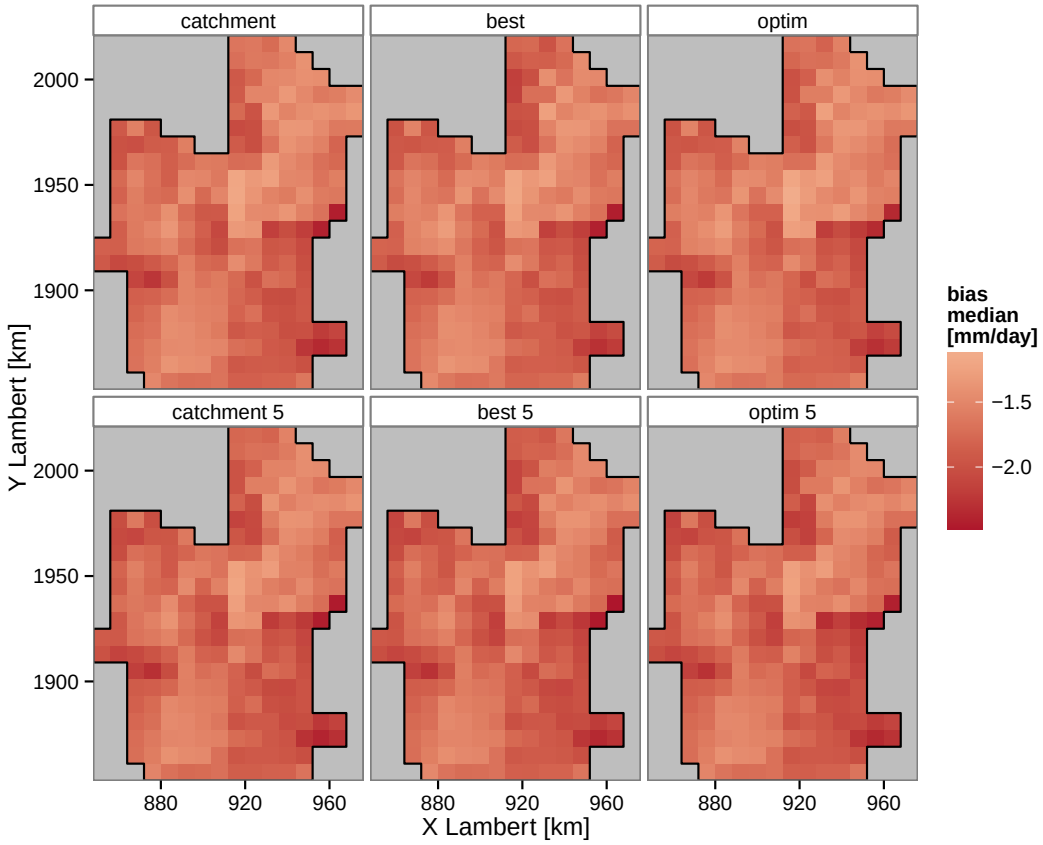

Figure 10.4: Bias of the median at the grid cells in the Durance catchment.

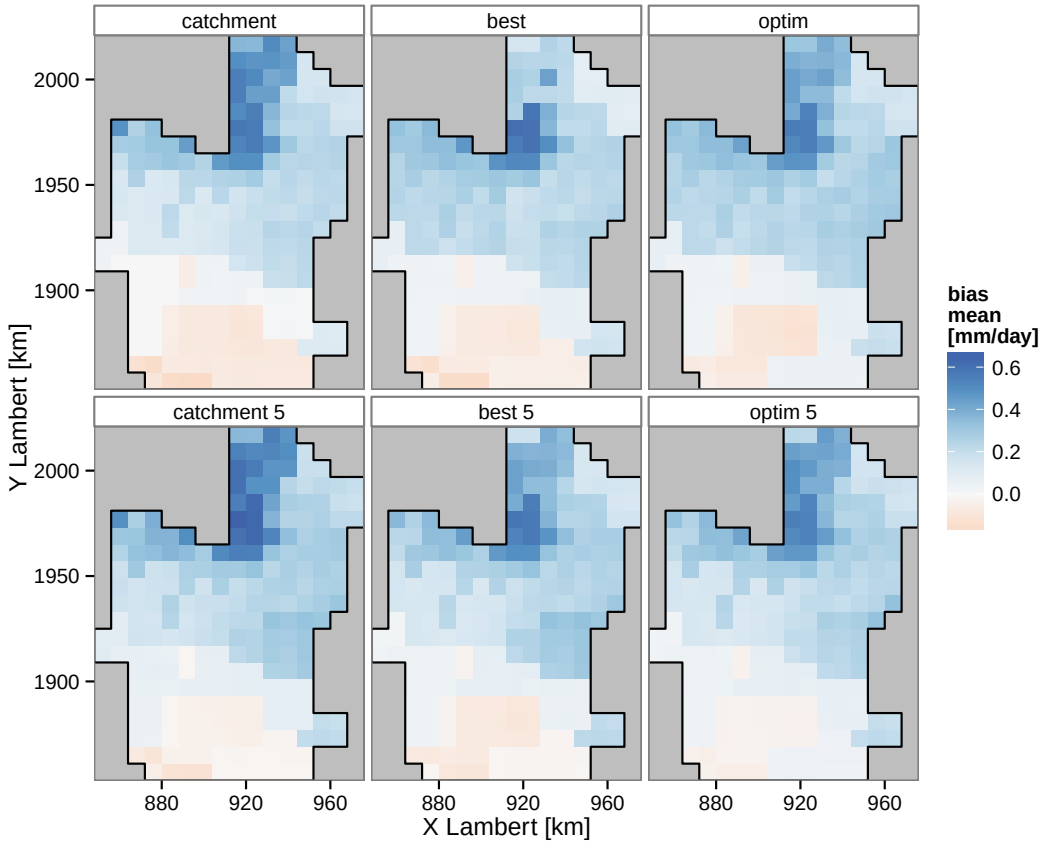

Figure 10.5: Bias of the mean at the grid cells in the Durance catchment.

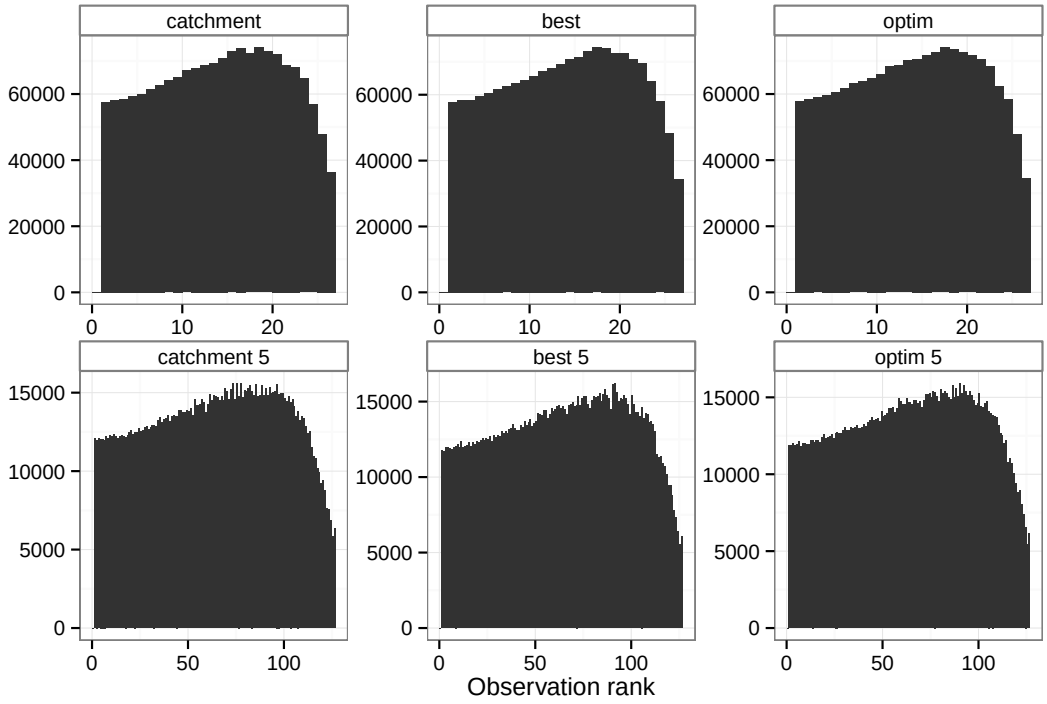

Figure 10.6: Rank histograms for all grid cells in the Durance catchment and the six main experiments.

10.4 A closer look at the improvement of the CRPSS using five domains

In Section 10.1 it was found that the local CRPSS are in general better for the experiments using the analogue days from five domains rather than only one for the studied catchments Rhône and Durance. But does this hold for the whole of France? Figure 10.7 shows the CRPSS for the 608 zones over the *late* period using only the first of the optimised domains on the x-axis (*optim* experiment) and using analogue precipitation from 5 domains on the y-axis (*optim 5* experiment). The CRPSS values using 5 domains are about 0.02 higher than using only one domain for all zones, so it can be concluded that the improvement of the CRPSS is a general feature.

In Figure 10.8 it can be seen that the differences in CRPS are larger for the zones that have large CRPS and that the CRPS for the *optim 5* configuration are systematically smaller than for the *optim* configuration. Note that better simulations have larger CRPSS but smaller CRPS. Talagrand et al. (1997) and

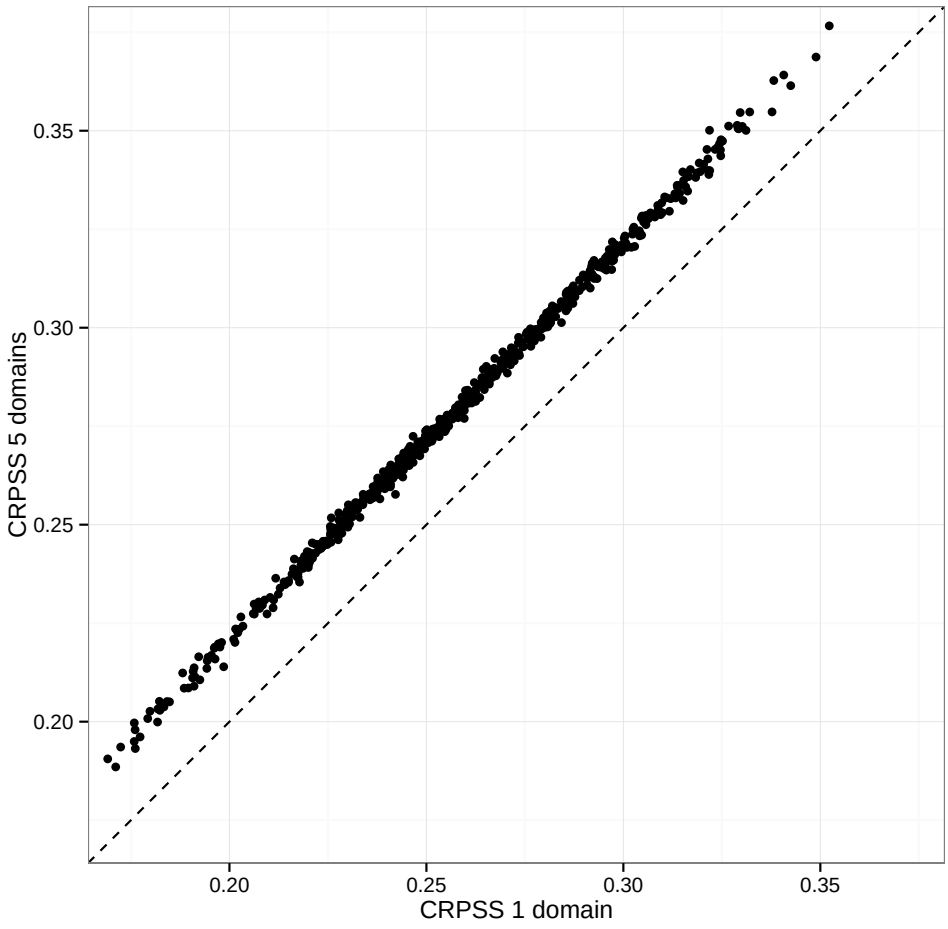

Figure 10.7: CRPSS using 1 optimised domain vs. using 5 optimised domains for 608 zones over the 1 Aug. 1982 - 31 July 2002 period.

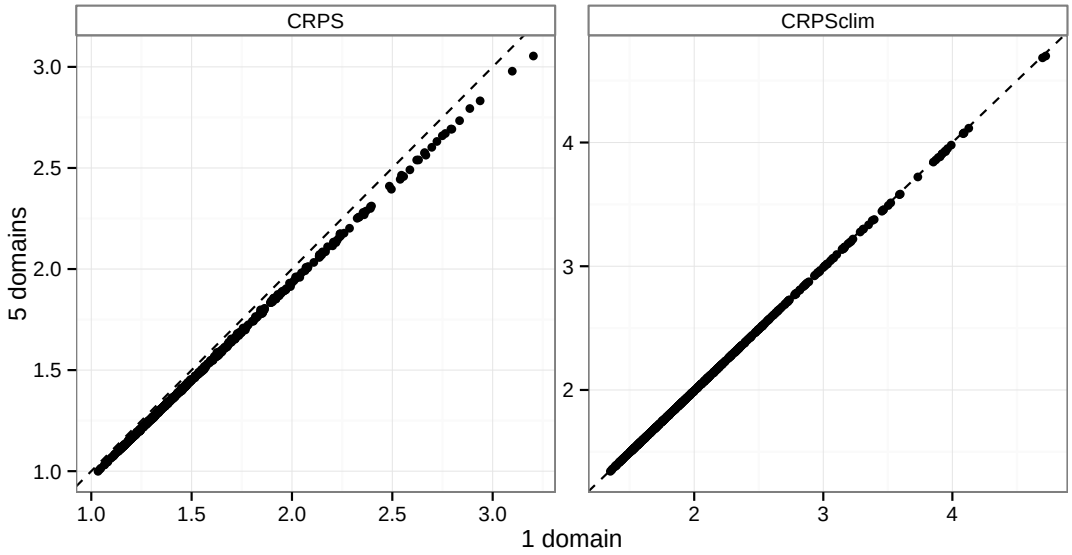

Figure 10.8: As Figure 10.7 but for the $CRPS_{clim}$ on the right hand side and the CRPS on the left hand side.

Richardson (2001) noted for the Brier score that it depends on the ensemble size due to sampling errors in small ensembles. This is likely to be the case for the CRPS as well, which is why the $CRPS_{clim}$ is calculated here using different sampling as well. The differences in the $CRPS_{clim}$ calculated from 25 and 125 values are shown in the right panel of Figure 10.8. It can be seen that these differences are very small compared to the ones in CRPS in the left panel of Figure 10.8. This means that the better sampling of the distribution when 125 values are used instead of 25 can not fully explain the smaller CRPS of the *optim 5* configuration.

Bontron (2004) introduced a decomposition of the CRPS in "finesse" and "justesse". "Finesse" corresponds to the CRPS obtained if the median of the predicted distribution was observed and thus is smaller for narrower distributions. "Justesse" refers to the part of the CRPS that is due to the difference between the median of the predicted distribution and the observation and thus corresponds to an absolute error of the median. Figure 10.9 shows the decomposition of the CRPS for the 608 zones in "finesse" on the left hand side and "justesse" on the right hand side. The two components have the same order of magnitude, the "finesse" being slightly smaller. Both components are slightly smaller in the *optim 5* configuration than in the *optim* configuration for most of the zones which means that the change in CRPS can be attributed to improvements of both characteristics.

A second decomposition is the one proposed by Hersbach (2000) in reliability and potential CRPS. The reliability tests if the frequency that the observation is

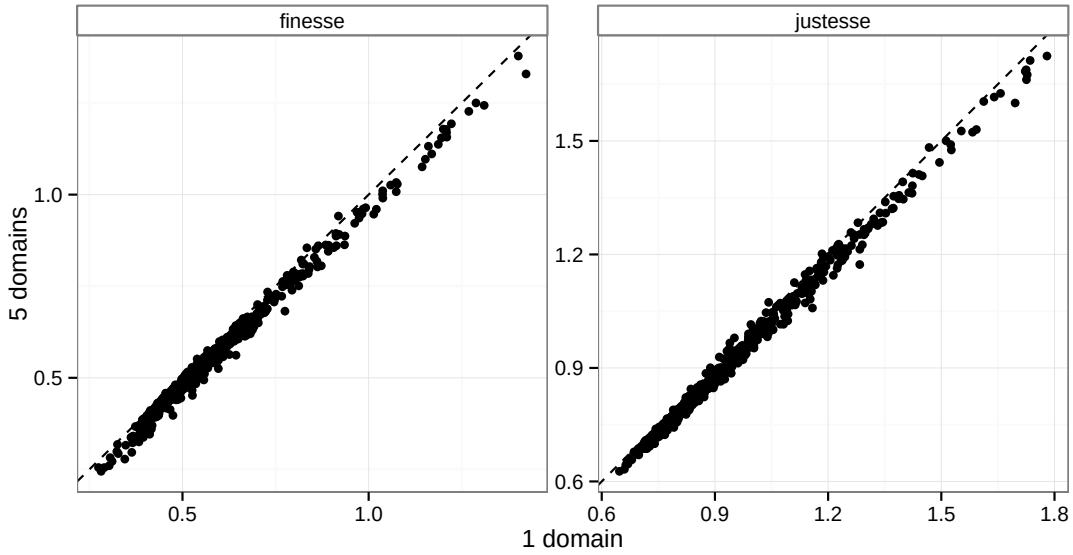

Figure 10.9: As figure 10.7 but for the "finesse" on the left hand side and the "justesse" on the right hand side.

found in a certain bin is proportional to the bin width devided by the number of bins. Thus it is closely related to the rank histogram with the difference that observations falling into large bins get higher weights in the reliability calculation while for the rank histogram it is just counted how often the observation falls into a given bin regardless its width. The potential CRPS is the CRPS that would be obtained for a forecast that is tuned to be perfectly reliable.

Figure 10.10 shows the decomposition of the CRPS following Hersbach (2000) in reliability (left hand side) and potential CRPS (right hand side) for 608 zones. In general the reliability component is rather small for all zones which means that the distributions given by the analogue method are quite reliable. For zones with a very low reliability component simulations with the *optim 5* experiment are more reliable than simulations from the *optim* experiment. For zones with slightly higher reliability components the reliability component tends to be higher for the *optim 5* than for the *optim* experiment which means that using more domains makes the estimate less reliable. The potential CRPS is much larger than the reliability component and is consistently smaller for the *optim 5* experiment than for the *optim* experiment.

Hersbach (2000) decomposed the potential CRPS further into uncertainty and resolution, uncertainty corresponding to what we call $CRPS_{clim}$ here. The resolution is related to the difference in steepness between the predicted cdf and the cdf of climatology. Given the small contribution from the reliability component and

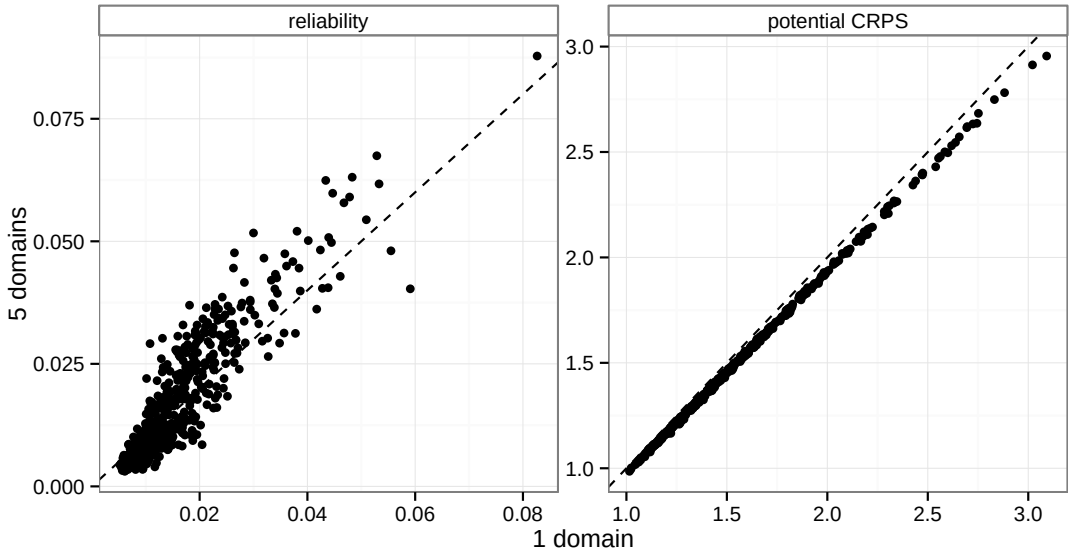

Figure 10.10: As figure 10.7 but for the reliability on the left hand side and potential CRPS on the right hand side.

the small decrease in $CRPS_{clim}$ compared to the total CRPS it can be concluded that the observed improvement essentially stems from increased resolution.

10.5 Conclusions

A probabilistic version of the spatial verification method SAL has been developed and applied on downscaled precipitation fields over the Rhône and the Durance catchment. Using a common threshold for defining the probability fields of the probabilistic SAL for the observation and the simulation based on observed precipitation amounts does not allow to evaluate the structure and the location characteristics for high precipitation events. As a consequence two different thresholds are used.

Probabilistic structure and location skill scores are used to compare downscaled precipitation from a set of experiments representing different predictor domain configurations. The selected experiments represent contrasting strategies: uniform configuration for the whole catchment or locally optimised predictor domains. The uniform predictor domain configurations lead to smoother simulated precipitation fields and more realistic structure, but less precise location than the fields glued together from spatially varying predictor domains. Many more experiments would have been possible, for example using the groups or clusters obtained in chapter

8, and using a set of predictor domains for each group.

The multi domain experiments lead to higher CRPSS compared to single domain experiments, which is not restricted to the studied catchments but true for all 608 zones in France. The higher CRPSS are due to higher resolution of the predictive distributions of the multi domain experiments.

The structure skill score and the location skill score show reversed results for the studied experiments and catchments. In the following chapter it will be explored which experiments lead to the most skillful streamflow simulations in the Durance catchment and if the scores for precipitation are related to the performance of the streamflow simulations.

11 Hydrological impacts of spatially coherent downscaling

In the previous chapter measures for the spatial coherence of probabilistic precipitation fields were developed. SANDHY was developed to be used for downscaling of atmospheric fields as input for hydrological models. Therefore the sensitivity of streamflow simulations with a distributed hydrological model on choices concerning the predictor domain configuration for simulated precipitation is studied here. The predictor domain configurations from the experimental setup described in the article in Section 10.1 are used to this end. The question is which configuration produces the most skillful streamflow simulations and which of the proposed skill scores for precipitation (CRPSS, SSS and LSS) might allow to predict this. The streamflow simulations and related performance calculations were conducted during an internship (Eeckman, 2014) supervised by Jean-Philippe Vidal, François Tilmant and myself. In the following the hydrological model used is introduced in Section 11.1, the treatment of the data is described in Section 11.2 and the performance of streamflow simulations is discussed in Section 11.3.

11.1 Hydrological model - J2000

For simulating streamflow from given atmospheric fields, a hydrological model is needed. The physically based distributed model J2000 has been developed at the University of Jena (Krause, 2002). The hydrological processes, notably snow accumulation and snow-melt, interception, surface runoff, infiltration, subsurface flow and various storages, runoff concentration and routing are explicitly simulated. J2000 is a modular system where the user can choose which modules, for example for snow-melt, he or she wants to use depending on the focus of the study and new modules can be added fairly easily. Labbas (2014) for example implemented a module for peri-urban catchments to simulate the behaviour of sewer overflow devices. J2000 has been used for streamflow simulations in various catchments, for

example the Mulde (Krause, 2002) and the Gera catchment (Krause et al., 2006) in Germany, the Dudh Kosi (Nepal et al., 2014) in Nepal, the Yzeron (Branger et al., 2013), the Chézine (Gudefin, 2013) and the Ardèche (Huza, 2013) in France. The version for the Durance catchment has been built during the Project "Risque, Ressource en eau et gestion durable de la Durance en 2050" (Sauquet et al., 2014) (R^2D^2-2050) and is now on the way to be extended to the Rhône catchment in the project "Modélisation Distribuée du Rhône" (MDR) (Tilmant et al., 2013).

Related to the present study, J2000 has been used to simulate streamflow for the Durance catchment using atmospheric forcing from the experiments described in Section 10.1 during an internship (Eeckman, 2014). J2000 operates on hydrological response units (HRU) (Flügel, 1995) that are defined to be coherent areas in terms of geology, slope, soil type and land use. The atmospheric forcing at each of the 1963 HRUs in the Durance catchment is taken from the nearest Safran grid cell.

11.2 Data preparation

So far, only precipitation has been considered in the downscaling exercise, but the hydrological model additionally needs minimum temperature, maximum temperature and potential evapotranspiration as inputs. Preferably all variables should belong to the same weather situation, which can easily be achieved with the analogue method by taking all variables from the same analogue date. Minimum temperature, maximum temperature and potential evapotranspiration are sampled from the DuO data set (Magand et al., 2014), which is a hybrid data product of the Safran (Vidal et al., 2010) and SPAZM (Gottardi et al., 2012) datasets for the Durance. The precipitation is taken from Safran for the sake of consistency with the downscaling experiments. The predictor domains used in the downscaling experiments were for example optimised using Safran data (Section 5.1) and the performance of the precipitation has been evaluated using Safran data as well. The solid precipitation from Safran is multiplied by 1.5 because snow amounts are underestimated in the Safran data due to a sparse station network at high altitudes and the undercatch of solid precipitation by rain gauges (Magand et al., 2014). Sensitivity tests from Eeckman (2014) showed that the best streamflow simulations are obtained using a factor close to 1.5.

The hydrological model can not directly deal with input that is in the form of an empirical predictive distribution at each grid point as delivered by the SANDHY method. Therefore an ensemble of deterministic scenarios has to be created for each experiment.

11.2.1 Random scenarios

For the experiments using one predictor domain per zone (*optim*, *best* and *catchment*, compare Section 10.1), random permutations of the 25 analogue dates are created for each zone and each day. The random permutations are necessary to create equiprobable scenarios since the 25 analogue dates are ordered from the closest to the least close one in terms of analogy. The same permutation is used for zones using the same predictor domain such that they use the same analogue dates, and different permutations are used where zones use different predictor domains.

For the experiments using five predictor domains (*catchment 5*, *optim 5* and *best 5*), the random permutations are done for the 125 analogue dates. For the *catchment 5* experiment the same permutations are used for all zones in the catchment, while for the *optim 5* and *best 5* experiments different permutations are used for different zones.

Later tests in the context of other works showed that in these random scenarios the autocorrelation of precipitation is slightly underestimated and the autocorrelation of temperature is severely underestimated (L. Caillouet, personal communication).

11.2.2 Reshuffled domains

In the case of the five predictor domains, it can happen that zones share some of the predictor domains, but not necessarily in the same order. Here the question is, how can we get them to use the same analogue date wherever possible since this is supposed to improve the spatial coherence. A recursive algorithm (J.-P. Vidal and L. Caillouet, personal communication) is used to reshuffle the order of the five domains for each zone, such that the number of different domains at each of the five positions (1 to 5) is minimised. Then the procedure is as for one domain per zone, but repeated five times (one for each position). This leads to a second set of realisations for the *best 5*, *optim 5* and *catchment 5* experiments named *shuffled* in the following.

11.3 Performance of streamflow simulations

As discussed above, several realisations of simulated atmospheric variables are available for each experiment. This allows to run ensemble simulations of streamflow. The ensemble can be evaluated as a whole using probabilistic performance measures, here the probabilistic version of the Nash-Sutcliffe criteria (Nash and Sutcliffe, 1970; Bulygina et al., 2009) is used, or using non-probabilistic measures applied to each ensemble member which gives a distribution of scores. Here the

corrected Kling-Gupta Efficiency (KGE) (Kling et al., 2012) and its three components are shown. The components of the KGE, β, γ and r can be used separately to diagnose different model behaviours which is why it has been chosen here. Formulas and details on the components are given in Section 3.4.

It is intended to evaluate which of the simulated atmospheric datasets leads to simulated streamflow closest to the one simulated using observed atmospheric data and how the streamflow simulation reacts on the different ways the atmospheric forcing is simulated. Therefore observed streamflow is not used as a reference for evaluating the streamflow simulation performance. A streamflow simulation using the reanalysis time series of the simulated period is used instead, similar to Lafaysse et al. (2014a).

Twenty-two years of streamflow at Cadarache from 1 Aug. 1980 to 31 July 2002 have been simulated. Cadarache is the station on the Durance river which is described in Section 2.2.2 of the article in Section 10.1. The performance of the simulations is evaluated over the 1 Aug. 1982 - 31 July 2002 period to allow for model spin-up. This time period corresponds to the optimisation period of the downscaling experiments. The simulations are performed with scenarii of simulated precipitation from six experiments. An example of such simulations is shown in Figure 11.1 for the *catchment*, *best* and *optim* experiment and the year 1988. The simulations using the simulated input data from the *catchment* experiment labelled Durance here show a larger ensemble spread than the simulations for the *best* and *optim* experiments.

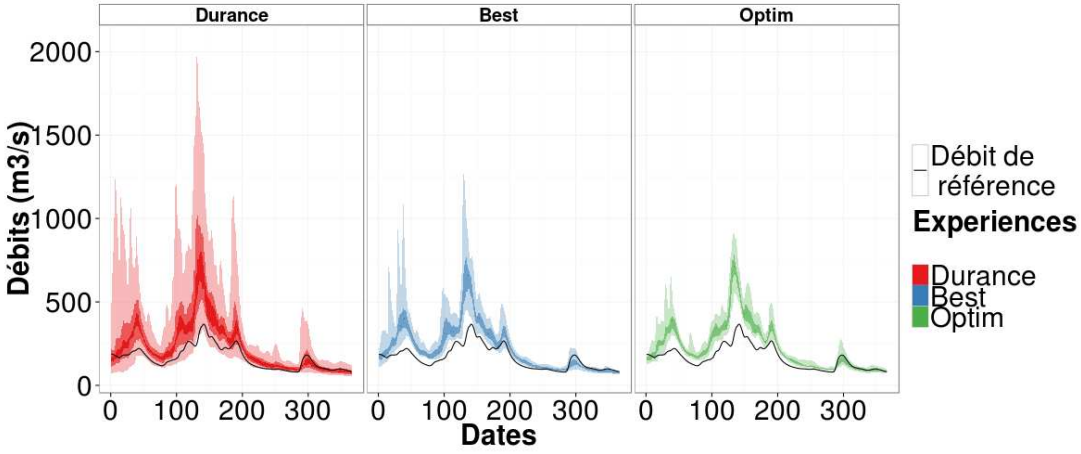

Figure 11.1: Ensemble of streamflow simulations with J2000 for the Durance catchment at Cadarache for the year 1988 and three experiments. Black: reference simulation. (Figure from Eeckman (2014))

Figure 11.2 shows the skill scores for precipitation (CRPSS, SSS and LSS)

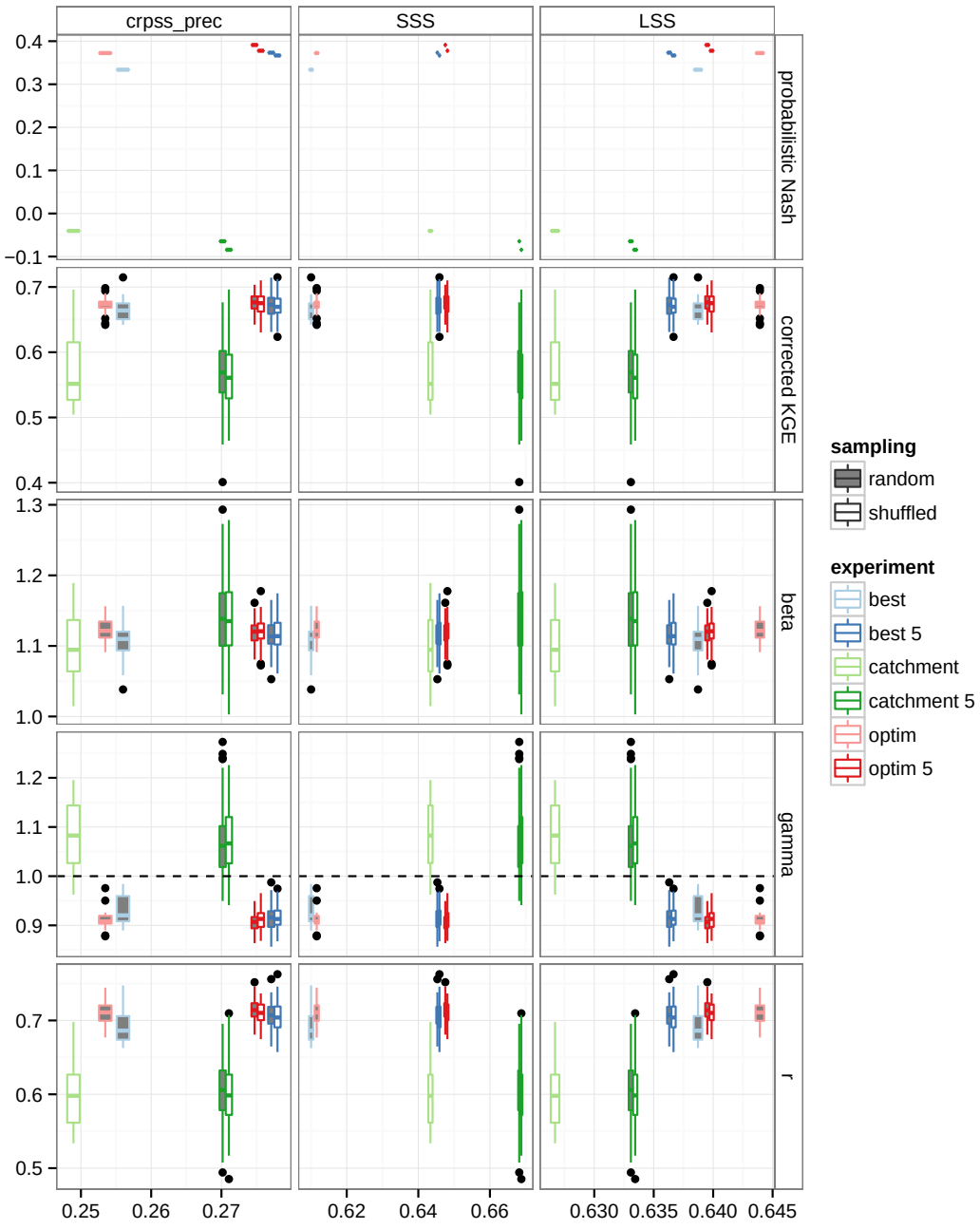

Figure 11.2: Performance scores for simulated precipitation and streamflow for different experiments.

versus the performance scores for streamflow at Cadarache (probabilistic Nash, corrected KGE and its components) for six downscaling experiments and two sampling methods for the multi domain experiments.

The probabilistic Nash and the KGE show qualitatively similar results with higher, that is better, scores for the *optim 5*, *optim*, *best 5* and *best* experiments and lower scores for the *catchment* and *catchment 5* experiments. The mean values, represented by the β component of the KGE, are slightly overestimated for all experiments. The variability ratio, γ, is larger than 1 for the *catchment* and *catchment 5* experiments and smaller than 1 for the others. This means that the temporal variability is overestimated for the *catchment* and *catchment 5* experiments and underestimated for the other experiments. The correlation coefficient is highest for the *optim*, *optim 5* and *best 5* experiments and lowest for the *catchment* and *catchment 5* experiments. The correlation coefficient is the component of the KGE that contributes the most to the total score, because it is farther to one than the other components. In addition almost all the difference between the experiments can be attributed to the correlation component, because for β there is not much difference between the experiments and for γ there are differences, but the absolute difference from 1 which is considered in the calculation (compare equation 3.13) is very similar between the experiments as well.

The differences between the two sampling methods (*random* and *shuffled*) are negligible and where they occur they tend to be in favor of the random sampling. As expected, the variance of the scores for different ensemble members is somewhat larger for the ensembles containing more members, that is the ones from the multi domain experiments. Interestingly the variance of the scores is clearly larger for the *catchment* and *catchment 5* experiments than for the other experiments. A possible reason is that the fields that are composed from different analogue dates are more similar to each other in terms of total precipitation over the catchment than the fields from the same analogue date for the whole catchment. This may explain the higher temporal variance as well.

The multi domain experiments lead to very similar runoff scores to those of their single domain counterparts despite the higher CRPSS of the precipitation simulation. In terms of the Nash criteria *optim 5* and *best 5* are somewhat better, and *catchment 5* somewhat worse than the respective single domain versions. For the γ component of the KGE it's the other way round, because the multi domain experiments have lower variability ratios and this brings the *catchment 5* closer to the ideal value of one, but moves the others further away from it.

Out of the three precipitation scores, the Location Skill Score (LSS) is the one that has the strongest tendency to vary in the same direction as the performance scores of the streamflow simulations Nash and KGE.

11.4 Conclusions

The best streamflow simulations in terms of the Nash and KGE citeria are obtained with the *optim* and *optim 5* experiments. Contrary to precipitation the *best* experiment does not outperform the *optim* experiment in the streamflow simulation and the multi domain experiments don't lead to systematically better streamflow simulations. There is definitely an interest in optimising predictor domains locally, because this leads to better streamflow simulations, notably in terms of temporal correlation (see the correlation component of the KGE).

Experiments having a high LSS for precipitation tend to lead to better streamflow simulations. The localisation of precipitation in the catchment is the most sensitive of the studied factors. The LSS seems to be the most meaningful of the three precipitation indices for streamflow simulations with J2000 despite its small variations. Note that this might be different for different hydrological models and different catchments.

Conclusions and Perspectives

Conclusions

The SANDHY method has been extended to the whole mainland of France and Corsica. The extention has been accomplished by optimising the predictor domains for the geopotential predictor individually for 608 climatologically homogeneous zones covering France (Chapter 5). An important spatial variability of the resulting predictor domains has been found. The predictor domain for the other three predictor variables has not been optimised which turned out to be a weakness of the chosen optimisation strategy. However, it can partially be overcome by considering predictor domains found for other locations because this can change the large scale grid cell used as predictor domain for the local predictor variables (Chapter 7).

The downscaling skill for precipitation depends on the precipitation climate and is higher for wetter climates than for drier ones which is in line with Fowler et al. (2007). The skill in terms of CRPSS is comparable to the one found in similar studies but given the poor skill of the streamflow simulations driven with downscaled data in chapter 11, the skill of raw output from the current method is probably not good enough for hydrological modelling. This is likely to be due on one hand to the downscaling method, that is optimised for precipitation but has seasonal biases for precipitation and temperature and on the other hand the chosen strategy for constructing scenarios that leads to less than observed temporal autocorrelation especially for temperature.

The validation experiments showed non-uniform skill loss in terms of CRPSS when a time period different from the optimisation period is simulated and a non-uniform change in bias if the archive period is different from the simulation period (Chapter 6). Skill loss in the out of sample validation and changes in the bias are particularly severe for a zone where the predictor-predictand relationship is not stationary over the two studied time periods due to an added station at higher altitude observing systematically higher precipitation amounts in the second period. Using transformed precipitation as predictand during the optimisation does not substantially change the optimisation and validation results.

There are two main philosophies regarding spatially coherent precipitation downscaling with analogues. One is to select the same analogue dates for a region as large as possible and the other one is to use analogue dates from locally optimised predictor domains accepting that they vary between locations but under the hypothesis that neighbour locations will have sufficiently similar parameters to ensure smooth transitions at least in a probabilistic setting. In the first case one can set a threshold in terms of performance loss between the locally optimised model and the common model to define the limits of the region that is supposed to use a common model configuration. Chardon et al. (2014) found such a region of transferability to be larger for an analogue method with only one selection step than for a two step method and concludes that the refined two step method increases the local skill but decreases the transferability. Such an analysis has not been performed with the four step SANDHY method but it is likely that the transferability is even lower given the higher number of steps and the limited aggregateability of predictand areas by equal predictor domain found in Chapter 8. The first approach seems therefore to be more suited for a single step analogue selection than for a multi step procedure. In terms of optimisation effort it has to be noted that following the first approach does not avoid local optimisation since the maximum local performance has to be known in order to calculate a performance loss.

The requirement of sufficiently similar parameters for neighbour locations to ensure smooth transitions in the case of the second approach is more easily met if small spatial entities are used in the local optimisation. For additional smoothing precipitation from analogue dates found for neighbour locations could be integrated in the local probabilistic precipitation estimate. The amount of information to integrate from specific neighbours should be spatially unisotropic and reflecting the similarity of parameters between two locations that depends on the similarity of the large scale ingredients that lead to precipitation at a given location. Indeed combining analogue precipitation found with multiple predictor domains was found to increase the local CRPSS of downscaled precipitation. The improvement is essentially due to a better resolution of the predictive distribution (Chapter 10).

If the target region of the downscaling is larger than the region of transferability, the first approach requires the definition of a border between two such regions. At these borders important discontinuities in the downscaled fields are likely to occur especially for large regions and high performance loss thresholds. The preferred location of the borders may vary with the application. In addition the borders may have to be redefined if the downscaling target region is enlarged. All this is avoided using the second approach which makes it independent of the application and more straight forward to extend spatially.

A spatial verification method has been adapted for probabilistic simulations

and was used to assess the spatial properties of the downscaled precipitation fields over two catchments of different sizes (Chapter 10). An experimental setup was designed that compares the two downscaling philosophies in terms of local skill, realistic spatial structure and realistic location of the downscaled precipitation. The spatially homogeneous predictor domains lead to more realistic spatial structure while the spatially varying ones better capture the location.

Simulations with a distributed hydrological model for the Durance catchment conducted during an internship (Eeckman, 2014) showed that there are indeed differences in the skill of the streamflow simulations due to the two downscaling philosophies, while combining analogue precipitation found with multiple predictor domains has a comparatively small influence (Chapter 11). Spatially varying predictor domains lead to higher skill of the streamflow simulations measured with the probabilistic Nash and the KGE criteria, notably the correlation component of the KGE is higher. The spatially varying predictor domains apparently provide enough spatial consistency of downscaled preciptation and are more precise in locating precipitation than uniform predictor domains. This leads to smaller errors in the streamflow simulations and shows that there is some interest in the local optimisation approach when an application in hydrology is intended.

Multiple variables are needed for streamflow simulations. Technically it is easy to achieve this inter-variable coherence using resampling methods by assuming realistic inter-variable relationships when all variables are taken from the same analogue dates. In practice it is far from sure that a method optimised for precipitation correctly simulates other variables. Indeed, Caillouet (2016) found that the seasonal cycle of simulated temperature is too flat and that the temporal autocorrelation of temperature is strongly underestimated in SANDHY simulations.

Some GCMs have a good representation of the climate system's response to greenhouse gas forcing, but location biases of the storm tracks. These biases in the predictor data with respect to the archive will lead to biases in the downscaled local variables as well, if the downscaling is applied to GCM data, since the predictors are assumed to be well simulated in perfect prognosis type downscaling like SANDHY. In this case the downscaling reveals non-trivial GCM biases and does not correct them. This may be a valuable diagnostic for the GCMs, but is not what most users would expect.

The SANDHY software developed during this thesis and described in Chapter 4 has already been successfully used by colleagues. Laurie Caillouet uses SANDHY to reconstuct historical droughts over France which implies the use of long reanalysis datasets as the twentieth Century Reanalysis 20CR (Compo et al., 2011) as predictors (Caillouet, 2016). Jean-Philippe Vidal downscaled precipitation for 81 Stations in Argentina using SANDHY for the second CORDEX-ESD workshop (Vidal and Radanovics, 2014), which implied to use ERA-Interim predictors and

predictands from station data instead of zone averages. This helped to identify and remove bugs that became visible due to the different set-ups in terms of data and time periods. An unresolved issue using station data as predictand is the treatment of missing data in the archive. The configuration file is rather long and somewhat complicated due to the different tasks and options that can be selected which limits the userfriendliness. On the other hand this makes it flexible enough to cope with different datasets for predictors and predictands, different regions of application and different simulation periods.

Perspectives

As always in research every result makes us ask at least one other question and as always at the end of a thesis or research project in general there are some questions that have not been answered yet.

Open questions

The validation of the SANDHY method has been carried out using the locally optimised domains but what would be the validation results for the *catchment* domains for the zones inside the respective catchment? Do they have more skill loss or less than the locally optimised ones? Does the skill loss vary considerably between the zones? A similar question can be asked about the *best* domains. They have better skill during the optimisation period, but how do they behave in an out of sample validation?

The fact that using the analogue dates from 5 domains instead of only one increases the skill, leads to the following questions that would be worth exploring. Until how many domains does the CRPSS increase? Simply retaining more analogue dates does not increase the skill because this adds information from days that are less similar. The use of analogue dates from different predictor domains is rather similar to a weighting as it is often used in k-nn methods. Therefore it would be interesting to look if a weighting with the rank of the analogue leads to the same improvement of the skill. Furthermore throughout the study CRPSS differences have been observed but which of these differences are actually significant? The significance of the differences in CRPSS could be tested using the paired tests proposed by Hamill (1999) but they can only be applied in cases where the same time series is simulated, that is the same zone and the same time period. This is the case for the question if the CRPSS increase using multiple predictor domains is significant but not in the validation experiments and for differences between zones.

In the development of the probabilistic SAL some choices have been made that would deserve to be explored in more detail. What is for example the sensitivity of the probabilistic SAL on the choice of the thresholds to define the probability fields and the threshold to define the probability objects? Furthermore the probability threshold for the observation is somewhat unsatisfying because it leads to entirely flat objects that limit the usefulness of the structure component. It would be interesting to do sensitivity experiments on the treatment of the observation data as well. The precipitation objects in the probabilistic SAL are completely flat due to the deterministic nature of the observations. This is not very satisfying because information on the observed precipitation structure is lost. One could for example use spatial- or temporal neighbourhoods to provide probabilistic observations or somehow take into account an amount difference between the threshold and the local precipitation amount.

A possible application of the SAL in a climate modelling and downscaling context would be to compare the spatial properties of observed and modelled climatologies. This would answer the question if the model simulates on average the highest precipitation in the right places, which is for example not obvious for RCMs in mountainous regions. The drawback of SAL in a climate modelling and downscaling context is that the SAL components are distances between two given situations rather than characteristics of fields. This means that it can not used to compare average spatial characteristics of simulated fields, only the characteristic of the average spatial fields can be compared. To compare average spatial characteristics the method of AghaKouchak et al. (2011) could be adopted.

In chapter 11 only the implication of choices in terms of predictor domains in the downscaling procedure on streamflow at the outlet of the catchment have been analysed. Beside that, the hydrological model simulates other variables like soil moisture and snow cover and it would be interesting to study the effect on these other variables as well.

Possible further developments of SANDHY

SANDHY has been optimised for precipitation, but for hydrological modelling other variables like temperature and evapotranspiration are needed as well. What is the performance of SANDHY for these other predictand variables? First results from Laurie Caillouet's thesis (Caillouet, 2016) show that the amplitude of the yearly cycle is underestimated for temperature (and thus likely for evapotranspiration as well). Furthermore biases in the seasonality of precipitation have been detected in areas with contrasted precipitation regimes, notably in the Cévennes where the precipitation regime is characterised by high precipitation in autumn (Ricard et al., 2012; Godart et al., 2011; Bresson et al., 2012). This is due to the temperature predictor that does not distinguish between spring and autumn

situations. An a posteriori selection using sea surface temperature removed the seasonal precipitation bias. A further development of SANDHY would be therefore to include sea surface temperature as a predictor along with temperature directly in the first analogue selection step. This could be helpful as well in regions with strong influence from the ENSO phenomenon to distinguish el Niño and la Niña years.

Lavers et al. (2013) showed that moisture flux is more correlated to precipitation than moisture, so it would be interesting to test moisture flux as humidity predictor instead of the product of relative humidity and total column water. Furthermore Lafaysse et al. (2014b) claimed that one single downscaling method with a single set of predictors even with probabilistic output will not describe the whole range of uncertainty due to the statistical downscaling and thus versions with different predictors should be used. Following this recommendation further developments could include to combine analogue dates obtained with different predictor variable combinations (with sea surface temperature or without, with vertical velocity or without, with geopotential fields or mean sea level pressure, with moisture flux or relative humidity and total column water as humidity predictor) in addition to different predictor domains as it has already been done in this study.

In the framework of the CORDEX initiative for coordinated downscaling, common predictand domains and experiments have been defined by the community for regional climate modelling and now a similar effort for statistical downscaling CORDEX-ESDM is on the way. For experiments in different climate zones like for example in the Tropics different predictors as for mid-latitude regions will be needed. But which predictors are the most suitable ones for different climate zones? The same question arises in the forecast context when forecasts are delivered for a different region. For example the Compagnie National du Rhône (CNR) wants to provide analogue precipitation forecast for the Mekong catchment. In the Tropics it is likely that predictors related to thermodynamics are more important and circulation related ones less important for daily precipitation than in the mid-latitudes.

When working with observation data directly instead of near surface reanalysis data we have to deal with missing values in the archive data. It may happen that for an analogue date the corresponding observation values are missing. But how to treat this case? Excluding these days from the analogue search makes the analogue search depending on a specific predictand location and contradicts the principle that the same configuration in terms of predictors lead to the same analogue dates. This is a major limitation for the spatially common analogue philosophy while for the local analogue philosophy it may be acceptable. The alternative is to keep the missing values and to accept to have not the same number of analogue precipitation values for each day. If there is occationally a missing value in the

archive, this should not have a dramatic influence on the downscaling results, but if there are longer periods missing, or the missing values are particularly frequent in a specific season, this becomes problematic. This is one example for the very strong dependence of the analogue method on the available data, that is even stronger than for other methods, except maybe the delta change method. Further developments of SANDHY should therefore include the treatment of missing data in the observations.

For historical reasons and notably due to the attempt to reuse code from earlier implementations the CRPS calculation as an objective function in the optimisation and as a validation measure in SANDHY is implemented using the empirical repartition function of Blom (Cunnane, 1978) for calculating the empirical probabilities from the sample. This repartition has in fact found to be rather suited for normally distributed data (Meylan et al., 2008) and is therefore not especially adapted for precipitation. In addition it would be nice to have comparable CRPS values when calculating them with R software and to be able to calculate the decomposition after Hersbach (2000). Recently a CRPS calculation as in the CrpsDecomposition function of the R package verification (NCAR - Research Application Program, 2012) has been implemented in SANDHY and can be used in the validation under certain circumstances. The aim is to implement it as an option in the optimisation as well.

Application for climate projections

Should the downscaling method be applied to projections of future climate there are some additional questions that come into play. For example there is quite some debate if downscaling is actually providing useful information for decision making. The usual top-down approach consists of running a multi-model ensemble of GCMs with possibly several runs each to account for internal variability (Allen et al., 2000), then do some downscaling of these using different methods and again several realisations for stochastic methods. Each of the resulting local scenarios is fed into several impact models. This aims at sampling the known sources of uncertainty at each step (see e.g. Lafaysse et al., 2014a) and results in huge amounts of data and simulations to do. At the end of this uncertainty cascade the uncertainties are very large (Mitchell and Hulme, 1999; Wilby and Dessai, 2010). Comparison studies of downscaling methods demonstrated that the choice of the downscaling method is a major contributor to uncertainty of downscaled climate (Schoof, 2013), at least in summer (Schmidli et al., 2007).

The top-down approach needs methods for intelligently subsetting the multitude of realisations at each step, where the most appropriate subset depends on the subject to be studied and the added challenge that some earth system models are systematically better than others (Brands et al., 2013). Subsetting renders

the modelling chain more parsimonious but requires to know the aim of the study and the sensitivity of the models at each of the steps. GCMs are typically tested for their sensitivity to doubled CO_2 concentrations, and indirectly to increased radiative forcing (Bellouin et al., 2011) in general via the representative concentration pathways (van Vuuren et al., 2011). Downscaling methods are not yet systematically tested for their sensitivity on changes in their input variables but Maraun et al. (2014) proposes a framework to validate downscaling approaches for climate change studies. Hwang and Graham (2013) found that results obtained with the same GCM but different downscaling techniques are more different than those obtained using different GCMs but the same downscaling technique. This is in line with Lafaysse et al. (2014a) who found substantial uncertainty related to the downscaling method.

Pielke and Wilby (2012) advocates the use of a risk oriented bottom-up approach to avoid the downscaling step (see Prudhomme et al. (2010); Ekström et al. (2013) for examples), that consists essentially of a sensitivity analysis of an impact model to different climate forcings. The result is a multidimensional response surface of the modelled output variable to combinations of changes in the input variables. The dimensionality of the response surface depends on the number of characteristics of the input variables that are considered to be important for the system. If thresholds can be defined for the response variable that correspond to a certain risk, like for example the maximum streamflow a given dam is constructed to resist to, it can be determined for which combinations of input characteristics the dam risks to break. The risk due to anthropogenic climate change can then be assessed by looking how many of the GCM runs simulate future input variable characteristics that fall into the dangerous zone. This assumes that the GCMs are able to realistically simulate those characteristics, which is very similar to the assumption done in downscaling that the GCM correctly simulate the large-scale predictors for statistical- or lateral boundary conditions for dynamical downscaling. Collins (2007) goes as far as saying that empirical approaches are not valid for climate change prediction as they could not be reliably used to make extrapolations outside their historical training period but on the other hand sees more use of probabilistic methods in impacts assessment and seamless (probabilistic) prediction of weather and climate from days to seasons and centuries in the future.

Hewitson et al. (2014) argues that in principle there is nothing wrong with research on downscaling, but we should be aware that climate change research and the downscaling that goes along with it is still fundamental research and therefore not ready to be used operationally. Climate scientists are pushed to provide information on climate change, which results in a multitude of sometimes contradictory and poorly documented data. This causes confusion and skepticism. It is

therefore important to provide the methodological limitations and the assumptions along with the data. To be able to provide information on the methodological limitations of a given method, rigorous diagnostics have to be performed under present climate conditions where observations are available.

Related research on analogue downscaling in France

Downscaling with an analogue method and questions about spatial coherence seems to be a topic of interest in France at the moment. Beside this thesis there are two others by Jeremy Chardon (Chardon, 2014) and Gildas Dayon (Dayon, 2015) who work on similar topics with complementary questions. At the same time, as hydroelectricity producers, such as the CNR, had done long time ago, the Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations (SCHAPI National center for hydrometeorology and flood forecasting support) seeks to implement an analogue downscaling model for operational use in flood forecasting (Marty and Gautheron, 2014). However the SCHAPI seeks a parsimonious implementation with only a few predictor domains for the mainland of France. The results of this thesis and the ones from Chardon et al. (2014) suggest to use several predictor domains, at least a separate one for the southeastern part of France with its Mediterranean climate if some skill for this region is of interest.

Bibliography

AghaKouchak, A., Nasrollahi, N., Li, J., Imam, B., and Sorooshian, S. (2011). Geometrical characterization of precipitation patterns. *Journal of Hydrometeorology*, 12:274–285.

Ailliot, P., Allard, D., Monbet, V., and Naveau, P. (2014). Stochastic weather generators : an overview of weather type models.

Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R., and Delworth, T. L. (2000). Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, 407:617–620.

Amengual, A., Homar, V., Romero, R., Alonso, S., and Ramis, C. (2012). A statistical adjustment of regional climate model outputs to local scales: Application to Platja de Palma, Spain. *Journal of Climate*, 25:939–957.

Arslan, Y. Z., Demirer, R. M., Palamar, D., Ugur, M., and Karamehmetoglu, S. S. (2012). Comparison of the data classification approaches to diagnose spinal cord injury. *Computational and Mathematical Methods in Medicine*, 2012:ID803980.

Auffray, A., Clavel, A., Jourdain, S., Ben Daoud, A., Sauquet, E., Lang, M., Obled, C., Panthou, G., Gautheron, A., Gottardi, F., and Garçon, R. (2011). Reconstitution hydrométéorologique de la crue de l'Isère de 1859 (reconstructing the hydrometeorological scenario of the 1859 flood of the Isère river). *La Houille Blanche*, (1):44–50. (in French).

Bannayan, M. and Hoogenboom, G. (2008a). Predicting realizations of daily weather data for climate forecasts using the non-parametric nearest-neighbour re-sampling technique. *International Journal of Climatology*, 28(10):1357–1368.

Bannayan, M. and Hoogenboom, G. (2008b). Weather analogue: A tool for real-time prediction of daily weather data realizations based on a modified k-nearest neighbor approach. *Environmental Modelling and Software*, 23(6):703–713.

Bárdossy, A., Stehlík, J., and Caspary, H.-J. (2002). Automated objective classification of daily circulation patterns for precipitation and temperature downscaling based on optimized fuzzy rules. *Climate Research*, 23:11–22.

Barnett, T. P. and Preisendorfer, R. W. (1978). Multifield analog prediction of short-term climate fluctuations using a climate state vector. *Journal of Atmospheric Sciences*, 35(10):1771–1787.

Bellone, E., Hughes, J. P., and Guttorp, P. (2000). A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate Research*, 15:1–12.

Bellouin, N., Rae, J., Jones, A., Johnson, C., Haywood, J., and Boucher, O. (2011). Aerosol forcing in the Climate Model Intercomparison Project (CMIP5) simulations by HadGEM2-ES and the role of ammonium nitrate. *Journal of Geophysical Research*, 116:D20206.

Ben Daoud, A. (2010). *Amélioratıons et développements d'une méthode de prévision probabiliste des pluies par analogie. Application à la prévision hydrologique sur les grands bassins fluviaux de la Saône et de la Seine.* Thèse de doctorat, Université Joseph Fourier, Grenoble. (in French).

Ben Daoud, A., Obled, C., Lang, M., Sauquet, E., and Bontron, G. (2009a). Comparison of 850-hPa relative humidity between ERA-40 and NCEP/NCAR re-analyses: detection of suspicious data in ERA-40. *Atmospheric Science Letters*, 10(1):43–47.

Ben Daoud, A., Sauquet, E., Lang, M., Bontron, G., and Obled, C. (2011a). Precipitation forecasting through an analog sorting technique: a comparative study. *Advances in Geosciences*, 29:103–107.

Ben Daoud, A., Sauquet, E., Lang, M., Obled, C., and Bontron, G. (2009b). La prévision des précipitations par recherche d'analogues : ''etat de l'art et perspectives (Precipitation forecasting through an analog sorting technique: state of the art and further investigations). *La Houille Blanche*, 6:60–65. (in French).

Ben Daoud, A., Sauquet, E., Lang, M., and Ramos, M.-H. (2011b). Peut-on étendre l'échéance de prévision des crues en optimisant la prévision de pluies par recherche d'analogues ? Application au bassin de la Seine à Paris (Can we extend flood forecasting lead-time by optimising precipitation forecasting based on analogs? Application to the Seine river basin). *La Houille Blanche*, (1):37–43. (in French).

Benestad, R. E. (2004). Tentative probabilistic temperature scenarios for northern Europe. *Tellus A*, 56A(2):89–101.

Benestad, R. E. (2007). Empirical-statistical downscaling. Technical report, Norwegian Meteorological Institute and Earth Sciences Centre, Gothenburg University.

Benestad, R. E. (2010). Downscaling precipitation extremes – Correction of analog models through PDF predictions. *Theoretical and Applied Climatology*, 100(1-2):1–21.

Bertin, J. (1967). *Semiology of Graphics*. Esri Press, 2010 edition.

Beuchat, X., Schaefli, B., Soutter, M., and Mermoud, A. (2012). A robust framework for probabilistic precipitations downscaling from an ensemble of climate predictions applied to Switzerland. *Journal of Geophysical Research Atmosphere*, 117:D03115.

Blenkinsop, S., Halllett, S., Truckell, I., and Fowler, H. J. (2010). The CREW project: Towards a toolkit for the use of probabilistic climate change projections. In *BHS Third International Symposium, Managing Consequences of a Changing Global Environment*.

Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. *Bioinformatics*, 27:2463–2464.

Boé, J. and Terray, L. (2008). A weather-type approach to analyzing winter precipitation in France: Twentieth-century trends and the role of anthropogenic forcing. *Journal of Climate*, 21(13):3118–3133.

Boé, J., Terray, L., Habets, F., and Martin, E. (2006). A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. *Journal of Geophysical Research*, 111(23):D23106.

Boé, J., Terray, L., Habets, F., and Martin, E. (2007). Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *International Journal of Climatology*, 27(12):1643–1655.

Boé, J., Terray, L., Martin, E., and Habets, F. (2009). Projected changes in components of the hydrological cycle in French river basins during the 21st century. *Water Resources Research*, 45(8):W08426.

Bompart, P., Bontron, G., Celie, S., and Haond, M. (2009). Une chaîne opérationnelle de prévision hydrométéorologique pour les besoins de la production hydroélectrique de la CNR (An operational hydrometeorological forecasting

chain for CNR's hydroelectric production needs). *La Houille Blanche*, (5):54–60. in French.

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. (in French).

Bontron, G., Djerboua, A., and Obled, C. (2002). Sélection de situations météorologiques analogues : applications en prévision opérationnelle de précipitations et en évolution climatique (Analog sorting of meteorological patterns: applications to operational precipitation forecast and climate evolution studies). *La Houille Blanche*, (8):46–51. in French.

Bontron, G. and Obled, C. (2005). L´adaptation probabiliste des prévisions météorologiques pour la prévision hydrologique. *La Houille Blanche*, (1):23–28. (in French).

Brands, S., Herrera, S., Fernandez, J., and Gutierrez, J. M. (2013). How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa? A performance comparison for the downscaling community. *Climate Dynamics*, 41:803–817.

Branger, F., Kermadi, S., Jacqueminet, C., Michel, K., Labbas, M., Krause, P., Kralisch, S., and Braud, I. (2013). Assessment of the influence of land use data on the water balance components of a peri-urban catchment using a distributed modelling approach. *Journal of Hydrology*, 505:312–325.

Bresson, E., Ducrocq, V., Nuissier, O., Ricard, D., and de Saint-Aubin, C. (2012). Idealized numerical simulations of quasi-stationary convective systems over the Northwestern Mediterranean complex terrain. *Quarterly Journal of the Royal Meteorological Society*, 138:1751–1763.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Briggs, W. M. and Levine, R. A. (1997). Wavelets and field forecast verification. *Monthly Weather Review*, 125:1329–1341.

Brown, T. A. (1974). Admissible scoring systems for continuous distributions. The Rand Paper Series, P-5235, The Rand Corporation, Santa Monica, California.

Brussolo, E., Von Hardenberg, J., Ferraris, L., Rebora, N., and Provenzale, A. (2008). Verification of quantitative precipitation forecasts via stochastic downscaling. *Journal of Hydrometeorology*, 9(5):1084–1094.

Buishand, T. A. and Brandsma, T. (2001). Multisite simulation of daily precipitation and temperature in the Rhine Basin by nearest-neighbor resampling. *Water Resources Research*, 37(11):2761–2776.

Bulygina, N., McIntyre, N., and Wheater, H. (2009). Conditioning rainfall-runoff model parameters for ungauged catchments and land management impacts analysis. *Hydrology and Earth System Sciences*, 13:893–904.

Burton, A., Glenis, V., Bovolo, C., Blenkinsop, S., Fowler, H. J., Chen, A. S., Djordjević, S., and Kilsby, C. (2010). Stochastic rainfall modelling for the assessment of urban flood hazard in a changing climate. In *BHS Third International Synposium, Managing Consequences of a Changing Globla Environment*.

Burton, A., Kilsby, C. G., Fowler, H. J., Cowpertwait, P. S. P., and O'Connell, P. E. (2008). RainSim: A spatial-temporal stochastic rainfall modelling system. *Environmental Modelling and Software*, 23:1356–1369.

Caillouet, L. (2016). *Hydrometeorological reconstruction of historical droughts and low flows in France: 1871–2010*. PhD thesis, Université Joseph Fourier, Grenoble. in progress.

Cannon, A. J. (2007). Nonlinear analog predictor analysis: A coupled neural network/analog model for climate downscaling. *Neural Networks*, 20(4):444–453.

Cannon, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers and Geosciences*, 37:1277–1284.

Carreau, J. and Vrac, M. (2011). Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resources Research*, 47:W10502.

Casati, B., Ross, G., and Stephenson, D. B. (2004). A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, 11:141–154.

Champeaux, J.-L. and Tamburini, A. (1996). Zonage climatique de la France à partir de précipitations (1971-1990) du réseau climatologique d'état (Climatological zoning of France from precipitation measurements (1971-1990) of the French climatological network). *La Météorologie*, (14):44–54.

Chandler, R. E. (2005). On the use of generalized linear models for interpreting climate variability. *Evironmetrics*, 16:699–715.

Chandler, R. E. and Weather, H. S. (2002). Analysis of rainfall variability using generalized linear models: A case study from the west of ireland. *Water Resources Research*, 38:10–1–10–11.

Chardon, J. (2014). *Intérêts de la méthode des analogues pour la génération de scénarios de précipitations à l'échelle de la France métropolitaine. Cohérence spatiale et adaptabilité du lien d'échelle.* PhD thesis, Université Joseph Fourier, Grenoble.

Chardon, J., Hingray, B., Favre, A.-C., Autin, P., Gailhard, J., Zin, I., and Obled, C. (2014). Spatial similarity and transferability of analog dates for precipitation downscaling over France. *Journal of Climate*, 27(13):5056–5074.

Charles, S. P., Bates, B. C., Smith, I. N., and Hughes, J. P. (2004). Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrological Processes*, 18:1373–1394.

Chen, J., Brissette, F. P., and Leconte, R. (2010). A daily stochastic weather generator for preserving low-frequency of climate variability. *Journal of Hydrology*, 388(3-4):480–490.

Chiew, F. H. S., Kirono, D. G. C., Kent, D. M., Frost, A. J., Charles, S. P., Timbal, B., Nguyen, K. C., and Fu, G. (2010). Comparison of runoff modelled using rainfall from different downscaling methods for historical and future climates. *Journal of Hydrology*, 387(1-2):10–23.

Christensen, J. H. and Christensen, O. B. (2003). Severe summertime flooding in Europe. *Nature*, 421(6925):805–806.

Christensen, J. H. and Christensen, O. B. (2007). A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Climatic Change*, 81, Supplement 1:7–30.

Collins, M. (2007). Ensembles and probabilities: a new era in the prediction of climate change. *Philosophical Transactions of the Royal Society Series A*, 365(1857):1957–1970.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Bröcker, J., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137(654):1–28.

Conway, D., Wilby, R. L., and Jones, P. D. (1996). Precipitation and air flow indices over the British Isles. *Climate Research*, 7:169–183.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels.

Cordano, E. and Eccel, E. (2012). RMAWGEN: A software project for a daily multi-site weather generator with R. In *Geophysical research abstracts*. poster.

Cunnane, C. (1978). Unbiased plotting positions — a review. *Journal of Hydrology*, 37(3–4):205 – 222.

Davis, C., Brown, B., and Bullock, R. (2006). Object-based verification of precipitation forecasts. part i: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, 134:1772–1784.

Davis, C. A., Brown, B. G., Bullock, R., and Halley-Gotway, J. (2009). The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Weather and Forecasting*, 24:1252–1267.

Dayon, G. (2015). *Évolution du cycle hydrologique sur la France au cours des prochaines décennies*. PhD thesis. in progress.

Dee, D. P., Balmaseda, M., Balsamo, G., Engelen, R., Simmons, A. J., and Thépaut, J.-N.paut, J.-N. (2014). Toward a consistent reanalysis of the climate system. *Bulletin of the American Meteorological Society*.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597.

Denis, B., Coté, J., and Laprise, R. (2002). Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Monthly Weather Review*, 130:1812–1829.

Diday, E., Ok, Y., and Schroeder, A. (1974). The dynamic clusters method in pattern recognition. In Rosenfeld, J. L., editor, *Proceedings of the IFIP Congress 74*, pages 691–697, New York. Elsevier.

Duc, L., Saito, K., and Seko, H. (2011). Application of spatial-temporal Fractions Skill Score to high-resolution ensemble forecast verification. In *5th International Verification Methods Workshop*.

Dunn, P. K. (2004). Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, 24:1231–1239.

Ebert, E. and McBride, J. (2000). Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology*, 239(1-4):179–202.

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, 15:51–64.

Eeckman, J. (2014). Impact de la cohérence spatiale des champs de précipitation sur les débits modélisés de la Durance. Master's thesis, Université de Nantes, Master 2 Carthographie et gestion de l'Environnement.

Ekström, M., Kuruppu, N., Wilby, R. L., Fowler, H. J., Chiew, F. H., Dessai, S., and Young, W. J. (2013). Examination of climate risk using a modified uncertainty matrix framework–Applications in the water sector. *Global Environmental Change*, 23(1):115–129.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8:985–987.

Fernández, A. and Gómez, S. (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, 25:43–65.

Fernández, J. and Sáenz, J. (2003). Improved field reconstruction with the analog method: searching the CCA space. *Climate Research*, 24(3):199–213.

Fernández-Ferrero, A., Sáenz, J., and Ibarra-Berastegi, G. (2010). Comparison of the performance of different analog-based bayesian probabilistic precipitation forecasts over Bilbao, Spain. *Monthly Weather Review*, 138:3107–3119.

Fernández-Ferrero, A., Sáenz, J., Ibarra-Berastegi, G., and Fernández, J. (2009). Evaluation of statistical downscaling in short range precipitation forecasting. *Atmospheric Research*, 94(3):448–461.

Flügel, W. A. (1995). Delineating hydrological response units by geographical information-system analyses for regional hydrological modeling using prms/mms in the drainage-basin of the river Brol, Germany. *Hydrological Processes*, 9:423–436.

Formayer, H., Haas, P., Hofstätter, M., Radanovics, S., and Kromp-Kolb, H. (2008). Räumlich und zeitlich hochaufgelöste Temperaturszenarien für Wien und ausgewählte Analysen bezüglich Adaptionsstrategien (spatially and temporally high-resolution temperature scenarios for Vienna and selected analysis regarding adaptation strategies. Technical report, from Universität für Bodenkultur Wien for Vienna environmental protection department, department for EU-Strategy and economical developement.

Fowler, H. J., Blenkinsop, S., and Tebaldi, C. (2007). Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, 27(12):1547–1578.

Fowler, H. J., Kilsby, C. G., O'Connell, P. E., and Burton, A. (2005). A weather-type conditioned multi-site stochastic rainfall model for the generation of scenarios of climatic variability and change. *Journal of Hydrology*, 308(1-4):50–66.

Fraley, C., Raftery, A. E., Sloughter, J. M., Gneiting, T., and of Washington., U. (2013). *ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging*. R package version 5.0.4.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315:972–977.

Friederichs, P. and Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135:2365–2378.

Frost, A. J., Charles, S. P., Timbal, B., Chiew, F. H., Mehrotra, R., Nguyen, K. C., Chandler, R. E., McGregor, J. L., Fu, G., Kirono, D. G., Fernandez, E., and Kent, D. M. (2011). A comparison of multi-site daily rainfall downscaling techniques under Australian conditions. *Journal of Hydrology*, 408:1–18.

Früh, B., Schipper, J. W., Pfeiffer, A., and Wirth, V. (2006). A pragmatic approach for downscaling precipitation in alpine-scale complex terrain. *Meteorologische Zeitschrift*, 15(6):631–646.

Gangopadhyay, S., Clark, M., and Rajagopalan, B. (2005). Statistical downscaling using K-nearest neighbors. *Water Resources Research*, 41:W02024.

Gibergans-Báguena, J. and Llasat, M. C. (2007). Improvement of the analog forecasting method by using local thermodynamic data. application to autumn precipitation in Catalonia. *Atmospheric Research*, 86(3-4):173–193.

Gilleland, E. (2013). *SpatialVx: Spatial Forecast Verification*. R package version 0.1-5.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24:1416–1430.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Godart, A., Anquetin, S., Leblois, E., and Creutin, J.-D. (2011). The contribution of orographically driven banded precipitation to the rainfall climatology of a Mediterranean region. *Journal of Applied Meteorology and Climatology*, 50(11):2235–2246.

Gomez, S., Fernandez, A., Granell, C., and Arenas, A. (2013). Structural patterns in complex systems using multidendrograms. *Entropy*, 15:5464–5474.

Goodess, C. and Palutikof, J. (1998). Development of daily rainfall scenarios for southeast Spain using a circulation-type approach to downscaling. *International Journal of Climatology*, 18:1051–1083.

Goodess, C. M. and Jones, P. D. (2002). Links between circulation and changes in the characteristics of Iberian rainfall. *International Journal of Climatology*, 22:1593–1615.

Gottardi, F., Obled, C., Gailhard, J., and Paquet, E. (2012). Statistical reanalysis of precipitation fields based on ground network data and weather patterns: Application over French mountains. *Journal of Hydrology*, 432:154–167.

Greene, A. M., Robertson, A. W., Smyth, P., and Triglia, S. (2011). Downscaling projections of Indian monsoon rainfall using a non-homogeneous hidden Markov model. *Quarterly Journal of the Royal Meteorological Society*, 137:347–359.

Gudefin, L. (2013). Modélisation hydrologique spatialisée du bassin versant périurbain de la Chézine avec le modèle J2000. Master's thesis, Mines d'Alès.

Guilbaud, S. and Obled, C. (1998a). L'approche par analogues en prévision météorologique : Idées générales et application la prévision journalière des précipitations = Meteorological forecasting with analogues; general ideas and application to the daily forecast of precipitation. *La Météorologie*, 24:21–35.

Guilbaud, S. and Obled, C. (1998b). Prévision quantitative des précipitations journalières par une technique de recherche de journées antérieures analogues : optimisation du critère d'analogie (Daily quantitative precipitation forecast by an analogue technique: optimisation of the analogy criterion). *Compte Rendus de l'Académie des Sciences - Series IIA - Earth and Planetary Science*, 327(3):181–188. in French.

Gupta, H. V., Kling, H., Yilmaz, K. K., and 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):90–91.

Gutiérrez, J. M., Cofiño, A. S., Cano, R., and Rodríguez, M. A. (2004). Clustering methods for statistical downscaling in short-range weather forecasts. *Monthly Weather Review*, 132(9):2169–2183.

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., and Herrera, S. (2013). Reassessing statistical downscaling techniques for their robust application under climate change conditions. *Journal of Climate*, 26:171–188.

Gutmann, E. D., Rasmussen, R. M., Liu, C., Ikeda, K., Gochis, D. J., Clark, M. P., Dudhia, J., and Thompson, G. (2012). A comparison of statistical and dynamical downscaling of winter precipitation over complex terrain. *Journal of Climate*, 25(1):262–281.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C. (2011). The integrated nowcasting through comprehensive analysis (INCA) system and its validation over the eastern Alpine region. *Weather and Forecasting*, 26:166–183.

Hamill, T. M. (1999). Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, 14:155–167.

Hamill, T. M. and Colucci, S. J. (1998). Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, 126:711–724.

Hamill, T. M. and Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, 134(11):3209–3229.

Hanssen-Bauer, I., Førland, E. J., Haugen, J. E., and Tveito, O. E. (2003). Temperature and precipitation scenarios for Norway: comparison of results from dynamical and empirical downscaling. *Climate Research*, 25(1):15–27.

Harpham, C. and Wilby, R. L. (2005). Multi-site downscaling of heavy daily precipitation occurrence and amounts. *Journal of Hydrology*, 312(1-4):235–255.

Haylock, M. R., Cawley, G. C., Harpham, C., Wilby, R. L., and Goodess, C. M. (2006). Downscaling heavy precipitation over the UK: a comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology*, 26(10):1397–1415.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M. (2008). A european daily high-resolution gridded dataset of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research*, 113(D20119).

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15:559–570.

Hertig, E., Paxian, A., Vogt, G., Seubert, S., Paeth, H., and Jacobeit, J. (2012). Statistical and dynamical downscaling assessments of precipitation extremes in the Mediterranean area. *Meteorologische Zeitschrift*, 21(1):61–77.

Hewitson, B., Daron, J., Crane, R., Zermoglio, F., and Jack, C. (2014). Interrogating empirical-statistical downscaling. *Climatic Change*, 122:539–554.

Hewitson, B. C. and Crane, R. G. (2006). Consensus between GCM climate change projections with empirical downscaling. *International Journal of Climatology*, 26(10):1315–1337.

Hidalgo, H. G., Dettinger, M. D., and Cayan, D. R. (2008). Downscaling with constructed analogues: daily precipitation and temperature fields over the United States. California Climate Change Center, Report Series 2007-027, California Energy Commission, Public Interest Energy Research Program.

Hingray, B. and Mezghani, A. (2007). Utilisation des réanalyses NCEP pour la génération de scénarios météorologiques. application pour la génération de scénarios de crues pour le Rhône à l´amont du Leman (Generation of meteorological scenarios from NCEP reanalyses. application for the generation of flood scenarios for the Rhône upstream to Lake Leman). *La Houille Blanche*, (6):104–110. (in French).

Hohenegger, C., Brockhaus, P., and Schär, C. (2008). Towards climate simulations at cloud-resolving scales. *Meteorologische Zeitschrift*, 17:383–394.

Horton, P. (2012). *Améliorations et optimisation globale de la méthode des analogues pour la prévision statistique des précipitations. Développement d'un outil*

de prévision et application opérationnelle au bassin du Rhône à l'amont du Léman. Thèse de doctorat, Université de Lausanne, Switzerland.

Horton, P., Jaboyedoff, M., Metzger, R., Obled, C., and Marty, R. (2012). Spatial relationship between the atmospheric circulation and the precipitation measured in the western Swiss Alps by means of the analogue method. *Natural Hazards and Earth System Sciences*, 12:777–784.

Hughes, J. P. and Guttorp, P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Applied Statistics*, 48:15–30.

Huza, J. (2013). Measuring soil moisture dynamics at multiple scales and understanding its response to precipitation characteristics in a small catchment in southern France. Master's thesis, Wageningen University.

Hwang, S. and Graham, W. D. (2013). Development and comparative evaluation of a stochastic analog method to downscale daily GCM precipitation. *Hydrology and Earth System Sciences*, 17:2141–2181.

Imbert, A. and Benestad, R. E. (2005). An improvement of analog model strategy for more reliable local climate change scenarios. *Theoretical and Applied Climatology*, 82(3-4):245–255.

Jolliffe, I. T., Stephenson, D. B., Broecker, J., Brown, B. G., Déqué, M., Ebert, E. E., Ferro, C. A. T., Gilleland, E., Hogan, R., Livezey, R. E., Mason, I. B., Mason, S. J., Pocernich, M., Potts, J. M., Richardson, D. S., and Weigel, A. P. (2012). *Forecast Verification: A Practioner's Guide in Atmospheric Science*. Wiley-Blackwell.

Kallache, M., Vrac, M., Naveau, P., and Michelangeli, P.-A. (2011). Nonstationary probabilistic downscaling of extreme precipitation. *Journal of Geophysical Research*, 116(5):D05113.

Kalnay, E., Kanamitsu, M., Kistler, R., W., C., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janoviak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471.

Keil, C. and Craig, G. C. (2007). A displacement-based error measure applied in a regional ensemble forecasting system. *Monthly Weather Review*, 135:3248–3259.

Keil, C. and Craig, G. C. (2009). A displacement and amplitude score employing an optical flow technique. *Weather and Forecasting*, 24:1297–1308.

Khan, M. S., Coulibaly, P., and Dibike, Y. (2006). Uncertainty analysis of statistical downscaling methods. *Journal of Hydrology*, 319:357–382.

Kilsby, C. G., Jones, P. D., Burton, A., Ford, A. C., Fowler, H. J., Harpham, C., James, P., Smith, A., and Wilby, R. L. (2007). A daily weather generator for use in climate change studies. *Environmental Modelling and Software*, 22(12):1705–1719.

Kioutsioukis, I., Melas, D., and Zanis, P. (2008). Statistical downscaling of daily precipitation over Greece. *International Journal of Climatology*, 28(5):679–691.

Kling, H., Fuchs, M., and Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424-425:264–277.

Krause, P. (2002). Quantifying the impact of land use changes on the water balance of large catchments using the J2000 model. *Physics and Chemistry of the Earth*, 27:663–673.

Krause, P., Bäse, F., Bende-Michl, U., Fink, M., Flügel, W., and Pfennig, B. (2006). Multiscale investigations in a mesoscale catchment – Hydrological modelling in the Gera catchment. *Advances in Geosciences*, 9:53–61.

Krzysztofowicz, R. (1983). Why should a forecaster and a decision maker use bayes theorem. *Water Resources Research*, 19(2):327–336.

Krzysztofowicz, R. (1998). Probabilistic hydrometeorological forecasts: Toward a new era in operational forecasting. *Bulletin of the American Meteorological Society*, pages 243–251.

Labbas, M. (2014). *Modélisation hydrologique de bassins versants périurbains et influence de l'occupation du sol et de la gestion des eaux pluviales - Application au bassin de l'Yzeron.* PhD thesis, University Joseph Fourier Grenoble. manuscript in preparation.

Lafaysse, M., Hingray, B., Terray, L., Gailhard, J., and Mezghani, A. (2014a). Sources of uncertainties in future climate and hydrological projections: the Alpine Durance basin. *Water Resources Research.* under review.

Lafaysse, M., Hingray, B., Terray, L., Mezghani, A., and Gailhard, J. (2014b). Internal variability and model uncertainty components in future hydrometeorological projections: The Alpine Durance basin. *Water Resources Research.*

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11:1267–1277.

Lall, U. and Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, 32(3):679–693.

Lavaysse, C., Vrac, M., Drobinski, P., Lengaigne, M., and Vischel, T. (2012). Statistical downscaling of the French Mediterranean climate: assessment for present and projection in an anthropogenic scenario. *Natural Hazards and Earth System Sciences*, 12(3):651–670.

Lavers, D., Prudhomme, C., and Hannah, D. (2013). European precipitation connections with large-scale mean sea-level pressure (MSLP) fields. *Hydrological Sciences Journal*, 58(2):310–327.

Leblois, E. and Creutin, J.-D. (submitted (2012)). Space-time simulation of intermittent rainfall with prescribed advection field: adaptation of the turnig band method. *Water Resources Research*.

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646.

Magand, C., Ducharne, A., Le Moine, N., and Gascoin, S. (2014). Introducing hysteresis in snow depletion curves to improve the water budget of a land surface model in an Alpine catchment. *Journal of Hydrometeorology*, 15(2):631–649.

Maraun, D. (2013). Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, 26:2137–2143.

Maraun, D., Osborn, T. J., and Rust, H. W. (2011). The influence of synoptic airflow on UK daily precipitation extremes. part I: Observed spatio-temporal relationships. *Climate Dynamics*, 36:261–275.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I. (2010). Precipitation downscaling under climate change. recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48:RG3003.

Maraun, D., Widmann, M., Benestad, R. E., Kotlarski, S., Hertig, E., Wibig, J., Gutierrez, J. M., Huth, R., Chandler, R. E., and Wilcke, R. A. (2014). VALUE - a framework to validate downscaling approaches for climate change studies. *Earth's Future*. in press.

Martin, E., Timbal, B., and Brun, E. (1997). Downscaling of general circulation model outputs: simulation of the snow climatology of the French Alps and sensitivity to climate change. *Climate Dynamics*, 13(1):45–56.

Marty, R. and Gautheron, A. (2014). Plate-forme analogues : Analyse des besoins et propositions techniques. Rapports chantier d'intérêt commun analogues, DREAL.

Marty, R., Zin, I., and Obled, C. (2013). Sensitivity of hydrological ensemble forecasts to different sources and temporal resolutions of probabilistic quantitative precipitation forecasts: flash flood case studies in the Cévennes-Vivarais region (Southern France). *Hydrological Processes*, 27:33–44.

Marty, R., Zin, I., Obled, C., Bontron, G., and Djerboua, A. (2012). Toward real-time daily PQPF by an analog sorting approach: Application to flash-flood catchments. *Journal of Applied Meteorology and Climatology*, 51:505–520.

Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, 30:291–303.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1095.

Matulla, C., Penlap, E. K., Haas, P., and Formayer, H. (2003). Comparative analysis of spatial and seasonal variability: Austrian precipitation during the 20th century. *International Journal of Climatology*, 23:1577–1588.

Matulla, C., Zhang, X., Wang, X. L., Wang, J., Zorita, E., Wagner, S., and von Storch, H. (2008). Influence of similarity measures on the performance of the analog method for downscaling daily precipitation. *Climate Dynamics*, 30(2-3):133–144.

Mehrotra, R. and Sharma, A. (2005). A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences. *Journal of Geophysical Research*, 110(16):D16108.

Meylan, P., Favre, A., and Musy, A. (2008). *Hydrologie fréquentielle: une science prédictive*. Science & ingénierie de l'environnement. Presses polytechniques et universitaires romandes.

Mezghani, A. and Hingray, B. (2009). A combined downscaling-disaggregation weather generator for stochastic generation of multisite hourly weather variables over complex terrain: Development and multi-scale validation for the Upper Rhône River basin. *Journal of Hydrology*, 377(3-4):245–260.

Micheas, A. C., Fox, N. I., Lack, S. A., and Wikle, C. K. (2007). Cell identification and verification of QPF ensembles using shape analysis techniques. *Journal of Hydrology*, 343:105–116.

Michelangeli, P.-A., Vrac, M., and Loukos, H. (2009). Probabilistic downscaling approaches: Application to wind cumulative distribution functions. *Geophysical Research Letters*, 36(L11708):L11708.

Mitchell, T. D. and Hulme, M. (1999). Predicting regional climate change: living with uncertainty. *Progress in Physical Geography*, 23(1):57–78.

Moron, V., Robertson, A. W., and Ward, M. N. (2006). Seasonal predictability and spatial coherence of rainfall characteristics in the tropical setting of senegal. *Monthly Weather Review*, 134:3248–3262.

Murphy, A. H. (1966). A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *Journal of Applied Meteorology*, 5:534–537.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105(7):803–816.

Murphy, Allan, H. (1969). On the "ranked probability score". *Journal of Applied Meteorology*, 8:988–989.

Nash, J. E. and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. part 1: A discussion of principles. *Journal of Hydrology*, 10(3):282–290.

NCAR - Research Application Program (2012). *verification: Forecast verification utilities.* R package version 1.35.

Nepal, S., Krause, P., Flügel, W.-A., Fink, M., and Fischer, C. (2014). Understanding the hydrological system dynamics of a glaciated alpine catchment in the Himalayan region using the J2000 hydrological model. *Hydrological Processes*, 28(3):1329–1344.

Nicholas, R. E. and Battisti, D. S. (2012). Empirical downscaling of high-resolution regional precipitation from large-scale reanalysis fields. *Journal of Applied Meteorology and Climatology*, 51:100–114.

Ning, L., Mann, M. E., Crane, R., and Wagener, T. (2012). Probabilistic projections of climate change for the Mid-Atlantic region of the United States: Validation of precipitation downscaling during the historical era. *Journal of Climate*, 25(2):509–526.

Nuissier, O., Ducrocq, V., Ricard, D., Lebeaupin, C., and Anquetin, S. (2008). A numerical study of three catastrophic precipitating events over southern France. I: Numerical framework and synoptic ingredients. *Quarterly Journal of the Royal Meteorological Society*, 134:111–130.

Obled, C., Bontron, G., and Garçon, R. (2002). Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach. *Atmospheric Research*, 63(3-4):303–324.

Obled, C., Panthou, G., Garçon, R., and Gottardi, F. (2009). Crue de l'Isère 1859 : Reconstitution d'un scénario de précipitations et de températures pour la période 1er Octobre - 5 Novembre 1859. In *Colloque scientificque Isère 1859-2009 – 150e anniversaire de la crue de référence*. PGRN and INP Grenoble and DDE 38.

Orlowsky, B., Bothe, O., Fraedrich, K., Gerstengarbe, F.-W., and Zhu, X. (2010). Future climates from bias-bootstrapped weather analogs: An application to the Yangtze River Basin. *Journal of Climate*, 23(13):3509–3524.

Orlowsky, B., Gerstengarbe, F.-W., and Werner, P. C. (2008). A resampling scheme for regional climate simulations and its performance compared to a dynamical RCM. *Theoretical and Applied Climatology*, 92(3-4):209–223.

Pfeiffer, A. and Zängl, G. (2011). Regional climate simulations for the European Alpine Region – sensitivity of precipitation to large-scale flow conditions of driving input data. *Theoretical and Applied Climatology*, 105:325–340.

Philandras, C. M., Nastos, P. T., Kapsomenakis, J., Douvis, K. C., Tselioudis, G., and Zerefos, C. S. (2011). Long term precipitation trends and variability within the Mediterranean region. *Natural Hazards and Earth System Sciences*, 11(12):3235–3250.

Philipp, A., Bartholy, J., Beck, C., Erpicum, M., Esteban, P., Fettweis, X., Huth, R., James, P., Jourdain, S., Kreienkamp, F., Krennert, T., Lykoudis, S., Michalides, S. C., Pianko-Kluczynska, K., Post, P., Álvarez, D. R., Schiemann, R., Spekat, A., and Tymvios, F. S. (2010). COST733CAT – a database of weather and circulation type classifications. *Physics and Chemistry of the Earth*.

Piani, C., Haerter, J. O., and Coppola, E. (2010). Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, 99(1-2):187–192.

Piazza, M., Pagé, C., Sanchez, E., and Terray, L. (2011). Comparaison des méthodes de désagrégation statistique et dynamique pour l'évaluation du changement climatique sur les zones de montagnes en France. Scampei rapport semestriel d'activité, CERFACS.

Pielke, R. A. and Wilby, R. L. (2012). Regional climate downscaling: What's the point? *EOS, Transactions of the American Geophysical Union*, 93(5):52–53.

Plaut, G., Schuepbach, E., and Doctor, M. (2001). Heavy precipitation events over a few Alpine sub-regions and the links with large-scale circulation, 1971–1995. *Climate Research*, 17:285–302.

Prudhomme, C., Wilby, R. L., Crooks, S., Kay, A. L., and Reynard, N. S. (2010). Scenario-neutral approach to climate change impact studies: Application to flood risk. *Journal of Hydrology*, 390(3-4):198–209.

Pushpalatha, R., Perrin, C., Le Moine, N., and Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, 420–421:171–182.

Quintana Seguí, P., Habets, F., and Martin, E. (2011). Comparison of past and future Mediterranean high and low extremes of precipitation and river flow projected using different statistical downscaling methods. *Natural Hazards and Earth System Sciences*, 11(5):1411–1432.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchistéguy, L., and Morel, S. (2008). Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France. *Journal of Applied Meteorology and Climatology*, 47(1):92–107.

Radanovics, S., Vidal, J.-P., Sauquet, E., Ben Daoud, A., and Bontron, G. (2013a). Optimising predictor domains for spatially coherent precipitation downscaling. *Hydrology and Earth System Sciences*, 17(10):4189–4208.

Radanovics, S., Vidal, J.-P., Sauquet, E., Bontron, G., and Ben Daoud, A. (2012). Optimising predictor domains for spatially coherent precipitation downscaling. Presentation at EGU General Assembly 2012, Vienna, Austria.

Radanovics, S., Vidal, J.-P., Sauquet, E., Bontron, G., and Ben Daoud, A. (2013b). Coherent predictand areas for spatially coherent precipitation downscaling. Presentation at EGU General Assembly 2013, Vienna, Austria.

Radanovics, S., Vidal, J.-P., Sauquet, E., Bontron, G., and Ben Daoud, A. (2013c). Defining predictand areas with homogeneous predictors for spatially coherent

precipitation downscaling of climate projections. Presentation at the International Meeting on Statistical Climatology, Jeju, South Corea.

Radanovics, S., Vidal, J.-P., Sauquet, E., Bontron, G., and Ben Daoud, A. (2013d). Spatio-temporal coherence in probabilistic downscaled precipitation for hydrological modelling. Poster at the Next Generation Climate Data Products Workshop, NCAR Boulder, Colorado, USA.

Radanovics, S., Vidal, J.-P., Sauquet, E., Bontron, G., and Ben Daoud, A. (2014). Stepwise analogue downscaling for hydrology (SANDHY): validation experiments over France. poster at the EGU General Assembly 2014, Vienna, Austria.

Rakovec, O., Hazenberg, P., Torfs, P. J. J. F., Weerts, A. H., and Uijlenhoet, R. (2012). Generating spatial precipitation ensembles: impact of temporal correlation structure. *Hydrology and Earth System Sciences*, 16(9):3419–3434.

Rebora, N., Ferraris, L., von Hardenberg, J., and Provenzale, A. (2006). Rainfall downscaling and flood forecasting: a case study in the Mediterranean area. *Natural Hazards and Earth System Sciences*, 6(4):611–619.

Ribalaygua, J., Torres, L., Pórtoles, J., Monjo, R., Gaitán, E., and Pino, M. R. (2013). Description and validation of a two-step analogue/regression downscaling method. *Theoretical and Applied Climatology*.

Ricard, D., Ducrocq, V., and Auger, L. (2012). A climatology of the mesoscale environment associated with heavily precipitating events over a northwestern Mediterranean area. *Journal of Applied Meteorology and Climatology*, 51:468–488.

Richardson, D. S. (2001). Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127:2473–2489.

Roberts, N. and Lean, H. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136:78–97.

Robertson, A. W., Moron, V., and Swarinoto, Y. (2009). Seasonal predictability of daily rainfall statistics over Indramayu district, Indonesia. *International Journal of Climatology*, 29:1449–1462.

Rust, H. W., Vrac, M., Lengaigne, M., and Sultan, B. (2010). Quantifying differences in circulation patterns based on probabilistic models: IPCC AR4 multimodel comparison for the North Atlantic. *Journal of Climate*, 23(24):6573–6589.

Sanchez-Gomez, E. and Terray, L. (2005). Large-scale atmospheric dynamics and local intense precipitation episodes. *Geophysical Research Letters*, 32(24):L24711.

Sauquet, E., Arma, Y., Ducharne, A., Hingray, B., Perrin, C., Hendrickx, F., Braud, I., Brun, J.-F., Vidal, J.-P., Datry, T., Mathevet, T., Thirel, G., Le Lay, M., Bouscasse, H., Cipriani, T., Magand, C., Samie, R., Rossi, A., Monteil, C., Blanc-Coutagne, E., Branger, F., Krowicki, F., Malerbe, F., Tilmant, F., Cherel, J., Poulhe, P., and Strosser, P. (2014). Projet R2D2 2050. risque, ressource en eau et gestion durable de la Durance en 2050. risk, water resources and sustainable development within the Durance river basin in 2050. Final project report, Irstea. in preparation.

Schmidli, J., Goodess, C. M., Frei, C., Haylock, M. R., Hundecha, Y., Ribalaygua, J., and Schmith, T. (2007). Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps. *Journal of Geophysical Research*, 112(D04105).

Schmith, T. (2008). Stationarity of regression relationships: Application to empirical downscaling. *Journal of Climate*, 21(17):4529–4537.

Schoof, J. T. (2013). Statistical downscaling in climatology. *Geography Compass*, 7(4):249–265.

Seibert, P., Frank, A., and Formayer, H. (2006). Synoptic and regional patterns of heavy precipitation in Austria. *Theoretical and Applied Climatology*, 87:139–153.

Souvignet, M. and Heinrich, J. (2011). Statistical downscaling in the arid central Andes: uncertainty analysis of multi-model simulated temperature and precipitation. *Theoretical and Applied Climatology*, 106:229–244.

Szczypta, C., Calvet, J.-C., Albergel, C., Balsamo, G., Boussetta, S., Carrer, D., Lafont, S., and Meurey, C. (2011). Verification of the new ECMWF ERA-Interim reanalysis over France. *Hydrology and Earth System Sciences*, 15(2):647–666.

Talagrand, O., Vautard, R., and Strauss, B. (1997). Evaluation of probabilistic prediction systems. In *ECMWF Workshop on Predictability*. ECMWF.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, 106(D7):7183–7192.

Teng, J., Chiew, F., Timbal, B., Wang, Y., Vaze, J., and Wang, B. (2012). Assessment of an analogue downscaling method for modelling climate change impacts on runoff. *Journal of Hydrology*, 472-473:111–125.

Termonia, P., Degrauwe, D., and Hamdi, R. (2011). Improving the temporal resolution problem by localized gridpoint nudging in regional weather and climate models. *Monthly Weather Review*, 139:1292–1304.

Teutschbein, C., Wetterhall, F., and Seibert, J. (2011). Evaluation of different downscaling techniques for hydrological climate-change impact studies at the catchment scale. *Climate Dynamics*, 37:2087–2105.

Teweles, S. and Wobus, H. B. (1954). Verification of prognostic charts. *Bulletin of the American Meteorological Society*, 35(10):455–463.

Themeßl, M. J., Gobiet, A., and Leuprecht, A. (2010). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, 31(10):1530–1544.

Tilmant, F., Cipriani, T., Branger, F., Sauquet, E., Braud, I., Leblois, E., and Gouttevin, I. (2013). Modélisation hydrologique distribuée du Rhône. Technical report, Irstea Lyon-Villeurbanne.

Timbal, B. (2004). Southwest Australia past and future rainfall trends. *Climate Research*, 26(3):233–249.

Timbal, B., Dufour, A., and McAvaney, B. (2003). An estimate of future climate change for western France using a statistical downscaling technique. *Climate Dynamics*, 20(7-8):807–823.

Timbal, B., Hope, P., and Charles, S. (2008). Evaluating the consistency between statistically downscaled and global dynamical model climate change projections. *Journal of Climate*, 21(22):6052–6059.

Timbal, B. and McAvaney, B. J. (2001). An analogue-based method to downscale surface air temperature: application for Australia. *Climate Dynamics*, 17(12):947–963.

Tödter, J. and Ahrens, B. (2012). Generalization of the ignorance score: Continuous ranked version and its decomposition. *Monthly Weather Review*, 140:2005–2017.

Torres, L., Ribalaygua, J., Gonzalez-Rouco, F., and Goodess, C. (2008). Recommendations for the modification of statistical downscaling methods for the construction of probabilistic projections. ENSEMBLES project deliverable D2B.14, University of East Anglia.

Toth, Z. (1991). Intercomparison of circulation similarity measures. *Monthly Weather Review*, 119:55–64.

Trigo, R. M. and Palutikof, J. P. (2001). Precipitation scenarios over Iberia: A comparison between direct GCM output and different downscaling techniques. *Journal of Climate*, 14(23):4422–4446.

Tryhorn, L. and DeGaetano, A. (2011). A comparison of techniques for downscaling extreme precipitation over the Northeastern United States. *International Journal of Climatology*, 31(13):1975–1989.

Turco, M., Quintana Seguí, P., Llasat, M. C., Herrera, S., and Gutiérrez, J. M. (2011). Testing MOS precipitation downscaling for ENSEMBLES regional climate models over Spain. *Journal of Geophysical Research*, 116:D18109.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. (2005). The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012.

Van den Dool, H. M. (1994). Searching for analogues, how long must we wait? *Tellus A*, 46:314–324.

van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., de Jeu, R. A. M., Liu, Y. Y., Podger, G. M., Timbal, B., and Viney, N. R. (2013). The millennium drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resources Research*, 49:1040–1057.

van Vliet, M. T. H., Blenkinsop, S., Burton, A., Harpham, C., Broers, H. P., and Fowler, H. J. (2012). A multi-model ensemble of downscaled spatial climate change scenarios for the Dommel catchment, Western Europe. *Climatic Change*, 111:249–277.

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., et al. (2011). The representative concentration pathways: an overview. *Climatic Change*, 109:5–31.

Vautard, R., Noël, T., Li, L., Vrac, M., Martin, E., Dandin, P., Cattiaux, J., and Joussaume, S. (2013). Climate variability and trends in downscaled high-

resolution simulations and projections over Metropolitan France. *Climate Dynamics*, 41(5-6):1419–1437.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M. (2010). A 50-year high-resolution atmospheric reanalysis over France with the Safran system. *International Journal of Climatology*, 30(11):1627–1644.

Vidal, J.-P. and Radanovics, S. (2014). SANDHY flies to Argentina. Invited talk at the second CORDEX-ESDM workshop.

Vidal, J.-P. and Wade, S. D. (2009). A multimodel assessment of future climatological droughts in the United Kingdom. *International Journal of Climatology*, 29(14):2056–2071.

Voisin, N., Schaake, J. C., and Lettenmaier, D. P. (2010). Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Weather and Forecasting*, 25:1603–1627.

Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L., and Somot, S. (2012). Dynamical and statistical downscaling of the French Mediterranean climate: uncertainty assessment. *Natural Hazards and Earth System Sciences*, 12(9):2769–2784.

Vrac, M. and Naveau, P. (2007). Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resources Research*, 43(7):W07402.

Vrac, M., Stein, M., and Hayhoe, K. (2007). Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate Research*, 34:169–184.

Vrac, M. and Yiou, P. (2010). Weather regimes designed for local precipitation modeling: Application to the Mediterranean basin. *Journal of Geophysical Research*, 115(D12):D12103.

Wernli, H., Hofmann, C., and Zimmer, M. (2009). Spatial forecast verification methods intercomparison project: Application of the SAL technique. *Weather and Forecasting*, 24:1472–1484.

Wernli, H., Paulat, M., Hagen, M., and Frei, C. (2008). SAL–a novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, 136(11):4470–4487.

Wetterhall, F., Bárdossy, A., Chen, D., Halldin, S., and Xu, C.-Y. (2006). Daily precipitation-downscaling techniques in three Chinese regions. *Water Resources Research*, 42:W11423.

Wetterhall, F., Halldin, S., and Xu, C.-y. (2005). Statistical precipitation downscaling in central Sweden with the analogue method. *Journal of Hydrology*, 306(1-4):174–190.

Wetterhall, F., Halldin, S., and Xu, C.-Y. (2007). Seasonality properties of four statistical-downscaling methods in central Sweden. *Theoretical and Applied Climatology*, 87(1-4):123–137.

Wilby, R. L. (1997). Non-stationarity in daily precipitation series: implications for GCM down-scaling using atmospheric circulation indices. *International Journal of Climatology*, 17:439–454.

Wilby, R. L. and Dawson, C. W. (2013). The Statistical DownScaling Model: insights from one decade of application. *International Journal of Climatology*, 33(7):1707–1719.

Wilby, R. L., Dawson, C. W., and Barrow, E. M. (2002). SDSM – a decision support tool for the assessment of regional climate change impacts. *Environmental Modelling and Software*, 17(2):145–157.

Wilby, R. L. and Dessai, S. (2010). Robust adaptation to climate change. *Weather*, 65(7):180–185.

Wilby, R. L., Hassan, H., and Hanaki, K. (1998a). Statistical downscaling of hydrometeorological variables using general circulation model output. *Journal of Hydrology*, 205(1-2):1–19.

Wilby, R. L., L., W. T. M., Conway, D., Jones, P. D., Hewiston, J. M., and Wilks, D. S. (1998b). Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, 34(11):2995–3008.

Wilby, R. L. and Wigley, T. M. L. (1997). Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21:530–548.

Wilby, R. L. and Wigley, T. M. L. (2000). Precipitation predictors for downscaling: observed and general circulation model relationships. *International Journal of Climatology*, 20:641–661.

Wilks, D. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210:178–191.

Wood, A. W., Leung, L. R., Sridhar, V., and Lettenmaier, D. P. (2004). Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, 62(1-3):189–216.

Yang, W., Bárdossy, A., and Caspary, H.-J. (2010). Downscaling daily precipitation time series using a combined circulation- and regression-based approach. *Theoretical and Applied Climatology*, 102:439–454.

Yiou, P. (2014). Anawege: a weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development*, 7(2):531–543.

Zappa, M., Beven, K. J., Bruen, M., Cofiño, A. S., Kok, K., Martin, E., Nurmi, P., Orfila, B., Roulin, E., Schröter, K., Seed, A., Szturc, J., Vehviläinen, B., Germann, U., and Rossa, A. (2010). Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 working group 2. *Atmospheric Science Letters*, 11:83–91.

Zepeda-Arce, J., Foufoula-Georgiou, E., and Droegemeier, K. K. (2000). Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *Journal of Geophysical Research*, 105:10129–10146.

Zhang, F. and Georgakakos, A. P. (2012). Joint variable spatial downscaling. *Climatic Change*, 111(3-4):945–972.

Zorita, E., Hughes, J. P., Lettemaier, D. P., and von Storch, H. (1995). Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *Journal of Climate*, 8(5):1023–1042.

Zorita, E. and von Storch, H. (1999). The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *Journal of Climate*, 12(8):2474–2489.

Abstract

Studying past and present day precipitation and its link to large scale circulation increases our understanding of precipitation characteristics and helps to anticipate their future behaviour. Downscaling techniques are being developed to bridge the gap between large-scale climate information from global reanalyses or GCM global projections and local meteorological information relevant for hydrology.

The stepwise analogue downscaling method for hydrology (SANDHY) is extended to the whole mainland of France by optimising the geopotential predictor domains for 608 zones covering France using a multiple growing rectangular domain algorithm that allows to take equifinality into account. A high diversity of predictor domains has been found. To increase the spatial coherence three ways are explored to reduce the parameter space: assessing the skill for predictor domains found for other zones, form groups of zones using cluster algorithms and using a less skewed predictand variable during optimisation. Using information from neighbouring zones allows to counterbalance in part limitations of the optimisation algorithm.

A feature based spatial verification method (SAL) is adapted for probabilistic precipitation simulation as provided by SANDHY. Skill scores derived from the probabilistic SAL are used to assess different strategies for spatially coherent precipitation downscaling at catchment scale. Locally optimised predictor domains lead to a better localisation of precipitation in the catchment and higher local skill while uniform predictor domains for the whole catchment lead to a more realistic spatial structure of the simulated precipitation. Streamflow simulations for the Durance catchment (Southern Alps) are most sensitive to the realistic localisation of precipitation which highlights the interest of locally optimising predictor domains.

Résumé

Étudier les précipitations et leur lien avec la circulation atmosphérique augmente notre connaissance de leurs caracteristiques et aide à anticiper leur comportement futur. Des méthodes de descente d'échelle statistiques sont développées pour fournir des informations météorologiques locales et importantes pour l'hydrologie à partir des informations issues des réanalyses ou des projections globales du climat.

La méthode SANDHY (Stepwise ANalogue Downscaling method for HYdrology) est étendue à l'ensemble de la France métropolitaine en optimisant les domaines pour le prédicteur géopotentiel pour les 608 zones climatiquement homogènes en France en utilisant un algorithme qui permet de prendre en compte l'équifinalité. Une grande diversité des domaines pour le prédicteur géopotentiel a été trouvée. Trois voies pour augmenter la cohérence spatiale et diminuer l'espace des paramètres sont explorés : prendre en compte les domaines optimisés pour des zones voisines, rassembler des zones en utilisant des algorithmes d'aggregation et utiliser un predictant moins asymétrique pendant l'optimisation. Utiliser de l'information issues de zones voisines permet de compenser certaines limitations de l'algorithme d'optimisation.

Une méthode de vérification spatiale (SAL) est ici adaptée pour les précipitations probabilistes simulées par SANDHY. Des mesures de performance derivées de cette version probabiliste du SAL sont ensuite utilisées pour évaluer différentes stratégies de descente d'échelle concernant la cohérence spatiale à l'échelle d'un bassin versant. Les domaines optimisés localement pour le prédicteur géopotentiel permettent de mieux localiser les précipitations dans le bassin tandis que des domains uniformes sur tout le bassin apportent une structure des précipitations plus réaliste. Les simulations de débit pour le bassin de la Durance sont le plus sensible à la localisation des précipitations, ce qui souligne l'interêt d'une optimisation locale des domaines des prédicteurs.