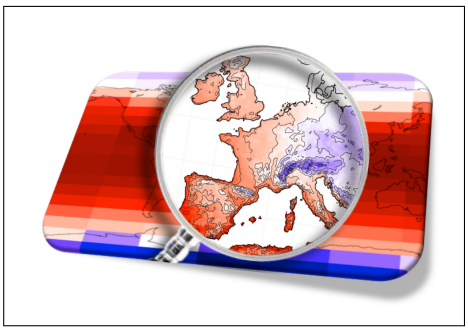
Mathieu Vrac

Modélisations statistiques à différentes échelles climatiques et environnementales

Habilitation à Diriger des Recherches



Soutenue le 30 mars 2012 devant le jury :

Dr. Denis Allard (INRA) : Rapporteur
Dr. Christophe Bouvier (HSM/IRD) : Rapporteur
Dr. Serge Planton (Météo-France) : Rapporteur
Prof. Philippe Bousquet (LSCE/UVSQ) : Président du Jury
Dr. Petra Friederichs (Univ. Bonn) : Examinatrice
Dr. Joël Guiot (CEREGE/CNRS) : Examinateur

À toi,
À moi,
À nous,
À rien.

PRÉFACE

Après un parcours universitaire en mathématiques appliquées et statistiques, la thèse à été pour moi l'occasion de découvrir (un peu par hasard) les apports et applications des statistiques dans les sciences du climat. Dès le début de cette thèse interdisciplinaire, je me suis pris au jeu d'essayer de regrouper ces deux disciplines. Les multiples dialogues et allers-retours entre climatologues et statisticiens – avec parfois la nécessité pressante d'être l'interface de traduction entre les deux – m'ont ouvert des horizons nouveaux que je n'ai pas cessés de scruter avec envie et curiosité depuis. Les recherches en statistiques et en climat m'ont permis d'apprendre quantité de concepts, méthodologies, théories, etc., que ce soit dans l'un ou l'autre de ces domaines. Les développements théoriques et pratiques en statistiques ont parfois permis de poser de nouvelles questions climatologiques. Inversement, les questionnements climatiques motivent bien souvent des recherches originales en statistiques. Si j'ai parfois longuement hésité entre une "carrière" en statistiques et une en climatologie (la catégorisation des personnes et de leur thématique est malgré tout encore un mal bien Français), cette thématique de *climatologie statistique*, au point de rencontre des deux, s'est progressivement imposée à moi comme une évidence. J'ai alors eu l'énorme chance de rencontrer des personnes qui croyaient en cette discipline hybride et, même mieux, ont cru en moi et m'ont donné la chance d'y participer. Je pense à mes directeurs de thèse (A. Chédin et E. Diday) mais aussi et surtout à P. Naveau, P. Yiou, M. Stein, D. Paillard et G. Ramstein, qui m'ont tous à un moment ou à un autre, permis de vivre de mes recherches.

Ce manuscrit est donc l'occasion pour moi de faire un petit bilan de mes principaux travaux dans ce vaste domaine. J'ai souhaité que cette HDR soit la plus représentative possible des quelques réflexions et collaborations que j'ai pu avoir ces dernières années. J'ai donc essayé de présenter à la fois la philosophie de mes travaux (souvent discutée en introduction des différents chapitres) mais également tous les éléments nécessaires à la compréhension technique de mes apports. Le juste milieu est certainement complexe à trouver mais j'espère que ce manuscrit offrira une double lecture, permettant à différents lecteurs avec différents intérêts d'y trouver leur compte sans ennui ni frustration.

Évidemment, j'ai du faire des choix dans la présentation de mes travaux et je prie par avance les collaborateurs, post-docs et étudiants, dont je n'ai pas cité les apports, de bien vouloir m'excuser de ces absences. Qu'ils sachent bien ici toutes les richesses scientifiques qu'ils ont su m'apporter et tout le plaisir que j'ai eu et ai parfois encore à les côtoyer.

Mathieu Vrac

Gif-sur-Yvette

Octobre 2011

TABLE DES MATIÈRES

CV

Travailler pour gagner sa vie, O.K.
Mais pourquoi faut-il que cette vie qu'on gagne, il faille la gaspiller à travailler pour gagner sa vie ?
Quino, Extrait de la bande dessinée Mafalda.

PARCOURS SCIENTIFIQUE

- Depuis 2008 : Chercheur CNRS permanent (CR2) au Laboratoire des Sciences du Climat et de l'Environnement, LSCE-IPSL, CNRS/CEA/UVSQ, Gif-sur-Yvette, France
- 2006 – 2008 : Chercheur CNRS CDD au LSCE
- 2004 – 2006 : Research Associate au "Center for Integrating Statistical and Environmental Science" (CISES) à l'Université de Chicago, USA
- Juillet-Août 2005 : Scientifique invité au "National Center for Atmospheric Research" (NCAR), Boulder, Colorado, USA
- 2002 – 2004 : Attaché Temporaire à l'Enseignement et à la Recherche (ATER) en Mathématiques Appliquées et Informatique, Université Paris Dauphine
- 1999 – 2002 : Doctorat en Mathématiques Appliquées, en collaboration entre le Laboratoire de Météorologie Dynamique (LMD-CNRS/IPSL, Ecole Polytechnique) et le Centre de Recherche de Mathématiques de la Décision (CEREMADE - Université Paris IX Dauphine) sous la direction d'Alain Chédin et d'Edwin Diday.

PROJETS ET RESPONSABILITÉS ACADÉMIQUES

Projets

- Co-coordinateur du projet international (Allemagne - France - Angleterre) "PLEIADES" ("Projections and predictions of Local prEcipitation Intensities. Advanced Downscaling using Extreme value Statistics") financé par la fondation Volkswagen (2011-2014) .
- Représentant Français au management commitee de l'Action Européenne COST "VALUE" ("Validating and Integrating Downscaling Methods for Climate Change Research", 2011 - 2015)
- Responsables des tâches "Predictions and their uncertainties" (WP 3.3) et "Upscaling-downscaling" (WP 4.2) dans l'ANR "McSim" ("Simulation conditionnelle multisupport de processus max-stables. Applications à la prédiction locale d'évènements extrêmes climatiques", 2011-2014)
- Participant au projet FP7 IMPACT2C ("Quantifying projected impacts under 2°C warming", 2011-2015)
- Participant au projet ANR CHEDAR ("Climate, Health and Environment : DAta Rescue and modelling", 2010-2013)
- Expert scientifique dans le projet PreviBoss ("Prévisibilité à courte échéance de la variabilité dans le cycle de vie du brouillard à partir de données d'observations sol et satellite", 2010-2013)
- Participant dans le projet GIS PEPER ("Plans d'Expérience appliqués à la Prévision des Extrêmes climatiques Régionaux", 2010-2013)
- Responsable du package "Liens grande-petite échelle" (WP1) du projet GIS REGYNA (2008-2010)
- Participant au projet ANR MedUP ("Forecast and projection in climate scenario of Mediterranean intense events : Uncertainties and Propagation on environment", 2008-2011)

– Expert scientifique dans l'ANR SHIVA (Socio-economic Assessment of the rural Vulnerability of water users under stressors of global change in the Hard Rock area of South India, 2008-2010)

Diverses responsabilités académiques

– Membre du conseil scientifique du pôle "Climat et Environnements Régionaux" (CER) de l'IPSL et responsable du groupe de modélisation statistique de ce pôle
– Organisateur de la session "Statistical downscaling" à la conférence internationale "Water 2010", Québec, Canada, Juillet 2010
– Examinateur dans le jury de thèse de J.C. Espinoza (13 janvier 2009)
– Animateur du groupe de lecture de l'IPSL sur le downscaling (régionalisation) statistique
– Reviewer pour les journaux : "Climate Dynamics", "Climate of the Past", "Climate Research", "Computers and Geosciences", "Data and Knowledge Engineering", "Extremes", "Journal of Applied Meteoroly and Climatology", "Journal of Climate", "Journal of Geophysical Research", "Journal of Hydrology", "Natural Hazards and Earth System Sciences", "Quarterly Journal of the Royal Meteorological Society", "Water Resources Research".

ENSEIGNEMENTS

– Années 2009/2010, 2010/2011 et 2011/2012 (Dernière année Ecole Centrale Marseille) : Responsable du module et chargé des cours et TDs de "Modélisation en Climatologie statistique et extrêmes environnementaux"

– Année 2003/2004 (Université Paris IX Dauphine) :
  • Responsable du module de calcul scientifique en DEUG MASS 1,
  • Responsable du module et des cours en amphithéâtre de Combinatoire, et des TD en IUP1 Miage,
  • Responsable du module et des cours en amphithéâtre de JAVA en DEUG MASS 1,
  • Chargé de TD de Statistiques (DEUG MASS 2),
  • Chargé de TD en "Management Scientifique Aide à la Décision" (Théorie des Graphes, Aide à la Décision, Prog. Linéaire) en Maîtrise de Science de Gestion.

– Année 2002/2003 (Université Paris IX Dauphine) :
  • Chargé de TD/TP d'algorithmique (langage JAVA) en DEUG MASS 1ère année,
  • Chargé de TD/TP de JAVA en Licence IUP Génie Mathématiques et Informatique,

– Année 2000/2001 (Université Paris IX Dauphine) : Chargé de TD/TP d'algorithmique et langage JAVA (DEUG MASS) avec Fabrice Rossi, en tant que vacataire.

ENCADREMENTS SCIENTIFIQUES

TABLE 0.1: Étudiants et postdocs que j'ai (co-)encadrés, ainsi que les principaux sujets et applications abordés.

Postdocs	**Downscaling**	**Régimes de temps**	**Extrêmes**	**Applications ou autres**
Thomas Noël (postdoc LSCE/LMD), 2011	Correction de biais			Prec., temp. etc. France (DRIAS)
Maalak Kallache (postdoc LSCE/Climpact), 2009-2010	MNH		XCDF-t	Precipitations en France
Tamara Salameh (postdoc LSCE), 2009-2011	GAM	Attrib./comparaison		Vents en Méditerranée
Nicolas Vigaud (postdoc BRGM), 2009-2010	CDF-t			Temp. & prec. en Inde
Armel Martin (postdoc LSCE), 2009-2010	GAM			Temp. & prec. en paléo
Pascal Oettli (postdoc LOCEAN), 2009-2010	CDF-t			Rendements en Afrique de l'ouest
Christophe Lavaysse (postdoc LMD), 2008-2009	CDF-t & RCMs			Temp., prec. & vent en Méditerranée
Henning Rust (postdoc LSCE), 2008-2009	GLM & MNH	EM & Eval. statistique		Régimes. en Europe Precip. Afr. de l'ouest
Katerina Goubanova (postdoc LMD), 2008	Fonctions de transfert			Vent Am. du sud
Thésards				
Aurélien Bechler (thésard LSCE-AgroPT) 2011-2014	Modèles spatiaux		Maxima	Précip. France
Guillaume Levavasseur (thésard LSCE), 2009-2012	GAM & GLM			permafrost & véget. en Europe
Clément Tisseuil (thésard U.P.S. Toulouse), 2007-2009	ABT, GLM, GAM & ANN	HAC		débits & biodiversité Sud de la France
Medhi Limam (thésard Dauphine, 2002-2004)		Classification hybride		Assurances
Étudiants de Master 1 & 2				
Pradeebane Vaittinada Ayar (Étudiant Master 2), 2010		EM		Régimes saisonniers en Europe
Guillaume Levavasseur (Étudiant Master2, 2009)	GAM			permafrost & véget. en Europe
Mohamed Azlim (Étudiant Master 2), 2004		Copules		Théorie
Medhi Limam (Étudiant Master2, 2002)		Classification hybride		Assurances
Elsa Bernard (Étudiante Master 1), 2011		PAM	F-Madogram	Precip. max France
Anthony Merlo (Étudiant Master 1), 2010	CDF-t & RCMs			Temp., prec. & vent sud de la France
Florian Hechner (Étudiant Magistère 2, 2003)		EM-copules		Simulations numériques

Abréviations : MNH = modèles non-homogènes ; CDF-t = Cumulative Distribution Function - transform ; XCDF-t = eXtreme CDF-t ; GAM = Generalized Additive Model ; GLM = Generalized Linear Model ; RCM = Regional Climate Model ; ABT = Aggregated Boosted Trees ; ANN = Artificial Neural Network ; HAC = Hierarchical Ascending Clustering ; PAM = Partitioning Around Medoids ; EM = Expectation-Maximization.

PUBLICATIONS À COMITÉ DE LECTURE ("PEER-REVIEWED")

La plupart de mes articles publiés sont téléchargeables sur mon site web : http ://www.lsce.ipsl.fr/Pisp/58/mathieu.vrac.html

Articles soumis ou en révision

- A. Colette, R. Vautard, M. Vrac : Regional climate downscaling with prior statistical correction of the global climate forcing. Submitted.
- G. Levavasseur, M. Vrac, D. Roche, D. Paillard, J. Guiot : An objective methodology for potential vegetation reconstruction constrained by climate. Submitted.
- M. Vrac, P. Drobinski, A. Merlo, M. Herrmann, C. Lavaysse, L. Li, S. Somot. 'Dynamical and statistical downscaling of the French Mediterranean climate : uncertainty assessment'. Submitted.
- F. C. Sperna Weiland, C. Tisseuil, H. H. Dürr, M. Vrac, and L. P. H. van Beek. 'Selecting the optimal method to calculate daily global reference potential evaporation from CFSR reanalysis data'. Hydrol. Earth Syst. Sci. Discuss., 8, 7355-7398, 2011, doi :10.5194/hessd-8-7355-2011.
- M. Vrac, P. Yiou and P. Vaittinada Ayar. 'Trends and variability of seasonal weather regimes'. Submitted.
- M. Troin, M. Vrac , M. Khodri, C. Vallet-Coulomb, E. Piovano, F. Sylvestre. 'Coupling statistically downscaled GCM outputs with a basin-lake hydrological model in subtropical South America : evaluation of the influence of large-scale precipitation changes on regional hydroclimate variability'. Hydrol. Earth Syst. Sci. Discuss., 7, 1–44, 2010, doi :10.5194/hessd-7-1-2010.
- N. Vigaud, M. Vrac, Y. Caballero. 'Probabilistic Downscaling of GCMs scenarios over southern India'. Submitted.

2012

35. C. Tisseuil, M. Vrac, G. Grenouillet, M. Gevrey, T. Oberdorff, A.J. Wade, S. Lek (2012) : Strengthening the link between hydro-climatic downscaling and species distribution modelling : Climate change impacts on freshwater biodiversity, accepted by Science of the Total Environment (STOTEN).
34. 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, Nat. Hazards Earth Syst. Sci., 12, 651-670, doi :10.5194/nhess-12-651-2012.
33. C. Tisseuil, F. Leprieur, G. Grenouillet, M. Vrac, S. Lek (2012). 'Projected impacts of climate change on spatio-temporal patterns of freshwater fish beta diversity : a deconstructing approach'. Accepted by Global Ecology and Biogeography.

2011

32. G. Levavasseur, M. Vrac, D. Roche, D. Paillard, A. Martin, J. Vandenberghe. 'Present and LGM permafrost from climate simulations : contribution of statistical downscaling'. Accepted in Climate of the Past.
31. J. Carreau & M. Vrac (2011) 'Stochastic downscaling of precipitation with neural network conditional mixture models'. Water Resources Research, 47, W10502, doi :10.1029/2010WR010128.

30. I. Bentaleb, C. Martin, M. Vrac, B. Mate, P. Mayzaud, R. de Stephanis, C. Guinet (2011). 'Foraging ecology of Mediterranean fin whales in a changing environment elucidated by satellite tracking and stable isotopes'. Marine Ecology Progress Series, Vol. 438 : 285-302, doi : 10.3354/meps09269.
29. M. Vrac, L. Billard, E. Diday, A. Chédin (2011). "Copula Analysis of Mixture Models". Accepted in Computational Statistics (in press).
28. M. Ghil, P. Yiou, S. Hallegatte, B. D. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, O. Mestre, C. Nicolis, H. Rust, P. Shebalin, M. Vrac, A. Witt, and I. Zaliapin (2011). "Extreme Events : Dynamics, Statistics and Prediction". Nonlinear Processes in Geophysics, 18 : 295–350, doi : 10.5194/npg-18-295- 2011.
27. P. Willems, M. Vrac (2011). Statistical precipitation downscaling for small-scale hydrological impact investigations of climate change. Journal of Hydrology, 402, 193-205, doi :10.1016/j.jhydrol.2011.02.030.
26. P. Oettli, B. Sultan, C. Baron, M. Vrac (2011). Are regional climate models relevant for crop yield prediction in West Africa ? Environ. Res. Lett., 6 (2011) 014008.
25. M. Kallache, M. Vrac, P. Naveau, P.-A. Michelangeli (2011). 'Non-stationary probabilistic downscaling of extreme precipitation'. J. Geophys. Res. - Atmospheres, VOL. 116, D05113, doi :10.1029/2010JD014892.

2010

24. H. Rust, M. Vrac, M. Lengaigne, B. Sultan (2010). Quantifying differences in circulation patterns based on probabilistic models : IPCC-AR4 multi-model comparison for the North Atlantic. Journal of Climate. Vol. 23, pp. 6573-6589.
23. P. Yiou, E. Bard, P. Dandin, B. Legras, P. Naveau, H. Rust, L. Terray, and M. Vrac (in alphabetical order after first author) (2010). Statistical issues about solar-climate relations. Climate of the Past, 6, 565-573, doi :10.5194/cp-6-565-2010
22. K. Goubanova, V. Echevin, B. Dewitte, F. Codron, K. Takahashi, P. Terray, M. Vrac (2010). 'Statistical downscaling of sea-surface wind over the Peru-Chile upwelling region : diagnosing the impact of climate change from the IPSL-CM4 model'. Climate Dynamics, 36 (7-8), 1365-1378, doi : 10.1007/s00382-010-0824-0
21. D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themeßl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, I. Thiele-Eich (2010). 'Precipitation downscaling under climate change. Recent developments to bridge the gap between dynamical models and the end user'. Reviews of Geophysics, 48, RG3003, doi :10.1029/2009RG000314
20. C. Tisseuil, M. Vrac, S. Lek, A.J. Wade (2010). Direct statistical downscaling of river flows. Journal of Hydrology. Vol. 385, Issues 1-4, pp. 279-291.
19. M. Vrac, P. Yiou (2010). Weather regimes designed for local precipitation modelling : Application to the Mediterranean basin. Journal of Geophysical Research – Atmosphere. Vol. 115, D12103, doi :10.1029/2009JD012871

2009

18. P.-A. Michelangeli, M. Vrac, H. Loukos (2009). Probabilistic downscaling approaches : Application to wind cumulative distribution function. Geophysical Research Letters, 36, L11708, doi :10.1029/2009GL038401.
17. V. Jomelli, D. Brunstein, M. Déqué, M. Vrac, D. Grancher (2009). Impacts of future climatic change (2000-2100) on the occurrence of debris flows : A case study in the Massif des Ecrins (French Alps). Climatic Change, Vol.97, numbers 1-2, pp. 171-191

16. G. Toulemonde, A. Guillou, P. Naveau, M. Vrac, F. Chevallier (2009). Autoregressive models for maxima and their applications to CH4 and N2O. Environmetrics, 21, 188-207.
15. T. Salameh, P. Drobinski, M. Vrac, P. Naveau (2009). Statistical downscaling of near surface wind field over complex terrain in southern France. Meteorology and Atmospheric Physics, 103, 253-265, doi : 10.1007/s00703-008-0330-7.

2008

14. A. Bernacchia, P. Naveau, M. Vrac, P. Yiou (2008), Detecting spatial patterns with the cumulant function. Part II : An application to El Nino, Nonlinear Processes in Geophysics, 15, 169-177.

2007

13. M. Vrac, P. Marbaix, D. Paillard, P. Naveau (2007), Non-linear statistical downscaling of present and LGM precipitation and temperatures over Europe, Climate of the Past, 3, 669–682.
12. M. Vrac, P. Naveau, P. Drobinski (2007), Modeling pairwise dependencies in precipitation intensities, Nonlinear Processes in Geophysics, 14, 789–797.
11. M. Vrac, M.L. Stein, K. Hayhoe, X.-Z. Liang (2007), A general method for validating statistical downscaling methods under future climate change, Geophysical Research Letters, 34, L18701, doi :10.1029/2007GL030295.
10. M. Vrac, and P. Naveau (2007) Stochastic downscaling of precipitation : From dry events to heavy rainfalls, Water Resources Research, 43, W07402, doi :10.1029/2006WR005308.
9. M. Vrac, K. Hayhoe, M. Stein (2007), Identification and inter-model comparison of seasonal circulation patterns over North America, International Journal of Climatology, 27 (5), 603-620, DOI : 10.1002/joc.1422.
8. M. Vrac, M. Stein, K. Hayhoe (2007), Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing, Climate Research, 34 : 169-184, doi : 10.3354/cr00696.

2006

Euh, bah non rien cette année là...

2005

7. M. Vrac, A. Chédin, E. Diday (2005), Clustering a Global Field of Atmospheric Profiles by Mixture Decomposition of Copulas, Journal of Atmospheric and Oceanographic Technology, 22 (10), 1445–1459.
6. E. Diday, M. Vrac (2005), Mixture decomposition of distributions by copulas in the symbolic data analysis framework, Discrete Applied Mathematics, 147, 27-41

2004

5. M. Vrac, E. Diday, A. Chédin (2004), Décomposition de mélange de distributions et application à des données climatiques, Revue de Statistique Appliquée, LII (1), 67-96.
4. P. Naveau, M. Vrac, M. G. Genton, A. Chédin and E. Diday (2004), Two statistical methods for improving the analysis of large climatic data sets : General skewed Kalman filters and distributions of distributions, Chapitre dans Vol. 13 of Quantitative Geology and Geostatistics : Geostatistics for Environmental Applications, pages 1-14, Kluwer Academic Publishers (Sanchez-Vila, Carrera and Gomez-Hernandez, eds).

2003

Celle là non plus apparemment...

2002

3. M. Vrac, E. Diday, M. M. Liman, S. Winsberg (2002), A top down binary tree method for symbolic class descriptions. Chapitre dans Information Processing and Management of Uncertainty, pages 99-109, Elsevier Science (Bouchon-Meunier, Foulloy and Yager, eds)
2. M. Vrac, E. Diday, A. Chédin (2002b), Mixture decomposition of copulas in climatology. Chapitre dans Information Processing and Management of Uncertainty, pages 36-51, Elsevier Science (Bouchon-Meunier, Foulloy and Yager, eds).
1. M. Vrac, E. Diday, A. Chédin (2002a), Mélange de lois de lois par copules, Chapitre dans Extraction et Gestion de connaissances, pages 105-117, Hermes Sciences (H. Briand, ed).

1 Introduction : Différentes échelles - Différents modèles

Bienvenue à toi, lent homme lié, poussif tresseur des vitesses.
Alain Damio, La horde du contrevent

Mes principales recherches concernent l'étude et la modélisation statistique de la variabilité du climat à différentes échelles spatiales et temporelles. En effet, ma formation en mathématiques appliquées m'a donné le goût de la modélisation statistique que j'ai cherché à partager au sein de la communauté des sciences liées au climat. À mes débuts de thèse, cette communauté, évidemment très orientée vers la physique et les modèles déterministes, n'avait que peu de culture statistique et ne voyait que rarement ce qu'une "modélisation statistique" signifiait ou pouvait apporter. La communauté "climatologie statistique" s'est depuis considérablement élargie et représente désormais une force indéniable pour la compréhension des variabilités climatiques et environnementales. Mes travaux s'insèrent donc dans cette mouvance qui cherche à établir un dialogue entre ces deux disciplines que sont les statistiques et les sciences du climat. Cette discussion a évidemment nécessité des efforts des deux côtés, ne serait-ce que pour parler un langage commun et compréhensible de part et d'autre. Je dois admettre que cet apprentissage des rudiments du climat, de sa problématique, de sa complexité, de ses grands questionnements, etc. a été pour moi particulièrement riche et passionnant. J'ose également espérer avoir, parfois, lors de séminaires ou de discussion plus ou moins informelles, montré en partie les apports originaux des statistiques en climat et environnement, et, pourquoi pas, donné envie à des climatologues de comprendre - si ce n'est même d'employer - certaines approches et modélisations statistiques.

Dans ce contexte que j'ai fortement souhaité multidisciplinaire, mes recherches se sont principalement concentrées sur trois axes inter-dépendants :

1. la caractérisation de **régimes de temps**,
2. le développement d'**approches statistique de régionalisation** (ou "downscaling"),
3. et la modélisation statistique d'**événements extrêmes**.

Chacune de ces trois thématiques fait l'objet d'un chapitre de ce manuscrit. Ces trois axes de recherche ont la particularité d'être associés à *différentes échelles spatiales* et d'être souvent utiles pour *différentes échelles temporelles*. Les régimes de temps et leurs propriétés fournissent des informations dites "à grande échelle" spatiale en caractérisant des structures atmosphériques de plusieurs centaines de km. Le downscaling statistique permet de simuler des phénomènes climatiques ou météorologiques à des échelles très "petites" (c-à-d., très locales, par ex., au niveau de stations météo) en les contraignant par diverses informations à grande échelle. Les événements

extrêmes, eux, peuvent à la fois être considérés à de grandes échelles spatiales - par exemple dans le cadre de vagues de chaleur ou de sécheresses - et à des échelles beaucoup plus locales - par exemple, les précipitations extrêmement intenses pouvant générer des crues dites “éclair”, doivent être modélisées à des résolutions assez fines (par ex., au niveau du bassin versant) pour être pertinentes.

Ces trois thématiques ont également des horizons temporels variés. Les régimes de temps, souvent définis journalièrement et possédant des persistances de quelques jours, peuvent être employés pour caractériser les principaux modes de variabilité du climat “présent” mais aussi pour évaluer leurs évolutions potentielles dans le futur ou depuis un climat passé plus ou moins lointain (par ex., dernier millénaire ou période du dernier maximum glaciaire). De même, la régionalisation ou modélisation statistique à haute résolution permet des études de processus continentaux et atmosphériques en climat présent mais également de réaliser des projections locales de variables climatiques [1] nécessaires aux études et modèles d’impacts (écologiques, hydrologiques, économiques, etc.) du changement climatique futur, ou par exemple, de faciliter la comparaison modèles-données dans un contexte d’études paléo-climatiques. Enfin, si les études sur les événements extrêmes sont pertinentes en soi pour mieux appréhender notre vulnérabilité au climat, celles-ci sont désormais souvent employées pour définir des cartes de risques (par ex., carte de niveau de retour pour un phénomène centennal, ou millenal), non-seulement à l’actuel mais aussi et surtout en contexte de changement climatique, par exemple pour la construction de digues maritimes ou plus généralement pour la construction d’ouvrages de protection contre les événements climatiques extrêmes. Cependant, les définitions de niveau de retour associé à une probabilité donnée (généralement faible) ou de temps de retour associé à un quantile donné (généralement élevé) reposent explicitement sur des phénomènes dont la distribution statistique est stationnaire dans le temps [2]. Ceci n’est évidemment plus vrai en contexte de changement climatique, ce qui a récemment motivé le développement de modèles *non-homogènes* faisant évoluer les distributions et leurs propriétés selon le contexte climatique ou temporel. Ces modèles font partie intégrante de mes recherches et seront discutés en sections 3.2, 4.2 et 4.3.

De manière globale, la compréhension du “risque” (climatique, environnemental) et des incertitudes associées passent par des concepts statistiques et des estimations de probabilités de divers événements. Le rôle de la modélisation statistique est donc centrale. Mon travail permet de faire un lien naturel et opérationnel entre climatologie et d’autres domaines influencés/impactés par le climat en développant des concepts et outils statistiques qui sont mis à la disposition de l’ensemble de la communauté climatique et des impacts.

Mes travaux concernant les régimes de temps sont présentés au chapitre 2 et s’appuient fortement sur le concept de *mélange de distributions*, permettant d’associer une incertitude à chaque régime, sous la forme de probabilités d’occurrence conditionnelle. Je présenterai tout d’abord en section 2.2 mes premiers résultats dans ce contexte en rappelant les principaux apports issus de ma

1. en s’appuyant évidemment sur des sorties de modèles de climat globaux
2. c-à-d., les phénomènes suivent la même loi statistique qui n’évolue pas dans de temps

thèse ([193, 192]) et permettant de définir des régimes (ou, plus précisément des types de masse d'air) à partir d'informations dites probabilistes correspondant à des données de type fonctions de répartition cumulée. Dans cette première illustration, il s'agissait de réaliser une classification non-supervisée (clustering) spatiale, au sens où l'on cherchait à grouper/séparer des lieux géographiques. Dans les illustrations suivantes de ce chapitre, le but est davantage un clustering "temporel", au sens où l'on cherchait à grouper/séparer des jours caractérisés chacun par un ou plusieurs champs de variables atmosphériques. Je présenterai, en section 2.3, comment la notion de mélange de distributions permet d'évaluer et surtout de quantifier la capacité des modèles de climat à reproduire les principaux modes de variabilité réelle, au travers d'une étude menée récemment ([156]). Je détaillerai également en section 2.4 comment la notion de saisons peut être revisitée grâce à des régimes de temps. Ce travail ([202]) nous a permis de montrer des évolutions significatives - en fréquences mais aussi en dates de début et de fin - de certaines saisons depuis le milieu des années 70, ainsi que d'identifier des sources des tendances en température. Enfin, je discuterai en section 2.5 comment des informations locales, issues par exemple de stations météorologiques, peuvent être insérées dans un processus de définition de régimes de temps par clustering portant sur des données à grande échelle, afin de disposer de régimes de temps fortement corrélés à des données locales. Différentes approches ayant été testées, une intercomparaison a été effectuée sur la base de critères originaux développés dans [201], fournissant ainsi un guide dans l'utilisation des principales méthodologies de classification supervisée et non-supervisée.

Le chapitre 3 présente un rappel, que j'ai souhaité être détaillé, de mes principaux apports en modélisation statistique pour la régionalisation climatique et l'environnement. La régionalisation climatique cherche à modéliser et à simuler des phénomènes climatiques et météorologiques à des échelles spatiales fines que ne peuvent pas atteindre les modèles globaux de climat. Ces simulations fines sont en effet nécessaires par exemple pour mener à bien des études d'impacts liés aux changements climatiques. Après une courte introduction aux principaux enjeux et une schématisation des principales grandes familles de downscaling statistique, je présenterai en section 3.2 certains avantages de l'approche "stochastique" en l'illustrant par mes travaux sur des densités *non-homogènes* (ou *conditionnelles*) de précipitation ([199, 27]). Je m'attarderai ensuite en section 3.3 sur les modèles de type "fonction de transfert" que j'ai pu aborder pour la régionalisation ou, plus généralement, la modélisation environnementale à haute résolution, et ceci dans divers contextes d'applications - tels que l'écologie ([179]), le climat et l'environnement passé lointain ([196]) ou le pergélisol ([114]) - que je brosserai brièvement. Une approche de régionalisation dite "Model Output Statistics" (MOS), permettant de générer des *distributions* locales de variables climatiques (et non de simples réalisations) à partir de *distributions* à grande échelle ([125]), sera ensuite présentée en section 3.4. Nous verrons que, de par sa simplicité et son efficacité, ce modèle original a eu un certain succès ces dernières années au travers de multiples applications ([131, 178, 189, 182]) que j'évoquerai en fin de chapitre 3.

Le volet concernant les événements extrêmes sera présenté au chapitre 4. Celui-ci sera essentiellement orienté vers la modélisation de ces événements dans un processus de régionalisation. Après un bref rappel des principales notions liées à la théorie des valeurs extrêmes, je préciserai en section 4.2 comment l'approche MOS détaillée précédemment en section 3.4 a été étendue au cas des distributions associées à des excès [3] ([98]) afin de tenir compte des spécificités de ces données en se basant sur un modèle asymptotique. Nous verrons également que, sous certaines conditions, des covariables [4] peuvent être introduites dans un tel modèle, permettant alors d'insérer davantage d'information pour guider la régionalisation des CDFs. Je détaillerai ensuite en section 4.3 deux modèles statistiques ([197, 27]) – basés sur une philosophie commune de mélange mais avec des techniques différentes – permettant de représenter, non-seulement les événements extrêmes mais aussi les valeurs les plus courantes (faibles et moyennes), fournissant ainsi deux modélisations "complètes" de la distribution de variables climatiques, ici les précipitations.

Enfin, les perspectives générales que je souhaite donner à mes travaux seront données au chapitre 5.

Pour terminer cette introduction, je souhaite mentionner que ce manuscrit a été rédigé avec l'objectif qu'il soit un document de travail pour toute personne (étudiant, thèsard, postdoc, chercheur) curieux de découvrir la climatologie statistique. Ce document n'a évidemment pas la prétention de couvrir la totalité de ce domaine mais j'ai souhaité qu'il puisse être employé comme référence la plus complète possible (au moins de mes travaux). La présentation de type "livre" est donc volontaire même si ce manuscrit n'a pas vocation à être publié mais bien distribué.

Enfin, tous ces travaux et apports scientifiques ont évidemment été menés et rendus possibles grâce aux multiples collaborations et interactions que j'ai pu avoir avec de nombreux collègues, étudiants et postdocs (Français ou étrangers), à qui je vais tacher de faire honneur dans les chapitres qui suivent.

3. c-à-d., aux valeurs dépassant un seuil fixé
4. prédicteurs (atmosphériques, etc.) autres que la variable grande échelle à régionaliser

2 Régimes de temps : un outil de modélisation et d'évaluation à grande échelle

Si les nuances infinies du langage ne s'accommodent point des classifications rigides qu'on veut faire, tant pis pour les classifications. La science doit s'accommoder à la nature. La nature ne peut s'accommoder à la science.
Ferdinand Brunot, La Pensée et la Langue (1922)

2.1 Introduction

Le choix d'entamer ce manuscrit par mes travaux sur les régimes de temps est essentiellement historique. En effet, ceux-ci ont constitué ma "porte d'entrée" vers les sciences du climat alors que j'étais jeune diplômé d'un DEA de mathématiques appliquées. Bien que je pense avoir depuis considérablement élargi ma vision et mes apports en climat et environnement, les études et modélisations sur les régimes de temps climatiques gardent pour moi ce goût de découverte initiale et constituent encore aujourd'hui une thématique dont je suis toujours curieux.

2.1.1 Quelques rappels sur le clustering *climatique*

Les régimes de temps (RTs, ou "weather regimes" en anglais), également nommés "types de temps" ou "régimes de circulation atmosphérique", sont une approche relativement simple pour caractériser les principaux modes de variabilité de l'atmosphère pour une région donnée du globe (par ex., [201]). Un régime de temps peut être défini comme une structure atmosphérique à grande échelle (généralement avec un rayon de plusieurs centaines de km) revenant de manière récurrente au-dessus de la région d'intérêt. Ces structures récurrentes sont habituellement décrites en termes de variables de circulation (hauteurs de géopotentiel, pression, etc.). Il est généralement supposé que les spécificités spatiales d'un RT induisent des conditions météorologiques "locales" récurrentes lors de l'apparition du régime [1]. Plusieurs méthodes peuvent être employées pour définir de tels régimes. Ces méthodes sont divisées en approches dites "subjectives" ou "objectives".

Dans la définition subjective, l'expert en climat ou météorologie décide lui-même (i) ce que sont les RTs les plus marqués et récurrents, et (ii) comment un ou des champs de variable(s) à grande échelle (par ex., caractérisant une journée) doivent être associés ou "attribués" à l'un des régimes.

1. Ce point sera évoqué au chapitre 3 sur la régionalisation statistique.

Les plus célèbres RTs subjectifs sont certainement ceux de [109] pour la Grande Bretagne ou les régimes de [77] pour l'Europe Centrale. Il est clair qu'une telle approche nécessite une excellente connaissance météorologique de la région d'intérêt.

Mathématicien de formation, je me suis davantage intéressé à l'approche objective. En effet, celle-ci est basée sur des méthodes mathématiques de *classification non-supervisée*, également appelées *clustering*. Ces méthodes cherchent à mettre dans un même groupe des champs atmosphériques qui sont proches les uns des autres, et à séparer dans des groupes différents des champs différents les uns des autres. La notion de "distance entre champs" est donc (explicitement ou implicitement) importante dans de tels algorithmes. Les groupes obtenus sont alors appelés des *clusters* en mathématiques et représentent ici les régimes de temps recherchés. De nombreuses méthodes ont été développées et appliquées pour réaliser une telle tâche. Elles proviennent généralement des études liées à la reconnaissance de forme.

La plus connue est certainement la méthode k-means ([42]), calculant itérativement le centre de chaque cluster (initialisé de manière aléatoire) et ré-attribuant les données au cluster dont le centre est le plus proche jusqu'à ce qu'une condition de fin fixée à l'avance soit atteinte. Les algorithmes de type k-means ont été popularisés par [124] qui décrivaient des régimes de temps récurrents quasi-stationnaires sur la région Nord Atlantique. Ils ont depuis été utilisés dans de très nombreuses études, par exemple pour caractériser des tendances climatiques au travers de RTs sur l'Europe par [87] ; comme base de définition dans un système d'alerte précoce pour les vagues de chaleur par [168, 169] ; pour relier des événements extrêmes de température et de précipitation à des RTs en Atlantique Nord par [219] ; ou pour étudier des structures de téléconnexion par [29].

Les méthodes de type "clustering hiérarchique agglomératif" (HAC, [205]) ont également été très populaires ces dernières décennies. À partir de tous les éléments (ici, les champs journaliers) considérés comme autant de clusters distincts, l'approche HAC génère un arbre (ou dendogramme) en groupant successivement à chaque étape deux clusters pour en former un nouveau, jusqu'à ce qu'il ne reste qu'un groupe contenant tous les éléments. L'arbre ainsi obtenu est alors "coupé" au niveau choisi pour avoir le nombre souhaité de clusters. Une approche similaire à HAC mais inverse (c-à-d., travaillant de manière divisive au lieu d'agglomérative) existe également mais est relativement peu employée de par ses coûts de calculs plus importants. En climat, HAC a été employée, par exemple, pour définir une climatologie des tempêtes violentes en Virginie aux USA par [37] ; pour définir des régions climatiques dans les grandes plaines du Nord Américain par [20] ; pour de la prévision à courts et moyens termes par [186] ; ou pour identifier des RTs dans le Pacifique-Nord Américain par [28].

Des techniques basées sur des réseaux de neurones artificiels peuvent aussi servir pour réaliser du clustering et ainsi définir des régimes de temps grâce aux approches dites par "cartes auto-organisatrices" (ou "Self-Organizing Maps", SOM) créées par [104, 105]. Par exemple, [79] ont utilisé SOM pour définir des RTs servant à conditionner un modèle statistique pour simuler des précipitations journalières ; et [112] pour comparer les capacités de différents GCMs à représenter les caractéristiques spatiales du phénomène "El Niño Southern Oscillation" (ENSO) durant le $20^{ème}$ siècle.

D'autres techniques de clustering issues de l'intelligence artificielle ont également vu le jour, telles que les *règles floues* (fuzzy rules) qui travaillent par optimisation de fonctions objectives choisies par l'utilisateur/modélisateur. Ce type d'approche a été utilisé avec succès par exemple par [2] pour identifier des RTs servant à conditionner des modèles de régionalisation par régressions linéaires ; ou par [145] pour déterminer des régimes mensuels de précipitation sur la Hongrie.

2.1.2 *mélange de distributions par "Expectation-Maximization" (EM)*

J'ai eu l'occasion d'appliquer de nombreuses méthodes de clustering (k-means [214], clustering hiérarchique [195, 179], etc.) et même de développer quelques distances pour le clustering (par ex., [197]) ou des approches originales basées sur des mélanges de corrélations ([201]). Cependant, une grande partie de mes recherches sur les régimes de temps est basée sur une autre approche de clustering mathématique : l'approche par modèle de mélange statistique.

Dans le contexte de cette approche, un RT peut être perçu comme un ensemble d'éléments (par ex., des champs journaliers) confinés dans un (hyper-) volume d'un espace d'états multi-dimensionnels défini par l'ensemble des points de grille. L'hypothèse sous-jacente aux RTs est ici basée sur l'existence d'un nombre fini (et relativement petit) de tels volumes qui représentent des maxima locaux dans la densité de probabilité globale des éléments, correspondant ainsi à une densité multi-modale ([124, 174]).

De manière plus pragmatique, cela signifie que l'on cherche à estimer f, la fonction de densité de probabilité (pdf) multivariée des champs atmosphériques X à classifier, comme un mélange paramétrique, c-à-d. une somme pondérée de K densités paramétriques f_k ([139]) avec paramètres $(\alpha_k)_{k=1,...,K}$:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x;\alpha_k), \tag{2.1}$$

où les π_k sont les poids du mélange et correspondent aux probabilités a priori d'appartenir à la composante k du mélange – c-à-d. au $k^{ème}$ régime de temps, disons RT_k. Dans cette formulation, RT_k est défini par la $k^{ème}$ pdf, f_k, suivant le principe du maximum a posteriori :

$$RT_k = \{x \text{ tq } \pi_k f_k(x;\alpha_k) \geq \pi_j f_j(x;\alpha_j), \forall j = 1,...,K\}. \tag{2.2}$$

Autrement dit, chaque jour (caractérisé par le champ ou vecteur x) est attribué au régime (ou cluster) RT_k pour lequel la pdf f_k maximise la probabilité a posteriori que x appartienne à ce régime. Dans la plupart des études basées sur des mélanges de densités, les pdfs f_k sont des densités Gaussiennes. Ainsi, $\alpha_k = (\mu_k, \Sigma_k)$, avec μ_k le vecteur de moyennes et Σ_k la matrice de variance-covariance de f_k. La Fig. 2.1 présente une illustration d'un mélange simulé de Gaussiennes 2-d à trois composantes.

À K fixé, les paramètres $(\pi_k, \alpha_k)_{k=1,...,K}$ de l'équation de mélange (2.1) sont généralement estimés par l'algorithme "Expectation-Maximization" (EM, [39, 122]). EM consiste en deux étapes

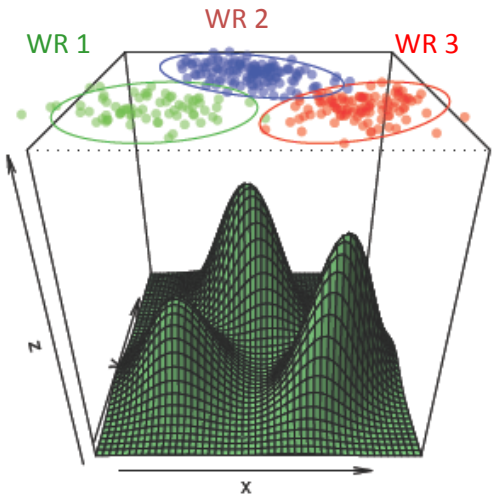


FIGURE 2.1: Illustration d'un mélange simulé de Gaussiennes 2-d à trois composantes.

successives répétées de manière itérative, l'une de calcul de l'espérance ("expectation" – étape E), l'autre de maximisation de la log-vraisemblance dite "complète" (étape M). Cet algorithme et ses variantes étant désormais considérés comme relativement classiques du point de vue statistique, son fonctionnement n'est pas davantage détaillé dans ce manuscrit. Le lecteur désireux d'en savoir plus sur ces aspects techniques pourra se référer à [39, 122, 191] par exemple, où tous les détails théoriques et techniques sont discutés.

Par ailleurs, quelle que soit la méthode de clustering employée, le choix du nombre K de régimes est toujours une question délicate. Différentes possibilités existent – telles que les critères du "coude" ou du "saut maximal" (voir [43] ou [195] par exemple) – mais cette question ne possède pas de réponse parfaite et reste donc ouverte et un sujet de recherche actif en reconnaissance de formes entre autres. L'approche par EM étant basée sur le modèle (2.1), elle permet de considérer cette question du nombre K comme un problème de sélection de modèle statistique. Autrement dit, le mélange (2.1) avec, disons, $K = 3$ composantes ne correspond pas au même modèle que le mélange (2.1) avec $K = 4$ composantes : les paramètres ne sont pas les mêmes et leur nombre est également différent. En utilisant EM pour définir les RTs, nous pouvons alors avoir recours aux critères classiquement employés en statistique pour la sélection de modèle, tels que le "Critère d'Information Bayésienne" (BIC, [164]). Pour cela, l'algorithme EM est utilisé pour estimer plusieurs fois les paramètres du mélange (2.1) pour différents nombres K. Le nombre K minimisant le BIC est alors retenu ([122]), avec :

$$BIC(K) = -2\log(L) + p\log(n), \tag{2.3}$$

où L est la vraisemblance du modèle paramétrisé à K composantes, n est la taille de l'échantillon (par ex., le nombre total de jours à partir desquels nous souhaitons définir les régimes) et p est le

nombre de paramètres à estimer. Ainsi, minimiser le BIC offre une façon automatique et statistiquement robuste de déterminer K et correspondant à un compromis entre une bonne vraisemblance (c-à-d., une bonne adéquation des données au modèle) et un nombre raisonnable de paramètre (c-à-d., pour éviter la sur-paramétrisation).

De plus, pour chaque nombre K de clusters testé, différentes contraintes sont également testées sur la structure de la matrice de variance-covariance afin, encore une fois, d'éviter la sur-paramétrisation. Cette matrice peut être sphérique, diagonale ou ellipsoïde et correspondant à des volumes égaux ou pouvant varier d'un régime à l'autre ([54]). Le BIC permet ainsi de sélectionner le nombre K de régimes et leur structure de covariance.

Les RTs définis par EM sont donc associés à des modèles statistiques. EM est donc un outil de *modélisation* des structures de circulation à grande échelle, ce qui n'est pas nécessairement le cas des autres méthodes de clustering qui caractérisent plus qu'elles ne modélisent.

Quelles hypothèses en changement climatique ?

Lorsque des RTs sont définis pour caractériser les principaux modes de variabilité temporelle de l'atmosphère, il est possible de projeter sur ces modes des simulations futures à grande échelle provenant de GCMs. Nous pouvons alors comparer les fréquences des différents régimes pour la période future choisie avec les fréquences pour le climat d'une période historique. Si cette information sur l'évolution des fréquences de RTs dans le temps est importante, elle n'en repose pas moins sur une hypothèse très forte : que la structure même des RTs ne change pas dans le temps. Ceci est une hypothèse sous-jacente à l'ensemble des méthodes de clustering. Cependant, l'approche EM permet en plus de connaitre l'évolution des probabilités d'occurrence des régimes dans le temps. En effet, il est possible de calculer la probabilité $\tau_k(x)$, la probabilité d'appartenance de la situation atmosphérique x au régime k par :

$$\tau_k(x) = \frac{\pi_k f_k(x;\alpha_k)}{\sum_{j=1}^{K} \pi_j f_j(x;\alpha_j)} \tag{2.4}$$

Cette probabilité donne ainsi une indication sur la qualité des associations à des RTs à partir de nouvelles situations atmosphériques, ou tout au moins sur la confiance que l'on peut avoir dans celles-ci. Une faible probabilité pour tous les régimes indiquerait alors un changement potentiel dans la structure des régimes définis en climat historique et donc une inadéquation de ces RTs pour catégoriser les nouvelles situations (par ex., futures).

Les sections suivantes de ce chapitre présentent mes principaux apports en termes de régimes de temps définis par l'approche de mélange. Nous verrons dans la section 2.2 une extension originale de cette approche au cas des données de type "fonction de répartition" réalisée durant ma thèse de doctorat. Dans la section 2.3, je discuterai comment l'utilisation de EM permet également

d'évaluer quantitativement la capacité des GCMs à reproduire les régimes de temps capturés dans les réanalyses. La section 2.4 consacrera quelques paragraphes sur une utilisation originale de EM pour redéfinir les saisons à partir de régimes de temps atmosphériques, et illustrera brièvement une analyse sur les évolutions en fréquence et en magnitude de ces saisons. Enfin, j'illustrerai en section 2.5 une méthode de clustering récemment développée et basée sur un mélange d'analyses des corrélations canoniques (CCA). Cette méthode, permettant d'introduire une dose d'information locale (par ex., provenant de stations) dans des RTs caractérisés à grande échelle, sera brièvement intercomparée avec d'autres méthodes de clustering plus classiques.

2.2 Types de masse d'air par *données CDFs*

Ma première expérience avec EM a sans doute également été la plus complexe du point de vue théorique. En effet, lors de ma thèse ([191]), j'ai développé un modèle de mélange travaillant, non pas sur des données classiques mais sur des données "fonctions de répartition" ("cumulative distribution functions" en anglais, CDFs par la suite). Autrement dit, chaque élément à classifier était caractérisé par une CDF. En effet, les bases de données (collectées ou simulées) sont désormais si imposantes en taille que certaines méthodes statistiques classiques sont inappropriées ou inefficaces ([14]). Les données doivent donc subir une étape "d'agrégation" initiale pour condenser l'information, par exemple sous la forme d'intervalles, d'histogrammes ou comme ici de CDFs. Plus précisément, les données initiales étaient ici des valeurs de température (T) et d'humidité spécifique (Q) provenant de l'ECMWF et réparties pour chaque degré de longitude $\times$ latitude en une trentaine de niveaux de pression sur la verticale (niveaux "sigma"). Pour chaque position longitude $\times$ latitude, une CDF non-paramétrique a été modélisée par la méthode des noyaux de Parzen ([138, 170]) à partir de la trentaine de valeurs de température (resp., d'humidité) sur la verticale. Ceci a été fait pour l'ensemble des points de grille du globe et uniquement (dans un premier temps et pour l'illustration dans ce manuscrit) pour le 15 décembre 1999 à 0h. En effet, le but ici n'était pas un clustering *temporel* – c-à-d., grouper/classifier des jours – mais un clustering *spatial* – c-à-d., définir des régions cohérentes à l'instant t. De plus, travailler sur T et Q au lieu des variables de circulation plus classiques, fait que nous ne définissons pas des RTs mais des "types de masse d'air". Cependant, pour des raisons de simplicité, le terme "RT" sera tout de même employé dans cette section.

Afin de travailler sur ces données CDFs, il a tout d'abord été nécessaire de définir la notion de CDF pour valeurs de CDFs. Si $\mathbb{F} = \{F_1, ..., F_N\}$ est un échantillon de N CDFs (par ex., des CDFs de température) provenant d'une population Ω_F, une *CDF de valeurs de CDFs au point Z* (Z étant par ex., une valeur donnée de température) est une fonction G_Z qui va de $[0,1]$ dans $[0,1]$ et définie par

$$G_Z(x) = \mathbb{P}(F(Z) \leq x), \forall x \in \mathbb{R} \tag{2.5}$$

qui peut être modélisée de manière empirique (par fonction en escalier) par G_Z^{emp} :

$$G_Z^{emp}(x) = \frac{card\{F_i \in \mathbb{F} \text{ tq } F_i(Z) \leq x, i = 1,...,N\}}{N} \tag{2.6}$$

où $card(A)$ est le cardinal de l'ensemble A. Ces deux équations (2.5) et (2.6) peuvent être généralisées pour $p \geq 2$ valeurs de CDF, avec une *CDF jointe de valeurs de CDFs au point* $Z = (Z_1,...,Z_p)$ (un p-uplet de valeurs de température) définie comme une fonction H_Z qui va de $[0,1]^p$ dans $[0,1]$ par :

$$H_Z(x_1,...,x_p) = \mathbb{P}(F(Z_1) \leq x_1,...,F(Z_p) \leq x_p). \tag{2.7}$$

Remarquons au passage que les fonctions définies en (2.5) et (2.7) sont elles-mêmes des CDFs et peuvent s'étendre au cas à n dimensions. Ces extensions ne sont pas détaillées ici, voir [192].

L'idée principale de l'approche développée pendant ma thèse ([191, 193, 192]) était de modéliser la CDF H_Z comme un mélange :

$$H_Z(x) = \sum_{k=1}^{K} \pi_k H_{k,Z}(x;\alpha_k), \tag{2.8}$$

où les π_k sont les proportions du mélange et α_k le vecteur paramètres de la $k^{èmes}$ CDF $H_{k,Z}$.

Pour modéliser les CDFs paramétriques $H_{k,Z}$, nous nous sommes basés sur la théorie des copules. Les copules fournissent une façon de décrire les liens qui existent entre une CDF jointe (c-à-d., multi-dimensionnelle) et les CDFs marginales associées (c-à-d., univariées). Les propriétés que doit posséder une fonction pour être une copule ne sont pas données ici mais peuvent être facilement trouvées, par exemple dans [129, 192]. Le coeur de la théorie des copules est certainement le théorème de Sklar ([171]) : Si H est une CDF à p dimensions et de marginales $G_1,...,G_p$, alors il existe une fonction copule C telle que

$$H(x_1,...,x_p) = C(G_1(x_1),...,G_p(x_p)). \tag{2.9}$$

Ainsi, en appliquant le théorème de Sklar aux $H_{k,Z}$ de l'Eq. (2.8), de marginales $G_{k,Z_1},...,G_{k,Z_p}$, il existe K fonctions copules de dimension p $C_1,...,C_K$ telles que

$$H_{Z_1,...,Z_p}(x_1,...,x_p) = \sum_{k=1}^{K} \pi_k C_k(G_{k,Z_1}(x_1;\alpha_{k,Z_1}),...,G_{k,Z_p}(x_p;\alpha_{k,Z_p})), \tag{2.10}$$

avec α_{k,Z_i} le(s) paramètre(s) de la marginale G_{k,Z_i}. Nous disposons alors d'un modèle de mélange de fonctions copules. En dérivant les termes des deux côtés de l'égalité (2.10) – qui sont tous les deux des CDFs –, nous nous ramenons à des densités de probabilité et donc à un mélange de densités plus classique tel qu'exprimé précédemment. L'équation dérivée n'est toutefois pas présentée ici pour ne pas alourdir le manuscrit. En effet, l'expression de la densité associée au terme de droite de (2.10) s'avère relativement complexe à écrire due aux nombreux indices employés. L'ensemble des détails se trouve évidemment dans les publications associées à ces travaux (par ex., [192]). Cependant, l'écriture du mélange de distributions (2.10) permet tout à fait d'appliquer l'algorithme

EM – ou même ses variantes. Différentes familles de copules ont été testées et pour diverses raisons techniques, nous avons retenu la copule de Frank ([55]) comme forme de copules paramétriques dont les paramètres sont à estimer par EM. Ces quelques points techniques ne représentent qu'une petite partie des travaux réalisés durant ma thèse de doctorat. Le lecteur comprendra aisément que les autres résultats théoriques ne soient pas davantage mentionnés ici.

En appliquant l'algorithme EM pour résoudre l'équation (2.10) à partir des données de température décrites au début de cette section et avec $Z = (225, 265)$ (unités en Kelvin), le BIC nous a indiqué que $K = 7$ RTs correspondait au meilleur compromis vraisemblance/complexité. La Fig. 2.2 présente, pour illustration, la carte des sept RTs ainsi obtenus. Les structures mises à jour ont été

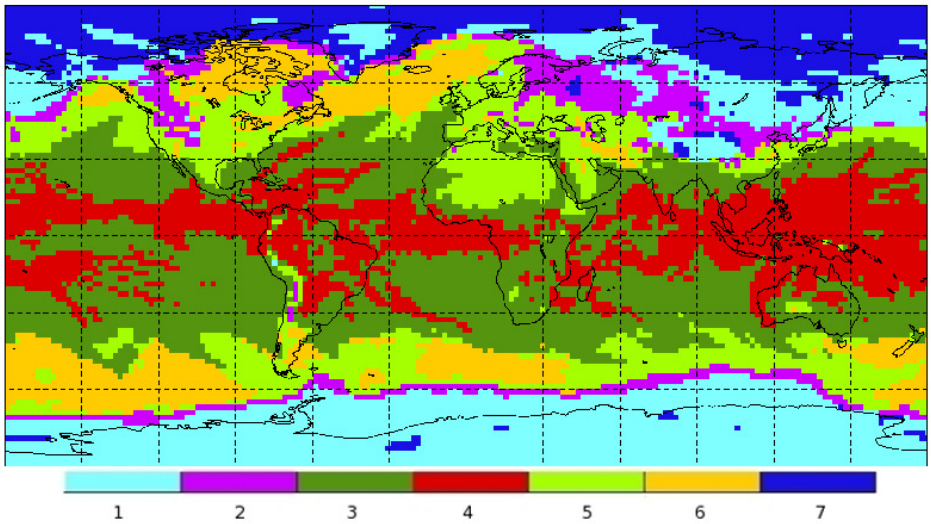


FIGURE 2.2: Les 7 régimes obtenus en résolvant l'Eq. (2.10) par EM sur les données de CDFs de température.

analysées et ont montré une grande cohérence et une excellente discrimination en terme de densité et ce, à différents niveaux de pression (non montré). Par ailleurs, une version multi-dimensionnelle – travaillant à la fois sur les CDFs de température et celles d'humidité – a également été mise au point (non détaillée, voir [192]). Celle-ci a permis de modéliser des structures extrêmement fines, par exemple associées à des dépressions prononcées mais très localisées ([193]). Plusieurs comparaisons à d'autres méthodes de clustering plus classiques ont été ménées (EM classique, HAC, k-means) et ont toutes montré la qualité des résultats par mélange de copules mais également l'intérêt de l'utilisation des données probabilistes.

Les apports de ce travail étaient multiples. Du point de vue pratique, à partir d'un modèle calibré de mélange de copules, il est par exemple possible de projeter les conditions prévues à grande échelle à un temps ultérieur pour disposer non seulement de nouvelles cartes présentant les régions

cohérentes (au sens des CDFs) mais aussi de la probabilité d'occurrence de chaque cluster en chaque localisation longitude $\times$ latitude. Du point de vue théorique et technique, j'ai mis au point diverses méthodes d'estimation des paramètres de copules, j'ai développé une nouvelle façon de définir une copule multi-dimensionnelle, et montré divers points théoriques tels que la convergence ou des propriétés asymptotiques. Je ne détaillerai pas davantage les apports de ce travail dans ce manuscrit.

2.3 Évaluation par modèles probabilistes

Les résultats obtenus dans cette section ont été obtenus – et concrétisés dans [156] – grâce à une collaboration avec H. Rust (post-doc que j'ai encadré au LSCE) et avec B. Sultan et M. Lengaigne (LOCEAN) dans le cadre du projet REGYNA. Dans [156], nous commencions par discuter les apports de l'approche EM qui permet de définir des régimes où la matrice de covariance n'est pas nécessairement sphérique, contrairement à l'approche par k-means avec distance Euclidienne. Je ne reviendrai pas dans ce manuscrit sur les comparaisons effectuées et les résultats obtenus sur ce point. Je m'attarderai en revanche sur l'utilisation que l'on a faite des sorties de l'approche EM (les pdfs associées à chaque RT) pour évaluer les GCMs en terme de régimes de temps caractérisés par des densités.

Il existe de nombreuses façons d'appliquer une méthode de clustering pour définir des RTs selon le contexte et le but à atteindre. Par exemple, si l'on s'intéresse aux changements de fréquences des régimes, ceux-ci sont généralement tout d'abord définis pour une période dite de contrôle (CTRL) du GCM et caractérisent alors la variabilité dans le monde du GCM. Puis, les simulations GCM pour une période future sont projetées sur ces RTs pour calculer les nouvelles fréquences associées à cette nouvelle période. Ce type d'analyse repose évidemment sur l'hypothèse que seules les fréquences évoluent dans le temps et que les structures des RTs ne changent pas (voir paragraphe "*Quelles hypothèses en changement climatique*" en page 19). Si l'on s'intéresse à la variabilité "observée" (par opposition à "simulée" par les GCMs), la définition des RTs se basera sur des données de réanalyses. À partir de ces RTs, il est également possible d'évaluer les GCMs. Une approche consiste pour cela à projeter les simulations des GCMs (pour la même période que celles des réanalyses) sur les RTs et à comparer les fréquences des RTs des réanalyses avec les fréquences des RTs attribués aux simulations. Cependant, une telle approche n'est pas totalement satisfaisante. En effet, l'attribution d'une situation se fait généralement au régime dont la distance, souvent au centroïde, est minimale. Or, pour un nombre fini de régimes, il existe toujours un RT le plus proche d'une situation donnée, aussi excentrique soit elle. Cela ne signifie pas pour autant (1) qu'ils soient réellement proches dans l'absolu, ni même (2) que la structure et la variabilité interne des RTs définis par attribution soit proches de celles du régime des réanalyses. Une autre approche consiste à définir des RTs directement à partir de simulations GCM et à comparer ces régimes à ceux obtenus à partir des réanalyses. Pour cela, il faut donc apparier (c-à-d., mettre par paire) un régime de réanalyses avec un régime GCM. Ceci est généralement effectué par minimisation d'une

distance (souvent la distance Euclidienne) appliquée sur les centroïdes. Autrement dit, ces deux approches d'évaluation de GCM par régimes de temps reposent principalement sur des associations GCM-réanalyses qui se font généralement uniquement en termes de valeurs moyennes des RTs. L'information sur la variabilité interne à chaque régime (par exemple, ce qu'on a pu nommer le volume en section 2.1.2) n'est pas du tout prise en compte. Il était donc nécessaire de développer des mesures de distance permettant de tenir compte de cette variabilité interne et incluant la taille et la forme des RTs, en plus de la simple valeur moyenne.

Afin de mener à bien ces développements, nous nous sommes basés sur les apports de l'approche par modèle de mélange avec EM. En effet, cette approche associe à chaque régime une densité de probabilité, c-à-d., pas une unique valeur moyenne mais également une caractérisation de la structure de sa variabilité interne grâce à la matrice de variance-covariance (dont la forme est sélectionnée par le BIC). L'idée était alors d'appliquer des mesures incorporant au minimum ce type d'information et même, si possible, travaillant directement sur les densités de probabilité pour plus de généralités (dans le cas de variables non Gaussiennes). Pour mémoire, la distance Euclidienne multi-dimensionnelle est :

$$d_{\text{Eucl}}(P,Q) := \|\mu_P - \mu_Q\| = \sqrt{(\mu_P - \mu_Q)^t(\mu_P - \mu_Q)} \tag{2.11}$$

où μ_P et μ_Q sont les vecteurs des moyennes des régimes de temps P et Q. Ces vecteurs peuvent évidemment provenir de EM, ou peuvent être calculés a posteriori de la définition des RTs par une autre méthode. Une généralisation de cette distance a été introduite par [115], la distance de Mahalanobis :

$$d_{\text{Maha}}(P,Q) := \|\mu_P - \mu_Q\|^2_{\Sigma_P^{-1}} = (\mu_P - \mu_Q)^t \Sigma_P^{-1}(\mu_P - \mu_Q), \tag{2.12}$$

avec Σ_P^{-1} la matrice de covariance associée au régime P. Ici, de la même manière que pour la distance Euclidienne, les paramètres peuvent provenir directement de EM ou être calculés a posteriori de la définition des RTs. Cette mesure symétrique fut créée pour des distributions Gaussiennes avec des matrices de covariance identiques et ne correspond à une distance au sens mathématique que dans ce cas. Sinon, elle dépend de la matrice de covariance Σ_P^{-1} (ou plus généralement de la pdf associée au régime P) utilisée comme référence dans la définition (2.12). Cette distance peut cependant aisément être symétrisée par $d_{\text{Maha,sym}}(P,Q) = d_{\text{Maha}}(P,Q) + d_{\text{Maha}}(Q,P)$. Dans la suite, si les RTs sont définis par EM, les régimes P et Q sont respectivement associés aux densités f et g.

Une autre approche populaire pour quantifier les différences entre pdfs est la *divergence* [2] de Kullback-Leibler (KL par la suite), également appelée la discrimination d'information ([107]), entropie relative ou coefficient I. Cette mesure est définie par :

$$d_{\text{KL}}(P,Q) := I(P \mid Q) = \int_{\mathbb{R}} \log\left(\frac{g(x)}{f(x)}\right) g(x)\, dx. \tag{2.13}$$

Elle mesure la différence entre la pdf f de référence et associée au régime P (la "vérité" provenant par exemple des réanalyses) et une autre pdf g associée au régime Q (par ex., provenant d'un

2. fonction de base pour définir des mesures entre densités, voir [36, 14].

GCM). Le KL est non-commutatif, c-à-d. que c'est une mesure non-symétrique et en tant que telle, ce n'est pas une distance au sens mathématique du terme. Il est relié aux mesures d'entropie de la théorie de l'information, par exemple à l'information mutuelle et à l'entropie de Shanon ([108]). Une particularité est que pour deux Gaussiennes de matrices de covariance identiques, le KL se réduit en $(1/2)d_{\text{Maha}}$ (voir [14, 156]). Une version symétrisée du KL, appelée coefficient J, peut être formulée ainsi :

$$d_{\text{J}}(P,Q) := J(P,Q) = I(P \mid Q) + I(Q \mid P). \tag{2.14}$$

Si les deux pdfs sont égales, le KL et le coefficient J valent zéro.

Une approche conceptuellement différente est basée sur le coefficient de Hellinger ([76]) de paramètre $0 < s < 1$:

$$d^{(s)}_{\text{Hell}}(f,g) := \int_{\mathbb{R}} g(x)^s f(x)^{(1-s)}\, dx. \tag{2.15}$$

Choisir $s = \frac{1}{2}$ crée une mesure symétrique avec valeurs entre 0 (f et g sont de supports disjoints) et 1 ($f \equiv g$). Ce coefficient peut ainsi être perçu comme une mesure de chevauchement ("overlapping") des deux pdfs. L'Eq. (2.15) définit en fait une similarité. Il est souvent utile d'utiliser le complément (c-à-d., $1 - d^{(s)}_{\text{Hell}}(f,g)$) pour être consistent avec les autres mesures.

La Table 2.1 résume les principales caractéristiques de ces mesures. La Fig. 2.3 présente quatre

Mesure	Notation	Range	Symétrique ?	Caractéristiques
Euclidienne	d_{Eucl}	$[0,\infty)$	oui	distance entre moyennes uniquement
Mahalanobis	d_{Maha}	$[0,\infty)$	non	distance entre moyennes avec métrique dépendant d'une matrice de covariance
Kullback-Leibler(KL)	d_{KL}	$[0,\infty)$	non	prend les deux pdfs (et donc matrices de covariance) en compte ; se réduit à Mahalanobis pour $\Sigma_f = \Sigma_g = I$
Coefficient J	d_J	$[0,\infty)$	oui	KL symétrisé ; $2d_J = d_{\text{KL}}$ sous certaines conditions de symétrie
Hellinger ($s = 0.5$)	d_{Hell}	$[0,1]$	oui	mesures de chevauchement

TABLE 2.1: Quelques caractéristiques des mesures entre pdfs suggérées dans [156].

illustrations du comportement de ces mesures calculées entre deux distributions Gaussiennes bivariées selon différentes formes et positions, où les grandes ellipses représentent 90% de la masse des distributions. Les barres sur la droite de chaque sous-figure montrent les valeurs des mesures. La distance de Mahalanobis est divisée par 2 afin que pour des covariances identiques nous obtenions la même valeur que celle du KL. Idem, le coefficient J est divisé par 2 afin que sous certaines conditions de symétrie il soit égale au KL (comparer Figs. 2.3b et 2.3d). Il est bon de rappeler que la comparaison des valeurs de différentes mesures n'est pas toujours sensée. Cependant, pour diverses situations, leur intercomparaison peut apporter certains éclairages. En Figs. 2.3a-c, les centres (c-à-d., moyennes) ne changent pas mais l'orientation et la forme évoluent. La distance Euclidienne reste donc la même contrairement à celle de Mahalanobis. Dans le panel (b), le KL et

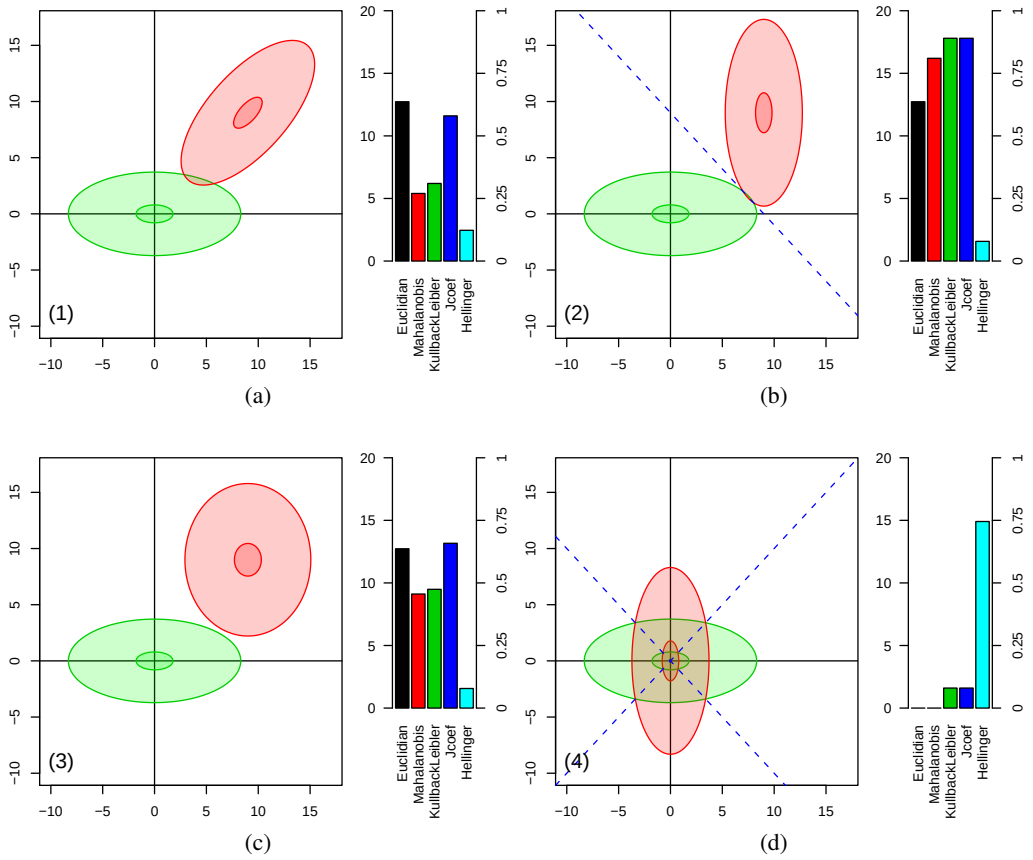


FIGURE 2.3: Illustrations des distances Euclidienne et de Mahalanobis, de la divergence de Kullback-Leibler (KL), et des coefficients J et de Hellinger, à l'aide de deux pdfs Gaussiennes bivariées. Les valeurs de la distance de Mahalanobis et du coefficient J ont été divisées par 2 pour faciliter les comparaisons avec la divergence de KL. La densité f de référence pour Mahalanobis et KL est la densité variant en position et en forme (rouge), voir Eq. (2.12). Les barres sur la droite de chaque figure montrent les valeurs des mesures. Noter que le coefficient de Hellinger est une mesure de similarité et réagit donc à l'opposé des changements dans les pdfs. (a) et (c) : Deux situations asymétriques différentes ($d_{KL} \neq d_J$) avec une pdf f variant en forme. (b) et (d) : Deux situations symétriques différentes ($d_{KL} = d_J$) avec une pdf f variant en position. Les lignes pointillées indiquent les axes de symétrie.

le coefficient J (divisé par 2) sont les mêmes à cause de la symétrie le long de la ligne tiretée. Dans le panel (c), un changement de forme casse cette symétrie, créant ainsi une différence entre KL et J. Le panel (d) montre deux distributions concentriques. Les distances Euclidienne et de Mahalanobis valent donc zéro. Toutefois, les matrices de covariance diffèrent en orientation, le KL et le J sont donc non nuls et le coefficient de Hellinger est inférieur à 1 (indicant une différence entre les pdfs). Due à la symétrie, KL et J sont encore égaux mais plus petits qu'en panels (a) et (b) à cause des moyennes identiques des pdfs. Le coefficient de Hellinger a fortement augmenté, signifiant que des parties importantes des pdfs se chevauchent. Ces illustrations simulées démontrent que les

mesures de Mahalanobis, de KL, J et de Hellinger fournissent des informations complémentaires à celles données par la distance Euclidienne.

Nous avons alors appliqué ces distances pour essayer d'évaluer différents GCMs en termes de régimes de temps. Pour cela nous avons extrait les champs journaliers des valeurs de pression au niveau de la mer (SLP) issues :

– des réanalyses NCEP/NCAR et des réanalyses ERA-40,

– ainsi que de 14 GCMs [3]

et couvrant la région Nord-Atlantique (définie comme $[-60^o,\ 40^oE] \times [30^o,\ 70^oN]$) pour tous les jours d'hiver (de novembre à mars – NDJFM) de la période 1975-2000. Pour faciliter les intercomparaisons et évaluations ultérieures, toute les données ont été bi-linéairement interpolées à la résolution spatiale de NCEP/NCAR de $2.5^o \times 2.5^o$. Pour chaque modèle, les anomalies de SLP sont ensuite calculées par rapport au cycle saisonnier moyen (en chaque point de grille) défini par une fonction spline de lissage calculée à partir de la moyenne des 26 années. Afin d'éviter une sur-représentation des dimensions (ici les positions des points de grille) hautement corrélées, il est classiquement avantageux de réduire la dimensionnalité des données (et donc du modèle statistique de clustering à appliquer par la suite) par une étape d'analyse en composante principale (ACP, [94]). Cette analyse a ici été appliquée à la base de données complète des anomalies de SLP, contenant l'ensemble des jeux de données de réanalyses et des simulations des 14 GCMS. Cette procédure garantit à tous les types de données une base commune de composantes principales (PCs) dans laquelle la variabilité de tous les modèles peut être représentée et intercomparée. Dans la suite, nous avions retenu les projections sur les 10 premières PCs, représentant plus de 85% de la variance globale (ce pourcentage avec ce nombre de PCs était d'ailleurs celui généralement obtenu lors d'ACPs appliquées séparément à chaque modèle). L'algorithme EM a tout d'abord été lancé uniquement à partir des projections des réanalyses NCEP/NCAR sur ces 10 PCs. Le nombre de $K = 5$ régimes a été choisi arbitrairement pour faciliter les comparaisons avec les cinq régimes (Atlantic Ridge, Blocking, Greenland Anticyclone, Western Blocking et Zonal) classiquement considérés par k-means sur cette région en hiver [4] (voir par exemple [219]). Le critère BIC (2.3) a cependant permis de sélectionner la structure des matrices de covariance qui s'est avérée être non-sphérique. Les régimes moyens obtenus (non présentés dans ce manuscrit, voir [156]) présentaient alors à la fois des similitudes et des différences avec les cinq régimes classiques.

EM a ensuite été appliqué *indépendamment* sur ERA-40 et sur chaque GCM (plus précisément à partir des projections de leurs anomalies respectives sur les 10 PCs de la base commune). Les différentes mesures décrites précédemment ont ensuite été calculées entre chacun de ces régimes et chacun des régimes NCEP/NCAR. Pour chaque mesure, le couple de RTs de mesure minimale a été associé. Il est alors possible de classer les GCMs dans l'ordre de leur capacité à reproduire chaque régime pour chaque mesure. Pour illustration, je présente en Fig. 2.4 une comparaison entre les valeurs de distance Euclidienne et celles du coefficient J pour les régimes associés au

3. cccma_cgcm3_1, cnrm_cm3, csiro_mk3_0, csiro_mk3_5, gfdl_cm2_0, gfdl_cm2_1, ingv_echam4, inmcm3_0, ipsl_cm4, miroc3_2_hires, miroc3_2_medres, miub_echo_g, mpi_echam5, mri_cgcm2_3_2a

4. Noter que le BIC n'était pas nécessairement d'accord sur ce nombre de régimes.

Greenland Anticyclone (GA). L'ordre des GCMs dans cette figure est déterminé par coefficients

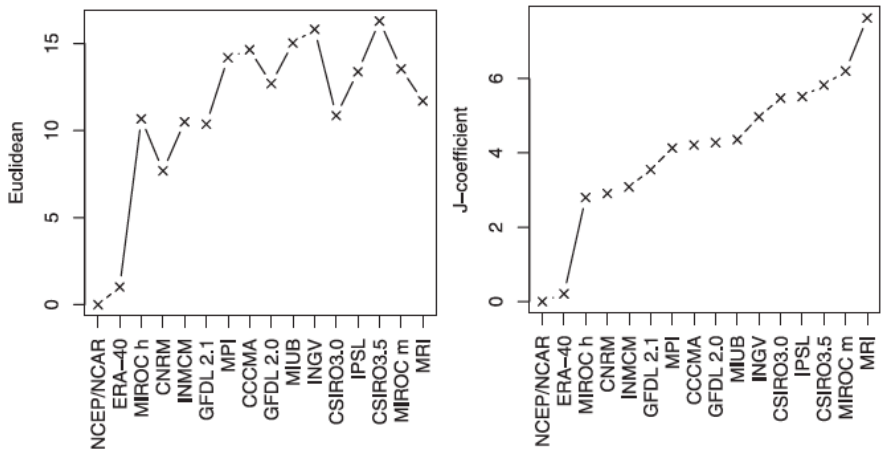


FIGURE 2.4: Valeurs de la distance Euclidienne (à gauche) et du coefficient J (à droite) des régimes GA de GCM par rapport au régime GA de NCEP/NCAR. Les GCMs sont ordonnés en abscisse par valeurs croissantes du coefficient J.

J croissants. Ce régime GA a été choisi pour cette illustration car il est le plus distinct des autres régimes au sens de NCEP/NCAR et ceci pour l'ensemble des mesures. Il est donc celui qui s'associe également le plus clairement. Une description approfondie de cette illustration n'est pas donnée ici mais se trouve bien sûr dans [156]. On peut tout de même remarquer que la distance Euclidienne classique n'entraine pas le même ordre pour les GCMs. Ceci signifie que "l'extension" (ou la variance) des pdfs associées à ce régime n'est pas négligeable comparée aux différences de valeurs moyennes, ce à quoi on peut s'attendre : nous obtenons des moyennes de RTs GCM proches des moyennes des RTs NCEP/NCAR mais avec des variabilités internes (formes, tailles) différentes. Ces mesures présentent ici l'intérêt de quantifier ces différences.

Le but de telles comparaisons de GCMs était de trouver un ou quelques GCMs qui soient très similaires aux réanalyses en termes de régimes de temps et de leur variabilité interne. Ceci devrait s'avérer utile pour des études sensibles aux comportements de modèles climatiques et de leur variabilité, telles que la régionalisation statistique ou dynamique des précipitations (voir chapitre 3), ou des analyses de vagues de chaleur généralement reliées au régime de blockage ([219]).

Enfin, je souhaite mentionner qu'un package R appelé "gaussDiff" a été développé pour le calcul des différentes mesures présentées ici. Ce package est librement et gratuitement disponible sur le site web du "Comprehensive R Archive Network" (CRAN) sur http ://cran.r-project.org/.

2.4 Régimes *saisonniers* (ou comment étudier l'assertion populaire "Il n'y a plus de saisons")

Dans l'étude présentée à la section 2.3 précédente, comme dans la plupart des études portant sur des régimes de temps, les premières étapes consistent généralement à *désaisonnaliser* les données (c-à-d., retirer le cycle saisonnier moyen) et à choisir la ou les saisons sur lesquelles les régimes vont être définis par clustering. Ces étapes permettent de caractériser les principaux modes de variabilité autour du cycle "normal" tout en différenciant les anomalies par exemple hivernales et estivales généralement causées par des phénomènes climatiques différents. Or, ne pas réaliser ces étapes de désaisonnalisation et de conditionnement saisonnier fait apparaitre des structures à grande échelle qui permettent de décrire les saisons et leurs évolutions dans le temps de manière originale. C'est tout l'objet de cette section, illustrée par certains des résultats de mes travaux [202] réalisés en collaboration avec P. Yiou (LSCE) et P. Vaittinada Ayar (stagiaire M2, LSCE), et poursuivant l'étude initiale [195].

Les quatre saisons sont classiquement définies selon des critères astronomiques : équinoxes pour le printemps et l'automne, solstices pour l'été et l'hiver. Ces dates correspondent à des périodes associées à des quantités différentes d'énergie solaire reçue au sommet de l'atmosphère. Cette définition donne les dates de début et de fin des saisons qui sont facilement prédictibles par des calculs astronomiques. Cependant, d'un point de vue météorologique, les caractéristiques des valeurs des variables de surface ne changent pas brutalement le jour de l'équinoxe ou du solstice. Les transitions de variabilité de ces variables (telles que la température, la pression ou la précipitation) d'une saison à une autre – et surtout la perception que nous en avons – varient donc d'une année à l'autre.

De nombreuses études existent sur la saisonnalité et ses changements passés ou futurs ([90]), la plupart étant orientée vers les températures ou les précipitations ([4, 223, 203, 143]). Ces changements ont souvent été étudiés à l'échelle mensuelle ([4, 93]) mais de récentes études ont portées sur des données journalières ([195]) ce qui a motivé notre étude [202]. Mes études sur les changements saisonniers s'appuient sur une définition originale des saisons basée sur des régimes de temps journaliers définis sur des données "brutes" (par opposition aux anomalies) à grande échelle et pour toute l'année (et non pour une unique saison). En effet, notre but étant de décrire la saisonnalité, il n'y aurait aucun sens à travailler sur les anomalies, ce qui correspondrait à retirer l'objet même d'intérêt de ces études. De même, travailler sur l'année entière permet de capturer les transitions continues d'une saison à une autre en ne contraignant pas les régimes à ne plus apparaitre à une date fixée. Cela pourrait s'avérer utile en contexte d'études sur les changements de saisons dans le futur où celles-ci pourrait avoir des périodes différentes (au moins au sens de la perception humaine).

Si dans l'étude [195], portant sur les régimes Nord-Américains, j'ai pu étudier ce type de RTs à partir de données de hauteurs de géopotentiel à différents niveaux de pression (non présentées dans ce manuscrit), dans l'étude [202] nous nous sommes essentiellement concentrés sur le géopotentiel

à 500 mb (Z500) qui est l'une des variables habituelles des études de RTs (par ex., [124, 219, 29]). Z500 est souvent choisie car elle intègre les fluctuations atmosphériques jusqu'au milieu de la troposphère et est moins "bruitée" que les pressions à la surface ou au niveau de la mer. Les données que nous avons utilisées sont les champs journaliers de 1975 à 2009 de Z500 des réanalyses NCEP/NCAR qui couvrent la région dite Nord-Atlantique [-77.5^{o}E ; 37.5^{o}E] $\times$ [22.5^{o}N ; 70^{o}N] avec une résolution spatiale de 2.5^o, soient 940 points de grilles pour la région. Une réduction de la dimension des données a été effectuée par une analyse en composantes principales où les valeurs de Z500 étaient pondérées par la racine carré de leur latitude afin de donner des poids proportionnels à la taille du point de grille ([124, 219, 29]). Par la suite, 11 PCs, représentant plus de 90% de la variance totale des champs Z500, ont été retenues. L'algorithme de clustering EM a alors été lancé sur ces composantes pour K variant de 2 à 15. Le BIC nous a indiqué $K = 6$ comme le nombre optimal de régimes. Ces régimes sont présentés en moyennes de Z500 "brutes" et en anomalies par rapport au cycle saisonnier dans les cartes de la Fig. 6.1 en annexe A. Les régimes 1 à 4 ne présentent pas de structures en anomalies très prononcées (anomalies moyennes proches de 0 : jaune). À l'inverse, les régimes 5 et 6 sont davantage marqués, signifiant que, en moyenne, les jours attribués à ces régimes ont des champs de Z500 différents du cycle saisonnier dans ces régions. Les régimes 1 et 3 ont un gradient d'anomalies Sud-Nord, avec un flux zonal correspondant à la phase positive de l'Oscillation Nord-Atlantique (NAO). Les régimes 4 et 5 ont des structures d'anomalies opposées avec une anomalie négative autour de 40^{o}N et positive sur le Groenland. Ceci correspond à la phase négative de la NAO. Le régime 6 est similaire au régime dit "d'Atlantic Ridge", avec une forte anomalie positive dans le centre de la région.

Nous avons calculé la fréquence moyenne de chaque régime pour chacun des 12 mois de l'année. Ces cycles saisonniers de fréquences sont présentés en Fig. 2.5. On voit parfaitement ici

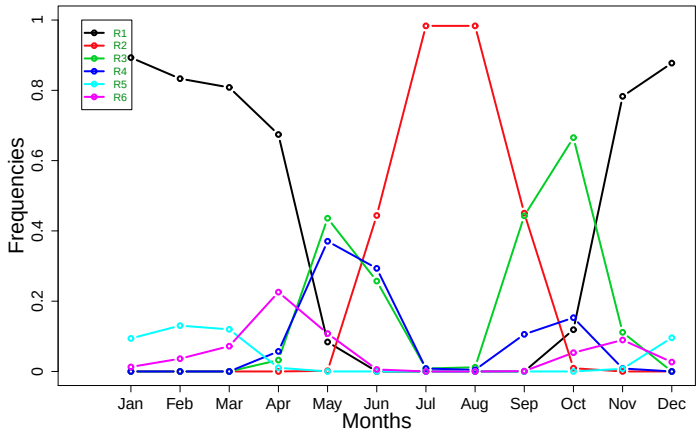


FIGURE 2.5: Cycles saisonniers des fréquences mensuelles des six RTs saisonniers définis par EM.

que la périodicité et les phases des régimes nous permettent de redéfinir les saisons par rapport à ces régimes, justifiant le terme de "régimes saisonniers". Le régime 1 (R1, en noir) correspond aux conditions d'hiver (novembre - avril), le régime 2 (R2, en rouge) à celles d'été (juin - septembre) et les régimes 3 et 4 (R3 et R4, en vert et bleu foncé) sont associés à la fois au printemps (mai - juin) et à l'automne (septembre - novembre) mais dans des proportions différentes. R3 et R4 sont respectivement associés aux phases positive et négative de la NAO. Les régimes 5 et 6 (R5 et R6, en bleu clair et rose) présentent des fréquences d'occurrences nettement moins élevées. Ils sont principalement en hiver pour R5 et pour les mois de "transition" avant et après l'hiver pour R6.

Les fréquences de ces RTs peuvent cependant varier et évoluer d'une année à l'autre. Pour tester ce point, l'évolution des fréquences de chaque régime a été calculée pour chaque mois à l'aide d'une moyenne glissante de cinq ans de 1977 à 2007. La Fig. 6.2 en annexe A présente ces évolutions temporelles mensuelles pour deux mois exemples : (a) septembre et (b) novembre. Les tendances semblent principalement linéaires (ce qui est également vrai pour les autres mois). Pour septembre, R2 (été en rouge) présente une augmentation de ses fréquences, contrairement à R3 dont les fréquences diminuent. Pour novembre, R1 (hiver, en noir) montre une décroissance de ses fréquences qui est compensée par l'augmentation des fréquences de R3 et de R6.

L'ensemble des valeurs des pentes significatives à 95% ($\alpha = 0.05$) des tendances linéaires des fréquences mensuelles de chaque régime sur 1975-2009 est donné en Table 6.1 en annexe A. Grâce à cette table, nous pouvons par exemple voir que la fréquence du RT d'hiver (R1) décroit en novembre – signifiant de moins en moins de jours avec des conditions d'hiver depuis le milieu des années 1970 – et augmente en janvier et février – plus de jours d'hiver. Le régime d'été (R2) présente des fréquences croissantes de mai à octobre, indiquant que les jours avec conditions d'été sont de plus en plus fréquents durant cette période depuis quelques décennies. La description des autres RTs n'est pas faite dans ce manuscrit (voir [202]).

Par ailleurs, afin d'étudier l'aspect temporel (et pas uniquement la fréquence) des évolutions d'occurrences des régimes saisonniers, nous avons aussi déterminé les premières et dernières dates (par an) d'apparition des régimes d'hiver et d'été. Ces dates sont tracées en Fig. 2.6. Elles sont cohérentes avec les fréquences et montrent comment ces régimes se décalent dans le temps [5]. On voit que les conditions hivernales ont tendance à apparaître de plus en plus tard (2.6a) et à se terminer de plus en plus tôt (2.6c), indiquant des hivers de plus en plus courts depuis les années 1970. À l'inverse, les régimes d'été arrivent de plus en plus tôt (2.6b) et s'arrêtent de plus en plus tard (2.6d), contribuant à des saisons estivales plus longues.

Enfin, nous avons également regardé comment les températures évoluaient conditionnellement à ces régimes. Une régression linéaire de température a été calculée avec le temps pour chaque point de grille conditionnellement à chaque régime. La Fig. 6.3 en annexe A présente les tendances de température (en ^{o}C) par régime saisonnier sur 1975-2009, ainsi que sans conditionnement par RTs. Seules les pentes significatives à 95% sont tracées dans cette figure. Notons que ces tendances locales conditionnelles aux six régimes ne peuvent pas être combinées, sommées ou moyennées pour

5. Les fréquences et dates présentées sont significatives à 95%.

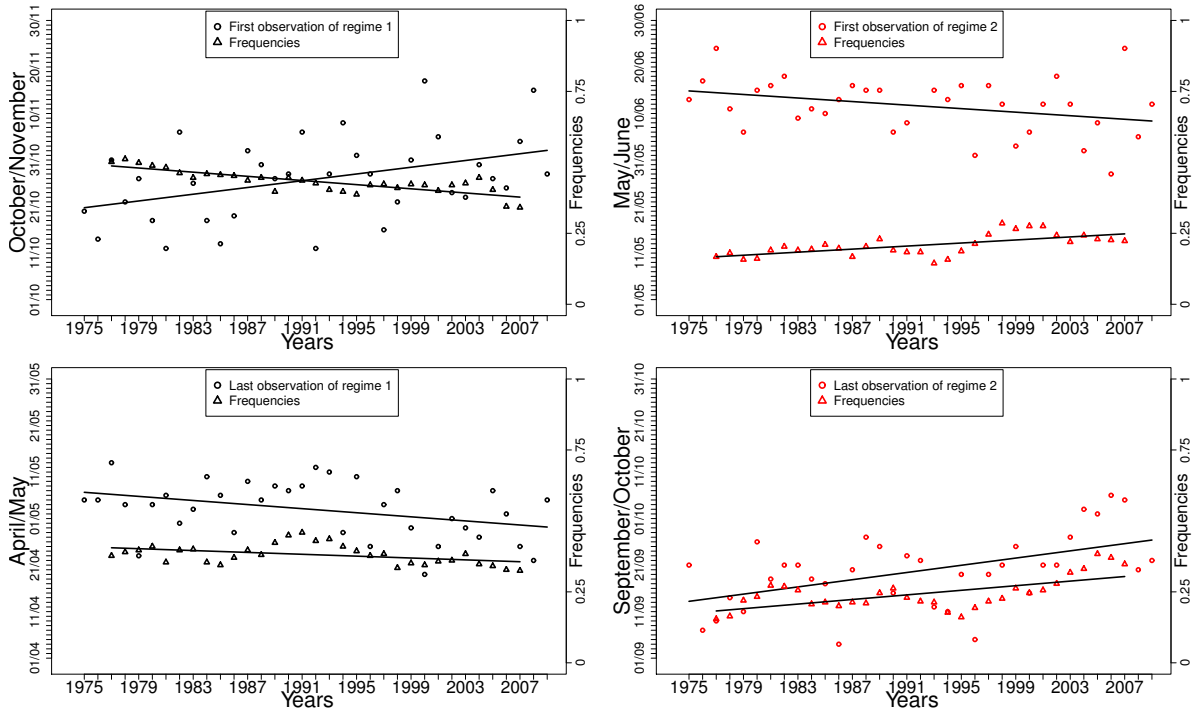


FIGURE 2.6: Évolution interannuelle des premières (a et b) et dernières (c et d) dates d'observations (cercles) pour les régimes d'hiver (R1, a et c) et d'été (R2, b et d). L'évolution des fréquences de ces régimes pour les périodes de deux mois de début ou de fin de saisons est indiquée par des triangles.

retrouver les tendances inconditionnelles. Ces résultats montrent l'effet saisonnier de l'évolution de température qui n'est pas capturé par une analyse inconditionnelle. Le même type de phénomène peut être observé en calculant les tendances significatives de température conditionnellement aux saisons classiques.

Dans cette étude, nous avons ainsi proposé une approche originale pour (re-) définir les saisons, suffisamment flexible pour permettre d'évaluer leurs changements d'une année sur l'autre et sur le long terme.

2.5 Régimes et précipitations locales : un exemple en Méditerranée Francaise

Si d'un point de vue théorique, ma méthode de clustering favorite est certainement l'algorithme EM, d'un point de vue pragmatique, selon le contexte et le but recherché, je suis évidemment prêt à appliquer une autre méthode si celle-ci montre de meilleurs résultats. Différents exercices d'intercomparaison ont été effectués pour mieux comprendre les différences entre RTs obtenus par diverses méthodes de clustering, selon la région d'intérêt, les variables atmosphériques, etc. ([85, 195] ou le projet COST 733 [6]). Les principales conclusions en sont généralement qu'il n'y a pas de "meilleure" méthode de clustering et le choix dépend de la région et de ses spécificités, de la variable d'intérêt et du but à atteindre. Cependant, les méthodes comparées utilisaient uniquement des variables atmosphériques et ne cherchaient pas à maximiser la corrélation de ces variables à des phénomènes météorologiques locaux. Ainsi, en collaboration avec P. Yiou (LSCE), je me suis intéressé aux questions suivantes :

- L'inclusion d'informations locales pour modéliser des RTs à grande échelle aide-t-elle à définir des régimes qui soient bien "discriminés" en termes de précipitation locale ?
- Est-ce que cette inclusion perturbe/complique l'attribution d'un nouveau jour (c-à-d., un nouveau champ) à l'un des régimes ?
- Quel est le coût de mauvaise classification, par ex., dans un contexte de simulations de précipitation basées sur les caractéristiques conditionnelles aux RTs ?

Pour répondre à ces questions nous avons comparé cinq méthodes de clustering où trois d'entre elles inséraient de l'information locale. Les résultats ont donné lieu à un article ([201]) pouvant être utilisé comme guide pour le choix d'une méthode de clustering climatique selon le but de l'application.

Les cinq méthodes que nous avons étudiées sont basées sur la méthodologie par "Correlation Clustering Model", l'approche par k-means, l'algorithme EM. Les deux dernières méthodes sont appliquées soit sur des composantes principales résultant d'une ACP, soit sur les variables canoniques (représentant la grande échelle) issues d'une analyse en corrélations canoniques (CCA, [84, 3]).

Les données à grande échelle de cette étude sont les champs journaliers de hauteurs de géopotentiel à 500 mb et de pression au niveau de la mer (SLP) issus des réanalyses NCEP/NCAR de 2.5^o de résolution pour la région Méditerranéenne [-15^oE ; 42.5^oE] $\times$ [27.5^oN ; 50^oN] associée à 240 points de grille pour chaque variable. Les données "locales" sont des cumuls de précipitation journalières pour sept stations météo réparties dans le sud de la France et provenant du projet "European Climate Assessment & Dataset" (ECA&D, [183]).

Je ne rentrerai pas ici dans les détails sur la CCA qui peuvent aisément être trouvés dans de nombreux ouvrages ou articles (par ex., [84, 3, 201]) mais j'en rappellerai simplement la philosophie. La CCA est fortement reliée à l'ACP. Cette dernière détermine les combinaisons linéaires

6. "harmonisation and applications of weather type classifications for European regions", http ://www.cost733.org/

des variables initiales d'un unique jeu de données pour générer de nouvelles variables (c-à-d., les composantes principales) qui maximisent la variance. De la même manière, une CCA calcule les combinaisons linéaires des variables initiales de deux jeux de données (ici les données à grande échelle d'un côté et les données "locales" de l'autre) pour créer de nouveaux couples (v,w) de variables, nommées variables canoniques (CVs), maximisant, cette fois-ci, la corrélation entre les CVs. Dans la suite, la nouvelle variable v provient de la combinaison des variables locales et w de la combinaison des variables à grande échelle. Ces deux approches sont utilisées pour réduire la dimension des données, tout en maximisant la variance (pour l'ACP) ou les corrélation (pour la CCA).

Ainsi, les approches k-means et EM appliquées sur les PCs des variables à grande échelle uniquement (approches nommées "k-means" et "EM" par la suite) n'insèrent pas d'informations locales pour définir les RTs. Les approches k-means et EM appliquées sur les CVs issues de la CCA sur les données à grande échelle (Z500 et SLP) et les données locales (précipitation aux sept stations) sont nommées "k-means(w)" et "EM(w)" par la suite. Plus précisément, k-means(w) et EM(w) ont été appliquées sur la nouvelle variable w issue de la CCA. k-means(w) et EM(w) permettent d'insérer une information locale dans les données puisque les valeurs de w à classifier sont définies pour maximiser les corrélations avec les données locales.

En plus de ces quatre méthodes, le "Correlation Clustering Model" (CCM) a également été appliqué. CCM a été initialement développé par [50] pour relier végétation et précipitation et adapté ici pour définir des régimes qui tirent parti de la CCA. Cette méthode consiste en un mélange de CCAs. En effet, si une CCA détecte la corrélation linéaire entre deux jeux de données, CCM introduit des non-linéarités grâce à des corrélations linéaires par morceaux. L'idée est de définir des clusters (régimes) qui ont les "meilleurs" modèles de CCA, c-à-d., la plus haute corrélation entre CVs pour chaque RT. Chaque régime est donc caractérisé par sa propre CCA qui n'est pas obtenue a posteriori des régimes : les régimes sont conçus pour optimiser les modèles de CCA. Afin de ne pas alourdir ce manuscrit avec trop de détails techniques, la description des différentes étapes constituant l'algorithme CCM est fournie en annexe B.

Selon différents critères non discutés ici, le nombre de $K = 7$ clusters a été choisi en commun pour les cinq méthodes de clustering dans un souci d'intercomparaison. À partir de ces régimes, plusieurs types d'analyses ont été menés. Nous avons, par exemple, étudié la persistance et la durée moyennes de chaque régime issu de chaque méthode ; nous avons cherché à voir quelle(s) méthode(s) capture(nt) le mieux les événements extrêmes de précipitation en termes d'occurrences et d'intensités ; nous avons comparé neuf méthodes d'attribution (classification *supervisée*) aux régimes des cinq méthodes de clustering à l'aide du taux de mauvaise classification calculé grâce à la mise en place d'une méthodologie originale ; j'ai ensuite développé diverses fonctions de coût quantifiant les implications de la mauvaise classification en termes d'erreurs de simulations de précipitation conditionnellement à chaque régime. Les résultats de toutes ces analyses ne sont pas présentés ici en détails. Seules les conclusions les plus marquées sont rappelées.

Principales conclusions sur CCM

Bien que les régimes obtenus par CCM sont bien discriminés en termes de précipitations locales, le pourcentage de mauvaise classification est généralement élevé quelle que soit la méthode d'attribution. De plus, ces mauvaises classifications peuvent avoir des coûts importants en termes de précipitations simulées (en occurrence et intensité). CCM n'est donc pas recommandé en contexte de modélisation et projection conditionnelles à des RTs. Cependant, les informations locales contenues dans les régimes définis font de CCM un outil très puissant pour l'analyse et l'exploration des liens entre échelles spatiales et particulièrement pour les extrêmes.

Principales conclusions sur k-means et EM

Les algorithmes k-means et EM appliqués aux composantes principales fournissent des régimes possédant à peu près les mêmes propriétés. Bien que les structures générées soient cohérentes en termes de variables atmosphériques, elles ne sont généralement pas bien discriminées au sens des précipitations locales. En effet, contrairement aux méthodes basées sur des CCAs (CCM, k-means(w), and EM(w)), aucune information locale n'est utilisée ici pour définir les régimes. En conséquence, retrouver le RT auquel un nouveau champ de variable atmosphérique doit être attribué, est plus facile et le pourcentage de mauvaise classification est donc plus faible. De plus, la similarité relative d'un régime à un autre, fait qu'une mauvaise classification ne coûte pas trop "cher" lors d'une modélisation ou simulation de précipitation. Plus précisément, ces régimes, s'ils sont utilisés pour conditionner un modèle statistique de précipitation, peuvent avoir des capacités pour les intensités moyennes mais peuvent grandement sous-estimer la variabilité.

Principales conclusions sur k-means(w) et EM(w)

Ces deux méthodes ont montré des caractéristiques locales bien discriminées d'un régime à un autre, particulièrement pour les précipitations extrêmes avec EM(w). Néanmoins, leur % de mauvaise classification est meilleur que celui de CCM. Par ailleurs, une mauvaise classification de nouveaux champs entraine un faible coût d'erreur de simulations, comparable aux k-means et EM classiques. EM(w) semble donc un excellent compromis : il permet une exploration fine des liens entre échelles et est tout fait pertinent en contexte de modélisations conditionnelles.

Pour terminer, je tiens à préciser que toutes les analyses et fonctions de coûts développées pour ces travaux ont été regroupées dans un package R nommé "CCM". Celui-ci est évidemment librement téléchargeable depuis le site web du "Comprehensive R Archive Network" (CRAN) sur http ://cran.r-project.org/, ou depuis mon propre site.

Cette étude [201] a ainsi fourni des outils pour aider les utilisateurs à choisir la méthode de clustering la plus adaptée selon le but à atteindre et l'utilisation des régimes de temps.

2.6 Deux ou trois mots de bilan sur les régimes de temps

Lors des sections de ce chapitre, j'espère avoir montré que les régimes de temps représentent une façon simple mais efficace de caractériser certaines formes de variabilité de l'atmosphère. Si de très nombreuses études ont porté sur les régimes pour différentes régions du globe, je pense que ces régimes restent néanmoins sous-employés, par exemple dans un contexte d'évaluation de modèles climatiques.

Dans ce dernier contexte, mes apports sont un premier pas vers une utilisation plus quantitative des RTs et surtout de leurs incertitudes et variabilités, que ce soit dans un contexte de régimes "classiques" ou "saisonniers". Pour cela, l'approche par modèle de mélange, grâce à l'algorithme EM, est un outil pertinent de *modélisation* des structures à grande échelle. Reposant sur des modèles statistiques (mélange de Gaussiennes par ex.), il fournit des bases théoriques pour évaluer la pertinence des régimes, leur probabilité d'occurrence, et quantifier leurs accords/désaccords avec des RTs issus de GCMs.

3 Downscaling statistique : vers des simulations à haute résolution spatiale

L'homme sait que le monde n'est pas à l'échelle humaine ; et il voudrait qu'il le fût.
André Malraux

La connaissance est une perpétuelle aventure à la lisière de l'incertitude.
Frank Patrick Herbert. Dune (1965)

En quelques décennies, notre regard sur le climat et la météo a fortement évolué. Passé du simple intérêt de savoir "comment s'habiller demain" à un véritable questionnement sur ce que nous coûterait - ou nous a déjà coûté - un changement climatique global, ce regard est aujourd'hui parfois inquiet des influences à plus ou moins longs termes du climat sur notre quotidien, nos sociétés et leurs économies. En effet, tout le monde garde à l'esprit nombre d'évènements météorologiques impressionnants - tels que tempêtes, cyclones, inondations dues à des précipitations intenses, ou à l'inverse, sécheresses particulièrement longues - qui possèdent des capacités de destructions et/ou de nuisances énormes. Cependant, sans même se référer spécifiquement à de tels évènements extrêmes (dont mes apports de modélisation sont présentés au chapitre 4), on estime qu'environ 30% des activités économiques mondiales sont affectés par les conditions climatiques qui règnent à la surface du globe ([91]). Les changements climatiques futurs peuvent donc considérablement impacter ces activités humaines. C'est dans ce contexte d'interrogations scientifiques et sociétales que le "Groupement d'experts Intergouvernemental sur l'Évolution du Climat" (GIEC ou IPCC en anglais pour "Intergovernmental Panel on Climate Change") a été créé dès 1988 à la demande du G7 (aujourd'hui G20), par deux organismes de l'Organisation des Nations Unies (ONU) : l'organisation météorologique mondiale (OMM) et le programme des Nations unies pour l'environnement (PNUE). Le but du GIEC est de faire une revue de l'état de l'art sur la compréhension du climat, de ses mécanismes physiques et de ses évolutions futures (Chapitre I des quatre premiers rapports, [90] [1]) mais aussi d'étudier leurs impacts et, si possible, comment les atténuer (Chapitres II et III des rapports, [91, 92]). Pour cela, un vaste exercice d'intercomparaison des simulations climatiques issues des principaux modèles de climat global (GCM) contraints par différents scénarios futurs d'émission de gaz à effet de serre ([90]) a été mené afin de documenter chaque rapport. L'un des problèmes récurrents de ces modèles climatiques est qu'ils fournissent des simulations sous forme de grilles dont la résolution spatiale de la maille (environ 250 km x 250 km) est trop faible pour être directement utilisées en entrées des modèles d'impacts (par ex., hydrologiques, écologiques, économiques, etc.) qui nécessitent des informations climatiques à beaucoup plus fine échelle pour

1. Ces rapports sont librement téléchargeables sur http ://www.ipcc.ch/. Le cinquième rapport est en cours.

être contraints de manière pertinente [2]. Il est donc nécessaire de “régionaliser” les simulations climatiques des GCMs, c’est à dire de générer des simulations pertinentes à haute résolution à partir des informations à grande échelle (schématisé en Fig. 3.1) : c’est toute la problématique du changement ou *réduction* d’échelle, couramment nommé “downscaling” dans ce Franglais si cher à toute discipline scientifique. Cependant, au-delà de cette motivation sociétale d’anticipation des

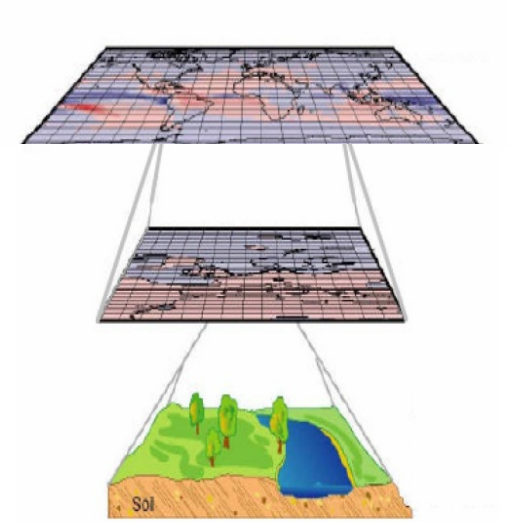


FIGURE 3.1: Schématisation du downscaling : toute la problématique consiste à générer des simulations pertinentes à haute résolution à partir d’informations à grande échelle.

changements climatiques potentiellement à venir, le downscaling est également nécessaire dans de nombreux autres contextes, et entres autres temporels. Appliqué à des données réanalysées caractérisant le climat actuel ou récent, le downscaling permet des études de processus physiques et atmosphériques à fine(s) échelle(s) améliorant notre compréhension du système climatique. Dirigé par des simulations caractérisant des climats lointains (par ex., mille dernières années, dernier maximum glaciaire, stade 3, etc.), ce “paléo-downscaling” offre, par exemple, l’opportunité de réaliser des comparaisons modèles - données aux échelles spatiales appropriées. Ces trois contextes temporels, passé - présent - futur, posent des questions différentes quant au développement des approches de downscaling, leurs conditions d’applicabilité, leur robustesse en climats différents du présent. Lors de ces dernières années, j’ai travaillé sur ces trois aspects temporels.

2. Dans tout ce document, nous parlerons de “grande” échelle ou d’échelle “large” pour une faible résolution et de “petite” échelle pour une fine ou haute résolution.

3.1 Différentes grandes familles et philosophies d'approches

Deux grandes familles d'approches existent pour répondre aux besoins de downscaling. Historiquement, la première approche à avoir vu le jour est appelée régionalisation "dynamique". Celle-ci repose sur des modèles climatiques régionaux (RCMs) qui peuvent être perçus comme une version régionale des modèles globaux, au sens où ils cherchent à résoudre les principales équations physiques de la dynamique de l'atmosphère mais pour des régions données et à des résolutions spatiales évidemment beaucoup plus fines (désormais de 50 à environ 5 km) que celles des GCMs (voir par ex., [110, 155] pour des revues récentes). Ces RCMs ont donc l'avantage de résoudre explicitement une partie des processus physiques impliqués aux échelles sous-maille des GCMs et sont donc cohérents du point de vue météorologique et hydrologique. Cependant, bien que les capacités informatiques ne cessent d'augmenter, les moyens et coûts de calculs nécessaires à la résolution de ces équations régionales et leur complexité également croissante (du fait de l'augmentation des paramètres désormais pris en compte) font que les RCMs ne sont encore appliqués que sur des régions et périodes de temps limitées ([7]). Un autre corollaire du coût des RCMs est que les simulations d'ensembles, nécessaires pour s'affranchir de la variabilité météorologique sont encore relativement prohibitives. Toutefois, depuis quelques années, un effort international assez considérable a été mis en oeuvre pour intercomparer ces différents modèles régionaux, par exemple :

- le programme NARCCAP (North America Regional Climate Change Assessment Program [3]) sur l'Amérique du Nord,
- le projet STARDEX (Statistical and Regional dynamical Downscaling of Extremes for European regions, FP5) essentiellement focalisé sur la régionalisation des événements extrêmes,
- le projet PRUDENCE (Prediction of Regional scenarios and Uncertainties for Defining EuropeaN Climate change risks and Effects, FP5) concernant la prédiction de scénarios régionaux de changements climatiques en Europe,
- suivi par le projet ENSEMBLE (FP6) portant sur les méthodes d'*ensemble* pour les modèles climatiques globaux et régionaux (dynamiques),
- ou plus récemment, l'exercice encore en cours CORDEX (COordinated Regional Downscaling EXperiment) portant sur la régionalisation de variables climatiques sur une douzaine de régions dans le monde.

Dans la plupart de ces exercices, les intercomparaisons sont réalisées (essentiellement, pas toujours uniquement) entre RCMs. Or, une autre approche de downscaling existe depuis maintenant plusieurs années et prend de plus en plus d'ampleur dans la communauté climatique et des études d'impacts. Il s'agit de l'approche de régionalisation "statistique", à laquelle j'ai contribué ces dernières années. Comme son nom l'indique, cette approche ne cherche pas à résoudre explicitement les équations de la dynamique atmosphérique régionale mais est basée sur des modélisations statistiques des relations et corrélations qui existent entre données à grande échelle (par ex., provenant de réanalyses ou de GCMs) et des variables observées à fine échelle

3. http ://www.narccap.ucar.edu/

(par ex., à une station météorologique). Ainsi, cette approche permet généralement de simuler des données locales de manières très rapides, et ceci sans moyens de calculs trop coûteux. Par ailleurs, ce faible coût ainsi que la formulation statistique des liens entre échelles, permet également à cette approche (selon la méthode utilisée, voir plus loin) de faciliter la modélisation des incertitudes associées aux projections régionales, ainsi que leur propagation au cours du processus de downscaling (par ex., [166, 167]). En pratique, l'approche statistique de downscaling (SDS dans toute la suite) recouvre une grande diversité de méthodes dont les philosophies peuvent être très différentes. Généralement, on considère trois ou quatre grandes familles de méthodologies afin de caractériser la régionalisation statistique : les méthodes dites par "fonctions de transfert", celles par "générateurs de temps", et celles dites par "types de temps" (Voir par ex., [210] qui ont été les premiers à définir ce découpage). Certaines méthodes issues de géostatistiques (par ex., le krigeage) participent aussi à ce downscaling. L'ensemble de ces familles est schématisé en Fig. 3.2.

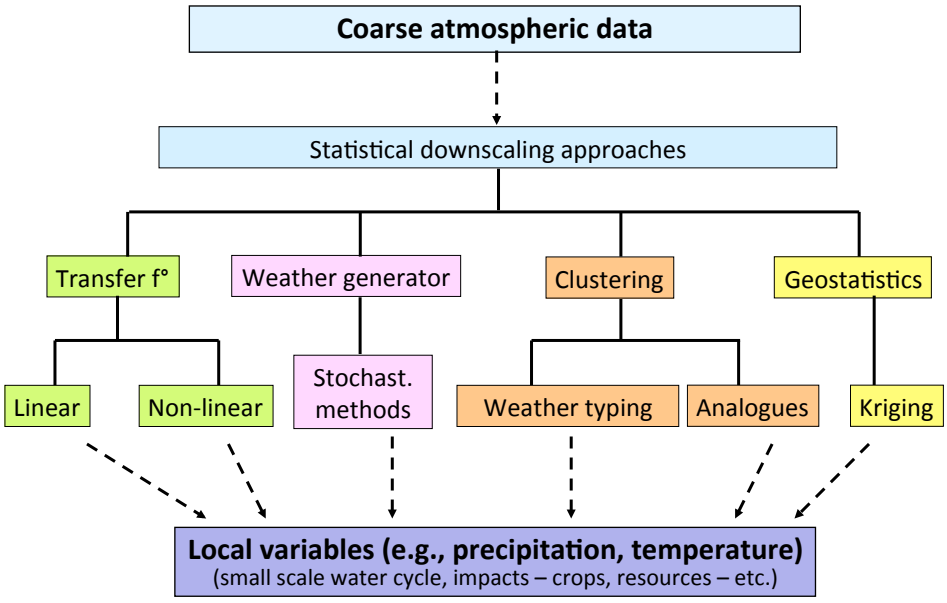


FIGURE 3.2: Les principales grandes familles de méthodologies de downscaling statistique.

Les fonctions de transfert ont pour but de "traduire" directement les données à grande échelle en valeurs locales grâce à des régressions linéaires ou non-linéaires. Étant donnée x, l'information à grande échelle, la réponse locale régionalisée y est estimée par une fonction $\hat{y}(x)$ qui est, habi-

tuellement, une estimation de $E[Y|X = x]$ ($\hat{y}(x)$ peut aussi correspondre à l'estimation de divers quantiles dans le cas de régression de quantiles, voir, par ex., [23, 58]). Ces fonctions de transfert peuvent être : des régressions linéaires sur données "brutes" ou issues d'analyses en composantes principales ou d'analyses en corrélations canoniques (par ex., [207, 88, 208, 22, 70]) ; des modèles paramétriques non-linéaires tels que des régressions polynomiales ([78, 157]) ; des régressions non-paramétriques basées sur des splines ou modèles additifs généralisés ([196, 159, 114]) ; et des réseaux de neurones artificiels ([173, 24, 75, 89, 66]).

Les méthodes par "générateurs de temps" (ou "générateurs stochastique") permettent de *simuler* des variables climatiques (telles que les précipitations ou la température) en se basant sur des fonctions de densité de probabilités (pdf) (par ex., voir [166, 167, 211]). Ces générateurs sont calibrés (c-à-d., leurs paramètres sont estimés) de manière à ce que les simulations reproduisent certaines propriétés statistiques des observations locales ([150, 151, 135, 167]). Dans un contexte de downscaling, les générateurs de temps simulent une variable locale Y conditionnellement à l'information à grande échelle X, en construisant un modèle pour la densité conditionnelle de $Y|X = x$. Pour atteindre ce but, initialement, les paramètres du générateur (par ex., les paramètres de la pdf) caractérisant les observations étaient proportionnels à ceux du générateur caractérisant la grande échelle. Ainsi, si les paramètres de ce dernier évoluaient dans le futur, ceux décrivant (et permettant de simuler) les valeurs de la petite échelle évoluaient proportionnellement (par ex., [208, 212]). Depuis quelques années, cette approche a été progressivement remplacée par une approche ou ce sont les paramètres de la pdf conditionnelle qui sont directement des fonctions d'une information grande échelle appropriée (et non des fonctions des paramètres d'un générateur à grande échelle). Cette dernière peut être des régimes de temps en Amérique du Nord (par ex., [199]), l'indice d'oscillation Nord-Américaine (NAO, par ex., [218]), ou tout autre variable climatique ou atmosphérique (par ex., [6, 27], ou [213, 212] pour deux reviews). Ainsi, les changements de variabilité des variables à grande échelle (les prédicteurs) sont "transférés" aux paramètres de la pdf conditionnelle à l'échelle locale, faisant alors évoluer dans le temps les simulations à fine échelle (par ex., [197]).

Les méthodes dans la catégorie dite par "types de temps" cherchent tout d'abord à classifier de manière non-supervisée (c-à-d., clustering) les situations (généralement journalières) de circulation atmosphérique à grande échelle en régimes (ou types) de temps récurrents. Ces méthodes supposent ensuite que chaque type de temps donné est toujours associé aux mêmes conditions météorologiques locales (par ex., [87, 195, 15, 136]). La composante de mes travaux liée à la définition des ces régimes de temps (en contexte de régionalisation ou plus généralement d'étude de la variabilité grande échelle du climat) est présentée dans le chapitre 2. En contexte de downscaling, le typage de temps est habituellement considéré comme une étape de pré-traitement des informations à grande échelle en les "condensant" en variables nominales et sert à conditionner des méthodes par fonctions de transfert (par ex., [89]) ou par générateurs stochastiques de temps (par ex., [162, 199, 197]).

La dernière catégorie de méthodes, regroupées sous le terme de "géostatistiques", est généralement davantage employée pour faire du "upscaling" plutôt que du downscaling, grâce à des outils de type krigeage . Toutefois, cette approche peut être tout à fait pertinente en contexte de downscaling, par

exemple, en interpolant spatialement des simulations réalisées à différentes stations par une autre méthode de downscaling - cela permet ainsi de générer des valeurs locales même à des stations où nous n'avons pas d'observations -, ou en interpolant les *paramètres* de modèles statistiques de régionalisation calibrés en différentes stations - cela permet ainsi de disposer d'un *modèle* même à des stations où nous n'avons pas d'observations. Notons toutefois que certaines approches emploient le krigeage, non pas comme post-traitement de valeurs locales simulées ou de paramètres locaux mais comme approche directe de downscaling. Par exemple, [11] réalisèrent un krigeage de valeurs mensuelles de précipitation dans un espace bivarié constitué des deux premières composantes principales de champs de pressions au niveau de la mer (SLP) sur la région Nord-Atlantique qui constitutent alors les prédicteurs de ce downscaling.

Il est intéressant de noter que d'autres "découpages" des méthodes statistiques existent. Par exemple, dans l'article de review [118] auquel j'ai participé, la distinction est faite, non pas sur les méthodologies statistiques à proprement parler, mais davantage sur la manière de les utiliser et plus précisément sur les données à grande échelle nécessaires à leur fonctionnement. Plus précisément, les approches nécessitant des données *réelles*[4] à grande échelle sont dites "perfect-prog"- par exemple les méthodes de fonctions de transfert appliquées sur des données journalières -, celles pouvant être directement calibrées sur des données *simulées* à grande échelle[5] - par exemple certains générateurs stochastiques - sont dites être de type "Model Output Statistics" (MOS). Ce dernier nom est généralement employé pour les méthodes statistiques de "correction de biais" ([44, 144, 118]). En pratique, le downscaling statistique et la correction de biais sont très proches l'un de l'autre et la frontière les séparant est souvent fonction du but à atteindre et du contexte de l'application (par ex., [182]). Mes apports dans ce contexte de méthodes MOS seront brièvement discutés en sections 3.4 et 4.2.

Quelles hypothèses en changement climatique ?

De manière générale, il est bon de rappeler un fait totalement incontournable : pour pouvoir calibrer un modèle statistique de régionalisation, il est indispensable de disposer de données observées à l'échelle à laquelle on souhaite réaliser des simulations. Autrement dit, il n'est pas possible de réaliser un downscaling statistique à une résolution plus fine que celle des observations utilisées pour la calibration du modèle.

De plus, dans toutes ces approches statistiques, l'un des points clés est évidemment la sélection des prédicteurs (c-à-d., l'information à grande échelle) utilisés en entrée de chaque approche. Ce point n'est pas discuté dans ce manuscrit de HDR mais reste un point crucial à toute application pertinente et efficace d'une approche de régionalisation, d'autant plus si celle-ci est employée pour régionaliser des projections futures issues de GCMs. Il existe par ailleurs certaines hypothèses

4. ou considérées comme réelles comme les réanalyses
5. telles que des sorties de GCM

implicites que doivent respecter les méthodes de downscaling statistique pour être appliquées de manière pertinente en contexte de changement climatique :

1. les prédicteurs sont pertinents et modélisés correctement dans les GCMs ;
2. les prédicteurs représentent *complètement* le changement climatique ;
3. le modèle statistique - calibré en contexte de climat présent ou proche présent - reste valide en climat modifié.

Il est à noter que, même si les RCMs sont basés sur des équations représentant les processus physiques et la dynamique atmosphérique (supposant ainsi une plus grande robustesse dans des conditions de climat différent du présent), certains choix de simplification ou de paramétrisation de ces équations et processus font que des hypothèses similaires sont sous-jacentes à l'utilisation de RCMs sous contraintes de changement climatique. Celles-ci peuvent être reformulées ainsi :

a. les informations à grande échelle utilisées par les RCMs pour effectuer un downscaling dynamique (conditions au bord du domaine, nudging, etc.) sont modélisées correctement dans les GCMs ;
b. ces informations représentent *complètement* le changement climatique ;
c. les RCMs et leurs modélisations des processus régionaux - évalués en climat présent - restent valides pour un climat différent.

L'évaluation de ces hypothèses implicites n'est que rarement effectuée et évidemment uniquement partiellement quand c'est le cas - que ce soit pour les approches statistiques ou dynamiques - avant de régionaliser des sorties futures de GCM. Certaines études ont cependant essayé d'évaluer la robustesse des approches statistiques dans un tel contexte. Par exemple, certains standards ont été définis pour l'évaluation de la capacité des modèles SDS à reproduire des observations historiques quand ils sont calibrés puis guidés par des données de réanalyses ([86, 152, 199]) ou guidés par des simulations "historiques" de GCMs (par ex., [31, 209, 30]). D'autres approches ont testé la validité de l'hypothèse 3 de robustesse dans le temps en comparant des projections futures de SDS avec celles de GCMs (par ex., [57]) ou de RCMs (par ex., [217, 21, 75]). Bien que je ne revienne pas spécifiquement sur ces travaux dans ce manuscrit par la suite, je tiens à signaler que - en collaboration avec des collègues de l'Université de Chicago (M. Stein), de la Texas Tech University (K. Hayhoe), et du Illinois State Water Survey (X.-Z. Liang) - j'ai moi-même participé à définir un standard d'évaluation de la capacité d'approches SDS à capturer le signal du changement climatique et à le restituer correctement à plus haute résolution (voir [200]). Pour cela, en plus d'une évaluation de SDS en contexte de climat historique (sur réanalyses et GCM), nous avons développé une méthodologie qui considère les données de RCMs simulées par un GCM historique comme étant des *pseudo-observations*. Nous pouvons alors évaluer les simulations statistiques régionales futures qui doivent se rapprocher au plus près des simulations dynamiques futures. Appliquée selon différents scénarios d'émissions de gaz à effet de serre, cette évaluation permet ainsi de voir les limites potentielles de l'approche testée selon le degré des changements climatiques modélisés dans le GCM. Bien que cette méthodologie possède évidemment diverses contraintes et hypothèses fortes - par exemple, le GCM modélise correctement le changement climatique ; ou le

lien entre grande et petite échelles modélisé par le RCM est correct - elle offre une vision du degré de confiance que l'on peut avoir dans les projections régionales futures, à mon sens nécessaire avant toute étude d'impact associée. Cette approche sera par ailleurs employée dans le cadre du projet international PLEIADES démarrant officiellement en Octobre 2011 pour l'évaluation des changements de précipitation en Europe au cours du 21ème siècle.

Pour terminer ces mots d'introduction au downscaling, j'aimerais souligner que les deux approches de régionalisation dynamiques et statistiques ne doivent pas être considérées comme des concurrentes. En fournissant des informations qui peuvent être utiles à l'autre, ces approches sont clairement complémentaires. En effet, les RCMs permettent des simulations où les dépendances spatiales sont implicitement représentées, par exemple lorsqu'un front météorologique génère des précipitations à des stations distantes. Cette caractéristique peut ainsi aider à la sélection des prédicteurs les plus appropriés (c-à-d., avec une interprétation physique) pour guider les modèles statistiques. Réciproquement, les modèles statistiques peuvent fournir une quantification des contributions des différents prédicteurs considérés, et ainsi donner une information pertinente sur les liens entre grande et petite échelles à potentiellement inclure et modéliser dans les simulations dynamiques.

Dans la suite de ce chapitre, je ne parlerai que de downscaling *spatial* et non *temporel*. En effet, bien que j'ai récemment eu l'occasion de contribuer ([214]) à ce contexte temporel (c-à-d., essayer de passer par exemple de données au pas de temps journalier à des données au pas de temps horaire) en développant avec P. Willems (Katholieke Universiteit Leuven) une méthode dite par "perturbation factor" [6], ce type de contexte n'est que récent dans mes travaux et ne constitue donc pas le coeur de mon expertise. Toutefois, les perspectives de développements méthodologiques en changement d'échelle temporelle sont particulièrement intéressantes. Ainsi, dans la suite de ce chapitre, la section 3.2 présente certains de mes apports concernant le développement de modèles stochastiques de downscaling. Mes apports reliés aux fonctions de transfert pour le downscaling et, plus généralement, pour la modélisation de variables environnementales, sont illustrés en section 3.3. Enfin, mes apports en contexte de "Model Output Statistics" sont discutés en sections 3.4. Des extensions aux cas de valeurs extrêmes de certains de ces apports seront présentés au chapitre suivant.

6. Cette méthode peut être perçue comme une méthode de type delta dans un contexte "probabiliste" puisque ne faisant pas uniquement évoluer la moyenne des valeurs, mais l'ensemble de la distribution statistique au travers de ses différents quantiles.

3.2 Downscaling des précipitations par approches stochastiques non-homogènes

Une partie de mes travaux sur le downscaling stochastique s'est focalisée sur des modèles de densité de probabilité (pdf) *non-homogènes*. En effet, la notion classique de pdf repose sur l'hypothèse de stationnarité des données qu'elle modélise, c-à-d. que ces données suivent la même loi statistique, sur toute leur période d'observation, de modélisation et de simulation. Modéliser une densité paramétrique f associée à une variable aléatoire Y telle que la température ou la précipitation, consiste alors à estimer ses paramètres (par méthode du maximum de vraisemblance, ou des moments, etc.) à partir d'un échantillon $(y_1, ..., y_n)$ de n réalisations (par ex., n observations journalières) de cette variable Y. Dans ce contexte classique, seules ces n réalisations sont utilisées pour déterminer les paramètres de la pdf qui sont alors des constantes.

Or, les données climatiques et environnementales ne sont pas toujours stationnaires, leur distribution statistique évolue parfois dans le temps ou en fonction du contexte météorologique synoptique. Par exemple, il est évident que la probabilité d'occurrence de précipitation journalière en un lieu donné ne sera pas la même en conditions anticycloniques ou en conditions dépressionnaires. De plus, si les fréquences des ces deux types de conditions, disons entre 1990 et 2010, ne sont pas les mêmes qu'entre, disons 2050 et 2070, les distributions globales [7] de probabilité d'occurrence de pluie en ce lieu ne seront pas les mêmes pour ces deux périodes de 20 ans. Il est donc souvent nécessaire – et particulièrement en contexte de changement climatique – de tenir compte d'informations "externes" pour définir un modèle stochastique pertinent, capable d'évoluer dans le temps et/ou en fonction de contraintes atmosphériques ou environnementales. Dans l'approche stochastique par modèle non-homogène, les paramètres ne sont pas constants mais dépendent de manière *analytique* d'autres données ou variables aléatoires X, appelées *prédicteurs* ou *covariables*, qui, en contexte de downscaling, sont des données à grande échelle qui peuvent caractériser la circulation atmosphérique ou la thermodynamique (par ex., hauteurs géopotentielles, humidité, température, etc.), ou même des régimes de temps définis sur une vaste région. L'introduction de ces covariables permet ainsi de faire évoluer la densité en fonction des valeurs quotidiennes prises par ces prédicteurs : on parle alors de *densités conditionnelles* de la variable Y sachant la réalisation x du prédicteur X. L'idée sous-jacente est que si la fréquence, l'intensité, ou de manière générale les propriétés statistiques des prédicteurs évoluent dans le temps, alors les paramètres – et donc la pdf associée $f(y|X = x)$ – modélisés conditionnellement à ces prédicteurs vont évoluer en conséquence. Notons ici qu'on s'appuie donc fortement sur les hypothèses 1 et 2 définies précédemment : les prédicteurs doivent représenter complètement le changement climatique d'intérêt et doivent être simulés correctement à grande échelle.

Par ailleurs, bien que ces modèles soient spécifiquement développés pour tenter de contourner ou au moins de réduire l'hypothèse de stationnarité des approches statistiques, dans la suite, je parlerai de modèles non-homogènes (MNH) et non de modèles non-stationnaires. En effet, bien

7. c-à-d., calculées sur l'ensemble des jours d'une période donnée

que ces MNHs font évoluer les pdf et leurs propriétés dans le temps, ils reposent malgré tout encore sur une hypothèse implicite de stationnarité qu'il est bon de garder à l'esprit : la stationnarité du lien entre grande et petite échelles, ici représentée par la fonction – ou le conditionnement – liant les paramètres de la pdf aux prédicteurs. Si la pdf peut alors être considérée comme non-stationnaire, cette fonction de lien, elle, repose encore sur la notion de stationnarité, au sens où elle ne peut pas évoluer dans le temps : elle est fixée par la calibration [8].

Mes travaux dans ce domaine sont illustrés successivement par les deux articles [199] et [27], basés sur la philosophie décrite ci-dessus avec deux techniques relativement différentes. Tous deux portent sur de la régionalisation de valeurs journalières de précipitation. En effet, bien que d'autres variables peuvent bénéficier des apports de la modélisation stochastique, la précipitation reste la variable clé en météorologie et climatologie et donc en régionalisation. Cette variable est nécessaire à de très nombreuses études et projections, allant de la prévision météorologique ou l'assimilation de données, en passant par nombre d'études et de modèles d'impacts portant par exemple sur l'hydrologie, les modèles de rendements agricoles, ou la biodiversité. Or, due à une très forte variabilité spatiale et temporelle, la précipitation reste très complexe à modéliser et à simuler aux échelles régionales, voire locales, que ce soit par approches dynamique ou statistique.

3.2.1 Conditionnement par régimes régionaux de précipitation

Lors de l'étude [199] réalisée en collaboration avec Michael Stein (University of Chicago) et Katharine Hayhoe (à l'époque à l'University of Illinois at Urbana-Champaign), le but était la régionalisation des précipitations journalières sur 37 stations météo réparties relativement uniformément dans l'état d'Illinois aux USA. Nous disposions de valeurs journalières (hauteurs géop., hum. spéc., écarts au point de rosée) de réanalyses NCEP/NCAR ([99]) et d'observations locales pour deux périodes de 20 ans : 1980-1999 – utilisée pour la calibration – et 1952-1971 – utilisée pour la simulation et l'évaluation [9]. Nous avons alors développé une alternative à l'approche par modèle de Markov caché de [6]. La méthodologie employée, et nommée NSWT pour "Non homogeneous Stochastic Weather Typing", est schématisée en Fig. 3.3. NSWT peut être considérée comme une méthode *hybride* entre les approches par "types de temps" et "générateurs stochastiques" puisque qu'elle consiste tout d'abord à définir des régimes de temps qui vont servir à conditionner les modèles stochastiques par la suite. Ces régimes peuvent être modélisés, soit par des données à grande échelle comme c'est traditionnellement réalisé (voir chapitre 2), soit par des données de précipitation observées quotidiennement sur nos 37 stations. En effet, si notre but est la modélisation des précipitations, il est logique de penser qu'un conditionnement par des régimes directement définis sur cette variable devraient permettre une modélisation plus efficaces que par des régimes de circulation potentiellement moins discriminants au niveau des stations. Ainsi, en plus de la modélisation développée, l'un des objectifs de l'étude [199] était de comparer la qualité des

8. et la sélection de modèle statistique
9. Calibration et évaluation ont également été inversés pour tests et fournissent des résultats de qualité tout à fait équivalente.

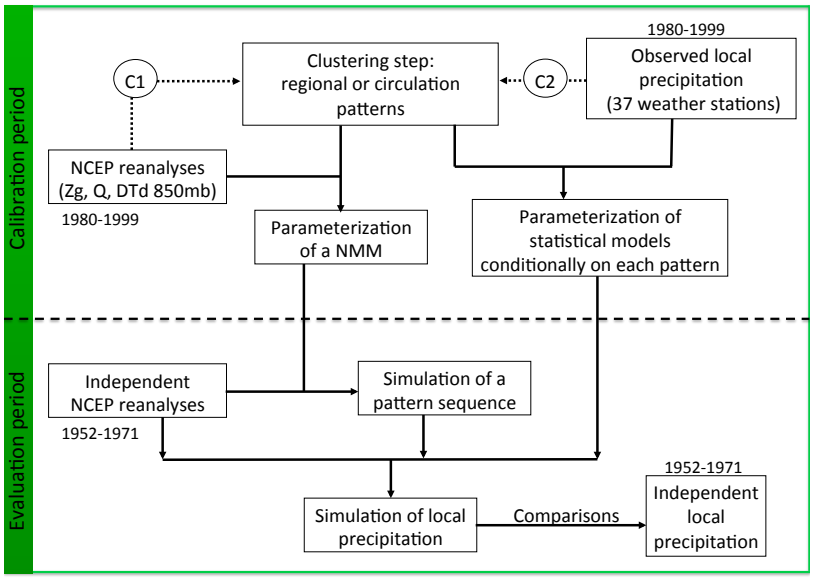


FIGURE 3.3: Schéma de calibration et d'évaluation de l'approche NSWT développé dans [199].

simulations locales lorsqu'elles sont conditionnées par l'un ou l'autre des types de régimes. Sans rentrer dans les détails, les régimes de circulation ont été définis à partir de données de réanalyse NCEP/NCAR à l'aide de l'algorithme EM (discuté au chapitre 2). Quant aux régimes *régionaux* de précipitation, une méthode de clustering de type "hiérarchique ascendante" (HAC) avec critère de Ward ([205]) a été utilisée, pour laquelle nous avons développé une distance originale permettant de travailler spécifiquement sur des valeurs journalières de précipitation [10]. Les détails ne sont pas fournis ici mais le lecteur intéressé par cette étape de clustering trouvera toutes les informations dans [199]. Toutefois, à partir d'une nouvelle série temporelle de champs de variables à grande échelle (par ex., NCEP ou GCM), il est nécessaire d'être capable de simuler une séquence de tels régimes pour réaliser, par exemple, des évaluations en climat présent, ou des projections en climat futur. Si les régimes sont définis uniquement à partir de telles données, la méthode de clustering possède généralement sa propre technique d'attribution d'un nouveau jour à l'un des clusters. Cependant, les régimes de précipitation ont été définis sur des observations. Celles-ci ne peuvent évidemment pas être utilisées en contexte de projection. C'est pourquoi nous avons mis en place un modèle de Markov non-homogène (NMM) du premier ordre, permettant de simuler l'occurrence de régimes journaliers en tenant compte de l'information grande échelle disponible et qui a été utilisé pour les deux types de régimes dans un souci d'intercomparabilité : sachant que le régime était dans l'état k (parmi K possibles) au temps $t-1$ ($R_{t-1} = k$) et que nous disposons de la

10. c-à-d., prenant en compte la caractérisation des jours secs, la différentiation des distances entre jours à faibles et plus fortes pluies, et les variabilités spatiales et temporelles.

réalisation x du prédicteur à grande échelle X au temps t ($X_t = x$), l'occurrence de chaque régime l ($l = 1, ..., K$) est définie par la probabilité

$$Pr(R_t = l | R_{t-1} = k; X_t = x) = \gamma_{kl} \times \frac{1}{C_{kl}} \exp(-\frac{1}{2}(x - \mu_{kl})\Sigma^{-1}(x - \mu_{kl})\prime), \tag{3.1}$$

où γ_{kl} correspond à la probabilité de base de transition du régime k vers le régime l – c-à-d., correspond à la probabilité de transition de k vers l si l'on ne tient pas compte de l'information apportée par $X_t = x$ comme dans un modèle de Markov homogène classique –, et où le terme suivant correspond à la densité de X calculée en x, avec $1/C_{kl}$ une constante d'intégration à 1 de cette densité, fonction de la transition de k vers l, et μ_{kj} et Σ respectivement le vecteur moyennes et la matrice de variance-covariance de X. Une fois les paramètres de ce NMM estimés (par ex., par maximum de vraisemblance), il est aisé de simuler l'occurrence de l'un des régimes pré-définis sachant $R_{t-1} = k$ et $X_t = x$ et donc $Pr(R_t = l | R_{t-1} = k; X_t = x)$. Ces régimes (qu'ils soient de circulation ou de précipitation) fournissent une première information au modèle de SDS : pour un jour donné, la connaissance du régime associé (simulé) à ce jour permet de conditionner les paramètres du modèle statistique pour la pluie journalière. Traditionnellement, la loi des intensités positives de pluie est représentée par une distribution Gamma ou log-normale (par ex., [215, 101, 218, 144, 33]). Ici, le modèle est la densité f de l'intensité de précipitation r modélisée conditionnellement au régime R_t au temps t également à l'information $X_t = x$. Cette pdf f correspondait à un produit sur l'ensemble des 37 stations météo de densités paramétriques, elles-mêmes dépendantes de la station i considérée :

$$f(r | R_t = k; X_t = x) = \Pi_{i=1}^{37} [g(r^i | \theta_{ki}; R_t = k; X_t = x)], \tag{3.2}$$

où $r = (r^1, ..., r^{37})$ est le vecteur des intensité de précipitation aux 37 stations et avec :

$$g(r^i | \theta_{ki}; R_t = k; X_t = x) = (p_{ki}(x)\ g(r^i | \alpha_{ki}; \beta_{ki}))^{1_{\{r^i > 0\}}}\ (1 - p_{ki}(x))^{1_{\{r_t^i = 0\}}}, \tag{3.3}$$

où

- r^i est une intensité de précipitation à la station i,
- $g(.|\alpha; \beta)$ est la pdf de la distribution Gamma de paramètres α et β,
- 1_A est une fonction indicatrice valant 1 si la condition A est vérifiée et 0 sinon,
- et $p_{ki}(x)$ est la probabilité d'occurrence de précipitation dans le régime k, à la station i et sachant l'information à grande échelle $X = x$.

Cette dernière probabilité d'occurrence a été testée constante (c-à-d., uniquement dépendante du régime k et de la station i) mais les résultats sont apparus assez peu réalistes. Ainsi, un modèle de régression logistique (par ex., [73]) a été mis en oeuvre :

$$p_{ki}(x) = \frac{\exp(x\prime\ \lambda_{ki})}{1 + \exp(x\prime\ \lambda_{ki})}, \tag{3.4}$$

où λ_{ki} correspond à un vecteur de coefficients à déterminer. Ainsi, dans les équations (3.2) et (3.3), θ_{ki} est le vecteur des paramètres à estimer, $\theta_{ki} = (\alpha_{ki}, \beta_{ki}, \lambda_{ki})$. Dans la formulation (3.2), le produit des densités paramétriques signifie que les précipitations aux 37 stations sont indépendantes.

Toutefois, on peut remarquer que cette indépendance n'est pas globale, au sens où elle est conditionnée par la connaissance du régime (circulation ou précipitation) : on parle alors d'*indépendance conditionnelle* à l'information disponible à grande échelle. La structure de dépendance modélisée est placée sur les paramètres des distributions marginales, et non sur les réalisations elles-mêmes. Pour six stations illustratives, la Fig. 3.4 présente des comparaisons en termes de probabilités de longueur de périodes sèches ou pluvieuses, pour les deux types de conditionnement. La Fig. 3.5

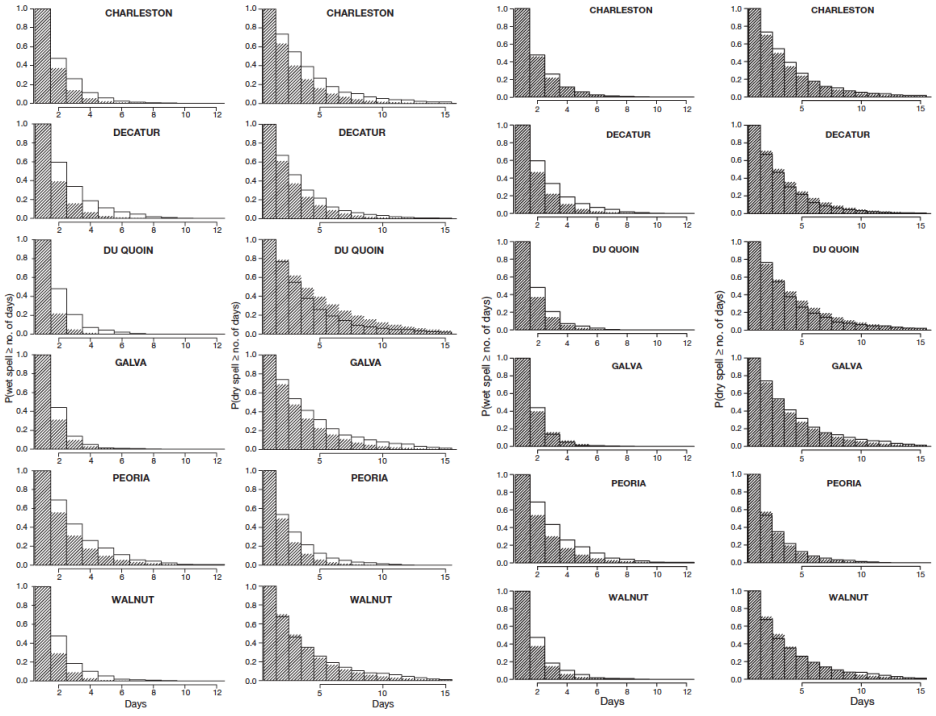
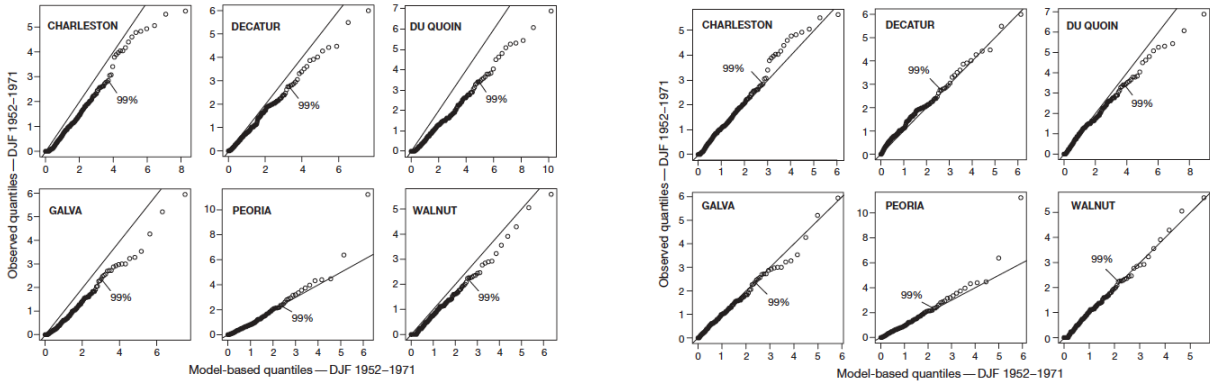


FIGURE 3.4: Pour six stations illustratives, probabilités de jours consécutifs pluvieux (colonnes 1 et 3) et secs (colonnes 2 et 4) à partir des simulations issues du conditionnement par régimes de circulation (2 colonnes de gauche) ou par régimes de précipitation (2 colonnes de droite). Les histogrammes grisés correspondent aux probabilités des simulations, les histogrammes blancs à celles des observations.

présente, pour les mêmes six stations illustratives, des comparaisons en termes de "quantile-quantile plots" (QQplots) de simulations vs. observations lorsque les modèles statistiques sont conditionnés par des régimes de circulation ou de précipitation. Ces comparaisons permettent d'évaluer la bonne reproduction de caractéristiques clés des précipitations régionalisées par rapport aux observations, mais également de de voir que les modèles statistiques fournissaient des simulations locales de bien meilleure qualité lorsqu'ils étaient conditionnés par des régimes de précipitation plutôt que par des régimes classiques de circulation. Cette étude a ainsi permis de



(a) (b)

FIGURE 3.5: Pour six stations illustratives, QQplots des simulations issues du conditionnement (a) par régimes de circulation (b) par régimes de précipitation.

mettre en avant une modélisation non-homogène, originale et plus efficace qu'une approche "classique" basée uniquement sur un conditionnement atmosphérique.

Je souhaite préciser qu'un package R nommé "NHMixt" et contenant les différentes fonctions R nécessaires pour calibrer un modèle de mélange non-homogène est disponible, soit depuis mon site web, soit sur demande. Ce package devrait prochainement être enrichi de diverses fonctionnalités (estimation de quantiles, simulations, etc.).

3.2.2 Conditionnement par prédicteurs continus et réseaux de neurones

La deuxième étude de développement d'un modèle statistique non-homogène de downscaling a été réalisée en collaboration avec J. Carreau ([27]). Pour celle-ci, nous avons travaillé sur des données issues de la base du "European Climate Assessment & Datasets" (ECA&D [11], [183]) et contenant des valeurs journalières de précipitation en diverses stations météorologiques de la région Méditerranéenne Française pour la période du 1er janvier 1959 au 31 décembre 2004. La philosophie de cette approche est relativement similaire à la précédente, au sens où la densité des précipitations journalières est modélisée conditionnellement à des informations à grande échelle. Cependant, ici, ces informations ne sont pas fournies sous la forme de régimes de temps discrets mais sous la forme de valeurs de pression au niveau de la mer (SLP) issues des données

11. http ://eca.knmi.nl/

de réanalyses NCEP/NCAR pour une région de 6×6 points de grille entourant les stations [12]. Bien que d'autres variables atmosphériques eussent pu être sélectionnées, seul ce jeu de SLP a été retenu. En effet, dans cette étude, notre but n'était pas de déterminer le meilleur ensemble de prédicteurs mais plutôt d'illustrer et de comparer les performances et avantages de quatre modèles non-homogènes de régionalisation statistique. Dans ce contexte, la SLP était donc appropriée car généralement considérée comme l'un des principaux prédicteurs de la précipitation (par ex., [69]).

De manière générique, le modèle développé pour représenter conjointement l'occurrence et l'intensité est le suivant :

$$\phi(y;\psi) = \underbrace{(1-\alpha)\delta(y)}_{\text{intensité}=0} + \underbrace{\alpha\phi_0(y;\psi_0)}_{\text{intensité}>0}, \tag{3.5}$$

où α est la probabilité d'occurrence de pluie ; $\delta(\cdot)$ est la fonction de Dirac qui est telle que $\int_{-\infty}^{\infty} f(z)\delta(z-a)dz = f(a)$ et $\delta(z-a) = 0$ pour $z \neq a$; $\phi_0(\cdot;\psi_0)$ est la densité de l'intensité > 0 de paramètre ψ_0 ; et $\psi = (\alpha, \psi_0)$ est le vecteur de paramètres du mélange (3.5) . Dans [215], $\phi_0(\cdot;\psi_0)$ était la densité Gamma. À la place, nous avons proposé d'employer des modèles de mélange :

$$\phi_0(y;\psi_0) = \sum_{j=1}^{m} \pi_j f(y;\theta_j), \tag{3.6}$$

où $f(\cdot;\theta_j)$ est une densité de paramètre (vecteur) θ_j, π_j est le "poids" de la composante j, et ψ_0 est le vecteur regroupant tous les paramètres du mélange $(\pi_1,...,\pi_m,\theta_1,...,\theta_m)$. Pour cette formulation (3.6), nous avons cherché à comparer trois types de distribution $f(;\theta)$ pour représenter les intensités positives de précipitation :

- une Gaussienne tronquée,
- une log-Normale,
- une Pareto "hybride".

Ces trois mélanges de la forme (3.6) ont été comparé au modèle de [215] – utilisé comme base de référence – défini comme (3.5) avec $\phi_0(\cdot;\psi_0)$ la densité Gamma.

Le but de cette étude était double : (1) proposer un nouveau modèle non-homogène et (2) qui était capable de représenter l'*ensemble de la distribution* des pluies, c-à-d. incluant une modélisation des extrêmes. Ce deuxième point (ainsi que la définition de la distribution de Pareto hybride) n'étant pas le sujet de ce chapitre, il sera discuté en section 4.3 du chapitre suivant. Pour ce qui est de la non-homogénéïté des modèles, contrairement à l'approche employée précédemment dans [199], les paramètres des mélanges de densités de l'équation (3.5) ne prenaient pas uniquement quelques valeurs possibles définies par les régimes et les stations, mais étaient des fonctions *continues*de nos prédicteurs x (c-à-d., les réalisations de SLP) : $\psi(x) = (\alpha(x), \pi_1(x),...,\pi_m(x), \theta_1(x),...,\theta_m(x))$. Or, si différentes fonctions de lien paramétriques auraient pu être mises en oeuvre (voir par ex., [218, 98]), diverses études ont montré que la relation liant variables atmosphériques et précipitations locales est non-linéaire (par ex., [80]). Nous avons alors

12. Diverses configurations et tailles de régions ont été testées : celle en 6×6 points de grille correspondait à un bon compromis permettant d'avoir une information régionale tout en capturant suffisamment de signal synoptique.

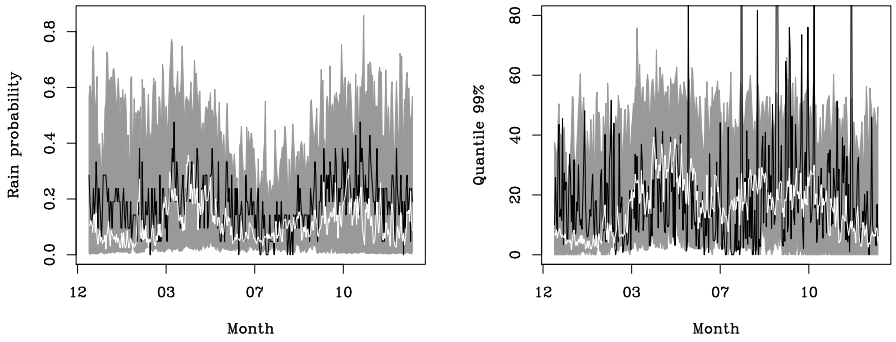
décidé de réaliser ce conditionnement continu par réseaux de neurones artificiels “feed-forward” à une couche cachée (voir par ex., [12, 13]). En effet, ces réseaux de neurones sont des modèles non-paramétriques très flexibles qui peuvent – en principe – approximer n’importe quelle fonction (voir [83]par exemple). Par ailleurs, nous avons ajouté à ce type de réseau une connexion linéaire supplémentaire entre les entrées et les sorties afin que la relation linéaire soit un cas particulier correspondant à zéro couche cachée. Plus de détails concernant la mise en place et l’optimisation de ce réseau sont donnés en annexe C de ce manuscrit.

Dans un *modèle de mélange conditionnel* (CMM en anglais), la précipitation peut être vue comme provenant d’un ou de plusieurs régimes (également appelés “états”) étant donnée l’atmosphère qui est caractérisée par nos prédicteurs SLP. Ces états ne sont pas directement observés – ils sont cachés – et peuvent être compris comme résultant de processus sous-maille qui ne sont pas explicitement pris en compte par les prédicteurs. Chaque état caché est modélisé par l’une des composantes du mélange. Le poids $\pi_j(x)$ donne la probabilité d’occurrence de l’état j, et $f(y;\theta_j(x))$ est la densité de l’intensité sachant l’état j. Cette vision des CMMs comporte des similitudes avec l’approche par “modèles de Markov cachés non-homogènes” (NHMM) développée par [6]. Dans un NHMM, les états cachés représentent des types de temps dont les occurrences sont supposées suivre une chaine de Markov du premier ordre où les probabilités de transition dépendent de prédicteurs (voir équation 3.1). Ceci constitue l’une des différences principales avec nos CMMs qui capturent les dépendances temporelles au travers des prédicteurs exclusivement.

Afin de modéliser des tendances potentielles, en plus des données de SLP, nous avons également inclu trois variables de dates, caractérisant l’année, le mois et la semaine. L’année est codée comme une différence par rapport à une année de référence (ici, 1970). Le mois est construit comme une différence “circulaire” par rapport au mois de janvier : cette différence va de 0 en janvier jusqu’à 6 en juillet et décroit jusqu’à 1 en décembre. Le même type de calcul est réalisé pour la semaine, avec une différence circulaire prise sur 52 semaines. Ces différences circulaires (mois et semaines) ont ensuite été normalisées pour appartenir à $[0;1]$. Elles représentent une alternative à l’utilisation de fonctions sinus-cosinus (par ex., [215]) pour caractériser la saisonnalité et une potentielle tendance dans une distribution conditionnelle. Par ailleurs, comme la saisonnalité et une tendance peuvent être présentes à la fois dans les données SLP et les variables de date, nous avons appliqué une analyse en composante principale (ACP, [94]) sur les 39 variables qui ont été centrées et réduites. Cette application d’ACP sert principalement à réduire la dimension et à supprimer la redondance des prédicteurs. Les quatre premières composantes principales ont été gardées, expliquant plus de 90% de la variance des prédicteurs.

Le résultat de la calibration de ces modèles conditionnels est que, sachant les prédicteurs à un instant donné, le réseau de neurone nous fournit les paramètres des densités. Autrement dit, pour chaque jour (c-à-d., 36 SLP et 3 variables de date), nous disposons, non pas d’une unique valeur locale de précipitation par station, mais d’une densité de probabilité (un mélange) pour l’intensité de précipitation locale par station. On voit là toute la richesse d’une approche stochastique conditionnelle par rapport à une approche plus classique par fonction de transfert. Nous pouvons, par exemple, comparer le cycle saisonnier des probabilités d’occurrence observée avec celui des

probabilités conditionnelles fournies par les modèles qui permettent de calculer des intervalles de confiance, par exemple à 90%. Nous pouvons faire de même avec les valeurs des quantiles conditionnels à 99%. Ces deux comparaisons sont présentées en Fig. 3.6(a, et b) pour la station d'Orange et le modèle de mélange conditionnel de log-Normales pour illustration. Ces quantiles



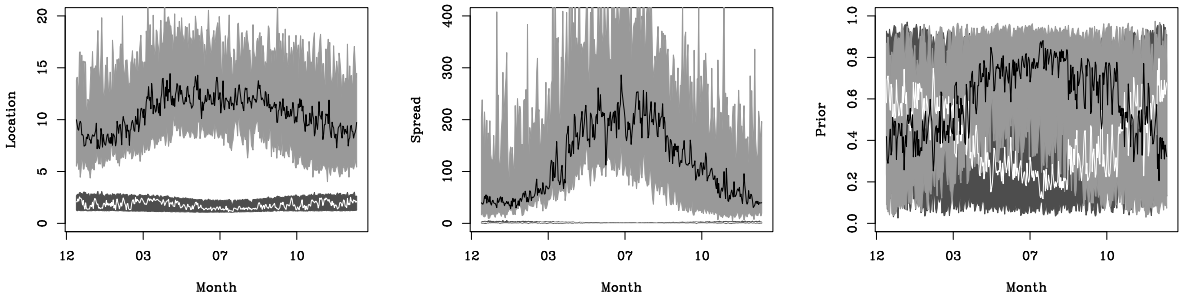
(a) (b)

FIGURE 3.6: Pour la station d'Orange avec le modèle de mélange conditionnel de lois log-Normales, illustration des cycles saisonniers journaliers (a) des probabilités d'occurrence de précipitation, (b) des quantiles de précipitation à 99%. Dans chaque panel, la ligne blanche correspond au cycle issu des observations de la période d'évaluation, la ligne noire au cycle issu des simulations et les zones grisées aux intervalles de confiance à 90% des simulations.

conditionnels et intervalles de confiance n'auraient pas été possibles avec une approche de type régression déterministe, et présentent une très bonne adéquation aux observations, bien meilleure que le modèle de [215] utilisé pour comparaison (non présenté). Les cycles saisonniers des paramètres eux-mêmes peuvent être étudiés grâce à une telle approche. La Fig. 3.7 montre les cycles saisonniers des paramètres de moyenne conditionnelles, de variance conditionnelle et de poids conditionnels du mélange à deux log-Normales, ainsi que l'intervalle de confiance à 90%. On voit alors que l'une des composantes est prédominante en été et l'autre en hiver, avec des propriétés évoluant dans l'année et surtout événement par événement. Une évaluation plus globale de la bonne reproduction de la distribution *inconditionnelle* des précipitations par CMM de log-Normales – meilleure que par le modèle de référence – est présentée en Fig. 3.8 au travers de QQplots.

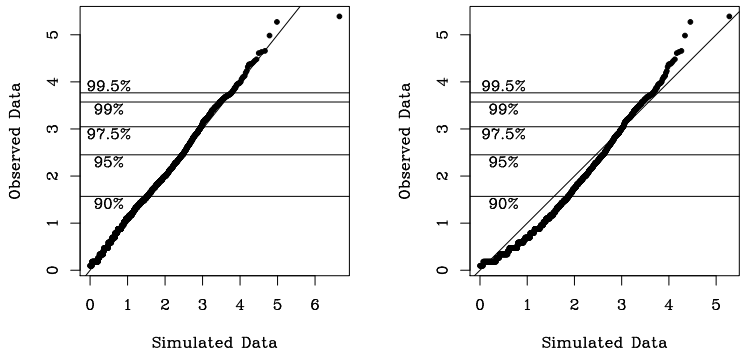
De manière générale, si le mélange conditionnel de Gaussiennes a présenté des résultats de meilleure qualité que le modèle de [215] utilisé en référence, il a toutefois montré quelques faiblesses dues à sa structure symétrique et à sa queue de distribution légère inadéquate pour représenter efficacement les événements extrêmes [13] (non présenté dans ce manuscrit). Cependant, le mélange conditionnel de log-Normales a présenté d'excellents résultats, que ce soit en terme de

13. ce modèle nécessitait par exemple un nombre relativement important de composante pour se rapprocher de la qualité des autres modèles testés.



(a) (b) (c)

FIGURE 3.7: Pour la station d'Orange avec le modèle de mélange conditionnel de lois log-Normales, illustration des cycles saisonniers des paramètres (a) de moyenne conditionnelles, (b) de variance conditionnelle et (c) de poids conditionnels du mélange à deux log-Normales. Les zones grisées (en plus ou moins foncé) correspondent aux intervalles de confiance à 90% de chaque composante.



(a) (b)

FIGURE 3.8: QQplots (en échelle logarithmique) pour la station d'Orange sur la période d'évaluation (a) entre observations et quantiles modélisés par le mélange conditionnel à deux lois log-Normales et (b) entre observations et quantiles modélisés par le modèle conditionnel de référence de [215] (Dirac - Gamma).

densité globale testée par QQplots, cycles saisonniers, ou même en terme "événementiel" [14]. Les résultats du modèle de mélange conditionnel de lois de Pareto hybrides seront discutés au chapitre 4.

Un package R (probablement nommée "CMM") permettant de calibrer et d'utiliser un modèle de mélange conditionnelle devrait prochainement être mis à disposition sur le site web du "Comprehensive R Archive Network" (CRAN) sur http ://cran.r-project.org/, ou depuis mon site web.

14. discuté au chapitre 4

3.3 Downscaling et modélisations environnementales par “fonctions de transfert”

Les méthodes de type “fonctions de transfert” (FT) ont eu un large succès en contexte de downscaling statistique. Leur facilité d’implémentation et d’utilisation – schématiquement, il s’agit de régressions – en ont fait des outils faciles à manipuler et à interpréter, même pour les non spécialistes des statistiques. Je ne reviendrai donc pas sur l’utilisation des méthodes par FT dans le contexte climatique “classique” (auquel j’ai moi-même contribué, par ex., dans [159, 70]) mais je vais tenter de présenter quelques applications et développements de celles-ci dans un cadre environnemental moins connu. En effet, si les méthodes par régressions paramétriques ou non-paramétriques, linéaires ou non-linéaires sont fréquemment employées en régionalisation statistique (voir courte introduction et références données en section 3.1), leur pratique peut également s’avérer fort utile pour lier *directement* des informations climatiques avec des variables environnementales diverses. Par exemple, [24] ont employé une méthode non-paramétrique basée sur un réseau de neurones pour prédire les débits de 21 bassins versants en Colombie-Britannique en fonction de données atmosphériques à l’échelle synoptique. On peut également citer [82] qui ont utilisé des données de réanalyses et des sorties de GCMs en entrées d’une régression multi-linéaire pour simuler les concentrations en ozone sur la région de Chicago et projeter leur évolution en contexte de changement climatique selon différents scénarios d’émission de gaz à effet de serre [15].

De la même manière que la régionalisation statistique est une alternative ou un complément à la régionalisation dynamique, ce genre d’approches est une alternative à des *modèles d’impacts* plus traditionnels (basés sur la physique, la chimie, l’hydrologie, etc.), apportant potentiellement d’autres informations complémentaires telles que la quantification de la contribution des différents prédicteurs atmosphériques.

3.3.1 Modélisation statistique directe de débits

Reprenant l’idée de [24], en collaboration avec C. Tisseuil (dont j’ai eu le plaisir de co-encadrer la thèse à l’Université Paul Sabatier de Toulouse), S. Lek (UPS, Toulouse) et A. Wade (University of Reading, UK), nous avons mené une étude d’intercomparaison de quatre modèles statistiques pour relier diverses variables climatiques à grande échelle provenant de réanalyses NCEP/NCAR ou de GCMs à des débits locaux de rivières répartis en 51 lieux de mesure dans le sud-ouest de la France et caractérisant différents systèmes hydrologiques, depuis un système “nival” (dominé par la fonte des neiges) jusqu’à un système “pluvial” (dominé par les pluies). Cette étude [179] avait pour but d’aider à sélectionner la méthode appropriée pour une modélisation hydrologique à fine échelle pour des études d’impacts des changements climatiques sur les ressources en eau dans cette région.

15. évidemment, l’évolution modélisée était uniquement due au changement climatique et non liée à un changement d’activités humaines.

La définition des prédicteurs résulte d'ACPs effectuées sur différents clusters de variables issues des réanalyses NCEP/NCAR et permettant de retenir comme prédicteurs les cinq composantes principales associées aux principaux processus identifiés qui sont : la précipitation, la température, les flux de chaleur, la pression et le rayonnement solaire. Les grandes lignes de cette sélection ainsi que le pré-traitement des données de débits sont fournis en annexe D. Une schématisation de l'ensemble des modèles statistiques et des résolutions (spatiales et temporelles) employés dans cette étude est donnée en Fig. 3.9. Les quatre modèles statistiques employés dans cette étude étaient :

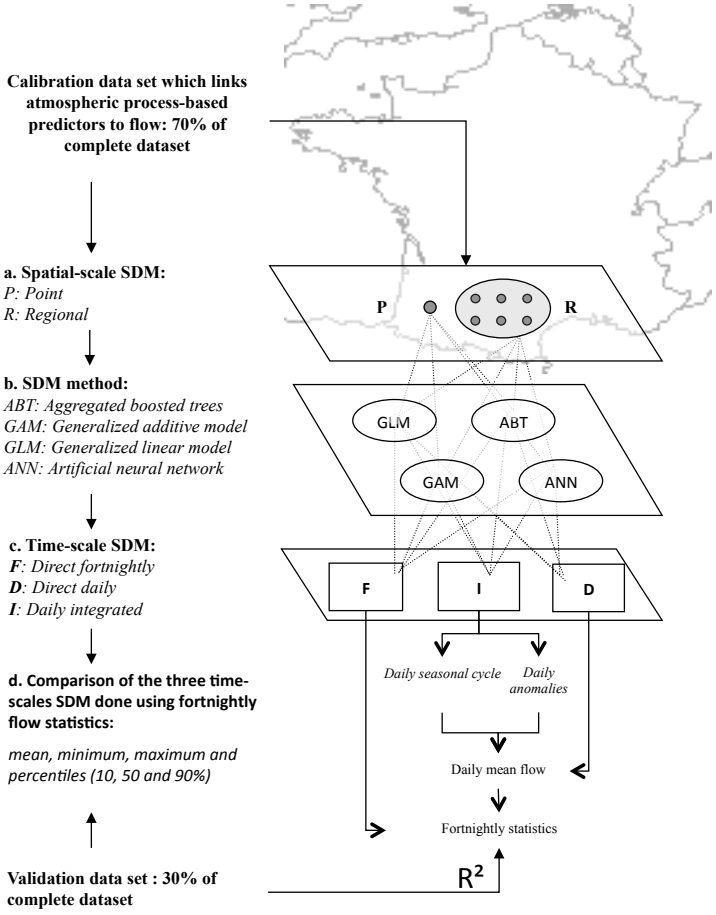


FIGURE 3.9: Schématisation de l'ensemble des modèles statistiques et des résolutions (spatiales et temporelles) employés pour la modélisation des débits dans l'étude [179].

– une modèle linéaire généralisé (GLM, [121, 73]),
– une modèle additif généralisé (GAM, [73]),

– un arbre “boosté” agrégé (“aggregated boosted tree” – ABT – en anglais, [61, 74]),
– un réseau de neurones perceptron multicouches (ANN, [12]).

Ces modèles sont brièvement rappelés en annexe E. Ces modèles ont été testés à deux résolutions spatiales différentes :

– Approche par “modèles locaux” : au niveau des stations elles-mêmes, signifiant que pour 51 stations, nous avons 51 modèles $\times$ 4 types de modèles ;
– Approche par “modèles régionaux” : au niveau régional défini par un groupement des stations en cinq classes (régimes) de comportements hydrologiques (depuis nival jusqu’à pluvial), signifiant que pour cinq classes, nous avons 5 modèles $\times$ 4 types de modèles, voir annexe D pour la définition de ces régimes. Dans cette approche régionale, les stations classées dans un même régime hydrologique sont supposées avoir le même modèle [16].

Par ailleurs, pour chaque modèle et chaque résolution spatiale, trois types de prédictants [17] (correspondant à trois résolutions temporelles) ont été considérés :

– les débits moyens journaliers,
– les pourcentiles 10%, 50% et 90% bi-mensuels (c-à-d., calculés sur des périodes de 15 jours consécutifs),
– les débits journaliers dits “intégrés”, consistant (1) à modéliser séparément le cycle journalier annuel des débits et les anomalies journalières par rapport à un cycle de référence, puis (2) à sommer les deux.

Le type d’approche par pourcentiles avait d’ailleurs déjà été employé, par exemple dans [41] pour améliorer l’estimation des extrêmes. On peut remarquer que cette approche est alors relativement proche (au moins philosophiquement) d’un modèle stochastique puisque le but est de modéliser des quantiles associés à des probabilités fixées, et non une réalisation déterministe du débit lui-même.

Chaque modèle a ensuite été calibré 500 fois sur 70% des données et les 30% restants ont été utilisés pour comparaison avec les simulations obtenues. Toutes les comparaisons sont effectuées en termes de coefficient de détermination R^2 des pourcentiles bi-mensuels par station (la définition du coefficient R^2 est redonnée en annexe D pour mémoire). Les résultats de l’analyse de variance sur les simulations de pourcentiles bi-mensuels par l’approche régionale ont montré que la sensibilité des débits modélisés aux prédicteurs atmosphériques était significativement différente entre régime nival et régime pluvial, le premier étant plus influencés par le rayonnement solaire ondes courtes et le deuxième par la température (voir Fig. 3.10a). Les modèles GLM, et GAM donnaient un poids prépondérant à la température et ANN au rayonnement solaire (Fig. 3.10b) même si ces deux variables avaient les contributions les plus fortes pour les quatre modèles. Les performances des modélisations ont ensuite été évaluées en terme de R^2 selon chaque modèle, approche et régime hydrologique. Les résultats sont illustrés en Fig. 3.11. En moyenne, les modèles non-linéaires (GAM, ABT and ANN) fournissent de meilleurs résultats que le linéaire GLM qui présentait les plus faibles résultats, le modèle ABT étant significativement le plus performant

16. leurs données ont été concaténées pour la calibration.
17. pour rappel : les prédictants sont les variables que l’on cherche à modéliser, simuler, ou reproduire (autrement dit les sorties du modèle statistique), par opposition aux prédicteurs qui servent à les contraindre (c-à-d., les entrées du modèle).

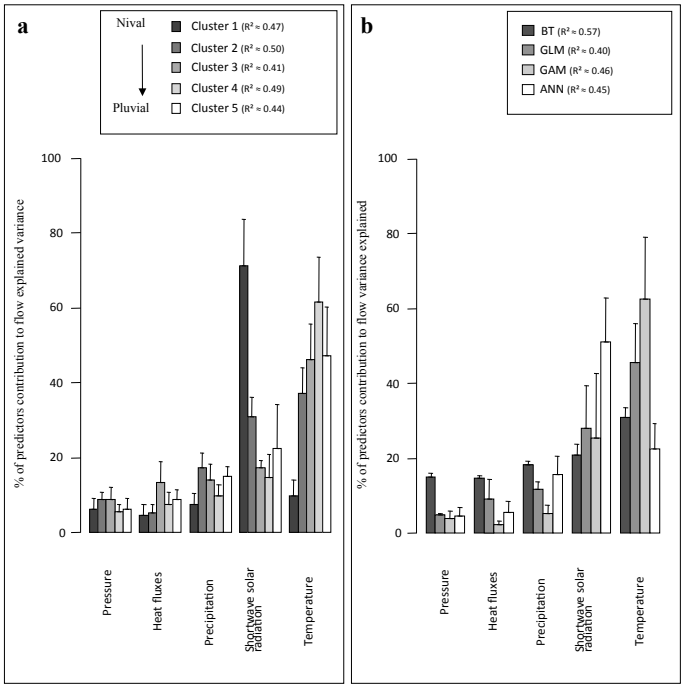


FIGURE 3.10: Résultats de l'analyse de sensibilité sur les simulations de pourcentiles bi-mensuels par l'approche régionale : pourcentage de contribution des prédicteurs atmosphériques (définis en annexe D) à la variabilité des débits simulés, (a) selon les régimes hydrologiques, de nival (en noir) à pluvial en blanc ; et (b) selon les quatre modèles statistiques.

(Fig. 3.11a). En global, la modélisation des pourcentiles bi-mensuels (B) surpasse légèrement les modélisations journalières et intégrées. Pour cette dernière, des analyses complémentaires (non présentées ici) montrent la qualité de la modélisation du cycle annuel mais la faiblesse des simulations des anomalies. Par ailleurs, les modèles locaux sont généralement plus efficaces que les modèles régionaux ; les performances sont globalement meilleures pour le régime pluvial que pour le nival ; et meilleures en été que dans les autres saisons.

Enfin, des projections de pourcentiles de débits bi-mensuels par l'approche régionale ont également été réalisées, basées sur les sorties du GCM CNRM-CM3 de Météo-France selon deux scénarios climatiques A2 et A1B. Les résultats (non présentés) suggèrent une diminution globale des débits pour l'ensemble des régimes hydrologiques, pour toutes les saisons sauf l'hiver, et quel que soit le scénario considéré.

Une extension de ce travail a déjà été lancée, au travers de la poursuite de la collaboration avec C. Tisseuil, sur une modélisation des présences/absences de diverses espèces de poissons d'eau douce dans le sud de la France et leurs évolutions spatiales et temporelles sous contraintes

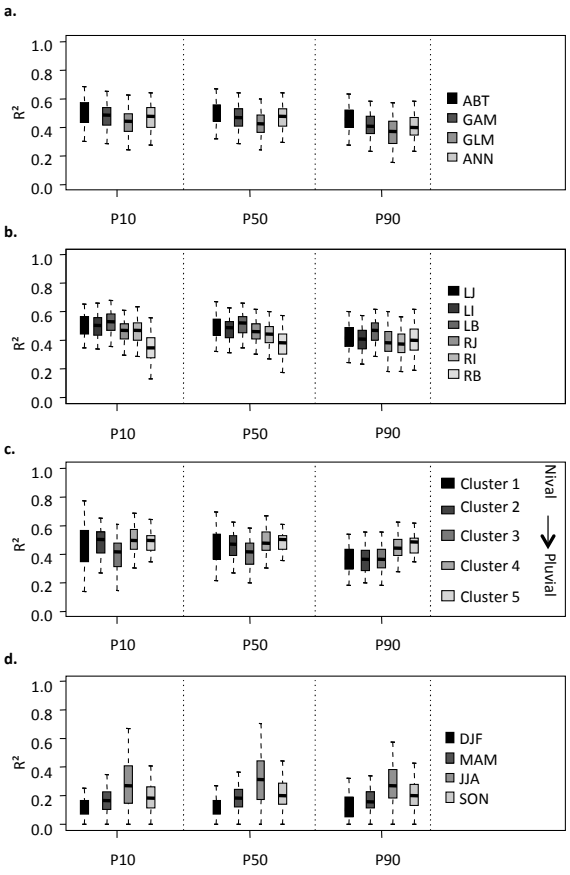


FIGURE 3.11: Boxplots des R^2 des trois pourcentiles 10% (P10), 50% (P50) and 90% (P90) bi-mensuels. Les comparaisons sont faites en termes : (a) de modèles statistiques ; (b) de résolution temporelle des prédictants (LJ = approche locale pour débit journalier, LI = approche locale pour débits “intégrés”, LB = approche locale pour pourcentiles bi-mensuels, RJ = approche régionale pour débit journalier, RI = approche régionale pour débits “intégrés”, et RB = approche régionale pour pourcentiles bi-mensuels) ; (c) de régimes hydrologiques ; (d) et de saisons.

de changements climatiques ([178, 177]). Les modèles ABT ont, dans ce contexte, été utilisés pour modéliser et simuler les conditions hydrologiques (débits) et climatiques (température et précipitation) en différents lieux ; et ces simulations locales (accompagnées de caractéristiques géographiques et environnementales) servent alors d’entrées à un modèle de niche écologique, lui-même de la forme ABT.

Cette étude et les poursuites en cours montrent bien les apports que peuvent avoir les statistiques dans un contexte de modélisation environnementale : intercomparaison efficace de modèles variés (linéaires ou non, paramétriques ou non, etc.) ; quantification de la contribution des prédicteurs ;

projections rapides et intelligibles pour les futures évolutions de débits ; variété des phénomènes climatiques et environnementaux modélisés ; etc. Bien sûr, seules les contraintes climatiques ont été prises en compte ici, et les différentes activités humaines (barrages, pompage, etc.) jouant sur les débits, ou les phénomènes de compétitions entre espèces par exemple n'ont pas (encore) été modélisées dans ces approches. Toutefois ces approches originales "climat vers débits" ou "climat vers biodiversité" constituent des alternatives (et des compléments !) désormais crédibles à des modèles hydrologiques ou écologiques, évidemment plus complexes et plus riches de par l'ensemble des processus mis en jeu, mais beaucoup plus "lourds" à faire tourner et donc moins flexibles pour des analyses de sensibilité et d'incertitude.

Un package R nommé "DS" a été créé afin de regrouper l'ensemble des méthodes testées et analyses effectuées. Ce package est disponible sur demande.

3.3.2 Downscaling de climats passés lointains

Si les développements de modèles pour la régionalisation sont majoritairement motivés par la compréhension de processus physiques à fine échelle au présent ou par des projections à haute résolution pour le futur, l'étude des climats régionaux passés n'en est pas moins une application pertinente. En effet, l'étude des climats passés repose essentiellement sur deux approches : la modélisation climatique globale à longue échéance passée lointaine ; et l'analyse de différents "proxies" (cernes d'arbres, carottes sédimentaires ou de glace, etc.) permettant des reconstitutions des principales caractéristiques de température et de précipitation. Or, si ces proxies sont nécessairement très locaux et très dépendants de leur environnement car basés sur des mesures en des lieux bien précis avec des caractéristiques géographiques données, les modèles climatiques pour les climats lointains possèdent bien souvent de très faibles résolutions spatiales et temporelles et des choix dans les processus atmosphériques représentés afin de pouvoir tourner en des temps raisonnables (par ex., [141]) : on parle alors de "modèles du système Terre de complexité intermédiaire" ("Earth system Models of Intermediary Complexity", EMIC). La comparaison entre sorties d'EMIC et reconstructions à partir de proxies est alors délicate, due à une différence d'échelle spatiale très importante. Le downscaling des sorties d'EMIC est donc une étape nécessaire pour évaluer la qualité de ces simulations par rapport à des reconstructions. C'est pourquoi, en collaboration avec P. Marbaix (Univ. catholique de Louvain) et D. Paillard et P. Naveau (LSCE), nous nous sommes intéressés dans l'étude [196] à la régionalisation statistique des températures et précipitations de l'EMIC CLIMBER ([142, 141]) sur l'Europe de l'ouest, durant le dernier maximum glaciaire (DMG, environ -21 000 ans).

La variabilité (journalière, inter-annuelle, etc.) étant difficilement mesurable pour cette période, des séries temporelles ne sont pas directement modélisables ici. Ainsi, l'une des originalités de notre objectif était que nous ne voulions pas simuler des séries temporelles mais des climatologies

mensuelles en température et précipitation à haute résolution. Pour cela, nous avons tout d'abord cherché à modéliser le lien entre :

- des sorties du modèle CLIMBER (température, précipitation vent, SLP, etc.), avec une résolution spatiale de 10^o en latitude et 51^o en longitude, pour le climat présent,
- et des données provenant du “Climate Research Unit” (CRU, [130]) à une résolution spatiale de $10'$ (c-à-d., environ $1/6$ de degré) pour l'Europe de l'ouest, et caractérisant en chaque point de grille les valeurs mensuelles et annuelles moyennes de température et de précipitation – autrement dit les “climatologies” mensuelles et annuelles – pour la période 1961-1990.

Une fois le modèle établi, des sorties de CLIMBER pour le DMG ont été utilisées pour projeter les climatologies à haute résolution pour cette période.

Afin de modéliser les valeurs de CRU (notre prédictant Y) en fonction des sorties de CLIMBER (nos prédicteurs X), le modèle statistique employé correspondait à un modèle additif généralisé (GAM, [73]) dont le principe a été rappelé en annexe E. L'intérêt d'une telle approche était double : la possibilité d'avoir un modèle fortement non-linéaire grâce aux combinaisons de splines cubiques ; et la visualisation des liens entre les données CLIMBER et CRU, fournissant alors une interprétation physique de ces fonctions splines.

La régionalisation des températures et précipitation du DMG ne peut, a priori, pas se faire avec les mêmes prédicteurs qu'une régionalisation plus classique dans un climat proche du présent. Notre objectif était de prendre en compte des faits climatologiques liés à la géographie (par ex., élévation par rapport au niveau de la mer, continentalité, pente du terrain) sans description physique explicite, et de les comparer à des prédicteurs physiques plus traditionnellement employés en climat proche présent. Par exemple, les montagnes et le vent influencent fortement la température et les précipitations mais la modélisation de ces influences d'une façon explicite peut rapidement devenir complexe. La principale idée était donc de “laisser” GAM le faire de manière statistique, en utilisant et comparant deux ensembles de prédicteurs.

Les neuf variables suivantes (issues de CLIMBER, donc à grande échelle) sont considérées comme les prédicteurs “physiques” : humidité spécifique (Q), humidité relative (RH), pression au niveau de la mer (SLP), temperature (T), composante u du vent (Wu), composante v du vent (Wv), température du point de rosée (Td), écart au point de rosée (DTd = Td - T), humidité spécifique intégrée verticalement (QI). Les variables Q, RH, T,Wu, Wv, Td et DTd sont prises à la surface. Dans cet ensemble, DTd représente le degré de saturation de la vapeur d'eau dans l'atmosphère. Les variables Td et DTd ont déjà montré de bonnes capacités pour le downscaling dans d'autres études ([31, 199, 197]).

Par ailleurs, quatre prédicteurs géographiques ont été testés : élévation par rapport au niveau de la mer (elv), continentalité advective (Aco), continentalité diffusive (Dco), et la variable W-slope (Wsl) résultant de l'intensité du vent zonal moyen multipliée par la pente est-ouest moyenne sur 100 km. Ces trois dernières variables sont définies en annexe F. Pour chaque ensemble de prédicteurs, deux modèles distincts ont été calibrés séparément pour la température et les précipitations selon l'approche dite de “step-wise screening”, signifiant que pour une variable donnée, un même modèle GAM est utilisé pour représenter chaque point de grille CRU pour

l'ensemble de la région. Pour les précipitations, plus précisément, GAM cherchait à modéliser la log(précipitation). En effet, GAM suppose que la famille de distribution du prédictant est connue. Or, le log des valeurs de précipitation peut généralement être considéré comme suivant une loi Gaussienne (voir section 3.2.2). Les précipitations régionalisées étaient donc le résultat de l'exponentielle appliquée aux log(précipitations) modélisées. En plus de diverses évaluations numériques telles que basées sur les pourcentages de variance expliquée, sommes des carrés des résidus (RSS), ou autres critères (non présentés ici, voir [196]), la qualité des modèles est illustrée en Fig. 3.12 en terme de température et précipitation reproduites par GAM pour le mois de janvier avec les prédicteurs physiques. Une évaluation rapide des prédicteurs géographiques ou physiques est

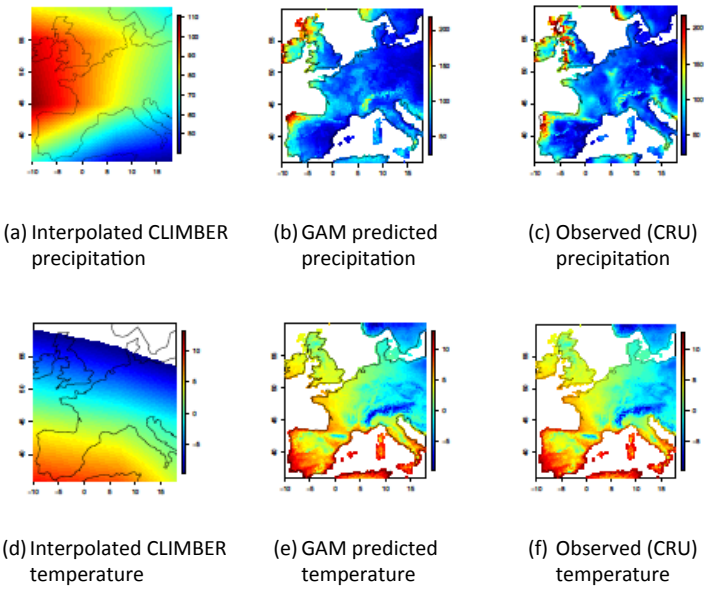


FIGURE 3.12: Pour le mois de janvier (climat présent), (a) précipitation CLIMBER interpolée à la résolution CRU, (b) précipitation régionalisée (résolution CRU) par GAM, (c) précipitation CRU (mm/mois). Les panels (d-f) sont les équivalents pour la température (^{o}C).

présentée en terme de cartes de résidus pour l'Europe de l'ouest pour les mois de janvier, avril, juillet et octobre dans les Figs. 3.13 (pour la log-précipitation) et 3.14 (pour la température). Afin de tester la généricité de ces modèles calibrés sur l'Europe de l'ouest et le rôle joué par les différents prédicteurs, d'autres tests ont été réalisés en appliquant ces modèles pour régionaliser le climat – au présent – d'autres régions : Amérique du nord et Europe du nord. Les cartes et évaluations numériques ne sont pas détaillées ici mais, en général, les résultats invalident les modèles basés sur les prédicteurs physiques et calibrés sur l'Europe de l'ouest pour projeter le climat dans les autres régions. Bien que les prédicteurs géographiques ne soient pas entièrement satisfaisants, ils fournissent de meilleures projections (c-à-d., par ex. avec de plus petits résidus)

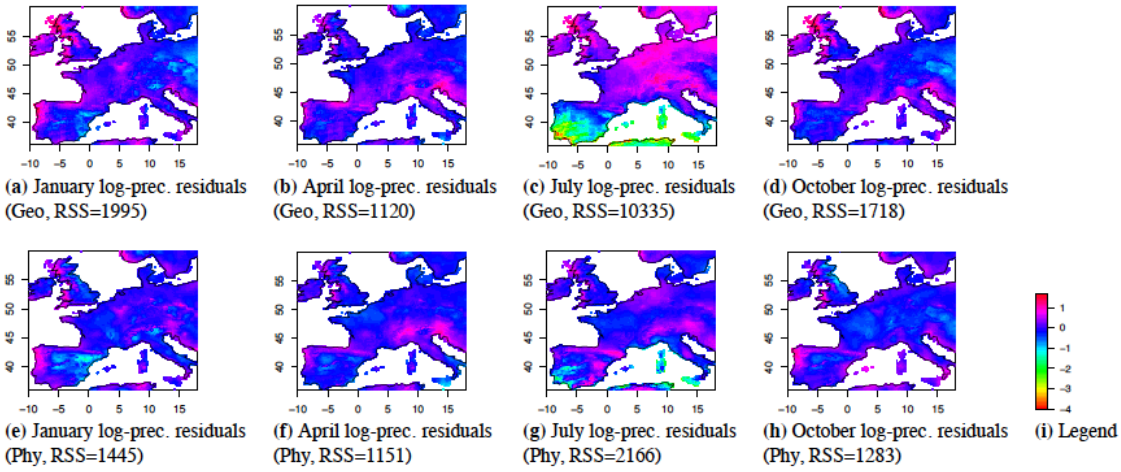


FIGURE 3.13: Résidus (obs-pred) des log-précipitations par GAM, (a-d) à partir des prédicteurs géographiques, (e-h) à partir de prédicteurs physiques, pour janvier, avril, juillet et octobre. Les valeurs des sommes des carrés des résidus (RSS) sont données pour chaque mois. La légende est fournie en log(mm/mois) en (i).

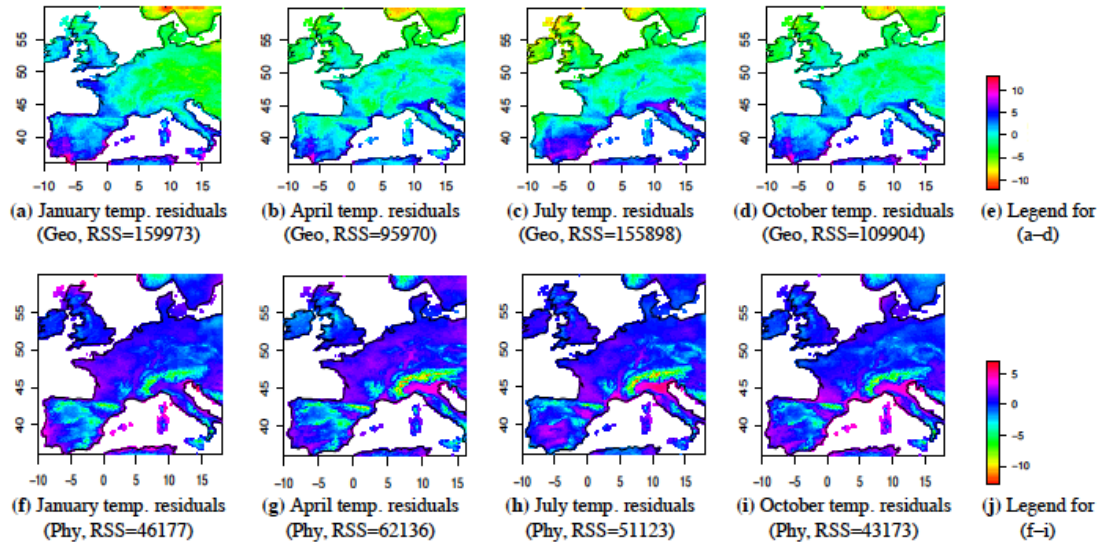


FIGURE 3.14: Comme pour Fig. 3.13 mais pour la température, avec les légendes données en ^{o}C.

que les prédicteurs physiques seuls sous des conditions de climat à grande échelle différentes de celles de calibration. Autrement dit, ils étaient plus “robustes” au changement de région pour la projection.

Nous avons alors cherché à régionaliser les températures et précipitations du dernier maximum glaciaire. Les prédicteurs sont les variables géographiques elv, Aco, et Wsl [18], auxquelles une variable physique a été ajoutée. Ce choix d'un unique prédicteur supplémentaire est une indication du critère BIC, calculé avec toutes les combinaisons possibles de prédicteurs. Cette variable a été choisie d'après les valeurs de BIC (au présent) et différents choix subjectifs. Les prédicteurs des deux modèles GAM sont donc :

- Pour la température : les prédicteurs géographiques elv, Aco, et Wsl + la température (T) CLIMBER

- Pour la (log) précipitation : les prédicteurs géographiques elv, Aco, et Wsl + SLP

Ces deux variables supplémentaires sont supposées fournir des informations à large échelle, utiles pour diriger physiquement la régionalisation. Les splines obtenues lors de la calibration (au présent bien sûr) sont essentiellement linéaires et/ou monotones mais jamais constantes, confirmant un sens physique sous-jacent aux variables sélectionnées. Les simulations de CLIMBER caractérisant le DMG ont donc été utilisées pour définir les prédicteurs Aco, Wsl, T, et , SLP. Comme le niveau de la mer au GMD est 120 m plus bas qu'aujourd'hui, le prédicteur elv du DMG est défini comme le prédicteur elv présent plus 120 m.

Afin d'illustrer les projections aux DMG, Fig. 3.15 présente les cartes de différences de température (en absolu) et de précipitations (en relatif) entre DMG et présent à haute résolution (CRU) pour janvier, avril, juillet et octobre. Ces résultats ne sont pas commentés ici. Cependant, mis à part l'augmentation de pluie dans le Nord-Est, ils sont en accord relatif avec les anomalies de précipitations et de températures au DMG telles que fournies par [96] pour l'Europe. D'autres évaluations de ces simulations DMG ont été effectuées en les comparant par exemple avec un ensemble de sorties de modèles climatiques issues du projet PMIP2 (par ex., [17]). Si ces évaluations ne sont pas non plus détaillées dans ce manuscrit, en général, le processus de régionalisation par le modèle GAM dirigé par les sorties de CLIMBER fournissait des climatologies de température et de précipitation réalistes et satisfaisantes, montrant ainsi la qualité et les apports de la méthode proposée.

J'ai depuis eu l'occasion de ré-employer GAM dans différents contextes, que ce soit un contexte climatique pour le downscaling des composantes du vent ([159]), ou par exemple pour la modélisation des occurrences de coulées de débris dans les Alpes en fonction du climat ([95]). Toutefois, ces travaux sur la régionalisation de climats lointains, nécessitant de définir des prédicteurs géographico-physiques, se sont récemment poursuivis en collaboration avec A. Martin (postdoc LSCE) et D. Paillard (LSCE) pour essayer de mieux comprendre le rôle de ces prédicteurs et leurs poids relatifs, le rôle de leur résolution, etc. Un article ([120]) est actuellement en finalisation

18. Dco n'est pas apparu très informatif dans les évaluations effectuées, non montrées ici, voir [196].

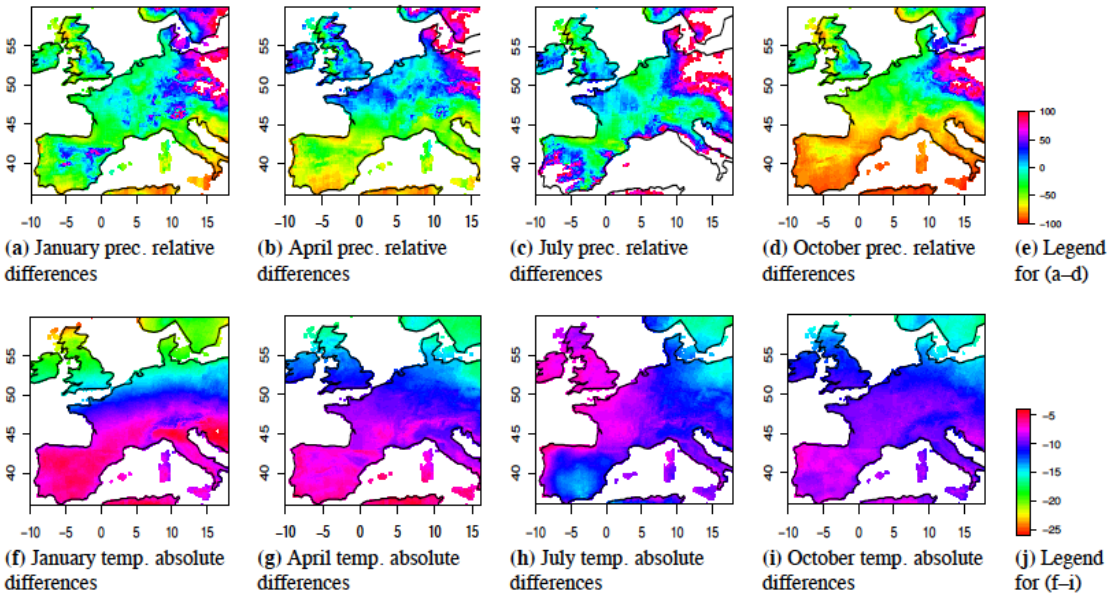


FIGURE 3.15: Cartes mensuelles (a-d) de différences relatives de précipitation (par rapport au présent) et (f-i) de différences absolues de températures, entre DMG régionalisé par GAM et présent pour janvier, avril, juillet et octobre. Les légendes sont en % pour (e) et en ^{o}C pour (j).

d’écriture à ce sujet et devrait être soumis très prochainement. Par celui-ci sera suivi par un article complémentaire [119] – également déjà en finalisation d’écriture – sur d’autres projections et comparaisons pour la régionalisation du climat passé (DMG) et du climat très lointain du futur million d’années.

3.3.3 Modélisation du pergélisol

La poursuite des études mentionnées en section 3.3.2 sur les climats lointains et les méthodologies associées de downscaling m’a alors permis de diversifier mes travaux de recherche. En effet, de nombreux phénomènes *environnementaux* liés au climat nécessitent d’être étudiés non seulement sur de longues échelles de temps (par ex., plusieurs milliers, voire millions d’années) mais également à un niveau spatial très local ou, mieux à haute résolution. Un modèle de régionalisation robuste en contexte de (fort) changement de climat à grande échelle est alors indispensable. C’est le cas, par exemple, du pergélisol. Le pergélisol (ou permafrost en anglais) est défini comme un sol dont la température est inférieure ou égale à 0^{o}C de manière permanente pendant au moins deux années consécutives ([56]). Ce phénomène possède des rétro-actions critiques sur le climat, entre autres, en contrôlant une partie des émissions de méthane ([221]), et ce sur plusieurs milliers d’années. Sa connaissance à différentes périodes – dont le DMG – est donc particulièrement importante pour améliorer notre compréhension du système Terre. Or, la modélisation du pergélisol

reste très délicate au DMG, la résolution des GCMs et EMICs étant très insuffisante pour fournir des simulations climatiques pertinentes aux différents modèles sensés représenter ce phénomène possédant une forte variabilité spatiale. On se retrouve alors bien souvent à devoir comparer des données très locales issues de reconstructions paléoclimatiques avec des sorties de modèles caractérisant plusieurs milliers de km^2. Par ailleurs, la variabilité des propriétés spatiales et temporelles des climats du DMG simulés par les modèles climatiques est relativement importante. Leur utilisation dans des modèles de pergélisol propage cette variabilité et génère ainsi une incertitude élevée sur la répartition spatiale de ce phénomène.

Une partie de la thèse (2009-2012) de G. Levavasseur que je co-encadre au LSCE a alors consisté à étudier si la régionalisation du climat DMG et la modélisation statistique à haute résolution du permafrost permettaient (1) d'améliorer la comparaison modèle - données et (2) de réduire la variabilité spatiale du pergélisol.

Pour cela, nous avons appliqué deux modèles statistiques de philosophies très différentes, et permettant tous les deux de passer de simulations à grande échelle fournies par les EMICs ou GCMs, à des données *catégorielles* à haute résolution (résolution CRU de $1/6$ de degré) sur l'ensemble de l'Eurasie et caractérisant en chaque point de grille l'absence ou la présence de permafrost, et dans ce dernier cas son type (continu [19] ou discontinu [20]). Comme dans la section précédente 3.3.2, nous travaillons sur des données climatologiques représentant des valeurs moyennes pour une période donnée et suffisamment longue où le climat est supposé être à l'équilibre. Nous supposons également que le permafrost est à l'équilibre avec le signal climatique. Autrement dit, dans la suite, nous ne prenons pas en compte les effets du climat sur le permafrost qui seraient décalés dans le temps. Les données à grande échelle proviennent des neuf modèles impliqués dans PMIP2. Les données de climat sont les données de climatologie mensuelle CRU. Les données de permafrost (nécessaires à l'évaluation des modèles statistiques ainsi qu'à la calibration du deuxième modèle) viennent de l' "International Permafrost Association" (IPA) et du "Frozen Ground Data Center" (FGDC) ([19]) pour les données au CTRL, et de [185] pour les données au DMG.

Le premier modèle testé est la continuité de mes travaux [196] et consiste à calibrer et appliquer GAM sur les prédicteurs physico-géoraphiques issues des modèles climatiques pour générer des valeurs de température *climatologique* à la résolution CRU. Puis, à partir de ces températures T, une simple relation définie par [148] (explicitement décrite dans [184] et nommée relation "RV") permet de catégoriser les types de pergélisol :

- Si (T moyenne annuelle est $\leq -8^oC$)& (T du mois le plus froid est $\leq -20^oC$), alors : pergélisol continu,
- Si ($8^oC \leq T$ moyenne annuelle $\leq -4^oC$), alors : pergélisol discontinu,
- Si aucune des conditions n'est vérifiée, alors : pas de pergélisol.

L'ensemble du modèle GAM allié à ces conditions sera nommé GAM-RV par la suite. Celui-ci étant basé sur un modèle GAM et des conditions déterministes, il se situe dans la catégorie des "fonctions de transfert".

19. couvrant plus de 80% du sous-sol
20. couvrant entre 30 et 80% du sous-sol

Le deuxième modèle testé se situe lui davantage dans la catégorie des “weather generators” que l’on pourrait renommer pour cette étude les “permafrost generators”. En effet, nous avons appliqué ici un modèle de régression logistique multinomiale (telle que par exemple présentée précédemment dans l’Eq. (3.4) dans le cas binomial) à l’aide d’un modèle GAM, et permettant de modéliser *directement* la probabilité d’être dans chacune des trois catégories de pergélisol (continu, discontinu, ou absence) à partir des prédicteurs physico-géographiques. Autrement dit, les paramètres de la loi multinomiale – qui sont les probabilités recherchées – sont fonctions des prédicteurs au travers d’un modèle GAM. Cette approche sera nommée ML-GAM (pour Multinomiale Logistic - GAM). À partir des probabilités modélisées, si plusieurs techniques peuvent être employées pour déterminer ou simuler une catégorie, dans cette étude, nous avons décidé de sélectionner la plus probable comme étant la réalisation.

Avant de passer à l’illustration de quelques résultats, je tiens à m’attarder quelques lignes sur les conditions RV mentionnées ci-dessus en les comparant aux relations que nous fourniraient ML-GAM à partir des mêmes informations (c-à-d., T moyenne annuelle et T du mois le plus froid) comme seuls prédicteurs. Les conditions RV peuvent être reformulées ainsi de manière probabiliste :

- Pr(pergélisol continue $\mid$ T moyenne annuelle $\leq -8^oC$ & T du mois le plus froid $\leq -20^oC$) = 1,
- Pr(pergélisol discontinu $\mid$ $8^oC \leq T$ moyenne annuelle $\leq -4^oC$) = 1,
- Pr(pas de pergélisol $\mid$ aucune des conditions n’est vérifiée) = 1.

Ces probabilités peuvent être visualisées dans la colonne de gauche de la Fig. 3.16. Par ailleurs, afin d’évaluer ces conditions RV, nous avons cherché les relations que fourniraient ML-GAM pour représenter les probabilités d’occurrence de pergélisol à partir des données fournies par l’IPA/FGDC et à partir des valeurs de température moyenne annuelle et de celle du mois le plus froid provenant des données CRU. En projetant les probabilités pour chaque couple possible (T moyenne, T la plus froide) à partir de la calibration ainsi établie, nous obtenons alors la colonne droite de la Fig. 3.16. Il est intéressant de voir que, bien que différentes, les deux colonnes de figures sont relativement similaires, celle de droite issue de ML-GAM étant plus “lisse” ou plus “continue”. Ces similitudes sont des indications positives à la fois pour les conditions RV et pour le modèle ML-GAM. En effet, le fait que ML-GAM fournisse des relations proches de celles définies “empiriquement” par [148] donne une certaine confiance dans la capacité de ce modèle à déterminer des relations “climat-pergélisol” pertinentes. À l’inverse, le fait que les conditions d’isothermes RV fournissent des relations proches de celles définies de manière indépendante et optimisées de façon totalement automatique par ML-GAM, est un indice d’une certaine robustesse de ces isothermes. Cependant, si la visualisation de ces relations est faisable lorsqu’on travaille avec uniquement deux prédicteurs, il n’en est pas de même avec davantage. Cette analyse était donc essentiellement illustrative du comportement de ML-GAM dans un contexte simple pouvant se comparer avec les isothermes de RV. Malgré cela, ces relations peuvent être établies pour chaque GCM indépendamment et seront normalement (légèrement ?) différentes d’un GCM à un autre. Bien que cela n’ait pour le moment pas été exploité davantage lors de la thèse de G. Levavasseur,

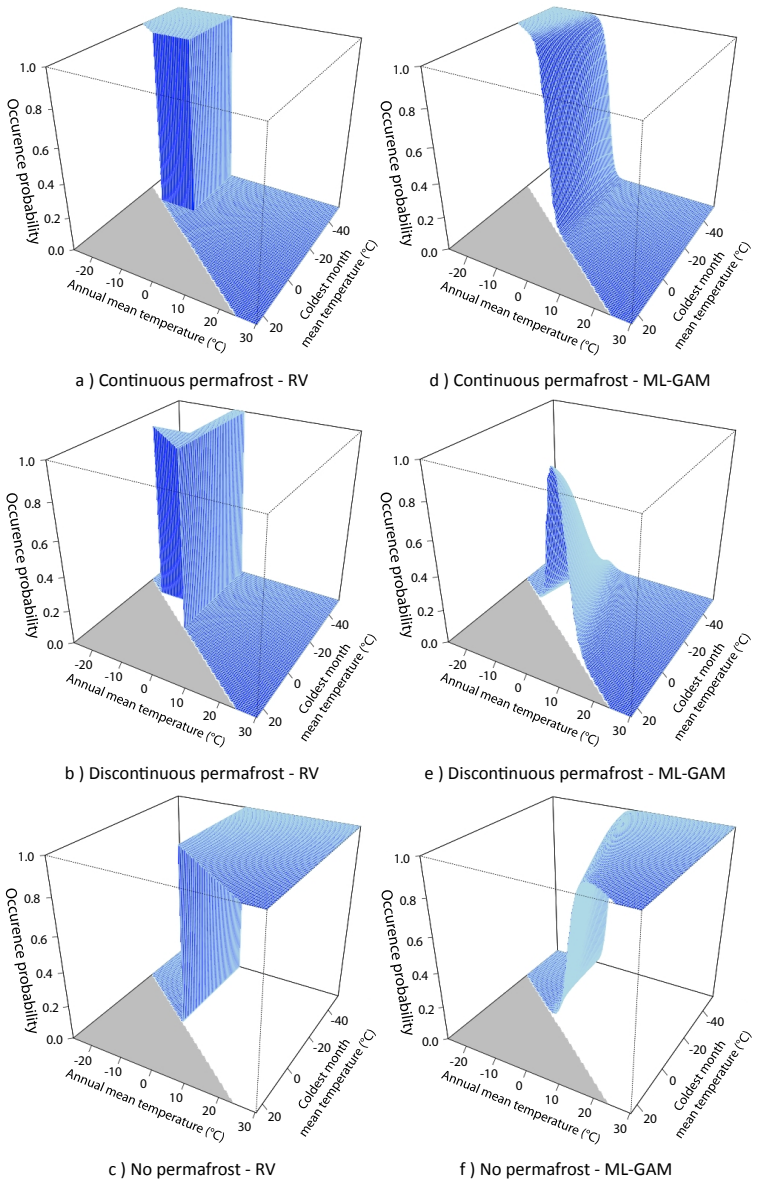


FIGURE 3.16: Probabilités d'occurrence de pergélisol basées sur la température moyenne annuelle et la température du mois le plus froid. La colonne de gauche (a-c) correspond aux conditions de température fixées par RV dans [148] utilisées pour GAM-RV La colonne de droite montre les relations de probabilités calibrées à partir des mêmes variables (ici provenant des données CRU)avec ML-GAM. Les zones grisées correspondent aux valeurs mathématiquement impossibles (c-à-d., quand la température moyenne annuelle est plus froide que la température du mois le plus froid).

ceci pourrait s’avérer particulièrement intéressant pour caractériser les différences entre GCMs dans leurs liens respectif avec le pergélisol. Ce type d’analyse pourrait évidemment s’étendre à bien d’autres phénomènes environnementaux.

Après différentes analyses préliminaires non présentées ici, quatre prédicteurs ont été sélectionnés. Pour les deux modèles statistiques, les prédicteurs sont les mêmes afin de pouvoir mener des comparaisons justes, et correspondent à : la température de l’air à la surface (TAS), les deux indices de continentalité advective et diffusive (Aco et Dco), et l’élévation locale (elv).

Une première étape d’évaluation des modélisations a été menée sur l’ensemble des neuf modèles climatiques impliqués dans PMIP2. Ici, seuls quelques résultats issus du modèle de l’IPSL seront présentés. Toutefois, un ensemble d’analyses basées sur des projections sous forme de cartes et des mesures quantitatives a été appliqué (statistique du Kappa, % de surface d’accord pour les différentes catégories, etc.) sur l’ensemble des modèles climatiques. Les résultats complets ne sont pas présentés ici mais peuvent évidemment être retrouvés dans [114]. La fig. 3.17 présente une comparaison des indices de pergélisol obtenus au CTRL à partir du modèle de l’IPSL (downscalé ou non) et les données IPA/FGDC utilisées en référence pour le CTRL. Chaque couleur montre un type d’accord ou de désaccord entre modèle et données. Les indices dérivés à partir du GCM sont obtenus en Fig. 3.17(a) par interpolation bilinéaire des températures du GCM puis par application des conditions RV ; en Fig. 3.17(b) par downscaling par GAM des températures du GCM puis application des conditions RV ; en Fig. 3.17(c) par modélisation directe des indices de permafrost selon ML-GAM. On voit parfaitement que l’approche GAM-RV n’apporte pas d’amélioration systématique par rapport aux indices dérivés des interpolations du GCM. Ce résultat se vérifie également sur les autres GCMs et sur l’ensemble des critères d’évaluations numériques mis en place (non montrés). Par contre, il est également assez net (de par ces figures et les critères numériques) que l’approche plus “probabiliste” ML-GAM entraine une amélioration de la distribution des catégories de pergélisol. Ceci confirme qu’une relation simple entre pergélisol et température de l’air (c-à-d., approche GAM-RV) n’est pas toujours suffisante pour une modélisation efficace du pergélisol. Cependant, pour les deux approches, l’information fournie à haute résolution a permis de réduire les variations inter-GCMs en termes de superficie d’erreur (non montré).

Afin de quantifier la capacité des GCMs à représenter le pergélisol au DMG, nous avons ensuite appliqué GAM-RV et ML-GAM à partir des sorties de GCMs pour cette période. Des cartes similaires à celles de la Fig. 3.17 ont été tracées pour comparer les simulations avec les données de [185] caractérisant le pergélisol DMG. Les projections par les deux approches n’étaient alors plus significativement en meilleur accord avec les données que ne l’étaient les projections issues de l’interpolation des modèles climatiques,f suivies des conditions RV (non montré). Si la variabilité inter-GCMs était réduite au CTRL, ce n’est le cas pour aucune des deux méthodes avec les sorties au DMG. De manière générale, les modèles climatiques sont trop chauds au DMG pour pouvoir guider correctement un modèle de pergélisol et réduisent ainsi la contribution des modèles statistiques au DMG. Ces résultats sont cependant à pondérer par le fait que les données de pergélisol

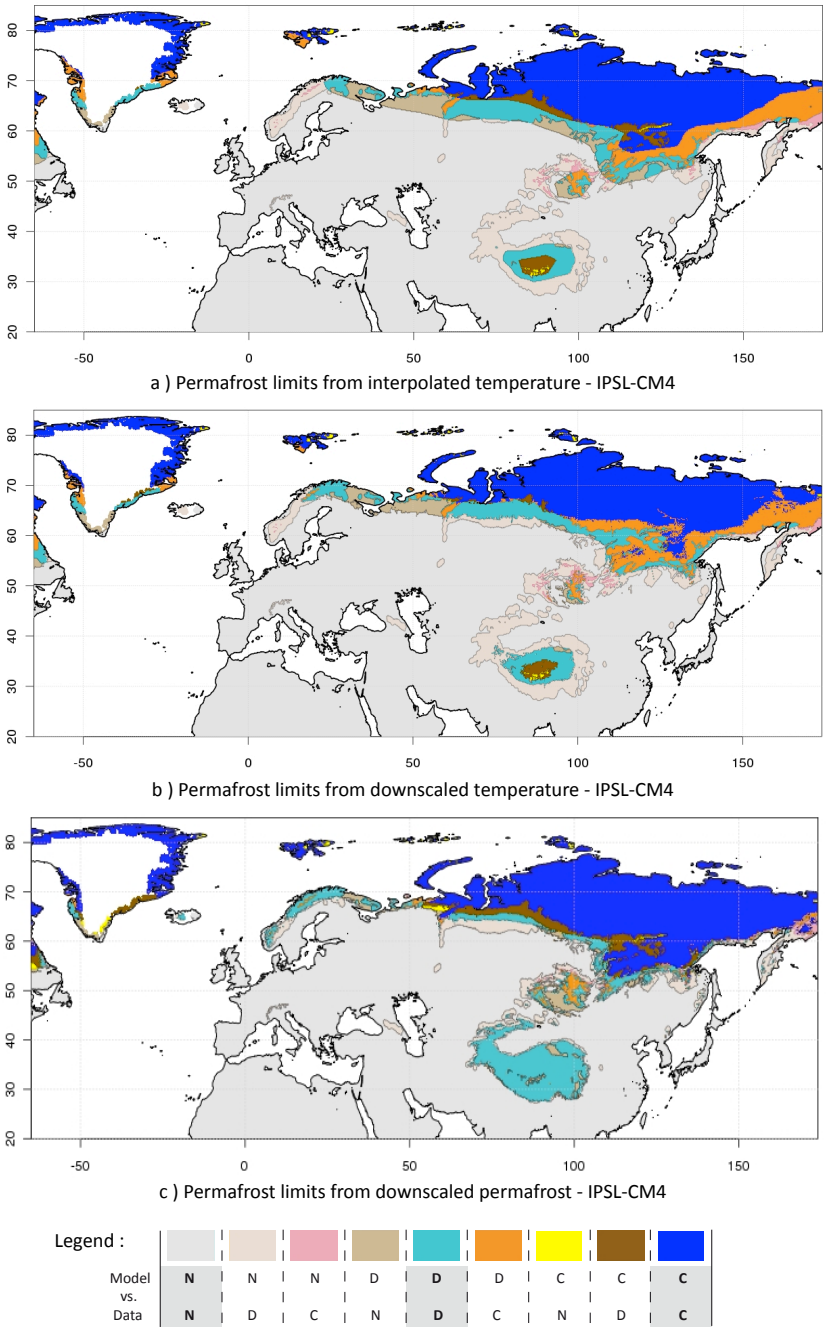


FIGURE 3.17: Comparaison des indices de pergélisol, au CTRL, entre le modèle de l'IPSL (downscalé ou non) et les données IPA/FGDC. Les indices dérivés à partir du GCM sont obtenus : (a) par interpolation bilinéaire des températures du GCM puis par application des conditions RV ; (b) par downscaling par GAM des températures du GCM puis application des conditions RV ; (c) par modélisation directe des indices de permafrost selon ML-GAM. Chaque couleur montre un type d'accord ou de désaccord entre modèle et données. Dans la légende, "N" correspond à "No permafrost", "D" à "Discontinuous permafrost" et "C" à "Continuous permafrost". Les couleurs dont la colonne est grisée et les lettres en gras dans la légende montrent un accord entre GCM et données. Les autres couleurs correspondent à des désaccords.

au LGM sont entachées d’une incertitude potentiellement importante et relativement difficile à estimer.

D’autres études et analyses sont nécessaires pour encore améliorer la modélisation des liens climat-pergélisol, par exemple au travers de davantage de prédicteurs, tels que température du sol (et non de l’air), type de sol, ou des prédicteurs plus géographiques tels que l’exposition au soleil, orientation du terrain, etc.

Si la synergie entre variables physiques et géographiques est visiblement très prometteuse, le choix et l’équilibre de ces variables sont certainement cruciaux pour obtenir une bonne représentation locale des liens prédicteurs-prédictants, et un modèle robuste, applicable à des climats lointains très différents du climat actuel.

3.4 Approche “Model Output Statistics” (MOS)

A la suite de ma thèse, j’ai gardé un goût particulier pour l’analyse et la modélisation des données probabilistes, par exemple de type fonctions de répartition cumulée (CDF en anglais). Dans le contexte de downscaling, cela correspond à chercher à directement modéliser/régionaliser une CDF à l’échelle locale, disons à une station, - en ce sens, le but est relativement similaire à l’approche stochastique par modèle non-homogène (voir section 3.2) - mais en considérant des données de type CDF également à grande échelle. Ici, nous ne cherchons pas à déterminer une densité (ou une CDF) dont les paramètres sont conditionnés par la réalisation d’un ensemble de variables atmosphériques à basse résolution pour un moment donné (par ex., aujourd’hui), mais nous cherchons le lien global qui existe pour l’ensemble de la période, par exemple entre la CDF d’une variable, disons l’intensité du vent à 10 m, pour un point de grille basse résolution (GCM ou réanalyses) et la CDF de cette même variable mais pour une station météo dans ce point de grille. Nous parlerons alors d’approche “probabiliste” ou “model output statistics” (MOS).

3.4.1 L’approche “Cumulative Distribution Function - transform” (CDF-t)

Dans ce contexte, lors d’une collaboration avec la société “ClimPact” [21] et plus particulièrement avec P.-A. Michelangeli, j’ai développé une approche permettant de modéliser la CDF d’une variable locale (par ex., à une station) à partir de la CDF de la même variable mais à une plus grande échelle (telle que celle d’une maille GCM). Le but était ici de mener une étude – financée par l’ADEME – sur l’évolution du potentiel éolien en France au cours du $21^{ème}$ siècle, à partir de simulations du vent issues d’un ou de plusieurs modèles climatiques. Le vent étant spatialement potentiellement hétérogène, il était donc nécessaire de réaliser un changement d’échelle. Par ailleurs, nous n’avions pas explicitement le besoin de générer des séries temporelles mais le but était ici

21. http ://www.climpact.eu/

d'avoir un aperçu de l'évolution des principales caractéristiques de l'ensemble de la distribution des valeurs d'intensités locales (c-à-d., pas uniquement les valeurs moyennes) pour une période donnée, par exemple 2040-2070. J'ai alors développé une nouvelle approche de régionalisation des CDFs de vent pour différentes périodes futures. Cette approche originale a conduit à la publication [125] et est basée sur l'hypothèse qu'il existe une transformation mathématique T qui, appliquée à la CDF F_G du vent à grande échelle, permet de définir une CDF qui est aussi près que possible de la CDF F_S du vent à la station d'intérêt. Cette approche a logiquement été nommée "Cumulative Distribution Function - transform" (CDF-t).

Plus précisément, si F_{Gh} et F_{Sh} représentent les CDFs des intensités de vent modélisées à un point de grille du GCM et à une station contenue dans ce point de grille, pour une période historique h de calibration. Nous supposons que la transformation T permet de passer de F_{Gh} à F_{Sh} :

$$T(F_{Gh}(x)) = F_{Sh}(x), \tag{3.7}$$

pour tout x dans le domaine des valeurs possibles de la variable à représenter (ici, une intensité de vent dans $[0,+\infty[$). En remplaçant x par $F_{Gh}^{-1}(u)$ dans cette équation, où u est une probabilité dans $[0,1]$ et F_{Gh}^{-1} la fonction inverse de F_{Gh}, nous obtenons :

$$T(u) = F_{Sh}(F_{Gh}^{-1}(u)), \tag{3.8}$$

qui correspond à une définition simple de T. Cette transformation T est supposée stationnaire dans le temps et peut donc être appliquée à F_{Gf}, la CDF de la variable à grande échelle pour une période future ou de validation, pour générer F_{Sf}, la CDF au niveau de la station pour la même période :

$$T(F_{Gf}(x)) = F_{Sf}(x), \tag{3.9}$$

ce qui est équivalent à

$$F_{Sf}(x) = F_{Sh}(F_{Gh}^{-1}(F_{Gf}(x))). \tag{3.10}$$

Evidemment, si cette méthodologie a été développée pour la variable vent, elle est applicable (et a d'ailleurs déjà été appliquée) pour d'autres variables, voir section 3.4.2.

L'approche CDF-t peut être perçue comme une extension de la méthode de correction de biais dite par "quantile-quantile" (QQ) connue depuis plus de 50 ans en statistiques ([137, 71]) mais appliquée au climat uniquement depuis quelques années (par ex., [44]). En reprenant les mêmes notations, cette méthode QQ fonctionne suivant le principe suivant : à partir d'une réalisation (c-à-d., un quantile) x_G de la variable (le vent) à grande échelle, la méthode QQ consiste à trouver la valeur de vent locale x_S à la station (c-à-d., le quantile après correction) telle que la probabilité d'avoir à grande échelle une valeur inférieure à x_G soit la même que celle d'avoir une valeur inférieure à x_S à la station, autrement dit :

$$F_G(x_G) = F_S(x_S), \tag{3.11}$$

ce qui équivaut à

$$x_S = F_S^{-1}(F_G(x_G)). \tag{3.12}$$

Bien que CDF-t et QQ aient une philosophie commune de travail sur des CDFs, lors de son application en contexte de changement climatique, l'approche QQ projette les simulations à grande échelle pour le futur sur la CDF pour le présent pour calculer et apparier les quantiles, alors que CDF-t prend en compte les évolutions de la distribution grande échelle entre les deux périodes (calibration et application) pour la retranscrire à la CDF au niveau de la station. Ceci est certainement un avantage de CDF-t lors d'une utilisation en contexte de changement climatique.

À partir de F_{Sf} modélisée suivant l'Eq. (3.10), il est évidemment possible de générer une série temporelle. Pour cela, nous pouvons générer des données aléatoirement suivant F_{Sf}. Ces données sont alors simulées indépendamment, c-à-d., que la valeur au temps $t+1$ est indépendante de la valeur au temps t (même si de même loi). Une autre solution, celle que nous avons retenue pour les différentes études menées avec CDF-t (voir aussi section 3.4.2), est de réaliser une transformation quantile-quantile entre F_{Gf} et F_{Sf}. Dans la méthode QQ décrite ci-dessus aux Eqs. (3.11-3.12), l'appariement est fait entre les CDFs au présent (F_{Gh} et F_{Sh}) avec des données issues de la CDF future (F_{Gf}). Ici, l'application de QQ pour générer des valeurs suivant F_{Sf} est directement faite entre les CDFs du futur.

De plus, si l'approche QQ peut être réalisée de manière empirique (par ex., [44]) ou paramétrique (par ex., [144]), c'est également le cas de CDF-t, c-à-d. que les trois distributions F_Gh, F_Sh et F_Gf de l'Eq. 3.10 peuvent être modélisées au choix comme des fonctions de répartitions empiriques ou comme appartenant à des familles paramétriques choisies par l'utilisateur et reflétant une connaissance a priori de la variable et/ou des phénomènes à modéliser. Dans la suite, toutes les illustrations de ce chapitre sont basées sur des modélisations empiriques (c-à-d., par fonctions de comptage en escalier) de ces trois CDFs.

Avant de présenter en section 3.4.2 quelques illustrations d'évaluation et de projection de CDF-t, je tiens à revenir sur la distinction entre downscaling statistique et corrections de biais. Comme on vient de la voir, ces deux familles sont souvent très similaires. La frontière séparant ces approches est fine et la distinction se fait souvent essentiellement par leurs objectifs : changement d'échelle ou correction de loi statistique. Toutefois, quelques distinctions plus méthodologiques peuvent être faites. Classiquement, en downscaling, les prédicteurs utilisés sont des variables ou caractéristiques atmosphériques à grande échelle, telles que des régimes de temps, ou des champs de SST, température, vent ou humidité à différents niveaux de pression généralement entre 1000 et 500 mb. Les méthodologies de correction de biais, quant à elles, utilisent en prédicteur essentiellement (si ce n'est uniquement) la variable simulée par le modèle (de prévision météo, régionale, etc.) à corriger par rapport aux observations, comme c'est le cas pour QQ. Ainsi, sous cet aspect, l'approche CDF-t peut être perçue comme une méthode de correction de biais statistique. Cependant, dans les études menées ou en cours (voir section 3.4.2) à l'aide de CDF-t, bien que nous ayons utilisé à chaque fois un unique prédicteur (la variable d'intérêt), un changement d'échelle a néanmoins été accompli puisque CDF-t était calibrée pour passer de la CDF des valeurs d'une maille GCM, à la CDF des valeurs observées en une station. Techniquement parlant, CDF-t peut donc bien réaliser un downscaling. Remarquons par ailleurs que, dans l'article [98] qui sera

présenté plus en détails dans la section 4.2, l'approche CDF-t a été étendue non seulement dans le contexte des valeurs extrêmes mais aussi au cas des CDFs non-stationnaires. Dans ce dernier cas, les CDFs à grande échelle sont paramétriques avec des paramètres qui dépendent de variables atmosphériques (comme celles mentionnées quelques lignes plus haut). Dans ce sens, cette version non-stationnaire de CDF-t remplit les conditions "classiques" du downscaling statistique. En le formulant autrement, on peut également dire que la version non-stationnaire de CDF-t réalise une correction de CDF *conditionnelle* à des covariables atmosphériques.

Par ailleurs, l'approche CDF-t – que ce soit dans sa version empirique ou paramétrique –, ainsi que la génération de séries temporelles suivant F_{Sf} par QQ ont été codées en R (ainsi que les métriques de Kolmogorov-Smirnov et de Cràmer-von Mises permettant d'évaluer des distances entre CDFs) et mises sous la forme d'un package R nommé "CDF-t" librement téléchargeable depuis www.r-project.org/ ou depuis mon site web [22]. Ainsi, tous les calculs liés à cette régionalisation par CDF-t dans la suite de ce manuscrit ont été effectués grâce au package "CDF-t".

3.4.2 Une large gamme d'applications

L'objectif de cette section est d'illustrer brièvement la diversité des applications et études menées grâce à la version stationnaire de CDF-t. La version non-stationnaire de CDF-t ayant été développée dans le cadre de la théorie des valeurs extrêmes, celle-ci sera présentée au chapitre 4. Certains détails techniques sont ici omis pour éviter trop de descriptions fastidieuses mais peuvent être évidemment retrouvés dans les articles mentionnés dans cette section.

Évolution des intensités de vent sur la France d'ici à la fin du $21^{ème}$ siècle

La première illustration que je souhaite présenter est évidemment liée à l'étude [125] ayant initialement motivé le développement de CDF-t et portant sur la régionalisation des CDFs d'intensité du vent à 10m à partir de simulations climatiques pour le $21^{ème}$ siècle. Nous avions à notre disposition :

- les intensités de vent journalier à 10m observées en 26 stations réparties en France pour 1958-2005,
- des données NCEP/NCAR de réanalyses de vent pour la même période,
- ainsi que les intensités simulées par divers GCMs depuis les années 1950 jusqu'à 2100. Pour l'étude [125], seul le modèle climatique de l'IPSL a été utilisé pour illustrer la méthodologie et ses capacités.

Une première partie d'évaluation de CDF-t a tout d'abord été menée en la comparant à l'approche QQ grâce aux critères (non présentés ici) de Kolmogorov-Smirnov (KS) et de Cramèr-von Mises

22. Certaines extensions et améliorations sont régulièrement apportées et mises en ligne sur mon site : il est donc préférable de prendre ces versions ou même de me demander la dernière version.

(CvM) permettant d'évaluer des distances entre CDFs. Pour cela, des techniques de validation croisée et de bootstrap ont été employées et ont montré, entre autres, que, de manière globale, les approches CDF-t et QQ permettaient toutes les deux d'améliorer les CDFs de vent des données NCEP/NCAR et IPSL, pour l'ensemble des stations, avec toutefois de meilleurs résultats à partir des données NCEP/NCAR. Ceci est certainement du à une résolution spatiale “moins faible” que celle du modèle de l'IPSL et à des forçages plus réalistes. Si ces améliorations étaient relativement équivalentes entre QQ et CDF-t à partir des réanalyses (avec tout de même une très courte préférence pour CDF-t), à partir des sorties IPSL, elles étaient tout de même significativement plus marquées pour CDF-t que pour QQ avec respectivement $\sim$80% et $\sim$70% de CDFs considérées comme correctement reproduites. Nous avons alors employé CDF-t calibrée sur les données IPSL et les observations pour 1958-2005 pour régionaliser aux 26 stations les CDFs de vent générées par l'IPSL selon le scénario A2 pour trois périodes : 2006-2040, 2041-2070 et 2071-2100. La climatologie de référence 1958-2006 et les résultats de ces projections sont présentés en Fig. 3.18. Pour les 26 stations, les intensités de vent à 10 m diminuent assez nettement – entre 0.5% à 9% – durant le $21^{ème}$ siècle par rapport à 1958-2005. Les stations au nord-ouest voient de plus grandes diminutions (surtout en Bretagne) que celles au sud-est. Les changements en anomalies (rayons des cercles) ne présentent pas de structure géographique, même si pour 2071-2100, les stations du nord-ouest ont des changement plus prononcés. Remarquons tout de même que, quelle que soit la période, peu de stations ont des différences significatives de CDFs d'anomalies par rapport à la période historique (cercles en gras). Malgré tout, il serait hasardeux de conclure catégoriquement uniquement à partir d'un unique scénario et d'un unique modèle. Toutefois, ces résultats sont en accord avec d'autres études parues depuis (par ex., [127, 187]).

Rendements agricoles en Afrique de l'ouest

Lors d'une collaboration avec P. Oettli et B. Sultan (LOCEAN) et C. Baron (CIRAD), nous avons cherché à évaluer la capacité de divers modèles régionaux à être employés dans un contexte d'application agricole en Afrique de l'Ouest. Dans cette étude [131], un ensemble de neuf RCMs [23] (dont huit provenant du projet ENSEMBLE) a été évalué sur 12 stations au Sénégal pour lesquelles nous disposions d'observations climatiques. Pour ces neuf modèles, les sorties utilisées par la suite étaient : (i) la précipitation, (ii) la température, (iii) le rayonnement solaire et (iv) l'évapotranspiration potentielle (PET) calculée à l'aide de l'équation de Penman–Monteith ([1]) à partir des sorties de valeurs minimales et maximales de température journalières, d'intensité moyenne du vent, d'humidité relative moyenne de l'air et de rayonnement solair à 2m des RCMs. Les sorties de ces RCMs pour les 12 stations ont été mises en entrées du modèle de rendement agricole SARRA-H pour simuler l'une des principales cultures de cette région : le sorgho. Bien qu'ils soient forcés par les mêmes conditions à grande échelle provenant des réanalyses ERA-Interim (ERA-I), les performances des RCMs pour reproduire les variables les plus cruciales à utiliser en entrées de modèles de rendement agricole, sont extrêmement variables. Ceci entraine alors une

23. HIRHAM (DMI), CLM (GKSS), HadRM3P (HC), RegCM (ICTP), RACMO (KNMI), HIRHAM (METNO), REMO (MPI), RCA (SMHI), PROMES (UCLM)

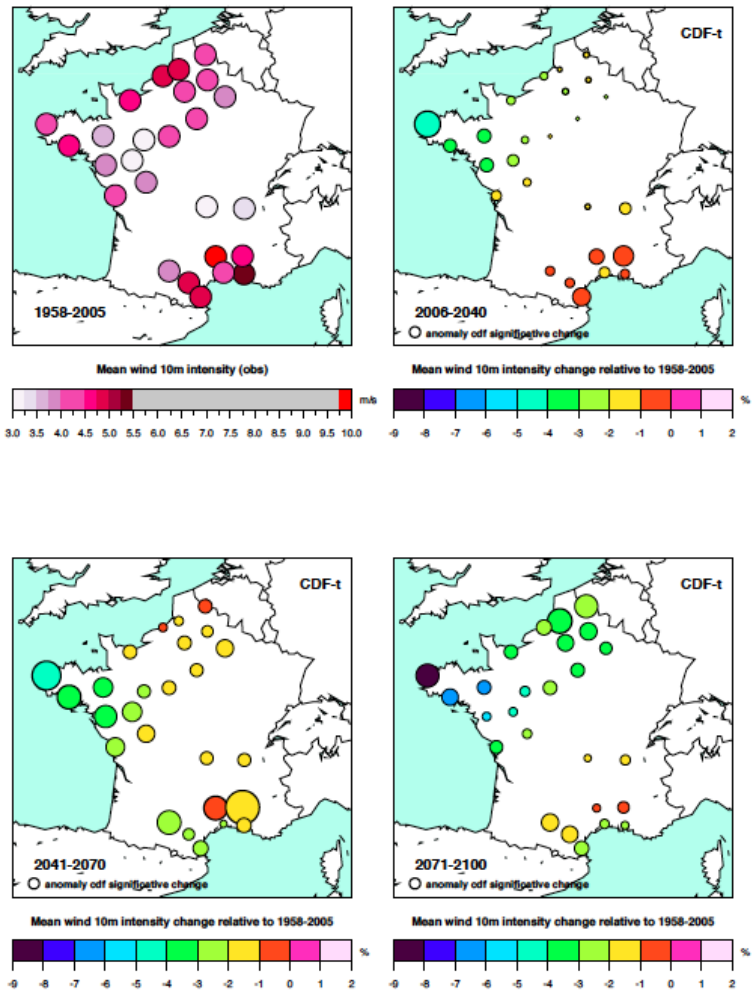
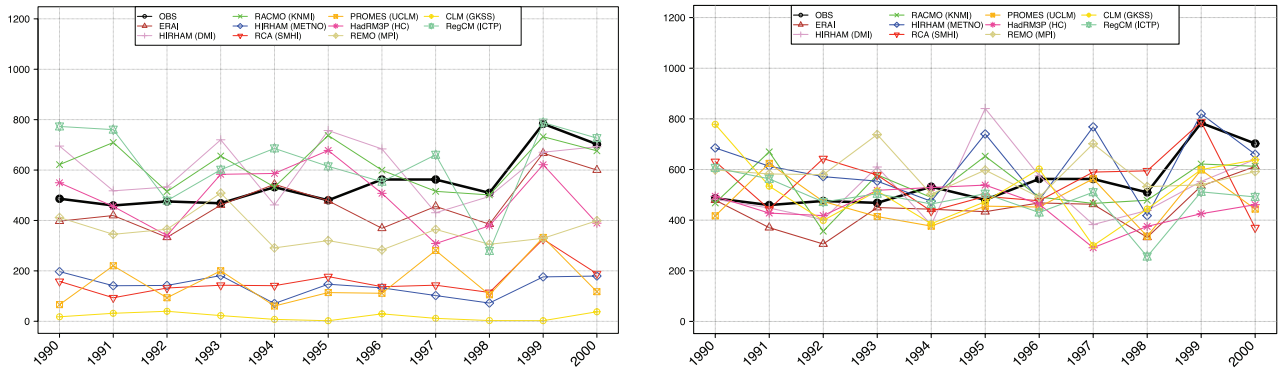


FIGURE 3.18: En haut à gauche : Climatologie des intensités de vent à 10m pour 1958-2005. Les autres panels montrent les changements entre 1958-2005 et les trois périodes futures (2006-2040, 2041-2070 and 2071-2100) : les couleurs correspondent au changement du vent moyen par rapport à 1958-2005 ; le rayon des cercles est proportionnel à la valeur du critère CvM entre les anomalies de vent pour 1958-2005 et les anomalies pour les périodes futures ; les cercles en gras correspondent aux stations où les CDFs des anomalies futures sont significativement différentes des CDFs pour 1958-2005 (α=0.05).

grande dispersion des simulations de rendement d'un RCM à un autre, visible par exemple dans la Fig. 3.19(a) présentant les rendements annuels moyennés sur les 12 stations pour chaque RCM (en couleurs), où le tracé noir correspond au rendement calculé par SARRA-H à partir des données météorologiques observées. Cette dispersion très nette provient de différentes physiques et des choix de paramétrisation dans chaque RCM.



(a) (b)

FIGURE 3.19: Rendements annuels moyennés sur les 12 stations pour chaque RCM (en couleur) ; le tracé noir correspond au rendement calculé par SARRA-H à partir des données météorologiques observées. (a) : Les rendements sont calculés à partir des sorties des RCMs ; (b) : Les rendements sont calculés à partir des sorties des RCMs corrigées statistiquement par CDF-t.

Nous avons alors appliqué CDF-t sur l’ensemble des variables climatiques nécessaires en entrées de SARRA-H afin d’essayer de corriger les principaux biais des RCMs. Les rendements simulés à partir des variables corrigées présentaient alors des comportements beaucoup plus réalistes, tels qu’illustrés en Fig. 3.19(b) pour les rendements annuels moyennés sur les 12 stations, où l’on voit une réduction marquée de la dispersion, correctement positionnée autour des simulations issues des observations. Bien que ce ne soit ni présenté, ni discuté dans ce manuscrit, d’autres évaluations (basées par exemple sur une analyse de sensibilité à la qualité des variables d’entrées ; les cycles saisonniers des variables climatiques et des simulations de rendements ; des diagramme de Taylor des variables climatiques ; les biais et déviations standards) ont confirmé l’apport et les résultats prometteurs de CDF-t dans un tel contexte d’application agricole.

Cependant, la variabilité inter-annuelle des simulations reste encore à améliorer : les corrections par CDF-t n’ont pas permis d’améliorer les variations “year-to-year” des variables climatiques et donc des rendements simulés. En effet, les transformations apportées par CDF-t portent sur la distribution statistiques des variables et non sur leur chronologie. Cette particularité n’est cependant pas propre uniquement à CDF-t et s’avère une faiblesse potentielle de très nombreuses méthodologies de downscaling, dynamiques ou statistiques.

Enfin, cette étude a également confirmé l’importance de l’approche multi-modèles pour quantifier les incertitudes dans les études d’impacts, et a souligné les bénéfices de combiner les approches dynamiques et statistiques de régionalisation.

Une grande diversité d'apports

Pour terminer cette section concernant CDF-t, je souhaite lister très brièvement les différentes études actuellement en cours ou soumises (au moment de l'écriture de ce manuscrit) et appliquant CDF-t. Celles-ci devraient évidemment donner lieu à publications dans les semaines/mois à venir.

Tout d'abord, comme mentionné en section 3.3.1, mes collaborations en modélisation statistique pour l'écologie et pour lier climat et écologie se poursuivent actuellement pour caractériser les présences/absences de diverses espèces de poissons d'eau douce dans la sud de la France et leurs évolutions spatiales et temporelles sous contraintes de changements climatiques. Or, si les modèles d'arbres "boostés" agrégés (ABT) en sont le coeur, leurs sorties nécessitent – tout comme les RCMs dans l'illustration précédente – des corrections de biais en température et précipitation, entres autres. Ces corrections sont effectuées à l'aide de CDF-t (ici dans sa version paramétrique) de manière saisonnière, avec une définition des saisons liée aux différentes périodes de vie des poissons (repos : octobre à février ; reproduction : mars à juin ; croissance : juillet à septembre). L'approche CDF-t permet ainsi de corriger les simulations des ABTs en se rapprochant au plus près des spécificités nécessaires à l'étude sur la bio-diversité aquatique. Ces travaux ont été soumis récemment pour détailler les méthodologies et modèles impliqués ([177]) et les résultats en termes d'impacts spatiaux et temporels des changements climatiques sur la bio-diversité aquatique ([178]).

CDF-t a également été récemment appliquée comme outil de régionalisation statistique lors d'une collaboration avec N. Vigaud et Y. Caballero (BRGM, Montpellier) dans le cadre du projet ANR SHIVA (Socio-economicAssessment of the rural Vulnerability of water users under stressors of global change in the Hard Rock area of South India). Le but était ici de simuler des champs de températures et de précipitations journalières à haute résolution, $1^o \times 1^o$ pour la température et $0.5^o \times 0.5^o$ pour la précipitation, sur l'ensemble de l'Inde, soient 210 et 729 points de grille respectivement. Après une évaluation de CDF-t par validation croisée sur une période historique, CDF-t a permis de régionaliser les changements climatiques selon le scénario A2, pour la période 2046-2065, à partir de cinq GCMs [24]. Les résultats ont montré une augmentation des précipitations durant la saison de mousson Indienne alors que les précipitations hivernales ont tendance à diminuer. Ces projections suggèrent également un réchauffement global de la région Indienne, particulièrement en hiver et dans la période post-mousson. Si cette étude ne cherche pas à aller très profondément dans l'analyse des processus climatiques et continentaux impliqués dans ces changements, elle montre la capacité de CDF-t à pouvoir traiter de très grandes quantités de données et surtout à pouvoir générer des cartes à haute résolution pour l'ensemble de la région Indienne, en des temps très raisonnables (quelques heures de calculs). De plus, l'article [189] soumis récemment est une étape préliminaire à des études d'impacts du changement climatique en Inde. En effet, les champs simulés sont actuellement utilisés en entrées de modèles hydrologiques sur l'Inde par les hydrologues Français et Indiens impliqués dans le projet SHIVA pour mener des études liées à

24. CGCM3, CNRM3, ECHAM5, BCCR2, MIROCMR

l’évolution des ressources en eau en Inde au cours du 21*ème* siècle.

Par ailleurs, dans une collaboration récente ([182]) avec M. Troin, F. Sylvestre, C. Vallet-Coulomb (CEREGE), M. Khodri (LOCEAN) et E. Piovano (Universidad Nacional de Córdoba), CDF-t a été employée pour régionaliser des simulations journalières de températures minimales et maximales et de précipitations issues de LMDz ou de réanalyses NCEP/NCAR. Cette régionalisation s’est faite sur six stations réparties dans le bassin Sali-Dulce en Argentine sur la période 1950-2005 pour deux stations et sur 1973-2005 pour quatre stations. Les simulations locales ainsi obtenues ont ensuite été mises en entrées du modèle hydrologique Soil Water Assessment Tool (SWAT) afin de simuler les décharges (c-à-d., les sorties ou volumes d’eau transportée par un fleuve en un certain temps) du fleuve Rio Sali-Dulce et de les comparer aux observations disponibles. Ces simulations de décharges par SWAT ont alors été utilisées pour forcer un modèle de lac simulant le niveau du lac Laguna Mar Chiquita qui a connu une montée abrupte de son niveau dans les années 1970, montée attribuée à une augmentation de décharge du Rio Sali-Dulce. Bien que ceci ne soit correctement reproduit dans aucune des simulations effectuées dans cette étude, les cycles hydrologiques clés des niveaux du lac étaient mieux capturés en utilisant CDF-t qu’en utilisant les données brutes. En plus de résultats d’analyses de processus climatiques et hydrologiques non détaillés ici, cette étude montre que cette approche intégrée “climat-downscaling-bassin-lac” originale est très prometteuse pour comprendre et simuler les variations à longs termes de niveaux de lac.

Enfin, CDF-t est également impliquée dans des études (en cours ou soumises) d’intercomparison de méthodes de régionalisation dynamiques et/ou statistiques :

- Intercomparaison de downscaling par CDF-t à partir de réanalyses ERA-40, et à partir de trois GCMs en période historique, puis pour le futur selon le scénario A2 pour une dizaine de stations dans le sud de la France ([111]),
- Intercomparaison de CDF-t avec trois RCMs (ALADIN, LMDz, WRF) pour le climat présent sur une dizaine de points de grille issus des données réanalysées du modèle SAFRAN ([113, 146]) dans le cadre du projet ANR MedUP ([194]),
- Corrections de biais statistiques par CDF-t de deux RCMs sur l’ensemble de la France, dans un but d’évaluation de la valeur ajoutée au présent et projections climatiques à haute résolution pour le scénario A2 ([188]).

L’ensemble des illustrations d’applications de CDF-t mentionnées ici montre bien l’attrait que possède CDF-t : facile d’utilisation grâce au package R “CDF-t”, rapide à tourner même sur de très gros jeux de données, fournissant des simulations fiables et tout à fait comparables (si ce n’est parfois meilleures) que des modèles statistiques et/ou dynamiques plus sophistiqués. Toutes ces qualités font que CDF-t est très flexible et peut donc être utilisée dans de très nombreux contextes.

3.5 Deux ou trois mots de bilan sur le downscaling statistique

Ces quelques sections ont montré – de manière bien sûr non-exhaustive – la diversité des approches possibles de régionalisation statistique mais aussi le large spectre d'applications de ces méthodes et modèles. Si la simulation de données à haute résolution est le but premier du downscaling, on a pu voir, par exemple aux sections 3.3.1 et 3.3.3, que les mêmes modèles statistiques liant climat à grande échelle et phénomènes météo locaux, peuvent aussi être pertinents pour représenter le lien entre climat et divers phénomènes environnementaux directement.

De manière très personnelle, il est vrai que les approches stochastiques et par MOS gardent ma préférence. Bien qu'elles reposent parfois sur des connaissances a priori ou des choix de l'utilisateur/modélisateur, par exemple la famille de distributions à calibrer, restreignant ainsi la marge de manoeuvre et de projections de ces approches [25], leur capacité à fournir, au final, une distribution de probabilité (conditionnelle à la grande échelle – comme en section 3.2 – ou caractéristique d'une période – comme en section 3.4 –, etc.) et non une unique valeur (souvent moyenne) comme la plupart des méthodes par fonction de transfert, est un avantage précieux qui permet d'avoir accès à de riches informations statistiques.

Enfin, je tiens à signaler que depuis maintenant quelques années, chaque fois qu'une étude a nécessité des développements méthodologiques ou de modèles originaux, ceux-ci ont été codés en R puis mis sous la forme d'un packages R au moment de la publication de l'étude. La mise à disposition gratuite et automatique à la communauté scientifique de ces packages R permet non seulement une vérification des modèles mis en oeuvre mais aussi une dissémination "culturelle" des méthodes et une utilisation facilitée – et même parfois optimale – des méthodes développées. Cette étape de mise à disposition du savoir-faire philosophique et technique est, pour moi, certainement aussi importante que la publication elle-même.

25. ce qui peut aussi être un avantage en évitant certaines projections aberrantes, par exemple en contexte de climat changeant.

4 Événements extrêmes et downscaling statistique

... la valeur d'un hasard est égale à son degré d'improbabilité.
Milan Kundera, L'immortalité (1993)

Mes contributions à l'étude des phénomènes extrêmes en environnement sont relativement récentes mais hétéroclites. Mes apports et collaborations dans ce domaine ont par exemple porté sur une méthode définissant des régimes de champs spatiaux à grandes déviations par rapport à la moyenne ([10]) ; un modèle statistique auto-régressif pour la modélisation de maxima journaliers et hebdomadaires de concentration en méthane ([181]) ; ou sur une participation à l'écriture d'un article "de review" sur les modèles statistiques pour les événements extrêmes dans le climat et l'environnement ([65]). Toutefois, je ne détaillerai pas ces travaux dans ce manuscrit afin de me concentrer davantage sur les développements que j'ai pu réaliser pour la régionalisation des valeurs extrêmes. En effet, nous avons pu voir au chapitre 3 que les méthodologies de descente d'échelle étaient particulièrement importantes pour pouvoir fournir des entrées climatiques à différents modèles dits d'impacts (hydrologiques, écologiques, économiques, etc.) à des échelles pertinentes. Or, ces modèles sont très souvent utilisés pour essayer de comprendre et même si possible anticiper les impacts multiples des événements rares mais souvent destructeurs ou très coûteux : les événements extrêmes. Il est alors nécessaire de disposer d'outils statistiques de modélisation de tels événements lors d'un processus de régionalisation. Je vais donc présenter dans ce chapitre trois extensions de modèles statistiques de downscaling (présentés au chapitre 3) au cas des événements extrêmes.

4.1 Brefs rappels sur la théorie des valeurs extrêmes

L'une des principales particularités de l'analyse des valeurs extrêmes est que, par définition, les événements extrêmes sont rares : nous disposons généralement de peu de données observées pour calibrer une distribution statistique. Cette dernière est par ailleurs différente des lois de valeurs plus courantes puisqu'elle doit pouvoir représenter de très fortes valeurs associées à des probabilités d'occurrence très faibles. De par la rareté des extrêmes, ces *quantiles* élevés doivent habituellement être extrapolés. C'est le cas pour l'exemple classique des infrastructures hydrologiques (barrages, digues) qui sont construites pour pouvoir résister à des événements théoriques pas nécessairement observés : il faut par exemple calculer un niveau de retour millenal, voire cinq-

millenal, à partir de données portant uniquement sur quelques décennies au mieux. Pour réaliser de telles extrapolations de manière solide et surtout associée à une évaluation de leurs incertitudes, la théorie des valeurs extrêmes (EVT en anglais) est une aide très précieuse en fournissant un cadre théorique mathématique développé pour la première fois par [51]. Cette théorie s'est considérablement développée (par ex., voir [35]) et est désormais utilisée dans de nombreux domaines allant de l'hydrologie ([100]) à l'économie ([48]) en passant bien sûr par le climat et l'environnement ([65]). Pour aider à comprendre mes apports dans ce contexte, je rappelle ici les deux distributions statistiques asymptotiques qui forment la base de l'EVT univariée ([35]).

Dans le cas de données représentant les valeurs maximales (par exemple pour un intervalle de temps fixe) de réalisations $(X_1,...,X_n)$, l'EVT indique que la fonction de répartition cumulée est :

$$\mathbb{P}(M_n \leq z) = \exp\left\{ -\left[1+\gamma\left(\frac{z-u}{\lambda}\right)\right]_+^{-1/\gamma} \right\} \tag{4.1}$$

où $M_n = \max(X_1,...,X_n)$, $(a)_+ = \max(0,a)$, λ est le paramètre d'échelle et γ est le paramètre de forme.

Dans cette Eq. (4.1), le signe du paramètres de forme γ indique comment la queue de cette distribution se comporte, autrement dit, si la distribution GEV peut représenter et générer des valeurs particulièrement grandes : on parle alors de *distribution à queue lourde*. Plus précisément, pour une GEV :

- Si $\gamma < 0$: la queue est dans le domaine d'attraction de la loi de Weibull, la distribution est à queue bornée ;
- Si $\gamma = 0$: la queue est dans le domaine d'attraction de la loi de Gumbel, la distribution est à queue "légère" (ou à décroissance rapide) ;
- Si $\gamma \geq 0$: la queue est dans le domaine d'attraction de la loi de Fréchet, la distribution est à queue lourde.

Dans le cas de données correspondant à des excès d'une variable aléatoire X au-delà d'un seuil u fixé suffisamment haut, la théorie des valeurs extrêmes indique que la loi de ces excès $Y = X - u$ est une distribution de Pareto Généralisée (GPD) définie par :

$$\mathbb{P}(X-u \leq y | X > u) = 1 - \left(1+\frac{\xi y}{\sigma_u}\right)_+^{-1/\xi}, \tag{4.2}$$

où σ_u est le paramètre d'échelle et ξ est le paramètre de forme. De la même manière que pour la GEV, le signe du paramètre ξ indique comment la queue de la GPD se comporte :

- Si $\xi < 0$: la distribution est à queue bornée ;
- Si $\xi = 0$: la distribution est à queue légère ;
- Si $\xi \geq 0$: la distribution est à queue lourde.

En travaillant sur le downscaling statistique des extrêmes, nous sommes confrontés à la diminution de la taille des échantillons observés et à l'affaiblissement du lien théorique à l'augmentation des gaz à effet de serre. Néanmoins, tout un corpus d'études à émergé pour tenter de comprendre les impacts du changement climatique sur les événements extrêmes locaux, par l'intermédiaire d'approches de type downscaling (par ex., [163, 133, 72, 41, 52, 59, 22]). Certaines de ces études étaient même basées sur l'EVT (par ex., [64, 8, 117]). Mes travaux en downscaling statistique des valeurs extrêmes se situent évidemment dans ce cadre théorique et, dans la suite de ce chapitre, mes apports portent sur la GPD et non la GEV.

Hypothèses sous-jacentes en contexte de changement climatique

Les sections suivantes portent sur l'insertion de distributions GPD dans des modèles stochastiques ou MOS de downscaling. Dans ces approches, les nouvelles distributions incorporant une ou des GPDs sont modélisées de manière non-homogène, c-à-d., avec des paramètres qui évoluent en fonction de *covariables* (prédicteurs atmosphériques), qui peuvent évidemment elles-mêmes évoluées avec le temps en contexte de changement climatique (voir section 3.2).

Le paramètre de forme étant généralement associé à une très lourde incertitude, celui-ci est souvent supposé ne pas évoluer dans le temps et est donc modélisé comme une constante (qui reste cependant à calibrer en climat présent). Dans un tel contexte, seul le paramètre d'échelle a ainsi la possibilité d'évoluer en fonction de covariables. Cette spécificité de modélisation relativement classique en EVT en contexte de changement climatique fait que les valeurs (par exemple de précipitation) associées à un niveau de retour donné peuvent changer selon la période future que l'on étudie. Cependant, le paramètre de forme n'évoluant pas, la structure même de la queue de distribution (bornée, légère ou lourde), elle, ne changera pas. Toutefois, certains modèles permettent au paramètre de forme d'évoluer dans le temps, ou tout au moins en fonction du moment de l'année où la densité est à estimer. C'est par exemple le cas dans l'approche [27] où l'ensemble des précipitations journalières de l'année sont représentées par le même modèle faisant varier les paramètres en partie grâce à des variables de dates. Cela autorise ainsi à avoir des propriétés de précipitations extrêmes (par ex., queues de distribution, niveaux de retour, etc.) distinctes en été et en hiver, capturant alors des caractéristiques climatiques différentes. De plus, modulo les tests nécessaires pour évaluer la significativité de leurs évolutions, il est parfois intéressant d'étudier des paramètres d'échelle pouvant évoluer – par ex., grâce à des covariables – en contexte de changement climatique. Cette évolution (si elle est significative malgré les incertitudes associées) peut amener des informations précieuses sur des changements potentiels de la structure même des queues de distributions de précipitation.

4.2 XCDF-t : Approche "CDF-t" pour les eXtrêmes

Je présente ici les principaux développements effectués lors de l'article [98] écrit en collaboration avec M. Kallache (postdoc LSCE), P. Naveau (LSCE) et P.-A. Michelangeli (ClimPact). Ces développements méthodologiques sont de deux ordres :

- Tout d'abord, nous avons étendu l'approche CDF-t (présentée en section 3.4) au cas des valeurs extrêmes de type excès au-dessus d'un seuil, avec une formulation paramétrique basée sur des GPDs.
- Nous avons ensuite introduit dans cette formulation la possibilité d'avoir des covariables. Bien que cette dernière extension puisse également être réalisée dans un contexte différent de celui de l'EVT, seul celui-ci est présenté dans cette section.

Ces deux extensions ont été illustrées sur cinq stations météorologiques dont les données de précipitation proviennent du projet "European Climate Assessment & Dataset" (ECA&D [1]) : Marseille, Perpignan, Mont-Aigoual, Nîmes et Sète. Ces données journalières couvraient les mois de novembre à mars depuis 1951 jusqu'à 1999. Dans la suite de cette section, 1951-1985 correspondra à la période de calibration et 1986-1999 à la période de projection pour l'évaluation.

La formulation de l'approche CDF-t est redonnée pour mémoire : Si F_{Gh} et F_{Sh} représentent les CDFs d'une variable, modélisées respectivement à un point de grille GCM et à une station contenue dans ce point de grille, pour une période historique h de calibration, la transformation T permettant de passer de F_{Gh} à F_{Sh} est formulée comme :

$$T(u) = F_{Sh}(F_{Gh}^{-1}(u)), \tag{4.3}$$

où u est une probabilité et F_{Gh}^{-1} est la fonction inverse de F_{Gh}. Appliquée à F_{Gf} – la CDF de la variable à grande échelle pour une période future (ou de validation) – T génère F_{Sf} – la CDF au niveau de la station pour la même période – par :

$$F_{Sf}(x) = F_{Sh}(F_{Gh}^{-1}(F_{Gf}(x))). \tag{4.4}$$

4.2.1 Modélisation par GPDs

Lors des travaux [98], nous avons considéré (a) une variable X à modéliser représentant des excès au-dessus d'un seuil et donc (b) que les trois distributions du membre de droite de la formulation (4.4) étaient des GPDs de paramètres (ξ_{Sh}, σ_{Sh}) pour F_{Sh}, (ξ_{Gh}, σ_{Gh}) pour F_{Gh} et (ξ_{Gf}, σ_{Gf}) pour F_{Gf}. Cette extension de CDF-t basée sur la théorie de valeurs extrêmes a été nommée "XCDF-t". Notons u_{Sh} le seuil associé aux observations à la station S pour la période historique h. De même, notons respectivement u_{Gh} et u_{Gf} les seuils associés aux simulations à grande échelle pour la période h et la période future (ou de projection). Ces seuils peuvent être déterminés séparément les uns des autres et doivent être suffisamment hauts pour que les excès suivent une GPD ([35])

1. http ://eca.knmi.nl

mais suffisamment bas pour avoir assez d'excès pour l'analyse. Pour cela, nous avons employé la méthode du “mean residual life plot” (voir [35] par exemple) pour déterminer les seuils et nous avons vérifié que les excès associés suivaient bien une GPD grâce au test d'Anderson-Darling ([34]). Le seuil u_{Sf} associé à la station pour la période de projection est déduit des trois autres seuils en sommant à u_{Sh} l'écart entre les deux seuils à grande échelle : $u_{Sf} = u_{Sh} + (u_{Gf} - u_{Gh})$. Dans le cas où les trois CDFs F_{Sh}, F_{Gh} et F_{Gf} ont les mêmes valeurs de paramètres, il est trivial que F_{Sf} est une GPD avec les mêmes paramètres. Sinon, F_{Sf} a une structure plus complexe, définie en insérant les trois GPDs dans l'équation (4.4) :

$$F_{Sf}(x) = 1 - \left(1 + \frac{\xi_{Sh}}{\xi_{Gh}} \frac{\sigma_{Gh}}{\sigma_{Sh}} \left[\left(1 + \frac{\xi_{Gf}}{\sigma_{Gf}} (x + \kappa) \right)^{\frac{\xi_{Gh}}{\xi_{Gf}}} - 1 \right] \right)^{-\frac{1}{\xi_{Sh}}}, \tag{4.5}$$

pour $x > 0$ et où κ est un facteur de correction correspondant à une “inflation” des données. En effet, il est classique de transformer les données à grande échelle par *translation* (shift) et *inflation* avant l'analyse. La translation permet de se ramener à $u_{Gh} = u_{Sh}$ et l'inflation permet d'avoir approximativement le même domaine pour F_{Sh} et F_{Gh}. Autrement, les bords de la CDF F_{Sf} peuvent ne pas être définis. Les détails techniques de cette transformation ne sont pas présentés dans ce manuscrit. Le lecteur intéressé pourra se référer à l'appendice B de [98]. Par ailleurs, la sensibilité des CDFs F_{Sf} obtenues avec ou sans cette transformation a été étudiée au travers de simulations (non montrées). Les différences finales étaient très petites et même négligeables dans le cas où les données à la station et à grande échelle avaient approximativement le même domaine de valeurs.

Dans la suite, nous avons choisi de poser $\xi_{Gf} = \xi_{Gh}$. La complexité du modèle est alors réduite et l'étape d'inflation des données n'est (techniquement) plus nécessaire [2]. nous obtenons alors :

$$F_{Sf}(x) = 1 - \left(1 + \frac{\xi_{Sh}}{\sigma_{Gf}} \frac{\sigma_{Gh}}{\sigma_{Sh}} (x + \kappa) \right)^{-\frac{1}{\xi_{Sh}}}. \tag{4.6}$$

Les résultats sont alors plus robustes : de petites erreurs d'estimations de ξ_{Gf} dans l'équation (4.5) pouvaient générer de larges différences pour le quotient ξ_{Gh}/ξ_{Gf} et donc dans la CDF finale. Cette erreur potentielle est ici supprimée dans l'équation (4.6). En comparant Eqs. (4.2) et (4.6), il est clair que F_{Sf} est une GPD de paramètre d'échelle $\sigma_{Sf} = \sigma_{Sh}(\sigma_{Gf}/\sigma_{Gh})$ et de paramètre de forme $\xi_{Sf} = \xi_{Sh}$. D'un point de vue pratique, nous avons restreint cette modélisation au cas des CDFs à queue lourde uniquement (c-à-d., $\xi > 0$). Bien que des paramètres de forme négatifs puissent occasionnellement être estimés (par ex., [180, 116]), cette condition est généralement vérifiée pour des données de débits hydrologiques ou de précipitations (par ex., [100]). Du point de vue de l'estimation, les paramètres de l'équation (4.6) sont estimés par maximum de vraisemblance. Une représentation de la GPD avec paramètres orthogonaux par rapport à la matrice d'information de Fisher a été utilisée pour des résultats plus stables, c-à-d., les paramètres (ν, ξ) avec $\nu = \sigma(1+\xi)$ (voir [32]).

2. En pratique, elle peut cependant s'avérer nécessaire pour éviter des CDFs finales trop “compressées”.

4.2.2 *Insertion de covariables*

L'insertion d'information supplémentaire a été effectuée au travers de covariables qui ont été reliées aux paramètres ν et ξ. Ces paramètres ont alors la possibilité d'évoluer avec les évolutions des covariables, définissant ainsi un modèle XCDF-t non-homogène. L'utilisation de distributions non-homogènes [3] de valeurs extrêmes pour des études liées au changement climatique devient relativement populaire. Par exemple, [53] ont employé ce type d'approche pour la détection de changements extrêmes dans les modèles climatiques ; [204] pour projeter les hauteurs de vagues dans l'Atlantique nord ; [117] ont intégré le temps comme covariable pour représenter la saisonnalité de précipitations extrêmes, ainsi que différentes covariables de circulation (force du vent, direction et vorticité) au travers de fonctions paramétriques ou non, linéaires ou non. De plus amples détails sur des GPDs dont les paramètres dépendent de covariables peuvent être trouvés par exemple dans [102, 128, 134]. Les covariables peuvent représenter différentes influences sur les précipitations extrêmes, telles que des influences synoptiques ou au contraire très locales comme la convection en montagne. Les covariables ont donc été sélectionnées séparément pour chaque station à partir d'un ensemble de covariables plausibles.

L'idée de cette approche non-homogène (ou non stationnaire) est d'appliquer XCDF-t à chaque pas de temps et non pour l'ensemble de la période comme précédemment dans l'approche sans covariable. En effet, l'insertion de covariables permet de définir, pour chacune des trois CDFs, des paramètres (σ, ξ) différents à chaque pas de temps, et donc des CDFs différentes à chaque pas de temps. Pour un instant donné, les covariables permettent de fixer les paramètres et XCDF-t est alors appliqué. Pour cela, les paramètres des CDFs F_{Sh} et F_{Gh} sont estimés par liaison avec les covariables de la période de calibration. Si $C^t = (c_1^t, ..., c_n^t)$ est un vecteur de n covariables au temps t de la période de projection, les paramètres de F_{Gf} sont directement estimés grâce à ce vecteur pour définir $F_{Gf}(.; \sigma_{Gf}(C^t), \xi_{Gf}(C^t))$. Pour appliquer XCDF-t au temps t de la période de projection, un Quantile-mapping ([71]) est tout d'abord effectué pour relier les valeurs des covariables en période de projection à leurs valeurs (quantiles) correspondantes en période de calibration. Les valeurs résultantes $\tilde{C}^t = (\tilde{c}_1^t, ..., \tilde{c}_n^t)$ sont les valeurs des covariables en période de calibration avec les mêmes conditions de probabilités que C^t dans la période de projection. Les fonctions de lien permettent alors d'obtenir $F_{Sh}(.; \sigma_{Sh}(\tilde{C}^t), \xi_{Sh}(\tilde{C}^t))$ et $F_{Gh}(.; \sigma_{Gh}(\tilde{C}^t), \xi_{Gh}(\tilde{C}^t))$. XCDF-t est ensuite appliquée à chaque pas de temps t de la période de projection, ce qui conduit à une CDF F_{Sf}^t variant dans le temps.

Nous avons choisi un ensemble de covariables potentielles provenant principalement de réanalyses NCEP et à partir duquel les covariables finales (qui peuvent être différentes d'une station à une autre) ont été sélectionnées. Cet ensemble comprend : des champs de pressions au niveau de la mer (SLP) et leurs composantes principales (PC), des champs de hauteurs géopotentielles à 850 mb (Z850) et leurs PCs, les valeurs de température maximale (tmax) à la station, les intensités et directions de vent à différents niveaux de pression (850, 100 et 10 mb) ainsi que la différence de ces intensité et directions entre 10 et 850 mb, la précipitation (à grande échelle), ainsi qu'une

3. Généralement nommées "non-stationnaires" dans la littérature.

variable de temps (dans le cas où une évolution des paramètres serait significative mais non reliée aux covariables choisies).

Les covariables finales ainsi que les fonctions de liens ont été sélectionnées par l'intermédiaire de la statistique de *déviance* (voir [38] par exemple) qui permet de trouver un modèle pertinent tout en évitant une sur-paramétrisation (overfitting). Après différents essais et différentes comparaisons basées sur la déviance, la fonction de lien exponentielle a été retenue pour le paramètre ν :

$$\nu_t = \exp(a_0 + a_1 c_1^t + ... + a_n c_n^t) \tag{4.7}$$

alors que les paramètres ξ ont été laissés constants dans le temps.

Pour les cinq stations, les covariables retenues sont : Marseille : précipitation, Z850, dir. du vent à 850 mb ; Perpignan : 3PC de SLP ; Mont-Aigoual : tmax, SLP ; Nîmes : SLP, 3PC de Z850 ; Sète : 3PC de SLP, int. du vent à 850 mb. Les différentes valeurs des paramètres pour les approches avec ou sans covariables ne sont pas données dans ce manuscrit (voir [98]). Cependant, afin d'illustrer la valeur ajoutée de l'utilisation de covariables, nous avons comparé en Fig. 4.1 les CDFs empiriques des observations de la période de projection avec les CDFs F_{Sf} obtenues par l'approche sans covariable (en noir) ainsi que les CDFs F_{Sf}^t obtenues par l'approche avec covariables (en gris) pour différents temps t de la période de projection. Les différences entre CDFs empiriques et celles modélisées par l'approche sans covariables peuvent être dues à une non-stationnarité des observations qui peut être capturée par une approche non-homogène telle que la notre. En effet, pour toutes les stations, les surfaces couvertes par les lignes grises capturent complètement les CDFs empiriques de vérification. Si ce point n'est en aucun cas une preuve de la supériorité de l'approche avec covariables par rapport à celle sans, c'est tout de même une indication de la valeur ajoutée potentielle de notre modèle non-homogène.

Afin de ne pas alourdir ce manuscrit avec trop de figures et de détails, j'ai décidé de ne pas présenter davantage de résultats issus de ces modèles XCDF-t (QQplots, comparaisons avec Quantile-mapping, critères numériques de qualité des simulations, etc.). En effet, le but de cette section était essentiellement de présenter les principaux développements théoriques et méthodologiques associés à ces travaux collaboratifs. Cependant, pour terminer cette section sur XCDF-t, j'aimerais attirer l'attention du lecteur – et peut-être futur utilisateur de cette approche – sur un aspect important de l'application de l'approche XCDF-t avec covariables. La version non-homogène de XCDF-t (ou d'un autre modèle CDF-t paramétrique non-homogène introduisant des paramètres conditionnels) nécessite que XCDF-t soit calibrée dans un contexte de “perfect-prog”, c-à-d. avec des prédicteurs à grande échelle pouvant être considérés comme *réels* afin d'avoir des correspondances journalières correctes entre grande et petite échelles, par ex., les réanalyses (voir introduction du chapitre 3 ou [118, 98]). Autrement dit, en fonction de l'objectif d'application de XCDF-t, cette approche non-homogène peut nous faire perdre l'un des avantages de CDF-t stationnaire. En effet, en contexte MOS (c-à-d., sans covariables supplémentaires), CDF-t peut être directement calibrée à partir des sorties de GCM. Ceci permet alors d'enlever une part importante de l'incertitude liée à la modélisation puisqu'aucune calibration sur des réanalyses n'est nécessaire avant d'appliquer CDF-t à des GCMs : l'inadéquation potentielle des sorties de

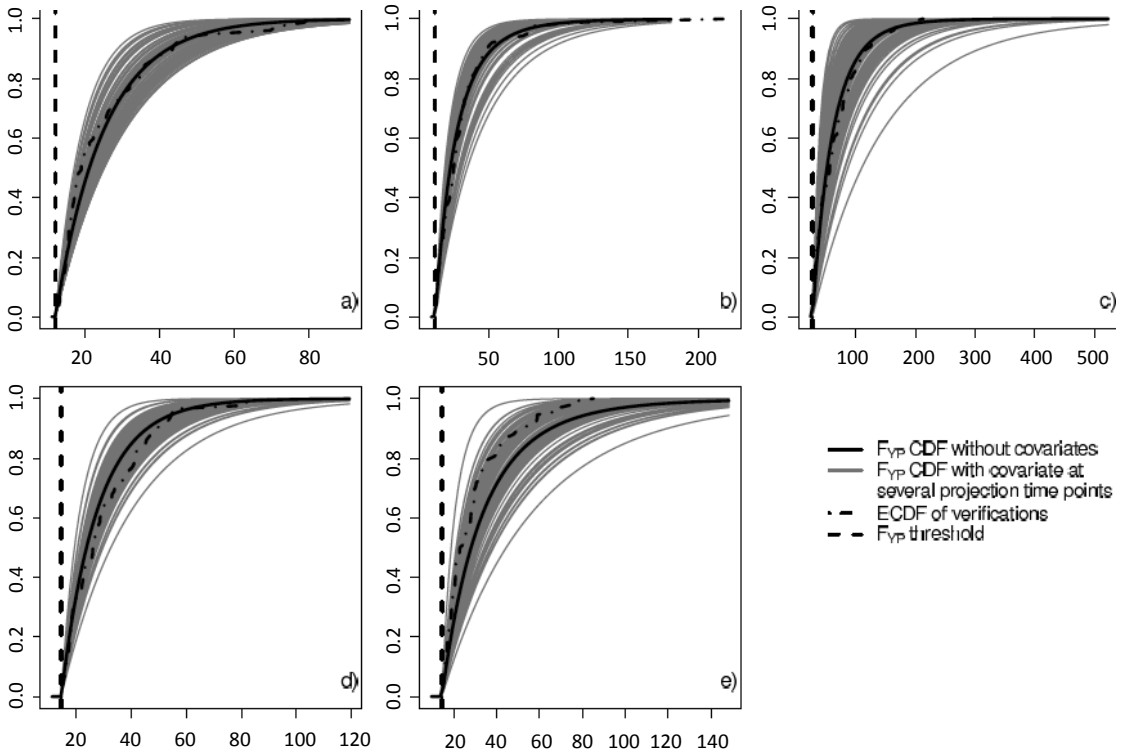


FIGURE 4.1: Pour chacune des cinq stations, CDFs F_{Sf} obtenues par l'approche sans covariable (en noir) et CDFs F^t_{Sf} obtenues par l'approche avec covariables (en gris) pour différents temps t de la période de projection. Pour comparaison, les CDFs empiriques des observations de la période de projection sont aussi présentées (lignes point-trait noire). Les abscisses sont en mm/jour et les ordonnées sont les valeurs des CDFS.

modèles climatiques à un modèle de downscaling calibré sur des réanalyses est alors supprimée, CDF-t étant spécifiquement défini pour capturer le lien statistique en GCM et observations. L'utilisation de CDF-t et de XCDF-t dans un cadre non-homogène doit donc se décliner avec précaution selon l'objectif et les contraintes des études.

Enfin, je souhaite mentionner que les extensions XCDF-t homogènes ou non sont inclues dans la dernière version du package R "CDF-t" et sont disponibles gratuitement sur demande en attendant d'être disponibles sur le site CRAN de R.

4.3 Deux mélanges stochastiques pour des distributions “complètes”

Si la théorie des valeurs extrêmes (EVT) fournit une base théorique pour modéliser les maxima (la GEV) et les excès au dessus d'un seuil donné (la GPD), elle ne s'occupe pas des valeurs en-dessous des maxima ou du seuil et ne peut donc pas être utilisée pour représenter les valeurs plus courantes, faibles ou moyennes. Pour ces valeurs non extrêmes, la modélisation peut être réalisée par des lois plus classiques, telles que celles discutées au chapitre 3 (par ex., loi Gamma pour les précipitations). Cependant, ces lois ne traitent pas, ou mal, les valeurs extrêmes. Une partie de mes travaux a alors consisté à développer des modèles groupant ces deux approches, afin de définir des lois statistiques caractérisant l'ensemble de la distribution, c-à-d., les valeurs faibles et moyennes (courantes) et les valeurs extrêmes (plus rares). Je présente ainsi deux approches pour disposer de telles distributions, toutes deux insérées dans des processus de régionalisation. Les deux approches cherchent à lier une loi classique et une GPD :

- soit au travers d'un mélange à poids variables en section 4.3.1,
- soit selon un mélange par “couture“ en section 4.3.2.

4.3.1 Mélange à poids non-homogènes

Pour cela, dans une première étude [197] menée en collaboration avec P. Naveau (LSCE), afin de modéliser les intensités de précipitation, nous avons adapté un modèle initialement développé par [62] et qui permet de coupler une loi classique avec une loi d'extrêmes par un mélange. Ce mélange associe une densité Gamma Γ de paramètre (γ,λ) pour représenter les valeurs usuelles et une densité de GPD gpd de paramètre (σ,ξ) où le seuil est fixé à 0 :

$$f(x|\beta) = c_\beta \left[(1-\omega(x|m,\tau)) \underbrace{\Gamma(x|\gamma,\lambda)}_{\text{pdf des valeurs “courantes”}} + \omega(x|m,\tau) \underbrace{gpd(x|\sigma,\xi,u=0)}_{\text{pdf des valeurs “extrêmes”}} \right], \quad (4.8)$$

où β est le vecteur des paramètres $(\gamma,\lambda,\sigma,\xi,m,\tau)$ et c_β est une constante de normalisation. Cette nouvelle densité f n'est pas un mélange statistique classique puisque le poids de chaque densité n'est pas constante mais prend une valeur différente pour chaque valeur x d'intensité de précipitation où nous souhaitons estimer la densité. Dans la formulation (4.8), ce poids est une fonction ω de paramètres (m,τ) et dépend donc de la valeur x de précipitation. Le but de ce poids fonctionnel est de faire varier l'importance de chaque pdf (et donc sa pertinence) selon que x est davantage dans la partie habituelle des précipitation ou davantage dans les extrêmes, et ceci avec une transition progressive de l'une à l'autre. Si plusieurs fonctions ω peuvent être employées, nous avons décidé d'utiliser celle initialement suggérée par [62] :

$$\omega(x|m,\tau) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-m}{\tau}\right). \tag{4.9}$$

La Fig. 4.2 illustre cette fonction de poids pour une valeur du paramètre m fixée (arbitrairement) à $m = 2$ et différentes valeurs du paramètre τ allant de $\tau = 10^{-12}$ (en noir) à $\tau = 5$ (en bleu). On voit

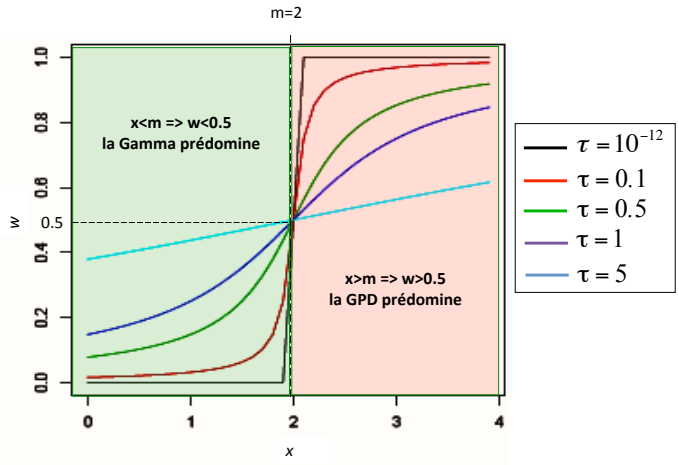


FIGURE 4.2: Illustration de la fonction de poids ω tracée ici pour une valeur du paramètre m fixée (arbitrairement) à 2 et différentes valeurs du paramètre τ allant de $\tau = 10^{-6}$ (en noir) à $\tau = 5$ (en bleu).

que la position de x par rapport à m définit quelle densité est prépondérante dans le mélange 4.8. Si $x = m$, la Gamma et la GPD ont des poids égaux et valant $1/2$. Avant m, la Gamma prédomine, après c'est la GPD. Quant au paramètre τ, il détermine la "vitesse" de transition de la Gamma vers la GPD. Pour un τ faible (par ex., 10^{-12}, en noir), la transition est très abrupte et tend vers une fonction "heavy side" quand $\tau \to 0$, ramenant ainsi la densité (4.8) à un mélange classique à poids constants. Pour des τ croissant progressivement, la transition est de plus en plus douce et lente, jusqu'à tendre vers des poids constants mais égaux pour la Gamma et la GPD quand $\tau \to +\infty$. Dans ce dernier cas, les deux pdf seraient donc aussi importante l'une que l'autre dans le mélange. Ce modèle (4.8) présente ainsi l'avantage de représenter l'ensemble de la distribution (des précipitations dans notre étude), c-à-d., valeurs extrêmes inclues et ceci sans avoir à sélectionner de seuil pour la GPD puisque celui ci est fixé à 0, le processus d'optimisation des paramètres (par maximum de vraisemblance) permettant de jouer sur les autres paramètres pour une représentation correcte des extrêmes.

Différentes comparaisons ont été menées avec d'autres distributions telles que, par exemple, la distribution "stretched exponential" ([216]) très en vogue pour représenter les valeurs extrêmes sans EVT. Toutes ces comparaisons ont montré le bon comportement du mélange (4.8), capable de s'adapter aux distributions à queue lourde et à celles classiques à queue plus légère. Nous avons ensuite inséré ce mélange dans le modèle de downscaling NSWT présenté en section 3.2.1, à la place de la distribution Gamma. Si la Gamma permettait d'obtenir de très bon résultats pour cer-

taines stations (voir Fig. 3.5b), ce n’était pas nécessairement le cas pour d’autres séries temporelles qui pouvaient présenter une sous-estimation des extrêmes, ou même parfois une sur-estimation des extrêmes [4], comme illustré par QQplots sur deux séries temporelles en Fig. 4.3 a et c. L’utilisation

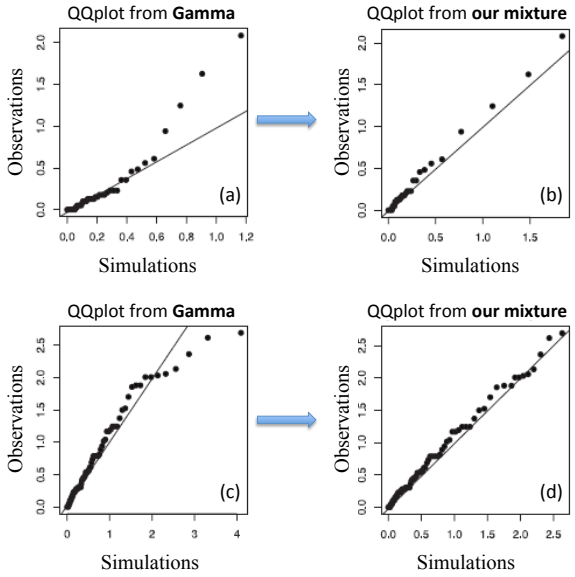


FIGURE 4.3: Pour deux séries temporelles différentes (les lignes), QQplots entre observations et downscaling des valeurs de précipitations avec une loi Gamma (panels de gauche) ou en incluant l’EVT selon le mélange (4.8)

.

du mélange Gamma/GPD à la place de la loi Gamma a alors permis de corriger convenablement la représentation des valeurs extrêmes (Fig. 4.3 b et c), sans pour autant détériorer la qualité de modélisation des événements plus faibles et plus courants, nous fournissant ainsi un modèle pour l’ensemble de la distribution. D’autres analyses complémentaires ont évidemment été menées, portant par exemple sur diverses paramétrisations possibles pour ce modèle, afin d’éviter l’over-fitting. Les résultats et aspects techniques ne sont pas rappelés ici, le but de cette section étant la présentation de ce mélange et l’illustration de ses qualités. Ce travail souligne l’importance des distributions issues de l’EVT pour améliorer la modélisation et le downscaling de précipitations locales extrêmes.

J’ai par ailleurs regroupé les différentes fonctions R permettant l’estimation des paramètres du mélange (4.8) dans un package R nommé “NHMixt” (pour “Non-Homogeneous Mixture”). Ce dernier est évidemment gratuitement disponible sur demande.

4. Je dois reconnaitre que ce phénomène est malgré tout assez rare, je ne l’ai rencontré que deux fois dans les dizaines, voire centaines de jeux de données que j’ai pu traiter.

Une extension en dimension 2

J'aimerais également mentionner brièvement que le modèle (4.8) a été étendu au cas bivarié ([198]), lors d'une collaboration avec P. Naveau (LSCE) et P. Drobinski (LMD). Cette extension est particulièrement nécessaire si l'on s'intéresse aux impacts des précipitations extrêmes. En effet, la prise en compte des dépendances d'extrêmes entre deux stations météorologiques permet une projection (en descente d'échelle, ou en contexte de changement climatique) plus réaliste, par exemple des crues potentiellement associées à ces extrêmes. Pour cela, nous avons repris le même principe que dans le cas univarié : un mélange entre une distribution pour les valeurs courantes et une pour les extrêmes.

Pour les valeurs les plus usuelles, il existe de nombreuses distributions bivariées classiques. Pour différentes raisons techniques liées à l'optimisation des paramètres et à son domaine de corrélations possibles, nous avons utilisé la version bi-dimensionnelle de la loi de Gamma Γ_{2d} dite de la famille "Cheriyan et Ramabhadran" (voir [106] par ex.) :

$$\Gamma_{2d}(x_1,x_2;\gamma) = \int_0^{\min(x_1,x_2)} \frac{e^{-z} z^{\gamma_0-1}}{\Gamma(\gamma_0)} \prod_{i=1}^{2} \left[\frac{e^{-(x_i-z)}(x_i-z)^{\gamma_i-1}}{\Gamma(\gamma_i)} \right] dz, \tag{4.10}$$

où $\gamma = (\gamma_0, \gamma_1, \gamma_2)$ et (x_1, x_2) sont les valeurs de précipitation aux stations 1 et 2.

Pour travailler sur des extrêmes bivariés, c'est un peu plus complexe. Tout d'abord, il faut noter que si l'EVT fournit un cadre paramétrique pour travailler dans le cas univarié, ce n'est pas le cas en dimension $n \geq 2$. Par ailleurs, encore faut-il que les "extrêmes bivariés" soient définis clairement. Il existe en effet au moins deux façons de définir un événement extrême en deux dimensions, au temps t, à partir de données à deux stations distinctes :

- soit les deux stations voient toutes deux des valeurs au-delà d'un seuil donné [5] (autrement dit, elles ont toutes deux des extrêmes univariés),
- soit la somme des deux événements est supérieure à une seuil.

Cette dernière définition est intéressante dans le cas des crues : si les deux stations sont assez proches, la somme des pluies aux deux stations peut engendrer une crue même si les deux valeurs prises séparément ne sont pas extrêmes. C'est bien la somme qui doit dépasser un seuil. Nous avons donc retenu cette deuxième définition et avons travaillé avec les coordonnées (R,Z) de Pickand (voir le chapitre 8 de [35]) :

$$R = x_1 + x_2, \quad \text{et} \quad Z = \frac{x_2}{x_1 + x_2}, \tag{4.11}$$

où R est appelé le "rayon" et Z l'angle. Nous dirons alors qu'un extrême bivarié a lieu si R a dépassé un seuil fixé suffisamment haut. L'angle Z permet de caractériser les dépendances entre extrêmes univariés. En effet, si Z est proche de 0 ou de 1, cela signifie que x_1 est beaucoup plus grand que x_2 (Z proche de 0) ou que x_2 est beaucoup plus grand que x_1 (Z proche de 1), correspondant ainsi à une certaine indépendance des extrêmes aux deux stations. Au contraire, si Z est proche de 0.5,

5. Ce n'est pas nécessairement le même seuil pour les deux stations.

cela signifie que x_1 et x_2 sont proches l'un de l'autre et varient en même temps, correspondant ainsi à une forte dépendance entre les deux stations. Les valeurs de Z étant entre 0 et 1, sa densité peut être modélisée par une loi Beta $h_{a,b}$ de paramètres (a,b). À partir de ces notations, l'extension en dimension 2 du mélange (4.8) est :

$$f(x_1,x_2|\theta) = c_\theta \left[(1-\omega(R|m,\tau)) \underbrace{\Gamma_{2d}(x_1,x_2|\gamma)}_{\text{pdf des valeurs“courantes” bivariées}} + \omega(R|m,\tau) \underbrace{gpd(R|\sigma,\xi,u=0) \times h_{a,b}(Z)}_{\text{pdf des valeurs “extrêmes” bivariées}} \right], \tag{4.12}$$

où θ est le vecteur contenant l'ensemble des paramètres à estimer, c_θ est une constante de normalisation et où ω est la fonction de poids définie en (4.9). Il est intéressant de noter que la densité jointe de (R,Z) est représentée par le produit des densités GPD et Beta associées respectivement au rayon (R) et à l'angle (Z). En effet, du point de vue théorique, dans le cas de valeurs R suffisamment élevées, R et Z sont indépendants (voir [149] ou p. 3 de [198]).

Je ne rentrerai pas dans davantage de détails sur cette extension afin de présenter deux brèves illustrations de son fonctionnement. La première permet de visualiser “l'influence” de la loi Beta des dépendances sur la loi bi-dimensionnelle. J'ai tracé en Fig. 4.4, trois densités Beta à partir de trois couples de paramètres (a,b) correspondant schématiquement à l'indépendance (a), la dépendance (c) et un cas intermédiaire (b). En-dessous de chacun de ces tracés, j'ai représenté les

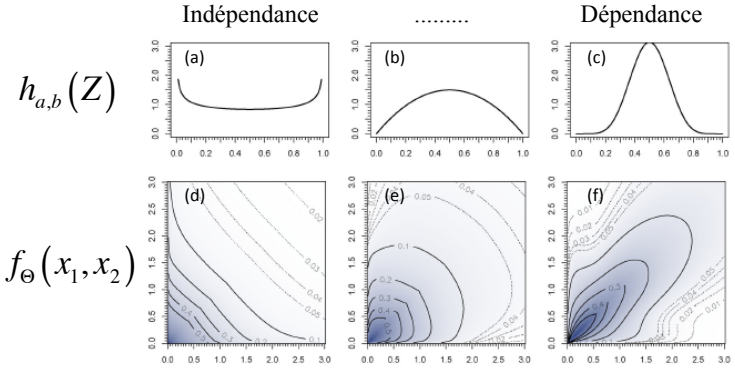


FIGURE 4.4: Panels du haut : trois densités Beta à partir de trois couples de paramètres (a,b) correspondant schématiquement à l'indépendance (a), la dépendance (c) et un cas intermédiaire (b). Panels du bas : densités 2-D associées aux lois Beta (les paramètres, autres que a et b des lois Beta, ont été fixés arbitrairement et sont identiques pour les trois panels).

densités 2-D (4.12) associées, où les autres paramètres a et b des lois Beta ont été fixés arbitrairement et sont identiques pour les trois densités. On voit ici parfaitement que la structure même de la distribution varie selon la dépendance présente dans la loi Beta. Le panel (f) présente une structure diagonale indiquant que si un extrême se produit à l'une des stations, la densité est forte pour qu'un

événement également extrême se produise au même instant à l'autre station. Pour le panel (d), les extrêmes ne sont pas reliés entre stations : une valeur élevée à l'une des stations n'est associée qu'à une (très) faible densité pour qu'une valeur élevée se produise à l'autre station, correspondant bien à une indépendance.

Appliqué à des jeux de données simulées et réelles (résultats non présentés ici), ce modèle de densité bivariée a montré des capacités très encourageantes, permettant de simuler des précipitations à dépendances complexes. Ce modèle étend ainsi non seulement le modèle univarié pour distribution dite "complète" (c-à-d., incluant aussi les extrêmes) au cas à deux dimensions mais permet également la modélisation d'une vaste gamme de dépendances 2-D, depuis l'indépendance jusqu'à la très forte dépendance.

Un package R associé à cette extension bivariée devrait être rendu disponible dans les mois à venir dans le cadre du projet VW "Pleiades".

4.3.2 Mélange de distributions Pareto hybrides

Cette deuxième approche pour régionaliser la distribution "complète" a été réalisée en collaboration avec J. Carreau (HSM). L'étude que nous avons menée ([27]) est basée sur le *modèle de mélange conditionnel* (CMM) présenté en section 3.2.2. Nous avons repris ce CMM mais au lieu de réaliser des mélanges conditionnels de lois Gaussiennes (tronquées) ou de lois log-Normales, la densité que nous avons étudiée est un mélange de *Pareto hybrides* dont la définition a été reprise de [26] et adaptée à un processus de downscaling pour les précipitations journalières. La densité de Pareto hybride est construite à partir d'une Gaussienne "cousue" avec une GPD selon des contraintes de continuité de la densité et de sa dérivée. Cette densité est donnée par :

$$h(y;\mu,\sigma,\xi) = \begin{cases} f(y;\mu,\sigma)/\gamma & \text{si } y \leq u, \\ g(y-u;\xi,\beta)/\gamma & \text{sinon,} \end{cases} \tag{4.13}$$

où $f(y;\mu,\sigma)$ est la densité Gaussienne de paramètres $\mu \in \mathbb{R}$ et $\sigma > 0$, $u \in \mathbb{R}$ est le "point de jonction" (il peut être vu comme le seuil pour la GPD), $\gamma = 1 + \int_{-\infty}^{\alpha} f(y;\mu,\sigma)dy$ est la constante de normalisation et $g(y-u;\xi,\beta)$ la densité de la GPD de paramètre d'échelle $\beta > 0$ et de paramètre de forme (ou indice de queue) $\xi \in \mathbb{R}$. Afin de me concentrer sur la philosophie de cette approche et de ne pas noyer mon message, l'ensemble des détails techniques concernant l'optimisation des paramètres de ce modèle (ainsi que ceux par la suite du mélange d'un tel modèle) n'est pas fourni ici mais peut évidemment être trouvé dans l'article [27] associé à cette étude.

Tout comme la densité (4.8), la densité de Pareto hybride permet de représenter l'ensemble de la distribution des précipitations. Pour faire une analogie avec le modèle de mélange non-homogène (4.8), cette distribution hybride peut être perçue comme un mélange à poids constants où tout le poids est sur la distribution Gaussienne jusqu'à la valeur du point de jonction u, puis tout le poids

est sur la GPD après u. Dans l'étude [27] nous avons défini un modèle de mélange de Pareto hybrides. Afin d'être inséré dans un cadre de régionalisation stochastique, comme expliqué en section 3.2.2 pour la Gaussienne tronquée et la log-Normale, ce mélange a été rendu conditionnel (ou non-homogène) en liant ses paramètres à divers prédicteurs atmosphériques : les paramètres sont alors des fonctions de ces prédicteurs et ces fonctions sont définies par réseaux de neurones artificiels (voir les détails et explications donnés en section 3.2.2). Ainsi, pour chaque jour (caractérisé par des valeurs de prédicteurs), nous disposons des paramètres calculés par le réseau de neurones à partir des prédicteurs et donc d'une densité et de toutes ses propriétés statistiques (moments, niveaux de retours, etc.).

Nous avons alors calibré le CMM d'hybrides à partir des mêmes données de précipitation et de prédicteurs que celles de la section 3.2.2 et avec les mêmes procédés de calibration-projection. Tout comme dans les Figs. 3.6 et 3.7 pour la loi log-Normale, nous pouvons visualiser les cycles saisonniers des différents paramètres conditionnels impliqués dans le mélange d'hybrides. Ceci est illustré pour la station d'Orange en Fig. 4.5, où l'on voit parfaitement l'évolution des paramètres tout au long de l'année.

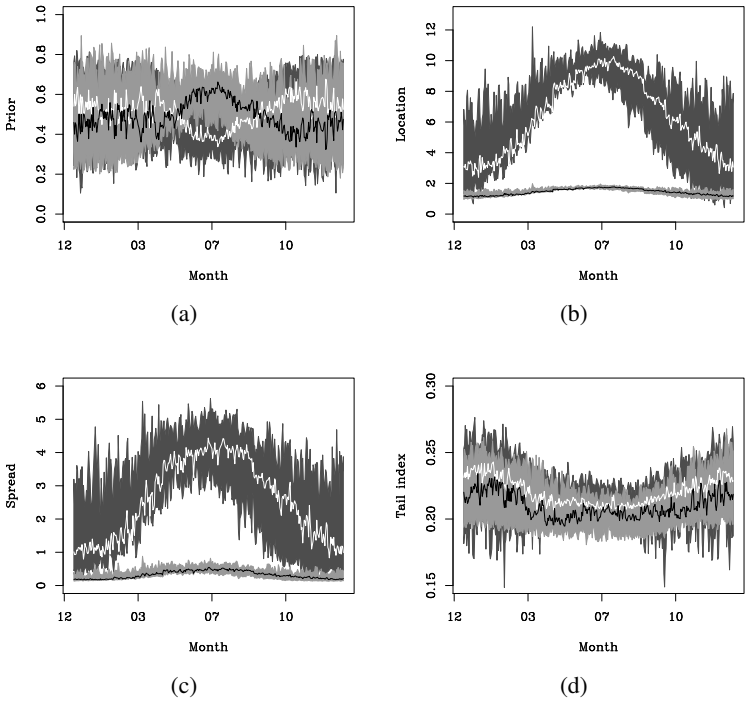
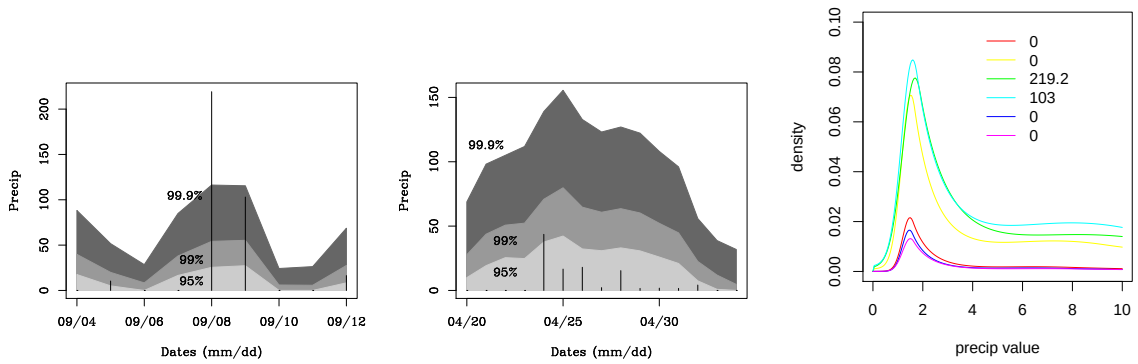


FIGURE 4.5: Pour la station d'Orange, avec le modèle de mélange conditionnel à deux lois Pareto hybrides, illustration des cycles saisonniers des paramètres conditionnels (a) de poids, (b) de points de jonction, (c) d'échelle et (d) de forme. Les zones grisées (plus ou moins foncées) correspondent aux intervalles de confiance à 90% de chaque paramètre pour chaque composante.

Si ces cycles fournissent des informations sur le comportement global des densités conditionnelles, l'un des intérêts des CMMs est qu'il est également possible de s'intéresser à ces densités du point de vue "événementiel", c-à-d., en regardant l'évolution des paramètres ou de quantiles associés à des probabilités fixées, pour des jours consécutifs d'une période donnée. Pour ce faire, la Fig. 4.6 (a et b) compare les observations de pluie et les quantiles conditionnels à 95, 99 et 99.9% du CMM de Pareto hybrides sur Orange pour deux périodes correspondant (a) à la période pluvieuse avec le plus gros cumul de pluie observé (322 mm les 08 et 09 août 2002) et (b) à la période pluvieuse observée la plus longue (9 jours entre le 24 avril et le 02 mai 1993) de la période de projection. Notre objectif était ici de voir si les modèles pouvaient simuler correctement les intensités de pluie lors de ces événements particulièrement sévères et a priori très difficiles à reproduire. On remarque que l'événement majeur du 08 août dépasse l'estimation du



(a) (b) (c)

FIGURE 4.6: Pour la station d'Orange, avec le modèle de mélange conditionnel à deux lois Pareto hybrides, illustration des quantiles conditionnels 95, 99 et 99.9% pour deux périodes correspondant (a) à la période pluvieuse avec le plus gros cumul de pluie observée (322 mm les 08 et 09 août 2002) et (b) à la période pluvieuse observée la plus longue (9 jours entre le 24 avril et le 02 mai 1993). Le panel (c) illustre les différences entre densités conditionnelles journalières pour la période pluvieuse du plus fort cumul (autour du 08 août 2002). Chaque densité journalière est associée à une couleur différente représentée dans la légende dans l'ordre chronologique (du haut vers le bas) avec cumul observé pour chaque jour.

quantile modélisé à 99.9%. Ce résultat n'est pas aberrant. En effet, un quantile de cet ordre doit être dépassé en moyenne une fois sur mille, ce qui est le cas au regard des statistiques globales sur l'ensemble de la période d'évaluation (non présenté dans ce manuscrit, voir [27]). Sur cette Fig. 4.6 (a et b), il est notable que l'évolution temporelle des quantiles conditionnels 95, 99 et 99.9% suit de près l'évolution des observations, montrant ainsi la capacité des CMMs à capturer le lien entre prédicteurs à grande échelle et les précipitations locales et ainsi à représenter de manière pertinente les densités conditionnelles, même pour des événements aussi importants. De manière similaire, on peut regarder en Fig. 4.6c, l'évolution de la densité elle-même (c-à-d., pas uniquement des quantiles 95, 99 et 99.9%), ici pour la période pluvieuse du plus fort cumul (autour du 08

août 2002) . Nous voyons que le modèle de mélange conditionnel est très réactif à un changement de conditions atmosphériques. Ses paramètres se modifient pour ajuster la forme de la distribution de manière réaliste, avec par exemple des queues de distributions beaucoup plus lourdes associées aux événements pluvieux les plus importants.

Les résultats de l'étude [27] ont montré que le mélange conditionnel de lois Pareto hybrides surpasse le modèle de [215] utilisé comme référence ainsi que le CMM de Gaussiennes tronquées qui, malgré un nombre plus important de composantes nécessaires pour caractériser les extrêmes, n'était pas adapté dans un tel contexte de régionalisation des précipitations. Cependant, le CMM de log-Normales a présenté des résultats tout à fait comparables à ceux du CMM d'hybrides. De manière générale, la modélisation conditionnelle permet des projections pertinentes de densités (et donc d'incertitudes des simulations journalières), et la flexibilité additionnelle des mélanges – par opposition aux distributions "uniques" – induit des distributions de précipitation de meilleure qualité, que ce soit dans la partie centrale ou la queue de distribution.

4.4 Deux ou trois mots de bilan sur les événements extrêmes

La théorie statistique des valeurs extrêmes fournit un cadre asymptotique robuste pour la modélisation d'événements rares. Mes travaux dans ce domaine cherchent à promouvoir cette théorie en l'appliquant et l'adaptant pour répondre au mieux à des problématiques climatiques telles que la régionalisation. En contexte de régionalisation du climat, plusieurs études ont essayé d'évaluer la capacité de quelques modèles de downscaling (dynamiques ou statistiques) à simuler correctement des événements extrêmes pour une période historique, essentiellement en température et précipitation. Ces études, dans le cadre de projets tels que PRUDENCE [6] ou STARDEX [7], cherchaient généralement à évaluer la gamme des valeurs possibles d'extrêmes pour un ensemble de modèles (souvent dynamiques) de régionalisation, que ce soit en climat présent (par ex., [175, 49]) ou futur ([45, 46, 9]).

Toutefois, les quelques modèles statistiques testés dans ces projets n'utilisaient généralement pas d'approches stochastiques telles que celles discutées dans ce chapitre et n'utilisaient généralement pas non plus la théorie des valeurs extrêmes.Mes contributions et collaborations dans ce domaine de downscaling stochastique des extrêmes, permettant des modélisations originales et efficaces ainsi que des simulations pertinentes, pourraient alors enrichir ces exercices d'intercomparaisons. Ceux-ci sont en effet capitales pour diriger correctement des modèles d'impacts des événements rares (mais possibles) et potentiellement dévastateurs, ainsi que leurs incertitudes.

6. http ://prudence.dmi.dk/
7. http ://www.cru.uea.ac.uk/projects/stardex/

5 Perspectives

On peut faire des prévisions sur tout, sauf sur l'avenir.
Lao Tseu

De manière générale, je souhaite poursuivre les travaux que j'ai présentés dans les précédents chapitres de ce manuscrit. J'ai regroupé ici certaines perspectives qui me tiennent toutefois plus à coeur que d'autres. Celles-ci peuvent schématiquement se classer en deux catégories s'imbriquant mutuellement :

- Le développement de modèles statistiques de régionalisation et d'extrêmes et la modélisation multidimensionnelle/spatiale,
- L'intercomparaison des modèles de régionalisation et l'évaluation de leurs incertitudes.

Dans chacune de ces catégories, certaines perspectives reposent sur mon expérience et s'appuieront donc sur mes travaux antérieurs. D'autres touchent à des domaines ou approches que je ne connais pas encore nécessairement aussi bien mais dont je suis curieux et pour lesquels je pense que l'intérêt scientifique général est fort. Après tout, si nous faisons de la recherche c'est bien pour continuer d'apprendre...

Dans les sections qui suivent, je ne mentionnerai que ponctuellement les zones géographiques sur lesquelles je travaillerai. En effet, ces travaux pourront s'effectuer (et s'effectueront probablement) sur différentes régions en fonction des besoins et des collaborations. Toutefois, je suis depuis quelques temps fortement intéressé par deux régions très différentes mais où les besoins d'études liées au climat sont notables : la région Méditerranéenne et l'Afrique de l'ouest.

- La région Méditerranéenne constitue un lieu privilégié pour l'étude des couplages océan-atmosphère-hydrologie-écosystèmes. Un bassin océanique quasi-fermé, une orographie marquée sur son pourtour, un climat très contrasté et une forte urbanisation sont des particularités géographiques qui donnent une complexité aiguë au domaine méditerranéen. Les évènements "violents" y sont fréquents : pluies intenses et crues affectant le pourtour méditerranéen occidental l'automne ; sécheresses avec des feux de forêt l'été, parfois accompagnés de vents violents. Si le contexte moyen à grande échelle est relativement bien identifié, la prévision de ces événements extrêmes demeure délicate en raison de la complexité des processus impliqués aux échelles les plus fines et de leurs interactions non-linéaires ([175]).
- Quant à l'Afrique de l'ouest, son développement économique actuel et sa sécurité alimentaire reposant en grande partie sur la culture du millet (principale ressource économique et alimentaire), les évolutions climatiques liées aux ressources en eau (sécheresses, déplacement des moussons, etc.) peuvent considérablement modifier les équilibres environnementaux et

humains futurs dans cette région. Différents travaux ont mis en évidence l'importance des échelles régionales auxquelles ces études doivent être menées (par ex., [5]).

Pour ces deux régions, les incertitudes climatiques, surtout en précipitation, rendent très délicates les études d'impacts pourtant indispensables. La compréhension de ces incertitudes doit alors passer par des simulations appropriées à des échelles locales, simulations pour lesquelles mes perspectives sont pertinentes.

5.1 Développements pour le downscaling et les extrêmes

Les approches statistiques de downscaling sont désormais des alternatives et des compléments reconnus aux approches dynamiques de downscaling (c-à-d. par RCMs). Toutefois, les modélisations statistiques restent bien souvent contraintes au cas univarié (c-à-d à une seule dimension). Par exemple, la régionalisation est faite pour une seule variable (telle que la précipitation journalière) et bien souvent station-par-station. Bien que les propriétés statistiques des simulations aux stations seules (une par une) peuvent alors être correctement modélisées, les structures et dépendances spatiales résultantes sont cependant généralement inappropriées et/ou sous-estimées. Ces "faibles" dépendances, lorsqu'elles sont utilisées dans des modèles d'impacts (hydrologiques, biologiques, économiques, etc.) liés au climat, peuvent alors engendrer des projections spatialement inadéquates et donc associées à une sous-estimation potentielle des impacts. Par exemple, les précipitations journalières possèdent souvent des structures spatiales complexes. Une modélisation inadaptée des dépendances spatiales entre différents lieux géographiques peut considérablement sous-estimer la représentation du ruissellement, des ressources en eau ou des risques d'inondations par exemple. Une démarche de régionalisation permettant de prendre en compte et de modéliser les caractéristiques et dépendances spatiales est par conséquent indispensable pour diriger convenablement de nombreux modèles d'impacts et évaluer correctement les impacts liés aux changements climatiques potentiels.

5.1.1 Modèles inter-sites et inter-variables

Pour ces raisons, je souhaite participer aux développements de modèles stochastiques spatiaux de downscaling pour répondre à ces besoins. Bien que ces modèles restent encore à définir de manière formelle, différentes approches pourront être mises en place et testées. La structure de dépendance inter-sites pourra par exemple être définie *sur les paramètres* du modèle, définissant ainsi un modèle spatial dit "d'indépendance conditionnellement aux paramètres" (on parle aussi de *dépendance latente*). Cette approche est certainement la plus classique en downscaling stochastique multi-sites des précipitations (par ex., [211, 206, 218]) ou en modélisation statistique spatiale (par ex., [190]). Une autre approche pourra définir la structure de dépendance directement sur les simulations locales. Cette approche est plus complexe et informatiquement plus couteuse mais

permettra de générer des projections avec des structures spatiales reconstruites plus cohérentes. La comparaison des propriétés de ces deux types de structures de dépendance n'a encore jamais été effectuée à ma connaissance et fournira un "guide" dans le choix d'utilisation de l'une ou l'autre de ces approches selon les besoins des modèles d'impacts.

Dans la même ligne d'idées, ces apports viendront enrichir les modèles multidimensionnels représentant plusieurs variables à la fois et leurs dépendances. En effet, peu de modèles statistiques de régionalisation permettent de projeter plus d'une variable physique, telle que la précipitation. En générale, mises à part les techniques basées sur des "analogues" (par ex., [222]), les approches statistiques de downscaling ne génèrent des projections locales que variable par variable (par ex., précipitation, puis vent ou température) même si certains générateurs stochastiques de temps (tels que ceux de [150, 151, 135]) commencent par simuler des précipitations et conditionnent les autres variables par les occurrences et intensités des précipitations générées. Ces générateurs ne disposent cependant pas d'une modélisation explicite des dépendances inter-variables qui sont alors généralement sous-évaluées, particulièrement pour les extrêmes.

Le développement d'approches stochastiques pour pallier ces besoins devrait permettre d'améliorer la cohérence inter-variables et spatiale des simulations statistiques régionales.

5.1.2 Les projet "PLEIADES" (VW) et "McSim" (ANR)

Le projet "PLEIADES" ("Projections and predictions of Local prEcipitation Intensities : Advanced Downscaling using Extreme value Statistics") sera l'un des cadres formels pour réaliser ces recherches. Je suis en effet impliqué en tant que co-proposant dans ce projet financé par la fondation Volkswagen pour la période 2011-2014. Ce projet international réunissant le LSCE (moi-même), *the University of Birmingham* (M. Widmann, UK), et *der Universität Kiel* (D. Maraun, Allemagne) a pour objectif de réunir des modélisateurs statisticiens et climatologues afin de développer des méthodes performantes d'estimation des précipitations régionales/locales futures, avec un accent particulier sur les extrêmes. D'un point de vue technique, nous nous baserons pour cela sur le modèle stochastique bi-dimensionnel que j'ai développé ([198]) et présenté en page 92. Dans un premier temps, afin d'affiner la structure spatiale des simulations, cette modélisation bivariée des dépendances sera employée successivement par paires de "lieux" (stations ou point de grille) voisins. Cette première méthodologie sera évaluée en comparaison d'autres approches de régionalisation de l'état de l'art. Par la suite, des efforts seront menés pour étendre cette approche bivariée au cas à $n > 2$ dimensions. Une fois évaluées (et potentiellement validées) sur le $20^{ème}$ siècle, ces méthodes seront appliquées à un ensemble de GCMs et RCMs simulant le climat du $21^{ème}$ siècle, afin de générer des scénarios "griddés" de précipitation pour différentes régions d'Europe (Allemagne, Angleterre, France) et d'étudier les incertitudes associées, en incluant des prédictions décennales jusqu'en 2030.

Un autre cadre tout aussi intéressant pour mener à bien ces développements est le projet "McSim" ("Multisupport conditional simulation of max-stable processes. Applications to the local pre-

diction of extreme climatic events") où je suis responsable des tâches "Predictions and their uncertainties" (WP 3.3) et "Upscaling-downscaling" (WP 4.2). Ce projet financé par l'ANR (2011-2014) et porté par J.-N. Bacro (Univ. Montpellier 2, I3M) s'intéresse essentiellement à la modélisation spatiale et temporelle d'extrêmes dans un contexte non-stationnaire (c-à-d., dont les propriétés statistiques évoluent dans le temps) et de régionalisation. Nous nous intéressons donc ici à essayer de simuler/projeter des événements extrêmes à fine échelle spatiale, conditionnellement à des informations de différentes sources et à différentes échelles. Ces informations peuvent bien sûr provenir de réanalyses et de différentes mesures (stations, satellites, obs griddées, etc.) pour la calibration, et de simulations par GCMs pour les évolutions dans le temps. L'hypothèse d'un processus stochastique spatio-temporel pour ces simulations conditionnelles présente divers avantages :

- L'"upscaling" à partir de mesures physiques, ou le downscaling à partir de GCMs sont consistants avec le modèle statistique, c-à-d., que les distributions aux différentes échelles sont compatibles entre elles (ce qui n'est pas nécessairement le cas de tous les modèles statistiques de régionalisation), quelle que soit l'opération considérée telle que moyenne ou passage au max.
- Les simulations conditionnelles peuvent être combinées pour estimer n'importe quelle statistique d'intérêt (quantiles, probabilités de dépassement de seuils, niveaux de retours, etc.) ainsi que son incertitude (par ex., intervalle de confiance).

Un certain nombre de défis théoriques et pratiques sont donc à relever, incluant la sélection de modèle, l'inférence statistique, la mise au point des algorithmes de simulations conditionnelles. Une attention particulière sera portée sur la famille des modèles max-stables, incluant le "processus tempête" ([172, 160] et le processus "multivariate maxima of moving maxima" (M4, [18, 220]) pour représenter des champs de précipitations convectives. Ces modèles sont essentiellement spatiaux et stationnaires. Le développement d'une version spatio-temporelle non stationnaire est au coeur de ce projet. Les méthodologies développées seront appliquées à la modélisation d'extrêmes de précipitation dans la région Française du Languedoc. Malgré l'apparente "étroitesse" du champ d'application (les extrêmes reliés aux précipitations convectives), les résultats de modélisation de ce projet pourront évidemment être étendus à d'autres variables et champs d'application où les phénomènes extrêmes spatio-temporels interviennent à différentes échelles.

5.1.3 MOS multidimensionnels

Ces modélisations inter-sites et inter-variables seront également poursuivies selon des approches de type "Model Output Statistics" (MOS). En effet, comme mentionné en section 3.4.1, les simulations des modèles climatiques (globaux et régionaux) sont associées à des biais statistiques plus ou moins forts. Les méthodes MOS corrigent ces biais afin d'utiliser ces simulations de manière efficace. Différentes méthodes de correction de biais existent pour cela, telle que le classique "quantile-mapping" (QM, [44]) ou son extension récente "CDF-t" ([125, 98], voir sections 3.4.1 et 4.2) prenant en compte l'évolution des distributions pour la correction. Cependant, la plupart (si ce n'est l'ensemble) des méthodes de correction reste dans un cadre univarié, c-à-d., à une seule variable ou à une seule station. Par exemple, si l'on souhaite corriger les simulations d'inten-

sité de vent, ces projections seront reliées statistiquement à des intensités observées pour réaliser les corrections appropriées. Si l'on souhaite corriger les simulations de précipitation, une procédure similaire sera alors appliquée indépendamment. Or, travailler sur chaque variable séparément, peut générer des corrections qui ne sont pas cohérentes du point de vue physique. Les mêmes problèmes se posent également en contexte spatial où les corrections effectuées indépendamment à deux stations peuvent parfois induire des corrections incohérentes (ou tout au moins irréalistes). Je souhaite ainsi répondre à ce problème en développant de nouvelles méthodologies de correction de biais, tout d'abord dans un contexte bivarié (par ex., vent et précipitation sont corrigés conjointement, ou deux stations sont corrigées ensembles), avant de me pencher sur la problématique plus générale du multivarié. Ceci permettra de corriger non seulement les distributions marginales (c-à-d, 1-d) des variables et des stations mais aussi les dépendances entre variables et entre stations.

Ces travaux pourraient s'effectuer entre autres en collaboration avec H. Rust (*Freie Universität*, Berlin) et M. Kallache (*instituto madrileño de estudios avanzados agua*, Madrid) avec lesquels des discussions ont déjà eu lieu sur ces points.

5.1.4 Vagues de chaleur - Sécheresses

Je souhaite également m'investir dans la modélisation statistique des vagues de chaleur et des périodes de sécheresse. Contrairement aux modèles pour les intensités de précipitation, de vent ou pour les températures qui sont basés sur des densités continues, ceux-ci sont "discrets" (au sens mathématique) : on cherche ici, par exemple, à caractériser l'occurrence ou la non-occurrence de pluie ou de température supérieure à un seuil, pour un nombre n de jours consécutifs. Bien que plusieurs études aient porté sur la modélisation de tels phénomènes – par exemple sous contraintes climatiques avec des modèles non-homogènes ([63]) – celles-ci sont encore trop peu nombreuses alors qu'il s'agit d'une problématique complexe et particulièrement importante pour nombre d'impacts (santé, sécurité alimentaire ou énergétique, etc.). Une partie de mes recherches sera donc consacrée à ces phénomènes. Outre les développements associés à des modèles stochastiques non-homogènes permettant de projeter (et si possible de comprendre) les changements en fréquences, durées (et en intensités pour les vagues de chaleur) sous changement climatique, je souhaite également explorer deux autres pistes.

- Dans le prolongement des modèles statistiques non homogènes, je souhaite tirer avantage des développements récents de la théorie des valeurs extrêmes dans un contexte discret (par ex., [132, 126]). Les distributions de valeurs extrêmes discrètes n'ont jamais été utilisées ni pour la modélisation, ni pour l'évaluation de durées de sécheresses ou de vagues de chaleur. Elles fourniraient des informations précieuses quant à la qualité des projections régionales en climat historique et quantifieraient également les évolutions des probabilités de longues sécheresses et durées de vagues de chaleur sous contraintes de changement climatique.
- Je souhaite par ailleurs poursuivre et mettre à profit mes recherches sur les régimes de temps. En effet, une part de la variabilité des vagues de chaleur peut être associée à des régimes précis. Les évolutions de ces régimes peuvent alors influencer les occurrences de

ces phénomènes. Il est donc nécessaire d'évaluer ces évolutions dans le temps (en fréquence mais aussi en "structure") et surtout d'évaluer la capacité des GCMs à représenter ces régimes (classiques ou saisonniers) et leurs évolutions déjà observées. Pour cela, les mesures statistiques de distance entre régimes (présentées en section 2.3) sont des outils informatifs dont je poursuivrai l'étude. Par ailleurs, la notion de régimes de temps peut également être employée comme outil de base dans un système d'alerte précoce pour les vagues de chaleur ([169]). Ces régimes sont généralement définis plus ou moins empiriquement pour être associés à des situations (atmosphériques, thermiques, etc.) potentiellement dangereuses pour la santé, avec des taux de mortalité élevés. Les apports de l'approche EM fourniraient une quantification de l'incertitude des occurrences de ces régimes (voir page 19) et donc de l'alerte. De plus, la définition même des régimes les plus corrélés à des phénomènes dangereux est un point particulièrement crucial. Les travaux [201] présentés en section 2.5 pourront être appliqués à ce contexte afin de disposer non-seulement de régimes à forte corrélation avec ces phénomènes mais également d'un modèle statistique liant les deux (par analyse des corrélations canoniques, voir section 2.5).

5.2 Intercomparaisons de modèles statistiques et incertitudes

5.2.1 Intercomparaisons

Les intercomparaisons de simulations climatiques sont devenues des exercices clés pour améliorer notre compréhension du système-Terre et sa modélisation. Les plus célèbres sont certainement les projets CMIP3 ("Coupled Model Intercomparison Project N°3", par ex., [123]) et CMIP5 (en cours) qui sont basés sur des simulations issues de GCMs et à partir desquelles les rapports de l'IPCC sont écrits ([90]). Ces dernières années, des projets d'intercomparaisons de simulations dynamiques régionales (c-à-d., par RCMs) sont également apparus. Par exemple, les projets PRUDENCE (FP5, 2001-2004) et ENSEMBLES (FP6, 2004-2009) ont analysé des simulations issues de multiples RCMs dirigés par plusieurs GCMs. Le projet ANR MedUP ("Forecast and projection in climate scenario of Mediterranean intense events : Uncertainties and Propagation on environment", 2008-2011) a comparé divers RCMs. J'ai moi-même participé à ce projet en participant à la comparaison de RCMs avec l'approche statistique CDF-t ([194]). Le projet STARDEX (FP5, 2002-2005) a également participé à ce type d'analyses en se concentrant davantage sur des indices d'extrêmes ([75]). Bien que ce projet incluait également quelques modèles statistiques de régionalisation, ceux-ci étaient essentiellement par fonctions de transfert linéaires. De plus, leur faible nombre ne permettait pas d'avoir une vision claire des possibilités d'évolutions futures du climat telles que modélisées par approches statistiques. Une intercomparaison *complète* reste ainsi encore à effectuer de manière rigoureuse et dans laquelle je souhaite m'investir. Je souhaiterai pour cela tester la globalité des principales approches statistiques telles que discutées en section 3.1 (et pas uniquement les fonctions de transfert). Cet exercice permettrait de comparer

et d'évaluer non seulement les avantages et faiblesses de ces approches entre elles mais aussi les différentes méthodes et variantes dans chaque catégorie.

Par ailleurs, le projet CORDEX ("COordinated Regional Climate Downscaling Experiment") est soutenu par le "World Climate Research Programme" (WCRP) pour réaliser des intercomparaisons de simulations régionales sur 12 régions du globe, essentiellement par des modèles dynamiques ([67]). Dans ce projet, d'importantes études et comparaisons avec des alternatives statistiques restent encore à faire, particulièrement avec des modèles incluant la théorie des valeurs extrêmes. Mes perspectives d'intercomparaisons des approches statistiques fourniront donc les points de comparaison statistiques nécessaires à ce projet international et seront donc structurées sous les mêmes contraintes (régionalisation d'ERA-Interim et des mêmes GCMs) pour des comparaisons directes avec les simulations dynamiques. L'accent sera cependant probablement mis essentiellement sur la région n° 12 (zone Méditerranéenne), coeur de l'action Med-CORDEX coordonnée entre les projets CORDEX et HYMEX ("HYdrological cycle in the Mediterranean EXperiment").

5.2.2 Indicateurs statistiques et Action COST "VALUE"

Le développement d'indicateurs permettant de caractériser la qualité des simulations (dynamiques et statistiques) et de quantifier leurs différences est aussi une perspective importante. Si de nombreux indicateurs ont déjà été proposés (par ex., dans STARDEX, [75], ou les indices Climdex, [140]) et seront bien sûr repris et potentiellement étendus, ils restent souvent limités à des critères très empiriques ou de comptage (par exemple, nombre de jours avec pluie ou température supérieure à un seuil). La mise au point et l'application d'indicateurs reposant sur des outils et modèles statistiques – portant par exemple sur les dépendances et variabilités spatiales et temporelles, les événements extrêmes, les relations inter-variables, etc., et potentiellement basés sur des "scores de probabilité" – renforceront notre capacité à évaluer et interpréter ces simulations. Cette partie de mes travaux futures s'insèrent dans le cadre de l'Action Européenne COST ES1102 VALUE ("Validating and integrating downscaling methods for climate change research", 2011-2015) dont je suis le représentant Français au comité de gestion. En effet, cette Action a pour but de définir des critères d'évaluation et de validation des méthodes de downscaling. VALUE fournira un réseau Européen pour améliorer les collaborations entre les différentes communautés de recherche (climatologues, statisticiens) et les décideurs économiques et politiques. À partir de ces critères et de leurs applications, VALUE délivrera l'ensembles des mesures de validation pour diverses méthodes de downscaling de l'état de l'art et fournira ainsi des pistes d'améliorations des scénarios régionaux de changement climatique sur l'Europe, pertinentes pour la société.

Il est bon aussi de se rappeler que, même si les simulations à haute résolution obtenues par ces approches statistiques sont intéressantes en soit du point de vue climatique, elles ont vocation à être employées pour nourrir des modèles d'impacts divers. Une composante qui me semble importante dans ce contexte est donc d'évaluer ces approches également en terme de "qualités" des impacts

générés par la suite. Cet angle d'évaluation permet ainsi de ne pas se concentrer à tort sur certains défauts ou biais des simulations qui n'affecteraient pas les impacts mais au contraire de pointer des propriétés statistiques qu'il est indispensable de modéliser correctement afin d'améliorer nos modèles statistiques de manière intelligente.

De manière plus institutionnelle, cette perspective d'intercomparaisons et de développement d'indicateurs devrait également participer à la dynamique de recherche en downscaling, en particulier en statistique, au sein de l'IPSL et devrait renforcer les différentes actions entreprises à l'IPSL dans ce contexte, telles que dans le "Labex IPSL" ou le pôle "Climat et Environnement Régionaux".

5.2.3 Modélisation des incertitudes

La connaissance des incertitudes des projections régionales est une composante essentielle à toute étude d'impacts à fine échelle. Pour cela, il est nécessaire de disposer d'un grand nombre de projections afin de couvrir l'ensemble de la distribution des réalisations possibles. Or, les méthodes de régionalisation dynamiques sont encore très couteuses en temps et moyens de calculs par opposition aux approches statistiques. Pour cette raison, bien que la communauté climatique internationale ait récemment lancé un vaste effort de simulations régionales dans le cadre de CORDEX (voir section 5.2.1), les méthodes d'évaluation des incertitudes sont plus facilement applicables sur des ensembles de simulations statistiques que dynamiques. Afin d'estimer les incertitudes associées aux résultats (de régionalisation et d'impacts associés), je souhaite appliquer des approches de type "model-merging". Ce type d'approches quantifie les incertitudes et variabilités globales associées à un ensemble de sorties de modèles et définissent une combinaison de ces modèles pour créer un nouveau jeu de données plus pertinent et réaliste par rapport aux observations. Ces combinaisons peuvent être linéaires ou non, pondérées ou non, fréquentistes ou Bayésiennes et ont été un sujet de recherche important ces dernières années (voir [97] et ses références). Si différentes techniques de model-merging ont déjà été appliquées avec plus ou moins de succès sur des sorties de GCMs et de RCMs ([176, 103, 161, 158, 97, 165]), aucune n'a jamais été appliquée (à ma connaissance) sur des ensembles de simulations statistiques. Une évaluation des gains apportés (ou pas) par une telle combinaison serait une première intéressante pour l'ensemble de la communauté climatique. De plus, des évaluations incluant également des membres dynamiques permettront des analyses complètes avec l'ensemble des approches de régionalisation existantes. Des combinaisons variées seront testées, depuis les plus simples (moyennes pondérées ou non) jusqu'aux plus complexes telles que l'approche Bayésienne par "Bayesian Model Averaging" (BMA, [81, 68]). Dans le cas de moyennes pondérées, les poids relatifs associés aux différents modèles peuvent être vus comme une évaluation de leur réalisme, donnant alors un "classement" aux modèles. En outre, dans le cadre Bayésien, les incertitudes et variabilités sont explicitement modélisées par des densités a posteriori des simulations obtenues, qualifiant ainsi l'incertitude sur les projections. Ceci est encore sous-exploité et mérite d'être développé davantage.

Par ailleurs, en termes de développements de modèles statistiques pour le downscaling, si mes travaux précédents se situent dans un contexte dit “fréquentiste”, mes perspectives sont, au moins pour partie, dans un cadre Bayésien. En effet, les modèles stochastiques discutés en sections 3.2 et 4.3 ont la particularité de régionaliser *les paramètres de densités de probabilité* sous contraintes climatiques. Nous disposons alors pour chaque jour d’une distribution caractérisant les valeurs, par ex., des précipitations possibles. Ce type de modèles fournit ainsi une information liée à l’incertitude (une pdf) et on voit bien la proximité de philosophie avec l’approche Bayésienne. Cette dernière va cependant un cran plus loin en définissant des pdf pour les paramètres eux-mêmes, explorant davantage l’espace de phase des paramètres. Ce contexte Bayésien est souvent considéré comme un atout dans un contexte d’évaluation et de modélisation des incertitudes. Cependant, il n’est pas encore tout à fait clair pour moi que ce type de modèle est le plus à même d’apporter les modélisations (et donc les simulations) les plus pertinentes en contexte de régionalisation. Une véritable comparaison est donc encore à mener sur ce point.

D’un point de vue plus technique, j’aimerais également brièvement soulever une question potentiellement intéressante, alliant downscaling statistique et modélisation des incertitudes. Dans un cadre de régionalisation par modèles stochastiques ou par MOS (tel que par CDF-t), il est possible de générer, non pas des simulations numériques classiques mais des densités de probabilité en sorties de downscaling. Si l’on a plusieurs modèles de la sorte, eux-mêmes potentiellement forcés par plusieurs GCMs ou RCMs, on dispose alors d’un ensemble de densités. On peut se demander comment, dans un processus d’évaluation des incertitudes, traiter un tel ensemble de pdf (ou CDFs) pour modéliser les incertitudes globales de ces projections probabilistes. Cette question relativement théorique pourrait trouver quelques réponses grâce aux apports de “l’analyse de données fonctionnelles” ([147]) et dans laquelle une partie de mes travaux précédents s’est inscrite (section 2.2, [191, 193, 192]). Dans la continuité de cette question, nous pouvons également nous interroger sur l’utilisation la plus adaptée de ces densités, qu’elles soient issues d’un processus de downscaling stochastique ou de modélisation des incertitudes, en entrées de modèles d’impacts. En effet, ces derniers n’utilisent pas directement ces pdfs en entrées mais, au mieux, génèrent un ensemble de valeurs à partir de celles-ci pour potentiellement générer un ensemble de scénarios “d’impacts”. Une évolution conceptuelle intéressante serait de développer de nouveaux types de modèles d’impacts, capables d’avoir en entrées la ou les densités elles-mêmes et non pas uniquement quelques valeurs la caractérisant. De telles modèles, alternatifs mais complémentaires, seraient directement adaptés pour travailler sur des objets mathématiques statistiques représentant les incertitudes et permettraient ainsi de comprendre leurs propagations en termes d’impacts.

5.2.4 Mises à disposition

J’aimerais terminer en rappelant l’importance de fournir à la communauté climatique et environnementale les outils et modèles que nous, chercheurs en climatologie statistique, développons. Depuis plusieurs années, j’essaie de mettre à disposition sous forme de packages R les principaux apports méthodologiques de mes travaux. Il faut l’avouer, cela prend un temps non négligeable

et représente parfois un travail d'ingénierie relativement fastidieux pour des chercheurs avides de passer aux études suivantes, sur d'autres domaines ou simplement d'aller un cran plus loin dans la compréhension et la modélisation des phénomènes climatiques et de leurs variabilités. Cependant, cet exercice ne doit, à mon sens, pas être perçu comme complémentaire aux résultats ou à la publication d'articles mais comme une composante à part entière du travail et de sa communication. La mise à disposition gratuite et automatique à la communauté scientifique de ces packages R permet non seulement une vérification totalement transparente des modèles développés mais aussi une dissémination "culturelle" des méthodes et une utilisation facilitée – et même parfois optimale – des méthodes mises en oeuvre. Cette étape de mise à disposition du savoir-faire philosophique et technique est, pour moi, certainement aussi importante que la publication elle-même.

Je tiens à préciser que c'est volontairement que je n'emploie pas le terme de "services climatiques". Si une partie de mes travaux pourrait s'inclure dans cette thématique, la limite entre recherche et services est délicate et pose des questions politiques, de structuration et de financement de la recherche et de ses liens avec le secteur industriel et plus généralement avec la société. Si les chercheurs (tout au moins dans le public) sont par définition au "service" de la société pour faire progresser la connaissance, on peut se demander où placer le curseur au niveau de l'application de ces connaissances dans un contexte de prise de décisions politiques et économiques. Ces questions ne trouveront pas de réponses dans ce manuscrit dont ce n'est évidemment pas l'objectif mais constituent certainement un point d'interrogation majeur dans ce contexte d'études scientifiques et sociétales sur les changements climatiques.

Un dernier mot ?

Ces perspectives ne sont bien sûr pas exhaustives du tout. J'ai également à l'esprit d'autres perspectives et d'autres collaborations (internes à l'IPSL, nationales ou internationales) que je ne mentionne pas ici car moins dans mon "coeur de cible" mais qui me font découvrir d'autres aspects des statistiques et de l'environnement (théorie, paléo-downscaling, biodiversité, agriculture, etc.) tout aussi passionnants. Par ailleurs, une partie de mes collaborations a vu le jour autour d'un café ou est due au hasard des rencontres, des discussions qui s'y créent et des idées qui en naissent. Ces perspectives là représentent un plaisir, souvent multi-disciplinaire, particulier sans doute car elles sont aussi inattendues. Je souhaite que cela soit encore le cas dans les années à venir.

6 Références et Annexes

RÉFÉRENCES

[1] R.G. Allen, L.S. Pereira, D. Raes, and M. Smith. Crop evapotranspiration—guidelines for computing crop water requirements. In *Irrigation and Drainage Paper 56 (Rome : FAO)*, page 328, 1998.

[2] A. Bardossy, J. Stehlik, and H.J. Caspary. Automated objective classification of daily circulation patterns for precipitation and temperature downscaling based on optimized fuzzy rules. *Clim Res*, 23 :11–22, 2002.

[3] T. Barnett and R. Preisendorfer. Origins and levels of monthly and seasonal forecast skill for united states surface air temperatures determined by canonical correlation analysis. *Mon. Weather Rev.*, 115 :1825–1850, 1987.

[4] A.G. Barnston and R.E. Livezey. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, 115 :1083–1126, 1987.

[5] C. Baron, B. Sultan, M. Balme, B. Sarr, S. Traore, T. Lebel, S. Janicot, and M. Dingkuhn. From gcm grid cell to agricultural plot : scale issues affecting modelling of climate impact. *Phil. Trans. R. Soc. B*, doi :10.1098/rstb.2005.1741, 2005.

[6] E. Bellone, J. P. Hughes, and P. Guttorp. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate Research*, 15 :1–12, 2000.

[7] R.E. Benestad. Empirical-statistical downscaling in climate modeling. *Eos*, 85(42), 2004.

[8] R.E. Benestad. Downscaling precipitation extremes. *Theoretical and Applied Climatology*, 100(1) :1–21, 2010.

[9] M. Beniston, D.B. Stephenson, O.B. Christensen, C.A.T. Ferro, C. Frei, S. Goyette, K. Halsnaes, T. Holt, K. Jylhä, B. Koffi, J. Palutikof, R. Schöll, T. Semmler, and K. Woth. Future extreme events in european climate : an exploration of regional climate model projections. *Climatic Change*, 81(doi : 10.1007/s10584-006-9226-z) :71–95, 2007.

[10] A. Bernacchia, P. Naveau, M. Vrac, and P. Yiou. Detecting spatial patterns with the cumulant function. part ii : An application to el niño. *Nonlinear Processes in Geophysics*, 15 :169–177, 2008.

[11] G. Biau, E. Zorita, H. von Storch, and H. Wackernagel. Estimation of precipitation by kriging in the EOF space of the sea level pressure field. *J. Clim.*, 12 :1070–1085, 1999.

[12] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford, 1995.

[13] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[14] H.-H. Bock and E. Diday (Eds). *Analysis of Symbolic Data*. Springer, 2000.

[15] J.L. Boé, L. Terray, F. Habets, and E. Martin. A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. *J. Geophys. Res.*, 111(D21106), 2006.

[16] G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(1) :211–252, 1964.

[17] P. Braconnot, B. Otto-Bliesner, S. Harrison, S. Joussaume, J.-Y. Peterschmitt, A. Abe-Ouchi, M. Crucifix, E. Driesschaert, Th. Fichefet, C.D. Hewitt, M. Kageyama, A. Kitoh, A. Laîné, M.-F. Loutre, O. Marti, U. Merkel, G. Ramstein, P. Valdes, S. L. Weber, Y. Yu, and Y. Zhao. Results of pmip2 coupled simulations of the mid-holocene and last glacial maximum - part 1 : experiments and large-scale features. *Clim. Past. (http ://www.clim-past.net/3/261/2007/)*, 3 :261–277, 2007.

[18] B.M. Brown and S.I. Resnick. Extreme values of independent stochastic processes. *J. Appl. Probability*, 14 :732–739, 1977.

[19] J. Brown, O. Ferrians, J. Heginbottom, and E. Melnikov. Circum-arctic map of permafrost and ground-ice conditions. In *Circum-Pacific Map Series CP-45*, volume Geological Survey in Cooperation with the Circum-Pacific Council for Energy and Mineral Resources, Washington, DC, U.S., 1997.

[20] M. Bunkers, J. Miller, and A. DeGaetano. Definition of climate regions in the northern plains using an objective cluster modification technique. *J Clim*, 9 :130–146, 1996.

[21] A. Busuioc, F. Giorgi, X. Bi, and M. Ionita. Comparison of regional climate model and statistical downscaling simulations of different winter precipitation change scenarios over romania. *Theor. Appl. Climatol.*, 86 :101–123, 2006.

[22] A. Busuioc, R. Tomozeiu, and C. Cacciamani. Statistical downscaling model based on canonical correlation analysis for winter extreme precipitation events in the emilia-romagna region. *Int. J. Climatol.*, 28(4) :449–464, 2008.

[23] A.J. Cannon. Quantile regression neural networks : Implementation in r and application to precipitation downscaling. *Computers & Geosciences*, (In press), 2011.

[24] A.J. Cannon and P.H. Whitfield. Downscaling recent streamflow conditions in british columbia, canada using ensemble neural network models. *Journal of Hydrology*, 259 :136–151, 2002.

[25] J. Carreau and Y. Bengio. A hybrid Pareto mixture for conditional asymmetric fat-tailed distributions. *IEEE Transactions on Neural Networks*, 20(7) :1087–1101, 2009.

[26] J. Carreau and Y. Bengio. A hybrid pareto model for asymmetric fat-tailed data : the univariate case. *Extremes*, 12(1) :53–76, 2009.

[27] J. Carreau and M. Vrac. Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resources Research*, 47(W10502, doi :10.1029/2010WR010128), 2011.

[28] J. Casola and J. Wallace. Identifying weather regimes in the wintertime 500-hpa geopotential height field for the pacific-north american sector using a limited-contour clustering technique. *J of Appl Meteo and Clim*, 46 :1619–1630, 2007.

[29] C. Cassou. Intraseasonal interaction between the madden-julian oscillation and the north atlantic oscillation. *Nature*, 455 :523–527, 2008.

[30] S. P. Charles, B. C. Bates, I. N. Smith, and J. P. Hughes. Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrol. Processes*, 18 :1373–1394, 2004.

[31] S.P. Charles, B.C. Bates, P.H. Whetton, and J.P. Hughes. Validation of downscaling models for changed climate conditions : case study of southern australia. *Clim. Res.*, 12 :1–14, 1999.

[32] V. Chavez ?Demoulin and A.C. Davison. Generalized additive modelling of sample extremes. *J. R. Stat. Soc., Ser. C.*, 54(1) :207–222, 2005.

[33] H.-K. Cho, K.P. Bowman, and G.R. North. A comparison of Gamma and Lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, 43 :1586–1597, 2004.

[34] V. Choulakian and M.A. Stephens. Goodness ?of ?fit tests for the generalized pareto distribution. *Technometrics*, 43(4) :478–484, 2001.

[35] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, 2001.

[36] L. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2 :299–318, 1967.

[37] R. Davis, R. Dolan, and G. Demme. Synoptic climatology of atlantic coast northeasters. *Int J Climatol*, 13 :171–189, 1993.

[38] A.C. Davison. *Statistical Models*. Cambridge Univ. Press, Cambridge, U. K., 2003.

[39] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc., Ser. B*, 39 :1–38, 1977.

[40] G. De'ath. Boosted trees for ecological modeling and prediction. *Ecology*, 88(1) :243–251, 2007.

[41] Y.B. Dibike and P. Coulibaly. Temporal neural networks for downscaling climate variability and extremes. *Neural Networks*, 19(2) :135–144, 2006.

[42] E. Diday, Y. Ok, and A. Schroeder. The dynamic clusters method in pattern recognition. In J.L. Rosenfeld, editor, *Proceedings of the IFIP Congress 74*, North-Holland, 1974.

[43] R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd ed.)*. Wiley & sons, New York, 2001.

[44] M. Déqué. Frequency of precipitation and temperature extremes over france in an anthropogenic scenario : Model results and statistical correction according to observed values. *Global Planet. Change*, 57 :16– 26, 2007.

[45] M. Déqué, R.G. Jones, M. Wild, F. Giorgi, J.H. Christensen, D.C. Hassell, P.L. Vidale, B. Rockel, D. Jacob, E. Kjellström, M. de. Castro, F. Kucharski, and B. van den Hurk. Global high resolution versus limited area model climate change projections over europe : quantifying confidence level from prudence results. *Climate Dynamics*, 25(doi : 10.1007/s00382-005-0052-1) :653–670, 2005.

[46] M. Déqué, D.P. Rowell, D. Lüthi, F. Giorgi, J.H. Christensen, B. Rockel, D. Jacob, E. Kjellström, M. de Castro, and B. van den Hurk. An intercomparison of regional climate simulations for europe : assessing uncertainties in model projections. *Climatic Change*, 81(doi : 10.1007/s10584-006-9228-x) :53–70, 2007.

[47] J. Elith, J.R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4) :802–813, 2008.

[48] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin, 1997.

[49] C. Goodess et al. Stardex. Final report of the STARDEX project (available at http ://www.cru.uea.ac.uk/projects/stardex/), 2005.

[50] X. Fern, C.E. Brodley, and M.A. Fried. Correlation clustering for learning mixtures of canonical correlation models. In H. Kargupta, editor, *Proceedings of the Fifth SIAM International Conference on Data Mining*, page 439–448, Newport Beach, Calif, 2005.

[51] R. Fisher and L. Tippett. Limiting forms of the frequency distribution in the largest particle size and smallest member of a sample. *Proc. Camb. Phil. Soc.*, 1928.

[52] H.J. Fowler, S. Blenkinsop, and C. Tebaldi. Linking climate change modelling to impacts studies : recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, 27(12) :1547–1578, 2007.

[53] H.J. Fowler, D. Cooley, S.R. Sain, and M. Thurston. Detecting change in uk extreme precipitation using results from the climateprediction.net bbc climate change experiment. *Extremes*, 13(doi :10.1007/s10687-010-0101-y) :241–267, 2010.

[54] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis and density estimation. *J Am Stat Assoc*, 97 :611–631, 2002.

[55] M.J. Frank. On the simultaneous associativity of f(x, y) and x + y ? f(x, y). *Aequationes Mathematicae*, 19 :194–226, 1979.

[56] H. French. *The periglacial environment, 3rd Edition*. Wiley, New York, 2007.

[57] M. D. Frías, E. Zorita, J. Fernández, and C. Rodríguez-Puebla. Testing statistical downscaling methods in simulated climates. *Geophys. Res. Lett.*, 33, 2006.

[58] P. Friederichs and A. Hense. Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135 :2365–2378, 2007.

[59] P. Friederichs and A. Hense. Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135 :2365–2378, 2007.

[60] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression : a statistical view of boosting. *The Annals of Statistics*, 28(2) :337–407, 2000.

[61] J.H. Friedman. Greedy function approximation : a gradient boosting machine. *The Annals of Statistics*, 29(5) :1189–1232, 2001.

[62] A. Frigessi, O. Haug, and H. Rue. A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5 :219– 235, 2003.

[63] E. M. Furrer, R.W. Katz, M.D. Walter, and R. Furrer. Statistical modeling of hot spells and heat waves. *Clim Res*, 43(doi : 10.3354/cr00924) :191–205, 2010.

[64] E.M. Furrer and R.W. Katz. Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, 44(W12439, doi :10.1029/2008WR007316), 2008.

[65] M. Ghil, P. Yiou, S. Hallegatte, B.D. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, O. Mestre, C. Nicolis, H. Rust, P. Sheba-

lin, M. Vrac, A. Witt, and I. Zaliapin. Extreme events : Dynamics, statistics and prediction. *Nonlinear Processes in Geophysics*, 18(doi : 10.5194/npg-18-295-2011) :295–350, 2011.

[66] S. Ghosh and P.P. Mujumdar. Statistical downscaling of gcm simulations to streamflow using relevance vector machine. *Advances in Water Resources*, 31(1) :132–46, 2008.

[67] F. Giorgi, C. Jones, and G.R. Asrar. Addressing climate information needs at the regional level : the cordex framework. *WMO Bulletin*, 58(3) :175–183, 2009.

[68] T. Gneiting, A.E. Raftery, A.H. Westveld III, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5) :1098–1118, 2005.

[69] J.F. González-Rouco, H. Heyen, E. Zorita, and F. Valero. Agreement between observed rainfall trends and climate change simulations in the southwest of europe. *J. Clim*, 13 :3057–3065, 2000.

[70] K. Goubanova, V. Echevin, B. Dewitte, F. Codron, K. Takahashi, P. Terray, and M. Vrac. Statistical downscaling of sea-surface wind over the peru-chile upwelling region : diagnosing the impact of climate change from the ipsl-cm4 model. *Clim. Dyn.*, (In press), 2010.

[71] Z.S. Haddad and D. Rosenfeld. Optimality of empirical z-r relations. *Q. J. R. Meteorol. Soc.*, 123 :1283–1293, 1997.

[72] C. Harpham and R.L. Wilby. Multi-site downscaling of heavy daily precipitation occurrence and amounts. *Journal of Hydrology*, 312(1-4) :235–255, 2005.

[73] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall 9, 41, 1990.

[74] T.J. Hastie, R.J. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

[75] M. R. Haylock, G. C. Cawley, C. Harpham, R. L. Wilby, and C. M. Goodess. Downscaling heavy precipitation over the United Kingdom : a comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology*, 26 :1397–1415, 2006.

[76] E. Hellinger. Neue begründung der theorie quadratischer formen von unendlich vielen veränderlichen. *J. Für Math.*, 136 :210–271, 1909.

[77] P. Hess and H. Brezowski. Katalog der grosswetterlagen europas. Berichte des Deutschen Wetterdienstes. 15 : 1-14, 1977.

[78] B. Hewitson. Regional climates in the giss general circulation model : surface air temperature. *J. Clim.*, 7(2) :283–303, 1994.

[79] B. Hewitson and R. Crane. Self organizing maps : applications to synoptic climatology. *Clim Res*, 26(10) :1315–1337, 2002.

[80] B. C. Hewitson and R. G. Crane. Climate downscaling : techniques and application. *Climate Research*, 7 :85–95, 1996.

[81] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian model averaging : a tutorial. *Statistical Sciences*, 14 :382–417, 1999.

[82] T. Holloway, S.N. Spak, D. Barker, M. Bretl, C. Moberg, K. Hayhoe, J. Van Dorn, and D. Wuebbles. Change in ozone air pollution over chicago associated with global climate change. *J. Geophys. Res.*, 113(D22306), 2008.

[83] K. M. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251–257, 1991.

[84] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(doi :10.1093/biomet/28.3-4.321) :321–377, 1936.

[85] R. Huth. An intercomparison of computer ?assissted circulation classification methods. *Int. J. Climatol.*, 16 :893–922, 1996.

[86] R. Huth. Statistical downscaling in central europe : Evaluation of methods and potential predictors. *Clim. Res.*, 13 :91–101, 1999.

[87] R. Huth. Disaggregating climatic trends by classification of circulation patterns. *Int. J. Climatol.*, 21 :135–153, 2001.

[88] R. Huth. Statistical downscaling of daily temperature in central europe. *J. Clim.*, 15 :1731 – 1742, 2002.

[89] R. Huth, S. Kliegrová, and L. Metelka. Non-linearity in statistical downscaling : does it bring an improvement for daily temperature in europe ? *Int. J. Climatol.*, 28(4) :465–477, 2008.

[90] IPCC. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change : The physical science basis. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

[91] IPCC. Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change : Impacts, adaptation and vulnerability. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

[92] IPCC. Contribution of working group III to the fourth assessment report of the intergovernmental panel on climate change : Mitigation of climate change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

[93] J. Jacobeit, H. Wanner, J. Luterbacher, C. Beck, A. Philipp, and K. Sturm. Atmospheric circulation variability in the north-atlantic-european area since the mid-seventeenth century. *Clim. Dyn.*, 20(4) :341–352, 2002.

[94] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[95] V. Jomelli, D. Brunstein, M. Déqué, M. Vrac, and D. Grancher. Impacts of future climatic change (2000-2100) on the occurrence of debris flows : A case study in the massif des ecrins (french alps). *Climatic Change*, 97(1–2) :171–191, 2009.

[96] A. Jost, D. Lunt, M. Kageyama, A. Abe–Ouchi, O. Peyron, P.J. Valdes G., and Ramstein. High-resolution simulations of the last glacial maximum climate over europe : a solution to discrepancies with continental palaeoclimatic reconstructions ? *Clim. Dynam.*, 24 :577–590, 2005.

[97] M. Kallache, E. Maksimovich, P.-A. Michelangeli, and P. Naveau. Multi-model combination by a bayesian hierarchical model : Assessment of ice accumulation over the oceanic arctic region. *Journal of Climate*, 23(20) :5421–5436, 2010.

[98] M. Kallache, M. Vrac, P. Naveau, and P.-A. Michelangeli. Non-stationary probabilistic downscaling of ex- treme precipitation. *Journal of Geophysical Research - Atmosphere*, 116(D05113, doi :10.1029/2010JD014892), 2011.

[99] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W.Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, Roy Jenne, and Dennis Joseph. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77 (3) :370–471, 1996.

[100] R. Katz, M. Parlange, and P. Naveau. Extremes in hydrology. *Advances in Water Resources*, 25 :1287–1304, 2002.

[101] R.W. Katz. Extreme value theory for precipitation : Sensitivity analysis for climate change. *Adv. Wat. Resour.*, 23(2) :133–139, 1999.

[102] R.W. Katz. Techniques for estimating uncertainty in climate change scenarios and impact studies. *Clim. Res.*, 20 :167–185, 2002.

[103] R. Knutti, R. Furrer, C. Tebaldi, J. Cermak, and G.A. Meehl. Challenges in combining projections from multiple climate models. *Journal of Climate*, 23(10) :2739–2758, 2010.

[104] T. Kohonen. Construction of similarity diagrams for phonemes by a self-organizing algorithm. report TKK-F-A463, Helsinki University of Technology, Espoo, 1981.

[105] T. kohonen. *Self-organizing maps*. Springer, Heidelberg, 2001.

[106] S. Kotz, N. Balakrishnan, and N.L. Johnson. *Continuous multivariate distributions, vol. 1, Models and applications*. Wiley, New York, 2000.

[107] S. Kullback. The kullback–leibler distance. *Amer. Stat.*, 41 :340–341, 1987.

[108] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22 :79–86, 1951.

[109] H. Lamb. British isles weather types and a register of daily sequence of circulation patterns. Geophysical Memoir 116, 1972.

[110] R. Laprise. Regional climate modelling. *Journal of Computational Physics*, 227 :3641–3666, 2008.

[111] C. Lavaysse, M. Vrac, P. Drobinski, M. Lengaigne, and T. Vischel. Statistical downscaling of the french mediterranean climate : assessment for present and projection in an anthropogenic scenario. *Natural Hazard and Earth System Science*, In revision, 2011.

[112] J. Leloup, M. Lengaigne, and J.P. Boulanger. Twentieth century enso characteristics in the ipcc database. *Clim Dyn*, 30 :277–291, 2008.

[113] P. LeMoigne. Description de l'analyse des champs de surface sur la france par le système safran (description of the analysis of near-surface atmospheric fields over france with safran system) (in french). CNRM/GAME report 77, CNRM/GAME, Météo-France/CNRS, Toulouse, France, 2002.

[114] G. Levavasseur, M. Vrac, D. Roche, D. Paillard, and A. Martin. Present and lgm permafrost from climate simulations : contribution of statistical downscaling. *Climate of the Past*, In press, 2011.

[115] P.C. Mahalanobis. On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India*, 2 :49–55, 1936.

[116] D. Maraun, T.J. Osborn, and H.W. Rust. The influence of synoptic airflow on uk daily precipitation extremes. part i : Observed spatiotemporal relationships. *Clim. Dyn.*, 36(1-2) :261–275, 2011.

[117] D. Maraun, H.W. Rust, and T.J. Osborn. Synoptic airflow and uk daily precipitation extremes : Development and validation of a vector generalised linear model. *Extremes*, 13(2) :133–153, 2010.

[118] D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themeßl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich. Precipitation downscaling uder climate change. recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3), RG3003, 2010.

[119] A. Martin, D. Paillard, M. Vrac, C. Dumas, M. Kageyama, and J. Brulhet. Potential climate change over the next million years : Regional projections over europe using an emic and a downscaling method. *(en finalisation d'écriture)*, 2011.

[120] A. Martin, M. Vrac, D. Paillard, C. Dumas, and M. Kageyama. Statistical-dynamical downscaling of earth models of intermediate complexity. *(en finalisation d'écriture)*, 2011.

[121] P. McCullagh. Generalized linear-models. *European Journal of Operational Research*, 16(3) :285–292, 1984.

[122] G. McLachlan and D. Peel. *Finite Mixture Model*. Wiley series in probability and statistics, New York, 2000.

[123] G.A. Meehl, C. Covey, T. Delworth, M. Latif, B. McAvaney, J.F.B. Mitchell, R.J. Stouffer, and K.E. Taylor. The wcrp cmip3 multimodel dataset : A new era in climate change research. *Bulletin of the American Meteorological Society*, pages 1383–1394, 2007.

[124] P.-A. Michelangeli, R. Vautard, and B. Legras. Weather regimes : Recurrence and quasi-stationarity. *J. Atmos. Sci.*, 52 :1237–1256, 1995.

[125] P.-A. Michelangeli, M. Vrac, and H. Loukos. Probabilistic downscaling approaches : Application to wind cumulative distribution function. *Geophysical Research Letters*, 36, 2009.

[126] S. Nadarajah and K. Mitov. Asymptotics of maxima of discrete random variables. *Extremes*, 5 :287 ?294, 2002.

[127] J. Najac, C. Lac, and L. Terray. Impact of climate change on surface winds in france using a statistical-dynamical downscaling method with mesoscale modeling. *Int. J. Climatol.*, doi :10.1002/joc.2075, 2010.

[128] P. Naveau, M. Nogaj, C. Ammann, P. Yiou, D. Cooley, and V. Jomelli. Statistical methods for the analysis of climate extremes. *C. R. Geosci.*, 337 :1013–1022, 2005.

[129] R.B. Nelsen. *An introduction to copulas*. Springer, New York, 1999.

[130] M. New, D. Lister, M. Hulme, and I. Makin. A high-resolution data set of surface climate over global land areas. *Clim. Res.*, 21 :1–25, 2002.

[131] P. Oettli, B. Sultan, C. Baron, and M. Vrac. Are regional climate models relevant for crop yield prediction in west africa ? *Environ. Res. Lett.*, 6, 2011.

[132] P. Olofsson. A poisson approximation with applications to the number of maxima in a discrete sample. *Statistics & Probability Letters*, 44 :23 ?27, 1999.

[133] J.C. Olsson, C. Uvo, and K. Jinno. Statistical atmospheric downscaling of short-term extreme rainfall by neural networks. *Physics and Chemistry of the Earth, Part B : Hydrology, Oceans and Atmosphere*, 26(9) :695–700, 2001.

[134] A. Overeem, A. Buishand, and I. Holleman. Rainfall depth-duration ?frequency curves and their uncertainties. *J. Hydrol.*, 348(doi :10.1016/j.jhydrol.2007.09.044) :124–134, 2008.

[135] Racsko P., Szeidl L., and M.A. Semenov. A serial approach to local stochastic weather models. *Ecological Modelling*, 57 :27–41, 1991.

[136] C. Pagé, L. Terray, and J. Boé. dsclim : A software package to downscale climate scenarios at regional scale using a weather-typing based statistical methodology. Technical Report TR/CMGC/09/21 1875, URA CERFACS/CNRS, Toulouse, France, 2009.

[137] H.A. Panofsky and G.W. Brier. *Some Applications of Statistics to Meteorology*. University Park. Penn. State Univ. Press, 1958.

[138] E. Parzen. On estimation of a probability density function and mode. *Ann Math Stat*, 33 :1065–1076, 1962.

[139] K. Pearson. Contributions to the theory of mathematical evolution. *Philos. Trans. R. Soc., Ser. A*, 185 :71–110, 1894.

[140] T.C. Peterson. Climate change indices. *WMO bulletin*, 54(2) :83–86, 2005.

[141] V. Petoukhov, M. Claussen, A. Berger, M. Cricifix, M. Eby, A. Eliseev, T. Fichefet, A. Ganopolski, H. Goose, I. Kamenkovich, M. Montoya, L.A. Mysak, A. Sokolov, P. Stone, Z. Wang, and A. Weaver. Emic intercomparison project (emip–co2) : comparative analysis of emic simulations of climate, and of equilibrium and transient responses to atmospheric co2 doubling. *Clim. Dynam.*, 25 :363–385, 2005.

[142] V. Petoukhov, A. Ganopolski, V. Brovkin, M. Claussen, A. Eliseev, C. Kubatzki, and S. Rahmstorf. Climber-2 : a climate system model of intermediate complexity. *Clim. Dynam.*, 16 :1–17, 2000.

[143] T. Petrow, J. Zimmer, and B. Merz. Changes in the flood hazard in germany through changing frequency and persistence of circulation patterns. *Natural Hazards and Earth System Sciences*, 9 (4) :1409–1423, 2009.

[144] C. Piani, J.O. Haerter, and E. Coppola E. Statistical bias correction for daily precipitation in regional climate models over europe. *Theor. Appl. Climatol.*, 99 :187–192, 2010.

[145] R. Pongracz, J. Bartholy, and I. Bogardi. Fuzzy rule-based prediction of monthly precipitation. *Phys Chem Earth*, 9 :663–667, 2001.

[146] P. Quintana-Seguí, P. Le Moigne, Y. Durand, E. Martin, F. Habets, M. Baillon, C. Canellas, L. Franchistéguy, and S. Morel. Analysis of near surface atmospheric variables : validation of the safran analysis over france. *J. Appl.Meteorol. and Climatol.*, 47 :92–107, 2008.

[147] J. Ramsay and B. Silverman. *Applied Functional Data Analysis, Methods and Case Studies*. Springer, 2002.

[148] H. Renssen and J. Vandenberghe. Investigation of the relationship between permafrost distribution in nw europe and extensive winter sea-ice cover in the north atlantic ocean du-

ring the cold phases of the last glaciation. *Quaternary Sci. Rev.*, 22(doi :10.1016/S0277-3791(02)00190-7) :209–223, 2003.

[149] S. Resnick. *Heavy-Tail Phenomena : Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering, 2007.

[150] C.W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resour Res*, 17 :182–190, 1981.

[151] C.W. Richardson and D.A. Wright. WGEN, a model for generating daily weather variables. ARS Bulletin ARS-8, USDA, Washington, DC, U.S.A. : Government Printing Office, 1984.

[152] A.W. Robertson, S. Kirshner, and P. Smyth. Downscaling of daily rainfall occurrence over northeast brazil using a hidden markov model. *J. Clim.*, 17 :4407–4424, 2004.

[153] P.J. Rousseeuw. Silhouettes – a graphical aid to the interpretation and validation of cluster-analysis. *Journal of Computational and Applied Mathematics*, 20 :53–65, 1987.

[154] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. *Parallel Distributed Processing : Explorations in the Macrostructure of Cognition, Bradford Books*, 1, 1986.

[155] M. Rummukainen. State-of-the-art with regional climate models. *WIRE (Wiley Interdisciplinary Reviews : Climate Change) Adv. Rev.*, 1(1) :82–96, 2010.

[156] H. Rust, M. Vrac, M. Lengaigne, and B. Sultan. Quantifying differences in circulation patterns based on pro- babilistic models : Ipcc-ar4 multi-model comparison for the north atlantic. *Journal of Climate*, 23 :6573–6589, 2010.

[157] D. J. Sailor and Xiangshang L. A semi-empirical downscaling approach for predicting regional temperature impacts associated with climatic change. *J. Clim.*, 12 :103–114, 1999.

[158] S.R. Sain and R. Furrer. Combining climate model output via model correlations. *Stochastic Environmental Research and Risk Assessment*, 24(6) :821–829, 2010.

[159] T. Salameh, P. Drobinski, M. Vrac, and P. Naveau. Statistical downscaling of near surface wind field over complex terrain in southern France. *Meteorology and Atmospheric Physics*, 103 :253–265, 2009.

[160] M. Schlather. Models for stationary max-stable random fields. *Extremes*, 5 :33–44, 2002.

[161] E. Schliep, D. Cooley, S.R. Sain, and J.A. Hoeting. A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. *Extremes*, 13 :219–239, 2010.

[162] R. Schnur and D. Lettenmaier. A case study of statistical downscaling in Australia using weather classification by recursive partitioning. *J. Hydrol.*, 212–213 :362–379, 1998.

[163] S. Schubert and A. Henderson-Sellers. A statistical model to downscale local daily temperature extremes from synoptic-scale atmospheric circulation patterns in the australian region. *Climate Dynamics*, 13(3) :223–234, 1997.

[164] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6 :461–464, 1978.

[165] C. Schölzel and A. Hense. Probabilistic assessment of regional climate change in southwest germany by ensemble dressing. *Climate Dynamics*, 36(9) :2003–2014, 2011.

[166] M. A. Semenov and E. M. Barrow. Use of a stochastic weathergenerator in the development of climate change scenarios. *Clim. Res.*, 35 :397–414, 1997.

[167] M. A. Semenov, R. J. Brooks, E. M. Barrow, and C. W. Richardson. Comparison of the WGEN and the LARS-WG stochastic weather generators in diverse climates. *Clim. Res.*, 10 :95–107, 1998.

[168] S. Sheridan and L. Kalkstein. Heat watch-warning systems in urban areas. *World Resource Review*, 10(3) :375–383, 1998.

[169] S.C. Sheridan and L.S. Kalkstein. Progress in heat watch ?warning system technology. *Bulletin of the American Meteorological Society*, 85 :1931 ?41, 2004.

[170] B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

[171] A. Sklar. Fonction de répartition a n dimensions et leurs marges. Compte-rendu, Institut de Statistiques, Université de Paris 8 : 229–231, 1959.

[172] R.L. Smith. Max-stable processes and spatial extremes. Unpublished, 1990.

[173] S.E. Snell, S. Gopal, and R.K.Kaufmann. Spatial interpolation of surface air temperatures using artificial neural networks : Evaluating their use for downscaling gcms. *J. Clim.*, 13 :886–895, 2000.

[174] D.B. Stephenson, A. Hannachi, and A. O'Neill. On the existence of multiple climate regimes. *Quart. J. Roy. Meteor. Soc.*, 130 :583–605, 2004.

[175] C. Tebaldi, K. Hayhoe, J.M. Arblaster, and G.A. Meehl. Going to the extremes : An intercomparison of model-simulated historical and future changes in extreme events. *Clim. Change*, 79 :185– 211, 2006.

[176] C. Tebaldi, L. O. Mearns, D. Nychka, and R. L. Smith. Regional probabilities of precipitation change : A bayesian analysis of multimodel simulations. *Geophysical Research Letters*, 31(L24231, doi : 10.1029/2004GL021276), 2004.

[177] C. Tisseuil, F. Leprieur, G. Grenouillet, M. Vrac, and S. Lek. Projected impacts of climate change on spatio-temporal patterns of freshwater fish beta diversity : a deconstructing approach. *Submitted*, 2011.

[178] C. Tisseuil, M. Vrac, G. Grenouillet, M. Gevrey, T. Oberdorff, A.J. Wade, and S. Lek. Climate change impacts on freshwater biodiversity : strengthening the link between hydroclimatic downscaling and species distribution modelling. *Submitted*, 2011.

[179] C. Tisseuil, M. Vrac, S. Lek, and A.J. Wade. Direct statistical downscaling of river flows. *Journal of Hydrology*, 385 :279–291, 2010.

[180] A. Toreti, E. Xoplaki, D. Maraun, F.G. Kuglitsch, H. Wanner, and J. Luterbacher. Characterisation of extreme winter precipitation in mediterranean coastal sites and associated anomalous atmospheric circulation patterns. *Nat. Hazards Earth Syst. Sci.*, 10(doi :10.5194/nhess-10-1037-2010) :1037–1050, 2010.

[181] G. Toulemonde, A. Guillou, P. Naveau, M. Vrac, and F. Chevallier. Autoregressive models for maxima and their applications to ch4 and n2o. *Environmetrics*, 21 :188–207, 2009.

[182] M. Troin, M. Vrac, M. Khodri, C. Vallet-Coulomb, E. Piovano, and F. Sylvestre. Coupling statistically downscaled gcm outputs with a basin-lake hydrological model in subtropical south america : evaluation of the influence of large-scale precipitation changes on regional hydroclimate variability. *Hydrol. Earth Syst. Sci. Discuss.*, 7, 2011.

[183] E.J.M. van den Besselaar, M.R. Haylock, G. van der Schrier, and A.M.G. Klein Tank. A european daily high-resolution observational gridded data set of sea level pressure. *Journal of Geophysical Research - Atmosphere*, 116(D11110), 2011.

[184] J. Vandenberghe, J. Lowe, G. Coope, T. Litt, and L. Züller. Climatic and environmental variability in the mid-latitude europe sector during the last interglacial-glacial cycle. In R. Battarbee, F. Gasse, and Springer Netherlands C. Stickley, editors, *Past Climate Variability through Europe and Africa*, volume 6, page 393–416, 2004.

[185] J. Vandenberghe, H. Renssen, D. Roche, H. Goosse, A. Velichko, A. Gorbunov, and G. Levavasseur. Eurasian permafrost instability constrained by reduced sea-ice cover. *Quaternary Sci. Rev.*, 2011.

[186] S. Vannitsem. Toward a phase-space cartography of the short- and medium-range predictability of weather regimes. *Dynamic meteorology and oceanography*, 53 :56–73, 2001.

[187] R. Vautard, J. Cattiaux, P. Yiou, J.N. Thépaut, and P. Ciais. Northern hemisphere atmospheric stilling partly attributed to an increase in surface roughness. *Nature Geoscience*, 3(doi :10.1038/ngeo979) :756–761, 2010.

[188] R. Vautard, T. Noël, L. Li, and M. Vrac et al. Climate variability and trends in long-term high-resolution downscaled simulations and projections over metropolitan france. *In preparation*, 2011.

[189] N. Vigaud, M. Vrac, and Y. Caballero. Probabilistic downscaling of gcms scenarios over southern india. *Submitted*, 2011.

[190] T. Vischel, T. Lebel, S. Massuel, and B. Cappelaere. Conditional simulation schemes of rain fields and their application to rainfall–runoff modeling studies in the sahel. *Journal of Hydrology*, 375 :273–286, 2009.

[191] M. Vrac. *Analyse et Modélisation de Données Probabilistes par Décomposition de Mélange de Copules et Application a une Base de Données Climatologiques*. Thèse de doctorat en sciences, spécialité mathématiques appliquées, Université Paris IX Dauphine & Ecole Polytechnique, Décembre 2002.

[192] M. Vrac, L. Billard, E. Diday, and A. Chédin. Copula analysis of mixture models. *Computational Statistics*, 2011.

[193] M. Vrac, A. Chédin, and E. Diday. Clustering a global field of atmospheric profiles by mixture decomposition of copulas. *Journal of Atmospheric and Oceanographic Technology*, 22(10) :1445–1459, 2005.

[194] M. Vrac, P. Drobinski, A. Merlo, M. Herrmann, C. Lavaysse, L. Li, and S. Somot. Dynamical and statistical downsca- ling of the french mediterranean climate : uncertainty assessment. *Natural Hazard and Earth System Science*, In revision, 2011.

[195] M. Vrac, K. Hayhoe, and M. Stein. Identification and inter-model comparison of seasonal circulation patterns over North America. *Int. J. Climatol.*, 27(5) :603–620, 2007.

[196] M. Vrac, P. Marbaix, D. Paillard, and P. Naveau. Non-linear statistical downscaling of present and LGM precipitation and temperatures over Europe. *Climate of the Past*, 3 :669–682, 2007.

[197] M. Vrac and P. Naveau. Stochastic downscaling of precipitation : From dry events to heavy rainfalls. *Water resources research*, 43, 2007.

[198] M. Vrac, P. Naveau, and P. Drobinski. Modeling pairwise dependencies in precipitation intensities. *Nonlinear Processes in Geophysics*, 14 :789–797, 2007.

[199] M. Vrac, M. Stein, and K. Hayhoe. Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Clim. Res.*, 34 :169–184, 2007.

[200] M. Vrac, M. Stein, K. Hayhoe, and X.-Z. Liang. A general method for validating statistical downscaling methods under future climate change. *Geophysical Research Letters*, 34, 2007.

[201] M. Vrac and P. Yiou. Weather regimes designed for local precipitation modelling : Application to the mediterranean basin. *Journal of Geophysical Research – Atmosphere*, 115, D12103(doi :10.1029/2009JD012871), 2010.

[202] M. Vrac, P. Yiou, and P. Vaittinada Ayar. Trends and variability of seasonal weather regimes. *Submitted*, 2011.

[203] H.L. Wang, S. Schubert, M. Suarez, J.Y. Chen, M. Hoerling, A. Kumar, and P. Pegion. Attribution of the seasonality and regionality in climate trends over the united states during 1950-2000. *J. Clim.*, 22 :2571–2590, 2009.

[204] X.L. Wang, F.W. Zwiers, and V.R. Swail. North atlantic ocean wave climate change scenarios for the twenty ?first century. *J. Clim.*, 17 :2368–2383, 2004.

[205] J. H. Ward. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58(301) :236–244, 1963.

[206] H. S. Wheater, R. E. Chandler, C. J. Onof, V. S. Isham, E. Bellone, C. Yang, D. Lekkas, G. Lourmas, and M.-L. Segond. Spatial-temporal rainfall modelling for flood risk estimation. *Stoch Environ Res Risk Assess*, 19(DOI 10.1007/s00477-005-0011-8) :403–416, 2005.

[207] T. M. L. Wigley, P. D. Jones, K. R. Briffa, , and G. Smith. Obtaining sub-grid scale information from coarse resolution general circulation model output. *Journal of Geophysical Research*, 95 :1943–1953, 1990.

[208] R.L. Wilby, C.W. Dawson, and E.M Barrow. SDSM - a decision support tool for the assessment of regional climate change impacts. *Environmental Modelling and Software*, 17(2) :145–157, 2002.

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

[210] R.L. Wilby and T.M.L. Wigley. Downscaling general circulation model output : A review of methods and limitations. *Prog. Phys. Geogr.*, 21 :530–548, 1997.

[211] D. S. Wilks. Multisite downscaling of daily precipitation with a stochastic weather generator. *Clim. Res.*, 11 :125–136, 1999.

[212] D.S. Wilks. Us of stochastic weather generators for precipitation downscaling. *Wiley Interdisciplinary Reviews : Climate Change*, 1 (6) :898–907, 2010.

[213] D.S. Wilks and R.L. Wilby. The weather generation game : a review of stochastic weather models. *Prog. Phys. Geogr.*, 23 (3) :329–357, 1999.

[214] P. Willems and M. Vrac. Statistical precipitation downscaling for small-scale hydrological impact investigations of climate change. *Journal of Hydrology*, 402(doi :10.1016/j.jhydrol.2011.02.030) :193–205, 2011.

[215] M. P. Williams. Modelling seasonality and trends in daily rainfall data. In *Advances in Neural Information and Processing Systems*, volume 10, pages 985–991, 1998.

[216] P.S. Wilson and R. Toumi. A fundamental probability distribution for heavy rainfall. *Geophys. Res. Lett.*, 32(L14812, doi :10.1029/2005GL022465), 2005.

[217] A.W. Wood, L.R. Leung, V. Sridhar, , and D. Lettenmaier. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Clim. Change*, 62 :189–216, 2004.

[218] C. Yang, R.E. Chandler, and V.S. Isham. Spatial-temporal rainfall simulation using generalized linear models. *Water Resources Res*, 41, 2005.

[219] P. Yiou and M. Nogaj. Extreme climatic events and weather regimes over the north atlantic : When and where ? *Geophys. Res. Lett.*, 31(L07202, doi :10.1029/2003GL019119), 2004.

[220] Z. Zhang and J. Huang. Extremal financial risk model and portfolio evaluation. *Comput. Stat. Data Anal.*, 51 :2313–2338, 2005.

[221] S. Zimov, E. Schuur, and F. Chapin. Permafrost and the global permafrost and the global carbon budget. *Science*, 312(doi :10.1126/science.1128908) :1612–1613, 2006.

[222] E. Zorita and H. von Storch. The analog method as a simple statistical downscaling technique :comparison with more complicated methods. *Journal of Climate*, 12 :2474–2489, 1999.

[223] I.I. Zveryaev. Seasonally varying modes in long-term variability of european precipitation during the 20th century. *Journal of Geophysical Research - Atmospheres*, 111, 2006.

ANNEXES

Annexe A : Quelques figures supplémentaires pour illustrer la section 2.4

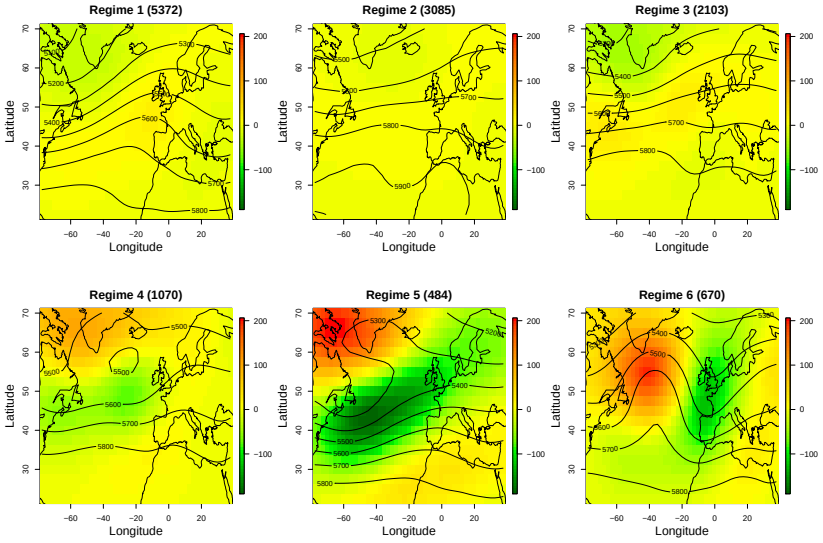


FIGURE 6.1: Cartes des six régimes de temps saisonniers définis par EM. Les couleurs correspondent aux anomalies moyennes de Z500 par régime. Les lignes noires représentent les hauteurs de géopotentiel (en m). Pour chaque carte, le nombre entre parenthèse indique combien de jours sont attribués aux régimes.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
WR 1	0.004	0.008									-0.006	
WR 2					3.10^{-4}	0.005	7.10^{-4}	7.10^{-4}	0.007	0.002		
WR 3				0.002		-0.006	-6.10^{-4}	-7.10^{-4}	-0.008	-0.004	0.004	
WR 4				0.003	0.005	0.005				0.003		
WR 5	-0.003	-0.009									8.10^{-4}	-0.002
WR 6				-0.004	-0.004	-0.004					0.004	0.001

TABLE 6.1: Valeurs des pentes (significatives à 95%, c-à-d., $\alpha = 0.05$) des tendances de fréquences de chaque régime saisonnier calculées sur 1975-2009.

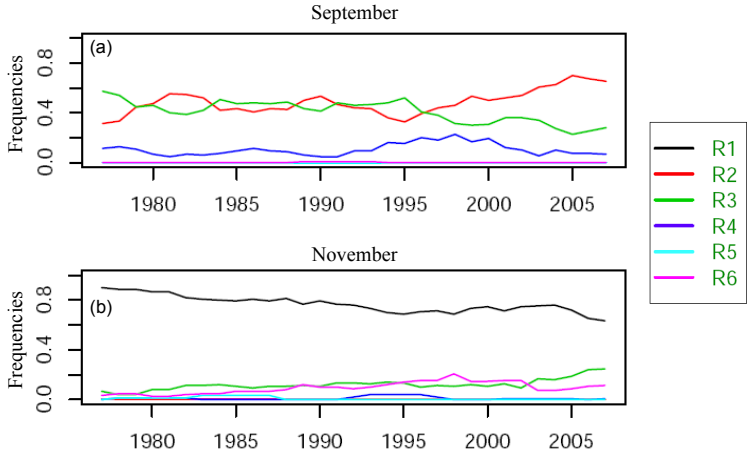


FIGURE 6.2: Évolution temporelle des fréquences mensuelles des régimes pour deux mois exemples : (a) septembre et (b) novembre. Ces évolutions sont calculées par moyenne glissante de cinq années sur la période 1977-2007.

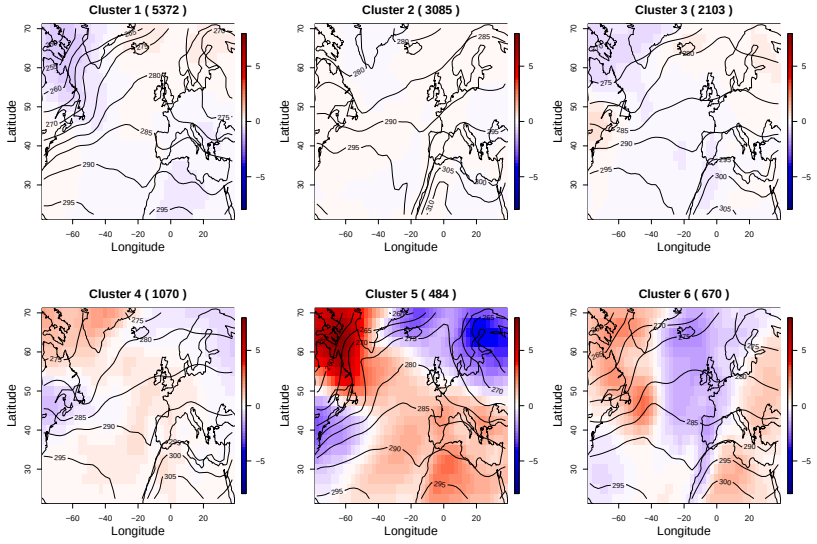


FIGURE 6.3: Tendances de température (en^{o}C) par régime saisonnier sur 1975-2009, ainsi que sans conditionnement par RTs (en bas).

Annexe B : Les étapes de l'algorithme "Correlation Clustering Model" (CCM)

Les étapes de CCM sont le suivantes :

1. Initialisation : les clusters (de jours) sont aléatoirement choisis, c-à-d., chaque jour est aléatoirement attribué à l'un des K clusters.
2. Modélisation : Pour chaque cluster $k = 1,\ldots,K$, une CCA est réalisée pour construire le $k^{ème}$ modèle de CCA $CM_k = \{(v_j^k, w_j^k), r_j^k, (A_j^k, B_j^k); j = 1,\ldots,M\}$, correspondant aux M paires de CVs (où, par construction, M est le minimum de la dimension des deux jeux de données, voir [3] par exemple ou tout livre d'analyses statistiques), la corrélation r_j des CVs de la $j^{ème}$ paire et les M paires (A_j^k, B_j^k) de vecteurs canoniques du cluster k (c-à-d., les vecteurs contenant les coefficients linéaires permettant de passer des données initiales aux CVs).
3. Attribution : Chaque jour est ré-attribué à un cluster, à partir des valeurs locales de précipitation (X) et des données atmosphériques à grande échelle (Y) et des K modèles de CCA($CM_{1,\ldots,K}$). Les détails de cette étape d'attribution sont donnés plus loin dans les Eqs. (6.1–6.3).
4. Si l'attribution a changé (c-à-d., si les clusters sont différents) depuis la dernière itération, retourner à l'étape 2. Sinon, l'algorithme s'arrêtent et les K clusters (RTs) et modèles de CCA sont ceux cherchés.

Plus précisément, l'étape d'attribution est réalisée ainsi : Pour chaque cluster k et son modèle de CCA CM_k, une régression linéaire est calculée pour chaque paire de CVs :

$$\hat{v}_j^k = a_j^k \times w_j^k + b_j^k, \qquad j = 1,\ldots,M. \tag{6.1}$$

Puis, pour chaque jour (caractérisé par X et Y) et chaque cluster k, les CVs sont calculés par le modèle CM_k :

$$v_j^k = A_j^k X'^{\{k\}} \quad \text{et} \quad w_j^k = B_j^k Y'^{\{k\}}, \quad j = 1,\ldots,M, \tag{6.2}$$

où $X'^{\{k\}}$ et $Y'^{\{k\}}$ sont les données centrées du $k^{ème}$ cluster. L'estimation $\hat{v}_j^k$ de v_j^k est ensuite calculée selon l'Eq. (6.1) et l'erreur pondérée err^k

$$err^k = \sum_{j=1}^{M} \frac{r_j^k}{r_1^k} \times (v_j^k - \hat{v}_j^k)^2, \tag{6.3}$$

où r_j^k / r_1^k est le poids de la $j^{ème}$ erreur. Le jour est attribué au cluster k qui minimise err^k. Notons que r_j^k (c-à-d, la corrélation entre v_j^k et w_j^k) est décroissante quand j augmente : le poids de la première erreur est 1 et les poids des autres sont plus petits et fonctions des corrélations entre les autres CVs. L'erreur pondérée concentre ainsi plus de poids sur les variables canoniques fortement corrélées.

Annexe C : Réseaux de neurones

Soit H le nombre d'unités cachées. Chaque unité cachée h ($h = 1, \ldots, H$) calcule une combinaison linéaire des prédicteurs (c-à-d., les entrées) x_i, combinaison qui subit ensuite une transformation non-linéaire par une tangente hyperbolique (tanh) :

$$z_h = \tanh\left(\sum_{i=1}^{d} v_{h,i} x_i + v_{h,0}\right), \tag{6.4}$$

où $v_{h,i}$ sont les poids liant les prédicteurs aux unités cachées. Une combinaison linéaire de ces valeurs z_h plus une combinaison linéaire des prédicteurs (c-à-d., la connexion linéaire supplémentaire mentionnée en sous-section 3.2.2) sont transformées par une fonction $g(\cdot)$ permettant d'assurer la positivité :

$$\psi_j = g\left(\underbrace{\sum_{h=1}^{H} w_{j,h} z_h}_{\text{partie non-linéaire}} + \underbrace{\sum_{i=1}^{d} \tilde{v}_{j,i} x_i}_{\text{partie linéaire}} + w_{j,0}\right), \tag{6.5}$$

où les $w_{j,h}$ sont les poids des unités cachés, les $\tilde{v}_{j,i}$ sont les poids linéaires de la connexion supplémentaire et $g(\cdot)$ est choisie selon les paramètres ψ_j du mélange (voir [25] par exemple pour les détails sur $g(\cdot)$). Soit ω les poids du réseau, c-à-d., les $v_{h,i}$ de l'équation (6.4) et les $w_{j,h}$ et $\tilde{v}_{j,i}$ de (6.5). Le mélange conditionnel peut alors s'écrire :

$$\phi_\omega(y|x) = \phi(y; \psi_\omega(x)), \tag{6.6}$$

où $\phi(\cdot;\psi)$ est défini dans l'équation (3.5) et $\psi_\omega(x)$ souligne que les paramètres du mélange dépendent des poids ω. Ceux-ci sont calibrés par minimisation de la log-vraisemblance négative du mélange conditionel sur un ensemble d'entrainement. L'optimisation est faite avec un algorithme de descente de gradient conjugué et le gradient est calculé par un algorithme de rétropropagation [154]. Afin d'éviter des minima locaux, l'optimisation est relancée plusieurs fois à partir de différentes valeurs initiales du réseau et les poids fournissant la plus faible erreur d'apprentissage sont gardés. Cette procédure aide à stabiliser l'optimisation et ainsi les performances de chaque algorithme. Le niveau de complexité de chaque CMM – c-à-d., leur degré d'adaptivité – est contrôlé à la fois par le nombre d'unités cachées et par le nombre de composantes du mélange : ce sont ce qu'on appelle des "hyper-paramètres". Ces hyper-paramètres doivent être sélectionnés précautionneusement pour avoir le meilleur compromis entre le "biais" (l'inadéquation du modèle aux données) et la variance (également appelée "sur-apprentissage"). Ceci a été réalisé grâce à une méthode de validation croisée. Le lecteur intéressé par ces détails de validation et méthodologiques peut se reporter à la lecture de [27].

Annexe D : Définition des prédicteurs et pré-traitement des données pour l'étude [179] de modélisation statistique des débits

La définition des prédicteurs pour cette étude consiste à synthétiser les 27 variables atmosphériques NCEP/NCAR initialement disponibles en un nombre réduit de prédicteurs modérément corrélés et interprétables physiquement, c-à-d., associés à des processus atmosphériques identifiés. Cette définition est schématisée en Fig. 6.4.

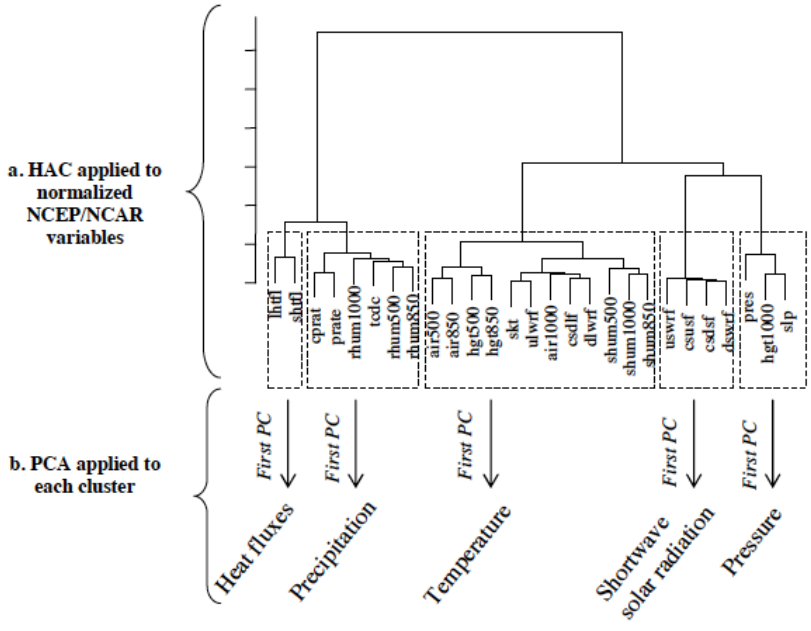


FIGURE 6.4: Schématisation du processus de pré-traitement des prédicteurs pour l'étude [179] de modélisation statistique des débits

Réduire les corrélations entre prédicteurs permet souvent d'améliorer la robustesse de la relation liant prédicteurs et *prédictants* (ici des débits). De plus, limiter le nombre de prédicteurs réduit le temps de calculs pour la calibration et les simulations. La définition des prédicteurs s'est faite en deux étapes :

1. Une méthode de clustering ascendante hiérarchique (HAC, voir par [199]) avec critère de Ward ([205]) a été appliquée à la matrice des distances Euclidiennes des 27 variables mensuelles NCEP/NCAR disponibles qui ont été normalisées auparavant. Les variables ayant des "comportements" similaires ont ainsi été regroupées et celles avec des comportements distincts ont été séparées. Cinq classes (ou clusters) ont été obtenues – ce nombre ayant été sélectionné par "in-

formation de silhouette" (voir [153, 179]) – et caractérisent des processus liés à la précipitation, la température, la pression, le rayonnement et les flux de chaleurs (Fig. 6.4a).

2. Une ACP a ensuite été appliquée sur chacun des cinq clusters de variables séparément pour dériver des descriptions synthétiques de chaque processus. La première composante principale (PC) de chaque classe (contenant à chaque fois plus de 80% de la variance) a été retenue comme prédicteur (Fig. 6.4b).

Avant le processus de calibration, les données de débits ont été standardisées par station. Pour chaque station, le débit annuel moyen a été soustrait de la série temporelle journalière afin de rendre comparables les "dimensions" des débits d'une station à une autre. Ces données standardisées ont ensuite "Gaussiannisées" par la transformation puissance de Box-Cox ([16]) afin que l'hypothèse de normalité des modèles GLM et GAM soit la plus valide possible.

Par ailleurs, cinq régimes hydrologiques ont été identifiés (voir Fig. 6.5) à partir des pourcentiles 10%, 50% et 90% des données mensuelles de débits standardisés. Pour cela, une méthode HAC fût appliquée sur la matrice de distance Euclidienne de ces pourcentiles mensuels. Notons qu'aucune autre données caractérisant les bassins n'a été utilisée pour la définition de ces régimes hydrologiques.

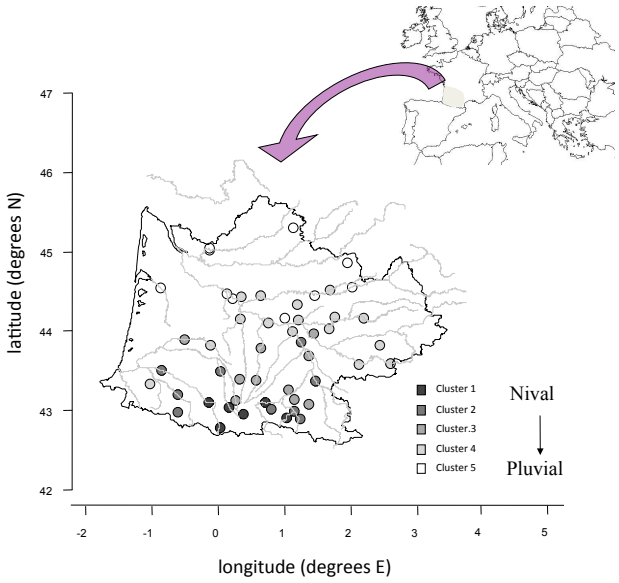


FIGURE 6.5: Les cinq régimes hydrologiques identifiés par HAC à partir des pourcentiles 10%, 50% et 90% des données mensuelles de débits standardisés.

Rappel de la définition du coefficient de détermination R^2 :

$$R^2 = \frac{\sum_{i=1}^{n}(O_i - \bar{O})(S_i - \bar{S})}{\sqrt{[\sum_{i=1}^{n}(O_i - \bar{O})^2] \times [\sum_{i=1}^{n}(S_i - \bar{S})^2]}}, \tag{6.7}$$

où O_i et S_i sont respectivement les observations et simulations pour l'année i et $\bar{O}$ et $\bar{S}$ les moyennes des observations et des simulations sur l'ensemble des n années.

Annexe E : Rappel de quelques modèles statistiques de fonction de transfert utilisés dans [179]

Les réseaux de neurones artificiels ayant déjà étant brièvement présentés en annexe A, seuls les modèles linéaires généralisés, les modèles additifs généralisés et les arbres "boostés" agrégés sont rappelés dans cette annexe.

Modèles Linéaires Généralisés

Les modèles linéaires généralisés (GLM) sont une généralisation de la régression ordinaire par moindres carrés, unifiant divers autres modèles statistiques tels que la régression logistique dans un unique cadre formel ([121]). Dans les GLMs employés dans l'étude [179], chaque sortie du prédictant Y est supposée générée à partir d'une distribution de la famille exponentielle incluant la loi normale, binomiale ou de Poisson. Les données de débits sont supposés gaussiennes après transformation de Box-Cox (voir annexe B). La moyenne μ de la distribution dépend des prédicteurs X. Le modèle était défini comme :

$$g(E(Y|X)) = \beta X + \alpha, \tag{6.8}$$

où $E(Y|X)$ est l'espérance conditionnelle de Y sachant X, β correspond au vecteur de paramètres à estimer et α à l'intercept. La fonction g est appelée la "fonction de lien" et peut prendre n'importe quelle forme (voulue par l'utilisateur) selon la famille de distribution choisie. Dans [179], comme Y est supposé gaussien, $E(Y|X)$ est directement relié au membre de droite de l'Eq. (6.8), autrement dit, g est la fonction identité, $g(x) = x$ (voir [73] pour les détails techniques et théoriques).

Modèles Additifs Généralisés

Les modèles additifs généralisés (GAM) ont été développés pour étendre les propriétés des GLMs à des relations non-linéaires entre X et Y ; grâce à des propriétés additives ([73]). Les GAMs modélisent l'espérance conditionnelle de Y sachant $X = (X_1, ..., X_m)$ comme étant la somme de m fonctions splines f_i appliquées aux prédicteurs X_i :

$$g(E(Y|X)) = \sum_{i=1}^{m} f_i(x_i) + \theta_0 \tag{6.9}$$

Comme pour les GLMs, GAM spécifie une distribution pour la variables réponse Y. Les fonctions f_i peuvent être paramétriques ou non, fournissant ainsi la capacité pour des ajustements non-linéaires, ce que GLM ne permet pas. Dans [179], les f_i sont des splines cubiques, construits par association de polynômes de degré trois par morceaux avec des conditions de continuité jusqu'à la dérivée seconde. Le paramètre θ_0 est une constante à estimer et g est la fonction identité.

Arbres "boostés" agrégés

Les arbres "boostés" agrégés (aggregated boosted regression trees – ABT) furent introduit par [60] et [74] en statistiques appliquées à l'écologie. Les arbres boostés (BT) sont basés sur une compilation de modèles d'arbres de classification et de régression (CART). Ces modèles CART expliquent les variations du prédictant Y en coupant successivement les données en groupes (noeuds) de plus en plus homogènes en utilisant des combinaisons de prédicteurs [1]. Chaque groupe final est caractérisé par une valeur typique du prédictant, le nombre d'observations dans ce groupe et les prédicteurs qui le définissent. Le but des BTs est d'améliorer les performances d'un unique modèle CART en ajustant m modèles ($m = 1000$ dans [179]) où chaque CART successif est construit pour modéliser les résidus de prédiction du précédent (détails techniques dans [47]). Pour limiter la sur-paramétrisation potentiellement causée par de trop nombreux modèles CART, chaque nouveau CART est construit sur un sous-ensemble de données choisi aléatoirement. Le nombre optimal d'arbres est ensuite sélectionné automatiquement de manière à minimiser la perte de pouvoir prédictif (détails dans [40]).

Les arbres "boostés" agrégés sont eux-mêmes une extension des arbres boostés et comprennent un ensemble de BTs générés sur des sous-ensembles par validation croisée. Ils permettent alors de réduire l'erreur de prediction par rapport à un unique arbre boosté (détails dans [40]).

Annexe F : Définition des prédicteurs géographiques utilisés dans l'étude [196] pour la régionalisation du climat du DMG

Continentalité diffusive

L'indice de continentalité diffusive Dco (entre 0 et 1) correspond à la plus courte distance à l'océan. Si un point est proche d'une mer ou d'un océan, Dco est proche de 0. Réciproquement, un point éloigné de la mer a un Dco proche de 1.

1. Ces prédicteurs peuvent être catégoriels et/ou numériques.

Continentalité advective

La continentalité advective Aco prend ses valeurs entre 0 et 1. Elle est associée aux intensités et directions de vent en étant basée sur les hypothèses suivantes. Une masse d'air devient progressivement continentale (ou inversement, maritime) quand elle voyage à l'intérieur des terres (océans). Le taux de ce changement est supposé être une fraction constante (τ) par unité de temps, c-à-d., le changement de continentalité durant un temps dt est

$$dC = [-C(1-i_{co}) + (1-C)(i_{co})]\tau dt, \tag{6.10}$$

avec C la continentalité (entre 0 et 1), i_{co} vaut 0 sur mer et 1 sur terre. Le paramètre τ satisfait $\tau dt = \tau \frac{dx}{dt}\frac{1}{U}dt = \frac{dx/U}{l_0/U_0} ln(2)$, où dx est la distance voyagée par la masse d'air pendant le temps dt, U est l'intensité moyenne du vent issu de CLIMBER et l_0/U_0 est le ratio distance/vent correspondant à un changement de continentalité de $1/2$. Ce ratio est posé comme étant $\frac{l_0}{U_0} = \frac{5.10^5 m}{5 m/s}$. Pour calculer la continentalité en un point, nous devons tout d'abord intégrer le changement de continentalité le long du "chemin" de chaque masse d'air arrivant en ce point :

$$C_d = \int_{chemin} dC = \int_{chemin} [-C(1-i_{co}) + (1-C)(i_{co})] \frac{ln(2)/U}{l_0/U_0} dx. \tag{6.11}$$

Il est ensuite nécessaire de décider du poids respectif de chaque direction de chemin. Pour cela, il est raisonnable de poser les simples règles suivantes : (1) donner plus de poids aux directions correspondant à celle du vent moyen, et (2) donner un poids nul aux chemins en opposition au vent moyen. Pour cela, nous utilisons le produit scalaire du vent moyen U et du vecteur unitaire $\hat{l}_p$ de la direction du chemin (intégré le long de chaque chemin) :

$$I_d = \int_{chemin} max(\hat{l}_p.U, 0) dC. \tag{6.12}$$

La moyenne pondérée des contributions de tous les chemins fournit alors la continentalité au point voulu :

$$C = \frac{\sum_d I_d C_d}{\sum_d I_d}. \tag{6.13}$$

W-slope

La variable W-slope prend en partie en compte l'impact des montagnes sur le climat régional. Elle est calculée séparément de la continentalité advective mais d'une manière similaire. Comme pour la continentalité, plusieurs masses d'air arrivant en un point sont considérées, avec la même pondération que précédemment, c-à-d., par (6.12). Le prédicteur W-slope correspond à la moyenne du vent zonal multipliée par la moyenne de l'angle de la pente est-ouest du terrain sur approximativement 100 km. En effet, le vent dominant de CLIMBER vient de l'ouest. Seules les tendances à la "montée" sont ici retenues : le W-slope augmente seulement quand la masse d'air monte, se

rafraichissant ainsi et précipitant potentiellement.

Remarquons que Aco et Wsl ne sont pas “purement géographiques” dans le sens où elles ne sont pas complètement statiques. La classification en prédicteurs “physiques” ou “géographiques”est partiellement arbitraire et reflète davantage une opposition entre prédicteurs “classiques” et ceux, plus originaux, proposés dans notre article [196].