Scénarios de variables hydrométéorologiques en région méditerranéenne : approches stochastiques et semi-paramétriques

Habilitation à Diriger des Recherches

Julie Carreau

15 avril 2019

Table des matières

Chapitre 1

CV et autres éléments

1.1 CV

Mission professionnelle à l'IRD

Développement et renforcement des capacités dans les pays du Sud, en particulier dans le bassin méditerranéen avec un focus sur la Tunisie

Situation professionnelle

- **Chargé de recherche en détachement** depuis 01/11/2018
  à l'Institut de Recherche pour le Développement (IRD)
- **Ingénieur de recherche** depuis 01/06/2010
  à l'Institut de Recherche pour le Développement (IRD)
- **Affectation structurelle et géographique actuelle :** depuis 01/06/2010
  HydroSciences Montpellier, Univ. Montpellier-CNRS-IRD, Montpellier, France
- **Affectation géographique à venir :** à partir de 01/09/2019
  École Supérieure des Communications de Tunis (Sup'Com), Tunis, Tunisie

Formation universitaire

- **Ph.D. en informatique**, spécialité apprentissage statistique 2002-2007
  Université de Montréal, Montréal, Canada
  Titre du mémoire : Modèles Pareto hybrides pour distributions asymétriques et à queues lourdes
  Directeur de thèse : Yoshua Bengio
- **M.Sc. en finance mathématique et computationnelle** 1999-2001
  Université de Montréal, Montréal, Canada
  Titre du mémoire : Système de transactions automatiques sur le marché des contrats à terme sur le taux d'intérêt
  Directeurs de maîtrise : Yoshua Bengio et René Garcia
- **B.Sc. en mathématiques fondamentales** 1995-1998
  Université de Montréal, Montréal, Canada
- **Validation de crédits en séjour inter-universitaire** 1996-1997
  Université de Bologne, Bologne, Italie

Bourses d'études d'excellence

- **Boursière M.Sc., Ph.D. et postdoctorale** 1999-2009
  Fonds de recherche sur la nature et les technologies,
  organisme subventionnaire du gouvernement du Québec, Canada
- **Boursière PRECARN** 2004-2005
  Institut de robotique et de systèmes intelligents, Canada
- **Boursière B.Sc. en milieu industriel** 1998
  Conseil de recherche en sciences naturelles et génie,
  organisme subventionnaire du gouvernement du Canada

Expériences professionnelles

- **Chercheuse postdoctorale** 2009-2010
  Équipe Mistis, INRIA Rhône-Alpes, Grenoble, France
  Encadrement : Stéphane Girard
- **Chercheuse postdoctorale** 2007-2009
  Équipe EstimR, Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ-IPSL, Gif-sur-Yvette, France
  Encadrement : Philippe Naveau
- **Consultante en recherche opérationnelle** 2007
  Ex Pretio, Montréal, Canada
- **Consultante en ingénierie financière** 2000-2003
  Département des marchés financiers de Hydro-Québec, Montréal, Canada
- **Analyste en recherche et développement** 2001
  Trésorerie de la Banque Nationale du Canada, Montréal, Canada
- **Stage de M.Sc. en ingénierie financière** 2000
  Département des marchés financiers de Hydro-Québec, Montréal, Canada
- **Assistante de recherche** 1999
  Centre interuniversitaire de recherche en analyse des organisations, Montréal, Canada

Activités de recherche

- **Enjeux sociétaux :** problématiques spécifiques en région méditerranéenne pour la gestion des ressources en eau et des risques naturels accentuées par le changement climatique (e.g. rareté de la ressource en eau, crues éclair) ;
- **Enjeux scientifiques :** contraindre des modèles à base physique qui permettent de développer des outils d'aide à la décision à l'aide de scénarios de variables hydrométéorologiques ;
- **Verrous méthodologiques :** forte variabilité spatio-temporelle, dépendances inter-variables, non-gaussiannité, présence d'événements extrêmes, configurations instrumentales spécifiques ;
- **Objectif principal :** développer, à l'aide d'approches stochastiques et semi-paramétriques, et évaluer des scénarios de variables hydrométéorologiques.

Activités d'enseignement

- **Introduction aux statistiques et probabilités** (20 h / an) depuis 2013
  École Polytech' Montpellier, France
- **Séminaire doctoral spécialisé en statistiques** (18 h / an) 2016-2017
  École Sup'Com Tunis, Tunisie
- **Modélisation du risque climatique** (24 h / an) 2013-2014
  École Centrale Marseille, France
- **Introduction aux statistiques pour l'hydrologie** (18 h / an) 2012
  Université de Montpellier, France
- **Introduction à l'apprentissage statistique et mathématiques pour l'informatique** 2003-2006
  Université de Montréal, Canada

Activités de valorisation

- **Initiation à R et méthodes de correction de biais** Banyuls-sur-mer, 1/2 journée en 2019
  École de printemps avec participants mixtes (étudiants, chercheurs et cadres techniques des organismes de gestion du sud de la Méditerranée), organisée par l'action transverse *Impacts-CC* du programme MISTRALS.
- **Méthodes de changement d'échelle avec R** Marrakech, 1/2 journée en 2017
  École d'hiver "Systèmes Hydro-Agricoles Méditerranéens" avec participants mixtes (étudiants, chercheurs et cadres techniques des organismes de gestion du sud de la Méditerranée), organisée par les LMI TREMA, MediTer et NAÏLA.
- **Statistiques pour l'hydrologie avec R** Tunis, 4 jours en 2014
  Direction Générale des Ressources en Eau, Tunis, Tunisie

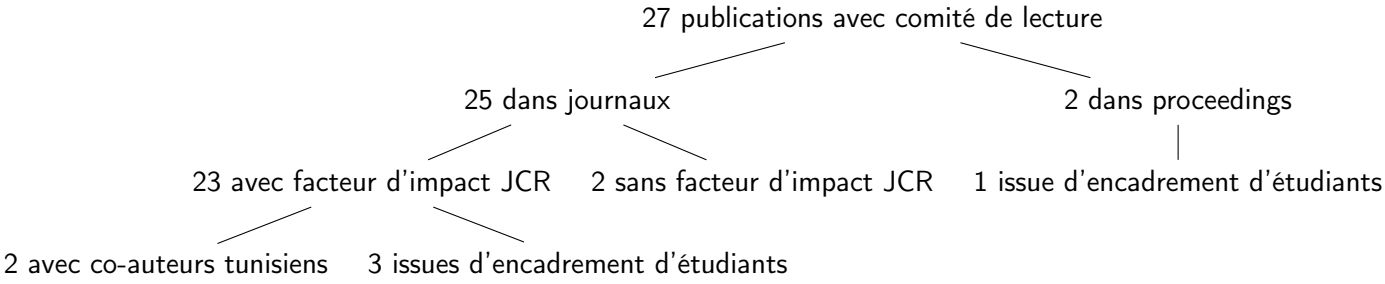
Responsabilités académiques

- **Activités d'évaluation**
  - Membre d'un jury de recrutement (CEA / IR)
  - Examinatrice pour jury de thèse (Sup'Com, Tunisie)
  - Membre de comités de suivi de trois thèses (UR Riverly, IRSTEA Lyon)
  - Évaluatrice pour appels à projets (Fonds National Suisse, Labex OSUG, LEFE/IMAGO)
- **Relectrice :** pour environ 15 journaux en statistiques appliquées et en sciences de l'environnement
- **Membre de sociétés savantes**
  - Comité *Probability and Statistics in the Physical Sciences* de la Société Bernoulli
  - Société Française de Statistiques
- **Conférences**
  - Responsable des subventions dans le comité d'organisation de METMA IX, colloque international ayant eu lieu à Montpellier en 2018
  - Organisatrice de 4 sessions dans des conférences internationales en statistiques dans les sciences de l'environnement

Développement de boîtes à outils logiciels

- **Développeuse et éditrice de la librairie *condmixt***
  Mise en oeuvre de la modélisation statistique issue de mes travaux de thèse et de postdoctorat.
  Dans l'environnement R disponible sur le dépôt officiel CRAN - impact difficile à mesurer.
- **Contributrice principale de l'interface graphique *Caterpillar***
  Extraction des simulations de modèles de climat globaux et descente d'échelle statistique.
  En `java` pour l'interface graphique avec externalisation vers R pour la descente d'échelle statistique.
  Déploiement à l'interne à HydroSciences Montpellier.

1.2 Liste des publications indexées



Sur les 27 publications avec comité de lecture :

- 2 avec co-auteurs tunisiens, soit 30 % depuis que j'assure un présentiel en Tunisie (noms en bleu)
- 4 issues d'encadrement d'étudiants (noms des étudiants au nord en mauve)

Avec facteur d'impact JCR

23. Palacios-Rodriguez, F., Toulemonde, G., Carreau, J., & Opitz, T. Generalized Pareto processes for simulating space-time extreme events. *JRSS-C*, soumis
22. Carreau, J., & Toulemonde, G. Extra-Parametrized Extreme Value Copula : Extension to a Spatial Framework. *Spatial Statistics*, soumis
21. Beaufort, A., Carreau, J., & Sauquet, E. A classification approach to reconstruct local daily drying dynamics at headwater streams. *Hydrological Processes*, 2019, 1–17
20. Carreau, J., Ben Mhenni, N., Huard, F., & Neppel, L. Exploiting the spatial pattern of daily precipitation in the analog method for regional temporal disaggregation. *Journal of Hydrology*, 2019, **568**, 780–791.
19. Carreau, J., Naveau, P., & Neppel, L. Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation. *Water Resources Research*, 2017, **53**(5), 4407–4426.
18. Carreau, J., & Bouvier, C. Multivariate density model comparison for multi-site flood-risk rainfall in the French Mediterranean area. *Stochastic Environmental Research and Risk Assessment*, 2016, **30**(6), 1591–1612.
17. Foughali, A., Tramblay, Y., Bargaoui, Z., Carreau, J., & Ruelland, D. Hydrological modeling in Northern Tunisia with regional climate model outputs : Performance evaluation and bias-correction in present climate conditions. *Climate*, 2015, **3**(3), 459–473.
16. Ayar, P. V., Vrac, M., Bastin, S., Carreau, J., Déqué, M., & Gallardo, C. Intercomparison of statistical and dynamical downscaling models under the EURO-and MED-CORDEX initiative framework : present climate evaluations. *Climate dynamics*, 2016, **46**(3-4), 1301–1329.
15. Soubeyroux, J.-M., Veysseire, J.-M., Gouget, V., Neppel, L., Tramblay, Y., & Carreau, J. Evolution of extreme rainfall in France with a changing climate. *La Houille Blanche*, 2015, **1**, 27–33.
14. Braud, I., Ayral, P. A., Bouvier, C., Branger, F., Delrieu, G., Le Coz, J., Nord, G., Vandervaere, J. P., Anquetin, S., Adamovic, M., Andrieu, J., Batiot, C., Boudevillain, B., Brunet, P., Carreau, J., Confoland, A., Didon-Lescot, J. F., Domergue, J. M., Douvinet, J. Dramais, G., Freydier, R., Gérard, S., Huza, J., Leblois, E., Le Bourgeois, O., Le Boursicaud, R., Marchand, P., Martin, P.,

Nottale, L., Patris, N., Renard, B., Seidel, J. L., Taupin, J.-D., Vannier, O., Vincendon, B., & Wijbrans, A. Multi-scale hydrometeorological observation and modelling for flash-flood understanding. *Hydrology and Earth System Sciences*, 2014, **18**(9), 3733–3761.

13. Lang, M., Arnaud, P., Carreau, J., Deaux, N., Dezileau, L., Garavaglia, F., Latapie, A., Neppel, L., Paquet, E., Renard, B., Soubeyroux, J. M., Terrier, B., Veysseire, J. M., Aubert, Y., Auffray, A., Borchi, F., Bernardara, P., Carre, J. C., Chambon, D., Cipriani, T., Delgado, J. L., Doumenc, H., Fantin, R., Jourdain, S., Kochanek, K., Paquier, A., Sauquet, E., & Tramblay, Y. Main results of a French project on extreme rainfall and flood assessment. *La Houille Blanche*, 2014, **2**, 5–13.
12. Neppel, L., Arnaud, P., Borchi, F., Carreau, J., Garavaglia, F., Lang, M., Paquet, E., Renard, B., Soubeyroux, J. M., & Veysseire, J. M. Comparison of extreme rainfall frequency analysis methods in France. *La Houille Blanche*, 2014, **2**, 14–19.
11. Renard, B., Kochanek, K., Lang, M., Garavaglia, F., Paquet, E., Neppel, L., Najib, K., Carreau, J., Arnaud, P., Aubert, Y., Borchi, F., Soubeyroux, J.-M., Jourdain, S., Veysseire, J.-M., Sauquet, E., Cipriani, T., & Auffray, A. Data-based comparison of frequency analysis methods : A general framework. *Water Resources Research*, 2013, **49**(2), 825–843.
10. Tramblay, Y., Neppel, L., Carreau, J., & Najib, K. Non-stationary frequency analysis of heavy rainfall events in southern France. *Hydrological sciences journal*, 2013, **58**(2), 280–294.
9. Ceresetti, D., Ursu, E., Carreau, J., Anquetin, S., Creutin, J.-D., Gardes, L., Girard, S., & Molinie, G. Evaluation of classical spatial-analysis schemes of extreme rainfall. *Natural hazards and earth system sciences*, 2012, **12**, 3229–3240.
8. Seghieri, J., Carreau, J., Boulain, N., De Rosnay, P., Arjounin, M., & Timouk, F. Is water availability really the main environmental factor controlling the phenology of woody vegetation in the central Sahel ? *Plant ecology*, 2012, **213**(5), 861–870.
7. Tramblay, Y., Neppel, L., Carreau, J., & Sanchez-Gomez, E. Extreme value modelling of daily areal rainfall over Mediterranean catchments in a changing climate. *Hydrological Processes*, 2012, **26**(25), 3934–3944.
6. Tramblay, Y., Neppel, L., & Carreau, J. Brief communication" Climatic covariates for the frequency analysis of heavy rainfall in the Mediterranean region". *Natural Hazards and Earth System Sciences*, 2011, **11**(9), 2463.
5. Carreau, J., & Vrac, M. Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resources Research*, 2011, **47**(10).
4. Carreau, J., Naveau, P., & Sauquet, E. A statistical rainfall-runoff mixture model with heavy-tailed components. *Water resources research*, 2009, **45**(10).
3. Carreau, J., & Bengio, Y. A hybrid pareto mixture for conditional asymmetric fat-tailed distributions. *IEEE transactions on neural networks*, 2009, **20**(7), 1087–1101.
2. Carreau, J., & Bengio, Y. A hybrid Pareto model for asymmetric fat-tailed data : the univariate case. *Extremes*, 2009, **12**(1), 53–76.
1. L'heureux, P.-J., Carreau, J., Bengio, Y., Delalleau, O., & Yue, S.Y. Locally Linear Embedding for dimensionality reduction in QSAR. *Journal of computer-aided molecular design*, 2004, **18**(7-9), 475–482.

Sans facteur d'impact JCR

2. Carreau, J., Neppel, L., Arnaud, P., & Cantet, P. Extreme rainfall analysis at ungauged sites in the South of France : comparison of three approaches. *Journal de la Société Française de Statistique*, 2013, **154**(2), 119–138.
1. Carreau, J., & Girard, S. Spatial extreme quantile estimation using a weighted log-likelihood approach. *Journal de la Société Française de Statistique*, 2011, **152**(3), 66–82.

Proceedings avec comité de lecture

2. Carreau, J., Dezetter, A., Aboubacar, H., & Ruelland, D. Evaluation and comparison of two downscaling methods for daily precipitation in hydrological impact studies. *IAHS-AISH publication*, 2013, 67–72.
1. Carreau, J., & Bengio, Y. A Hybrid Pareto Model for Conditional Density Estimation of Asymmetric Fat-Tail Data. *Pages 51–58 of :* Meila, Marina, & Shen, Xiaotong (eds), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research, 2007, vol. 2. San Juan, Puerto Rico : PMLR.

1.3 Direction d'étudiants

Encadrement : postdoctorat

Fatima Palacios-Rodriguez, (UM/IMAG, Inria/LEMON) Montpellier 2017 - 2018
Simulation de processus spatio-temporels intégrant des extrêmes pour mesurer le risque inondation : approches semi et non-paramétriques
co-encadrement avec Gwladys Toulemonde (UM/IMAG, Inria/LEMON) et Thomas Opitz (INRA/BioSP).

Encadrement : thèses

3 doctorantes provenant de cursus tunisiens (noms en bleu)

3. Hela Hammami, école Polytechnique de Tunis 2019-2021
   Générateur stochastique spatial de conditions météorologiques en rive sud de la Méditerranée
   co-encadrement avec Sadok El Asmi (Sup'Com).
2. Amal Saadi, INAT 2019-2021
   Étude de la variabilité spatiale et temporelle des sécheresses en climat sub-humide à semi-aride
   co-encadrement avec Zohra-Lili Chaabane (INAT), Haifa Feki (ESIM) et Sadok El Asmi (Sup'Com).
1. Nesrine Farhani, cotutelle INAT - Université Toulouse 3 2018-2020
   Apport de la Télédétection et des variables auxiliaires dans l'étude de l'évolution des périodes de sécheresses
   co-encadrement avec Zohra-Lili Chaabane et Zeineb Kassouk (INAT), Gilles Boulet (IRD/CESBIO) et Rim Zitouna (INRGREF).

Encadrement : masters

5 étudiantes provenant de cursus tunisiens (noms en bleu)

9. Rahma Timouni, INAT 2019
   Développement d'un outil d'analyse spatio-temporelle des observations agro-météorologiques. Cas du CapBon,
   co-encadrement avec Rim Zitouna (INRGREF) et Nesrine Farhani (INAT).
8. Aïcha Ilmi-Ahmed, Agrocampus Rennes 2019
   Typologie des champs d'événements extrêmes : application aux précipitations de la région méditerranéenne en contexte de changement climatique
   co-encadrement avec Gwladys Toulemonde (UM/IMAG, Inria/LEMON).
7. Nesrine Farhani, INAT 2017
   Spatialisation de variables météorologiques à l'aide de variables auxiliaires sur le bassin de la plaine du Merguellil, Tunisie
   co-encadrement avec Gilles Boulet (IRD/CESBIO) et Zeineb Kassouk (INAT).
6. Nada Neji, ESIM 2017
   Evaluation et comparaison des produits satellitaires de précipitations avec les observations au sol dans le centre tunisien pour différentes résolutions temporelles
   co-encadrement avec Haifa Feki (ESIM).

5. Cléo Campion, Université de Montpellier 2015
   Modélisation et simulation d'extrêmes spatiaux : application aux hauteurs de vagues dans le golfe de Lion
   co-encadrement avec Jean-Noël Bacro (UM/IMAG) et Gwladys Toulemonde (UM/IMAG, Inria/LEMON).
4. Nada Ben Mhenni, ENIT 2015
   Désagrégation temporelle des cumuls de pluie journaliers
   co-encadrement avec Emna Gargouri et Rim Chérif (ENIT/LMHE) et Frédéric Huard (INRA/Agroclim).
3. Salah Neffaa, Université d'Aix-Marseille 2013
   Conception et réalisation d'une boîte à outils logicielle pour la génération de scénarios climatiques
   co-encadrement avec Alain Dezetter (IRD/HSM).
2. Nesrine Hamdi, ENIT 2013
   Désagrégation spatiale des précipitation issues de modèles climatiques régionaux : Application au bassin versant de la Koshi, Himalaya
   co-encadrement avec Luc Neppel (UM/HSM).
1. Hattoibe Aboubacar, Polytech' Clermond-Ferrand 2011
   Désagrégation spatio-temporelle de variables climatiques issues des modèles de circulation générale en Méditerranée
   co-encadrement avec Alain Dezetter (IRD/HSM).

Encadrement : projets de fin d'études

6 étudiants provenant de cursus tunisiens (noms en bleu)

9. Sonia Naffouti, ESSAI 2018
   Krigeage en présence de distribution non-gaussienne : application à la pluie et à la température journalière sur le Cap Bon, Tunisie
8. Zakaria Zayen, Sup'Com 2017
   Étude de la variabilité spatiale et temporelle de la pluie et du vent sur le bassin du Lebna, Tunisie
   co-encadrement avec Riadh Abdelfattah (Sup'Com) et Rim Zitouna (INRGREF).
7. Boudour Kharrat, ESIM 2017
   Modélisation hydrologique et hydrodynamique du bassin versant de l'oued Mellegue
   co-encadrement avec Haifa Feki (Sup'Com) et Jalel Aouissi (INAT).
6. Sabri Galai, ESSAI 2016
   Analyse des tendances temporelles des précipitations extrêmes en Méditerranée française
   co-encadrement avec Luc Neppel (UM/HSM).
5. Yasmine Ebnibrahim, ESSAI 2016
   Analyse multi-échelles des précipitations extrêmes en Méditerranée française
   co-encadrement avec Luc Neppel (UM/HSM).
4. Omar Charfeddine, ESSAI 2014
   Comparaison des modèles de processus spatiaux des pluies extrêmes pour l'estimation du risque hydrologique
   co-encadrement avec Gwladys Toulemonde (UM/IMAG, Inria/LEMON).
3. Florent Blondot, Polytech' Lyon 2013
   Sélection de variables atmosphériques pour le downscaling statistique de la précipitation
   co-encadrement avec Mathieu Vrac (CNRS/LSCE).

2. Mounia Dkhissi, Polytech' Clermond-Ferrand 2013
   Typologie des pluies cévenoles.

1. Ali Mohammed Nassur, IUT Nice Côte d'Azur 2012
   Générateur de pluie multi-sites à pas de temps journalier sur le bassin versant du Gardon à Anduze.

1.4 Collaborations

1.4.1 Collaborations doctorat et postdoctorat

Lors de mon doctorat à l'université de Montréal, j'ai travaillé essentiellement avec mon directeur de thèse. J'ai développé lors de mes deux postdoctorats au Laboratoire des Sciences du Climat et de l'Environnement (LSCE) à Gif-sur-Yvette et, plus brièvement, à l'Inria Rhône-Alpes, un réseau de collaborations dans le milieu de la statistique appliquée à l'environnement que j'ai entretenu et renforcé après ma prise de poste à Montpellier en 2010, en particulier avec le LSCE et l'UR Riverly, Irstea Lyon.

1.4.2 Collaborations IRD - Tunisie

Mes collaborations en Tunisie ont pour cadre le laboratoire mixte international (LMI) NAÏLA. Le LMI est un dispositif de l'IRD formalisant la coréalisation de projets de recherche, de formation et d'innovation entre des unités sous tutelle de l'IRD et leurs partenaires dans les pays en développement. Le LMI NAÏLA, ayant démarré en 2016 pour une durée de 5 ans renouvelable une fois, se consacre à la gestion des ressources en eau dans les milieux ruraux tunisiens.

Via les dispositifs de l'IRD pour la projection au sud (missions longues durées et affectations à l'étranger), j'ai obtenu le financement de trois séjours de deux mois de 2016 à 2018 à Sup'Com (voir le tableau 1.1 pour les acronymes). En septembre 2019, je partirai à Tunis pour une expatriation de deux ans, potentiellement quatre ans.

Suite à mon rapprochement avec le LISAH en 2014, HSM, notamment l'équipe EvExt à laquelle j'appartiens, s'est associé à la création du LMI NAÏLA. En conséquence, j'ai articulé mes activités de recherche en Tunisie avec le partenariat existant dans le cadre des actions IRD en Tunisie, avec, côté tunisien, l'INRGREF et l'INAT et côté français, le LISAH et le CESBIO (voir le tableau 1.1). De plus, j'ai apporté un nouveau partenariat avec une expertise complémentaire sur la modélisation statistique et la télédétection avec Sup'Com et l'ESIM (voir le tableau 1.1).

Ces collaborations franco-tunisiennes m'ont permis de m'impliquer dans de nombreux encadrements de stages de niveau PFE et M2 et dans celui de trois doctorats qui sont en cours, voir la section 1.3. Cette dynamique a également donné lieu au projet AMANDE qui est décrit à la section 1.5.

1.4.3 Collaborations IRD - France

Le pilier de mes collaborations en France est l'Institut de Mathématiques Alexander Grothendiek (IMAG) à l'université de Montpellier. Cette collaboration a donné lieu à des encadrements de stages d'étudiants (niveaux PFE et M2) et d'une postdoctorante, voir la section 1.3. J'ai participé activement à la création de projets communs qui ont été financés via le Labex NUMEV, le programme LEFE MANU de l'INSU / CNRS et le programme inter-institutionnel HYMEX / MISTRALS. Autour de cette collaboration avec l'IMAG, gravitent d'autres collaborations en statistiques, notamment avec le LSCE et l'équipe de Biostatistics and Spatial Processes (BioSP) de l'INRA à Avignon. Ces collaborations font partie intégrante des projets AMANDE et FRAISE décrits à la section 1.5.

Table 1.1 – *Partenariat et structures collectives dans le cadre de l'IRD.*

Affectation IRD / HydroSciences Montpellier (HSM)	Membre équipe Événements Extrêmes (EvExt)
Collaborations autres UMR IRD	• LISAH (Laboratoire des Interactions Sol-Agrosystème-Hydrosystème) • CESBIO (Centre d'Etudes Spatiales de la BIOsphère)
Programmes structurants IRD	• LMI NAÏLA (Tunisie) • ORE OMERE (Tunisie et France) • Système d'Observation Merguellil (Tunisie)
Partenariat dans le cadre des actions IRD en Tunisie	Intégration dans le partenariat existant • INRGREF (Institut National de Recherche en Génie Rural Eaux et Forêts) • INAT (Institut National Agronomique de Tunis) Création de nouveaux partenariats • Sup'Com (École Supérieure des Communications de Tunis) • ESIM (École Supérieure des Ingénieurs de Medjez El Beb)

1.5 Projets

Projet AMANDE, France-Tunisie, porteur français, $\sim$ 10 000 €en 2019

Le projet AMANDE (**A**pproches stochastiques et se**M**i-p**A**ramétriques combinées à la télédétectio**N** pour l'étude du stress hy**D**riqu**E**) est la concrétisation de mes activités de recherche menées dans le cadre du LMI NAÏLA depuis 2016. AMANDE est un projet franco-tunisien, dont je suis le porteur côté français et Sadok El Asmi (Sup'Com) est le porteur côté tunisien, qui est financé par le programme Hubert-Curien Utique sur la période 2019-2021. Ces fonds servent principalement à la mobilité d'étudiants entre la Tunisie et la France avec deux thèses en co-tutelle dont l'une en démarrage. AMANDE se décline en deux objectifs, dont les activités sont résumées dans la figure 1.1, ayant chacun des questions de recherches associées :

1. Générer et la valider des scénarios de variables hydrométéorologiques selon des approches stochastiques afin d'alimenter des modèles à base physique en contexte méditerranéen ;
   (a) Comment adapter un générateur stochastique spatial de conditions météorologiques en rive sud de la Méditerranée ?
   (b) Comment exploiter des données provenant des modèles de climat régionaux pour estimer les structures de dépendance dans un générateur stochastique de façon non-paramétrique ?
   (c) Quels apports pour la gestion des ressources en eau ?
2. Estimer des indicateurs du stress hydrique et étudier leurs tendances spatiales et temporelles ;
   (a) Comment exploiter les données de la télédétection active (radar) pour caractériser le stress hydrique ?
   (b) Comment calculer des trajectoires d'évolution d'indicateurs de stress hydrique pertinents pour l'agriculture ?

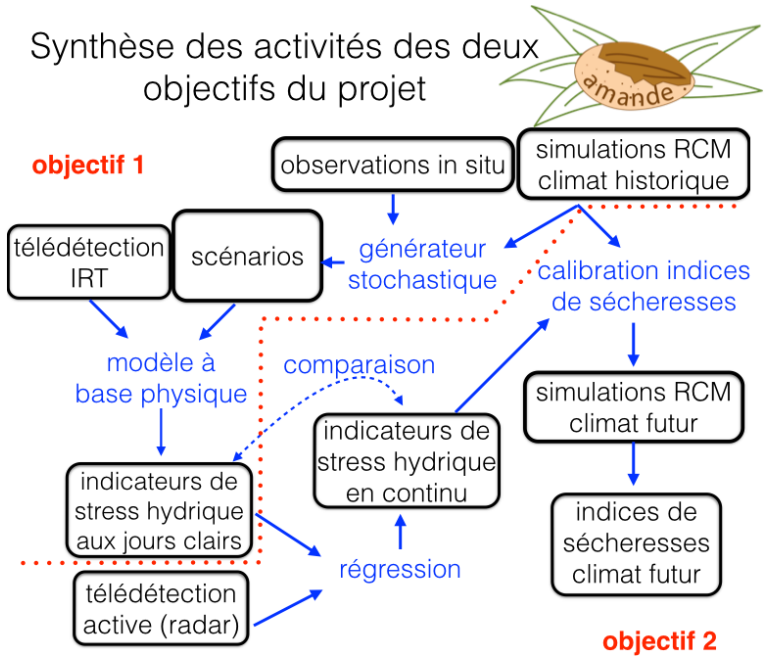


Figure 1.1 – *Résumé des activités du projet AMANDE : **A**pproches stochastiques et se**M**i-p**A**ramétriques combinées à la télédétectio**N** pour l'étude du stress hy**D**riqu**E**.

Projet FRAISE, France, co-porteur, $\sim$ 4 000 €

Le projet FRAISE (**F**o**R**ç**A**ges de précipitations par simulation stochastique pour études d'**I**mpacts hydrologiques : des périodes **S**èches aux événements **E**xtrêmes) fait suite au projet CERISE (simulation de scenarii intégrant des **C**hamps **E**xt**R**êmes spatio-temporels avec éventuelle indépendance asymptotique pour des études d'**I**mpact en **S**cience de l'**E**nvironnement), tous deux financés par le programme LEFE MANU de l'INSU / CNRS, le premier sur 2016-2018 et le deuxième sur 2019-2021. Étant déjà fortement impliquée dans CERISE, je suis co-porteur de FRAISE avec Gwladys Toulemonde de l'IMAG en tant que porteur.

L'objectif principal du projet, dont les activités sont résumées dans la figure 1.2, est de concevoir et développer des scénarios de variables hydrométéorologiques qui prennent en compte des événements extrêmes et démontrer leur intérêt en termes d'impacts hydrologiques. Il se décline en trois sous-objectifs spécifiques et questions de recherche associées :

1. Génération de scénarios de forçages de précipitations ;
   (a) Comment simuler des champs spatio-temporels extrêmes à l'échelle de l'événement, intégrant éventuellement l'indépendance asymptotique, anisotropes ?
   (b) Comment modéliser les transitions dans l'espace et dans le temps entre l'absence de pluie, les événements pluvieux communs et extrêmes ?
   (c) Comment caractériser et reproduire les structures de dépendance spatio-temporelle de façon non-paramétrique, en particulier, quels apports possibles de l'apprentissage statistique et des réseaux de neurones profonds ?
   (d) Comment définir des scénarios types qui varient en termes de motifs spatio-temporels (i.e. en termes de structures de dépendance) afin d'en étudier la non-stationnarité ?
   (e) Comment reproduire la non-stationnarité temporelle dans les intensités (i.e. les lois marginales univariées) mais aussi dans la structure de dépendance ?
2. Adaptation, développement et évaluation de techniques de descente d'échelle pour la simulation des écoulements en milieu urbain ;
   (a) Comment modéliser les distributions des hauteurs d'eau et des vitesses d'écoulement sur un maillage à haute résolution conditionnellement aux valeurs moyennes de ces mêmes variables sur un maillage à basse résolution ?
   (b) Peut-on quantifier de façon fiable le risque d'inondation en appliquant une méthode de changement d'échelle aux simulations de modèles d'écoulement à basse résolution ?
3. Évaluation de l'apport des scénarios générés dans des études d'impacts hydrologiques :
   (a) Quelles mesures de risque spatiales considérer pour quantifier l'effet des forçages extrêmes en fonction des différents objectifs des applications ?
   (b) En quoi les scénarios intégrant des événements extrêmes améliorent l'estimation du risque ?

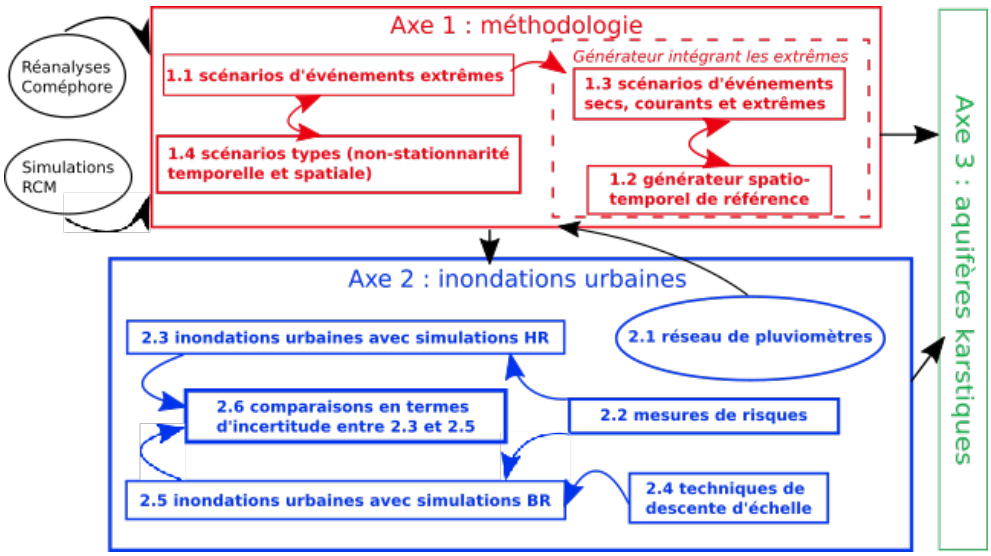


Figure 1.2 – *Résumé des activités du projet FRAISE : **FoRçAges** de précipitations par simulation stochastique pour études d'**I**mpacts hydrologiques : des périodes **S**èches aux événements **E**xtrêmes.*

Chapitre 2

Analyse des travaux scientifiques

Après avoir obtenu un bagage en mathématiques fondamentales avec un B.Sc., j'ai pris une tangente plus financière et informatique en maîtrise pour finalement, m'immerger complètement en informatique en complétant un doctorat en apprentissage statistique. J'ai bénéficié d'un cadre stimulant à l'université de Montréal sous la supervision de Yoshua Bengio. Yoshua m'a donné beaucoup de liberté sur l'élaboration de mon sujet de thèse tout en me fournissant en appui son expertise et son dynamisme. L'objectif de mes travaux de thèse était d'adapter des approches d'apprentissage statistique à la présence de valeurs extrêmes. En effet, l'apprentissage statistique cherche à tirer profit de grandes quantités de données alors que par définition, les valeurs extrêmes sont des événements rares. Il s'agit **du premier thème** que je souhaite présenter concernant mes travaux scientifiques : **des développements méthodologiques alliant l'apprentissage statistique à la théorie des valeurs extrêmes**, voir la section 2.1.

Le deuxième thème que je souhaite aborder, section 2.2, traite **des premières applications en sciences du climat et de l'environnement** avec lesquelles je me suis familiarisée lors de mon stage postdoctoral de deux ans avec Philippe Naveau au LSCE en banlieue parisienne. En effet, mes travaux de thèse ont été appliquées dans les domaines de la finance et de l'assurance et j'ai pu proposer des adaptations des méthologies issues de ces travaux à deux applications en sciences du climat et de l'environnement - la modélisation pluie-débit et la descente d'échelle statistique (*downscaling* en anglais). Ce faisant, j'ai pu acquérir des connaissances et des compétences liées à ces domaines d'applications ainsi que de me créer un réseau de collègues experts dans ces problématiques et dans les questions statistiques qui en découlent.

Le troisième et dernier thème concerne des travaux, toujours en sciences du climat et de l'environnement, mais **qui ont préludé mon projet de recherche en faisant émerger des questions de recherche en statistiques liées au cadre spatial**, voir la section 2.3. Ces travaux ont pris place après ma prise de poste en tant qu'ingénieur de recherche à l'IRD à Montpellier, au laboratoire HydroSciences Montpellier. La première question aborde la problématique de l'interpolation spatiale des précipitations extrêmes, i.e. des niveaux de retour ayant des périodes de retour élevées. J'ai pu proposer des développements méthodologiques qui, comme dans mes travaux de thèse, allient les atouts d'approches non-paramétriques et de la théorie des valeurs extrêmes. La deuxième question présentée dans ce troisième thème s'intéresse à la dépendance inter-sites des précipitations intenses, donc dans un cadre multivarié mais où chaque élément représente un site dans une région donnée. Cela m'a permis de faire une première étude des dépendances des pluies intenses dans l'espace.

2.1 Apprentissage statistique et théorie des valeurs extrêmes

Les questions scientifiques motivant les travaux décrits dans ce premier thème sont par essence méthodologiques. Les applications en finance et en assurance viennent illustrer ces travaux et sont dans la continuité de mon passage par le domaine financier lors de la maîtrise et de l'année pendant laquelle j'ai travaillé en tant qu'analyste à la trésorerie d'une banque.

Les approches développées en apprentissage statistique sont, généralement, non-paramétriques, i.e. elles font peu d'hypothèses sur la structure sous-jacente aux données car cette structure est apprise grâce aux données. Par ailleurs, les méthodes proposées dans le cadre de la théorie des valeurs extrêmes sont conçues de façon à permettre d'extrapoler de manière rigoureuse au-delà de la plage des valeurs observées. L'extrapolation est possible grâce à des modèles paramétriques qui sont issus de comportements asymptotiques. Je présente d'abord, au § 2.1.1, la loi Pareto hybride que j'ai proposé d'utiliser dans un mélange de distributions pour faire l'estimation de densité univariée non-paramétrique en présence de valeurs extrêmes. Puis, au § 2.1.2, je décris comment ce mélange de lois Pareto hybrides peut être couplé avec un réseau de neurones artificiel pour estimer la densité conditionnelle, i.e. en présence de covariables.

2.1.1 Des valeurs centrales aux valeurs extrêmes

Je résume ici les travaux présentés dans Carreau & Bengio (2009a) dont un exemplaire est disponible dans l'annexe A.1.1. Il s'agit d'une partie des travaux dans le cadre de mon Ph.D. En finance, l'estimation des profits et des pertes d'une porte-feuille est centrale pour la gestion du risque. Les praticiens utilisent le plus souvent la Valeur-à-Risque (VaR) qui est en fait un quantile élevé de la distribution des pertes et profits. L'information apportée par la VaR est une borne sur une perte possible avec une certaine probabilité mais elle n'indique pas l'ampleur des pertes possibles au-delà de cette borne. Une mesure de risque qui contient cette information est la VaR conditionnelle qui mesure la perte attendue étant donné que la VaR est dépassée. Pour estimer la VaR conditionnelle, il faut avoir recours à un estimateur de densité qui puisse s'adapter aux valeurs extrêmes.

L'objectif des travaux dans Carreau & Bengio (2009a) était de présenter un estimateur de densité non-paramétrique qui peut prendre en compte l'asymmétrie, la multi-modalité et les ailes lourdes de la densité sous-jacente. Ces caractéristiques de la distribution peuvent se retrouver dans différents domaines d'application tels que la finance et l'assurance mais aussi l'hydrologie et le climat. On entend, par non-paramétrique, un estimateur dont la complexité, souvent décrite par le nombre de paramètres, peut augmenter avec la taille du jeu de données d'apprentissage. Les méthodes issues de la théorie des valeurs extrêmes permettent d'extrapoler vers les valeurs fortes (ou faibles) de la plupart des distributions. Ces méthodes s'appuient sur le comportement asymptotique de la distribution sous-jacente. Or, le niveau auquel le comportement asymptotique entre en jeu dépend de la distribution sous-jacente et il ne sera pas une approximation valide si on s'intéresse pas à une plage de valeurs suffisamment élevées. Par ailleurs, l'estimation de la distribution complète, pas seulement la partie des valeurs fortes, est souvent utile. L'estimateur de densité non-paramétrique que nous avons proposé fournit une approximation de la densité pour toute la plage de valeurs, que le comportement asymptotique soit présent ou non.

Les modèles de mélanges de Gaussiennes sont des estimateurs non-paramétriques souvent utilisés car ils permettent une estimation flexible de la densité faisant peu d'hypothèses. Cependant, les mélanges de Gaussiennes peuvent être en difficulté en présence de valeurs extrêmes, en particulier si l'ensemble d'apprentissage est petit. La loi de Pareto hybride est une extension continue à l'ensemble des valeurs réelles de la loi de Pareto généralisée, adaptée uniquement pour la partie supérieure de la distribution,

obtenue en la juxtaposant à une loi Gaussienne. La Pareto hybride peut être utilisée dans un mélange. Le mélange de Pareto hybrides hérite des propriétés non-paramétriques pour la partie centrale de la distribution et des propriétés asymptotiques pour les valeurs fortes.

Mélange de Gaussiennes

Je m'intéresse ici à l'estimation de densité dans un cadre univarié (Bishop, 2006). Plus précisément, soit Y une variable aléatoire qui prend ses valeurs dans $\mathbb{R}$, et soit $p(\cdot)$, la densité de Y qui doit être une fonction positive, $p : \mathbb{R} \to \mathbb{R}^+$, et qui intègre à 1, $\int p(y)dy = 1$. Un estimateur de $p(\cdot)$ est une fonction $\hat{p}(\cdot;\theta)$ de paramètres θ satisfaisant les mêmes propriétés. La procédure d'inférence ou d'apprentissage consiste à identifier des valeurs de paramètres $\hat{\theta}$. Une stratégie courante cherche à déterminer $\hat{\theta}$ de façon à maximiser un critère d'adéquation entre $\hat{p}(\cdot;\hat{\theta})$ et $\{y_1, \ldots, y_n\}$, un jeu de données observées issues de $p(\cdot)$. Un critère d'adéquation souvent utilisé est la log-vraisemblance $\mathcal{L}(\theta; \{y_1, \ldots, y_n\})$ qui repose sur la probabilité que le modèle attribue aux données observées :

$$\mathcal{L}(\theta; \{y_1, \ldots, y_n\}) = \sum_{i=1}^{n} \log\left(\hat{p}(y_i;\theta)\right). \tag{2.1}$$

Un estimateur de densité non-paramétrique classique est le mélange de Gaussiennes (Bishop, 2006). Il s'agit d'une somme pondérée de densités de la loi Gaussienne (aussi appelée loi Normale) qui sont les composantes du mélange :

$$\hat{p}(y;\theta) = \sum_{j=1}^{m} \pi_j f(y;\mu_j,\sigma_j), \tag{2.2}$$

où $f(\cdot;\mu_j,\sigma_j)$ est la densité d'une loi Normale de paramètres μ_j et σ_j et $0 \leq \pi_j \leq 1$ sont les poids du mélange tels que $\sum_{j=1}^{m} \pi_j = 1$. Les paramètres du mélange de Gaussiennes sont $\theta = (\pi_1, \ldots, \pi_m, \mu_1, \ldots, \mu_m, \sigma_1, \ldots, \sigma_m)$, soit un vecteur de longueur $3m$, avec $3m-1$ paramètres libres étant donnée la contrainte sur les poids du mélange.

Un mélange de Gaussiennes est un estimateur non-paramétrique dans la mesure où le nombre de composantes m est choisi en fonction du jeu de données et peut augmenter avec la taille de celui-ci. Autrement dit, plus un jeu de données est important au sens de l'information qu'il contient, plus on pourra se permettre de prendre un nombre de composantes élevées et en conséquence, plus grand sera le nombre de paramètres du mélange. Dans ces conditions, le mélange de Gaussiennes est en mesure d'approximer toute densité continue, pourvu que le jeu de données soit assez informatif.

L'apprentissage d'un mélange de Gaussiennes se fait généralement en maximisant la log-vraisemblance avec une procédure itérative en deux étapes appelées *Expectation-Maximisation* (EM). Cette procédure contourne le recours à l'optimisation numérique nécessaire pour maximiser directement la log-vraisemblance car la maximisation dans l'étape *M* a une solution analytique dans le cas du mélange de Gaussiennes. La figure 2.1 illustre la densité d'un mélange de Gaussiennes.

Théorie des valeurs extrêmes

Un des objets principaux d'étude de la théorie des valeurs extrêmes est le comportement statistique du maximum d'une variable aléatoire Y (Coles, 2001). Soit $Y_1, \ldots, Y_n$ une suite indépendante de

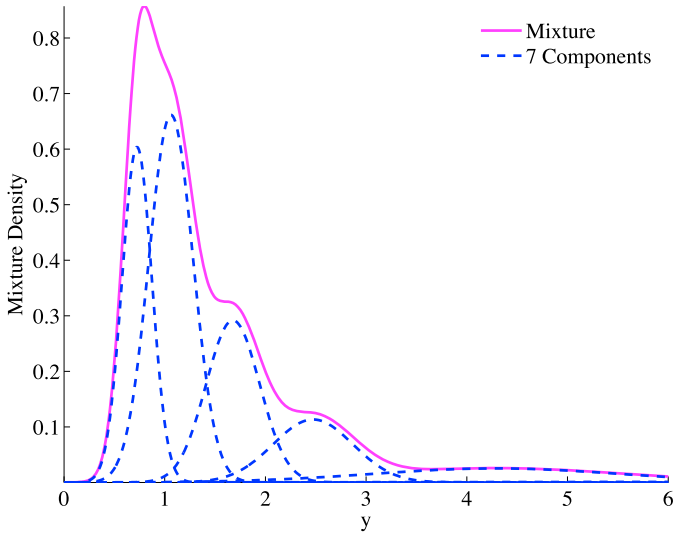


Figure 1. Gaussian mixture density (solid line) with seven components trained on heavy-tailed data. The dashed lines represent the contribution of each component to the density. Five components model the central part, and the other two components contribute to the density in the upper tail.

Figure 2.1 – *Tirée de Carreau* et al. *(2009).*

copies de Y et soit leur maximum :

$$M_n = \max\{Y_1, \ldots, Y_n\}.$$

Le théorème des types extrêmaux stipule que si la loi de M_n, lorsque n devient grand, tend vers une loi non-dégénérée, alors cette loi est dans la famille de l'une des lois des valeurs extrêmes. Il existe trois lois des valeurs extrêmes ; chacune correspondant à un comportement différent du maximum. La loi de Fréchet est dite à queue lourde car la densité décroît avec une vitesse polynomiale pour les grandes valeurs. Le loi de Gumbel est dite à queue légère car la vitesse de décroissance de la densité est exponentielle. Enfin, la loi de Weibull a une densité bornée supérieurement.

Une autre approche pour l'étude des valeurs extrêmes se focalise sur le comportement statistique des dépassements d'une variable aléatoire Y. Soit u un seuil suffisamment élevé et supposons que le maximum M_n converge vers l'une des lois des valeurs extrêmes, alors la loi des dépassements $Y - u|Y > u$ peut être approximée par la loi de Pareto généralisée. Les paramètres de cette loi sont reliées à ceux de la loi des valeurs extrêmes vers laquelle converge M_n et en particulier, le type de décroissance de la densité - polynomiale, exponentielle ou bornée - sera le même.

Loi Pareto hybride

La théorie des valeurs extrêmes établit donc que la loi de Pareto généralisée peut approximer la queue supérieure de presque toute distribution continue, pourvu qu'un seuil suffisamment élevé soit fixé. Plusieurs propositions ont été avancées pour le choix du seuil bien que cela reste un exercice délicat. Par ailleurs, si l'on s'intéresse à l'ensemble de la distribution, pas seulement à la queue supérieure, comment modéliser la partie de la distribution sous le seuil tout en tirant profit de la Pareto généralisée au-delà du seuil ?

Pour répondre à cette question, nous avons développé la loi de Pareto hybride qui juxtapose la densité

d'une loi Gaussienne avec la densité de la loi de Pareto généralisée, voir la figure 2.2. La jonction entre ces deux densités correspond au seuil de la loi de Pareto généralisée. Pour faciliter l'apprentissage avec des méthodes de descente de gradient, deux contraintes de continuité au niveau du point de jonction - continuité de la densité et de sa dérivée - sont imposées. Initialement, il y avait cinq paramètres pour la loi de Pareto hybride - deux paramètres provenant de la loi Gaussienne, deux paramètres provenant de la loi de Pareto généralisée et le seuil. Avec la prise en compte des contraintes de continuité, il reste trois paramètres libres - μ et σ, les deux paramètres de la loi Gaussienne et ξ, le paramètre de forme de la loi de Pareto généralisée qui caractérise le comportement de la queue de la distribution, i.e. le type de décroissance de la densité.

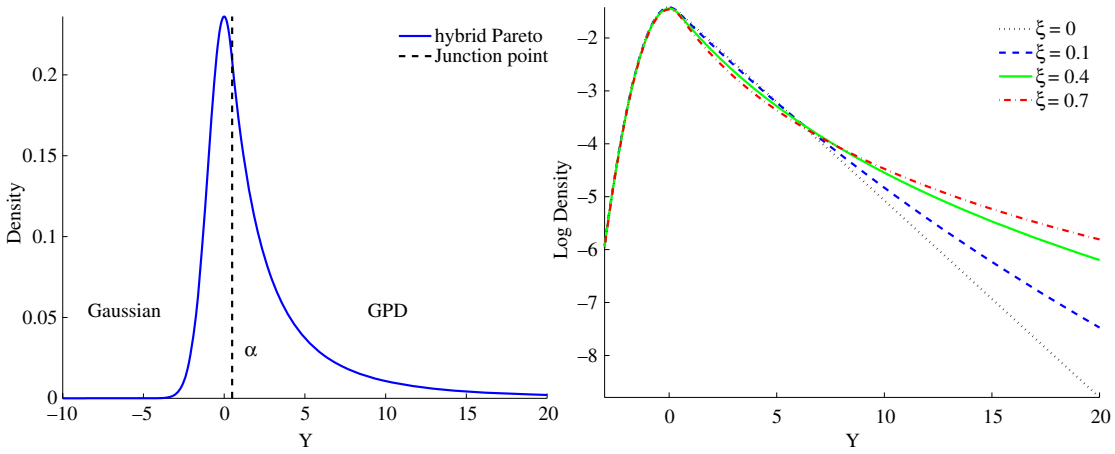


Fig. 1 *Left panel* hybrid Pareto density with parameters $\xi = 0.4$, $\mu = 0$ and $\sigma = 1$. *Right panel* hybrid Pareto log-density for various tail parameters and in all cases $\mu = 0$ and $\sigma = 1$.

Figure 2.2 – *Tirée de Carreau & Bengio (2009a). GPD = generalized Pareto distribution.*

Apprentissage

La loi de Pareto hybride contourne la question de la sélection du seuil de la loi de Pareto généralisée car celui-ci est exprimé comme une fonction de trois paramètres libres de la loi de Pareto hybride. Afin de permettre une approche non-paramétrique plus flexible pour la partie centrale de la distribution, nous avons proposé d'utiliser la loi de Pareto hybride dans un mélange de distributions, i.e. en tant que composantes à la place de la densité de la loi Gaussienne dans l'équation (2.2). La partie centrale de l'estimateur est donc le conventionnel mélange de Gaussiennes. Pour la queue supérieure, il s'agit d'un mélange de lois de Pareto généralisée dont la composante avec la queue la plus lourde qui domine et détermine le comportement de la queue supérieure de la distribution.

Les paramètres du mélange de lois de Pareto hybride avec m composantes sont $\theta = (\pi_1, \ldots, \pi_m, \mu_1, \ldots, \mu_m, \sigma_1, \ldots, \sigma_m, \xi_1, \ldots, \xi_m)$, soit un vecteur de longueur $4m$, avec $4m - 1$ paramètres libres en raison de la contrainte sur les π_j. Ces paramètres sont appris en maximisant la log-vraisemblance telle qu'exprimée dans l'équation (2.1). Dans ce cas, l'algorithme *EM* ne présente pas d'intérêt car il n'y a pas de solution analytique de la maximisation de l'étape *M*. La log-vraisemblance est donc maximisée directement en ayant pris soin de trouver des valeurs initiales raisonnables pour les paramètres du mélange. Ces valeurs initiales sont obtenues en appliquant un algorithme de classification non-supervisée robuste, les K-Médianes, et en estimant les paramètres de la loi de Pareto hybride sur chacune des m classes identifiées. Un algorithme de descente de gradient de type gradients conjugués est utilisé et le gradient de $\mathcal{L}(\theta) = \mathcal{L}(\theta; \{y_1, \ldots, y_n\})$, la log-vraisemblance l'équation (2.1), est calculé analytiquement :

$$\left(\frac{\partial}{\partial \pi_1} \mathcal{L}(\theta), \ldots, \frac{\partial}{\partial \mu_1} \mathcal{L}(\theta), \ldots, \frac{\partial}{\partial \sigma_1} \mathcal{L}(\theta), \ldots, \frac{\partial}{\partial \xi_1} \mathcal{L}(\theta), \ldots, \right) \tag{2.3}$$

Principaux résultats

Nous avons montré à l'aide de simulations de jeux de données provenant de la loi de Pareto hybride que les estimateurs par maximum de vraisemblance de la loi Pareto hybride convergent vers les valeurs des paramètres ayant généré les données lorsque la taille de l'échantillon augmente.

Nous avons ensuite comparé le mélange de lois de Pareto hybrides à plusieurs autres estimateurs sur des données synthétiques provenant d'une loi de Fréchet (asymmétrique et à queue lourde). Les autres estimateurs considérés sont : la loi de Pareto généralisée avec sélection du seuil, un mélange de Gaussiennes, un mélange de lois Log-Normales et l'estimateur de la fenêtre de Parzen (un estimateur non-paramétrique classique). Dans tous les cas où cela est nécessaire, les hyper-paramètres (le nombre de composante m pour les mélanges et la largeur de la fenêtre de Parzen) sont choisis sur un ensemble de validation. Les conclusions générales sont que le mélange de lois de Pareto hybrides a une perfomance supérieure aux autres estimateurs en particulier pour les petits jeux de données et en présence de queue de distribution très lourde. L'estimateur le plus proche en performance est le mélange de lois Log-Normales. La figure 2.3 illustre ces comparaisons entre estimateurs.

Enfin, le mélange de lois de Pareto hybrides et les autres estimateurs considérés ont été appliqués à des données d'assurance. La performance du mélange de lois hybrides était supérieure aux autres types de mélanges et à l'estimateur de la fenêtre de Parzen qui réagit particulièrement mal en présence de valeurs extrêmes. Cependant, le mélange de lois Paretos hybrides tend à trouver des queues de distribution plus lourdes que l'approche classique avec la loi de Pareto généralisée après sélection d'un seuil.

Perspectives

Les inconvénients principaux du mélange de Pareto hybrides sont :

- la présence de densité dans l'aile inférieure de la distribution donnant des probabilités généralement faibles mais non nulles à des plages de valeurs négatives, ce qui est irréaliste dans de nombreuses applications ;
- le recours à plusieurs lois de Pareto pour modéliser l'aile supérieure de la distribution alors qu'en théorie, une seule est nécessaire, en particulier alors que le paramètre de forme est difficile à estimer ;
- les contraintes de continuité au point de jonction ne sont pas forcément cohérentes avec les propriétés asymptotiques, bien que ces contraintes soient relaxées dans le mélange.

De nombreuses autres propositions ont émergées depuis pour modéliser l'ensemble de la distribution tout en employant la loi de Pareto généralisée pour les valeurs fortes, voir par exemple Li *et al.* (2012) et Naveau *et al.* (2016).

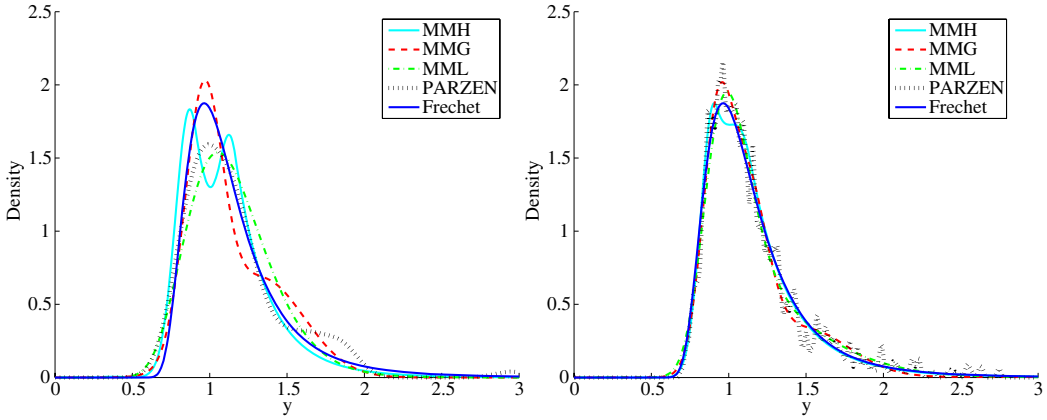


Fig. 4 Estimated density in the central part (99% of training points) for the Fréchet data with $\xi = 1/5$. *Left panel* 100 training points and *right panel*, 1,000 training points.

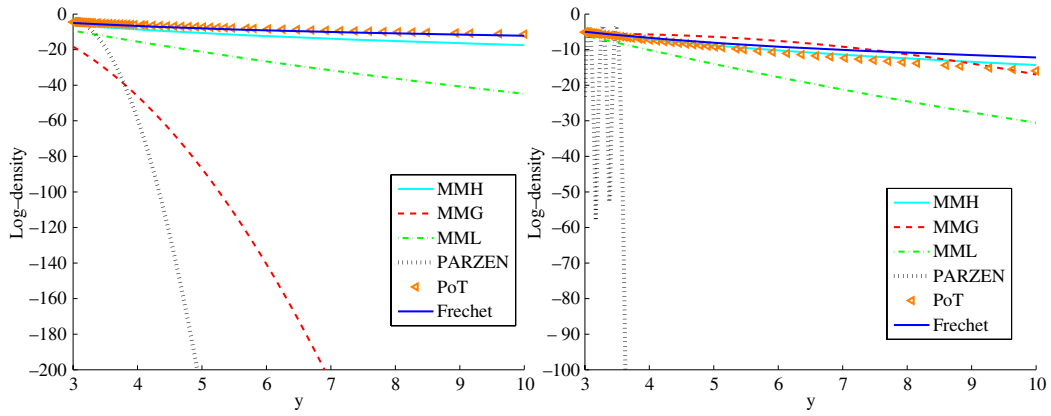


Fig. 5 Estimated log-density in the upper tail (<1% of training points) for the Fréchet data with $\xi = 1/5$. *Left panel* 100 training points and *right panel*, 1,000 training points.

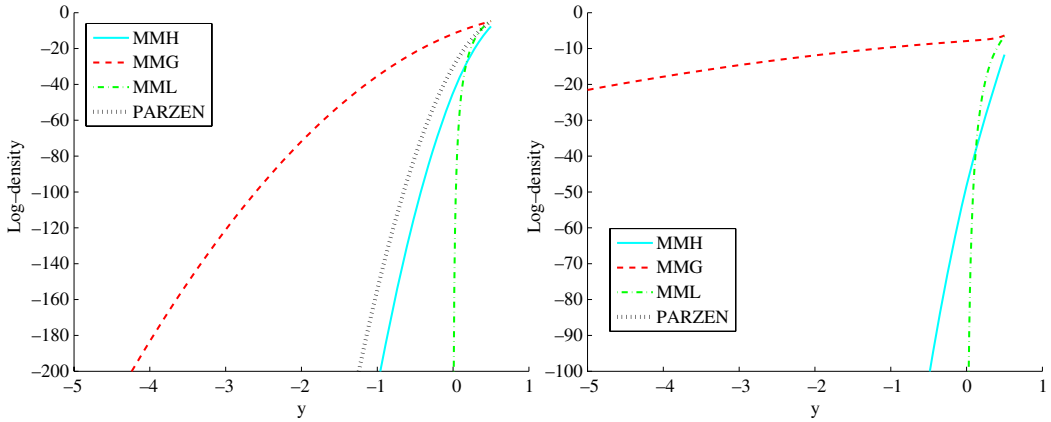


Fig. 6 Estimated log-density in the lower tail (no training points) for the Fréchet data with $\xi = 1/5$. *Left panel* 100 training points and *right panel*, 1,000 training points.

Figure 2.3 – *Tirées de Carreau & Bengio (2009a). MMH (MMG, MML) = mixture model with hybrid Pareto (Gaussian, Log-Normal) components. PoT = peaks-over-threshold (i.e. la loi de Pareto généralisée avec sélection du seuil). PARZEN = parzen window (estimateur non-paramétrique classique).*

2.1.2 Modélisation conditionnelle à l'aide de réseau de neurones

Je résume ici les travaux présentés dans Carreau & Bengio (2009b) dont un exemplaire est disponible dans l'annexe A.1.2. Il s'agit d'une partie des travaux dans le cadre de mon Ph.D. Une des applications visées par ces travaux était le calcul des primes d'assurance pour un client donné. Ce calcul se base sur la distribution conditionnelle des réclamations étant donné le profil du client. Pour l'assurance automobile, le profil du client contient de l'information sur le conducteur, sur la voiture et sur les options sélectionnées par l'assuré dans le contrat d'assurance. Typiquement, des valeurs extrêmes peuvent être observées dans la distribution des réclamations, selon le profil du client.

Notons $\boldsymbol{X}$, un vecteur aléatoire prenant ses valeurs dans $\mathbb{R}^d$ qui contient de l'information pertinente, par exemple le profil du client, sur Y, la variable aléatoire représentant les réclamations. Alors les compagnies d'assurance s'intéressent à l'estimation de la densité conditionnelle de $Y|\boldsymbol{X} = \boldsymbol{x}$. Une façon d'y parvenir est de définir un modèle $\hat{p}(y;\theta)$ pour la densité de Y et de relier les paramètres θ à $\boldsymbol{X}$ à l'aide de fonctions $\theta(\boldsymbol{x};\omega)$ de paramètres ω. Le modèle pour la densité conditionnelle est donc

$$p(y|\boldsymbol{X} = \boldsymbol{x}) = \hat{p}(y;\theta(\boldsymbol{x};\omega)). \tag{2.4}$$

L'apprentissage repose sur la maximisation de la log-vraisemblance conditionnelle à l'aide d'un jeu de données de paires $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$:

$$\mathcal{L}(\omega; \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}) = \sum_{i=1}^{n} \log\left(\hat{p}(y_i;\theta(\boldsymbol{x}_i;\omega))\right). \tag{2.5}$$

Réseaux de neurones à une couche cachée

La stratégie que nous avons suivie consiste à connecter un réseau de neurones à une couche cachée de type *feed-forward* pour approximer les fonctions $\theta(\boldsymbol{x};\omega)$ dans l'équation (2.4) dans le cas où le modèle de densité est un mélange, voir l'équation (2.2). Les paramètres ω représentent donc les poids du réseau de neurones. Les réseaux de neurones à une couche cachée de type *feed-forward* ont la capacité d'approximer arbitrairement bien toute fonction continue à condition que le jeu de données servant à l'apprentissage soit suffisamment riche (Bishop, 2006).

Voici l'architecture générale de ce type de réseaux de neurones (voir la figure 2.4 pour le cas du mélange conditionnel de Pareto hybrides). Soit $\boldsymbol{x} = (x_1, \ldots, x_d)$ le vecteur dit d'entrées du réseau de neurones. Chaque neurone h de la couche cachée, avec $1 \leq h \leq H$, transforme non-linéairement, grâce à la tangente hyperbolique, une combinaison linéaire des entrées :

$$z_h = z(\boldsymbol{x};\boldsymbol{v}_h) = \tanh\left(\sum_{i=1}^{d} v_{h,i}x_i + v_{h,0}\right), \tag{2.6}$$

où $\boldsymbol{v}_h = (v_{h,1}, \ldots, v_{h,d})$ sont les poids reliant les entrées aux neurones cachés. Les neurones de la couche de sortie sont associés aux paramètres de la densité conditionnelle $\theta(\boldsymbol{x};\omega)$. De façon semblable à la couche cachée, une combinaison linéaire des z_h, résultant des calculs des neurones de la couche cachée dans l'équation (2.6), est affectée à chaque neurone de la couche de sortie et transformée avec

une fonction $g(\cdot)$ de sorte à satisfaire des contraintes sur les plages de valeurs du paramètre $\theta_j(\boldsymbol{x};\omega)$:

$$\theta_j(\boldsymbol{x};\omega) = g\left(\underbrace{\sum_{h=1}^{H} w_{j,h} z_h}_{\text{non-linéaire}} + \underbrace{\sum_{i=1}^{d} \tilde{v}_{j,i} x_i}_{\text{linéaire}} + w_{j,0}\right), \tag{2.7}$$

où $\boldsymbol{w}_j = (w_{j,1}, \ldots, w_{j,H})$ sont les poids reliant les neurones cachés au $j^{\text{ème}}$ paramètre de la densité conditionnelle et $\tilde{\boldsymbol{v}}_j = (\tilde{v}_{j,1}, \ldots, \tilde{v}_{j,H})$ sont les poids d'une connexion linéaire supplémentaire de sorte que la régression linéaire est un cas particulier du réseau de neurones lorsque $H = 0$. Lorsqu'il n'y a pas de contraintes sur le paramètre $\theta_j(\boldsymbol{x};\omega)$, $g(a) = a$ est l'identité. Pour imposer des valeurs positives, on peut utiliser la *softplus*, $g(a) = \log(1 + e^a)$, et la transformation dite *softmax*, $g(a_1, \ldots, a_p) = (e^{a_1}, \ldots e^{a_p}) / \sum_{k=1}^{p} e^{a_k}$, permet d'encoder des probabilités.

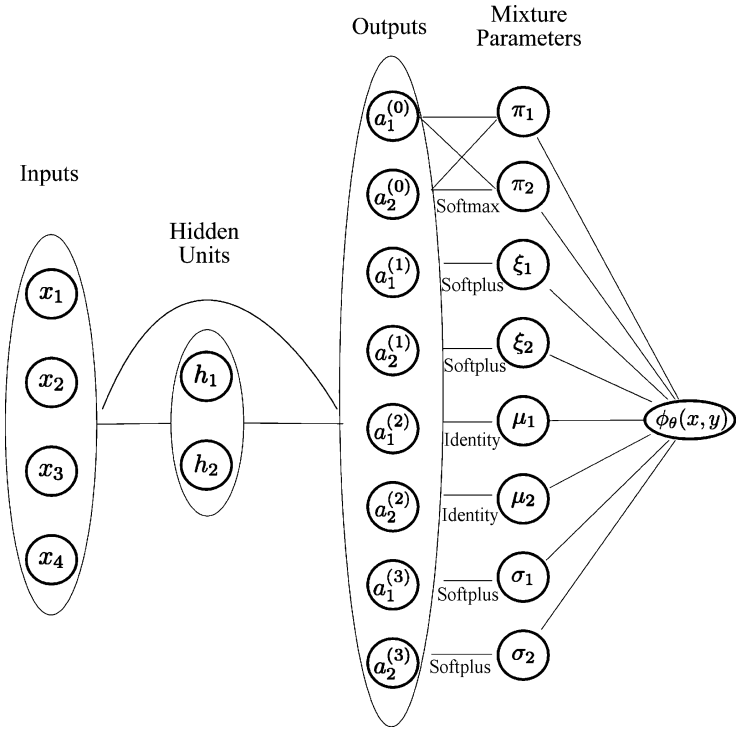


Fig. 4. Conditional mixture model. A feedforward neural network with one hidden layer and hyperbolic tangent activation function is used to predict input dependent mixture parameters. Appropriate transfer functions at the network outputs are used to impose range constraints [see (10)].

Figure 2.4 – *Tirée de Carreau & Bengio (2009b).*

Adaptation du réseau de neurones et apprentissage

La contribution méthodologique principale dans Carreau & Bengio (2009b) est la proposition d'un mélange conditionnel de Pareto hybrides, i.e. un mélange de Pareto hybrides dont les paramètres sont des fonctions de $\boldsymbol{x}$ telles qu'implementées par le réseau de neurones décrit ci-haut. L'adaptation du réseau de neurones pour que les fonctions $\theta_j(\boldsymbol{x};\omega)$ calculées en sortie puissent être interprétées comme les paramètres d'un mélange de lois de Pareto hybrides se fait en deux étapes (voir la figure 2.4).

La première étape consiste à modifier la phase avant (*forward* en anglais), pour que le nombre de neurones de la couche de sortie soit de $4m - 1$ et les fonctions de transformation $g(\cdot)$ soient appropriées. La deuxième étape concerne la phase arrière (*backward* en anglais) pour le calcul du gradient de la log-vraisemblance de l'équation (2.5) en fonction des paramètres $\omega = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_H, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{4m-1}, \tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_{4m-1})$. Pour cela, il faut brancher le gradient de l'équation (2.3) au début de la rétro-propagation du gradient du réseau de neurones décrit par les équations (2.6)-(2.7).

Les valeurs initiales du mélange inconditionnel de lois de Pareto hybrides servent à initialiser les valeurs des paramètres $w_{j,0}$ dans l'équation (2.7) qui sont indépendants des entrées $\boldsymbol{x}$. Les autres valeurs initiales sont choisies aléatoirement de façon uniforme dans certaines plages de valeurs de sorte à ce que l'apprentissage se passe bien (i.e. en tenant compte du nombre total de connexions d'une couche à l'autre et en veillant à ne pas saturer les tangentes hyperboliques dans les calculs de la couche cachée, voir l'équation (2.6)).

Principaux résultats

J'ai comparé plusieurs estimateurs au mélange conditionnel de Pareto hybrides : un mélange conditionnel de Gaussiennes, un mélange conditionnel de Log-Normales ainsi qu'un approche type fenêtre de Parzen avec un double noyau. J'ai également implementé une variante conditionnelle de l'approche classique en théorie des valeurs extrêmes où la loi de Pareto est ajustée à des résidus d'une régression estimée par un réseau de neurones.

Tous les estimateurs considérés ont d'abord été comparés sur des jeux de données synthétiques provenant d'une loi de Fréchet ayant une queue de distribution plus ou moins lourde, i.e. à décroissance plus ou moins rapide. Les paramètres de la loi dépendent soit linéairement soit non-linéairement (via un sinus) d'une variable $X \in \mathbb{R}$, voir la figure 2.5. Les hyper-paramètres (le nombre de composantes dans un mélange, le nombre de neurones cachés et les largeurs de fenêtres des noyaux) sont choisis sur des ensembles de validation distincts des ensembles d'apprentissage. Tout comme dans le cas inconditionnal décrit au § 2.1.1, le mélange conditionnel de Pareto hybrides surpasse les autres estimateurs dans toutes les configurations de jeux synthétiques considérées, voir la figure 2.6 pour une illustration des comparaisons. L'estimateur le plus proche en performance est le mélange conditionnel de Log-Normales.

Les estimateurs de densité conditionnelle ont ensuite été comparés sur deux jeux de données réels, l'un provenant du milieu de l'assurance et l'autre de la compétition KDD Cup 98 organisée lors d'un colloque (Fourth International Conference on Knowledge Discovery and Data Mining). Dans ces deux cas aussi, la performance du mélange conditionnel de Pareto hybrides est supérieure à celle des autres estimateurs considérés.

Le code développé pour l'estimation de densité inconditionnelle et conditionnelle à base de mélanges, incluant les mélanges de Pareto hybrides, a été rendu librement accessible sous forme d'une librairie `R` appelée `condmixt` sur le dépôt officiel, voir https://CRAN.R-project.org/package=condmixt.

Perspectives

Les travaux présentés dans cette section proposent une façon de combiner des méthodes non-paramétriques, les mélanges de distributions et les réseaux de neurones, avec des approches paramétriques issus de la théorie des valeurs extrêmes. Avec les développements récents, se pose maintenant la question des apports potentiels des réseaux de neurones profonds à la modélisation des valeurs extrêmes (Goodfellow *et al.* , 2016).

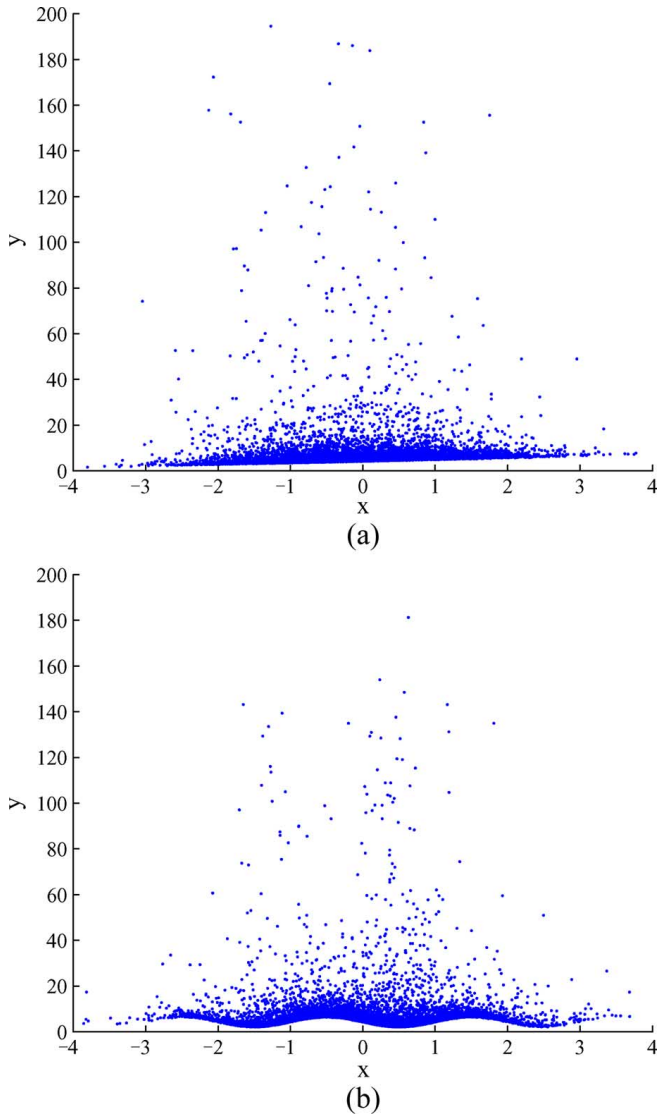


Fig. 5. Conditional Fréchet generated data with very heavy-tail index. (a) Linearly dependent parameters. (b) Sinusoidal dependent parameters. The y-axis is shortened so that the shape of the dependence can be seen ($\max y \approx 10^6$).

Figure 2.5 – *Tirée de Carreau & Bengio (2009b).*

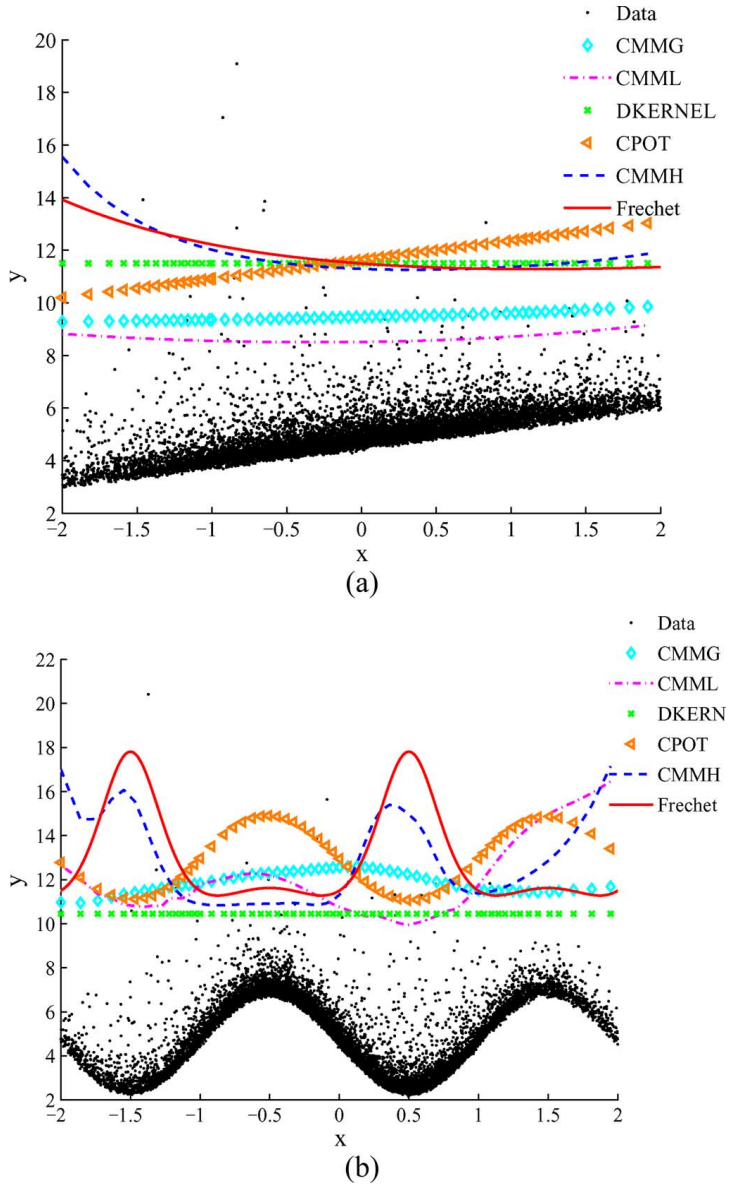


Fig. 11. Average estimated conditional quantiles from trained models on 2000 points generated from a conditional Fréchet with moderately heavy tail. The dependence functional is (a) linear and (b) sine-shaped. Quantile level is 0.999. Test data are also illustrated. CMMH recovers fairly well the shape of the conditional quantile whereas the other estimators fail to do so.

Figure 2.6 – *Tirée de Carreau & Bengio (2009b). CMMH (CMMG, CMML) = conditional mixture model with hybrid Pareto (Gaussian, Log-Normal) components. CPoT = conditional peaks-over-threshold (i.e. la loi de Pareto généralisée avec seuil estimé par régression). DKERN = double kernel estimator (estimateur non-paramétrique classique).*

2.2 Applications en sciences du climat et de l'environnement

À la différence du thème 1 en section 2.1, dans ce thème, les applications sont le moteur des questions scientifiques. Comment adapter les méthodologies du thème 1 à ces applications pour fournir des solutions réalistes et pertinentes ? Pour cela, il m'a fallu acquérir des connaissances sur ces applications en sciences du climat et de l'environnement et consolider mes connaissances théoriques notamment en théorie des valeurs extrêmes.

Les premiers travaux présentés au § 2.2.1 concernent une application en hydrologie pour modéliser la relation pluie-débit. Il s'agit d'adapter le mélange conditionnel de Pareto hybrides pour estimer la distribution du débit à un pas de temps futur étant donné de l'information passée sur les conditions hydrologiques et météorologiques. Puis les travaux abordés au § 2.2.2 proposent une application en climat du mélange conditionnel de Pareto hybrides. La problématique considérée est celle de la descente d'échelle statistique, i.e. comment, à partir d'information à basse résolution spatiale telle que fournie par les modèles de climat globaux, passer à une résolution plus élevée requise pour les études d'impact.

2.2.1 Prévision probabiliste du débit

Je résume ici les travaux présentés dans Carreau *et al.* (2009) dont un exemplaire est disponible dans l'annexe A.2.1. Ces travaux se sont déroulés lors de mon premier postdoctorat en France, en collaboration avec Philippe Naveau, expert en statistiques appliquées au climat au LSCE à Gif-sur-Yvette, et avec Eric Sauquet, hydrologue à IRSTEA Lyon. La modélisation des écoulements fluviaux sert à la planification hydro-électrique, l'irrigation et la prévention des crues. C'est un fait bien connu parmi les hydrologues que la distribution des écoulements a des ailes lourdes, ce qui signifie que de grandes valeurs d'écoulements, très éloignées de la valeur médiane, peuvent se produire. Un autre fait bien connu est que les précipitations dans le bassin hydrographique influence les écoulements fluviaux. Cependant, de nombreux autres mécanismes en jeu, tels que les nappes phréatiques souterraines et la perméabilité des sols, sont spécifiques à un bassin hydrographique donné. La plupart des modèles hydrologiques tentent de reproduire la dynamique du bassin en modélisant les mécanismes en termes de réservoirs. Une approche alternative consiste à utiliser un modèle stochastique fournissant une distribution complète des écoulements.

Nous nous sommes donc intéressés à la distribution du débit Q_t à un pas de temps t de 1, 6, ou 12 heures. En s'inspirant des modèles pluie-débit, e.g. les modèles conceptuels à réservoirs, l'idée était de proposer une alternative probabiliste pour la prévision de Q_t étant donné $\mathcal{F}_{t-1}$, des covariables contenant de l'information sur les conditions hydrologiques et météorologiques disponibles au pas de temps précédent. Dans la notation introduite précédemment, $Y = Q_t$ et $X = \mathcal{F}_{t-1}$ et l'on cherche à modéliser la distribution conditionnelle de $Q_t|\mathcal{F}_{t-1}$.

La distribution du débit se caractérise par de l'asymmétrie, la présence de crues qui sont des événements extrêmes et différents régimes saisonniers qui peuvent entraîner de la multi-modalité. De plus, la relation entre la pluie, une information primordiale sur les conditions hydrologiques, et le débit est hautement non-linéaire. Le mélange conditionnel de Pareto hybrides est en mesure de prendre en compte l'ensemble de ces caractéristiques. Dans ce travail, nous avons déterminé comment adapter cet estimateur et évalué ses apports pour modéliser $Q_t|\mathcal{F}_{t-1}$ notamment pour tenir compte de l'incertitude liée à la prévision.

Pénalité sur les paramètres de forme de la Pareto hybride

La première adaptation de la méthodologie présentée au § 2.1.2 concerne l'ajout d'un terme de pénalité à la log-vraisemblance, équation (2.5), pour introduire de l'information a priori sur les valeurs prises par les paramètres de forme de la loi de Pareto hybride. Cette information a priori repose sur le raisonnement suivant.

Le paramètre de forme de la loi de Pareto généralisée est particulièrement difficile à estimer car il caractérise le comportement des valeurs extrêmes dont, par définition, il y a peu d'occurrences. Dans le mélange de Pareto hybrides, cette difficulté est amplifiée car il y a m paramètres de forme alors que selon la théorie des valeurs extrêmes, un seul est vraiment nécessaire. De plus, dans le mélange conditionnel, les paramètres de forme ont encore plus de flexibilité, étant autorisés à varier avec des covariables, et leur estimation est d'autant plus entachée d'incertitude.

Pour réduire cette incertitude, nous avons conçu une distribution a priori des valeurs prises par les paramètres de forme qui reflète deux idées. La première est que la plupart des paramètres de forme du mélange doivent prendre de faibles valeurs, près de zéro, car ils sont associés à des composantes qui sont dans la partie centrale de la distribution. La deuxième est que quelques uns, voire un seul, des paramètres de forme doivent prendre des valeurs correspondant au comportement des valeurs extrêmes de la distribution sous-jacente que l'on cherche à estimer.

La distribution a priori que nous avons proposée est donc un mélange à deux composantes, dont la première, prépondérante, concerne de faibles valeurs positives et la deuxième est centrée sur la valeur attendue du paramètre de forme de la distribution sous-jacente. En ce qui concerne la distribution du débit, la valeur attendue du paramètre de forme est fixée à 0.5, selon l'information d'experts hydrologues. La figure 2.7 illustre cette distribution a priori.

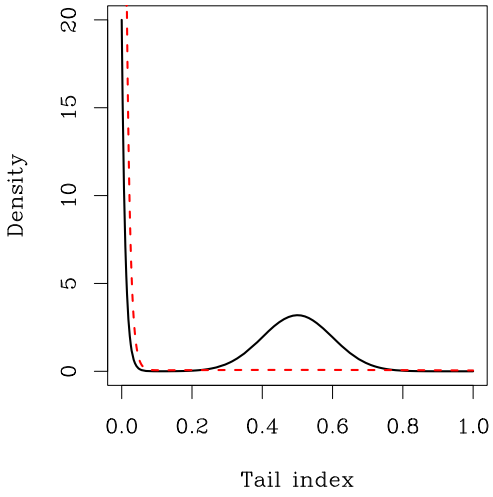


Figure 3. The distribution represented by the solid line has one mode at zero and one mode at 0.5, while the distribution represented by the dashed line has significant density only around zero. The former distribution reflects our prior information about how the tail indexes of a hybrid Pareto mixture should be distributed when the data are heavy tailed, and the latter distribution, reflects the situation when the data are light tailed.

Figure 2.7 – *Tirée de Carreau* et al. *(2009).*

La pénalité proposée dans Carreau *et al.* (2009) s'ajoute à la log-vraisemblance afin de favoriser des paramètres ω qui donnent lieu à des valeurs de paramètres de forme en accord avec la distribution a priori :

$$\mathcal{L}(\omega; \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}) = \sum_{i=1}^{n} \log\left(\hat{p}(y_i; \theta(\boldsymbol{x}_i; \omega))\right) + \frac{\lambda}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \log f(\xi_j(\boldsymbol{x}_i); \tau, \eta, \rho), \quad (2.8)$$

où λ est un hyper-paramètre contrôlant le compromis entre la log-vraisemblance et la pénalité, $\xi_j(\boldsymbol{x}_i)$ est le paramètre de forme de la $j^{\text{ième}}$ composante évaluée en $\boldsymbol{x}_i$ et $f(\cdot; \tau, \eta, \rho)$ est la densité de la distribution a priori des valeurs des paramètres de forme, voir la figure 2.7, avec trois paramètres τ, η et ρ. En plus du nombre de neurones cachés et du nombre de composantes du mélange, il faut choisir, en validation, quatre hyper-paramètres liés à la pénalité (λ, τ, η et ρ).

Covariables liées au conditions hydrologiques et météorologiques

La deuxième adaptation de la méthodologie présentée au § 2.1.2 se focalise sur la définition d'un ensemble de covariables pouvant résumer les conditions hydrologiques et météorologiques au temps $t-1$ et ayant potentiellement un pouvoir prédictif sur Q_t. L'application portait sur le bassin de l'Orgeval, à l'est de Paris, pour lequel, en plus des mesures horaires de débit, des mesures horaires de précipitations sont disponibles en quatre sites sur le bassin et des mesures journalières moyennes de température sont disponibles en un site.

Les conditions météorologiques sont résumées à l'aide des précipitations et de la température. Concernant les précipitations, plutôt que de les agréger a priori en prenant, par exemple, la moyenne spatiale, le choix a été fait de laisser le réseau de neurones apprendre comment agréger les données de précipitations. Il y a donc quatre covariables qui sont les précipitations au pas de temps précédent. Une cinquième covariable est fournie par la température journalière moyenne du jour précédent.

Des effets mémoire sur le débit synthétisent les conditions hydrologiques en cherchant à recréer de la persistence. Tout d'abord, pour des effets mémoire à court terme, les mesures de débit au deux pas de temps précédents sont inclus parmi les covariables. Puis pour des effets mémoire à plus long terme, sont considérés des moyennes et des écarts-types mobiles avec des fenêtres d'une journée, d'une semaine et d'un mois.

Enfin, pour tenir compte de la présence éventuelle de cycle et de tendance annuels, trois covariables représentant l'année, le mois et la semaine associés à une date sont ajoutés à $\mathcal{F}_{t-1}$. Les covariables sont donc au nombre de 16.

Principaux résultats

L'efficacité et la pertinence de la pénalité proposée pour la log-vraisemblance dans l'équation (2.8) est évaluée sur un jeu de données synthétiques provenant d'un loi de Fréchet conditionnelle. L'objectif était de vérifier que la pénalité a bien pour effet de guider le choix des paramètres du mélange conditionnel vers des valeurs de paramètres de forme en accord avec la distribution a priori et avec le modèle générateur. L'histogramme des paramètres de forme estimés par le mélange conditionnel permet d'évaluer visuellement l'effet de la distribution a priori. L'information a priori est ensuite validée en comparant cet histogramme avec les paramètres de forme du modèle générateur. Une évaluation supplémentaire de l'effet de la pénalité est effectuée en comparant les quantiles extrêmes estimés par

le mélange conditionnel avec les quantiles extrêmes du modèle générateur. Ces analyses ont démontré l'utilité de la pénalité.

Les analyses sur les données du bassin de l'Orgeval ont porté sur un ensemble de test, distinct de l'ensemble d'entraînement, lequel a servi a la sélection d'hyper-paramètres à l'aide d'une validation croisée à cinq plis. Tout d'abord, la performance est évaluée quantitativement. La première mesure de performance est le pourcentage de valeurs de débit observé qui tombent dans les intervalles de confiance conditionnels de niveau 90 % prédits par le modèle. La deuxième mesure est le R^2 qui compare le débit observé au débit prédit par la médiane conditionnelle du modèle. Selon ces deux mesures, le mélange conditionnel de Pareto hybrides fournit une bonne performance, bien que celle-ci se détériore légèrement au pas de temps le plus long (12h).

Le deuxième volet d'évaluation de la performance est qualitatif et s'appuie sur des analyses graphiques de trois types. D'abord, le débit observé est comparé à la médiane prédite par le mélange conditionnel en fonction du temps, voir la figure 2.8. Cela permet de mettre en évidence les dynamiques qui sont bien reproduites par le mélange conditionnel. Cependant, les valeurs extrêmes sont un peu sous-estimées puisque l'on a recours à la médiane pour la prévision. Ensuite, les intervalles de confiance conditionnels de niveau 90 % tels qu'estimés par le mélange sont tracés en fonction du temps et le débit observé y est superposé. Cette fois, ce sont les dynamiques de l'incertitude telles qu'apprises par le mélange conditionnel qui sont mises en évidence. Enfin, l'effet de la pénalité peut être vérifié en examinant les histogrammes des paramètres de forme du modèle, qui sont tels que prescrits par la distribution a priori.

Perspectives

Le modèle de prévision du débit pourrait être utilisé de façon stochastique, à la façon des générateurs stochastiques de conditions météorologiques. De cette façon, les événements extrêmes seraient vraisemblablement plus à même d'être reproduits.

Une autre perspective concerne l'apport des réseaux de neurones profonds (Goodfellow *et al.* , 2016). Par exemple, permettraient-ils de nous affranchir d'injecter de l'information a priori sous forme de pénalité ?

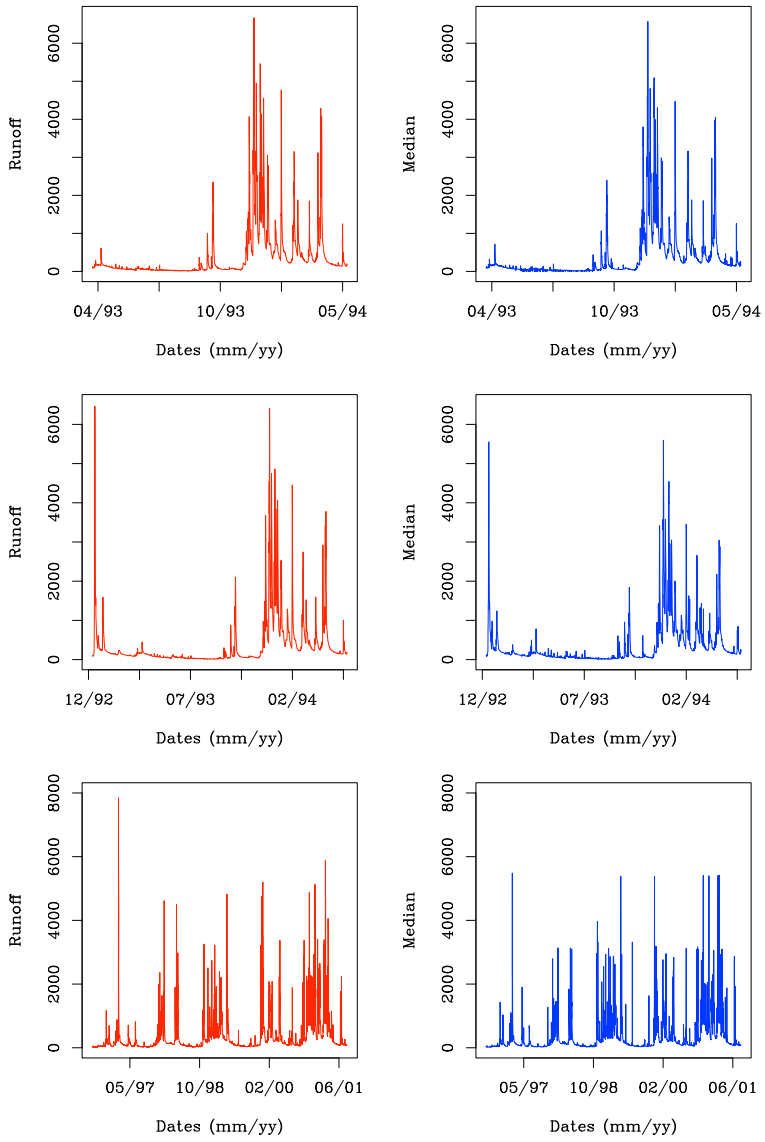


Figure 4. (left) Observed runoff of the Avenelles subbasin for the test period corresponding to a given time step: (top) 1 h, (middle) 6 h, and (bottom) 12 h. (right) Predicted median on the test set from the learned hybrid Pareto conditional mixture for the three time steps.

Figure 2.8 – *Tirée de Carreau* et al. *(2009).*

2.2.2 Descente d'échelle statistique

Je résume ici les travaux présentés dans Carreau & Vrac (2011) dont un exemplaire est disponible dans l'annexe A.2.2. Il s'agit d'une collaboration avec Mathieu Vrac, statisticien spécialisé dans la descente d'échelle statistique, qui a débuté lors de mon postdoctorat au LSCE. Les méthodes dites de *downscaling* statistiques, i.e. de descente d'échelle, cherchent à adapter et développer des approches statistiques pour estimer l'information à haute résolution spatiale à partir d'information à basse résolution spatiale.

Les modèles de climat globaux simulent des variables climatiques en résolvant des équations de la physique sur des grilles en trois dimensions qui couvrent le globe terrestre. Ces simulations permettent l'étude des mécanismes du climat passé, présent et futur de la Terre. Les impacts du changement climatique sur l'économie, l'agriculture, l'écologie et l'hydrologie dans les décennies à venir sont aussi cruciaux. Or, ces études d'impact requièrent des simulations de variables climatiques à haute résolution spatiale, allant de quelques kilomètres à l'échelle d'une station de mesure. En particulier, les précipitations, une variable d'importance majeure pour l'agriculture et l'hydrologie, se caractérisent par une très forte variabilité spatiale. En conséquence, la résolution spatiale des modèles de climat, avec des mailles de l'ordre de la centaine de km de côté, n'est pas du tout adaptée pour capturer une telle variabilité spatiale.

Dans la notation introduite précédemment, nous nous sommes intéressés au cas où Y représente une variable climatique en un site donné, il s'agit donc de l'information à haute résolution spatiale. De plus, $\boldsymbol{X}$ représente un vecteur issu des variables climatiques simulées par les modèles de climat globaux résumant la circulation atmosphérique à basse résolution spatiale.

Il existe plusieurs familles de méthodes de descente d'échelle statistiques. En particulier, les approches stochastiques, grâce à la distribution conditionnelle de $Y|\boldsymbol{X}$, peuvent simuler des scénarios plausibles de la variable à haute résolution spatiale étant donné une réalisation d'un scénario à basse résolution spatiale. Dans Carreau & Vrac (2011), nous avons adapté le modèle de mélange conditionnel au contexte de la descente d'échelle statistique des précipitations. De plus, nous avons évalué et comparé plusieurs types de mélanges conditionnels avec une approche qui fait référence dans la littérature sur des données de précipitations en trois sites dans le sud de la France.

Particularités de la distribution des précipitations

Les précipitations sont un élément clé pour les études d'impact, notamment en agriculture et en hydrologie, et les modèles de climat globaux n'en reproduisent pas bien les caractéristiques. En effet, les précipitations présentent une forte variabilité temporelle et spatiale. De plus, leur distribution peut avoir la queue lourde, comme c'est le cas dans la région méditerranéene française, connue pour ses épisodes de pluies intenses qui se produisent le plus souvent à l'automne.

Les précipitations ont la particularité d'être composées d'intensités positives, aux pas de temps où il pleut, et de zéros, aux pas de temps où il ne pleut pas. L'absence de précipitation introduit une masse de probabilité discrète dans la distribution qui peut être prise en compte de la façon suivante. Soit α la probabilité de précipitation, $\delta(\cdot)$ la fonction de Dirac et $\hat{p}(y;\theta)$ un modèle de densité pour les intensités. Alors, un estimateur de densité incluant une masse de probabilité en zéro peut s'écrire :

$$\hat{p}(y;\alpha,\theta) = (1-\alpha)\delta(y) + \alpha\hat{p}(y;\theta). \tag{2.9}$$

L'estimateur de densité $\hat{p}(y;\theta)$ modélise la distribution des intensités de précipitations qui sont stric-

tement positives. La densité doit donc prendre des valeurs positives uniquement sur $\mathbb{R}^+$. Or, les mélanges de distributions, voir l'équation (2.2), selon le choix de densité utilisé pour les composantes, ne sont pas forcément restreints à $\mathbb{R}^+$. Dans ce cas, nous avons choisi de tronquer la densité sous zéro de la façon suivante :

$$\begin{cases} 0 & \text{si } y \leq 0 \\ \hat{p}(y;\theta)/(1-\hat{P}(Y \leq 0;\theta)) & \text{si } y > 0, \end{cases}$$

avec $\hat{P}(Y \leq 0;\theta)$ la masse de densité sous zéro associée à $\hat{p}(y;\theta)$, i.e. la fonction de répartition évaluée en zéro.

Modèles comparés

Puisque l'on s'intéresse à la distribution conditionnelle de $Y|\boldsymbol{X} = \boldsymbol{x}$, les paramètres de l'estimateur de densité de l'équation (2.9) sont définis comme des fonctions telles que calculées par un réseau de neurones, voir les équations (2.6) et (2.7). Autrement dit, $\alpha(\boldsymbol{x};\omega)$ et $\theta(\boldsymbol{x};\omega)$ sont des fonctions de $\boldsymbol{x}$ de paramètres ω. Le vecteur $\boldsymbol{x}$ est déterminé de la même façon pour toutes les méthodes de descente d'échelle considérées. Nous avons choisi d'inclure dans $\boldsymbol{x}$ les valeurs de pression moyenne au niveau de la mer, fournies par un modèle de climat global en six points de grille autour du site étudié, et de trois variables décrivant la date (année, mois et semaine). La dimension de $\boldsymbol{x}$ est ensuite réduite à l'aide d'une analyse en composantes principales.

Le modèle pour la densité des intensités, $\hat{p}(y;\theta)$, dans le cas de l'approche de référence est la loi de Gamma qui a deux paramètres. Afin de mettre en évidence l'apport des mélanges conditionnels pour la descente d'échelle des précipitations, nous avons comparé des mélanges avec des composantes gaussiennes, log-normales et Pareto hybrides. Dans ce dernier cas, nous avons utilisé le même type de pénalité que proposée dans Carreau *et al.* (2009) mais en faisant l'hypothèse que le paramètre de forme de la distribution des précipitations est centré autour de 0.2, en accord avec des connaissances d'experts.

Principaux résultats

Les quatre modèles de descente d'échelle qui diffèrent au niveau du modèle pour la densité des intensités de précipitations ont été évalués et comparés par validation croisée. Leur performance est évaluée en termes de log-vraisemblance conditionnelle sur un jeu de données de test (non utilisé ni pour l'apprentissage ni pour la sélection des hyper-paramètres). Les résultats montrent que le mélange conditionnel de Pareto hybrides surpasse le modèle de référence (densité Gamma) et le mélange conditionnel de gaussiennes. En revanche, les mélanges conditionnels de Pareto hybrides et de Log-Normales ont une performance équivalente.

Nous avons, de plus, proposé une série d'analyses des modèles de descente d'échelle considérés à partir d'ensemble de simulations sur le jeu de données de test. Pour les différents modèles, nous avons comparé

— la distribution des intensités à l'aide de graphiques quantiles-quantiles qui ont mis en évidence la difficulté de la loi Gamma à fournir un ajustement adéquat autant pour la partie centrale que pour la queue de la distribution, voir la figure 2.9 ;

— la fréquence des durées sèches et pluvieuses qui sont estimés de façon équivalente par tous les modèles puisque le même mécanisme est utilisé pour tenir compte de la probabilité de pluie et

d'absence de pluie ;

— les cycles saisonniers de la probabilité de pluie et du quantile 99% des intensités ainsi que des intervalles de confiance de niveau 90% qui peuvent être validés en fonction des connaissances d'experts, voir la figure 2.10 ;

— les cycles saisonniers des paramètres des modèles conditionnels ainsi que des intervalles de confiance de niveau 90% qui fournissent un éclairage sur les mécanismes sous-jacents à la modélisation, voir la figure 2.11.

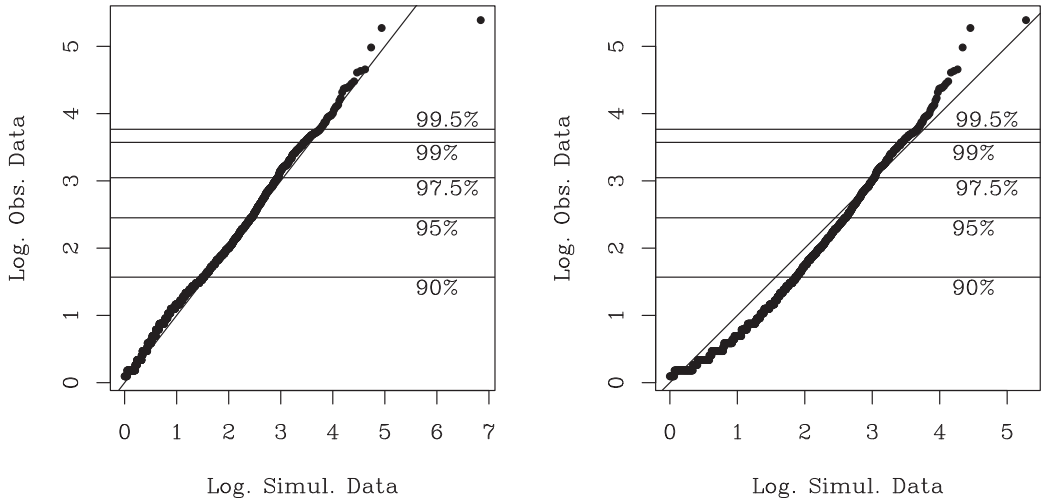


Figure 2. QQ-plots on a logarithmic scale of the simulated precipitation versus observations >1 mm on the Orange test set for (left) the hybrid Pareto conditional mixture and (right) the benchmark model. The horizontal lines are the empirical unconditional quantiles from observations of the test set.

Figure 2.9 – *Tirée de Carreau & Vrac (2011).*

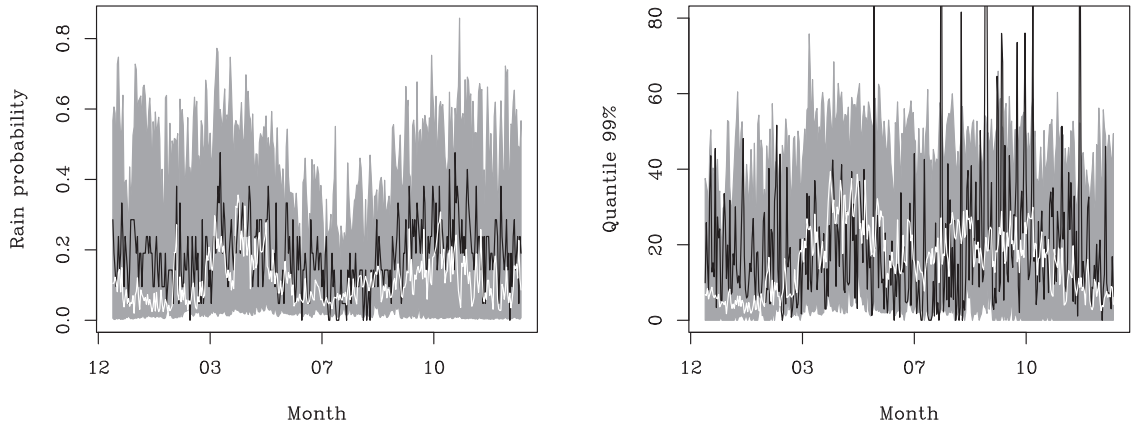


Figure 4. Daily seasonal cycles of (left) the rain occurrence probability and (right) the 99% quantile from the observations (black line) together with an empirical 90% confidence interval (gray band) and median (white line) from the hybrid Pareto conditional mixture for the Orange station test data. Peaks in the seasonal cycle of the occurrence process are visible in spring and fall.

Figure 2.10 – *Tirée de Carreau & Vrac (2011).*

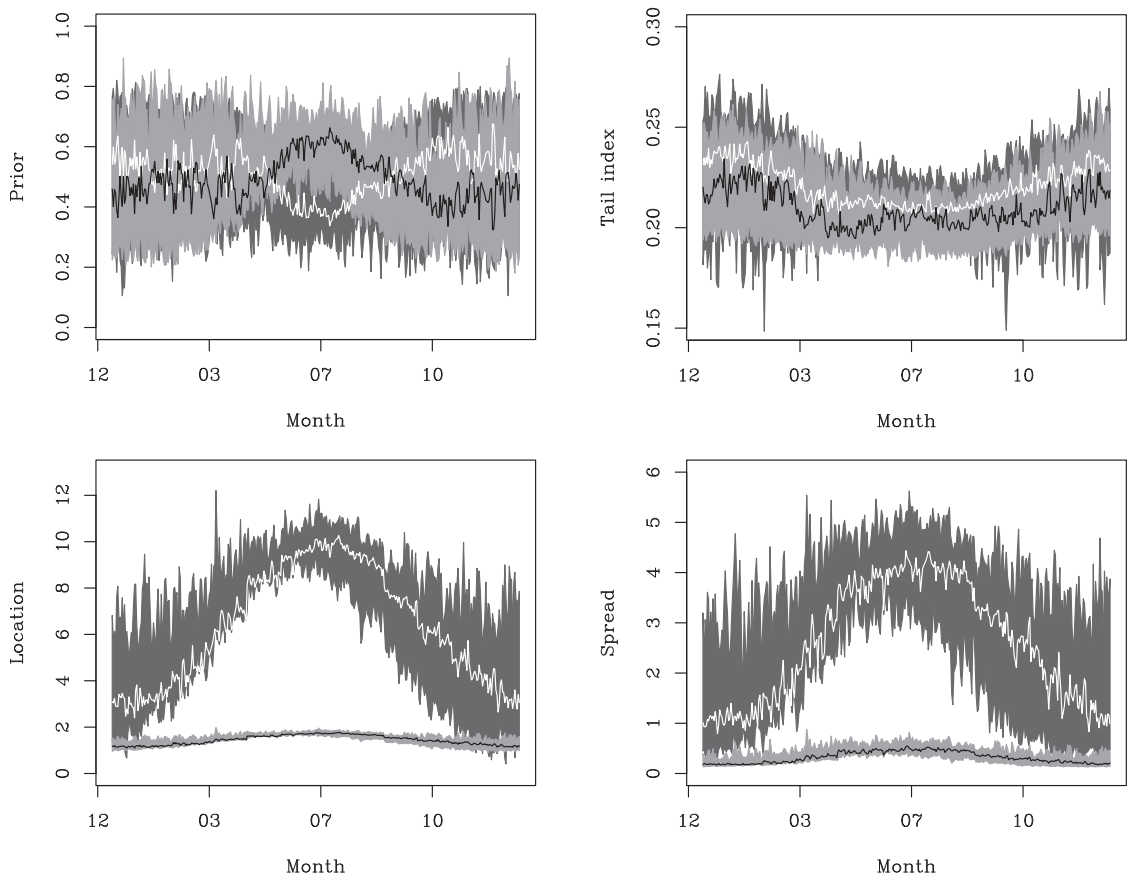


Figure 5. Daily seasonal cycles of the hybrid Pareto conditional mixture parameters (top left to bottom right: mixture weights π_j, tail index parameters ξ_j, location parameters μ_j, and scale parameters σ_j) together with an empirical 90% confidence interval. The mixture has two components whose parameters are represented by the black and white lines.

Figure 2.11 – *Tirée de Carreau & Vrac (2011).*

Pour finir, nous avons examiné les deux événements pluvieux les plus intenses du jeu de données de test, l'un en termes de volume, l'autre en termes de durée. Les précipitations observées ont été comparées aux quantiles de niveau 95%, 99% et 99.9% estimés par le mélange conditionnel de Pareto hybrides, voir la figure 2.12. Nous avons également tracé les densités estimées, dans la partie centrale et de la queue, pour chaque jour de chaque événement. Cela permet de mettre en évidence les dynamiques estimées et l'adaptation du modèle aux conditions atmosphériques.

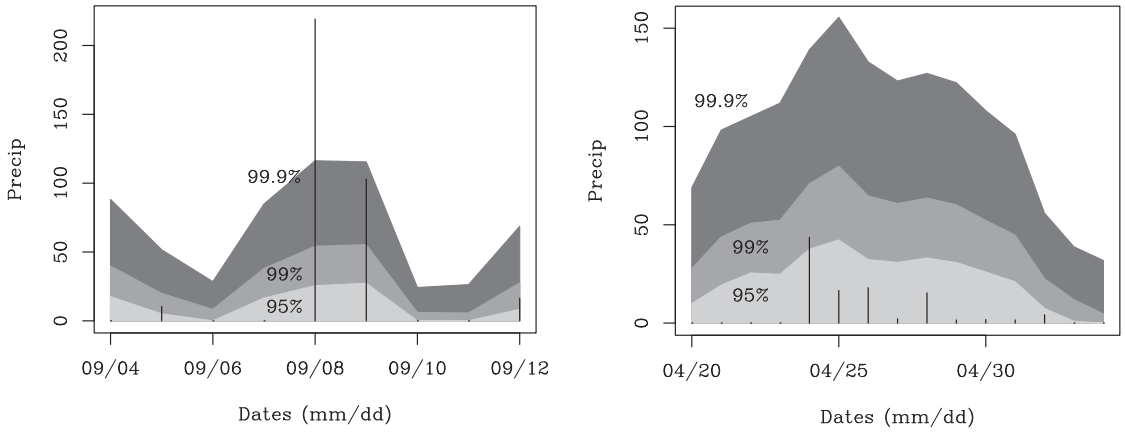


Figure 7. Hybrid Pareto conditional mixture: conditional quantiles $y_{0.95}(\mathbf{x})$ (light gray), $y_{0.99}(\mathbf{x})$ (medium gray), and $y_{0.999}(\mathbf{x})$ (dark gray) on (left) the wet spell with the highest volume of rain and (right) the longest wet spell from the Orange test data. Vertical lines represent the observed precipitation. The largest observed rainfall, 220 mm on 8 September (Figure 7, left), corresponds to a quantile level of 99.99% for the hybrid Pareto CMM.

Figure 2.12 – *Tirée de Carreau & Vrac (2011).*

Perspectives

La question de la dépendance spatiale, temporelle et inter-variables est au coeur de plusieurs travaux récents dans la communauté du downscaling (Cannon, 2018; Vrac, 2018). Comment le modèle de mélange conditionnel pourrait-il être étendu au cadre spatial (voir par exemple Williams (1996)) ?

2.3 Caractérisation spatiale des précipitations intenses

Dans ce troisième et dernier thème, les questions scientifiques sur lesquelles j'ai travaillé sont à la fois méthodologiques et issues d'applications. Les aspects méthodologiques, tout en étant complètement distincts des travaux présentés aux thèmes 2.1 et 2.2, s'appuient sur la même base, l'alliance d'approches non-paramétriques pour les valeurs centrales et paramétriques pour les valeurs extrêmes. De plus, les applications se focalisent sur de nouvelles problématiques en sciences du climat et de l'environnement tournées vers la spatialisation des précipitations intenses. Ces problématiques constituent les préliminaires des questions concernant la spatialisation de variables hydrométéorologiques pour générer des scénarios, au coeur de mon projet de recherche présenté au chapitre 3.

Je me suis intéressée à la question de l'interpolation spatiale des pluies intenses dans le sud de la France dans quelques travaux (Carreau & Girard, 2011; Ceresetti *et al.* , 2012; Carreau *et al.* , 2013, 2017). Je me focalise au § 2.3.1 sur les travaux les plus récents concernant le développement méthodologique d'une approche de type régionale. À partir de la loi de Pareto généralisée, nous avons proposé un mélange conditionnel qui permet de partionner une région en sous-régions en fonction du risque lié à l'occurrence d'événements extrêmes. Au § 2.3.2, je résume des travaux qui ont étudié l'impact du choix de modèle de densité multivariée pour les précipitations intenses mesurées sur un bassin-versant. L'objectif était de mettre en évidence les forces et les faiblesses de modèles proposés dans la littérature selon plusieurs critères statistiques et hydrologiques.

2.3.1 Approche régionale basée sur la loi de Pareto généralisée

Je résume ici les travaux présentés dans Carreau *et al.* (2017) dont un exemplaire est disponible dans l'annexe A.3.1. Il s'agit d'une collaboration avec Philippe Naveau du LSCE, avec qui j'ai fait mon premier postdoctorat en France, et avec Luc Neppel, expert en hydrologie statistique, avec qui je travaille au sein de l'équipe "Evénements extrêmes" d'HSM.

En Méditerranée française, des événements de fortes précipitations, appelés *épisodes cévenols*, se produisent principalement en automne et ont tendance à se regrouper dans des endroits très spécifiques. Les facteurs expliquant la localisation de ces événements sont la présence de montagnes et les trajectoires généralement empruntées par des masses d'air contrastées, humides et chaudes venant de la mer Méditerranée et froides du Nord (voir la figure 2.13, panneau de gauche). Ces épisodes de fortes précipitations peuvent déclencher des crues éclair, le principal danger naturel en région méditerranéenne, susceptibles de causer des morts et d'importants dégâts matériels.

Bien que les hydrologues effectuent généralement l'évaluation des risques en se basant sur des niveaux de retour tels que le niveau de retour sur 100 ans, la probabilité d'événements extrêmes, et donc les niveaux de danger, dépendent fortement du paramètre de forme de la loi de Pareto généralisée. Plusieurs approches de régression ont été développées pour interpoler la distribution des extrêmes au sein d'une région. Comme l'estimation du paramètre de forme est difficile, on fait souvent l'hypothèse qu'il est constant, ce qui entraîne un niveau de danger constant dans la région. L'approche régionale que nous proposons fait un compromis en laissant le paramètre de forme varier par sous-régions.

Soient $\{Y_1, \ldots, Y_d\}$, des variables aléatoires représentant les pluies intenses en d sites d'une région donnée. Plus précisément, définissons Y_s comme les excès, au site s, au-delà d'un seuil suffisamment élevé pour que l'approximation par la loi de Pareto généralisée soit raisonnable, voir § 2.1.1. Soit $\mu_s = \mathbb{E}[Y_s]$, la valeur espérée des excès au site s. Alors la distribution des variables standardisées Y_s/μ_s est toujours la loi de Pareto généralisée mais elle ne dépend que du paramètre de forme ξ_s qui caractérise le comportement des valeurs extrêmes. Une région est dite *homogène* si les Y_s/μ_s sont

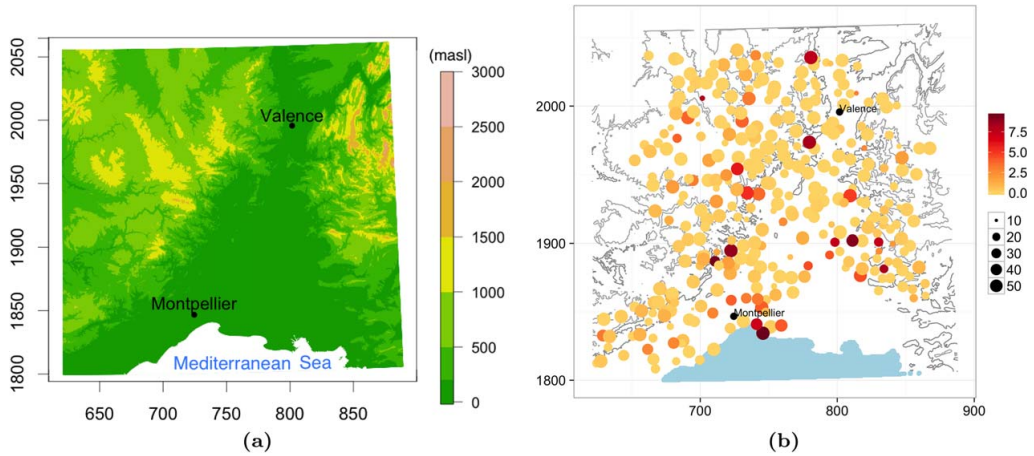


Figure 7. (a) Digital elevation map of the area of interest in the French Mediterranean. (b) 332 rain gauge stations of the Météo-France network (French weather service) covering the period 1 January 1958 to 31 December 2014 (57 years). The size of the symbol is proportional to the length of the observation period (10–57 years) and the color shade (light orange to dark red) indicates the percentage of missing values (0–10%).

Figure 2.13 – *Tirée de Carreau* et al. *(2017).*

identiquement distribuées. De façon équivalente, cela signifie que les paramètres de forme sont les mêmes à tous les sites.

Le principe de base de l'approche régionale est l'identification des régions homogènes (Hosking & Wallis, 2005). Cela permet, à partir des observations standardisées, d'obtenir des échantillons plus grands pour l'estimation du paramètre de forme et d'en réduire l'incertitude. De plus, aux sites non-jaugés, il suffit d'interpoler μ_s pour déduire les paramètres de la loi de Pareto généralisée. L'approche régionale peut être vue comme une méthode d'estimation de la loi conditionnelle $Y|\boldsymbol{x}$ où $\boldsymbol{x}$ contient des covariables spécifiques à chaque site, e.g. de l'information géographique.

Inférence

Pour une région homogène donnée, nous avons proposé une méthode d'inférence qui repose sur les moments pondérés afin d'estimer les paramètres de la loi de Pareto généralisée. Le premier moment pondéré est la valeur espérée μ qui sert à standardiser les observations. Nous avons recours à la régression par noyaux pour estimer non-paramétriquement $\mu(\boldsymbol{x}_s)$ à partir des observations en chaque site s. Pour estimer le paramètre de forme, nous utilisons le deuxième moment pondéré de la variable standardisée (le premier moment étant égal trivialement à 1 par construction). Dans ce cas, les observations standardisées sur l'ensemble des sites de la région contribuent à l'estimation puisque le paramètre de forme est supposé constant dans un région homogène. Le paramètre d'échelle de la loi de Pareto est estimé en combinant les estimateurs du paramètre de forme et de la valeur espérée.

Pour une région hétérogène, comme la région d'étude de la figure 2.13, nous faisons l'hypothèse qu'elle puisse être partitionner en K sous-régions homogènes. Ceci permet d'exprimer la distribution conditionnelle de $Y|\boldsymbol{x}$ comme un mélange conditionnel de loi de Paretos généralisées et de proposer un algorithme en deux étapes analogue à l'algorithme *Expectation-Maximization* mentionné au § 2.1.1. L'étape *E* consiste à déterminer la partition. Pour ce faire, nous appliquons l'algorithme *KMeans*, un algorithme de classification non-supervisée non-paramétrique, au deuxième moment pondéré de la variable standardisée estimée site par site. Pour l'étape *M*, il s'agit d'employer l'inférence pour chaque sous-région homogène telle que décrite au paragraphe précédent.

Pour estimer les paramètres de la loi de Pareto généralisée en un site non-jaugé, il faut d'abord établir

dans quelle sous-région de la partition il appartient. Il s'agit maintenant d'un problème de classification supervisée pour lequel nous avons utilisé les k plus proches voisins, un classifieur non-paramétrique. Le paramètre de forme du site non-jaugé est celui de la sous-région auquel le site appartient. Il suffit ensuite d'interpoler $\mu(\boldsymbol{x})$ au site en question pour en déduire la paramètre d'échelle.

L'identification du nombre de sous-régions homogènes dans la partition donne lieu à un compromis biais-variance. Plus il y a de sous-régions, plus le modèle est en mesure d'estimer finement les variations spatiales du paramètre de forme. En revanche, le nombre d'observations pour estimer chaque paramètre de forme diminue et en conséquence, la variance de l'estimation est plus élevée. Nous avons donc proposé de déterminer la taille de la partition avec une procédure de validation-croisée qui permet de minimiser simultanément le biais et la variance.

Résultats principaux

Dans un premier temps, nous avons effectué des analyses sur 18 jeux de données synthétiques, issus du mélange conditionnel de lois de Pareto généralisée. Les jeux de données variaient en taille d'échantillon, nombre de sites et taille de partition. Chaque jeu de donnée a été répété 1000 fois pour évaluer l'incertitude. Trois critères ont été employés pour mesurer la performance : la log-vraisemblance, la statistique d'Anderson-Darling et la somme des erreurs carrés des quantiles.

Ces analyses ont permis, d'une part, d'évaluer la procédure de sélection de la taille de partition. Celle-ci a une légère tendance à la sur-estimation, en particulier pour les jeux de données plus grands (en nombre de sites ou en taille d'échantillon). D'autre part, les analyses ont mis en évidence la performance de la stratégie d'inférence des paramètres de la loi de Pareto généralisée. Même si la taille de la partition n'est pas toujours celle du modèle générateur, les paramètres estimés convergent vers les valeurs des paramètres du modèle générateur lorsque la taille du jeu de données augmente.

Dans un deuxième temps, des analyses sur un jeu de données de 332 pluviomètres dans le sud de la France ont permis de comparer trois approches d'interpolation des pluies basées sur la loi de Pareto généralisée, voir la figure 2.14. La première est l'approche régionale avec une seule région homogène, i.e. qu'elle fait l'hypothèse qu'un seul indice de forme suffit à caractériser le comportement des valeurs extrêmes pour tous les sites. La deuxième est le mélange conditionnel de Pareto généralisées, avec une partition dont la taille est fixée en validation croisée. Cela revient à une approche régionale qui découpe la région en sous-régions homogènes contiguës. Cette approche autorise le paramètre de forme à varier par sous-région, voir le panneau de gauche de la figure 2.15. La troisième approche interpole, avec la régression à noyaux, les paramètres de la loi de Pareto estimés en chaque site. Le paramètre de forme peut donc varier de façon très libre à travers la région, voir le panneau de droite de la figure 2.15.

De ces analyses, il ressort clairement que le mélange conditionnel de Pareto généralisées, avec quatre sous-régions homogènes sélectionnées, surpasse l'approche régionale avec une seule région homogène selon les trois critères de performance. Toutefois, sa performance n'est pas significativement différente de l'approche d'interpolation des paramètres par régression à noyaux, ce qui peut s'expliquer par le fait que le jeu de données est très dense. Les figures 2.16 et 2.17 illustrent une partie de ces analyses.

Perspectives

Une extension intéressante de ces travaux concerne la possibilité de prendre en compte toute la distribution des précipitations, pas seulement les extrêmes, comme dans Naveau *et al.* (2016).

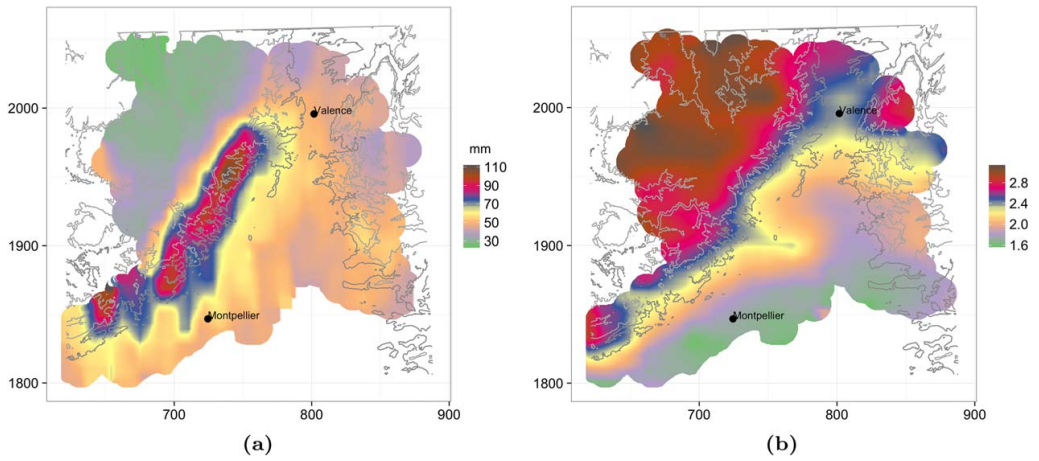


Figure 8. French Mediterranean precipitation data: (a) Threshold defined as the 98% quantile of precipitation intensities (b) Average number of excesses above the threshold per year. Interpolation onto the grid is performed with kernel regression.

Figure 2.14 – *Tirée de Carreau* et al. *(2017).*

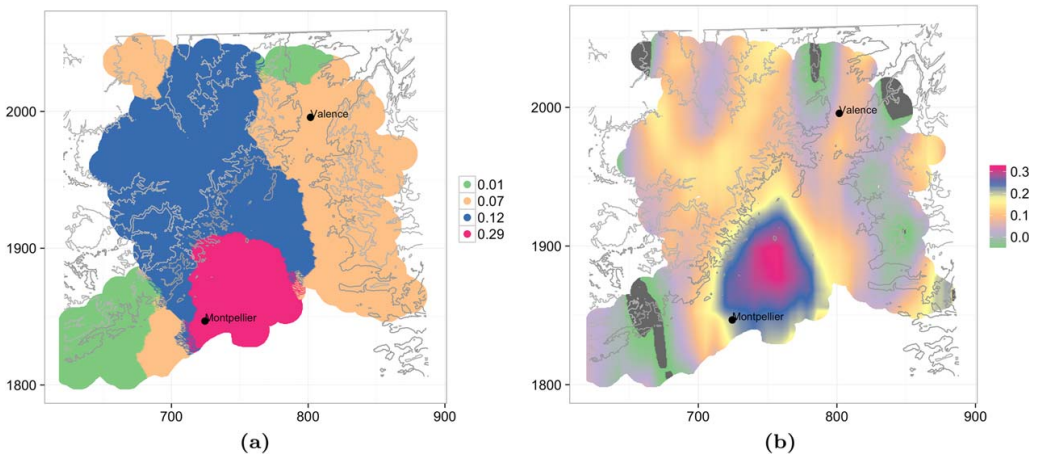


Figure 9. Interpolated shape parameter of the GP distribution onto the grid with (a) the regional peaks-over-threshold model with four subregions and (b) kernel regression applied to at-site shape parameter estimates. Less than 5% of the values are below −0.05 and are shown in dark gray. The shape parameter estimate of the regional peaks-over-threshold model with a single region is $\hat{\xi}_1 = 0.11$.

Figure 2.15 – *Tirée de Carreau* et al. *(2017).*

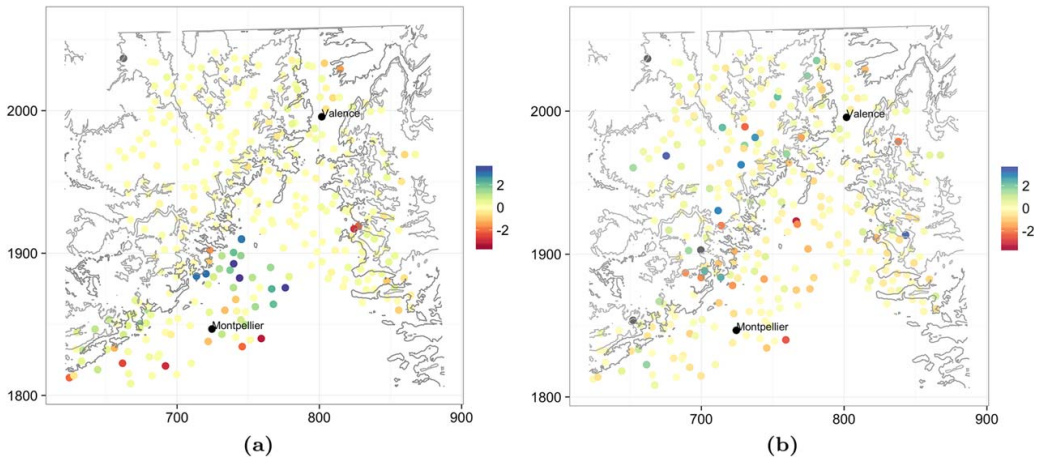


Figure 11. Negative log-likelihood relative to the regional model with four subregions. (a) For the regional model with a single region and (b) for the kernel regression interpolation of the at-site estimates. Positive values (blue shades) indicate that the regional model with four subregions outperforms the other interpolation approach in terms of log-likelihood.

Figure 2.16 – *Tirée de Carreau* et al. *(2017).*

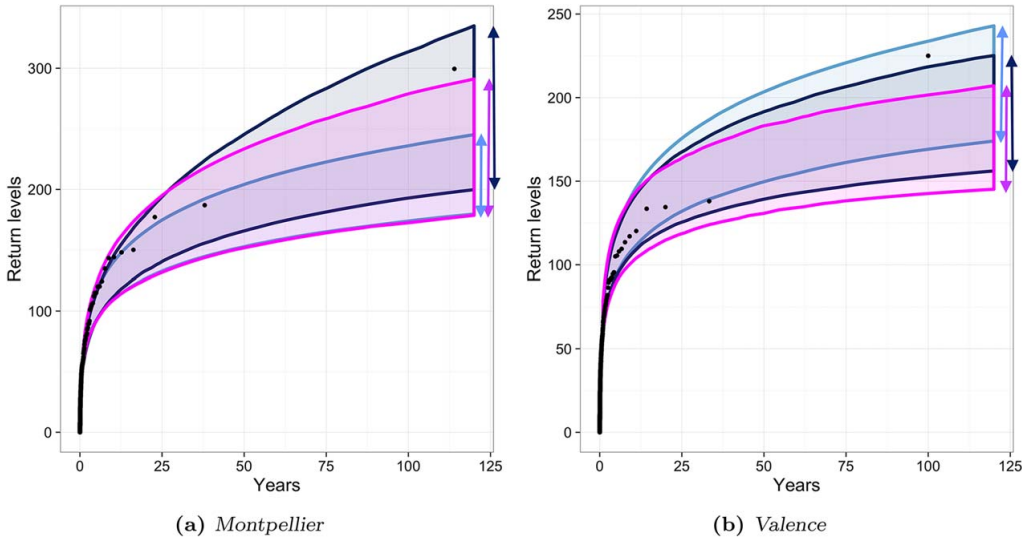


Figure 12. Return level curves from the regional peaks-over-threshold model with a single region (light blue), with four hazard subregions (dark blue) and the kernel regression interpolation of at-site estimates (magenta). The points represent the empirical return levels.

Figure 2.17 – *Tirée de Carreau* et al. *(2017).*

2.3.2 Étude comparative des choix de modèles de densité multivariée

Je résume ici les travaux présentés dans Carreau & Bouvier (2016) dont un exemplaire est disponible dans l'annexe A.3.2. Ces travaux résultent d'une collaboration avec Christophe Bouvier, hydrologue, dans la même équipe que moi, "Événements extrêmes", à HSM. L'apport principal de cette étude comparative est de mettre en relation des questions scientifiques du point de vue hydrologique et statistique. Plus précisément, nous avons cherché à ce que les analyses statistiques soient le plus possible au diapason de la problématique hydrologique.

Nous nous sommes intéressés aux cumuls de précipitation journaliers en huit sites du bassin-versant d'Anduze situé dans la chaîne de montagne des Cévennes dans le sud de la France, un lieu connu pour les événements de pluies intenses appelés *épisodes cévénols*. Bien que les crues éclair soient souvent associées avec des épisodes de pluie courts - de quelques heures - et localisés, elles peuvent aussi être la conséquence de pluie aux cumuls modérés mais sur une certaine durée et affectant le bassin au complet.

Selon l'expertise hydrologique, le risque de crue peut se définir en termes de l'importance de la lame d'eau, i.e. la moyenne spatiale de pluie, reçue par le bassin-versant un jour donné. Nous avons donc défini notre objet d'étude comme des vecteurs, représentant le cumul journalier en chaque site du bassin-versant, dont la moyenne est supérieure à un seuil. De ce fait, différents types d'événements - courts et intenses ou longs et modérés - peuvent être inclus.

Les analyses statistiques proposées, autant dans le volet exploratoire que pour la modélisation, mettent l'accent sur la structure de dépendance spatiale de ces pluies pouvant mener à des crues. Il s'agit d'une élément essentiel dans la conception d'un générateur stochastique multi-sites de précipitations. L'approche méta-gaussienne, qui a recours à la structure de dépendance gaussienne avec des lois marginales transformées, est souvent utilisée. L'objectif de ces travaux est de mettre en évidence les forces et les faiblesses de différents choix, dont l'approche méta-gaussienne, pour la structure de dépendance.

Analyse exploratoire

Il y a moins de 2% des jours sur la période de 43 ans considérée pour lesquels la moyenne spatiale est supérieure au seuil prescrit. Bien que cela se produise principalement en automne, il y a des occurrences tout au long de l'année et plusieurs mécanismes météorologiques peuvent en être responsables. Nous avons traité cette possibilité d'un point de vue statistique en autorisant des mélanges de distributions. En supposant de plus que la dépendance temporelle était négligeable, nous avons fait l'hypothèse que les différentes observations sélectionnées étaient indépendantes et identiquement distribuées.

Nous avons d'abord étudié la dépendance entre chaque paire de stations à l'aide du coefficient τ de Kendall. Il s'agit d'une mesure de dépendance qui s'appuie sur les rangs, plus générale que le traditionnel coefficient de corrélation de Pearson mais qui s'interprète de la même manière. On remarque que la force de la dépendance décroît linéairement avec la distance entre les sites, voir la figure 2.18. Nous avons également calculé le χ, une mesure de dépendance pour les événements extrêmes. Pour des paires de stations rapprochées, les graphiques du χ permettent d'identifier de la dépendance dans les valeurs extrêmes et inversement pour des paires de stations éloignées.

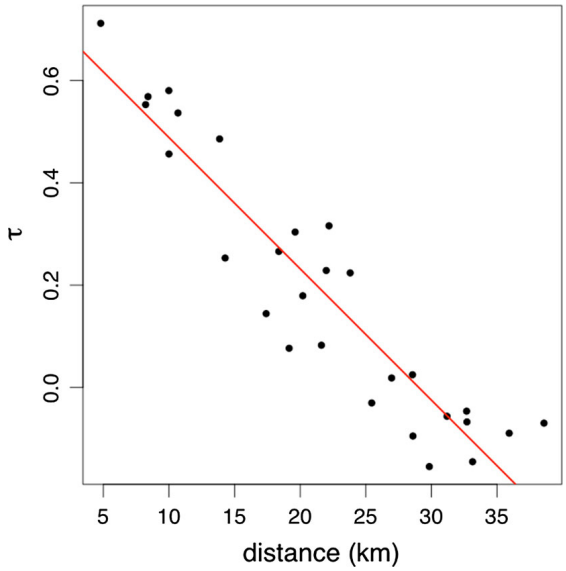


Fig. 3 Plot of the estimated Kendall's τ coefficients with respect to horizontal distances together with a regression line

Figure 2.18 – *Tirée de Carreau & Bouvier (2016).*

Modèles de densité multivariée

Les modèles de densité multivariée se décomposent en deux parties. D'une part, un modèle pour les lois marginales univariées sert à décrire la distribution en chaque site. D'autre part, un modèle de structure de dépendance permet de caractériser l'inter-dépendance spatiale entre les huit sites. Nous avons croisé deux possibilités pour les lois marginales avec quatre pour les structures de dépendance et donc huit modèles de densité multivariée sont inclus dans la comparaison.

Concernant les lois marginales univariées, les deux modèles considérés sont la loi Gamma et le mélange de lois Log-Normales. La loi Gamma est classiquement utilisée pour modéliser les précipitations mais peine souvent à reproduire les fortes valeurs. Deux composantes ont été sélectionnées avec le critère BIC (Bayesian Information Criterion) dans le mélange de lois Log-Normales. Ces deux composantes peuvent être pensées comme résultant de différentes populations dues, par exemple, à des mécanismes météorologiques distincts. De plus, le mélange, ayant plus de paramètres et plus de flexibilité, est plus à même de représenter les valeurs extrêmes.

En ordre croissant de complexité, les structures de dépendance issues des lois multivariées Normale, Student t, Skew Normal et Skew t ont été employées dans la comparaison. Les deux premières structures de dépendance sont symétriques alors que les deux dernières permettent l'asymétrie. Les structures de dépendance issues de la loi Normale et de la Skew Normal impliquent l'indépendance dans les valeurs extrêmes, ce qu'on appelle l'*indépendance asymptotique* alors que celles issues de la Student t et de la Skew t autorisent la *dépendance asymptotique*.

Résultats principaux

Nous avons mené une première série d'analyses afin d'évaluer de façon qualitative l'ajustement des modèles statistiques en termes des lois marginales univariées et bivariées. Des graphiques quantiles-quantiles accompagnés d'intervalles de confiance calculés par bootstrap, voir la figure 2.19, ont mis en évidence une meilleure reproduction du comportement des fortes valeurs en plusieurs stations du

mélange de deux lois Log-Normales. Cependant, ceci vient au prix d'intervalles de confiance plus larges, en raison sans doute du plus grand nombre de paramètres du mélange. Les structures de dépendance bivariées sont capables, de façon approximativement équivalentes, de reproduire la force de la dépendance telle que mesurée par le τ de Kendall, voir la figure 2.20. Néanmoins, les formes des structures de dépendance bivariées ajustées sont très différentes, comme on peut le voir notamment à l'aide d'histogrammes bivariés.

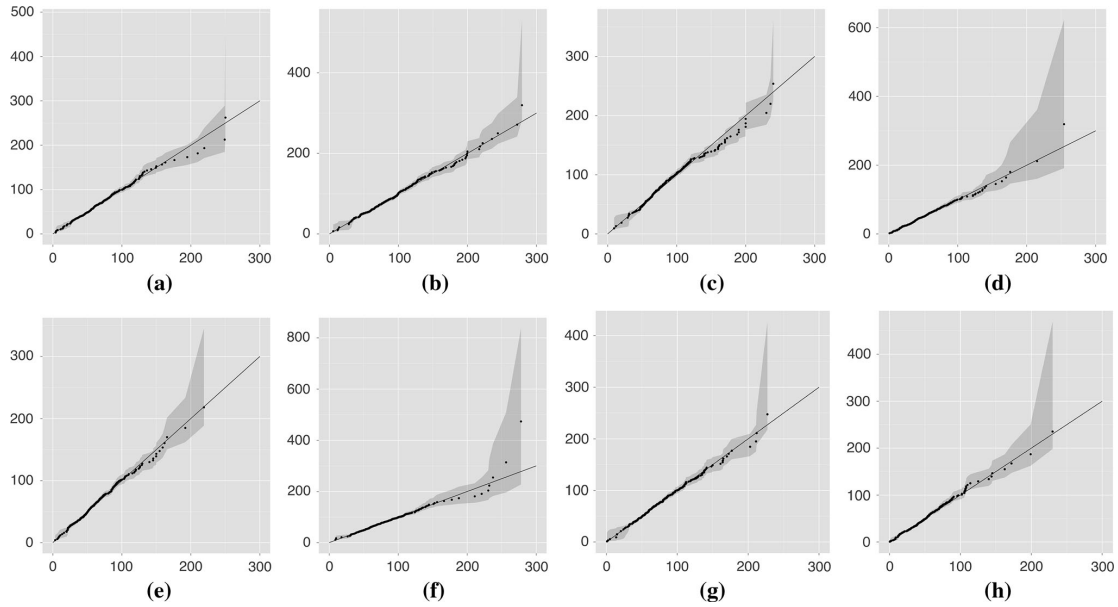


Fig. 9 Quantile–quantile plots of the 2-component Log-Normal mixture distribution at each station with parametric bootstrap 95 % confidence interval. Empirical quantiles (theoretical quantiles) are on the *x*-axis (*y*-axis).The range of the first diagonal covers [0,300] on both axes in all plots. **a** Barre-des-Cevennes. **b** Cassagnas. **c** Lecollet-de-Deze. **d** Ales. **e** Generargues. **f** Lasalle. **g** Saint-Andre-de-Val. **h** Saint-Christol-les-A

Figure 2.19 – *Tirée de Carreau & Bouvier (2016).*

Une évaluation quantitative systématique est ensuite conduite à l'aide d'une validation croisée. Les statistiques de Cramer-Von mises et d'Anderson-Darling sont utilisés comme mesures de performance avec des intervalles de confiance calculés avec l'erreur standard, voir la figure 2.21. Il ressort que le modèle de densité multivariée combinant le mélange de deux Log-Normales avec la structure de dépendance de la loi Skew Normale surpasse significativement les sept autres modèles par rapport aux deux mesures de performance. De plus, l'emploi du mélange de Log-Normales pour les lois marginales univariées améliore, dans presque tous les cas, la performance des modèles de densité. Les structures de dépendance autorisant la dépendance asymptotique ne donnent pas de meilleures performances que les celles qui impliquent l'indépendance asymptotique.

La dernière série d'analyses s'appuie sur deux quantités qui s'interprêtent plus directement en hydrologie : les niveaux de retour de la moyenne des huit cumuls journalier, i.e. la moyenne spatiale, et des probabilités conditionnelles de dépassement des niveaux de retour en deux paires de sites. Les analyses concernant les courbes de niveaux de retour de la moyenne spatiale, voir la figure 2.22, confirment les conclusions précédentes sur les différences entre les deux types de lois marginales univariées et sur les structures de dépendance autorisant ou non la dépendance asymptotique. De plus, elles font ressortir l'un des avantages de la structure de dépendance de la loi Skew Normale qui est de mieux reproduire les niveaux de retour de la moyenne spatiale pour les plus faibles valeurs (inférieures à 100 mm). Les analyses sur les probabilités conditionnelles de dépassement mettent essentiellement en évidence des différences entre les différents modèles théoriques. Il reste toutefois difficile de tirer des conclusions en raison de la forte incertitude liée aux estimations empiriques.

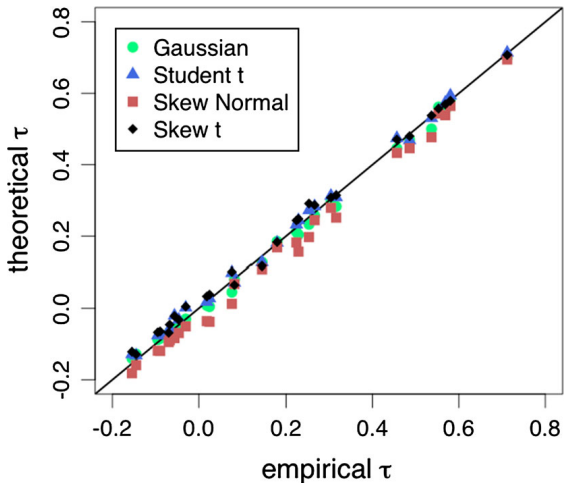


Fig. 10 Theoretical Kendall's τ coefficients of the fitted four spatial dependence structures with respect to the empirical Kendall's τ coefficients for all pairs of stations

Figure 2.20 – *Tirée de Carreau & Bouvier (2016).*

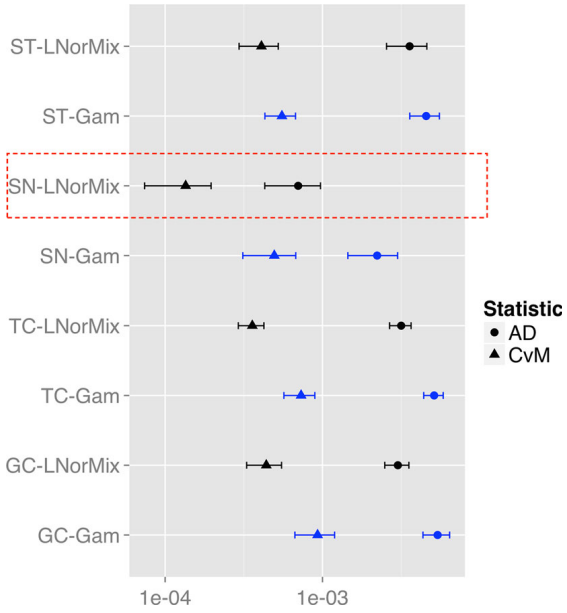


Fig. 13 Model selection based on the leave-one-out evaluation of the Cramer-Von Mises (CvM) and the Anderson-Darling (AD) goodness-of-fit statistics: average value and 95 % confidence interval are shown on a logarithmic scale. In blue, models with Gamma margins (Gam), and in black, with 2-component Log-Normal mixture margins (LNorMix). GC and TC stand for Gaussian and Student t copulas and SN and ST for Skew Normal and Skew t. The model SN-LNorMix outperforms significantly the other models

Figure 2.21 – *Tirée de Carreau & Bouvier (2016).*

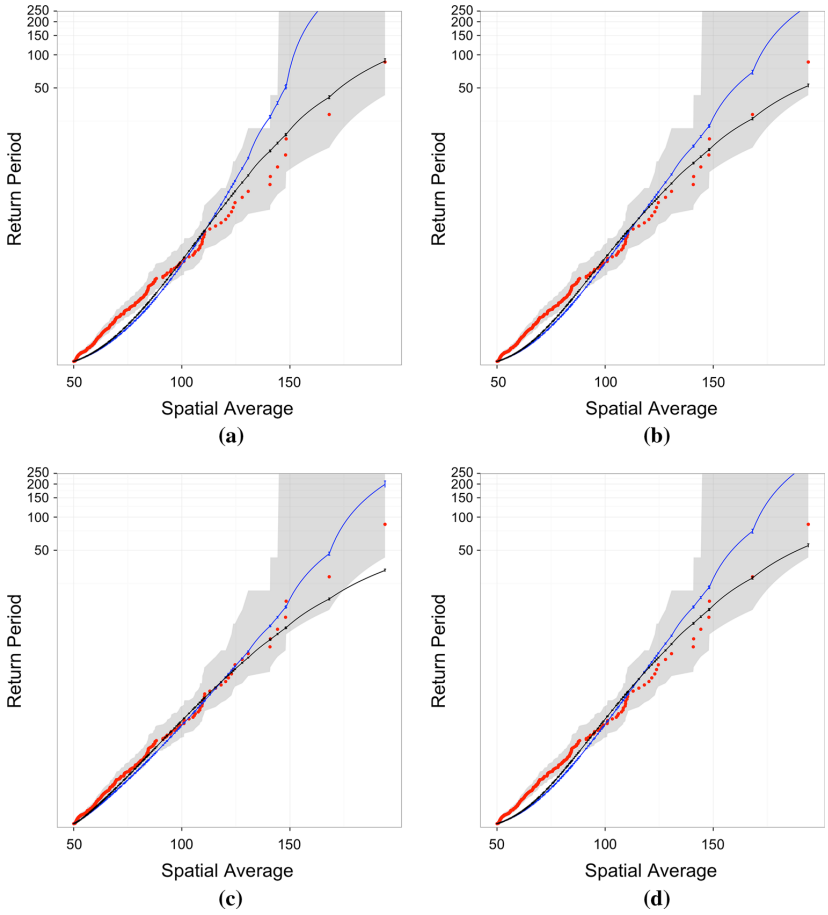


Fig. 14 Empirical (*red dots*) and theoretical return periods (*blue curve* for the Gamma margins and *black curve* for the 2-component Log-Normal mixture margins) in logarithmic scale of the observed spatial average are plotted against the observed spatial averages. The 95 % confidence band of the empirical estimates are shown in *grey* while those of the theoretical estimates are the tiny *vertical bars* along the *blue* and *black curves*. **a** Gaussian copula. **b** Student t copula. **c** Skew Normal. **d** Skew t

Figure 2.22 – *Tirée de Carreau & Bouvier (2016).*

Perspectives

Ces analyses peuvent servir de base pour définir un générateur stochastique de précipitation multi-sites, voire spatial, adapté pour les événements extrêmes. Il faut établir une stratégie pour que la modélisation statistique prenne en compte aussi les pluies ordinaires et l'absence de pluie.

Chapitre 3

Projet de recherche

Le projet de recherche que j'envisage s'intitule **Scénarios de variables hydrométéorologiques en région méditerranéenne : approches stochastiques et semi-paramétriques**. Il met à profit la base de connaissances et d'expériences acquises depuis mes études doctorales. Sur le plan méthodologique, je m'appuie d'une part sur des approches non-paramétriques provenant de l'apprentissage statistique pour modéliser les valeurs centrales et d'autre part, sur des approches paramétriques provenant de la théorie des valeurs extrêmes qui permettent d'extrapoler vers de grandes valeurs non observées. Sur le plan des applications, j'ai développé une expertise sur les questions reliées à l'hydrologie, notamment sur la caractérisation des précipitations en Méditerranée française.

Mon projet de recherche met également à profit un réseau de collaborations avec deux communautés l'une centrée sur l'IMAG et l'autre, sur le LMI NAÏLA. Ce réseau m'est indispensable pour stimuler et alimenter mes connaissances et mes questions de recherche. En conséquence, mes questions de recherche sont à deux sens : nourries autant par les études d'impact que par les développements méthodologiques.

3.1 Enjeux et objectifs

La région méditerranéenne se caractérise par un climat subhumide à semi-aride ainsi que des reliefs montagneux et collinaires. Elle est soumise aux influences maritime et désertique selon les champs de vents. Elle connaît des épisodes de pluies extrêmes qui, en raison des propriétés hydrodynamiques du sol et du sous-sol dans les bassins versants, peuvent entraîner des crues éclair, l'un des risques naturels les plus destructeurs de la région (Braud *et al.* , 2014). De plus, la distribution inégale dans le temps et dans l'espace de la ressource en eau combinée à une anthropisation marquée crée des problèmes de rareté de la ressource en eau, en particulier pour l'agriculture (Alazard *et al.* , 2015). La Méditerranée étant l'une des régions particulièrement sensibles au changement climatique, ces phénomènes iront vraisemblablement en s'accentuant. Pour faire face à ces problématiques spécifiques concernant la gestion des ressources en eau et des risques naturels, des outils d'aide à la décision sont nécessaires.

Pour développer des outils d'aide à la décision en soutien aux politiques de gestion, on peut avoir recours à des modèles à base physique qui permettent de simuler les possibles évolutions. Ces modèles cherchent à reproduire, à partir des connaissances sur les processus, les transferts sol - plante - atmosphère, la consommation en eau des cultures, la production culturale, la répartition ruissellement / infiltration, les débits emboîtés et leur cumul aux exutoires. Ils sont contraints en entrée par de l'information sur les variables hydrométéorologiques telles que les précipitations, la température, le vent, l'humidité de l'air et le rayonnement. Cette information peut provenir d'un réseau de sites instrumentés, de produits de la télédétection ou de simulations des modèles de climat.

Des scénarios spatialisés intégrant **la forte variabilité spatiale et temporelle du climat méditerranéen** doivent être générés à partir de l'information disponible sur les variables hydrométéorologiques. Pour ce faire, des résolutions spatio-temporelles fines (de l'ordre du km et de l'heure) sont requises. Il est aussi essentiel de respecter **les dépendances entre les différentes variables** lorsque plusieurs variables contraignent un modèle physique (e.g. pour les transferts sol - plante - atmosphère). Certaines lois de probabilités qui caractérisent les variables hydrométéorologiques sont **non-gaussiennes** ce qui demande une adaptation des méthodes classiques de statistiques. En particulier, les précipitations présentent **des événements extrêmes**. Des lois de probabilités classiques (telles que les lois Normale ou Gamma) et les processus gaussiens pour décrire la structure de dépendance sont mis en défaut et il faut faire appel à des approches statistiques qui s'appuient sur **la théorie des valeurs extrêmes**. Enfin, **les configurations instrumentales**, notamment la faible densité des sites instrumentés, la disparité entre les sites au niveau des variables mesurées et les périodes d'observation courtes et souvent non-concomitantes, mettent au défi les approches statistiques conventionnelles.

L'**objectif principal** de mon projet de recherche est de développer, à l'aide **d'approches stochastiques et semi-paramétriques**, et d'évaluer **des scénarios de variables hydrométéorologiques**. Les approches stochastiques autorisent une prise en compte l'incertitude en générant plusieurs scénarios pour une situation donnée. De plus, elles ont la possibilité d'extrapoler, i.e. de sortir de la gamme des valeurs observées. Une approche semi-paramétrique combine des aspects paramétriques et non-paramétriques. Cela permet de relaxer les hypothèses sous-jacentes à la modélisation et d'exploiter des données disponibles sur des grilles, e.g. les produits de la télédétection ou les simulations des modèles de climat. Cet objectif principal se décline en **deux objectifs spécifiques :**

(OS.1) la génération de scénarios multi-variables en rive sud de la Méditerranée pour l'étude des ressources en eau ;
(OS.2) la génération de scénarios de précipitation (une seule variable) intégrant des événements extrêmes pour l'étude des risques naturels.

3.2 État de l'art

Un générateur stochastique est, dans la plupart des cas, un assemblage de plusieurs approches statistiques qui cherche à reproduire différents aspects de la distribution des variables hydrométéorologiques à partir de laquelle des scénarios seront tirés aléatoirement. Je me focalise ici sur les générateurs spatiaux, i.e. qui génèrent des champs de valeurs en tout point de la zone d'étude, et à pas de temps discret, par exemple à la journée ou à l'heure, car ils sont, à ma connaissance, plus à même de répondre aux objectifs du présent projet.

Le regroupement des pas de temps en situations météorologiques typiques appelées **types de temps** permet de prendre en compte en partie la variabilité temporelle et saisonnière des variables hydrométéorologiques. Souvent, les types de temps sont définis à partir de description de la circulation atmosphérique (Garavaglia *et al.* , 2010). Il est aussi possible de les déterminer à partir des variables hydrométéorologiques mêmes, voir Leblois (2012) et Monbet (2018). Pour tenir compte de la variabilité spatiale, on peut établir un lien entre les distributions des intensités des variables hydrométéorologiques et des covariables qui varient dans l'espace. En plus des coordonnées géographiques, on peut définir **des caractéristiques du paysage** à partir d'un modèle numérique de terrain ce qui permet de prendre en considération l'influence de l'orographie (Arnaud *et al.* , 2006). Ces techniques pour modéliser la variabilité temporelle et spatiale n'ont pas été développées spécifiquement pour la gestion des ressources en eau ou l'étude des risques naturels en contexte méditerranéen et devront donc être adaptées. Par ailleurs, peu de travaux portent sur des générateurs stochastiques à des résolutions infra-journalières nécessaires pour rendre compte de la variabilité du climat méditerranéen (Benoit *et al.* , 2018). Pour pallier au manque de données à la résolution requise, **des techniques de descente d'échelle temporelle** peuvent être employées (Allard & Bourotte, 2015; Carreau *et al.* , 2019).

Beaucoup de générateurs stochastiques spatiaux reposent sur des processus gaussiens pour modéliser les structures de dépendance. Lorsque les lois de probabilités caractérisant les intensités des variables hydrométéorologiques sont non-gaussiennes, **la calibration du processus gaussien** demande des méthodes spécifiques dont la mise en oeuvre peut être exigeante (Kleiber *et al.* , 2012; Baxevani & Lennartsson, 2015; Bourotte *et al.* , 2016). D'autre part, il existe peu de modèles théoriques de fonction de covariance croisée qui décrivent **la dépendance multivariée**, i.e. inter-variables, en plus de la dépendance spatio-temporelle (Bourotte *et al.* , 2016). Une autre solution consiste à décomposer la loi de probabilité multivariée en produit de lois conditionnelles univariées (Chandler, 2014). Ces modèles multivariés n'ont pas été appliqués à des résolutions infra-journalières ni en contexte méditerranéen. Enfin, les structures de dépendance gaussiennes sont souvent inadaptés pour modéliser des événements extrêmes (Carreau & Bouvier, 2016). Les premières approches spatiales issues de la théorie des valeurs extrêmes sont basées sur les processus max-stables théoriquement justifiés lorsque l'on s'intéresse à des données de maxima pris site par site (Schlather, 2002). Des approches plus récentes reposant sur **les processus de Pareto** cherchent à caractériser **des dépassements au-delà d'un seuil** ce qui fournit des simulations à l'échelle de l'événement (Ferreira & de Haan, 2014). Toutefois, il n'y a pas, à ma connaissance, de proposition de générateur stochastique spatial intégrant des structures de dépendances spécifiques pour les événements extrêmes (Ailliot *et al.* , 2015). De plus, la région méditerranéenne connaît des épisodes extrêmes très localisés dans le temps et dans l'espace. Il est donc indispensable d'être en mesure de simuler, dans un même champ, **des zones avec des événements ordinaires et des zones avec des événements extrêmes**. Néanmoins, il n'existe pas actuellement d'approches permettant de simuler des transitions dans les structures de dépendance entre ces deux types d'événements dans le cadre spatial ou spatio-temporel (Ailliot *et al.* , 2015).

L'utilisation **de données sur grille** (télédétection ou simulations de modèles à base physique) peut

servir à développer des approches semi-paramétriques et à compléter l'information fournie par le réseau d'observations. **Les réanalyses** sont produites en fusionnant les deux types d'information. Par exemple, Era-Interim est une réanalyse globale d'une résolution spatiale d'environ 80 km^2 et à pas de temps tri-horaire basée sur un modèle de prévision (Dee *et al.* , 2011). À l'échelle de la France, des réanalyses horaires de résolution spatiale de 1 km^2 ont été élaborées en s'appuyant sur des données radar (Tabary *et al.* , 2012; Delrieu *et al.* , 2014). Or, les techniques utilisées pour obtenir les réanalyses - l'assimilation ou le krigeage - exigent que les données soient disponibles sur les mêmes périodes ce qui limitent leur applicabilité. **Les approches de correction de biais et de descente d'échelle** offrent une autre façon de combiner l'information des données sur grille et du réseau d'observations. Ces approches cherchent à corriger et / ou augmenter la résolution des simulations des modèles de climat globaux pour étudier les effets du changement climatique selon différents scénarios sur des périodes futures (Ayar *et al.* , 2015). À ma connaissance, ces approches sont encore peu appliquées dans le but d'obtenir une reconstruction passée des séries d'observations, en particulier à la résolution infra-journalière. Concernant les événements extrêmes, une approche semi-paramétrique a été proposé dernièrement pour générer des tempêtes de vagues, i.e. des champs spatio-temporels extrêmes (Chailan *et al.* , 2017). De plus, Palacios-Rodriguez *et al.* (2018) ont montré les liens entre cette méthode et les processus de Pareto. La technique consiste à définir des tempêtes à partir d'une fonctionnelle de risque, e.g. le maximum du champ dépasse une certaine valeur seuil, et à les rendre plus extrême par un facteur d'échelle tiré d'une loi de Pareto. Bien que les tempêtes obtenues satisfassent des propriétés théoriques, il reste encore de nombreuses questions, par exemple **comment choisir les formes des tempêtes et comment perturber les formes observées**, pour en faire des scénarios utilisables en pratique.

3.3 Structuration des activités

Les stratégies pour répondre à l'objectif principal de mon projet ainsi que les questions de recherche, identifiées à partir de l'état de l'art, qui y sont associées et auxquelles les pistes de recherche proposées chercheront à répondre sont organisées autour des deux objectifs spécifiques.

3.3.1 Générateur stochastique spatial de conditions météorologiques

Je propose ici trois stratégies pour répondre à l'objectif spécifique (OS.1) concernant la génération de scénarios multi-variables en rive sud de la Méditerranée pour l'étude des ressources en eau. Ces stratégies sont complémentaires car elles reposent sur des approches faisant des hypothèses différentes.

Questions de recherche associées

(Q.1) Comment définir des types de temps adaptés pour la gestion des ressources en eau en rive sud de la Méditerranée ?
(Q.2) Comment adapter les méthodes existantes pour prendre en compte l'influence de l'orographie au contexte sud méditerranéen ?
(Q.3) Comment calibrer des processus gaussiens en présence de distribution non-gaussienne avec une procédure intuitive semblable au variogramme empirique ?
(Q.4) Comment adapter les approches multivariées à la résolution infra-journalière et au contexte méditerranéen ? Est-ce que les modèles théoriques de fonction de covariance restent plausibles ou les hypothèses sous-jacentes sont-elles trop fortes ?

(Q.5) Est-ce que les approches de correction de biais et de descente d'échelle existantes fournissent des reconstructions passées acceptables des séries d'observations ? Quel type de modifications faut-il y apporter pour les adapter à l'échelle infra-journalière ?

(Q.6) Comment extraire des données sur grille de l'information sur les structures de dépendance sans faire d'hypothèse paramétrique afin de combiner deux sources d'information disponibles sur des périodes différentes ?

Stratégie basée sur les processus gaussiens

Cette première stratégie repose sur une classification des pas de temps en types de temps au sein desquels un processus gaussien modélise les structures de dépendance spatio-temporelle multivariée. Deux alternatives sont considérées, l'une paramétrique et l'autre non-paramétrique, pour modéliser le fonction de covariance croisée. Des lois de probabilité caractérisent les intensités des variables hydrométéorologiques pour chaque type de temps avec une prise en compte des effets orographiques. Les différentes étapes de la stratégie sont détaillées ci-dessous.

i Définition des types de temps

Afin d'introduire des connaissances a priori plus facilement, je propose d'employer des méthodes de classification non-supervisée, i.e. *clustering*, plutôt que des méthodes latentes telles que les modèles de Markov cachés. Il faudra inclure l'information provenant de plusieurs variables hydrométéorologiques pertinentes pour la gestion des ressources en eau et déterminer la façon la plus judicieuse de les utiliser. En particulier, la direction principale du vent est souvent importante - par exemple, en Tunisie, venant du sud , il apporte la chaleur du Sahara et potentiellement de l'humidité provenant de la Méditerranée, alors que du nord-ouest, il apporte le froid des montagnes. Il s'agit d'une variable particulière qui prend ses valeurs sur le cercle unitaire. La température de l'air pourrait servir d'indicateur de la saison et guider les transitions entre les types de temps au lieu de découper l'année en saisons. Il est sans doute approprié de considérer ces deux variables au niveau synoptique plutôt que provenant d'observations in situ. Néanmoins, des caractéristiques du motif spatial de la pluie apporteraient certainement de l'information utile à la classification en types de temps (Leblois, 2012).

ii Prise en compte de l'orographie

L'influence de l'orographie, au travers par exemple des caractéristiques de paysages sus-mentionnées, peut varier d'un type de temps à l'autre. Il faut, d'une part, calibrer des paramètres - résolution spatiale du modèle numérique de terrain, taille de la fenêtre de voisinage - qui affectent la précision des caractéristiques. D'autre part, il faut identifier comment les lois de probabilité des variables hydrométéorologiques varient en fonction de ces caractéristiques, i.e. par le biais d'un ou des paramètres de la loi ou via un facteur d'échelle comme dans l'approche régionale (Carreau *et al.* , 2017). Cela nécessite donc d'ajuster des lois de probabilité à chacune des variables hydrométéorologiques.

iii Ajustement de processus gaussiens multivarié

Au sein de chaque type de temps, la dépendance spatio-temporelle multivariée pourra être modélisée à l'aide d'un processus gaussien ce qui permettra d'obtenir des scénarios spatialisés. Toute la difficulté réside dans l'identification d'un modèle théorique de la fonction de covariance croisée. Je propose d'utiliser le modèle flexible et non-trivial développé dans Bourotte *et al.* (2016) qui semble prometteur. Il faudra, dans un premier temps, adapter la procédure de calibration pour tenir compte des configurations instrumentales. Puis, dans un deuxième temps, la capacité du modèle à reproduire les motifs de dépendance inter-variables et spatio-temporels en région sud méditerranéenne devra être évaluée.

Il sera aussi intéressant d'élaborer et de tester une nouvelle méthode de calibration des processus

gaussiens qui s'appuie sur le variogramme empirique, comme l'approche classique en géostatistique, mais estimé à l'aide du coefficient de correlation τ de Kendall. En effet, le τ de Kendall est invariant sous transformation monotone et permet donc d'estimer la force de la dépendance en présence de distribution non-gaussienne. J'ai déjà testé cette méthode de calibration sur des données synthétiques dans le cadre spatial. Il sera ensuite nécessaire d'évaluer son applicabilité à des données réelles et de l'étendre au cadre spatio-temporel et multivarié.

Une autre piste consiste à développer une approche non-paramétrique pour la fonction de covariance croisée qui s'appuie sur des fonctions de base (Cressie & Johannesson, 2008). Ce type d'approches autorise des structures de dépendance très flexibles - par exemple, présentant de la non-stationnarité et de l'anisotropie. Il faudra explorer les extensions possibles au cadre multivarié.

iv Descente d'échelle temporelle

Dans le cas où la base de données à l'échelle infra-journalière souhaitée est sur une période trop courte, le générateur pourra être calibré à l'échelle journalière et couplé à une technique de descente d'échelle temporelle. À partir de l'approche proposée dans Carreau *et al.* (2019), il sera nécessaire d'étudier des adaptations au cadre multivarié pour faire le choix des analogues.

Stratégie basée sur les modèles linéaires généralisés

Cette deuxième stratégie se propose d'adapter l'approche proposée dans Chandler (2014) au contexte sud méditerranéen et au pas de temps infra-journalier. La dépendance inter-variables est modélisée comme un produit de lois conditionnelles univariées. Chaque loi de probabilité conditionnelle repose sur un modèle linéaire généralisé pour lesquels des covariables introduisent différents effets, e.g. cycles saisonnier et diurne, orographie. La dépendance spatiale résiduelle peut être prise en compte via un processus gaussien. Les étapes de cette stratégie sont présentées ci-dessous.

i Décomposition de la loi de probabilité multivariée

La loi de probabilité jointe des variables $Y_1, Y_2, \ldots, Y_p$ peut toujours se décomposer en produit de lois de probabilité conditionnelles univariées :

$$\mathbb{P}(Y_1, Y_2, \ldots, Y_p) = \mathbb{P}(Y_1) \prod_{i=2}^{p} \mathbb{P}(Y_i | Y_{i-1}, \ldots, Y_1) \tag{3.1}$$

En se basant sur le graphe de dépendance inter-variables de Chandler (2014) ci-dessous, Fig. 3.1, les lois conditionnelles des variables hydrométéorologiques peuvent être simplifiées en ne conditionnant que selon certaines variables.

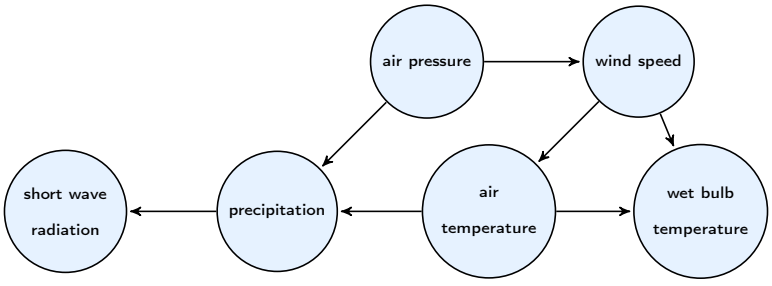


Figure 3.1 – *Graphe de dépendance inter-variables.*

ii Choix des lois de probabilité conditionnelle

Pour chacune des variables hydrométéorologiques, il faut choisir une loi de probabilité qui sera dépendante des variables conditionnantes via un ou plusieurs de ses paramètres. Pour un modèle linéaire classique, la loi de probabilité est gaussienne. Pour les modèles linaires généralisés, beaucoup d'autres lois de probabilité peuvent être utilisées, par exemple la loi Gamma pour des variables positives ou la loi Binomiale pour modéliser des occurrences telles que "pluie" versus "non-pluie" (McCullagh & Nelder, 1989).

iii Sélection des covariables

Outre les variables hydrométéorologiques conditionnantes de l'Eq. (3.1), d'autres covariables peuvent être ajoutées afin d'apporter de l'information complémentaire. D'une part, des covariables contenant de l'information géographique, e.g. les coordonnées des sites et les caractéristiques de paysages, peuvent être employées. Le cycle saisonnier peut être modélisé en ayant recours à des covariables de la forme :

$$\cos\left(\frac{2\pi d}{k}\right) \quad \sin\left(\frac{2\pi d}{k}\right) \tag{3.2}$$

où d est le jour de l'année associé au pas de temps considéré et k est la période, e.g. $k = 366$ donne une période annuelle. Pour modéliser le cycle diurne, la période k doit être ajustée pour osciller sur la journée. La persistence peut être prise en compte à l'aide d'effets retards, e.g. la valeur observée de la variable hydrométéorologique sur une période précédente. Enfin, des covariables à basse résolution spatiale, e.g. sur la circulation atmosphérique, peuvent aussi être intégrées.

iv Dépendance spatiale

On peut définir des résidus des modèles linéaires généralisés de sorte à ce qu'ils soient, au moins approximativement, normalement distribués (Chandler, 2014). On peut alors employer des processus gaussiens pour modéliser la dépendance spatiale, si nécessaire.

Stratégie basée sur les données sur grille

Cette troisième stratégie se distingue en ce qu'elle s'appuie entièrement sur des données sur grille pour capturer les structures de dépendance de façon non-paramétrique. L'ambition est d'étendre les séries d'observations des différentes variables hydrométéorologiques dans le temps et dans l'espace.

Afin de reconstruire les séries d'observations sur une période passée où elles ne sont pas disponibles, je propose d'utiliser des approches de correction de biais et de descente d'échelle. Ces approches cherchent à établir une relation statistique entre des simulations à basse résolution spatiale générées par les modèles de climat globaux et des données à plus haute résolution spatiale. Le but, le plus souvent, est d'augmenter la résolution spatiale des simulations sur des périodes futures à l'aide de la relation statistique estimée sur une période historique. Ici, au contraire, ces approches serviront à augmenter la résolution spatiale des simulations sur une période passée. Il faudra adapter les approches existantes pour le cas multivarié, e.g. Cannon (2018), à la résolution infra-journalière.

Pour pallier à la faible densité du réseau d'observations, on peut tirer profit de simulations à haute résolution spatiale de modèles de climat régionaux, lorsque disponibles sur la zone d'étude, pour interpoler dans l'espace. Comme ce type de simulations est coûteuse en temps de calcul, elles ne sont pas forcément disponibles exactement sur la même période que les observations. Une piste possible consistera à extraire des motifs de dépendance des simulations de façon non-paramétrique - par exemple, à l'aide des fréquences empiriques - afin de les utiliser pour interpoler spatialement. Ces motifs de dépendance devront être associés à des lois de probabilité pour les intensités des variables hydrométéorologiques qui pourront être ajustées à partir du réseau d'observations. La génération de scénarios pourra être obtenue à partir d'une approche de type analogues (Yiou, 2014).

Comparaison des stratégies

Les trois stratégies pour générer des scénarios multi-variables seront comparées en termes de critères statistiques et en termes d'impact pour l'étude des ressources en eau. Parmi les critères statistiques à considérer, il faut inclure la reproduction de la distribution des intensités, des dépendances spatiales, temporelles et inter-variables. Pour la validation des scénarios en termes d'impact, je prévois de vérifier que les scénarios permettent de restituer le bilan d'énergie dans les transferts sol - plante - atmosphère (Boulet *et al.* , 2015).

3.3.2 Générateur stochastique spatial de précipitation intégrant des phénomènes extrêmes

Je propose également deux stratégies pour répondre à l'objectif spécifique (OS.2) concernant la génération de scénarios de précipitation intégrant des événements extrêmes pour l'étude des risques naturels. Les stratégies diffèrent au niveau de la catégorie de scénarios envisagée : **la stratégie événementielle** considère des scénarios composés uniquement d'événements extrêmes réalistes alors que **la stratégie continue** s'intéresse à des scénarios comprenant toute la gamme des valeurs d'événements, extrêmes ou non.

Questions de recherche associées

(Q.7) Comment définir un type de temps identifiant les événements extrêmes non seulement en termes d'intensité mais aussi en termes de motifs spatiaux ?
(Q.8) Comment simuler des champs contenant des zones non-extrêmes et des zones extrêmes avec une transition dans la structure de dépendance ?
(Q.9) Comment mettre en oeuvre une approche non-paramétrique de rééchantillonnage de type noyau pour le motif de dépendance des événements extrêmes, i.e. en perturbant légèrement les motifs observés ?
(Q.10) Quels sont les apports d'une prise en compte explicite de la structure de dépendance spatio-temporelle des événements extrêmes ?

Stratégie continue

Je vise ici à proposer un générateur spatial de précipitation continu, i.e. qui comprend toute la gamme des valeurs d'événements - secs, ordinaires et extrêmes. La première étape consistera à établir un générateur de référence en s'appuyant sur les propositions existantes dans la littérature (Ailliot *et al.* , 2015; Baxevani & Lennartsson, 2015). La deuxième étape cherchera à modifier le générateur de référence en employant des approches provenant de la théorie des valeurs extrêmes. Enfin, les différents modèles considérés seront évalués et comparés.

i Générateur spatial de précipitation de référence

Étant donné la forte variabilité du climat méditerranéen, il me semble nécessaire de regrouper les pas de temps en types de pluie. La distribution de la pluie en termes d'intensité et en termes de répartition spatiale devra être identique au sein de chaque type de pluie. Tout comme pour les types de temps décrits précédemment, de l'information sur la circulation atmosphérique, notamment sur la force et la direction du vent, ainsi que sur le motif spatial de la pluie pourront servir à définir les types de pluie.

La distribution de la pluie, conditionnellement au type de pluie, doit modéliser les transitions entre les zones avec et sans pluie. Une approche répandue consiste à avoir recours à un processus gaussien spatio-temporel latent qui, combiné à des valeurs de seuils, définira les zones pluvieuses et sèches. Lorsque le champ gaussien se trouve au-dessus du seuil, donc en zone pluvieuse, une transformation est appliquée pour s'ajuster aux lois de probabilités des intensités de pluie (Allard & Bourotte, 2015; Baxevani & Lennartsson, 2015).

ii Générateur spatial avec prise en compte des événements extrêmes

Il s'agira ici d'adapter ou de développer des alternatives pour la modélisation conditionnelle du ou des types pluvieux correspondant aux événements extrêmes. Au sein de ce type pluvieux, en plus des événements secs et ordinaires, se trouvent des événements extrêmes. Il faut donc être en mesure de faire des transitions entre ces trois types d'événements. Une première piste consiste, à partir du processus gaussien latent, à établir un deuxième seuil identifiant la zone pluvieuse extrême et à modifier la loi de probabilité des intensités en ayant recours à la loi de Pareto généralisée qui est justifiée théoriquement. Une deuxième piste repose sur une approche qui emploie des processus max-stables ou max-stables inversés pour modéliser des séries de dépassements (Thibaud *et al.* , 2013). Pour ces deux premières pistes, il y a un seul processus, soit gaussien, soit max-stable ou max-stable inversé, qui porte toute la structure de dépendance pour les trois types d'événements - secs, ordinaires et extrêmes.

Je chercherai ensuite à développer et évaluer des alternatives où il y a une transition entre les types d'événements au niveau de la structure de dépendance spatiale. Une première piste dans ce sens pourra s'appuyer sur des constructions qui combinent plusieurs processus sous forme de mélange, soit de façon classique avec une somme pondérée soit en prenant leur maximum avec une puissance ce qui est permet de conserver la propriété de max-stabilité et donc d'être compatible avec la théorie des valeurs extrêmes (Carreau & Toulemonde, 2018). Une deuxième piste pourra exploiter les constructions à partir de fonctions de base. En effet, lorsque ces fonctions sont combinées avec une variable aléatoire gaussienne, on obtient un processus gaussien (Cressie & Johannesson, 2008) alors que si une variable aléatoire α-stable est utilisée, un processus max-stable en résulte (Reich & Shaby, 2012).

iii Évaluation et comparaison des générateurs spatiaux de précipitation

Le générateur spatial de référence sera comparé avec les différentes alternatives en simulant des chroniques de champs de pluie. Je vérifierai, d'une part, si les champs simulés reproduisent les propriétés statistiques des champs observés, e.g. de nombreux critères spatio-temporels sont analysés dans Baxevani & Lennartsson (2015). D'autre part, je chercherai à évaluer l'apport des différents générateurs en termes d'impact hydrologique. Pour ce faire, les simulations de pluie alimenteront des modèles pluie-débit et le débit simulé sera comparé au débit observé et / ou au débit simulé lorsque le modèle pluie-débit est alimenté par les pluies observées.

Stratégie événementielle

L'objectif est, à partir de données sur grille, de développer une approche semi-paramétrique permettant de générer des tempêtes, i.e. uniquement des événements extrêmes. Ce type de scénarios peut servir à des analyses de risque notamment pour les inondations en milieu urbain. Je propose de poursuivre les travaux de Palacios-Rodriguez *et al.* (2018) dans lesquels les lois marginales, i.e. des intensités de précipitations, sont modélisées de façon semi-paramétrique et les formes des tempêtes qui sont caractérisées par les structures de dépendance spatio-temporelle sont extraites des formes observées. L'idée est d'identifier des descripteurs non-paramétriques de la structure de dépendance spatiale. Ces descripteurs pourront fournir une base pour définir des similarités entre les formes de tempête ce qui permettra, d'une part, de sélectionner des formes particulières et d'autre part, de géné-

rer des perturbations dans les formes observées à l'aide d'une approche de type analogues (Yiou, 2014).

i Descripteurs non-paramétriques de la structure de dépendance spatiale
Les descripteurs conventionnels des structures de dépendance spécifiques aux événements extrêmes tels que le coefficient extrêmal, bien que non-paramétriques, reposent sur l'hypothèse de max-stabilité. De plus, ils sont, le plus souvent, basés uniquement sur des dépendances bivariées, voire trivariées (Erhardt & Smith, 2012). À partir des descripteurs utilisés en géostatistique, je chercherai à proposer des descripteurs qui puissent rendre compte de l'anisotropie, i.e. du fait que la dépendance se comporte de façon différente selon les directions. Par ailleurs, des descripteurs de l'asymmétrie dans la structure de dépendance sont encore, à ma connaissance, pratiquement inexistants.

ii Évaluation des scénarios
Les scénarios de tempêtes pourront être évalués en termes de mesures de risques telles que des niveaux de retour régionaux associés à de longues périodes de retour (Thibaud *et al.* , 2013). De plus, ils pourront être évalués en termes d'impact en alimentant des modèles hydrauliques pour les inondations en milieu urbain (Guinot *et al.* , 2017).

3.4 Milieux modèles

Dans un premier temps, je prévois des développements en Tunisie - un cas emblématique pour les enjeux sociétaux et scientifiques. De plus, les deux zones d'étude du LMI NAÏLA représentent un panel de situations représentatives de la rive sud de la Méditerranée. Dans un deuxième temps, des travaux seront envisagés en collaboration avec le LMI TREMA dans le Haut-Atlas au Maroc. Il sera aussi possible de considérer par la suite d'autres pays du Sud tels que l'Inde. En parallèle, je continuerai des développements dans l'arrière-pays montpellierain - incluant le massif des Cévennes - qui est sujet à des épisodes de pluies extrêmes. Ceci permet de développer et tester des méthodologies dans un cadre plus favorable avant de les adapter au sud de la Méditerranée.

Les deux zones d'étude du LMI NAÏLA en Tunisie sont marquées par la chaîne de la dorsale tunisienne : le bassin du Lebna au Cap Bon (210 km^2) en climat sub-humide et la plaine et la partie amont du Merguellil (400 km^2 et 1200 km^2 respectivement) en climat semi-aride. L'ORE OMERE et le SO du Merguellil collectent des données sur plusieurs variables hydrométéorologiques sur les deux zones : les précipitations, la vitesse et la direction du vent, l'humidité de l'air, la température de l'air et le rayonnement. Le pas de temps des observations varie entre 30 min et la journée et les périodes d'observations varient entre 5 et 10 ans. En complément, une base de données de cumuls journaliers de précipitations est fournie par la Direction Générale des Ressources en Eau sur la partie amont du Merguellil. Côté français, Météo-France met à disposition les observations provenant de ses réseaux de pluviomètres et de pluviographes et des données radar à une résolution spatiale de 1 km.

Des données sur grille provenant de modèles de climat régionaux sont disponibles sur toutes les zones (des simulations "sur mesure" de WRF à une résolution spatiale de 3 km sur la plaine et l'amont du Merguellil et des simulations du programme EuroCordex à une résolution spatiale de 12 km sur l'ensemble des zones).

Chapitre 4

Bibliographie

Ailliot, P., Allard, D., Monbet, V., Naveau, P. Stochastic weather generators : an overview of weather type models. *J. de la Société Française de Statistique*, 2015, **156**(1), 101–113.

Alazard, M., Leduc, C., Travi, Y., Boulet, G., & Salem, A. B. Estimating evaporation in semi-arid areas facing data scarcity : example of the El Haouareb dam (Merguellil catchment, Central Tunisia). *J. of Hydrology : Regional Studies*, 2015, **3**, 265–284.

Allard, D., & Bourotte, M. Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process. *Stoch. Environ. Res. Risk. Assess.*, 2015, **29**(2), 453–462.

Arnaud, P., Lavabre, J., Sol, B., & Desouches, C. Cartographie de l'aléa pluviographique de la France. *La Houille Blanche*, 2006, **5**, 102–111.

Ayar, P. V., Vrac, M., Bastin, S., Carreau, J., Déqué, M., & Gallardo, C. Intercomparison of statistical and dynamical downscaling models under the EURO- and MED-CORDEX initiative framework : present climate evaluations. *Climate Dynamics*, 2015, **46**(2), 1301–1329.

Baxevani, A., & Lennartsson, J. A spatiotemporal precipitation generator based on a censored latent Gaussian field. *Water Resour. Res.*, 2015, **51**(6), 4338–4358.

Benoit, L., Allard, D. & Mariethoz, G. Stochastic Rainfall Modeling at Sub-kilometer Scale. *Water Resour. Res.*, 2018, **54**(6), 4108–4130..

Bishop, C. *Pattern recognition and machine learning*. Springer-Verlag New York, 2006.

Boulet, G., Mougenot, B., Lhomme, J. P,, Fanise, P., Lili-Chabaane, Z., Olioso, A., Bahir, M., Rivalland, V., Jarlan, L., Merlin, O., Coudert, M., Er-Raki, S. & Lagouarde, J. P. The SPARSE model for the prediction of water stress and evapotranspiration components from thermal infra-red data and its evaluation over irrigated and rainfed wheat. *Hydrologic and Earth System Science*, 2015, **19**, 4653–4672.

Bourotte, M., Allard, D., & Porcu, E. A flexible class of non-separable cross-covariance functions for multivariate space–time data. *Spatial Statistics*, 2016, **18**, 125–146.

Braud, I., Ayral, P.A., Bouvier, C., Branger, F., Delrieu, G., Le Coz, J., Nord, G., Vandervaere, J.P., Anquetin, S., Adamovic, M., *et al.* . Multi-scale hydrometeorological observation and modelling for flash-flood understanding. *Hydrology and Earth System Sciences*, 2014, **18**(9), 3733–3761.

Cannon, A. Multivariate quantile mapping bias correction : an N-dimensional probability density function transform for climate model simulations of multiple variables. *Climate Dynamics*, 2018, **50**(2), 31–49.

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

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

Carreau, J.,Ben Mhenni, N., Huard, F. & Neppel, L. Exploiting the spatial pattern of daily precipitation in the analog method for regional temporal disaggregation. *J. of Hydrology*, 2019, **568**, 780–791.

Carreau, J., & Bouvier, C. Multivariate density model comparison for multi-site flood-risk rainfall in the French Mediterranean area. *Stoch. Environ. Res. Risk. Assess.*, 2016, **30**(6), 1591–1612.

Carreau, J., & Girard, S. Spatial extreme quantile estimation using a weighted log-likelihood approach. *Journal de la Société Française de Statistique*, 2011, **152**(3), 66–82.

Carreau, J., Naveau, P., & Neppel, L. Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation. *Water Resour. Res.*, 2017, **53**, 4407–4426.

Carreau, J., Naveau, P., & Sauquet, E. A statistical rainfall-runoff mixture model with heavy-tailed components. *Water Resour. Res.*, 2009, **45**, 1–11.

Carreau, J., Neppel, L., Arnaud, P., & Cantet, P. Extreme rainfall analysis at ungauged sites in the South of France : comparison of three approaches. *Journal de la Société Française de Statistique*, 2013, **154**(2), 119–138.

Carreau, J. & Toulemonde G. *Extra-Parametrized Extreme Value Copula : Extension to a Spatial Framework*. Proceedings of the 9th Workshop on Spatio-temporal modeling (METMA IX), Montpellier, 2018.

Carreau, J. & Vrac, M. Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resour. Res.*, 2011, **47**, 1–15.

Ceresetti, D., Ursu, E., Carreau, J., Anquetin, S., Creutin, J.-D., Gardes, L., Girard, S., & Molinie, G. Evaluation of classical spatial-analysis schemes of extreme rainfall. *Natural hazards and earth system sciences*, 2012, **12**, 3229–3240.

Chailan, R., Toulemonde, G., & Bacro, J.-N. A semiparametric method to simulate bivariate space–time extremes. *Ann. Appl. Stat.*, 2017, **11**(3), 1403–1428.

Chandler, R. *Rglimclim : A multisite, multivariate weather generator based on generalized linear models*. R package, 2014.

Coles, S. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

Cressie, N., & Johannesson, G. Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc. B*, 2008, **70** Part 1, 209–226.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., *et al.* . The ERA-Interim reanalysis : Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 2011, **137**(656), 553–597.

Delrieu, G., Wijbrans, A., Boudevillain, B., Faure, D., Bonnifait, L., & Kirstetter, P.-E. Geostatistical radar–raingauge merging : A novel method for the quantification of rain estimation accuracy. *Adv. in Water Resour.*, 2014, **71**, 110–124.

Erhardt, R. J., & Smith, R.L. Approximate Bayesian computing for spatial extremes. *Comp. Stat. & Data Ana.*, 2012, **56**(6), 1468–1481.

Ferreira, A., & de Haan, L. The generalized Pareto process ; with a view towards application and simulation. *Bernoulli*, 2014, **20**, 1717–37.

Garavaglia, F., Gailhard, J., Paquet, E., Lang, M., Garçon, R., Bernardara, P. Introducing a rainfall compound distribution model based on weather patterns sub-sampling. *Hydrology and Earth System Sciences*, 2010, **14**, 951–964.

Goodfellow, I., Bengio, Y., & Courville, A. *Deep learning*. MIT press, 2016.

Guinot, V., Sanders, B. F. & Schubert, J. E. Dual integral porosity shallow water model for urban flood modelling. *Adv. in Water Resour.*, 2017, **103**, 16–31.

Hosking, J. R. M., & Wallis, J. R. *Regional Frequency Analysis : An Approach Based on L-Moments*. Cambridge Univ. Press, Cambridge, UK, 2005.

Kleiber, W., Katz, R. W., & Rajagopalan, B. Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour. Res.*, 2012, **48**(1).

Leblois, E. *Le bassin versant, système spatialement structuré et soumis au climat*. Habilitation à Diriger les Recherches, Université de Grenoble, 2012.

Li, C., Singh, V. P., & Mishra, A. K. Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water Resour. Res.*, 2012, **48**(3).

McCullagh, P., & Nelder, J. A. Generalised linear models 2nd edn. *Monographs on statistics and applied probabilitity*, 1989.

Monbet, V. *Markov Switching Multivariate Space Time model for weather variables*. Proceedings of the 9th Workshop on Spatio-temporal modeling (METMA IX), Montpellier, 2018.

Naveau, P., Huser, R., Ribereau, P., & Hannart, A. Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resour. Res.*, 2016, **52**(4), 2753–2769 .

Palacios-Rodriguez, F., Toulemonde, G., Carreau, J. & Opitz, T. *Space-time extreme processes. simulation for flash floods in Mediterranean France*. Proceedings of the 9th Workshop on Spatio-temporal modeling (METMA IX), Montpellier, 2018.

Reich, J. R., & Shaby, B. A. A hierarchical max-stable spatial model for extreme precipitation. *Ann. Appl. Stat.*, 2012, **6**(4), 1430–1451.

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

Tabary, P., Dupuy, P., L'Henaff, G., Guenguen, C., Moulin, L., Laurantin, O., Merlier, C., & Souberoux, J.-M. A 10-year (1997-2006) reanalysis of Quantitative Precipitation Estimation over France : methodology and first results. *IAHS-AISH publication*, 2012, 255–260.

Thibaud, E., Muzner, R. & Davison, A.C. Threshold modeling of extreme spatial rainfall. *Water Resour. Res.*, 2013, **49** 4633–4644.

Vrac, M. Multivariate bias adjustment of high-dimensional climate simulations : the Rank Resampling for Distributions and Dependences (R 2 D 2) bias correction. *Hydrology and Earth System Sciences*, 2018, **22**(6), 3175.

Williams, P. M. Using Neural Networks to Model Conditional Multivariate Densities. *Neural Computation*, 1996, **8**, 843-854.

Yiou, P. AnaWEGE : a weather generator based on analogues of atmospheric circulation. *Geosci. Model Dev.*, 2014, **7** 531–543.

Annexe A

Principaux travaux scientifiques

A.1 Apprentissage statistique et théorie des valeurs extrêmes

A.1.1 Des valeurs centrales aux valeurs extrêmes

Extremes (2009) 12:53–76
DOI 10.1007/s10687-008-0068-0

A hybrid Pareto model for asymmetric fat-tailed data: the univariate case

Julie Carreau · Yoshua Bengio

Received: 15 April 2007 / Revised: 2 July 2008 /
Accepted: 30 July 2008 / Published online: 30 August 2008

Abstract Density estimators that can adapt to asymmetric heavy tails are required in many applications such as finance and insurance. Extreme value theory (EVT) has developed principled methods based on asymptotic results to estimate the tails of most distributions. However, the finite sample approximation might introduce a severe bias in many cases. Moreover, the full range of the distribution is often needed, not only the tail area. On the other hand, non-parametric methods, while being powerful where data are abundant, fail to extrapolate properly in the tail area. We put forward a non-parametric density estimator that brings together the strengths of non-parametric density estimation and of EVT. A hybrid Pareto distribution that can be used in a mixture model is proposed to extend the generalized Pareto (GP) to the whole real axis. Experiments on simulated data show the following. On one hand, the mixture of hybrid Paretos converges faster in terms of log-likelihood and provides good estimates of the tail of the distributions when compared with other density estimators including the GP distribution. On the other hand, the mixture of hybrid Paretos offers an alternate way to estimate the tail index which is comparable to the one estimated with the standard GP methodology. The mixture of hybrids is also evaluated on the Danish fire insurance data set.

Keywords Mixture model · Fat-tailed data · Extreme quantiles · Generalized Pareto distribution

AMS 2000 Subject Classifications 62G07 · 62G32

J. Carreau ( ) · Y. Bengio
Dept. IRO, University of Montreal, Montreal, Canada
e-mail: carreau@iro.umontreal.ca

Y. Bengio
e-mail: bengioy@iro.umontreal.ca

1 Introduction

In this paper, we introduce a new non-parametric density estimator that can take into account asymmetry, multimodality and tail heaviness of the underlying density. Data exhibiting such characteristics can be found in application domains such as finance and insurance. By non-parametric, we mean an estimator whose complexity, that is the number of free parameters, grows with the size of the training set. Methods from extreme value theory (EVT) allow the estimation and the extrapolation of the tails of most distributions. These methods rely on the asymptotic behaviour of the underlying distribution. However, it is difficult to pinpoint at which level the asymptotics on which EVT is based kick in because it depends on the distribution we are dealing with. If we are interested in an area that is not sufficiently far away in the tail, EVT techniques might lead to an estimation bias. On the other hand, estimation of the full range of an heavy-tailed distribution is often useful. The non-parametric density estimator that we propose is able to provide approximation of the density in any area, whether the asymptotic assumptions of EVT are valid or not.

In finance, estimating the predictive distribution of the profit and loss (P&L) of a portfolio is central for portfolio and risk management. Finance practitioners most often use the so-called value-at-risk (VaR) which is a large quantile of the P&L distribution. However, the only information carried by the VaR is a bound on the lowest loss attainable with a given probability. It doesn't provide any information as how bad can things go below that bound. The Conditional VaR (CVaR) has been proposed to resolve this issue (Rockafellar and Uryasev 2002); it measures the expected loss given that the VaR has been exceeded. A density estimator providing a model of the tail of distribution is thus required to compute the CVaR. However, since several authors (Fama 1965; Mandelbrot 1963) have provided strong evidence for the presence of fat tails in stock returns data, this density estimator must be able to handle tail heaviness.

In insurance applications, companies are interested in modeling the distribution of the claims of their clients. Insurers cover losses that fall in a given interval, called the reinsurance layer, by resorting to reinsurance companies. This reinsurance layer is meant to protect against a range of large claims; therefore, good estimates of the tail of the claim distribution are needed to evaluate the probability of losses in the reinsurance layer. Claim distributions are also typically fat-tailed (McNeil 1997).

The mixture of hybrid Paretos is a new non-parametric density estimator that builds on EVT and non-parametric modelling. The hybrid Pareto is a smooth extension of the generalized Pareto distribution to the whole real axis. In a simulation study, benefits from using the mixture of hybrid Paretos with respect to other density estimators are highlighted. The approximation abilities of the proposed estimator in the tail area are shown to be equivalent to the so-called *Peaks over Thresholds* (PoT) method based on the generalized Pareto

distribution. We then evaluate comparatively the proposed estimator on a real data set, the Danish fire insurance data set.

2 Extreme value theory

EVT (Embrechts et al. 1997) has put forward sound mathematical methods to estimate the tail of univariate distributions. The so-called *Peaks over Thresholds* (PoT) method (Davison and Smith 1990) consists of fitting the generalized Pareto distribution to the exceedances of a random variable above a suitable threshold. This is justified by Pickands theorem (Pickands 1975) which states that, for most random variables, the distribution of the exceedances converges to a GP as the threshold tends to the right endpoint of the support of the underlying distribution.

The tail index of a distribution, usually denoted ξ, characterizes the tail heaviness of the distribution. In case of asymmetry, the upper and the lower tails could have different tail indexes. When $\xi > 0$, the distribution is heavy-tailed, that is the tail decreases at a polynomial speed like the Pareto, the α-stable or the Student t distributions. When $\xi = 0$, the distribution is light-tailed, that is the tail decreases exponentially fast; examples of such distributions are the Gaussian, the Exponential and the Log-Normal distributions. When $\xi < 0$, the tail is finite, this is the case for the Uniform and the Beta distributions. The rth moment of the distribution is finite if and only if $\xi < 1/r$. The GP distribution has two parameters, one of which is the tail index. Thus, fitting a GP distribution to the exceedances provides an estimator of the tail index of the underlying distribution.

2.1 Tail estimation

Let F be the d.f of Y and let $F_u(y) = P(Y - u \leq y | Y > u)$ be the distribution of the exceedances of Y given that Y exceeds the threshold u. We have the following relationship between F and F_u, $\forall y > 0$:

$$F(u + y) = F(u) + (1 - F(u))F_u(y). \tag{1}$$

The PoT approximation to the tail of F consists of replacing in Eq. 1 $F_u(y)$ with the GP distribution function and $1 - F(u)$ with the proportion of excesses in the sample.

The parameters of the GP can be estimated by maximizing the log-likelihood on the training data. Smith (1985) showed that maximum likelihood estimators (MLEs) exist when $\xi > -1$ and are asymptotically normal and efficient when $\xi > -1/2$. The probabilistic method of moments proposed by Hosking and Wallis (1987) is a viable alternative to MLE when $\xi > 0$. However, MLE allows for more flexible modelling such as temporal or covariate dependency. In the PoT method, the GP parameters are estimated on the normalized exceedances of Y over a given threshold. Smith (1987) showed

that in a number of cases, the approximation made by the GP introduces a nonnegligible bias in the MLEs.

2.2 Threshold selection

The choice of threshold above which the exceedances are used for inference of the GP parameters is subject to a bias-variance trade-off. If the threshold is too high, very few points enter in the estimation of the GP parameters, making the estimator subject to high variance. If the threshold is too low, the GP approximation of the tail will have a large bias since, according to Pickands theorem, convergence occurs as the threshold approaches the right endpoint of the distribution. Embrechts et al. (1997) suggest to estimate the tail index for various levels of threshold and to plot the tail index estimator against the level of threshold. The threshold is chosen in a region of the graph that shows stability of the tail index estimator. McNeil and Frey (2000) use a random threshold: the $k+1$ largest observation so that k observations are used in the estimation. The number k is chosen in a simulation study based on the Student t distribution. Danielsson et al. (2001) propose a bootstrap method that selects the threshold that minimizes the asymptotic mean squared error of the tail index estimator. Choulakian and Stephens (2001) introduced goodness-of-fit tests for the GP distribution that can be used for threshold selection. Dupuis (1998) developed a robust selection method based on optimal bias robust estimator ($OBRE$). This method assigns weights between 0 and 1 to each observation. The weights measure the model accuracy on each observation. The threshold is taken high enough so that all the exceedances receive a weight close to 1. Frigessi et al. (2002) propose a method similar to ours: a two-component mixture with a GP component located at 0 and a light-tail component. The proportion of the GP in the mixture increases with y and thus acts as a smooth threshold. In contrast, the model we propose uses hybrid Pareto components that have scaled GP tails. The number of components can be adapted to the data at hand and the threshold is defined implicitly as a function of the mixture parameters.

3 Hybrid Pareto distribution

A popular model in density estimation is the mixture of Gaussians. When the number of components is well chosen according to the number of observations, the mixture of Gaussians has nice convergence properties as a non-parametric estimator (see Priebe 1994 for instance). If the tail of the generative distribution is heavy, i.e. extreme observations can occur far away in the tail, good empirical results can often be obtained by considering a mixture of Gaussians in which one of the Gaussians has a very large standard deviation, that serves to capture the points far away. One disadvantage of this approach is that the density model will only account for observed extremes and will still underestimate the density of the upper tail (this is specially true for small

training sets). Another drawback is that by using symmetric components in the mixture, the lower tail tends to be overestimated. These two problems will be illustrated in the simulation study.

We circumvent those shortcomings by introducing the hybrid Pareto distribution and using it in a mixture model. In this way, we combine the advantages of non-parametric modeling for the bulk of the data and EVT in regions where data are too scarce. Using directly the GP as a component of a mixture model requires to set the location of the GP a priori (as in Frigessi et al. 2002). It is thus difficult to integrate threshold selection within mixture learning.

3.1 Derivation

The hybrid Pareto distribution is a smooth extension of the GP to the whole real axis. This new distribution is built by stitching a GP tail to a Gaussian, while enforcing continuity of the resulting density and of its derivative. The threshold is then defined as the junction point of the Gaussian and the GP and is computed implicitly as a function of the hybrid parameters. Let α be the threshold and let $f_{\mu;\sigma}(y) = 1/(\sqrt{2\pi}\sigma)\exp(-(y-\mu)^2/(2\sigma^2))$ be the Gaussian density function with parameters μ and σ. The GP density located above α is given in Eq. 2 where $y \geq \alpha$ when $\xi \geq 0$ and $\alpha \leq y \leq \alpha - \beta/\xi$ when $\xi < 0$.

$$g_{\xi;\beta}(y-\alpha) = \begin{cases} \frac{1}{\beta}\left(1+\frac{\xi}{\beta}(y-\alpha)\right)^{-1/\xi-1} & \text{if } \xi \neq 0, \\ \frac{1}{\beta}e^{-\frac{y-\alpha}{\beta}} & \text{if } \xi = 0. \end{cases} \tag{2}$$

There are initially five parameters: α, μ, σ, ξ and β. Since two continuity constraints must be satisfied, there are thus three free parameters. We set ξ, μ and σ as the free parameters and we let α and β be functions of these. This choice is intuitively reasonable since those parameters are easily interpretable: ξ as measuring the tail heaviness of the distribution, μ its location and σ its spread.

We show how the equations for the hybrid Pareto are derived for the case $\xi > 0$, the other cases are similar and need only minor adjustments. The continuity constraint on the density at α means that $f_{\mu;\sigma}(\alpha) = g_{\xi;\beta}(0)$ which gives:

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = \frac{1}{\beta} \Leftrightarrow \exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = \frac{\sqrt{2\pi}\sigma}{\beta}. \tag{3}$$

Continuity of the derivative of the density at α means that $f'_{\mu;\sigma}(\alpha) = g'_{\xi;\beta}(0)$, which yields:

$$-\frac{(\alpha-\mu)}{\sqrt{2\pi}\sigma^3}\exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = -\frac{(1+\xi)}{\beta^2}. \tag{4}$$

Combining Eqs. 3 and 4, we get that:

$$\frac{1+\xi}{\beta} = \frac{\alpha-\mu}{\sigma^2} \Leftrightarrow \alpha = \mu + \frac{\sigma^2}{\beta}(1+\xi). \tag{5}$$

By replacing α in Eq. 3 by the expression obtained in Eq. 5 and re-arranging we get:

$$\frac{(1+\xi)^2}{2\pi} = \frac{\sigma^2(1+\xi)^2}{\beta^2} \exp\left(\frac{\sigma^2(1+\xi)^2}{\beta^2}\right). \tag{6}$$

To solve Eq. 6, we make use of the Lambert W function: given an input z, $w = W(z)$ is such that $z = we^w$. We use a numerical algorithm of order four to find the zero of $z - we^w$ (Corless et al. 1996). We let $z = (1+\xi)^2/2\pi$ in Eq. 6 and thus we have an expression for β:

$$W(z) = \frac{\sigma^2(1+\xi)^2}{\beta^2} \Leftrightarrow \beta(\xi,\sigma) = \frac{\sigma(1+\xi)}{\sqrt{W(z)}}. \tag{7}$$

To obtain an expression for α in terms of the free parameters, we replace β in Eq. 5 by its expression of Eq. 7:

$$\alpha(\xi,\mu,\sigma) = \mu + \sigma\sqrt{W(z)}. \tag{8}$$

Let $\theta = (\xi,\mu,\sigma)$ be the parameter vector of the hybrid Pareto. The hybrid Pareto density function is given by:

$$h_\theta(y) = \begin{cases} \frac{1}{\gamma} f_{\mu;\sigma}(y) & \text{if } y \le \alpha, \\ \frac{1}{\gamma} g_{\xi;\beta}(y-\alpha) & \text{if } y > \alpha \end{cases}$$

where γ is the appropriate re-weighting so that the density integrates to one and is given by:

$$\gamma(\xi) = 1 + \frac{1}{2}\left(1 + Erf\left(\sqrt{W(z)/2}\right)\right),$$

where $Erf(\cdot)$ is the error function $Erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$, which can be readily approximated numerically to high precision in standard ways. Figure 1 illustrates the density and the log-density of the hybrid Pareto distribution.

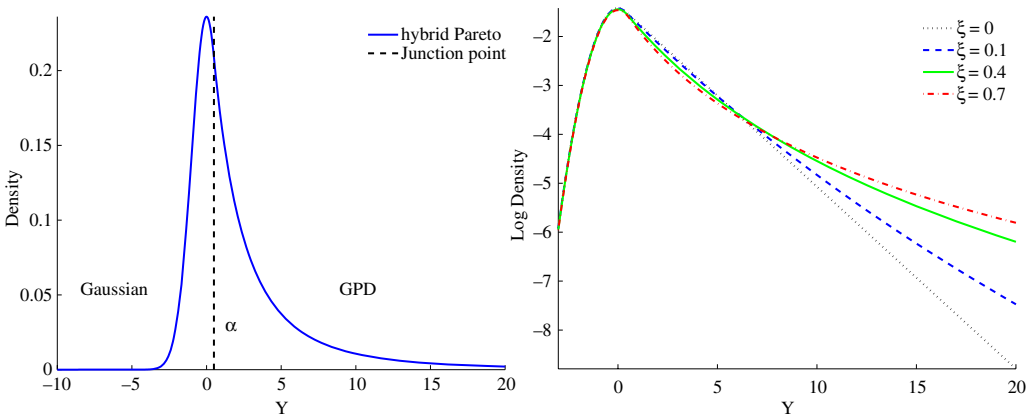


Fig. 1 *Left panel* hybrid Pareto density with parameters $\xi = 0.4$, $\mu = 0$ and $\sigma = 1$. *Right panel* hybrid Pareto log-density for various tail parameters and in all cases $\mu = 0$ and $\sigma = 1$.

3.2 Properties of the hybrid Pareto

Let $H_\theta(y)$ be the d.f. of the hybrid Pareto. When $y > \alpha$, we have the following relationship with $G_{\xi;\beta}(y)$, the d.f. of the GP:

$$1 - H_\theta(y) = \frac{1}{\gamma}(1 - G_{\xi;\beta}(y - \alpha)).$$

The tail of the hybrid Pareto is thus the same as the GP tail apart from the multiplicative factor $1/\gamma$. It can be seen that $P(Y > \alpha) = 1/\gamma$ and thus it acts as $P(Y > u) = 1 - F(u)$ in Eq. 1. Therefore, the tail approximation properties of the GP transfer to the hybrid Pareto.

Besides maximum likelihood estimators of θ, it is possible to develop estimators based on an initial estimate of ξ and quantiles of the hybrid Pareto distribution. The Hill estimator (Embrechts et al. 1997) can be used to estimate the tail index when $\xi > 0$. This estimator is based on the kth largest observations of the data set. Moments estimators (Hosking and Wallis 1987) can be used for more general values of ξ. These estimators require to select a threshold as well. Since our purpose is to obtain a rough initial estimate of ξ, we simply fix the threshold such that the 10% percent largest observations are used to compute the estimator. Two quantiles have a particularly simple form:

$$P(Y \leq \alpha) = \Phi\left(\sqrt{W(z)}\right) / \left[1 + \Phi\left(\sqrt{W(z)}\right)\right] \tag{9}$$

$$P(Y \leq \mu) = 1/\left[2\left(1 + \Phi\left(\sqrt{W(z)}\right)\right)\right], \tag{10}$$

where $\Phi(\cdot)$ is the standard Normal d.f. and $z = (1+\xi)^2/2\pi$. Given $\hat{\xi}$, an estimate of ξ, we can compute the sample quantiles corresponding to levels of equations Eqs. 9 and 10 and those could serve as initial estimators of α and μ respectively. An estimator for σ could be find by solving Eq. 8 for σ. These quantiles estimators are then used as a starting point for maximum likelihood estimation.

3.3 Preliminary Monte Carlo simulation

We tested the asymptotic behaviour of MLEs of the hybrid Pareto parameters with a Monte Carlo simulation. We generated a training set $\mathcal{D}_n$ and a test set $\mathcal{D}_l$ according to a hybrid Pareto of parameters θ. Increasing sizes of training set n are used while the test set size l is fixed at 10,000. Let $\hat{\theta}_n$ be the MLE of θ estimated on $\mathcal{D}_n$. We evaluated the goodness of fit of the estimated density $h_{\hat{\theta}_n}(\cdot)$ by measuring the log-likelihood relative to the generative distribution on the test set:

$$\mathcal{R}_l\left(\hat{\theta}_n; \theta\right) = \frac{1}{l}\sum_{i=1}^{l}(\log h_\theta(y_i) - \log h_{\hat{\theta}_n}(y_i)).$$

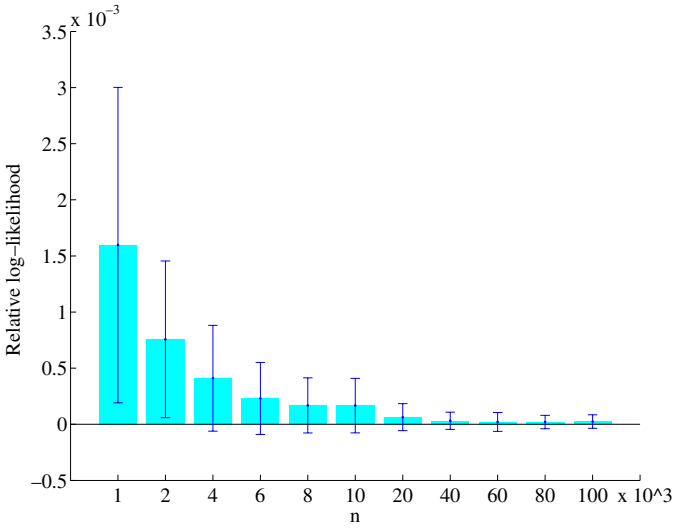


Fig. 2 Log-likelihood relative to the data generative density on the test set for hybrid Pareto MLE as the training set size n increases and $\theta = (0.7, 0, 1)$.

This performance criterion is equivalent to the empirical Kullback-Leibler divergence up to a scaling factor. The smaller the $\mathcal{R}_l(\hat{\theta}_n; \theta)$ is, the better the estimator is performing. For each value of n, 100 training sets are generated and MLE $\hat{\theta}_n^i$ is computed on each training set. We are thus able to report the average relative likelihood $\left(1/100 \sum_{i=1}^{100} \mathcal{R}_l(\hat{\theta}_n^i; \theta)\right)$ along with confidence intervals and squared bias and variance of the MLE. The average relative log-likelihood on the test set is illustrated in the bar plot of Fig. 2 when the generative parameters are $\theta = (0.7, 0, 1)$. On the top of each bar, a confidence interval of level 95% is drawn. The relative log-likelihood is steadily decreasing to zero and the confidence intervals become narrower as n increases, as expected. Squared bias and variance for each of the estimated hybrid Pareto parameters are shown in Fig. 3. The squared bias fluctuates a little bit for small values of n but then decreases. Variance shows a clear decreasing trend.

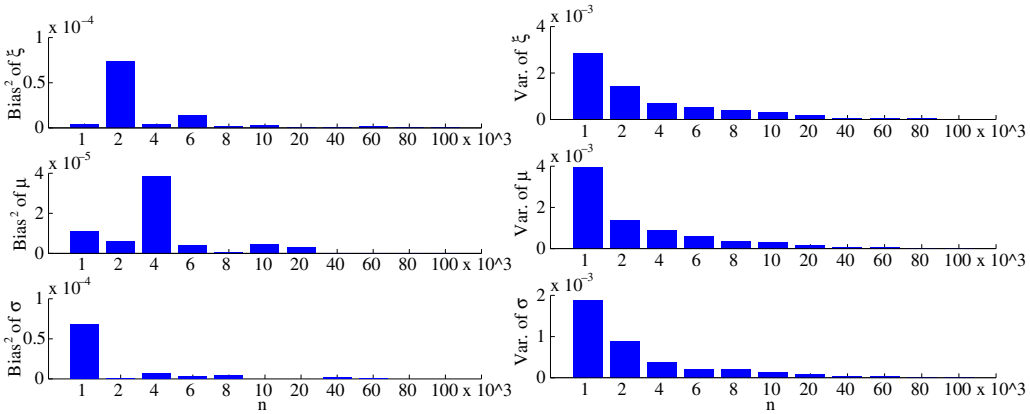


Fig. 3 For $\theta = (0.7, 0, 1)$, as the training set size n increases, squared bias and variance of the hybrid Pareto MLEs are shown in the *left and right panel* respectively.

Experiments with other value of θ give similar results. MLEs thus behave as expected.

3.4 Mixture of hybrid Paretos

The hybrid Pareto distribution can be used in a mixture in several ways. When the underlying distribution is right-skewed, only positive extremes will be present in the data. A mixture with hybrid Pareto components can be used in this case. Conversely, when the underlying distribution is left-skewed, only negative extremes will be observed. In that case, the hybrid Pareto distribution can be reversed and a mixture of such reverse hybrid Paretos can be used. Let $\tilde{h}_\theta(\cdot)$ and $\widetilde{H}_\theta(y)$ be the density and distribution functions respectively of the reverse hybrid Pareto. These relates to the hybrid Pareto in the following way:

$$\tilde{h}_\theta(y) = h_\theta(-y) \quad \text{and} \quad \widetilde{H}_\theta(y) = 1 - H_\theta(-y).$$

When the underlying distribution has heavy tails on both sides such that one can observe positive and negative extremes, one could use a mixture with hybrid Pareto and reverse hybrid Pareto components. This type of mixture model would easily adapt to the case where the underlying density is asymmetric and thus has different upper and lower tail indexes.

We could consider using two types of components in the mixture: several Gaussians and one hybrid Pareto (or reverse hybrid Pareto). In principle, one hybrid Pareto should be enough to model the tail of the distribution. This would reduce the complexity of the model and accelerate learning. However, preliminary experiments with Fréchet generated data showed that a mixture of hybrid Pareto components outperforms a mixture with Gaussian and hybrid Pareto components. It might simply be that the asymmetric shape of the hybrid Pareto does better than the bell-shaped Gaussian at modeling an asymmetric distribution such as the Fréchet. Another possibility is that using several hybrids might reduce the bias of the GP approximation and the bias introduced by the continuity constraints imposed at the junction point of the hybrid. In this paper, we focused on a mixture with hybrid Pareto components only, reverse or not.

Mixture parameters are learned by maximizing the log-likelihood on the training data. In the case of hybrid Pareto components (reverse or not), the *EM* algorithm is not efficient because there is no analytical solution at the maximization stage. Extensions of *EM* such as *ECME* used to learn the parameters of a mixture of Student *t*'s (Liu et al. 1994) cannot be used either. This is because the hybrid Pareto parameters are interrelated and there is no way to separate the Gaussian parameters μ and σ for which there is an analytical solution from the GP parameter ξ for which there is no analytical solution. Therefore, we maximize the log-likelihood directly by means of a numerical optimizer, such as a conjugate gradient optimizer.

3.5 Dominant tail and implicit threshold

The component in the mixture that has the heaviest tail is said to have the dominant tail. More precisely, following Kang and Serfozo (1999), let $F_1, \ldots, F_m$ be the d.f associated with the mixture components. The d.f F_{i^*} is said to have a dominant tail if, for $i = 1, \ldots, m$:

$$\lim_{y \to \infty} \frac{1 - F_i(y)}{1 - F_{i^*}(y)} = r_i, \text{ for } 0 \leq r_i < \infty.$$

Kang and Serfozo (1999) have shown that the component of the mixture that has the dominant tail determines the tail index of the mixture. For a mixture of Gaussians, the dominant tail is the component i^* such that $\sigma_{i^*} = \max_i \sigma_i$ and $\mu_{i^*} > \mu_i$ for all i such that $\sigma_i = \sigma_{i^*}$. This means that a mixture of Gaussians has an exponential tail with tail index $\xi = 0$.

For a mixture of hybrid Paretos, the dominant tail is the component i^* such that $\xi_{i^*} = \max_i \xi_i$ and $\beta_{i^*} > \beta_i$ if $\xi_{i^*} = \xi_i$. This entails that the tail index of the mixture of hybrid Paretos is ξ_{i^*}. Hence the mixture of hybrid Paretos can have any type of tails. In particular, if the distribution to be modelled has a heavy tail with $\xi > 0$, it would be expected that a mixture of hybrid Paretos would be a more appropriate estimator than a mixture of Gaussians.

The hybrid Pareto mixture model bypasses the need for threshold selection inherent in the PoT method (see Section 2). The threshold can be defined as the junction point α_{i^*} of the dominant component. It becomes a function of the dominant component parameters and is thus determined implicitly by the data through learning.

4 Simulation study

We used synthetic data to study the properties of the mixture of hybrid Paretos with respect to other estimators. We generated pairs of training and test sets $(\mathcal{D}_n, \mathcal{D}_l)$ according to a Fréchet distribution. Most heavy-tailed distributions asymptotically have the same tail behaviour as the Fréchet distribution. The d.f. of the standard Fréchet is: $\Phi_\xi(y) = \exp(-y^{-1/\xi})$ when $y > 0$. We can change the location and scale of the Fréchet d.f. by replacing y by $(y - \mu)/\sigma$ in the preceding equation. We chose two different values of the tail index, $\xi = 1/5$ and $\xi = 1/2$, that correspond to different degrees of tail heaviness. The location and scale parameters are taken to be zero and one respectively in all cases. Increasing training set sizes (from 100 to 20,000) are used while the test set size is fixed to 10,000. For each training set size, a number b_n of training and test sets are generated. We are thus able to report the average performance of each model along with its variance. Since there is more variability in smaller training sets and that computations are longer for larger training sets, we use more replications for the former than the latter. Specifically, $b_n = 100$ for $100 \leq n \leq 800$, $b_n = 60$ for $1{,}000 \leq n \leq 8{,}000$ and $b_n = 30$ for $n > 10{,}000$.

4.1 Training and performance criteria

The goal of this simulation study is twofold. Firstly, we want to compare the **m**ixture **m**odel with **h**ybrid Pareto components (MMH) to other density estimators. We consider mixtures with Gaussian (MMG) and log-normal (MML) components and the Parzen window estimator (PARZEN). The Log-Normal is often used in the presence of extreme events since it is asymmetric and has a heavier tail than the Gaussian. However, like the Gaussian, the Log-Normal has a tail index equals to zero, which means that the upper tail eventually decreases exponentially fast. All the density models are trained by maximizing the log-likelihood on the training set. Since optimization may lead to local minima, during learning, mixture models are randomly initialized five times and the optimization is re-started accordingly. We keep the parameters that give the smallest training error. To initialize a mixture model with m components, we use either k-means or k-medians (Pollard 1981) to group the data into m clusters. The parameters of the ith component are then computed from the data in the ith cluster. The mixture weights (priors) are initialized as the proportion of data in each cluster.

The trained models are compared in terms of relative log-likelihood on the test set:

$$\mathcal{R}_l(\phi_\theta) = -\frac{1}{l}\sum_{i=1}^{l} \log\left(\frac{\phi_\theta(y_i)}{p(y_i)}\right),$$

where $p(\cdot)$ is the generative density and $\phi_\theta(\cdot)$ a density estimator. The relative log-likelihood is, up to a scaling factor, the empirical counterpart of the Kullback-Leibler divergence $(-\int p(y)\log(\phi_\theta(y)/p(y))\,dy)$. The smaller the relative log-likelihood is, the better the estimator is performing. We also compute estimated extreme quantiles from the density estimators. A quantile of level q of the d.f. F is defined as the value y_q such that $F(y_q) = q$ given that F is increasing. For the density models that we consider, the inverse of the d.f. does not have a closed form, so we use a numerical approximation to find the zero of $F(y) - q$. We compute the estimated quantiles for levels $q = 0.99$, $q = 0.999$ and $q = 0.999$. These estimated quantiles are compared in terms of root mean squared error ($RMSE$):

$$RMSE(\hat{z}_q) = \sqrt{bias^2 + variance} = \sqrt{(E[\hat{z}_q] - 1)^2 + E[(\hat{z}_q - E[\hat{z}_q])^2]},$$

where $\hat{z}_q = \hat{y}_q / y_q$ is the standardized estimated quantile, y_q being the true quantile of level q and $\hat{y}_q$ the estimated quantile. By working with standardized quantiles, the quantity being estimated becomes one for all quantile levels.

Secondly, we want to evaluate how the mixture of hybrid Paretos performs with respect to the PoT method. For this, we compare estimators of the tail index and of extreme quantiles in terms of RMSE for the two methods. The threshold of the PoT method is taken to be a sample quantile of level q_{PoT} and q_{PoT} is treated as a hyper-parameter. For a given threshold, the GP parameters are learned by maximizing the log-likelihood. A goodness-of-fit

test (Choulakian and Stephens 2001) is then used to assess the validity of the chosen threshold. If the test fails, the threshold is increased. All other hyper-parameters (number of mixture components and Parzen window width) are chosen on a validation set which is fixed to 20% of the training set. An asymmetric t test of confidence level of 5% is then used to assess a significant improvement in performance between two mixture models. For the Parzen window estimator, this model selection procedure did not work well so we simply chose the new model over the previous one when the average performance was better.

4.2 Simulation results

Complete results for the relative log-likelihood statistic can be found in Tables 1 and 2 for the Fréchet generated data with tail index $\xi = 1/5$ and $\xi = 1/2$ respectively. The corresponding selected hyper-parameters are given in Table 3. All mixture models converge in terms of relative log-likelihood. However, *the mixture with hybrid Pareto components performs significantly better*, this being particularly true for small data sets. For both data sets, the Parzen window estimator performs poorly. This highlights the limitation of classical non-parametric estimators in the presence of extreme observations. On the Fréchet data with tail index $\xi = 1/5$, the mixture of hybrid Paretos outperforms the other estimators whereas when the tail is heavier, with the corresponding tail index $\xi = 1/2$, the performances of the mixture with hybrid Pareto and Log-Normal components are nearly indistinguishable. On both data sets, the mixture of hybrid Paretos uses less components. As expected, since the tail is heavier, more components are required for all mixture models for the case $\xi = 1/2$. The simulation results for the PoT methodology are given in Table 4. The quantile level determining the threshold is selected higher as more training data become available, therefore reducing the bias of the GP approximation while keeping the variance stable (see Section 2.2 on threshold selection).

Table 1 Log-likelihood relative to the data generative density (std. err.) on the test set for the Fréchet generated data with tail index $\xi = 1/5$ and training set size n

n	MMH	MMG	MML	PARZEN
100	0.047 (0.0028)	0.22 (0.023)	0.1 (0.017)	2.6 (1.2)
200	0.025 (0.0012)	0.11 (0.0053)	0.049 (0.0042)	1.5 (0.82)
400	0.013 (0.0007)	0.08 (0.0031)	0.026 (0.0024)	1.5 (0.57)
800	0.0071 (0.00041)	0.056 (0.0033)	0.017 (0.00044)	2.1 (0.68)
1,000	0.0059 (0.00047)	0.042 (0.0044)	0.017 (0.00073)	0.46 (0.21)
2,000	0.0035 (0.0002)	0.018 (0.0025)	0.013 (0.00078)	0.66 (0.39)
4,000	0.0021 (0.00011)	0.011 (0.0015)	0.0079 (0.00088)	1.5 (0.79)
8,000	0.0016 (9e-05)	0.0066 (0.0012)	0.0026 (0.00045)	0.85 (0.53)
10,000	0.0015 (0.00014)	0.0061 (0.0019)	0.002 (0.00055)	0.72 (0.22)
20,000	0.00092 (0.00013)	0.0025 (0.00039)	0.001 (0.0001)	0.56 (0.31)

Smaller values mean better estimators

Table 2 Log-likelihood relative to the data generative density (std. err.) on the test set for the Fréchet generated data with tail index $\xi = 1/2$ and training set size n

n	MMH	MMG	MML	PARZEN
100	0.07 (0.005)	1.6 (0.23)	0.089 (0.012)	91 (15)
200	0.036 (0.0034)	0.84 (0.12)	0.038 (0.0028)	1.3e+02 (38)
400	0.019 (0.00093)	0.55 (0.13)	0.024 (0.00085)	1.3e+02 (49)
800	0.011 (0.00052)	0.29 (0.062)	0.019 (0.0006)	1.1e+02 (21)
1,000	0.0098 (0.00055)	0.23 (0.073)	0.017 (0.00069)	2.3e+02 (1.9e+02)
2,000	0.0066 (0.00031)	0.21 (0.063)	0.012 (0.00076)	40 (12)
4,000	0.0048 (0.00026)	0.21 (0.083)	0.0071 (0.00083)	1.3e+02 (83)
8,000	0.0032 (0.00017)	0.13 (0.066)	0.0027 (0.00041)	51 (34)
10,000	0.0026 (0.00028)	0.072 (0.057)	0.002 (0.00055)	2.1e+03 (2e+03)
20,000	0.0021 (0.00023)	0.1 (0.068)	0.0011 (0.00011)	49 (33)

Smaller values mean better estimators

Figures 4, 5 and 6 show the estimated density compared to the generative model in the central, upper tail and lower tail area respectively for one sample experiment with tail index 1/5 and training set size $n = 100$ and $n = 1{,}000$. It can be seen from Fig. 4 that the trained models do reasonably well at approximating the density in the central area, except maybe for the Parzen window estimator which is rather bumpy. The approximation gets better for the larger training set. However, as shown in Fig. 5, the upper tail is underestimated by the Gaussian and the log-normal mixtures and the Parzen window estimator. The Gaussian mixture benefits from seeing more data

Table 3 Average hyper-parameters selected for the Fréchet generated data with tail index $\xi = 1/5$ and $\xi = 1/2$ and training set size n

n	m_{MMH}	m_{MMG}	m_{MML}	σ_{PARZEN}
$\xi = 1/5$				
100	2.5	2	2	0.15
200	2	2.1	2.3	0.17
400	2	2.9	2.3	0.17
800	2	3.3	2.5	0.12
1,000	2.2	4	2.5	0.095
2,000	2	4.1	2.9	0.11
4,000	2.3	4.8	3.3	0.085
8,000	2.2	5.9	4.4	0.092
10,000	2.3	6.8	4.2	0.094
20,000	2.8	7.5	4.3	0.055
$\xi = 1/2$				
100	2.1	2.6	2.1	0.77
200	2	3.2	2	1.1
400	2	4.5	2.1	0.9
800	2	5	2.3	1.8
1,000	2.2	5.5	2.3	0.71
2,000	2.1	6	2.9	0.93
4,000	2.4	7	3.6	0.69
8,000	3	8.5	3.9	0.91
10,000	3.9	9.1	4.6	1.2
20,000	3.9	9.3	4.5	0.63

For the mixture models (MM), m_{MM} is the number of components and for the Parzen window estimator (PARZEN), σ_{PARZEN} is the window width

Table 4 Log-likelihood relative to the data generative density $\mathcal{R}_l$ and hyper-parameter selected for the PoT method on the Fréchet generated data with tail index $\xi = 1/5$ and $\xi = 1/2$ and training set size n

n	u	q_{PoT}	$\mathcal{R}_l$ (std.err)
$\xi = 1/5$			
100	0.92	0.2	0.034 (0.0028)
200	0.95	0.26	0.02 (0.0013)
400	1	0.34	0.01 (0.00051)
800	1	0.39	0.0086 (0.00033)
1,000	1	0.42	0.0066 (0.00035)
2,000	1.1	0.52	0.0013 (0.00024)
4,000	1.2	0.59	0.002 (0.00019)
8,000	1.2	0.64	0.004 (0.00015)
10,000	1.2	0.66	0.003 (0.00018)
20,000	1.2	0.7	0.0012 (0.00012)
$\xi = 1/2$			
100	0.67	0.096	0.034 (0.0031)
200	0.82	0.2	0.019 (0.0011)
400	0.85	0.24	0.012 (0.00081)
800	0.97	0.32	0.005 (0.0003)
1,000	0.96	0.32	0.0053 (0.0003)
2,000	1.1	0.41	0.0021 (0.00026)
4,000	1.2	0.5	0.00068 (0.00019)
8,000	1.4	0.57	0.00037 (0.00012)
10,000	1.4	0.57	0.0043 (0.00015)
20,000	1.6	0.64	0.0031 (0.00011)

The selected threshold is u and the corresponding quantile level is q_{PoT}

during training but the Parzen window estimator is not able to learn the upper tail properly. The mixture of Log-Normal is not performing too bad but still the upper tail is under-estimated by a significant amount . The PoT estimator and the mixture with hybrid Paretos behave similarly and approximate the upper tail reasonably well even for the smaller data set. On the other hand, the density of the data generative model is zero in the lower tail. By design, the mixture of Log-Normals has zero density in the lower tail and thus models adequately the generative model. Since the Parzen window estimator generally predicts little density outside the range of the data, it also performs well in that case at modelling the lower tail. The main observation from Fig. 6 is that the lower tail of mixture of hybrid Paretos drops faster than the lower tail of the mixture of Gaussians. The latter is thus over-estimating badly the lower tail. Similar observations can be drawn when the tail index of the generative model is 1/2.

The standardized quantiles are shown in Figs. 7, 8 and 9 for the quantile levels $q = 0.99$, $q = 0.999$ and $q = 0.9999$ respectively and the corresponding RMSE are given in Tables 5, 6 and 7. The results for the case $\xi = 1/2$ are not shown here because they give a similar insight. When the training size is smaller than 1,000, the mixture of hybrid Paretos gives highly variable quantile estimates. Few components are used and they are centered where most of the data lie. To account for extremes, the tail index of one of the components has to be very large. This gives rise to largely overestimated quantiles, particularly as we move further in the tail of the distribution (when $q = 0.9999$). We have to

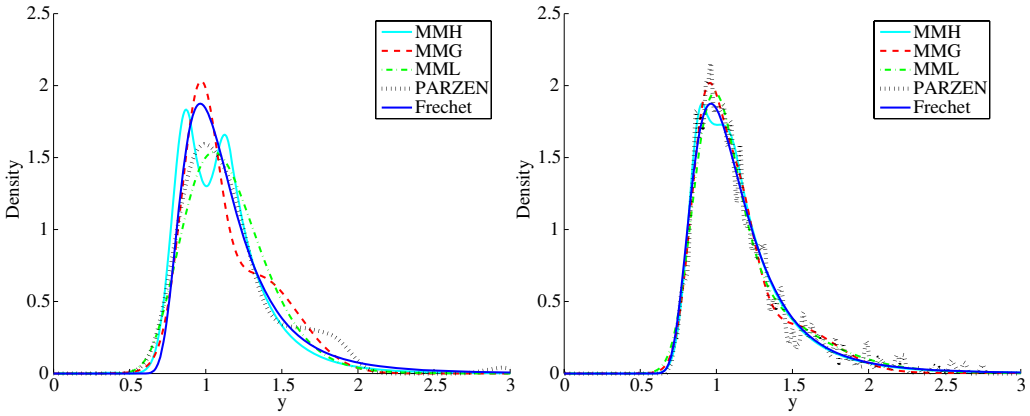


Fig. 4 Estimated density in the central part (99% of training points) for the Fréchet data with $\xi = 1/5$. *Left panel* 100 training points and *right panel*, 1,000 training points.

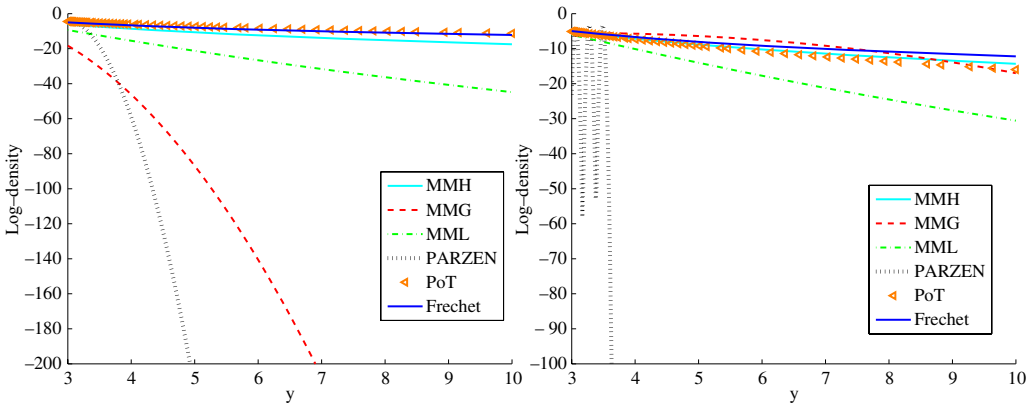


Fig. 5 Estimated log-density in the upper tail (<1% of training points) for the Fréchet data with $\xi = 1/5$. *Left panel* 100 training points and *right panel*, 1,000 training points.

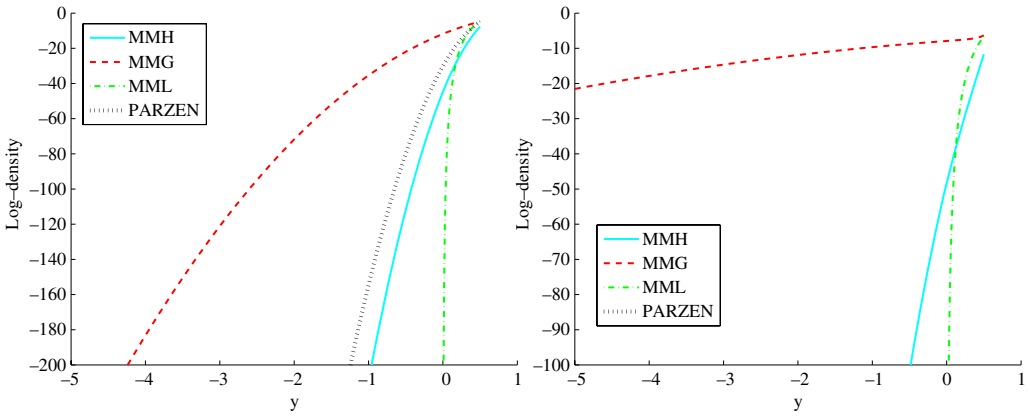


Fig. 6 Estimated log-density in the lower tail (no training points) for the Fréchet data with $\xi = 1/5$. *Left panel* 100 training points and *right panel*, 1,000 training points.

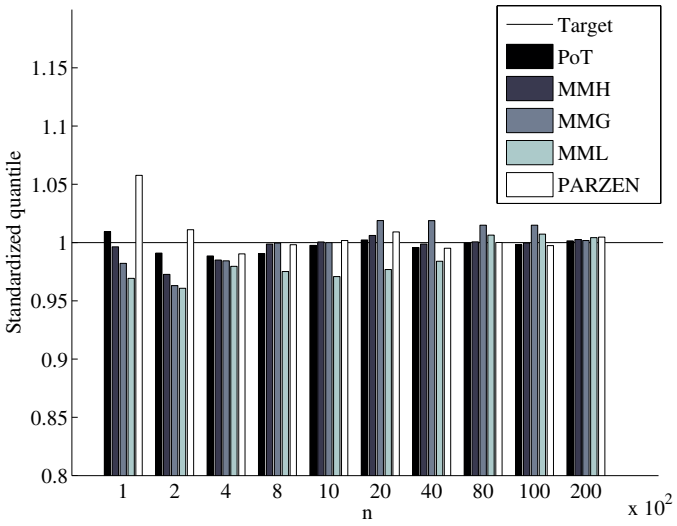


Fig. 7 Standardized estimated quantiles of level 0.99 for Fréchet data with $\xi = 1/5$.

keep in mind that we trained the models by maximizing the likelihood criterion and in this regard, the mixture of hybrid Paretos outperforms the other models in all circumstances. Some authors (Embrechts et al. 1997) advise against estimating too extreme quantiles since such estimates require extrapolation far away from the observations and are thus generally unreliable. However, for medium to large data sets, the quantiles estimated from the mixture of hybrid Paretos are in general more accurate then those produced from other models, PoT included. As the training set size increases, the Gaussian and log-normal mixtures produce quantile estimates with *RMSE* similar to the hybrid Pareto mixture estimates. The Parzen window estimator produces quantile estimates with large *RMSE* in most cases.

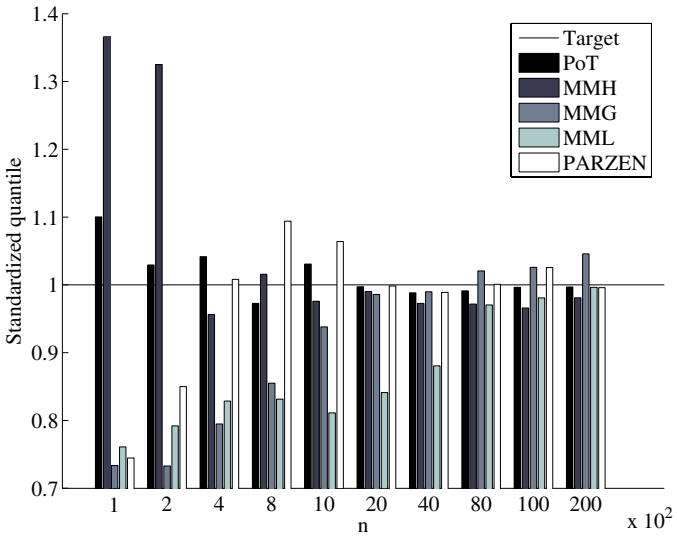


Fig. 8 Standardized estimated quantiles of level 0.999 for Fréchet data with $\xi = 1/5$.

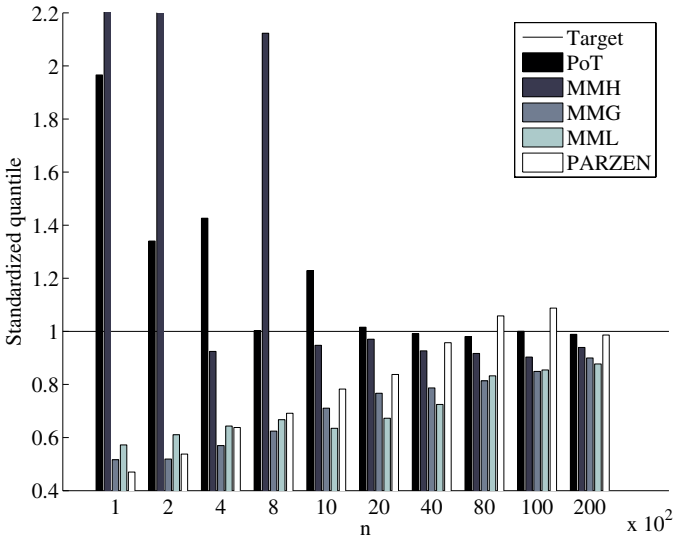


Fig. 9 Standardized estimated quantiles of level 0.9999 for Fréchet data with $\xi = 1/5$.

Tail index estimators provided from the mixture of hybrid Paretos and the PoT method are given in Fig. 10 along with their corresponding *RMSE* for the Fréchet data with tail index equal to 1/5. The tail index estimated with the mixture of hybrid Paretos gives on average a better estimate of the tail index although, the *RMSE* is sometime large, particularly on smaller training sets. This is due to the variance factor in the *RMSE*. An estimator of the tail index serves to determine the heaviness of the tail of the underlying distribution. It is also crucial in the PoT method to obtain tail and quantile estimates. For the mixture of hybrid Paretos, accurate estimation of the tail index is not so central because tail estimation is given by the combination of components, not just the dominant one. Regarding extreme quantile estimation, the mixture of hybrid Paretos provides more accurate estimates than the PoT method in all but a few instances. On the other hand, when comparing the density estimators on the upper tail and on the tail index estimates, the two methods perform similarly. *The mixture of hybrid Paretos is thus a valid alternative to the PoT methodology as far as the estimation of the tail of the distribution is concerned.*

Table 5 RMSE of the estimated quantiles of level 0.99 for Fréchet data with $\xi = 1/5$. The true quantile is $y_{0.99} = 2.5094$

n	PoT	MMH	MMG	MML	PARZEN
100	0.17	0.18	0.28	0.29	0.28
200	0.13	0.11	0.15	0.14	0.14
400	0.092	0.081	0.11	0.11	0.098
800	0.063	0.059	0.091	0.076	0.071
1,000	0.051	0.049	0.071	0.077	0.063
2,000	0.041	0.046	0.074	0.052	0.042
4,000	0.032	0.031	0.055	0.041	0.031
8,000	0.019	0.017	0.043	0.032	0.021
10,000	0.02	0.017	0.032	0.028	0.018
20,000	0.013	0.013	0.024	0.021	0.016

Table 6 RMSE of the estimated quantiles of level 0.999 for Fréchet data with $\xi = 1/5$. The true quantile is $y_{0.999} = 3.9807$

n	PoT	MMH	MMG	MML	PARZEN
100	0.78	2	0.36	0.35	0.36
200	0.48	2.4	0.31	0.29	0.35
400	0.36	0.18	0.3	0.24	0.39
800	0.18	0.28	0.25	0.24	0.29
1,000	0.21	0.13	0.27	0.22	0.31
2,000	0.13	0.13	0.19	0.19	0.12
4,000	0.081	0.084	0.14	0.16	0.081
8,000	0.06	0.058	0.097	0.087	0.076
10,000	0.066	0.063	0.11	0.089	0.094
20,000	0.041	0.043	0.095	0.063	0.047

Experiments with other generative distributions, namely the Pareto with $\xi = 1/3$, the Gaussian, the Cauchy and the Student t with degree of freedom equal to four, were carried out but detailed results are not shown here. Since the last three generative distributions have density on both tails, we make use of both hybrid and reverse hybrid Pareto components in the mixture. The Pareto distribution is up to an affine transformation the same as the GP distribution for positive tail indexes. This gives a great advantage to the PoT method. Nevertheless, the hybrid Pareto mixture, which models the upper tail as a combination of GPs, provides a sensible estimator of the tail for the Pareto generated data. The tails of the Gaussian distribution are eventually decreasing exponentially fast. The tail index of the Gaussian distribution is thus zero. This is a limit case of the PoT method and of the hybrid Pareto mixture since the tail indexes are parametrized in such a way that they are guaranteed to be positive. The average relative log-likelihood results show that the hybrid Pareto mixture's performance is very similar to the Gaussian mixture's performance although it gets slightly worst for larger training set size. Regarding extreme quantile estimation, the *RMSE* of the PoT method is much larger than the *RMSE* of the hybrid Pareto mixture. This might be explained by the fact that the approximation of the Gaussian tail by the GP starts to make sense for very high threshold values, that is, very far away in the tail. The

Table 7 RMSE of the estimated quantiles of level 0.9999 for Fréchet data with $\xi = 1/5$. The true quantile is $y_{0.9999} = 6.3095$

n	PoT	MMH	MMG	MML	PARZEN
100	6.8	1.1e+03	0.52	0.48	0.55
200	2	5.5e+02	0.5	0.44	0.5
400	1.8	0.36	0.47	0.38	0.44
800	0.5	7.6	0.42	0.41	0.35
1,000	0.99	0.25	0.4	0.38	0.37
2,000	0.34	0.27	0.32	0.35	0.29
4,000	0.19	0.17	0.27	0.31	0.27
8,000	0.14	0.13	0.22	0.2	0.25
10,000	0.16	0.13	0.22	0.19	0.26
20,000	0.098	0.11	0.17	0.16	0.17

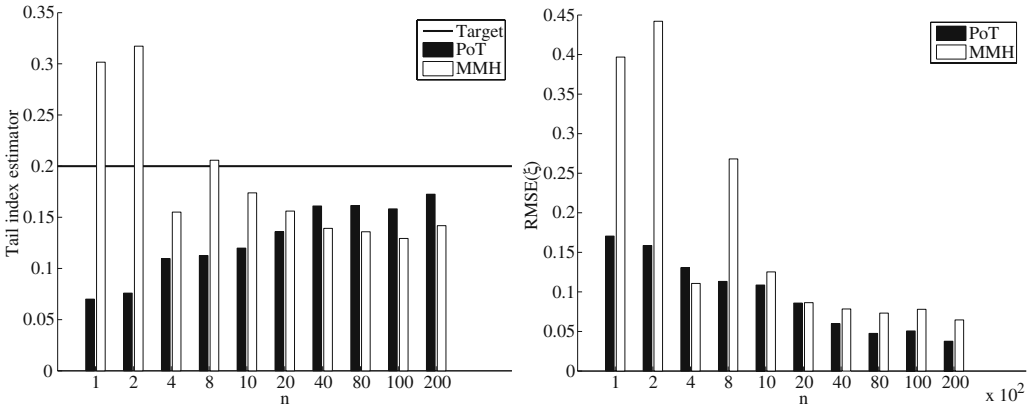


Fig. 10 Estimated tail index on the *left panel* and root-mean-squared error (RMSE) on the *right panel*. The generative distribution is the Fréchet with $\xi = 1/5$.

hybrid Pareto mixture might take advantage of using many components, thus many GPs, to improve the approximation of the Gaussian tail. The Gaussian case demonstrates that the hybrid Pareto mixture is a reasonable estimator even in the case of light-tailed data. The Cauchy distribution is heavy-tailed with tail index equal to one and the Student t with ν degrees of freedom has tail index equal to $1/\nu$. The Cauchy generated data were particularly difficult to learn for the Gaussian mixture and the Parzen window estimator, even for the largest data set. In this last two instances, the hybrid Pareto mixture proved to be an appropriate density estimator for heavy-tailed data on small training sets.

5 Danish fire insurance data

We apply the mixture of hybrid Pareto estimator on the Danish fire insurance data. These data were used by McNeil (1997) to illustrate the PoT methodology on insurance data. The Danish fire insurance data set is available within the software *R*; it consists of an irregular time series with 2167 observations. The values of the claims were re-scaled for commercial reasons. Figure 11 shows the histogram of the logarithm of the claims.

Since the generative model is unknown, we measure the performance by comparing how a competing estimator performs on test data relative to the proposed mixture of hybrid Paretos. For a given point y, the performance measure is given by the relative log-likelihood:

$$\mathcal{R}(y) = \log\left(\phi_\theta^{\mathrm{MMH}}(y)\right) - \log\left(\phi_\theta^{\mathrm{other}}(y)\right),$$

where $\phi_\theta^{\mathrm{MMH}}$ is the mixture of hybrid Pareto estimator and $\phi_\theta^{\mathrm{other}}$ is a competing density estimator. Positive values of $\mathcal{R}$ mean that the hybrid Pareto mixture performs better than the competing estimator. Increasing training set sizes which correspond to 50%, 75% and 90% of the original data set were used. Just

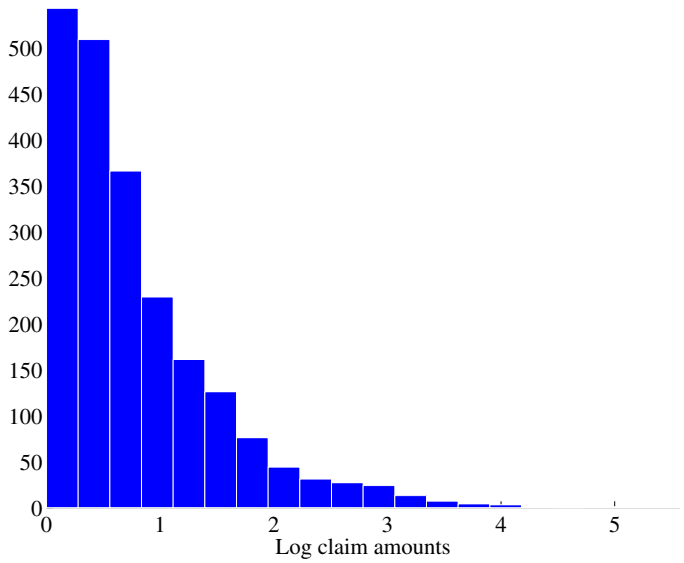


Fig. 11 Histogram of the logarithm of the Danish fire insurance data.

as in the simulation study, hyper-parameters (number of mixture components and Parzen window width) were selected on a validation set which is fixed to 20% of the training set. An asymmetric t test of confidence level of 5% is used to assess a significant improvement in performance between two mixture models. For the Parzen window estimator, this model selection procedure did not work well so we simply chose the new model over the previous one when the average performance was better. The quantile level which determines the threshold of the PoT methodology is chosen again with the goodness-of-fit test by Choulakian and Stephens (2001).

Table 8 gives the average relative log-likehood on the test set along with its standard error. The hyper-parameters selected are given in Table 9. *The mixture of hybrid Paretos out-performs significantly the other mixture models for all training set sizes.*

McNeil (1997) tried fitting a GP at various threshold levels. The lower the threshold is, the larger the quantile estimates are. The threshold selection method we used, based on the goodness-of-fit test provided by Choulakian and Stephens (2001), yields rather low thresholds as can be seen in Table 9. The goal of the method is to select the lowest threshold that gives an adequate fit. Figure 12 shows the quantile estimates of level $q = 0.99$, $q = 0.999$ and $q = 0.9999$ for all models together with the tail index estimates of the PoT method

Table 8 Average log-likelihood relative to the mixture of hybrid Paretos (MMH) on the test set for the Danish fire insurance data with training set sizes n

n	MMG	MML	PARZEN
1084	0.075 (0.0098)	0.038 (0.01)	6.7 (4.5)
1625	0.13 (0.02)	0.12 (0.021)	13 (9)
1950	0.13 (0.031)	0.13 (0.035)	1.1 (0.077)

Positive values indicate MMH performs better

Table 9 Hyper-parameters selected for the Danish fire insurance data with training set sizes n

n	m_{MMH}	m_{MMG}	m_{MML}	σ	u	q_{PoT}
1,084	2	4	2	1	1.4694	0.3
1,625	2	4	2	1	1.4417	0.3
1,950	2	4	2	4	1.4201	0.3

and of the mixture of hybrid Paretos. The tail index and the quantiles estimates for the levels $q = 0.999$ and $q = 0.9999$ provided by our implementation of the PoT method are within the range of the estimates obtained by McNeil (1997). The mixture of hybrid Paretos gives larger tail index estimates and as a result, the quantile estimates, specially for the higher level and the larger training sets, are larger than for the PoT method. The quantile estimates provided by the other density estimators are generally smaller thus more conservative.

It is possible to develop a binomial test to measure the performance of the quantile estimators based on the number of violations (number of times the quantile estimators are exceeded), see McNeil and Frey (2000). Under the null hypothesis, the indicator for a violation follows a Bernouilli: $1_{\{x_i > x_q\}} \sim$

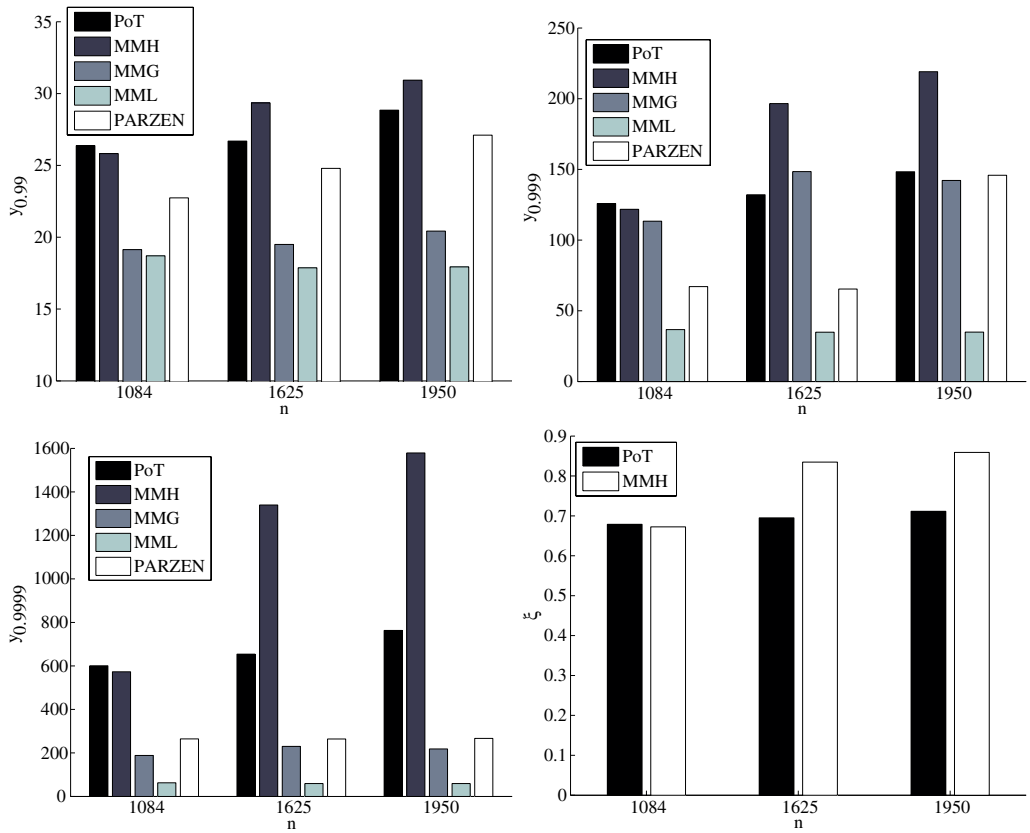


Fig. 12 Danish fire insurance data: Estimated quantiles of level 0.99, 0.999 and 0.9999 on the *top left*, *top right panel* and on the *bottom left panel* respectively. Estimated tail indexes on the bottom right panel.

Table 10 *P*-values and 95% confidence interval for the binomial test of the number of violations of quantile estimators. Under the null hypothesis, the number of violations should follow a binomial distribution $B(n_{\text{test}}, 1-q)$, where n_{test} is the test set size and q is the quantile level

n_{test}	$1-q=0.01$	$1-q=0.001$	$1-q=0.0001$
MMH			
1,083	0.3554 (0.007 0.0216)	0.2948 (0.0002 0.0067)	1 (0.00 0.0034)
542	0.6672 (0.0041 0.0239)	1 (0.00 0.0068)	1 (0.00 0.0068)
217	0.73 (0.0001 0.0254)	1 (0.00 0.0169)	1 (0.00 0.0169)
MMG			
1,083	**0.0020 (0.0128 0.0306)**	0.2948 (0.0002 0.0067)	1 (0.00 0.0034)
542	**0.0038 (0.0128 0.041)**	0.4186 (0.00 0.0010)	1 (0.00 0.0068)
217	0.4824 (0.0029 0.0399)	0.1952 (0.0001 0.0254)	1 (0.00 0.0169)
MML			
1,083	**0.0020 (0.0128 0.0306)**	**0.0051 (0.0015 0.0107)**	**0.005453 (0.0002 0.0067)**
542	**0.0005 (0.0156 0.0452)**	**0.0023 (0.0020 0.0188)**	**0.001414 (0.0004 0.0133)**
217	0.1740 (0.0050 0.0465)	0.1952 (0.0001 0.0254)	0.02147 (0.0001 0.0254)
PARZEN			
1,083	0.04422 (0.0099 0.0261)	0.2948 (0.0002 0.0067)	1 (0.00 0.0034)
542	0.2721 (0.0064 0.0289)	0.1031 (0.0004 0.0133)	1 (0.00 0.0068)
217	1 (0.0011 0.0329)	1 (0.00 0.0169)	1 (0.00 0.0169)
PoT			
1,083	0.4454 (0.0064 0.0204)	0.2948 (0.0002 0.0067)	1 (0.00 0.0034)
542	0.5105 (0.0052 0.0264)	0.1031 (0.0004 0.0133)	1 (0.00 0.0068)
217	0.73 (0.0001 0.0254)	1 (0.00 0.0169)	1 (0.00 0.0169)
108	0.2938 (0.0023 0.0653)	1 (0.00 0.0336)	1 (0.0 0.0336)

If the null hypothesis is adequate, the true proportion of violations $1-q$ should be inside the confidence interval

$Be(1-q)$. The number of violations on the test set is thus a Binomial: $\sum_{i=1}^{l} 1_{\{x_i > x_q\}} \sim B(l, 1-q)$. In Table 10, we provide *P*-values for the quantile estimators of all models for the three quantile levels. Besides the *P*-value, we give a 95% confidence interval of the proportion of violations. If the null hypothesis is adequate, we should expect the P-value to be larger then a given confidence level, usually taken as 5%. We would also expect that the true proportion of violations $1-q$ be included in the 95% confidence interval. When no violations occur, as it is often the case for the largest quantile level or the smallest test set size, not much conclusions can be drawn. The P-value is 1 and the confidence interval is wide. However, when violations do occur, the Binomial test never rejects the hybrid Pareto mixture estimator. The same is true for the PoT estimator. Over-estimation of the number of violations which corresponds to under-estimation of the quantiles are in bold in Table 10. The binomial test rejects the Gaussian mixture estimator in two instances. The Log-Normal mixture estimator is rejected quite often while the Parzen window estimator performs surprisingly well according to this test.

6 Conclusion

Extremes occur in a variety of contexts. Extreme sea conditions can lead to flooding. Insurance companies need to evaluate the probability of large claims

for re-insurance purposes. Financial portfolio managers have to estimate the risk of wide variations in the price of the financial instruments in their portfolio. Extreme value theory has put forward sound mathematical methodologies to estimate the tail of a distribution when extreme events are more frequent than for a Gaussian distribution. One such method is the *Peaks-over-Thresholds* (PoT) that fits a generalized Pareto (GP) distribution above a given threshold.

The mixture of Gaussians with adaptive number of components is a popular non-parametric estimator with good convergence properties. However, it performs poorly in small data sets when the underlying distribution is fat-tailed. In order to combine the advantages of the PoT method and of the mixture of Gaussians, we build a new distribution, the hybrid Pareto, which stitches together a Gaussian and a GP while enforcing continuity constraints. The hybrid Pareto is thus a smooth extension of the GP to the whole real axis and it can be used in a mixture model. Moreover, this estimator circumvents the need for threshold selection inherent in the PoT methodology since the threshold becomes a function of the mixture of hybrid Pareto parameters and is therefore implicitly learned by maximizing the log-likelihood.

In the mixture of hybrid Paretos, the whole data set participates in the tail estimation, not only the exceedances above the threshold. The tail is now estimated by a combination of GP, instead of a single GP and this might help in alleviating the bias of the approximation of the tail by the GP. This is supported by the results in the simulation study which shows that the mixture of hybrid Paretos does generally better than the PoT method at estimating extreme quantiles. At the same time, the mixture of hybrid Paretos provides an alternate way to estimate the tail index of the underlying distribution. Finally, *the proposed non-parametric estimator outperforms systematically other density estimators in terms of log-likelihood when the underlying distribution is fat-tailed, this being particularly striking in smaller data sets.* We applied the proposed methodology on the Danish fire insurance data and obtained similar quantile and tail index estimates than in McNeil (1997) for the PoT method although the mixture of hybrid Paretos generally gives larger estimates. A Binomial test for the number of violations of the estimated quantiles confirms that the mixture of hybrid Paretos gives as reliable quantile estimates as the PoT method. This test also warns against the mixture of Gaussians and Log-Normals that in some occasions give quantiles estimates that are too small.

Acknowledgements The authors thank the following funding organizations: NSERC, MITACS, and the Canada Research Chairs.

References

Choulakian, V., Stephens, M.A.: Goodness-of-fit tests for the generalized Pareto distribution. Technometrics **43**, 478–484 (2001)

Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J., Knuth, D.E.: On the Lambert W function. Adv. Comput. Mat. **5**, 329–359 (1996)

Danielsson, J., de Haan, L., Peng, L., de Vries, C.G.: Using the Bootstrap method to choose the sample fraction in tail index estimation. J. Multivar. Anal. **76**, 226–248 (2001) (Sample: Extreme Quantile and Probability Estimation, Financial Markets Group)
Davison, A.C., Smith, R.L.: Models for exceedances over high thresholds. J. R. Stat. Soc., Ser. B Stat. Methodol. **52**, 393–442 (1990)
Dupuis, D.J.: Exceedances over high thresholds: a guide to threshold selection. Extremes **1**, 251–261 (1998)
Embrechts, P., Kluppelberg, C., Mikosch, T.: Modelling Extremal Events. Springer-Verlag, Berlin (1997)
Fama, E.F.: The behavior of stock market prices. J. Bus. **38**, 34–105 (1965)
Frigessi, A., Haug, O., Rue, H.: A dynamic mixture model for unsupervised tail estimation without threshold selection. Extremes **5**, 219–235 (2002)
Hosking, J.R.M., Wallis, J.R.: Parameter and quantile estimation for the generalized Pareto distribution. Technometrics **29**, 339–349 (1987)
Kang, S., Serfozo, R.F.: Extreme values of phase-type and mixed random variables with parallel-processing examples. J. Appl. Probab. **36**, 194–210 (1999)
Liu, C., Rubin, D.B., Liu, C., Rubin, D.B.: Weighted finite population sampling to maximize entropy. Biometrika **81**, 633–648 (1994)
Mandelbrot, B.: The variation of certain speculative prices. J. Bus. **36**, 394–419 (1963)
McNeil, A.J.: Estimating the tails of loss severity distributions using extreme value theory. ASTIN Bull. **27**, 117–137 (1997)
McNeil, A.J., Frey, R.: Estimation of tail-related risk measures for heteroscedatic financial time series: an extreme value approach. J. Empir. Finance **7**, 271–300 (2000)
Pickands, J.: Statistical inference using extreme order statistics. Ann. Stat. **3**, 119–131 (1975)
Pollard, D.: Strong consistency of the K-means clustering. Ann. Stat. **9**, 135–140 (1981)
Priebe, C.E.: Adaptive mixtures. J. Am. Stat. Assoc. **89**, 796–806 (1994)
Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Bank Finance **26**, 1443–1471 (2002)
Smith, R.L.: Maximum likelihood estimation in a class of nonregular cases. Biometrika **72**, 67–90 (1985)
Smith, R.L.: Estimating tails of probability distributions. Ann. Stat. **15**, 1174–1207 (1987)

A.1.2 Modélisation conditionnelle à l'aide de réseau de neurones

A Hybrid Pareto Mixture for Conditional Asymmetric Fat-Tailed Distributions

Julie Carreau and Yoshua Bengio

***Abstract*—In many cases, we observe some variables X that contain predictive information over a scalar variable of interest Y, with (X, Y) pairs observed in a training set. We can take advantage of this information to estimate the conditional density $p(Y|X = x)$. In this paper, we propose a conditional mixture model with hybrid Pareto components to estimate $p(Y|X = x)$. The hybrid Pareto is a Gaussian whose upper tail has been replaced by a generalized Pareto tail. A third parameter, in addition to the location and spread parameters of the Gaussian, controls the heaviness of the upper tail. Using the hybrid Pareto in a mixture model results in a nonparametric estimator that can adapt to multimodality, asymmetry, and heavy tails. A conditional density estimator is built by modeling the parameters of the mixture estimator as functions of X. We use a neural network to implement these functions. Such conditional density estimators have important applications in many domains such as finance and insurance. We show experimentally that this novel approach better models the conditional density in terms of likelihood, compared to competing algorithms: conditional mixture models with other types of components and a classical kernel-based nonparametric model.**

***Index Terms*—Conditional density estimation, extreme events, fat-tailed data, generalized Pareto distribution (GPD), mixture models, neural nets.**

I. Introduction

INSURANCE companies need to compute insurance premia for a given client. This computation is based on the conditional distribution of the claims given a client profile. For auto insurance data that will be used in the experiments below, the client profile is described by the observed variables X which contain information about the driver, the car, and the options selected by the insured in the insurance contract. The dependent variable Y represents the claim amounts. The conditional distribution of Y given X can be heavy-tailed, that is large claim amounts can be observed, depending on the client profile X. We propose a new nonparametric conditional density estimator that can account for asymmetry, multimodality, and heavy tails. Such an estimator will prove useful for the auto insurance data.

Extreme value theory (EVT) is a branch of probability that has put forward techniques to model the tail of univariate distributions [1]. One such technique is based on the generalized Pareto distribution (GPD). Let Y be a scalar variable such as the one representing the claim amounts, then the probability density of Y given that Y exceeds a threshold u, $p(Y|Y > u)$, can be modeled by the GPD. For all well-known distributions, the approximation by the GPD can be made arbitrarily good by choosing the threshold u to be high enough [2]. In practice, let $\{Y_1, \ldots, Y_n\}$ be the observations and let u be a given threshold. Then, the exceedances $\{Y_j - u|Y_j > u\}$ can be used to estimate the parameters of a GPD unconditional density model. The selection of an adequate threshold gives rise to a bias-variance tradeoff: the higher the threshold is, the better the approximation by the GPD becomes (smaller bias) but at the same time, the variance of the estimated parameters increases because less exceedances are available. In insurance, the GPD has been used to model the density of large claims [3]. The case we will study in this paper is complicated by the dependence of the density of Y on some observed variables X. For this purpose, we need a way to represent the density not just in the tail. Therefore, we propose to use the hybrid Pareto [4], illustrated in Fig. 1. It is a continuous density which shares the tail approximation properties of the GPD. The GPD is discontinuous: its density is zero below the threshold. The hybrid Pareto is built by stitching a Gaussian to the left of the threshold of the GPD while enforcing continuity constraints. The threshold, which is the junction point between the Gaussian and the GPD, becomes a function of the hybrid Pareto parameters. The hybrid Pareto thus circumvents the need for threshold selection. The hybrid Pareto was used in a mixture model in the context of unconditional density estimation [4], i.e., estimating $p(Y)$. For the central part, the hybrid Pareto mixture boils down to a mixture of Gaussians. The Gaussian mixture is a popular nonparametric density estimator [5]. It has good convergence properties (see [6], for instance). However, when few data points are available and the tail is heavy, the Gaussian mixture underestimates the tail of the underlying distribution, as illustrated in Fig. 2. In contrast, the hybrid Pareto mixture has the advantage of the GPD approximation properties to handle heavy-tailed data. The tail of the hybrid Pareto mixture is a convex combination of GPD tails.

In [4], the introduced model addressed unconditional density estimation with asymmetric, multimodal, and heavy-tailed data. In this paper, we consider an extension of that model which allows conditional density estimation $p(Y|X = x)$. The model is a hybrid Pareto mixture whose parameters are defined as *functions* of the input value x. These functions are implemented

Manuscript received March 12, 2008; revised November 08, 2008; accepted February 07, 2009. First published May 26, 2009; current version published July 09, 2009. This work was supported by the National Sciences and Engineering Research Council of Canada (NSERC), the Mathematics of Information Technology and Complex Systems (MITACS), and the Canada Research Chairs.

J. Carreau is with the Laboratoire des Sciences du Climat et de l'Environnement, UMR CEA-CNRS-UVSQ, Gif-sur-Yvette 91191, France (e-mail: julie.carreau@lsce.ipsl.fr).

Y. Bengio is with the Department of Computer Science and Operations Research, University of Montreal, Montreal, QC H3C 3J7 Canada (e-mail: bengioy@iro.umontreal.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2009.2016339

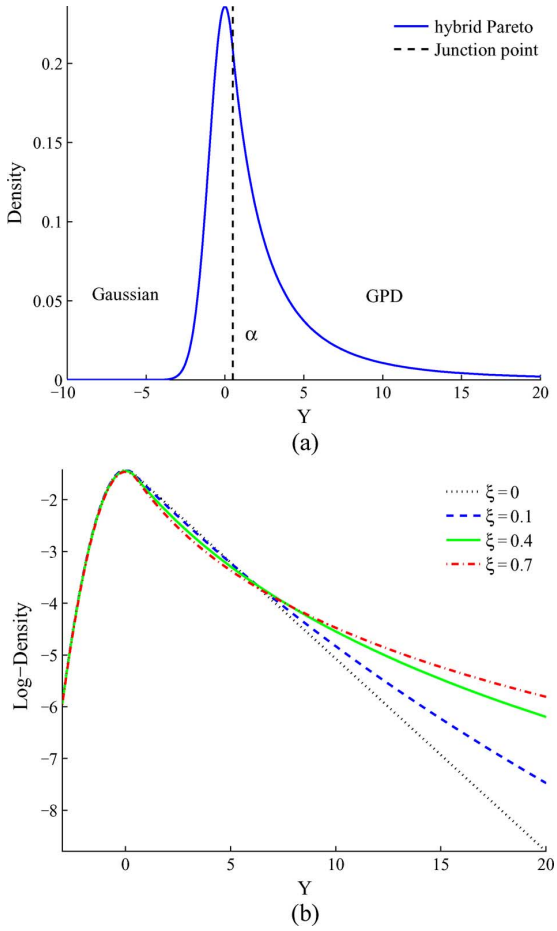


Fig. 1. (a) Hybrid Pareto density with parameters $\xi = 0.4, \mu = 0$, and $\sigma = 1$. On the left of the junction point, the density is Gaussian shaped, on the right it is GPD shaped. (b) Hybrid Pareto log-density for various tail parameters and in all cases $\mu = 0$ and $\sigma = 1$.

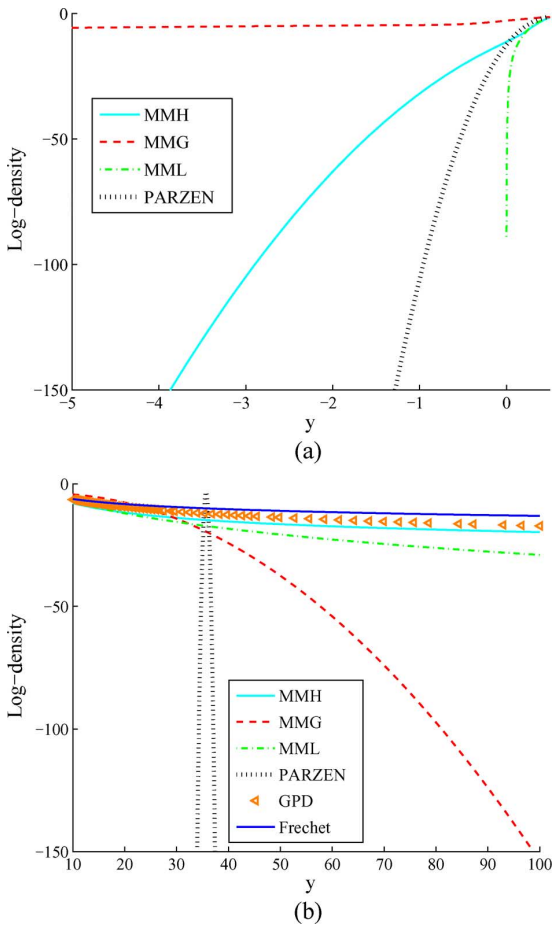


Fig. 2. Unconditional density estimation by a mixture with either hybrid Pareto (MMH), Gaussian (MMG), or log-normal (MML) components and the Parzen window estimator (PARZEN). One hundred training points were generated from a standard Fréchet distribution with tail index $\xi = 0.5$. (a) Lower tail (no training points), MMG overestimates the density whereas MMH drops quickly. (b) Upper tail (<1% of training points), MMG underestimates the density while MMH and the GPD are close to the Fréchet.

with a one hidden-layer neural network which is a convenient function approximator [7] that can be optimized in this context. Provided enough data and hidden units, and appropriate optimization, the neural network can capture any smooth dependencies of the parameters on the input, i.e., given the input, it can, in principle, capture any conditional continuous density, be it asymmetric, multimodal, or heavy-tailed. Synthetic data sets are used to highlight the tail approximation properties of the proposed conditional density estimator. We then evaluate its performance on real data sets: an insurance data set and the KDD cup 98 data set.

II. Hybrid Pareto Distribution

The hybrid Pareto distribution was proposed as a way to bridge the gap between nonparametric density estimation and EVT methods [4]. Whereas the GPD has density 0 for negative values (or more generally for values below a threshold), the hybrid Pareto extends smoothly the GPD to the whole real axis and enables the use of the GPD within mixture models. The density function of the GPD is given in (1) where $y \geq 0$ when $\xi \geq 0$ and $0 \leq y \leq -\beta/\xi$ when $\xi < 0$. The location of the GPD can be changed by replacing y by $y - \alpha$ in

$$g_{\xi;\beta}(y) = \begin{cases} \frac{1}{\beta}\left(1 + \xi\frac{y}{\beta}\right)^{-1/\xi - 1}, & \text{if } \xi \neq 0, \\ \frac{1}{\beta}e^{-\frac{y}{\beta}}, & \text{if } \xi = 0. \end{cases} \quad (1)$$

The parameter ξ of the GPD, called the *tail index*, controls the thickness of the tail while β is a scale parameter. When $\xi > 0$, the GPD can account for heavy tails (e.g., Pareto, α-stable, and student t distributions). When $\xi = 0$, the GPD can model exponential tails (e.g., Gaussian, exponential, and log-normal distributions). Finally, when $\xi < 0$, the GPD has a finite tail (e.g., uniform or Beta distributions). The approximation of the tail of a distribution by the GPD is based on the following decomposition. Let u be a given threshold and let Y be the scalar variable

of interest. The tail of the distribution of Y can be written as $\forall y > 0$

$$P(Y \leq u+y) = P(Y \leq u)+P(Y > u)P(Y-u \leq y|Y > u). \tag{2}$$

The following approximations are then used when building unconditional distribution models with the GPD: $P(Y > u) \approx$ (# excesses)/(# observations) and $P(Y - u \leq y|Y > u) \approx G_{\xi;\beta}(y)$, where $G_{\xi;\beta}$ is the distribution function of the GPD and $F_u(y) = P(Y-u \leq y|Y > u)$ is called the excess distribution function of Y. The approximation of $F_u(\cdot)$ by the GPD is theoretically justified by Pickands theorem [2] which states that, for most distributions, F_u converges to $G_{\xi;\beta}$ as the threshold increases. The GPD ξ and β parameters can be learned from the observations that exceed the threshold u. Maximum-likelihood estimators (MLEs) of the GPD parameters exist when $\xi > -1$ and are asymptotically normal and efficient when $\xi > -0.5$ [8].

As mentioned in the introduction, the choice of threshold u in such models [8] is subject to a bias-variance tradeoff. The higher u, the better the approximation of the tail by the GPD becomes. Lower u yields a smaller variance of the GPD MLEs. Several methods have been proposed for threshold selection; see, for instance, [1]. However, most methods require hand-tuning and do not extend easily to *conditional* density estimation.

A. Hybrid Pareto Derivation

The hybrid Pareto is built by stitching together a Gaussian distribution and a GPD while enforcing continuity of the resulting density and of its derivative at the junction point. Let α be the junction point and let $f_{\mu;\sigma}(y) = 1/(\sqrt{2\pi}\sigma)\exp(-(y-\mu)^2/(2\sigma^2))$ be the Gaussian density function with parameters μ and σ. The GPD density located above α is given in (1) by replacing y by $y - \alpha$ in the equations.

There are initially five parameters: α, μ, σ, ξ, and β. Since two continuity constraints (on the density function and its derivative) must be fulfilled, we are left with three free parameters. We set ξ, μ, and σ as the free parameters and we let α and β be functions of these. This choice is intuitively reasonable since those parameters are easily interpretable: ξ as measuring the tail heaviness of the distribution, μ its location, and σ its spread. We show how the equations for the hybrid Pareto are derived for the case $\xi > 0$; the other cases are similar and need only minor adjustments. The continuity constraint on the density at α means that $f_{\mu;\sigma}(\alpha) = g_{\xi;\beta}(0)$, which gives

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = \frac{1}{\beta} \Leftrightarrow \exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = \frac{\sqrt{2\pi}\sigma}{\beta}. \tag{3}$$

Continuity of the derivative of the density at α means that $f'_{\mu;\sigma}(\alpha) = g'_{\xi;\beta}(0)$, which yields

$$-\frac{(\alpha-\mu)}{\sqrt{2\pi}\sigma^3}\exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) = -\frac{(1+\xi)}{\beta^2}. \tag{4}$$

Combining (3) and (4) to get rid of the exponential, we get that

$$(1+\xi)/\beta = (\alpha-\mu)/\sigma^2 \Leftrightarrow \alpha = \mu + \sigma^2(1+\xi)/\beta. \tag{5}$$

By replacing α in (3) by the expression obtained in (5) and rearranging, we get

$$\frac{(1+\xi)^2}{2\pi} = \frac{\sigma^2(1+\xi)^2}{\beta^2}\exp\left(\frac{\sigma^2(1+\xi)^2}{\beta^2}\right). \tag{6}$$

To solve (6), we make use of the Lambert W function: given an input z, $w = W(z)$ is such that $z = we^w$. We use a numerical algorithm of order four to find the zero of $z - we^w$ [9]. We let $z = (1+\xi)^2/2\pi$ in (6) and thus we have an expression for β

$$W(z) = \sigma^2(1+\xi)^2/\beta^2 \Leftrightarrow \beta(\xi,\sigma) = \sigma(1+\xi)/\sqrt{W(z)}. \tag{7}$$

To obtain an expression for α in terms of the free parameters, we replace β in (5) by its expression of (7)

$$\alpha(\xi,\mu,\sigma) = \mu + \sigma\sqrt{W(z)}. \tag{8}$$

Let $\psi = (\xi, \mu, \sigma)$ be the parameter vector of the hybrid Pareto. The hybrid Pareto density function is given by

$$h_\psi(y) = \begin{cases} f_{\mu;\sigma}(y)/\gamma, & \text{if } y \leq \alpha, \\ g_{\xi;\beta}(y-\alpha)/\gamma, & \text{if } y > \alpha \end{cases}$$

where γ is the appropriate reweighting factor so that the density integrates to one and is given by

$$\gamma(\xi) = 1 + \frac{1}{2}\left(1 + \mathrm{Erf}\left(\sqrt{W(z)/2}\right)\right)$$

where $\mathrm{Erf}(\cdot)$ is the error function $\mathrm{Erf}(z) = 2/\sqrt{\pi}\int_0^z e^{-t^2}dt$, which can be readily approximated numerically to high precision in standard ways. Fig. 1 illustrates the density and the log-density of the hybrid Pareto distribution.

Let H_ψ and $G_{\xi,\beta}$ be the distribution function of the hybrid Pareto of the GPD, respectively. Let $Y \sim H_\psi$ and $y > \alpha$, then

$$\begin{aligned} P\left(Y > y\right)) &= \frac{1}{\gamma}\left(1 - G_{\xi;\beta}(y-\alpha)\right) \\ &= P(Y > \alpha)P(Y - \alpha > y - \alpha|Y > \alpha). \end{aligned}$$

It can be seen that the right tail of the hybrid Pareto is the same as the GPD tail apart from the multiplicative factor $1/\gamma$, which is equal to $P(Y > \alpha)$, and that α acts as the threshold u in (2). Therefore, the tail approximation properties of the GPD transfer to the hybrid Pareto. The threshold is then defined as α, the junction point of the Gaussian and the GPD, and is computed implicitly as a function of the hybrid parameters [see (8)]. The hybrid Pareto thus circumvents the need for threshold selection.

B. Mixture of Hybrid Paretos

Nonparametric density estimation is a way to extract general features from large amount of data without making specific distributional assumptions. Features in which we are inter-

ested in our applications include multimodality, tail behavior, and asymmetry. An estimator is termed nonparametric [5] if its complexity, that is, the number of free parameters, grows with the size of the training set. In addition to kernel-based models in which the predictor involves a sum with $O(n)$ terms when there are n training examples, nonparametric models include mixture models when the number of components is data selected, and neural networks when the number of hidden units is data selected [5]. When the number of components is well chosen according to the number of observations, the mixture of Gaussians is convergent as a nonparametric estimator (see [6], for instance). If the tail of the generative distribution is heavy, i.e., extreme observations can occur far away in the tail, good empirical results can often be obtained by considering a mixture of Gaussians in which one of the Gaussians has a very large σ that serves to capture the points far away. One disadvantage of this approach is that the density model will only account for observed extremes and will still underestimate the density of the upper tail. This is specially true for small training sets. Another drawback is that by using symmetric components in the mixture, the lower tail tends to be overestimated. These two phenomena are illustrated in Fig. 2.

Although other heavy-tailed distributions could be used, the hybrid Pareto is a convenient density to work with when dealing with extreme observations. It shares the tail approximation properties of the GPD: heavy, light, or finite tails of all well-known distributions can be handled. Since the hybrid Pareto is continuous over the real axis, it can be used in a mixture. It can also be reversed so that the GPD part is in the lower tail and negative extremes can then be modeled. In principle, one hybrid Pareto should be enough to model the tail of the distribution and the other components could be Gaussian. This would reduce the complexity of the model and accelerate learning. However, preliminary experiments with Fréchet generated data that has a heavy upper tail showed that a mixture of hybrid Pareto components outperforms a mixture with Gaussian components and a single hybrid Pareto component. Therefore, we use a mixture with hybrid Pareto components only.

A sample result of the experiments on the unconditional density estimation with a mixture of hybrid Paretos is shown in Fig. 2. More results can be found in [4]. A training set of a 100 points was generated from a standard Fréchet distribution with tail index $\xi = 0.5$ (see Fig. 3). We trained three mixture models on these data with either hybrid Pareto (MMH), Gaussian (MMG), or log-normal (MML) components. The log-normal is asymmetric and has a heavier tail than the Gaussian distribution. We also tested the Parzen window estimator (PARZEN). We implemented the GPD with the threshold selected from the goodness-of-fit test of Choulakian and Stephens [10] to compare the estimators in the upper tail area. The lower tail of the mixture estimators is plotted in Fig. 2(a). The Fréchet has no density in that region. We clearly see that the Gaussian mixture overestimates the lower tail while the hybrid Pareto mixture decreases quickly. In Fig. 2(b), the upper tail of the estimators is shown. The Gaussian mixture underestimates the density in the upper tail area. The hybrid Pareto mixture is close to the GPD approximation and follows fairly well the generative model.

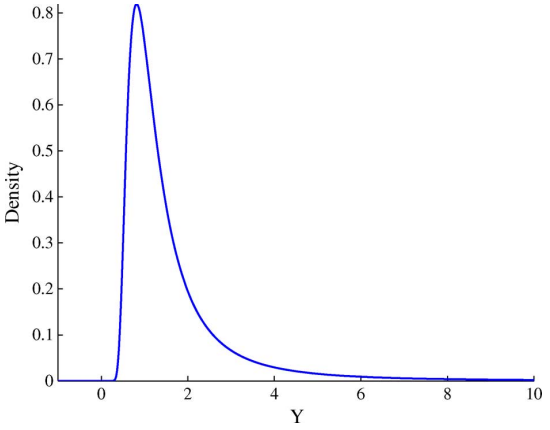


Fig. 3. Standard Fréchet density with tail index $\xi = 0.5$. The distribution function of the standard Fréchet is given by $\Phi_\xi(y) = \exp(-y^{-1/\xi})$ when $y > 0$ and is zero otherwise.

III. Conditional Density Estimation

When observations come into pairs (X, Y), with Y the scalar variable of interest and X containing inputs with predictive information over Y, it is common practice to perform regression of Y over X. This amounts to estimate the conditional expectation $E[Y|X = x]$. Robust methods have been developed to deal with extreme events [11] in the context of regression. However, we are interested in the case where the distribution of Y given X might be asymmetric and multimodal in addition to having heavy tails. In that case, the conditional expectation $E[Y|X = x]$ has no meaningful interpretation. This is why we propose to model the whole conditional distribution. Several quantities of interest such as quantiles can then be computed from the estimator of the conditional distribution.

We propose to use a new nonparametric conditional density estimator which extends the mixture of hybrid Paretos. In the nonparametric literature, few models have been developed for conditional density estimation. One class of models takes the form of double kernel estimators (see [12]) which use one kernel in the input space and one kernel in the output space. The width of each kernel controls the influence of the neighborhood in each space, and thus, bandwidth selection in both spaces is crucial as far as efficiency and convergence of the estimator are concerned [12].

A second class of conditional density models takes the form of a convex combination of component densities. The jth component density is $p(y|x, j)$ and its weight in the combination is $p(j|x)$, which may depend on the input $X = x$. The jth component density assigns the density $p(y|x, j)$ to the value y of Y, given that $X = x$. The combination weight $p(j|x)$ gives the proportion of component j in the mixture, and it can also be considered a prior for the component j, before seeing y, but given x. A generic formulation for this class of models is given by: $\hat{p}(y|x) = \sum_j p(j|x)p(y|x, j)$. The mixture of experts [13] belongs to this class. A variant of this type of model is the conditional mixture of Gaussians proposed by Bishop [14]. The main difference between these two models is that in the mixture of experts, two different functions are used to compute $p(j|x)$ and $p(y|x, j)$ whereas the conditional mixture of Gaussians uses the same function [with outputs associated with $p(j|x)$ and output

associated with $p(y|x, j)$] to compute both quantities. Bishop [14] uses a neural network to implement this function. The conditional mixture can, in principle, exploit a common representation of the inputs X (as captured by the hidden layer). The parameters of the conditional mixture are learned by minimizing the conditional negative log-likelihood by gradient descent.

A. Conditional Mixture of Hybrid Paretos

The conditional mixture of hybrid Paretos extends the mixture of hybrid Paretos to perform conditional density estimation, building on Bishop's architecture for the conditional mixture of Gaussians [14]. The parameters of the mixture of hybrid Paretos are defined as functions of the input x and the conditional estimator can then be written as follows, where m is the number of components:

$$p_\theta(y|x) = \sum_{j=1}^{m} \pi_j(x) h_{\psi_j(x)}(y) \tag{9}$$

where $h_{\psi_j(x)}(y) = p(y|x, j)$ is a hybrid Pareto density with parameter vector $\psi_j(x) = (\xi_j(x), \mu_j(x), \sigma_j(x))$ that depends on the input x and $\pi_j(x) = p(j|x)$ is the mixture proportion of component j given that x is observed. A feedforward neural network with input x and one hidden layer is used to predict the mixture proportions $\pi_j(x)$ and components parameters $\psi_j(x)$. We add a linear connection between the input and the output so that the linear case is a limit of the neural network (when there are no hidden units).

Because they only need a gradient on their output in order to be trained, neural networks are convenient classes of functions to compute the parameters of the output density, that is, to implement the functions $\pi_j(\cdot)$, $\xi_j(\cdot)$, $\mu_j(\cdot)$, and $\sigma_j(\cdot)$, given an x. Learning of the conditional mixture is performed by minimizing the conditional negative log-likelihood and this is efficiently implemented by the backpropagation algorithm for neural networks [14], albeit minimizing the average $-\log p_\theta(y|x)$ rather than the traditional squared prediction error. Moreover, by increasing the number of hidden units, neural networks can, in principle, approximate any continuous function [7]. However, any other parametrized class of functions which can be trained using the gradient with respect to mixture parameters could be used to predict the mixture parameters.

Fig. 4 depicts the architecture of the conditional mixture of hybrid Paretos. Let n_h be the number of hidden units, which controls the capacity (the number of parameters and the flexibility of the model) and let d be the dimension of the inputs. For illustration purposes, in Fig. 4, $m = 2$, $n_h = 2$, and $d = 4$. The inputs are linearly combined with weights v_{hk} and bias c_h and then nonlinearly transformed which yields the hidden unit output z_h

$$z_h = \tanh\left(c_h + \sum_{k=1}^{d} v_{hk} x_k\right).$$

The outputs of the neural network are labeled $a_j^{(i)}$, where j designates one component of the mixture and $i \in \{0, 1, 2, 3\}$ indicates that the outputs are dedicated to either $\pi_j (i = 0)$, $\xi_j (i = 1)$, $\mu_j (i = 2)$, or $\sigma_j (i = 3)$. The neural network outputs

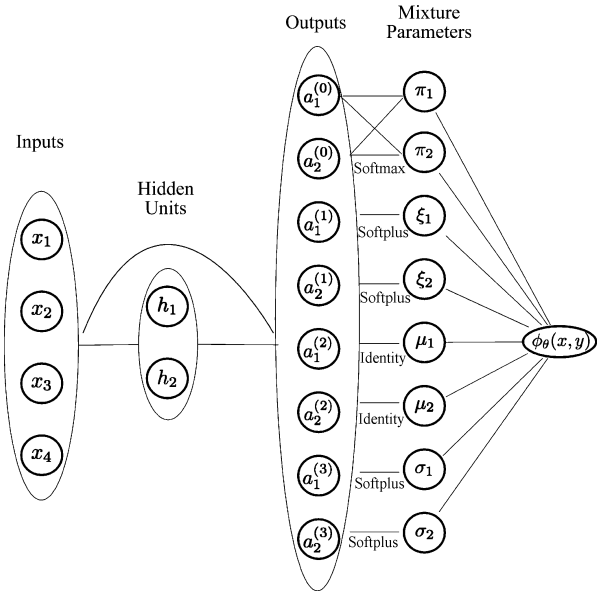


Fig. 4. Conditional mixture model. A feedforward neural network with one hidden layer and hyperbolic tangent activation function is used to predict input dependent mixture parameters. Appropriate transfer functions at the network outputs are used to impose range constraints [see (10)].

$a_j^{(i)}$ are obtained by linear combination of the hidden unit outputs [with weights $w_{jh}^{(i)}$] and of the inputs [with weights $\widetilde{w}_{jk}^{(i)}$], with additive bias $b_j^{(i)}$

$$a_j^{(i)} = b_j^{(i)} + \sum_{h=1}^{n_h} w_{jh}^{(i)} z_h(x) + \sum_{k=1}^{d} \widetilde{w}_{jk}^{(i)} x_k.$$

The transfer functions at the output of the neural network are chosen so as to impose range constraints on the parameters of the mixture. These are given explicitly in (10). The mixture component proportion $\pi_j(x) = P(j|X = x)$ is the probability that the jth component is responsible for generating y given x. It must, therefore, be positive and all the $\pi_j(\cdot)$'s must sum to one. This is ensured by a *softmax* function. A *softplus* function $(\mathrm{softplus}(x) = \log(1 + e^x))$ is used to guarantee the positivity of the $\sigma_j(\cdot)$'s and of the $\xi_j(\cdot)$'s. The *softplus* has been introduced by [15]; like the exponential, the *softplus* has a positive range but it grows slower than the exponential (asymptotically linear with slope 1) which makes numerical optimization more stable.[1] The $\mu_j(\cdot)$'s are unconstrained $a_j^{(2)}(\cdot)$'s

$$\pi_j = \exp\left(a_j^{(0)}\right) \Big/ \sum_k \exp\left(a_k^{(0)}\right) \tag{10a}$$

$$\xi_j = \log\left(1 + \exp\left\{a_j^{(1)}\right\}\right) \tag{10b}$$

$$\mu_j = a_j^{(2)} \tag{10c}$$

$$\sigma_j = \log\left(1 + \exp\left\{a_j^{(3)}\right\}\right). \tag{10d}$$

[1]Note that if $x > 0$, we have $\mathrm{softplus}(x) = x + \log(1 + e^{-x})$, and asymptotically, we have that $\lim_{x \to \pm\infty} \mathrm{softplus}(x) \to x^+$, where x^+ denotes the positive part of x.

B. Learning of the Hybrid Pareto Conditional Mixture

The free parameters of the conditional mixture model are the neural network parameters $\theta = (b, c, v, w, \widehat{w})$. These are determined by minimizing the empirical negative log-likelihood: $l(\theta) = -\sum_i^n \log p_\theta(y_i|x_i)$. We use a conjugate gradient-descent algorithm for the optimization. For each example, we obtain the gradient of the empirical negative log-likelihood with respect to θ in two steps.

1) First, compute derivatives of l with respect to $a_j^{(i)}$, $j = 1, \ldots, m$, and $i = 0, 1, 2, 3$ (the outputs of the neural network before the output transfer function; see Fig. 4).
2) Backpropagate gradients as usual through the neural network in order to obtain $\partial l/\partial\theta$. Implicitly, the resulting gradient is, therefore, obtained through

$$\frac{\partial l}{\partial \theta} = \sum_i \sum_j \frac{\partial l}{\partial a_j^{(i)}} \frac{\partial a_j^{(i)}}{\partial \theta}.$$

Since the derivative in step 2) is standard in neural network applications (see [14]), we describe only the derivative in step 1). Let $l = -\log(p_\theta(y|x)) \stackrel{\text{def}}{=} -\log(\phi_{\theta(x)}(y))$ be the value of the error function, for example, (x, y) where $p_\theta(y|x)$ is given in (9) and $\theta(x) = (\pi_1(x), \ldots, \pi_m(x), \psi_1(x), \ldots, \psi_m(x))$ are the mixture parameters for input x. From Fig. 4, we see that the derivative $\partial l/\partial a_j^{(i)}$ can be separated into two different cases depending on the value of i.

— If $i = 0$ (component proportions), $a_j^{(0)}$ is one of the outputs controlling the proportions (for component j) and its derivative can be expressed as

$$\frac{\partial l}{\partial a_j^{(0)}} = \frac{\partial l}{\partial \phi_{\theta(x)}(y)} \sum_{k=1}^{m} \frac{\partial \phi_{\theta(x)}(y)}{\partial \pi_k} \frac{\partial \pi_k}{\partial a_j^{(0)}}. \tag{11}$$

— On the other hand, if $1 \leq i \leq 3$, $a_j^{(i)}$ governs one of the hybrid Pareto component parameter and its derivative is simpler

$$\frac{\partial l}{\partial a_j^{(i)}} = \frac{\partial l}{\partial \phi_{\theta(x)}(y)} \frac{\partial \phi_{\theta(x)}(y)}{\partial \psi_{j,i}(x)} \frac{\partial \psi_{j,i}(x)}{\partial a_j^{(i)}} \tag{12}$$

where $\psi_{j,i}(x)$ is the ith element of $\psi_j(x)$, the parameter vector of the jth component.

Each partial derivative in (11) and (12) is developed next. In both equations, we have $\partial l/\partial \phi_{\theta(x)}(y) = -1/\phi_{\theta(x)}(y)$. When the derivative is taken with respect to one of the mixture component proportions, we have, for $1 \leq j \leq m$

$$\frac{\partial \phi_{\theta(x)}(y)}{\partial \pi_j} = h_{\psi_j(x)}(y).$$

Differentiating with respect to the hybrid Pareto parameters $\psi_{j,i}(x)$, we get

$$\frac{\partial \phi_{\theta(x)}(y)}{\partial \psi_{j,i}(x)} = \pi_j(x) \frac{\partial}{\partial \psi_{j,i}(x)} h_{\psi_j(x)}(y).$$

The derivatives of the proportions and of the mixture parameters with respect to the network outputs are

$$\frac{\partial \pi_k}{\partial a_j^{(0)}} = \begin{cases} \pi_j(1 - \pi_j), & \text{if } j = k \\ -\pi_k \pi_j, & \text{if } j \neq k \end{cases}$$
$$\frac{\partial \xi_j}{\partial a_j^{(1)}} = 1 - \exp(-\xi_j)$$
$$\frac{\partial \mu_j}{\partial a_j^{(2)}} = 1$$
$$\frac{\partial \sigma_j}{\partial a_j^{(3)}} = 1 - \exp(-\sigma_j).$$

IV. Experiments

We first ran experiments on synthetic data in order to evaluate how the conditional mixture of hybrid Paretos performs when the generative model is known. We use the Fréchet distribution as the basis to generate synthetic data. This is motivated by the fact that most heavy-tailed distributions asymptotically have the same tail behavior as the Fréchet distribution. We then evaluate the hybrid Pareto conditional mixture on two real data sets.

A. Experimental Setup

We compare the conditional mixture of hybrid Paretos with conditional mixture models that have a different type of component density: Gaussian or log-normal. The log-normal is an asymmetric distribution with a heavier tail than the Gaussian. We also compare the double kernel estimator [16] and [12] with Gaussian kernels and diagonal covariance matrices. We use the abbreviated labels CMMH, CMMG, and CMML to refer to a conditional mixture model with, respectively, hybrid Pareto (CMMH), Gaussian (CMMG), and log-normal (CMML) components. We use the DKERNEL label to refer to the double kernel estimator [12].

All conditional mixture parameters (such as means, variances, and hybrid Pareto parameters for all the mixture components) are predicted by means of a one-layer feedforward neural network [17], as per Fig. 4. The parameters of the neural network are learned by minimizing the negative conditional log-likelihood $-\log p(y|x)$ averaged over training set examples (x, y). Regardless of the type of component, a conditional mixture has two hyperparameters: the number of hidden units (in the neural network) and the number of components (in the mixture). The double kernel estimator has two hyperparameters as well: the bandwidth in the input space and the bandwidth in the output space. All hyperparameters are selected on a validation set that is obtained by removing 20% of the original training set (which is a standard way of proceeding). A large independent test set is used to estimate generalization performance. Different values of each hyperparameter are considered and all combinations of these values are tried, in such a way that the finally selected values are not on the boundary of the set of values tried, so that these selected values are roughly near a minimum (possibly local) in the

TABLE I
AVERAGE RELATIVE LOG-LIKELIHOOD (STD. ERR) ON TEST DATA WITH TRAINING SET SIZE n. THE SMALLER THIS CRITERION IS, THE BETTER THE ESTIMATOR IS PERFORMING. THE BEST PERFORMANCES ARE IN BOLD. THE DATA WAS GENERATED ACCORDING TO A CONDITIONAL FRÉCHET DISTRIBUTION

	n	$\overline{\mathcal{R}}_{CMMH}$ (std. err)	$\overline{\mathcal{R}}_{CMMG}$ (std. err)
lin-mod	200	**0.1081 (0.01621)**	0.6401 (0.07695)
	2000	**0.006835 (0.0002546)**	0.1919 (0.02726)
sin-mod	200	**0.7014 (0.06141)**	2.338 (0.9337)
	2000	**0.1834 (0.005721)**	0.3024 (0.02447)
lin-heavy	200	**0.4176 (0.03305)**	6.583e+07 (5.612e+07)
	2000	**0.01875 (0.0003348)**	1450 (1027)
sin-heavy	200	**0.597 (0.04247)**	8.112e+05 (5.067e+05)
	2000	**0.1783 (0.006356)**	7.403e+06 (7.128e+06)
	n	$\overline{\mathcal{R}}_{CMML}$ (std. err)	$\overline{\mathcal{R}}_{DKERNEL}$ (std. err)
lin-mod	200	0.3059 (0.02426)	2.54 (0.3703)
	2000	0.1255 (0.002318)	1.459 (0.2688)
sin-mod	200	0.9909 (0.07663)	2.038 (0.2649)
	2000	0.3028 (0.005838)	1.399 (0.2735)
lin-heavy	200	**0.476 (0.04296)**	8.377e+04 (4.578e+04)
	2000	0.1761 (0.008315)	3.929e+05 (3.841e+05)
sin-heavy	200	2.335 (0.9769)	9.563e+05 (5.717e+05)
	2000	0.2628 (0.006283)	1.807e+06 (1.278e+06)

TABLE II
AVERAGE HYPERPARAMETERS SELECTED ON THE VALIDATION SET FOR THE CONDITIONAL FRÉCHET GENERATED DATA WITH TRAINING SET SIZE n. FOR CONDITIONAL MIXTURE MODELS, h_{CMM} IS THE NUMBER OF HIDDEN UNITS AND m_{CMM} IS THE NUMBER OF COMPONENTS. FOR THE DOUBLE KERNEL ESTIMATOR, λ_x AND λ_y ARE THE BANDWIDTH IN INPUT SPACE AND OUTPUT SPACE, RESPECTIVELY

	n	$(h_{\mathrm{CMMH}}, m_{\mathrm{CMMH}})$	$(h_{\mathrm{CMMG}}, m_{\mathrm{CMMG}})$
lin-mod.	200	(0, 1.94)	(0.1, 2.47)
	2000	(0, 2)	(0.02, 2.54)
sin-mod.	200	(2.53, 2.34)	(2.34, 3.07)
	2000	(3.97, 4.36)	(4.35, 5.81)
lin-heavy	200	(0.1, 1.64)	(0, 2.11)
	2000	(0, 2)	(0.02, 3.75)
sin-heavy	200	(1.71, 2.43)	(0.78, 3.1)
	2000	(2.16, 6.18)	(2.73, 5.3)
	n	$(h_{\mathrm{CMML}}, m_{\mathrm{CMML}})$	(λ_x, λ_y)
lin-mod.	200	(0.01, 2.2)	(0.1106, 0.2494)
	2000	(0.05, 2.2)	(0.0586, 0.1495)
sin-mod.	200	(2.45, 2.8)	(1.006, 0.541)
	2000	(3.51, 4.5)	(0.00325, 0.1882)
lin-heavy	200	(0.12, 2.3)	(7.448, 24.77)
	2000	(0.13, 3.34)	(5.896, 41.9)
sin-heavy	200	(0.74, 2.83)	(16.87, 31.72)
	2000	(1.27, 6.98)	(35.7, 58.15)

space of hyperparameter values. The selected values are shown in Tables II and III for the artificial data experiments, and in Tables V and VII for the real data experiments.

In the EVT community, two methods have been proposed to adapt the peaks-over-threshold (PoT) method, which is based on the approximation of the exceedances by the GPD, to conditional density estimation. Davison and Smith [8] modeled the GPD parameters as functions of the input x. They chose to use a linear combination of the input followed by an exponential to ensure positivity of both GPD parameters. They use an unconditional threshold which is chosen by successive trials. McNeil and Frey [18] used the PoT method in a time-series framework. They used a time-series model on financial data [a AR(1)-GARCH(1,1) model] and then applied the PoT method on the residuals. Building on this idea, we propose to use a

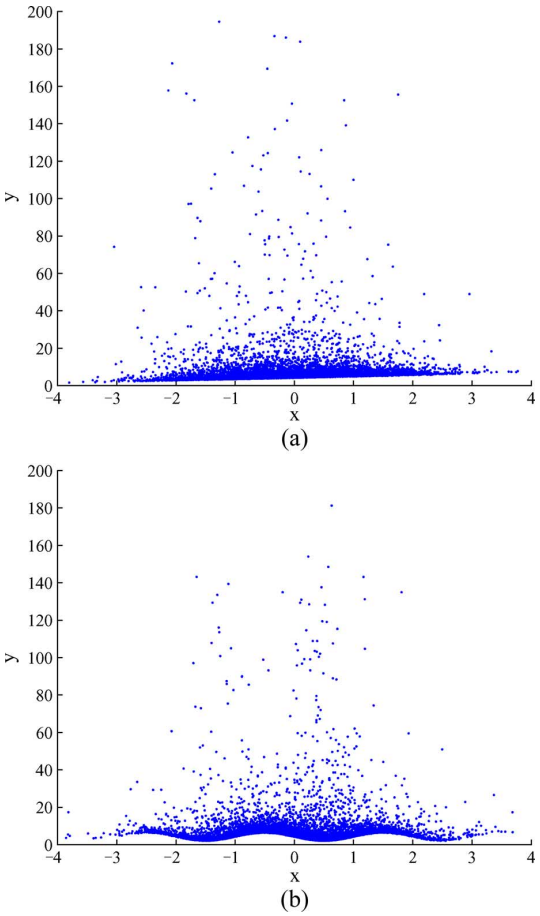


Fig. 5. Conditional Fréchet generated data with very heavy-tail index. (a) Linearly dependent parameters. (b) Sinusoidal dependent parameters. The y-axis is shortened so that the shape of the dependence can be seen ($\max y \approx 10^6$).

neural network to model the dependence of the output variable with respect to the input variable and then to apply the PoT method to the residuals. Let $f(x, \theta)$ be the function computed by the neural network with parameter vector θ and input x. The threshold u for the residuals is taken to be a sample quantile of level q_{PoT}. The threshold for the dependent variable y is thus given by $u(x) = f(x, \theta) + u$. Above $u(x)$, exceedances are modeled by a GPD. We call this model the conditional PoT (labeled with the abbreviation CPOT). The neural network parameters are learned by minimizing the sum-of-square error. The number of hidden units is chosen on a validation set that consists of 20% of the training set. The quantile level which determines the threshold is chosen with the GPD goodness-of-fit test of Choulakian and Stephens [10]. We start with a small quantile level and increase it until the GPD parameter estimators satisfy the goodness-of-fit test. This threshold method does not require any manual tuning and is easy to implement. It is thus convenient for the purpose of large scale experiments.

B. Simulation Study

To obtain (x, y) data in which the y density depends on the value of a variable x, we generated y examples from a Fréchet

TABLE III
AVERAGE HYPERPARAMETERS SELECTED FOR THE CONDITIONAL POT METHOD (NUMBER OF HIDDEN UNITS h AND QUANTILE LEVEL q_{PoT} DETERMINING THE THRESHOLD) ON THE CONDITIONAL FRÉCHET GENERATED DATA AND TRAINING SET SIZE n. THE AVERAGE ESTIMATED TAIL INDEX $\hat{\xi}$ IS ALSO GIVEN

	n	h	q_{PoT}	$\hat{\xi}$
lin- mod	200	0	0.7075	0.2936
	2000	0	0.7095	0.3131
sin-mod	200	2.93	0.685	0.2643
	2000	3.92	0.747	0.3736
lin-heavy	200	0.19	0.8165	0.7728
	2000	0.02	0.8865	0.9615
sin-heavy	200	0.46	0.751	0.8983
	2000	0.02	0.893	1.1427

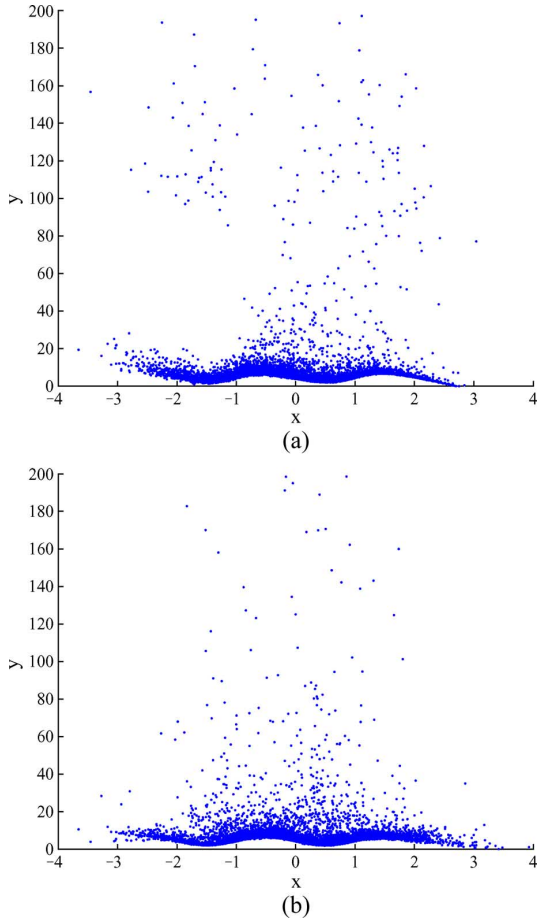


Fig. 6. Ten thousand data points generated from CMMH with a training set of (a) 200 points and (b) 2000 points.

distribution whose three parameters (the tail index ξ, the location μ, and the spread σ)[2] are conditionally dependent of the input x. To compute the Fréchet parameters from x, we used either a linear ($l(x) = ax + b$) or a sine-shaped ($s(x) = c\sin(ax + b) + d$) functional. The input variable x is sampled from a standard normal distribution. The parameters of the functional dependency (e.g., a and b in the linear case) are chosen so that each parameter of the Fréchet belongs to a given interval most of the time. In particular, the tail index ξ is chosen to lie either in the interval $[0.25\ 0.5]$, which corresponds to a moderately heavy tail or in the interval $[0.66\ 1.33]$, which corresponds to a heavier tail. Considering the two types of functional dependencies and the two types of conditional tails, this gives us four distinct generative models. The generative models with the heavier tail indexes are illustrated in Fig. 5. Again, we use shortcut labels to refer to these generative models: "lin" and "sin" stand, respectively, for linear and sine-shaped dependency of the Fréchet parameters, "mod" and "heavy" refers to moderately heavy and heavier tail of the conditional Fréchet. We considered two training set sizes 200 and 2000; the test set size in both cases is 10 000. Each experiment is repeated 100 times so that an average performance measure can be computed along with its standard error.

[2]The distribution function of the Fréchet with three parameters is $\Phi_{\xi,\mu,\sigma}(y) = \exp\{-((y-\mu)/\sigma)^{-1/\xi}\}$ when $y > \mu$ and zero otherwise.

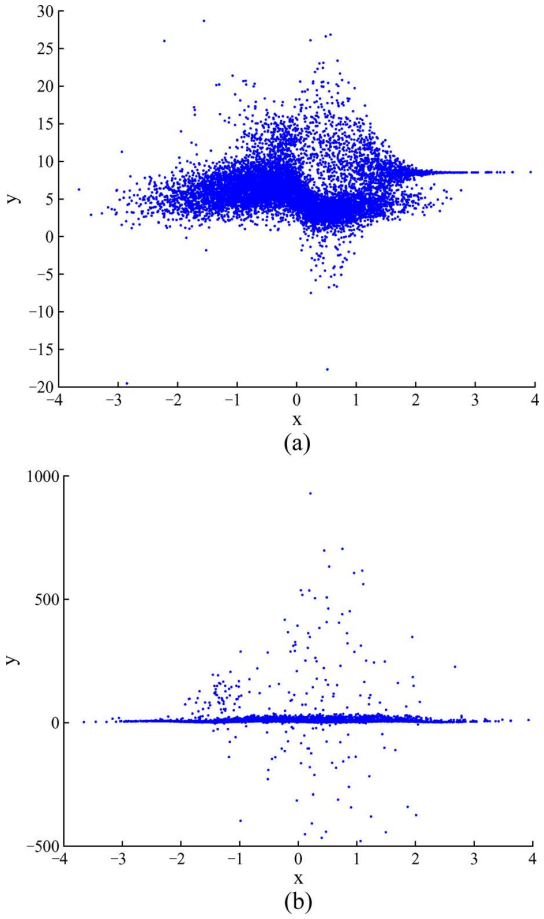


Fig. 7. Ten thousand data points generated from CMMG with a training set of (a) 200 points and (b) 2000 points.

1) Test Results: The conditional density estimators are compared in terms of out-of-sample relative log-likelihood

$$\mathcal{R}_{p_\theta}(\mathcal{D}_l) = -\frac{1}{l}\sum_{i=1}^{l}\log\left(\frac{p_\theta(y_i|x_i)}{\mathcal{P}(y_i|x_i)}\right) \tag{13}$$

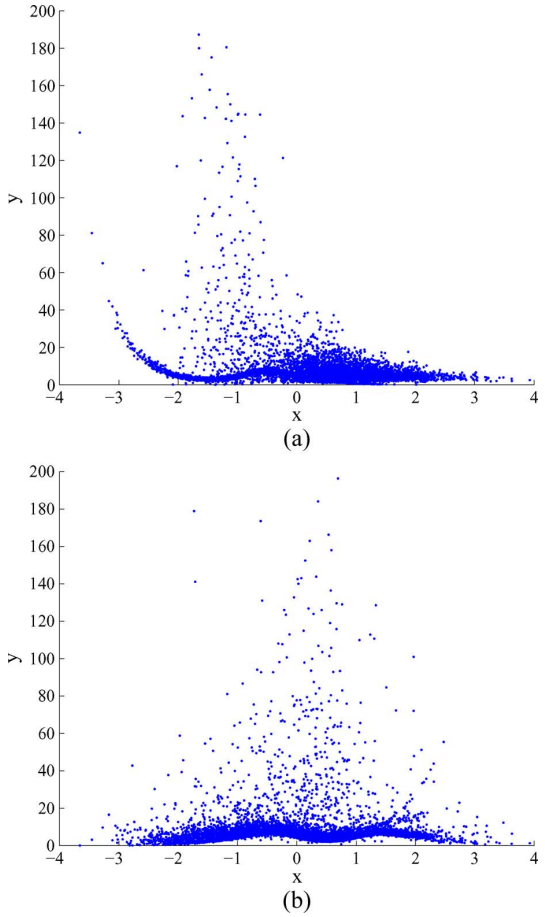


Fig. 8. Ten thousand data points generated from CMML with a training set of (a) 200 points and (b) 2000 points.

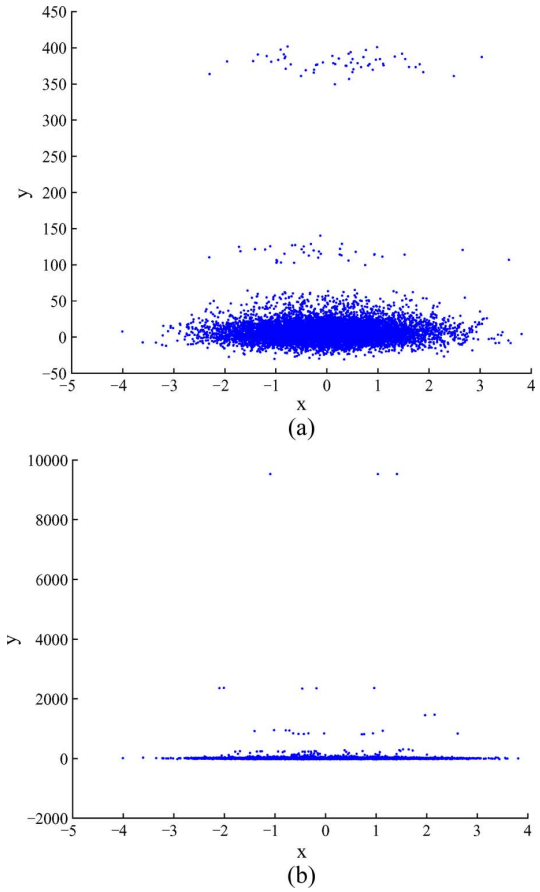


Fig. 9. Ten thousand data points generated from DKERNEL with a training set of (a) 200 points and (b) 2000 points.

where $\mathcal{P}(\cdot|\cdot)$ is the conditional density function of the data generating model (i.e., the target density), $p_\theta(\cdot|\cdot)$ is the conditional density function of the estimator, and the sum is over the test set $\mathcal{D}_l$. The smaller the $\mathcal{R}_{p_\theta}$ criterion is, the better the estimator is performing. The results for the $\mathcal{R}_{p_\theta}$ criterion for the four generative models are presented in Table I and the average selected hyperparameters are presented in Table II. There are two types of difficulty in learning these synthetic data: the heaviness of the tail and the shape of the functional dependency. In all cases but one, the hybrid Pareto conditional mixture outperforms significantly the other models considered in terms of relative log-likelihood. The log-normal conditional mixture is in one instance comparable to the hybrid Pareto conditional mixture and offers, in general, a reasonable model. On the other hand, the Gaussian conditional mixture is not able to model the heavier tailed conditional Fréchet properly. Similarly, the double kernel estimator's performance degrades severely in the heavier tail case. The interesting point to notice regarding the hyperparameter selection is that, for the conditional mixtures, fewer hidden units are selected for the sine-shaped dependency when the tail is heavier. This is particularly striking for the lighter tailed component mixtures such as CMMG and CMML. It seems that the heaviness of the tails interferes with the ability of the neural network to capture the dependency relationship between x and the density of y.

Table III presents the average selected hyperparameters for the conditional PoT method for the conditional Fréchet generated data and the average estimated tail index in the fourth column. As for the conditional mixture models, when the tail is heavier, the number of hidden units selected for the sine-shaped dependency is lower. The threshold selected generally increases with the training set size, as expected from trading off optimally bias and variance. The average tail index estimates are within the range of the conditional index of the generative model.

2) Estimators as Generative Models: We generated data from the trained estimators for one sample experiment of each of the four generative models. We present the results only for the case of the conditional Fréchet with sine-shaped dependency and very heavy tail (referred to as sin heavy). This case highlights well the differences among conditional density estimation models. Figs. 6–8 show the data generated from the conditional mixtures with hybrid Pareto, Gaussian, and log-normal components, respectively. The hybrid Pareto conditional mixture seems to capture well both the dependency functional and the conditional tail of the conditional Fréchet.

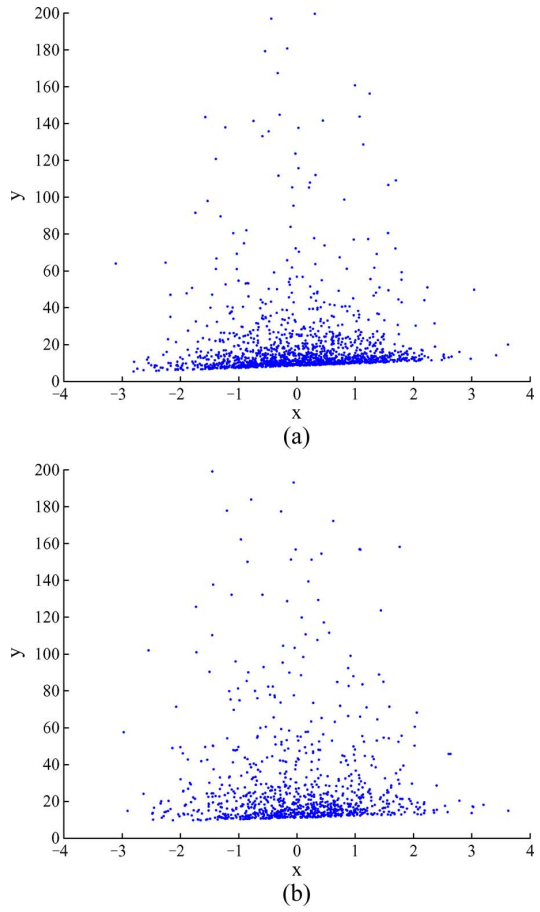


Fig. 10. Ten thousand data points generated from the conditional PoT model with a training set of (a) 200 points and (b) 2000 points.

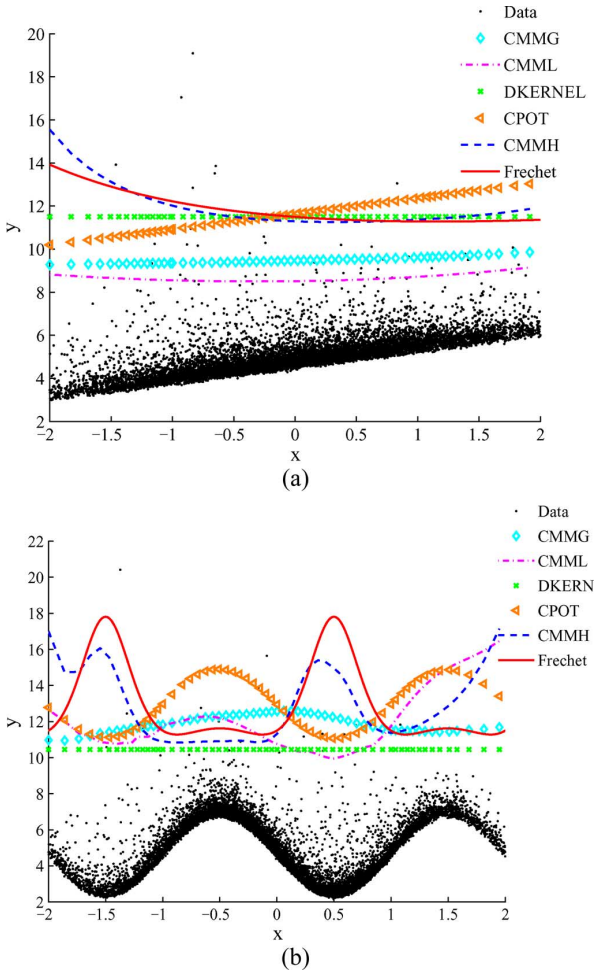


Fig. 11. Average estimated conditional quantiles from trained models on 2000 points generated from a conditional Fréchet with moderately heavy tail. The dependence functional is (a) linear and (b) sine-shaped. Quantile level is 0.999. Test data are also illustrated. CMMH recovers fairly well the shape of the conditional quantile whereas the other estimators fail to do so.

Besides not capturing very well the dependency functional, the Gaussian conditional mixture exhibits the problem described in Section II-B on nonparametric density estimation. One component has a large standard deviation to account for the largest observations, but since the Gaussian is symmetric, there is undesirable density in the lower tail. This is particularly apparent when the training set size is larger. The log-normal conditional mixture is behaving more reasonably. In this case, it can be seen from Fig. 9 that the double kernel estimator does not capture well the dependency relationship, whereas for the moderately tail case, it was able to uncover the sinusoidal shape of the data. The double kernel estimator puts density around extremes seen in the training set. In Fig. 10, we see that the conditional PoT method does not capture the sine-shaped dependency whereas it was able to detect it for the moderately tail case.

3) Quantile Estimation: In order to evaluate how the conditional density estimators perform in the tail area, we compute from each model estimated quantiles of level 0.99 and 0.999. Since the quantiles do not exist in closed form for the models considered (except for the conditional PoT method), we approximate them numerically. When there is little data, the hybrid Pareto conditional mixture accounts for extreme observations by relying on large tail indexes of some of the components. This gives rise to a good performance in terms of log-likelihood on test data but it also results in overestimation of the extreme quantiles. When the generative model is the conditional Fréchet with moderately heavy tail and the training set has 2000 points, the hybrid Pareto conditional mixture offers an interesting performance. We present the quantile estimations for the quantile level of 0.999 in Fig. 11. The hybrid Pareto conditional mixture is able to recover the conditional quantile shape fairly well whereas the other estimators fail to do so. However, to target specifically conditional quantile estimation, the proposed estimator would need to be adjusted.

C. Real Data Sets

For the experiments with real data sets, since the generative model is unknown, we compute a comparative measure, the performance that is the difference in average out-of-sample nega-

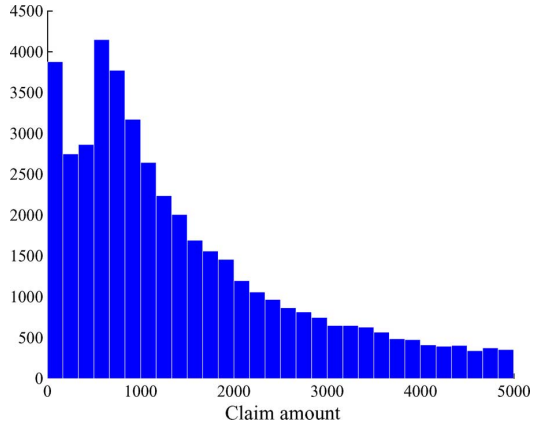


Fig. 12. Histogram of the unconditional distribution of the positive claims divided by the duration of the policy of an insurance company smaller than 5000 dollars. The multimodality (at least two modes) and heaviness of the tail are obvious.

TABLE IV
AVERAGE LOG-LIKELIHOOD (STD. ERR.) RELATIVE TO THE HYBRID PARETO CONDITIONAL MIXTURE (CMMH) ON TEST DATA FOR THE INSURANCE DATA SET. THE TRAINING SET SIZE IS n. POSITIVE VALUES OF THE RELATIVE LOG-LIKELIHOOD INDICATE THAT CMMH PERFORMED BETTER. THE RESULT IN ITALIC INDICATES A PERFORMANCE NOT SIGNIFICANTLY POSITIVE

n	$\overline{\mathcal{R}}_{\text{CMMG}}$	$\overline{\mathcal{R}}_{\text{CMML}}$	$\overline{\mathcal{R}}_{\text{DKERNEL}}$
200	2.774 (0.1312)	0.1983 (0.01412)	78.2 (56.23)
2 000	0.3562 (0.06264)	0.7903 (0.02752)	68.45 (55.18)
20 000	*0.0687 (0.03117)*	0.03902 (0.01603)	54.55 (50.44)

TABLE V
SELECTED HYPERPARAMETERS ON THE VALIDATION SET FOR THE INSURANCE DATA SET AND TRAINING SET SIZE n. FOR CONDITIONAL MIXTURES, h_{CMM} IS THE NUMBER OF HIDDEN UNITS AND m_{CMM} IS THE NUMBER OF COMPONENTS. FOR THE DOUBLE KERNEL ESTIMATOR, λ_x AND λ_y ARE THE BANDWIDTHS IN INPUT AND OUTPUT SPACE, RESPECTIVELY

n	$(h_{\text{CMMH}}, m_{\text{CMMH}})$	$(h_{\text{CMMG}}, m_{\text{CMMG}})$
200	(5, 2)	(15, 2)
2 000	(0, 2)	(0, 4)
20 000	(0, 8)	(0, 16)
n	$(h_{\text{CMML}}, m_{\text{CMML}})$	(λ_x, λ_y)
200	(15, 16)	(1e+05, 1e+06)
2 000	(5, 1)	(1e+05, 1e+06)
20 000	(0, 2)	(1, 1e+06)

tive log-likelihood. Let (x, y) be a particular data point; the relative performance measure is then written as

$$\mathcal{R}_{\text{other}}(x, y) = \log\left(p_\theta^{\text{CMMH}}(y|x)\right) - \log\left(p_{\hat{\theta}}^{\text{other}}(y|x)\right)$$

where $p_\theta^{\text{CMMH}}(\cdot|\cdot)$ is for the conditional mixture of hybrid Paretos and $p_{\hat{\theta}}^{\text{other}}(\cdot|\cdot)$ is for a competing estimator. Positive $\mathcal{R}_{\text{other}}(x, y)$ values indicate that the conditional mixture of hybrid Paretos performs better than the competing estimator.

In order to get an idea of the conditional densities estimated by the various models considered, we plot for 20 input points $\{x_{j_1}, \ldots, x_{j_{20}}\}$ selected randomly from the test set, the conditional density curves $p(Y|X = x_{j_i})$ for each model. In this case, we do not give the results from the PoT model since what we are really interested in is a conditional density model of the whole distribution, and the PoT model can only provide tail probability estimates.

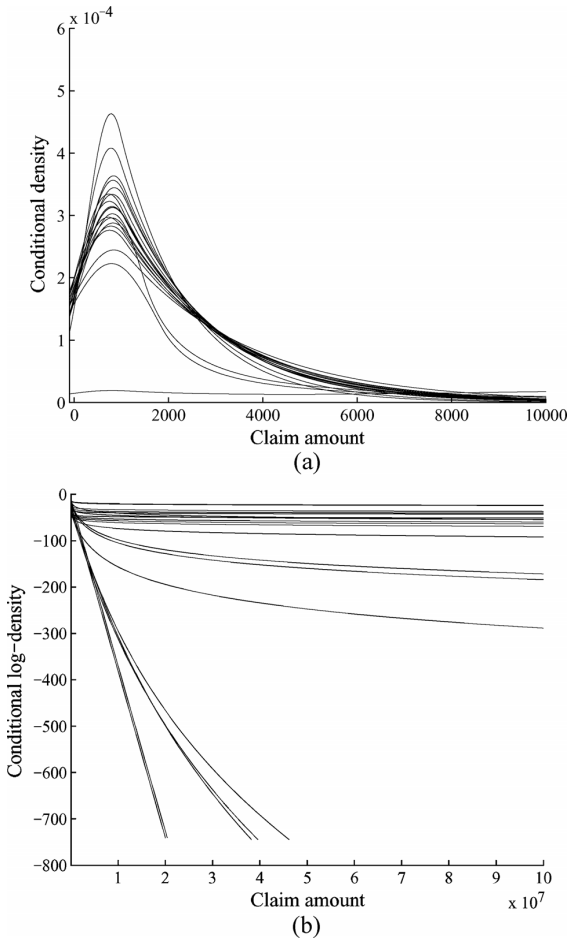


Fig. 13. Conditional densities for the hybrid Pareto conditional mixture models trained on the insurance data set with a training set of 200 observations. (a) Densities in the central area and (b) in the upper tail in logarithmic scale. The hybrid Pareto conditional mixture is able to predict various types of tail decay and is less affected by extremes for the estimation of the central area.

1) Insurance Data: We first used insurance data graciously provided by an anonymous insurance company. The complete distribution of the claims includes a mass point at zero. One way to deal with this is to use a probabilistic classifier that predicts, given a client profile X, the most probable class ($\text{claim} = 0$ or $\text{claim} > 0$). For the second class, we need to estimate $p(\text{claim}|X, \text{claim} > 0)$ and this is the part of the problem we addressed here. This is why the records used in the experiments are only for policies that had a nonzero claim. The data set consists of data from one year with 54 119 records with positive claims. The dependent variable Y is the claim amount divided by the duration of the policy. The largest claim is in the order of a million dollars while the average claim is in the order of a thousand dollars. The 75% of the claims are smaller than the average claim. The input variable X is a vector of 140 numbers, mostly binary indicators, describing the client profile. In Fig. 12,

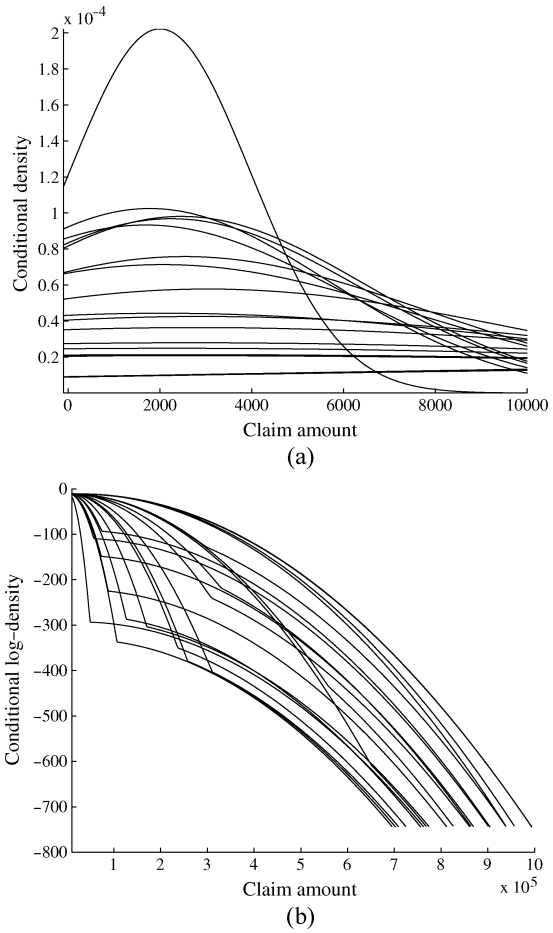


Fig. 14. Conditional densities for the Gaussian conditional mixture models trained on the insurance data set with a training set of 200 observations. (a) Densities in the central area and (b) in the upper tail in logarithmic scale.

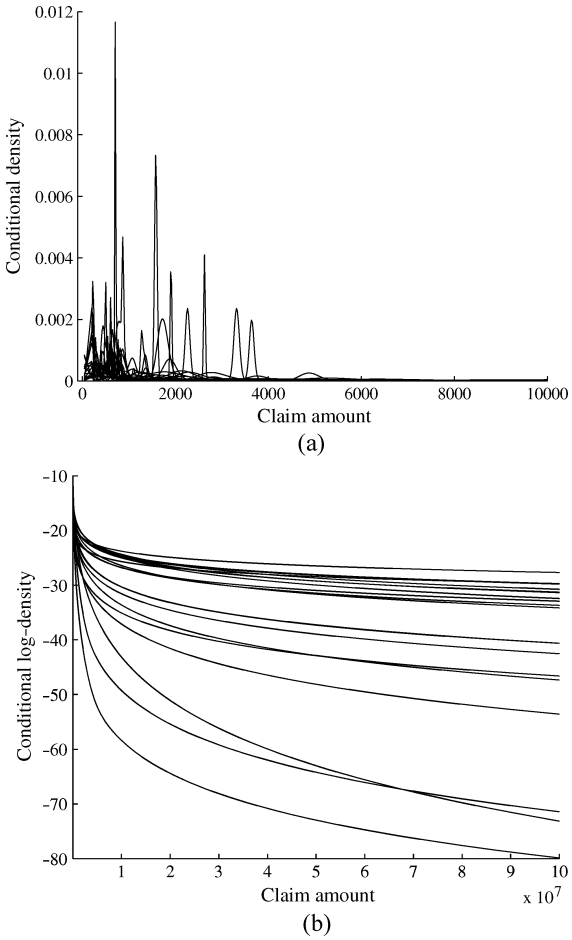


Fig. 15. Conditional densities for the log-normal conditional mixture models trained on the insurance data set with a training set of 200 observations. (a) Densities in the central area and (b) in the upper tail in logarithmic scale.

the unconditional distribution of the positive claims below 5000 dollars is illustrated by means of a histogram. We can clearly see that the unconditional distribution is multimodal and heavy tailed.

To study the effect of the training set size on the estimators considered, we trained the estimators on 200, 2000, and 20 000 data points. The numeric inputs have been standardized. Principal component analysis has been applied on the input variables to reduce dimensionality: enough components (between 49 and 69 depending on the training set size) were retained to explain 90% of input variance. The training, hyperparameter selection, and testing procedure are as before, with hyperparameters selected using 20% of the training set for cross validation. The test set is what remains from the 54 119 examples from the original data set after having removed the above 200, 2000, or 20 000 examples (for training and validation).

The relative log-likelihood results are presented in Table IV. The hybrid Pareto conditional mixture outperforms the other models considered in all instances although, in one occasion, the performance is not significantly positive. The difference in performance, indicated by a large positive value of the relative log-likelihood, is particularly important for the smaller training set sizes. The selected hyperparameters are shown in Table V. We

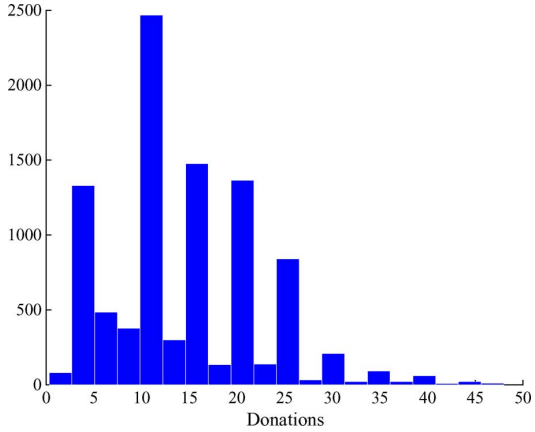


Fig. 16. Histogram of the unconditional distribution of the positive donations below 50 dollars of the KDD cup 98 data set. Donations that are multiple of five dollars are particularly frequent.

TABLE VI
AVERAGE LOG-LIKELIHOOD (STD. ERR.) RELATIVE TO THE HYBRID PARETO CONDITIONAL MIXTURE (CMMH) ON TEST DATA FOR THE KDD CUP 98 DATA SET. THE TRAINING SET SIZE IS n. POSITIVE VALUES OF THE RELATIVE LOG-LIKELIHOOD INDICATE THAT CMMH PERFORMED BETTER

n	$\overline{\mathcal{R}}_{\text{CMMG}}$	$\overline{\mathcal{R}}_{\text{CMML}}$	$\overline{\mathcal{R}}_{\text{DKERNEL}}$
200	0.5818 (0.04452)	2.045 (0.02841)	108.3 (47.93)
2 000	1.03 (0.0332)	2.001 (0.02726)	431.2 (405.3)

TABLE VII
SELECTED HYPERPARAMETERS ON THE VALIDATION SET FOR THE KDD CUP 98 DATA SET AND TRAINING SET SIZE n. FOR CONDITIONAL MIXTURES, h_{CMM} IS THE NUMBER OF HIDDEN UNITS AND m_{CMM} IS THE NUMBER OF COMPONENTS. FOR THE DOUBLE KERNEL ESTIMATOR, λ_x AND λ_y ARE THE BANDWIDTHS IN INPUT AND OUTPUT SPACE, RESPECTIVELY

n	$(h_{\text{CMMH}}, m_{\text{CMMH}})$	$(h_{\text{CMMG}}, m_{\text{CMMG}})$
200	(20, 20)	(5, 40)
2 000	(20, 50)	(10, 20)
n	$(h_{\text{CMML}}, m_{\text{CMML}})$	(λ_x, λ_y)
200	(10, 8)	(10, 0.1)
	(15, 30)	(10, 0.01)

note that the selected complexity level is higher for the smaller training set sizes. This counterintuitive result is explained by the variability of hyperparameter selection for the smaller training set sizes. The double kernel estimator is not able to extrapolate to extremes unseen in the training set. It has a high density near the training set points and low density elsewhere resulting in large standard errors of the relative log-likelihood.

Figs. 13–15 illustrate the conditional densities of the conditional mixture models with hybrid Pareto, Gaussian, and log-normal components, respectively. The training set had 200 points and 20 random test points were chosen to compute the conditional densities. From the variety of the curves, we can conclude that these models take advantage of the relationship between the input and the output. Multimodality is captured either by the number of hidden units which allow flexible modeling of the location and the height of the components or by the number of components itself. The hybrid Pareto conditional mixture densities are unimodal with various shapes and locations. There is a variety of type of tails, from light to heavy. Given the client profile, the hybrid Pareto conditional mixture is thus able to predict various types of behavior, ranging from low risk (light tail) to high risk (very heavy tail) from the insurer's perspective. The Gaussian conditional mixture densities are often spread out to model the extremes and all the upper tails decrease rapidly. Notice in Fig. 14 how one of the Gaussian components typically tries to capture the tail (with a much larger variance), creating an apparent discontinuity of the first derivative of the log-density (beyond that point the component with large variance dominates). The log-normal conditional mixture densities present a lot of artifacts in the central part due to the large number of components selected. The tails are all in the same range; a lot heavier than their Gaussian counterpart. The double kernel estimator gives the same conditional density curves (not shown here) whatever the value of the input. This means that this estimator does not capture any dependency between the input and the output variables. Also, the upper tails

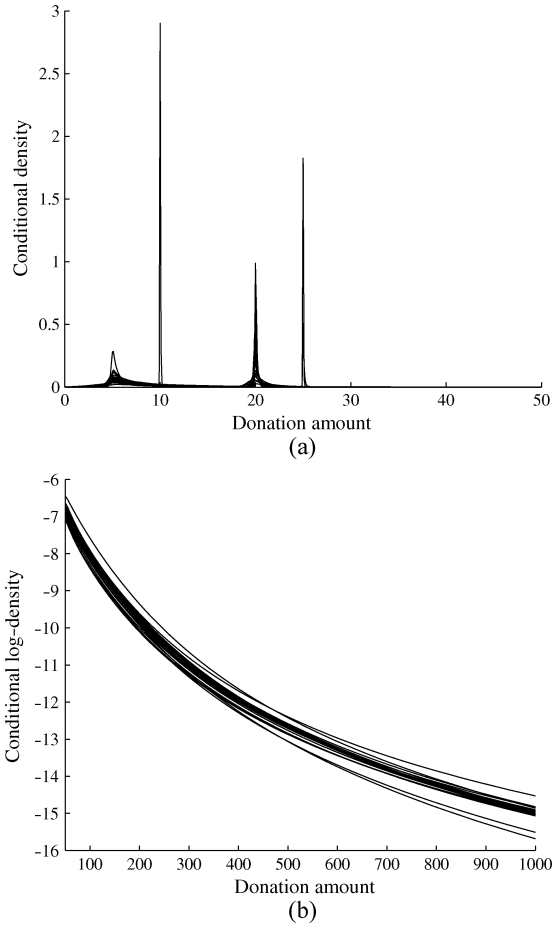


Fig. 17. Conditional densities for the hybrid Pareto conditional mixture models trained on the KDD cup 98 data set with a training set of 200 observations. (a) Densities in the central area and (b) in the upper tail in logarithmic scale.

of the conditional densities decrease quickly after the largest claim of the training set.

2) KDD Cup 98 Data: In the last set of experiments, we used the data set provided by the Fourth International Conference on Knowledge Discovery and Data Mining (KDD Cup 98). The dependent variable Y is the amount donated to a national veterans organization. The input variable X has 479 fields describing the donor profile. A binary variable indicates whether a person responded to the promotion; the donation amount is only observed when this variable is on. Just like for the insurance data set, a probabilistic classifier could be used to predict, given X, the probability that the person will make a positive donation. However, we addressed only the problem of estimating $p(Y|X, Y > 0)$. We thus have a total of 9716 positive donation records. We note that 75% of the donations are less or equal to 20 dollars although some donations go all the way up to 500 dollars. The target variable takes value in a discrete set containing mainly integer numbers; the amounts corresponding to multiple of 5 dollars are especially frequent as can be seen from the histogram of the donation amounts of Fig. 16.

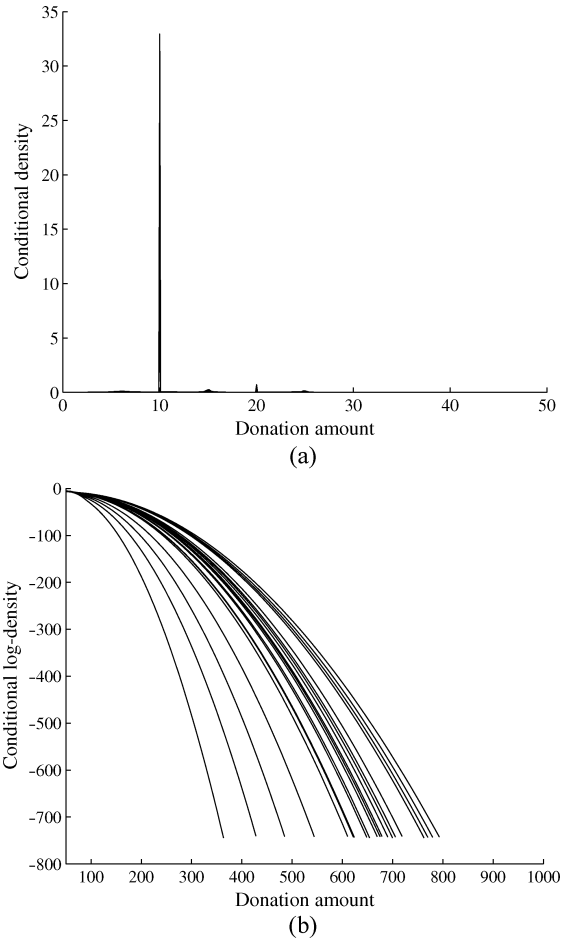


Fig. 18. Conditional densities for the Gaussian conditional mixture models trained on the KDD cup 98 data set with a training set of 200 observations. (a) Densities in the central area and (b) in the upper tail in logarithmic scale.

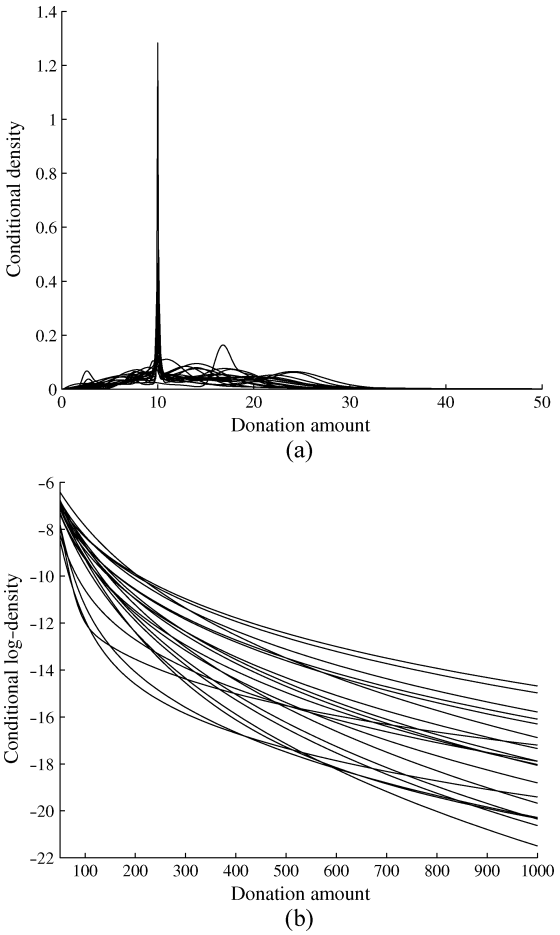


Fig. 19. Conditional densities for the log-normal conditional mixture models trained on the KDD cup 98 data set with a training set of 200 observations. (a) Densities in the central area and (b) in the upper tail in logarithmic scale.

We followed [19] for preprocessing, yielding five input variables. The inputs have been standardized. We trained the estimators on 200 and 2000 data points. The hyperparameters have been selected as before using 20% of the training set as a validation set. The remaining examples (from the total of 9716) are used for the test set. The relative log-likelihood results are presented in Table VI. The hybrid Pareto conditional mixture outperforms the other conditional mixture models. The difference in log-likelihood is very large on average between the hybrid Pareto conditional mixture and the double kernel estimator. The standard error is also very large because of the inability of the double kernel estimator to extrapolate beyond the data seen during training. The hyperparameters selected in validation are presented in Table VII. Overall, the complexity level chosen is rather high.

Figs. 17–19 depict the conditional densities for 20 random test points for the conditional mixture models with hybrid Pareto, Gaussian, and log-normal components, respectively. Multimodality is well captured. Depending on the donor profile, some amounts of donation are more probable than others. The hybrid Pareto conditional mixture and the log-normal conditional mixture produce heavy-tailed densities whereas the Gaussian conditional mixture produces light-tailed densities. However, the double kernel estimator (not shown here), despite detecting the multimodality of the data, does not predict any variation in the density given the donor profile. This estimator is, again, unable to extrapolate beyond the range of the data seen during training.

V. Conclusion

Fat-tailed data are prominent in several commercial applications of statistical machine learning, such as finance and insurance. Research on extreme events has been mainly concerned with unconditional density estimation of the tail of the distribution whereas in many such applications it is required to consider a conditioning variable, which can be very high dimensional. On the other hand, existing tools for representing conditional densities are not always appropriate in the presence of heavy-tail behavior, multimodal and asymmetric conditional densities. The main contributions of this paper are thus the following.

The hybrid Pareto is a new fat-tailed density which stitches together the generalized Pareto with the Gaussian distribution.

This distribution serves as a way to combine EVT methods and nonparametric density estimators such as the mixture of distributions. We have proposed a new conditional density estimator based on a mixture of hybrid Paretos whose parameters are learned functions of the conditioning variable. These functions can be parametric or nonparametric and we have worked with a simple neural network formulation, which is flexible enough for many applications.

The hybrid Pareto conditional mixture offers a way to model conditional distributions with arbitrary tail behavior, multimodality, or asymmetry. The proposed estimator inherits the tail approximation properties of the GPD while allowing for automatic and implicit conditional threshold selection. Through experiments on synthetic data, we have shown that the conditional mixture of hybrid Paretos outperforms competing algorithms based on conditional mixtures and nonparametric density estimation. Moreover, the proposed estimator gives a sensible model of the tail of the underlying distribution. Finally, through experiments on two real data sets, the hybrid Pareto conditional mixture has proven to provide significant advantages over the other conditional density estimators considered.

Current challenges in EVT concern multivariate extreme modeling [20]. The crucial problem is the estimation of the extremal dependence function. Here, we have focussed on predicting the conditional density of a scalar variable Y. By resorting to machine learning models, it might be possible to propose novel ways to also tackle these problems with multivariate Y.

REFERENCES

[1] P. Embrechts, C. Kluppelberg, and T. Mikosch, *Modelling Extremal Events*, ser. Applications of Mathematics, Stochastic Modelling and Applied Probability. New York: Springer-Verlag, 1997.

[2] J. Pickands, "Statistical inference using extreme order statistics," *Ann. Statist.*, vol. 3, pp. 119–131, 1975.

[3] A. J. McNeil, "Estimating the tails of loss severity distributions using extreme value theory," *Astin Bull.*, vol. 27, pp. 117–137, 1997.

[4] J. Carreau and Y. Bengio, "A hybrid Pareto model for asymmetric fat-tailed data: The univariate case," *Extremes*, vol. 12, pp. 53–76, 2009.

[5] H. White, "Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings," *Neural Netw.*, vol. 3, no. 5, pp. 535–549, 1990.

[6] C. E. Priebe, "Adaptive mixtures," *J. Amer. Statist. Assoc.*, vol. 89, pp. 796–806, 1994.

[7] K. M. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, no. 2, pp. 251–257, 1991.

[8] A. C. Davison and R. L. Smith, "Models for exceedances over high thresholds," *J. R. Statist. Soc. B*, vol. 52, no. 3, pp. 393–442, 1990.

[9] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert w function," *Adv. Comput. Math.*, vol. 5, pp. 329–359, 1996.

[10] V. Choulakian and M. A. Stephens, "Goodness-of-fit tests for the generalized Pareto distribution," *Technometrics*, vol. 43, no. 4, pp. 478–484, November 2001.

[11] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics, The Approach Based on Influence Functions*. New York: Wiley, 1986.

[12] D. M. Bashtannyk and R. J. Hyndman, "Bandwidth selection for kernel conditional density estimation," *Comput. Statist. Data Anal.*, vol. 36, no. 3, pp. 279–298, May 2001.

[13] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixture of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[14] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[15] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, "A universal approximator of convex functions applied to option pricing," in *Adv. Neural Inf. Process. Syst.*, 2001, vol. 13, pp. 1–8.

[16] M. Rosenblatt, "Conditional probability density and regression estimators," in *Multivariate Analysis II*, P. R. Krishnaiah, Ed. New York: Academic, 1969, pp. 25–31.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

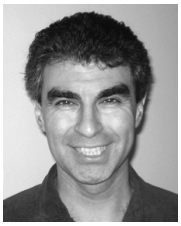
[18] A. J. McNeil and R. Frey, "Estimation of tail-related risk measures for heteroscedatic financial time series: An extreme value approach," *J. Empirical Finance*, vol. 7, pp. 271–300, 2000.

[19] J. Georges and A. H. Milley, "Kdd'99 competition: Knowledge discovery contest," *SIGKDD Explorations*, vol. 1, no. 2, pp. 79–84, Jan. 2000.

[20] T. Mikosch, "How to model multivariate extremes if one must?" The Danish National Research Foundation, Copenhagen, Denmark, Res. Rep. 21, 2004.

Julie Carreau received the Ph.D. degree from University of Montreal, Montreal, QC, Canada, in 2008.

Currently, she has a postdoctoral position at the Laboratoire des Sciences du Climat et de l'Environnement, Gif-sur-Yvette, France. Her research interests focus on the development of algorithms at the junction between machine learning and extreme value theory. Many stimulating problems arise in the field of climate and environment for such algorithms. She has been working on application in river runoff modeling, downscaling of precipitation, and bivariate extreme estimation.



Yoshua Bengio received the Ph.D. degree from McGill University, Canada, in 1991.

Currently, he is Professor at the Department of Computer Science and Operations Research, Université de Montréal, Montréal, QC, Canada. His main ambition is to understand how learning can give rise to intelligence. He has been an early proponent of deep architectures and distributed representations as tools to bypass the curse of dimensionality and learn complex tasks. He contributed to many machine learning areas: neural networks, recurrent neural networks, probabilistic graphical models (especially for temporal data), kernel machines, semisupervised learning, unsupervised learning and manifold learning, pattern recognition, data-mining, natural language processing, machine vision, and time-series prediction.

Dr. Bengio is Canada Research Chair in Statistical Learning Algorithms, as well as NSERC-CGI Chair, and Fellow of the Canadian Institute for Advanced Research.

A.2 Applications en sciences du climat et de l'environnement

A.2.1 Prévision probabiliste du débit

A statistical rainfall-runoff mixture model with heavy-tailed components

J. Carreau,[1] P. Naveau,[1] and E. Sauquet[2]

Received 18 February 2009; revised 13 July 2009; accepted 24 July 2009; published 29 October 2009.

[1] We present a conditional density model of river runoff given covariate information which includes precipitation at four surrounding stations. The proposed model is nonparametric in the central part of the distribution and relies on extreme value theory parametric assumptions for the upper tail of the distribution. From the trained conditional density model, we can compute quantiles of various levels. The median can serve to simulate river runoff, quantiles of level 5% and 95% can be used to form a 90% confidence interval, and, finally, extreme quantiles can estimate the probability of large runoff. The conditional density model is based on a mixture of hybrid Paretos. The hybrid Pareto is built by stitching a truncated Gaussian with a generalized Pareto distribution. The mixture is made conditional by considering its parameters as functions of covariates. A neural network is used to implement those functions. A penalty term on the tail indexes is added to the conditional log likelihood to guide the maximum likelihood estimator toward solutions that are preferred. This alleviates the difficulties encountered with the maximum likelihood estimator of the tail index on small training sets. We evaluate the proposed model on rainfall-runoff data from the Orgeval basin in France. The effect of the tail penalty is further illustrated on synthetic data.

Citation: Carreau, J., P. Naveau, and E. Sauquet (2009), A statistical rainfall-runoff mixture model with heavy-tailed components, *Water Resour. Res.*, *45*, W10437, doi:10.1029/2009WR007880.

1. Introduction

[2] River runoff modeling is relevant for hydroelectricity planning, irrigation and flood prevention. It is a well-known fact among hydrologists that the river runoff is fat tailed, meaning that sudden large values of runoff can occur which are three or four standard deviations away from the sample mean [*Bernadara et al.*, 2008]. Taking into account those large values is essential since they understandably have a very large impact. Another well-known fact is that precipitation in the hydrographic basin influences the river runoff. However, there are many other mechanisms at work such as underground water tables and soil permeability that are specific to a given hydrographic basin. Most hydrological models try to reproduce the dynamics of the basin by modeling the mechanisms in terms of reservoirs. An alternative approach is to use a stochastic model which provides a full distribution of the river runoff. For example, such a model has been proposed by *Lu and Berliner* [1999]. They assume three states or regimes of the runoff process: rising, falling and normal. Transitions probabilities between the states are modelled depending on past runoff values and on rainfall data. Given the current state, the distribution of the river runoff is assumed to follow an autoregressive process which depends on the past runoff values and the observed precipitation. We propose to model the distribution of the runoff at a future time step $t + 1$ given covariate information available at time t with another stochastic model, the conditional mixture of hybrid Paretos presented by *Carreau and Bengio* [2009a]. This model bears some similarities to the model of *Lu and Berliner* [1999]. In the conditional mixture, we can see the number of components as the number of states, which is determined by model selection instead of being set a priori. The state selection which is controlled by the mixture weights depends on all the covariates but not on the previous state. The distribution of the river runoff given the current state is given by the corresponding component density, that is a hybrid Pareto density. The parameters of this density are modeled as function of covariates which include past runoff and precipitation. The conditional mixture can adapt to a more general shape of the underlying distribution, including asymmetry and multimodality. Also, the hybrid Pareto enables the stochastic model to take explicitly extreme values into account. Moreover, a neural network computes, given the covariates, the mixture weights (or state probabilities) and the component density parameters. In contrast to *Lu and Berliner* [1999], we don't need to assume a specific form for the relationship between the covariates and the model parameters since such a neural network can in principle approximate any continuous mapping. The model will be further detailed in section 2.

[3] Neural networks have been popular models for a good while in hydrology (see *Maier and Dandy* [2000] for a survey). They were used to predict river runoff but, to our knowledge, not within a conditional mixture framework. Such traditional neural networks are generally not apt at

[1]Laboratoire des Sciences du Climat et de l'Environnement, UMR 1572, CEA, UVSQ, CNRS, Gif-sur-Yvette, France.
[2]Cemagref Lyon, Unité de Recherche Hydrologie-Hydraulique de Lyon, Lyon, France.

0043-1397/09/2009WR007880

capturing extreme observations. On the other hand, standard models to tackle extremes are drawn from extreme value theory (EVT) [*Embrechts et al.*, 1997]. These models consider either maxima over a given period, in which case the generalized extreme value (GEV) distribution is used, or observations that exceed a selected threshold and a generalized Pareto distribution (GPD) models the distribution of the exceedances. The EVT models thereby mean to estimate the upper tail of the underlying distribution. The choice of the GEV and the GPD is motivated by the fact that these are the limiting distributions of the maxima and the exceedances, respectively, under some fairly general conditions. Although extreme runoff behavior is utterly important, hydrologists need to model the whole runoff distribution. One way to extend the GPD model to the whole distribution has been proposed by *Frigessi et al.* [2002]. Their model is a two-component mixture with one light-tailed component and one GPD component. The hybrid Pareto mixture can be seen as a different way to include the GPD into a mixture model. The hybrid is built by stitching together a Gaussian and a GPD while ensuring continuity at the junction point. In the hybrid Pareto mixture, the number of components is chosen according to the data at hand. The central part of the hybrid Pareto mixture consists of a Gaussian mixture which is a flexible nonparametric estimator. The upper tail of the hybrid Pareto mixture is made of a linear combination of GPDs. Through experiments, this approach has shown to perform well on heavy-tailed data [*Carreau and Bengio*, 2009b].

[4] *Vrac and Naveau* [2007] have incorporated covariates in the Frigessi mixture [*Frigessi et al.*, 2002] in order to predict the distribution of rainfall. The covariates help discriminating between different sorts of rainfall regimes: no rainfall, regular rainfall and extreme rainfall. A particular distribution is used according to which regime prevails. Another way to include covariates into an EVT model has been developed by *Chavez-Demoulin and Davison* [2004]. Covariates are assumed to influence the value taken by the GPD parameters. This relationship is modeled by spline smoothers. In the conditional hybrid Pareto model, the mapping between the hybrid Pareto mixture and the covariates is modeled by a neural network. In this case, the whole conditional distribution is estimated, not just the conditional upper tail, as in the model of *Chavez-Demoulin and Davison* [2004].

[5] The tail index parameter is the most difficult parameter to estimate, whatever model is used, be it the GPD, the GEV distribution or some other method which one could think of for tail index estimation. This is because the tail index parameter, also termed the shape parameter, gives a sense of the overall shape of the distribution and in particular, of the tail behavior. Typically, few observations will occur in the tail which makes the estimation of the tail index very sensitive. Despite the good asymptotic properties of maximum likelihood estimators (MLEs), they are not very reliable in small samples given their high variance. Estimators of moments show a better behavior in small samples, however they assume that the expectation of the underlying distribution is finite (equivalently, that the tail index is smaller than one). *Coles and Dixon* [1999] introduced a penalty term in the MLEs of the GEV parameters. The intuition behind the penalty term is to include a similar range restriction on the tail index estimator as for the moment estimator. *Coles and Dixon* [1999] show that the penalized MLE of the tail index performs better in small samples than the classical MLE.

[6] The hybrid Pareto is one such model with a tail index parameter, which is inherited from the GPD. When density estimation is performed with a hybrid Pareto mixture, the tail index of the underlying distribution can be estimated from the tail index of the dominant component in the mixture, that is the component with the largest tail index (and consequently, the heaviest tail). In this case, the MLEs sensitivity in small samples appears in the following way: large tail indexes are assigned to components with negligible mixture weights. To prevent this, we add a penalty term to the log likelihood based on a prior distribution of the mixture tail indexes. This is similar in spirits to the penalty proposed by *Coles and Dixon* [1999]. We devised a prior distribution of the mixture tail indexes based on the following intuitive idea. We would expect that most components would take care of modeling the central part of the distribution and therefore, have a tail index close to zero. If the tail of the underlying distribution is heavy, we would then expect that some components would have a tail index close to the tail index of the underlying distribution.

[7] We evaluate the conditional hybrid Pareto mixture on rainfall-runoff data from the Orgeval basin in France. The conditional median of the learned conditional hybrid Pareto mixture serves to generate river runoff at a future time step $t + 1$. A 90% confidence interval is also computed as the quantiles of level 5% and 95%. This is in contrast with the work of *Frigessi et al.* [2002] and *Vrac and Naveau* [2007] who did not use their model for prediction at a future time step. We also look at the distribution of the conditional tail indexes on the test set; the effect of the tail penalty term in the maximum likelihood estimator can be seen. We gain then more insight into the effect of the new penalty by looking at experiments on synthetic data.

2. Statistical Model of the Rainfall-Runoff Process

[8] We propose to model the rainfall-runoff process with the conditional hybrid Pareto mixture [see *Carreau and Bengio*, 2009a]. This model combines the flexibility of nonparametric modeling and the extrapolation capability of the GPD methodology. Given a vector of covariates which describe meteorological and hydrological conditions, the conditional distribution of the river runoff is modeled by a mixture of hybrid Paretos whose parameters depend on covariates. Such a mixture is able to adapt to asymmetry, multimodality and tail heaviness that might be present in the conditional distribution of the runoff. The neural network which learns the relationship between the covariates and the mixture parameters is able to approximate properly the highly nonlinear relationship between rainfall and runoff. The conditional hybrid Pareto mixture provides a conditional density model that has proven to perform well on many kind of data sets [see *Carreau and Bengio*, 2009a]. The model is explained in detail in sections 2.1–2.3.

2.1. Hybrid Pareto Mixture

[9] Suppose we want to model the distribution of Y, a variable representing the river runoff, with no additional

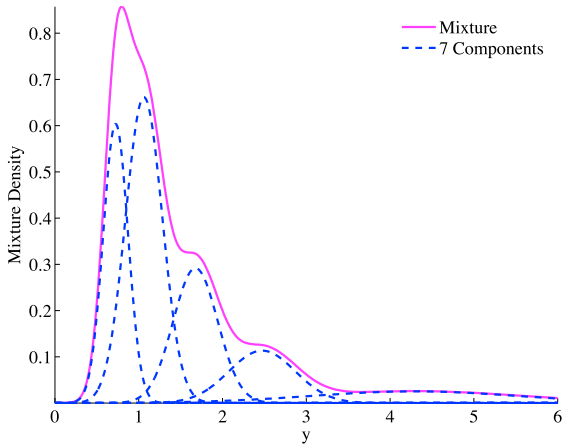


Figure 1. Gaussian mixture density (solid line) with seven components trained on heavy-tailed data. The dashed lines represent the contribution of each component to the density. Five components model the central part, and the other two components contribute to the density in the upper tail.

predictive information. We could estimate the distribution of Y with a mixture of Gaussians, which is a popular nonparametric estimator [*Bishop*, 1995]. This type of approach circumvents the need to choose a specific parametric form for the distribution of the runoff and can take into account multimodality and asymmetry. Mixtures of Gaussians approximate a density by adding up weighted Gaussians or "bumps" (see Figure 1). The density estimator is formally given by $\sum_{j=1}^{m} \pi_j \phi_{\mu_j,\sigma_j}(y)$, where the π_j are the mixture weights and $\phi_{\mu_j,\sigma_j}(\cdot)$ is the Gaussian density with parameters μ_j and σ_j. The weights must sum to one, that is, $\sum_{j=1}^{m} \pi_j = 1$, to ensure that the estimator is a proper density.

[10] A Gaussian mixture approximates the distribution of heavy-tailed data, such as runoff data, by locating one component with a large standard deviation around the largest observations. However, its capacity to extrapolate beyond the sample range might be poor.

[11] The hybrid Pareto distribution was put forward as a way to transfer the extrapolation properties of the GPD [*Embrechts et al.*, 1997] to mixture models. The hybrid Pareto distribution is a smooth extension of the GPD to the whole real axis. This new distribution is built by stitching a GPD tail to a Gaussian, while enforcing continuity of the resulting density and of its derivative. In this work, we focus on runoff data which is heavy tailed so we let $\xi > 0$ in the GPD density where the scale parameter β is positive and the location parameter α is real:

$$g_{\xi;\beta}(y-\alpha) = \frac{1}{\beta}\left(1+\frac{\xi}{\beta}(y-\alpha)\right)^{-1/\xi-1} \qquad \xi > 0, \quad y > \alpha.$$

Let α be the junction point and $\phi_{\mu;\sigma}(y) = 1/(\sqrt{2\pi}\sigma)\exp(-(y-\mu)^2/(2\sigma^2))$ be the Gaussian density function with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. The two constraint equations (equality of the density and of its derivative at α) are solved so that α and β, the GPD scale parameter, become functions of ξ, the GPD tail index and of μ and σ, the Gaussian parameters. Let $\theta = (\xi, \mu, \sigma)$ be the parameter vector of the hybrid Pareto. The hybrid Pareto density is given by

$$h_\theta(y) = \begin{cases} \frac{1}{\gamma}\phi_{\mu;\sigma}(y) & \text{if } y \le \alpha, \\ \frac{1}{\gamma} g_{\xi;\beta}(y-\alpha) & \text{if } y > \alpha, \end{cases}$$

where the dependent parameters are $\alpha(\xi, \mu, \sigma) = \mu + \sigma\sqrt{W\left((1+\xi)^2/2\pi\right)}$, $\beta(\xi, \sigma) = (\sigma(1+\xi))/\left(\sqrt{W\left((1+\xi)^2/2\pi\right)}\right)$ and W is the Lambert W function defined by $w = W(we^w)$ [see *Corless et al.*, 1996]. The reweighting factor γ ensures that the density integrates to one and is given by

$$\gamma(\xi) = 1 + \frac{1}{2}\left(1 + Erf\left(\sqrt{W\left((1+\xi)^2/2\pi\right)/2}\right)\right),$$

where $Erf(\cdot)$ is the error function $Erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t2}\,dt = 2\,\Phi(z\sqrt{2}) - 1$ and Φ is the standard Gaussian distribution function [see *Press et al.*, 1992]. The hybrid Pareto, while inheriting the approximation properties of the GPD, bypasses the need for threshold selection inherent in the classical GPD methodology [*Embrechts et al.*, 1997] since α, the junction point of the Gaussian and the GPD is computed implicitly as a function of the hybrid parameters.

[12] With a hybrid Pareto mixture $\sum_{j=1}^{m} \pi_j h_{\theta_j}(y)$ to model the distribution of the river runoff, we get the best of both worlds: the central part is a mixture of Gaussians which benefits from flexible approximation properties and the upper tail is a linear combination of GPD densities that are capable of extrapolating in areas of unseen data under sound parametric assumptions.

2.2. Conditional Density Model

[13] Our goal is to provide a model of the river runoff at a future time step. We have at our disposal rainfall data in the hydrographic basin of interest which influences river runoff. We therefore look into modeling the distribution of the runoff at time $t + 1$ given covariate information at time t, which includes rainfall observations and past runoff. The hybrid Pareto mixture can be turned into a conditional density model by thinking of the parameters of the mixture as function of covariates [*Bishop*, 1995]. These functions can be implemented in many ways. The simplest model would be a linear model. However, the relationship between rainfall and runoff is highly nonlinear. A one-layer feed forward neural network of which the linear model is a special case (no hidden units) is able, if the number of hidden units is well chosen, to approximate any continuous relationship between covariates and mixture parameters. Data-driven selection of the number of hidden units provides a proper level of complexity (or nonlinearity). A representation of the conditional mixture model with a neural network is given in Figure 2. The covariates, or inputs, are combined linearly and either fed to the hidden units or directly connected to the neural network outputs. We took the hyperbolic tangent as the activation function of the hidden layer. The neural network outputs are then

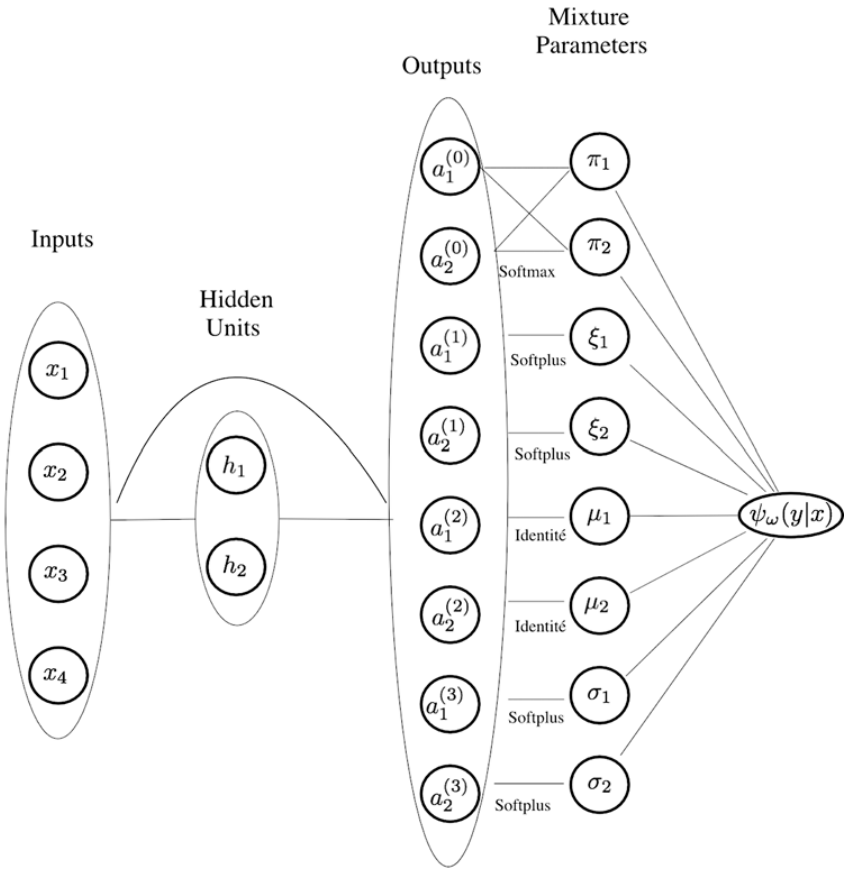


Figure 2. Representation of a conditional mixture model with hybrid Pareto components $\psi_\omega(y|x)$. Inputs are fed to a one-layer feed forward neural network with an extra linear connection directly to the outputs. The outputs are then transformed into the mixture parameters so as to fulfill range constraints.

transformed into the mixture parameters. Different transformation functions constrain the range of each mixture parameter. The $a_j^{(0)}$ in Figure 2 are dedicated to the mixture weights. The transformation function, the softmax, ensures that these weights are positive and sum to one: $\pi_j = \exp(a_j^{(0)})/\sum_k \exp(a_k^{(0)})$. The $a_j^{(1)}$ and $a_j^{(3)}$ control the tail index and the spread parameter, respectively, of the jth component. They are guaranteed to be positive by using a softplus [*Dugas et al.*, 2001], a slow-growing version of the exponential: $y = \text{softplus}(x) = \log(1 + \exp x)$. Finally, the $a_j^{(2)}$ are assigned to the location parameters and need no range constraint.

[14] There are two hyperparameters to adjust the level of complexity in the conditional hybrid Pareto mixture: the number of hidden units in the neural network and the number of components in the mixture. The former controls the degree of nonlinearity of the mapping between the covariates and the mixture parameters and the latter accounts for the complexity of the conditional density (in particular, the multimodality and asymmetry). Given the approximation capabilities of the neural network and of the mixture model, if the complexity level is well chosen, the conditional mixture should be able to approximate any type of conditional density. The hyperparameters are chosen so as to maximize the conditional log likelihood on a validation set, distinct from the training set and thus, should be reasonably close to the ones that give the best generalization performance (the capacity to perform well on unseen data). Because there are many sources of variability (training data, optimization process), the hyperparameter selection can be variable as well. Overall, the conditional hybrid Pareto mixture gave a better performance than other conditional density estimator in the presence of heavy-tailed data [*Carreau and Bengio*, 2009a].

2.3. Learning and Regularization

[15] The conditional mixture parameters are the neural network parameters ω. These are learned by minimizing the negative conditional log likelihood on the training data:

$$\mathcal{L}(\omega) = -\sum_{i=1}^{n} \log(\psi_\omega(y_i|x_i)),$$

where the sum is over the training set $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and $\psi_\omega(y_i|x_i)$ is the hybrid Pareto conditional mixture model evaluated at the data point i.

[16] *Carreau and Bengio* [2009a] observed empirically that maximum likelihood estimation of the hybrid Pareto

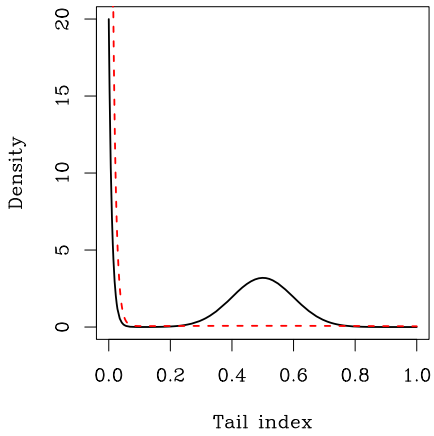


Figure 3. The distribution represented by the solid line has one mode at zero and one mode at 0.5, while the distribution represented by the dashed line has significant density only around zero. The former distribution reflects our prior information about how the tail indexes of a hybrid Pareto mixture should be distributed when the data are heavy tailed, and the latter distribution, reflects the situation when the data are light tailed.

mixture, conditional or not, can lead to overestimation of the tail indexes. This is especially striking for small training sets. The overestimation of the tail index, even by a small amount, leads to gross overestimation of the extreme quantiles. In order to guide maximum likelihood estimation and avoid the overestimation of the tail indexes, we use a penalty term based on the prior density of equation (1):

$$f(x;\tau,\eta,\rho) = \tau\eta\exp\{-\eta x\} + (1-\tau)\frac{\exp\left\{-(x-0.5)^2/(2\rho^2)\right\}}{\sqrt{2\pi}\rho}. \quad (1)$$

[17] Figure 3 illustrates two typical shapes of the prior density. In the case of runoff data, we can safely assume that the distribution has a tail index around 0.5 [*Bernadara et al.*, 2008]. This implies that a variant of the solid line density in Figure 3 will hold. Most components will be light tailed, with tail indexes close to zero. These components will take care of modeling the central part of the distribution. Some components will be heavy tailed, with a tail index value close to the one of the underlying density and these will estimate the upper tail of the distribution. Hence, the solid line density is bimodal, with one mode at zero and the other one, smaller, around 0.5. On the other hand, if the data are light tailed, then we assume that all the components will have tail indexes close to zero. The prior density in this case would look like the dashed line density in Figure 3.

[18] The two-component mixture of equation (1) can generate densities such as those illustrated in Figure 3. The exponential component with parameter η controls the density assigned to the small tail indexes and the Gaussian component centered at 0.5 with standard deviation ρ determines how wide the range of the larger tail indexes can be. The mixture weight τ establishes the trade-off between the two components. When τ is equal to zero, we are in the light-tail case.

[19] The conditional mixture parameters ω are now learned by minimizing a new cost function, the negative conditional log likelihood minus the penalty term:

$$\mathcal{L}(\omega) = -\sum_{i=1}^{n}\log(\psi_\omega(y_i|x_i)) - \frac{\lambda}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\log f\left(\xi_{i,j};\tau,\eta,\rho\right),$$

where the first sum is over the training set $\mathcal{D}_n$, the second sum in the penalty term is over the number of components m, $\psi_\omega(y_i|x_i)$ is the hybrid Pareto conditional mixture model evaluated at point i and $f(\xi_{i,j};\ \tau,\ \eta,\ \rho)$ is the prior density evaluated at the tail index of the jth component of the conditional mixture at point i. The penalty term introduces four other hyperparameters: λ which controls the weight of the penalty with respect to the conditional log likelihood and τ, η and ρ from the prior density (see equation (1)). A restricted set of values for the prior density parameters was selected so as to ensure that the prior density follows our prior information about the shape of the distributions of the tail indexes. The model is trained for several combinations of hyperparameters (which include the number of hidden units and the number of components of the conditional hybrid Pareto mixture and the hyperparameters attached to the penalty term). The set of hyperparameters which gives the smallest cost in terms of negative conditional log likelihood on data unseen during training (the validation set) is selected.

3. Experiments

[20] We evaluate the conditional hybrid Pareto mixture on the rainfall-runoff data from the Orgeval basin in France. Synthetic data experiments help to gain more insight into the role of the new penalty term in the cost function. Since the generative model is known, the predicted tail indexes can be compared with the tail indexes of the generative model. We also compare the conditional quantiles of the generative versus learned model.

3.1. Orgeval Basin Data

[21] The Orgeval Basin is located in France, east of Paris. There is no snow accumulation in the area that could affect the river runoff. Therefore, we focus on rainfall as a predictor of the river runoff. In order to capture the mechanisms of the basin, moving averages and moving standard deviations of various window lengths of the river runoff are included in the covariates. The river runoff Q_t from the Avenelles subbasin and the precipitations at four surrounding stations, P_t^j, j = 1, ..., 4, are available at a hourly time step for over 30 years but we use approximately 10 years of data, from 1986 to 1996 (see http://www.antony.cemagref.fr for more details on the data and the basin). We also have daily average temperatures at this site for the same time period. Date variables serve to capture the cycles and trends in the data. Precisely, there are 16 covariates to predict the river runoff distribution: rainfall from the four precipitation stations at the previous time step, the runoff at the two previous time steps, moving averages and standard deviations with daily, weekly and monthly window widths,

Table 1. Three Periods With No Missing Value in the Orgeval Basin Data in Order of Decreasing Lengths

Data Set	Time Period	Hourly Observations
1	26 Mar 1986 1800:00 to 22 May 1994 0800:00	71,487
2	22 Jul 1996 1500:00 to 24 Aug 2001 1600:00	44,618
3	30 May 1994 1800:00 to 18 Jun 1996 0300:00	17,987

three date variables concerning the year, the month and the week and the daily average temperature at the previous day. Three time periods where there is no missing data are split into training and test sets. The data sets are summarized in Table 1. For this experiment, we set $Y_t = Q_{t+1}$ and $X_t = [Q_t, Q_{t-1}, P_t^1, \ldots,]$ which means that given information available at time t, we model the distribution of the runoff at time $t + 1$. With the hourly data, we thus model the conditional distribution of the runoff at the next hour. In order to increase the prediction horizon to 6 and 12 h, the hourly data are aggregated to form 6 h and 12 h time steps. To this end, we take the average of the runoff and the sum of the rainfall over the appropriate time period. This means that the lengths of our initial data sets in Table 1 are divided by the length of the time steps. We thus have three different models, one for each time step.

[22] We assume that given the covariate vector X_t, the Y_t are independent and identically distributed. It is thus possible to perform model selection via fivefold cross validation (as opposed to sequential cross validation which is more computationally intensive; see *Bishop* [1995] for details). Model selection works as follows. The training set is divided into five subsets or folds. The conditional hybrid Pareto mixture is first trained on four of those folds by minimizing the penalized negative conditional log likelihood for each set of hyperparameters considered and the performance in terms of conditional log likelihood of each trained model is evaluated on the left out fold. This process is repeated five times, so that each fold in turn was left out and that the model performance was evaluated on all the data of the training set. The hyperparameters that gave the best performance in validation are selected. The model with the selected hyperparameters are trained again this time on the whole training set. The generalization ability, that is how well the model does on unseen data, is then evaluated on the test set, which is distinct from the training set. Results from the experiments on the Orgeval basin data are summarized in Table 2 for each time step (1 h, 6 h, 12 h). The selected hyperparameters for the penalty term, (λ, τ, η, σ), correspond to the prior belief that the distribution is heavy tailed. The confidence interval is computed from the conditional quantiles of level 0.05 and 0.95; therefore, the observed runoff should fall into that interval nine times out of ten. The percentage given on the confidence interval row is the actual percentage of observed runoff on the test set which fall into the confidence interval. We can see that it is pretty close to the expected one. A measure of goodness of fit is the so-called R square given as $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \overline{y})^2$, where y_i is the observed runoff, $\hat{y}_i$ is the prediction and $\overline{y}$ is the sample average. The closer R^2 is to one, the better the prediction is. The R square is computed on the test set and the conditional median of the trained model is used to predict the runoff. We can see from the last row of Table 2 that the R square for all time steps are very good, although the accuracy of the prediction decreases with the length of the time step. Prediction at longer time steps are understandably more difficult. A different test set is used for the 12 h time step data (data set number 2 in Table 1) in order to leave more data for the training set. The prediction is possibly more challenging on that time period and at least, not directly comparable with the other two models, 1 h and 6 h, which uses a similar test set.

[23] The river runoff for the test period is illustrated in the left plots of Figure 4. The top, middle, and bottom plots each correspond to one time step. The model prediction, which is the conditional median of the trained model, is plotted for each test set in the right plots of Figure 4. For all time steps, we can see that the model captured very well the dynamics of the river runoff. In the left plots of Figure 5, we have plotted the confidence intervals in light grey with quantiles of level 0.05 and 0.95 for the first 100 points of the test set. The black line is the observed runoff. Sometimes, the confidence interval is very narrow while it grows larger where the model perceives more uncertainty. We can check the effect of the tail penalty by looking at the distribution of the tail indexes of the conditional hybrid Pareto mixture on the test set. This is illustrated the histograms in Figure 5. Except for a few cases in which the tail index exceeds one (which is allowed by the prior), the largest tail index values vary between 0.2 and 0.6 while most tail indexes take on values near zero. The distribution of the tail indexes is thus consistent with our prior belief.

3.2. Synthetic Data

[24] We generate synthetic data which resemble the runoff data in the sense that there are cycles and that the tail indexes are in the same range. Let Y be a random variable distributed according to a Fréchet distribution whose parameters are functions of an input variable X. Then the distribution function of $Y|X = x$ is given by

$$P(Y \le y|X = x) = \begin{cases} 0 & \text{si } y \le \mu(x), \\ \exp\left\{ -\left(\dfrac{y - \mu(x)}{\sigma(x)} \right)^{-1/\xi(x)} \right\} & \text{si } y > \mu(x). \end{cases}$$

Table 2. Experiments for the Orgeval Basin Data for Each Time Step[a]

	Hourly	6 h	12 h
Training data	52 846 (1)	9 913 (1)	7 455 (1,3)
Test data	10,000 (1)	2000 (1)	3 717 (2)
h, m	(4,4)	(4,8)	(4,12)
λ, τ, η, ρ	(0.01,0.5,50,0.1)	(0.1,0.1,50,0.2)	(1,0.1,50,0.1)
Confidence interval (%)	91.94	92.1	87.6
R^2	0.99	0.92	0.73

[a]Shown are the sizes of the training and test sets (with data set number from Table 1 in parentheses), the selected number of hidden units and components (h, m) followed by the selected penalty hyperparameters (λ, τ, η, σ), the percentage of the runoff in the test set which falls in the predicted 90% confidence interval, and the R^2 of the predicted median on the test set.

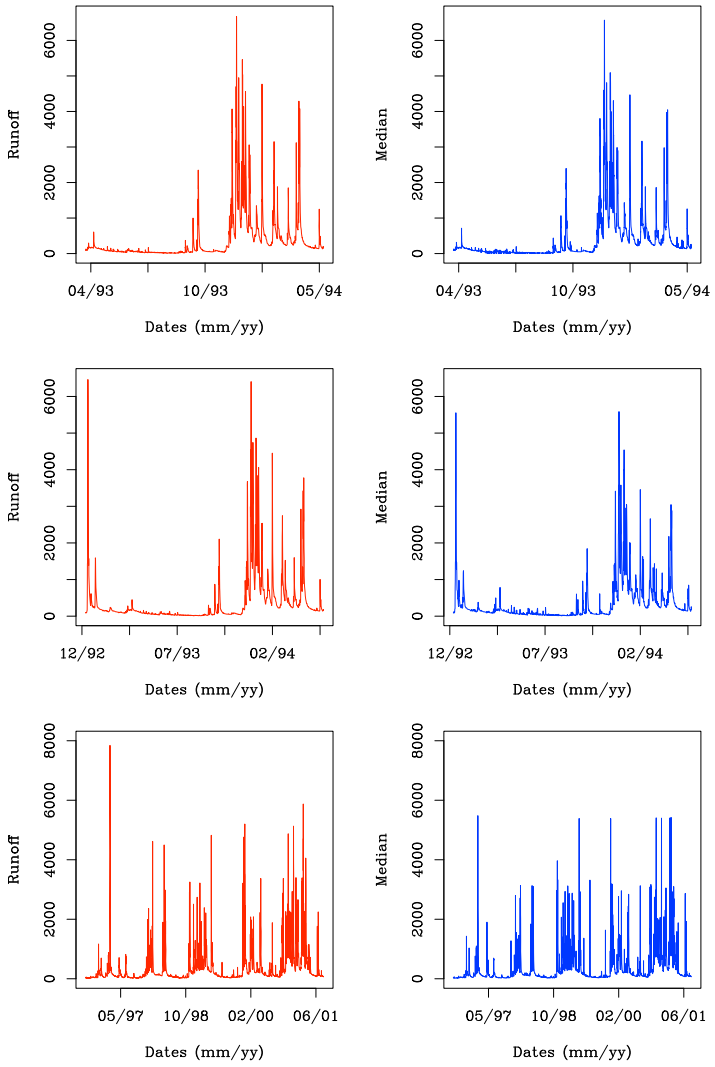


Figure 4. (left) Observed runoff of the Avenelles subbasin for the test period corresponding to a given time step: (top) 1 h, (middle) 6 h, and (bottom) 12 h. (right) Predicted median on the test set from the learned hybrid Pareto conditional mixture for the three time steps.

The Fréchet distribution is a canonical heavy-tail distribution: the tail of most heavy-tailed distribution eventually behaves like the Fréchet tail. The input variable X is distributed according to a standard Normal distribution. We chose the following sine-shaped functional form for the dependence function $\xi(\cdot)$:

$$\xi(x) = \beta_1 + \beta_2 \sin(\gamma_1 + \gamma_2 x).$$

Since $X \sim \mathcal{N}(0, 1)$, we select the parameters of $\xi(\cdot)$ so that $\xi(X) \in [0.25, 0.5]$ with probability 0.99. The dependence function $\mu(\cdot)$ and $\sigma(\cdot)$ have a similar sine-shaped form but their parameters are chosen so that $\mu(X) \in [2, 6]$ and $\sigma(X) \in [0.5, 1]$ with probability 0.99. We generated pairs of observations (X_i, Y_i) according to this generative model. Figure 6 (left) illustrates the training set which is made of 2000 such pairs of observations. Figure 6 (right) shows the corresponding tail indexes. Model selection (the choice of the proper set of hyperparameters) is performed via fivefold cross validation on the training set. Results are presented on a test set, distinct from the training set, which consists of 10,000 pairs of observations generated according to the conditional Fréchet distribution described above.

[25] The model selected via fivefold cross validation for the training set of Figure 6 has eight hidden units and two mixture components. The hyperparameters for the tail penalty are the following: $\lambda = 0.1$, $\tau = 0.45$, $\eta = 50$ and $\sigma = 0.05$. This corresponds to the shape of a prior density for heavy tails in Figure 3. The effect of the tail penalty can be seen in Figure 7 (left): the histogram of the conditional

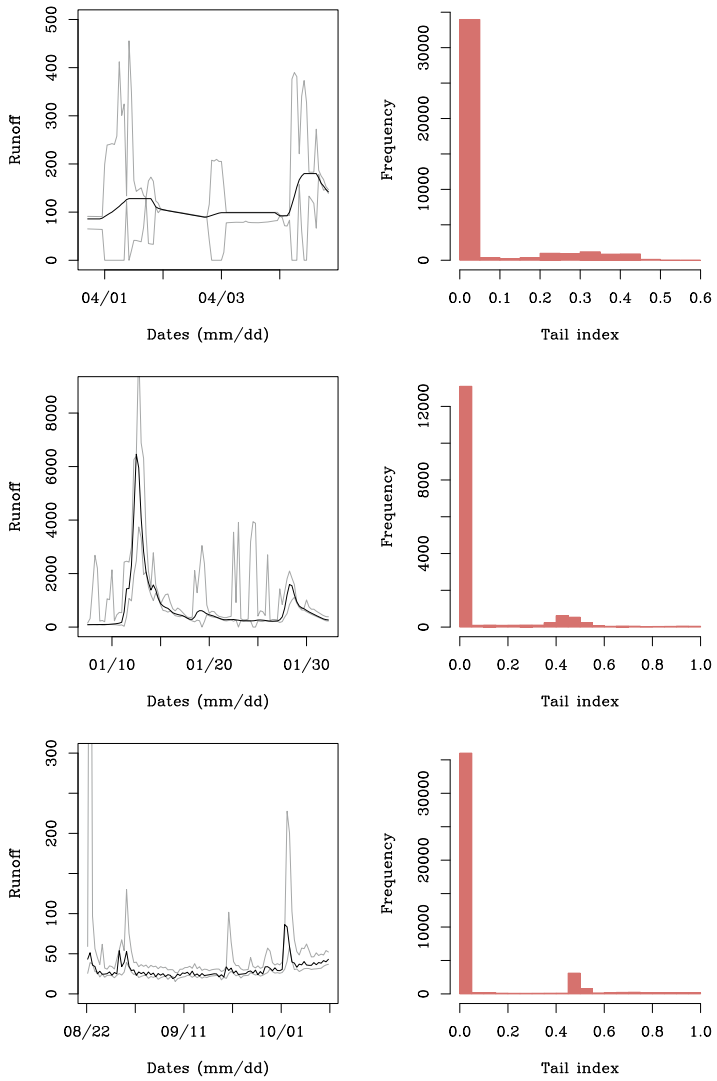


Figure 5. (left) The observed runoff for the first 100 points of the test set illustrated in Figure 4 (black) together with a 90% confidence interval (light grey) predicted from the conditional mixture. (right) Histogram of the tail indexes of the conditional hybrid Pareto mixture on the test set.

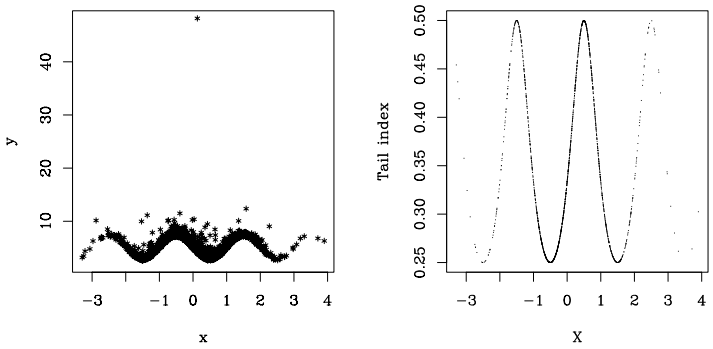


Figure 6. (left) Training set of 2000 data points distributed according to the conditional Fréchet distribution with a sine-shaped functional for the dependent parameters. (right) The corresponding conditional tail indexes of the generative conditional Fréchet model.

tail indexes of the conditional hybrid Pareto mixture on the test set reflects the shape of the prior density. Note that less than 1% of the tail indexes are larger than 1 and are thus not shown in Figure 7, this is due to the upper tail of the prior which still has some significative density in that area. For the generative model, the conditional tail indexes $\xi(X)$ vary between 0.25 and 0.5 (see Figure 6, right). According to our prior belief, there should be a small subset of tail indexes from the conditional hybrid Pareto mixture which take care of modeling the upper tail and thus should take values in the same interval [0.25, 0.5]. The histogram of Figure 7 is consistent with this prior belief. In Figure 7 (right) we have plotted the test set together with the quantiles of level 0.05% and 0.95% which form a 90% confidence interval as predicted from the trained conditional hybrid Pareto mixture. Among the test set, 89% of the data points fall into the confidence interval.

[26] In order to check how well the conditional density is learned in the upper tail, we compare three conditional quantiles of levels 0.9, 0.95 and 0.99 as computed from the generative model and the learned model. These are plotted in Figure 8: the black line is the quantile as computed from the trained conditional hybrid Pareto mixture and the light grey line is the quantile from the generative model. For the levels 0.9 and 0.95 (Figure 8, top), the two lines are almost indistinguishable from one another except for the lower and upper ends. The data density is much lower in these areas (see Figure 6) because the X variable follows a standard Normal distribution and this makes learning more difficult. The conditional quantile of level 0.99 is less well approximated. This is also due to data scarcity and shows that the model is less reliable in that case. Table 3 compares the percentage of the data in the test set which fall below the conditional quantiles of the generative model and the trained model for the three quantile levels. The picture is pretty similar for both models. Overall, the performance of the conditional hybrid Pareto mixture with the new tail penalty proves to be satisfying.

4. Conclusion

[27] We have propose a new stochastic model based on the conditional hybrid Pareto mixture [*Carreau and Bengio*, 2009a], in order to model the distribution of the river runoff at a future time step given rainfall observations in the hydrographic basin. This model relies on nonparametric algorithms, namely a feed forward neural network and a

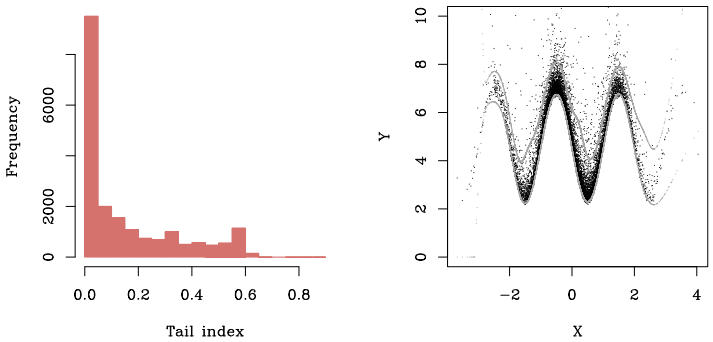


Figure 7. (left) Histogram of the conditional tail indexes of the trained conditional hybrid Pareto mixture on the test set. (right) The 90% confidence interval from the trained model on the test set together with the data points (89% of the data fall into the confidence interval).

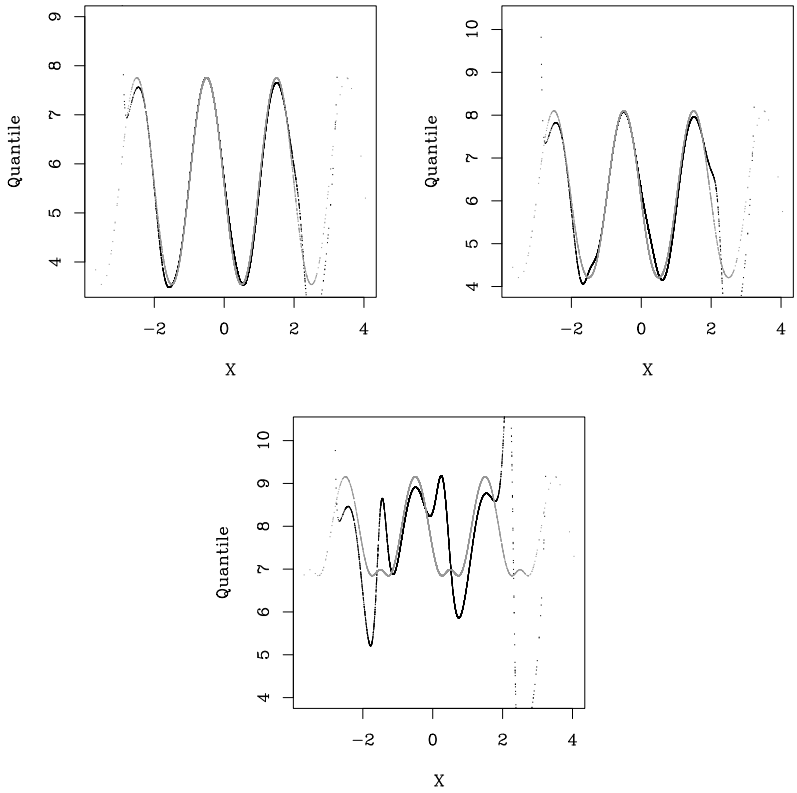


Figure 8. Conditional quantiles of (clockwise) level 90%, 95%, and 99% as computed from the mixture model (black) and from the generative conditional Fréchet model (light grey).

mixture of distributions, from which it gains flexibility. Moreover, the component of the mixture, the hybrid Pareto, inherits the tail approximation properties of the generalized Pareto distribution which are thus transmitted to the conditional hybrid Pareto mixture. Therefore, the conditional hybrid Pareto mixture has good approximation properties, as much in the central part of the distribution as in the upper tail area.

[28] We have introduced a penalty term in the maximum likelihood estimator in order to yield more realistic conditional tail index estimation. The penalty is based on a bimodal density which captures our prior knowledge of the distribution of the tail index. A hybrid Pareto mixture has as many tail indexes as there are components in the mixture. In the conditional case, the number of tail indexes is further multiplied by the number of data points. Our intuition is that the distribution of the tail indexes should have two modes, one around zero and one around the value of the tail index of the underlying distribution, if the latter is heavy tailed. Most components would be light tailed and take care of modeling the central part of the distribution whereas few components would have a heavier tail, near the value of the tail index of the generative model, and would thus approximate the upper tail of the underlying distribution.

[29] The conditional hybrid Pareto mixture has been trained on data from the Orgeval basin in France. Rainfall at four surrounding stations and the river runoff are available at hourly time step. These data were aggregated to obtain 6 h and 12 h time steps. The stochastic model was trained on three data sets, the hourly, six and 12 h time steps. Each model can then be used to forecast the river runoff at the next hour, 6 or 12 h later. Our experiments have shown that the conditional hybrid Pareto mixture is able to capture the dynamics of the basin for the three predictive time horizons. In addition, the model provides reliable confidence intervals. The tail index penalty introduces the expected distribution of the conditional tail indexes, with one mode at zero and the second mode around 0.5, more or less sharp depending on the data set.

[30] Finally, the conditional hybrid Pareto mixture was trained on synthetic conditional data based on the Fréchet distribution. The distribution of the tail indexes is consistent

Table 3. Experiments With the Conditional Fréchet Data[a]

	0.9 Quantile	0.95 Quantile	0.99 Quantile
Generative model	89.64	94.54	98.97
Trained model	89.16	94.1	98.39

[a]Shown are percentage of the data in the test set which fall below the conditional quantiles of levels 0.9, 0.95 and 0.99 for the generative and the trained models.

with the values of the conditional tail indexes of the generative model. On the test set, 89% of the data points falls into the 90% confidence interval predicted by the model. Moreover, the trained model compares favorably with the generative model in terms of extreme quantiles.

[31] The conditional hybrid Pareto mixture with the new penalty term has proven to be effective at modeling the rainfall-runoff process for various time steps on the Orgeval basin and more insight into the model was gain by looking at an experiment on synthetic data. This model is very flexible and could be useful to model the rainfall-runoff process in other hydrographic basins, by using appropriate covariates.

[32] **Acknowledgments.** The authors thank the following funding organizations: FQRNT, CNRS, and CEA and the AssimileX and ACQWA projects.

References

Bernadara, P., D. Schertzer, E. Sauquet, I. Tchiguirinskaia, and M. Lang (2008), The flood probability distribution tail: How heavy is it?, *Stochastic Environ. Res. Risk Assess.*, *22*, 107–122.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Clarendon, Oxford, U. K.

Carreau, J., and Y. Bengio (2009a), A hybrid Pareto mixture for conditional asymmetric fat-tailed distributions, *IEEE Trans. Neural Networks*, *20*, 1087–1101.

Carreau, J., and Y. Bengio (2009b), A hybrid Pareto model for asymmetric fat-tailed data: The univariate case, *Extremes*, *12*, 53–76.

Chavez-Demoulin, V., and A. C. Davison (2004), Generalized additive modelling of sample extremes, *Appl. Stat.*, *54*, 207–222.

Coles, S. G., and M. J. Dixon (1999), Likelihood-based inference for extreme value models, *Extremes*, *2*, 5–23.

Corless, R. M., G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth (1996), On the Lambert W function, *Adv. Comput. Math.*, *5*, 329–359.

Dugas, C., Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia (2001), A universal approximator of convex functions applied to option pricing, in *Advances in Neural Information Processing Systems*, vol. 13, edited by M. I. Jordan, Y. LeCun, and S. A. Solla, pp. 1–8, MIT Press, Cambridge, Mass.

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997), *Modelling Extremal Events*, *Appl. Math.*, vol. 33, Springer, New York.

Frigessi, A., O. Haug, and H. Rue (2002), A dynamic mixture model for unsupervised tail estimation without threshold selection, *Extremes*, *5*, 219–235.

Lu, Z.-Q., and L. M. Berliner (1999), Markov switching time series models with application to a daily runoff series, *Water Resour. Res.*, *35*(2), 523–534.

Maier, H. R., and G. C. Dandy (2000), Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications, *Environ. Modell. Software*, *15*, 101–124.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1992), *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd ed., Cambridge Univ. Press, Cambridge, U. K.

Vrac, M., and P. Naveau (2007), Stochastic downscaling of precipitation: From dry events to heavy rainfalls, *Water Resour. Res.*, *43*, W07402, doi:10.1029/2006WR005308.

J. Carreau and P. Naveau, Laboratoire des Sciences du Climat et de l'Environnement, UMR 1572, CEA, UVSQ, CNRS, CE Saclay l'Orme des Merisiers, bat. 701, F-91191 Gif-sur-Yvette, France.

E. Sauquet, Cemagref Lyon, Unité de Recherche Hydrologie-Hydraulique de Lyon, 3 bis quai chauveau, F-69336 Lyon CEDEX, France.

A.2.2 Descente d'échelle statistique

Stochastic downscaling of precipitation with neural network conditional mixture models

Julie Carreau[1] and Mathieu Vrac[2]

Received 15 October 2010; revised 13 July 2011; accepted 4 August 2011; published 4 October 2011.

[1] We present a new class of stochastic downscaling models, the conditional mixture models (CMMs), which builds on neural network models. CMMs are mixture models whose parameters are functions of predictor variables. These functions are implemented with a one-layer feed-forward neural network. By combining the approximation capabilities of mixtures and neural networks, CMMs can, in principle, represent arbitrary conditional distributions. We evaluate the CMMs at downscaling precipitation data at three stations in the French Mediterranean region. A discrete (Dirac) component is included in the mixture to handle the "no-rain" events. Positive rainfall is modeled with a mixture of continuous densities, which can be either Gaussian, log-normal, or hybrid Pareto (an extension of the generalized Pareto). CMMs are stochastic weather generators in the sense that they provide a model for the conditional density of local variables given large-scale information. In this study, we did not look for the most appropriate set of predictors, and we settled for a decent set as the basis to compare the downscaling models. The set of predictors includes the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalyses sea level pressure fields on a 6×6 grid cell region surrounding the stations plus three date variables. We compare the three distribution families of CMMs with a simpler benchmark model, which is more common in the downscaling community. The difference between the benchmark model and CMMs is that positive rainfall is modeled with a single Gamma distribution. The results show that CMM with hybrid Pareto components outperforms both the CMM with Gaussian components and the benchmark model in terms of log-likelihood. However, there is no significant difference with the log-normal CMM. In general, the additional flexibility of mixture models, as opposed to using a single distribution, allows us to better represent the distribution of rainfall, both in the central part and in the upper tail.

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

1. Introduction

[2] General circulation models (GCMs) solve the principal physics equations of the dynamics of the atmosphere and of the oceans together with their interactions on a 3-D grid over the globe. GCMs allow us to simulate climate variables and to study the mechanisms of the present, past, and future climate of the Earth [e.g., *Gladstone et al.*, 2005; *Intergovernmental Panel on Climate Change* (*IPCC*), 2007a]. In the last two decades, the IPCC has compared and studied the outputs of different GCMs. These analyses attempt to understand the many processes involved in the current and upcoming climate changes resulting from different greenhouse gas emission scenarios [*IPCC*, 2007a]. In addition, the IPCC seeks to evaluate the potential impacts of climate changes on economy, agriculture, and ecology in the next decades [*IPCC*, 2007b]. Such impact studies require climate simulations at high spatial resolution (small scale), ranging from a few kilometers down to station locations. In particular, precipitation, which is of major importance in agriculture, vegetation, and flood risk assessment, has a strong spatial variability. However, the spatial resolution at which GCMs operate (about 200×200 km) is typically too low to capture such spatial variability. Other reasons why GCMs struggle to reproduce precipitation are related to the features of the distribution of precipitation, namely, boundedness at zero, nonnormality, and the presence of extreme values at local scale with a potential destructive power.

[3] In this context, downscaling techniques have been developed to bridge the gap between large- and small-scale variables [*Hewitson and Crane*, 1996]. There are two different approaches to downscaling. The dynamical approach consists in refining GCMs over a higher-resolution grid. These refined GCMs, called regional climate models (RCMs), operate at a resolution down to about 10 km. RCMs have a high computational cost and thus are often limited in their uses to restricted regions and periods of time. On the other hand, the statistical approach to downscaling proposes statistical models that relate large-scale GCM outputs to local-scale climate

[1]HydroSciences Montpellier, UMR 5569, CNRS/IRD/UM1/UM2, Université de Montpellier 2, Montpellier, France.
[2]Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CNRS/CEA/UVSQ, Gif-sur-Yvette, France.

0043-1397/11/2010WR010128

variables. These statistical downscaling models (SDMs) are, in general, computationally more tractable and can be easily applied to many GCM runs covering large regions and long time periods. Moreover, SDMs offer a great modeling flexibility that has proved useful, for example, in extreme event modeling [e.g., *Vrac and Naveau*, 2007] and for uncertainty assessment [e.g., *Semenov*, 2007]. SDMs generally borrow from one or more of the following sets of methods: transfer functions, stochastic weather generators, and weather typing [e.g., *Maraun et al.*, 2010]. Transfer functions aim at directly translating large-scale data into local-scale values by performing linear or nonlinear regressions. Given large-scale information x, the downscaled local response y is a function $\hat{y}(x)$, which usually estimates $E[Y|\mathbf{X} = x]$ ($\hat{y}(x)$ could estimate another quantile in the case of quantile regression [e.g., see *Cannon*, 2011; *Friederichs and Hense*, 2007]). Among the regression-based transfer functions, we find linear regression [e.g., *Wigley et al.*, 1990; *Huth*, 2002; *Wilby et al.*, 2002; *Busuioc et al.*, 2008; *Goubanova et al.*, 2010], nonlinear parametric models such as polynomial regression [*Hewitson*, 1994; *Sailor and Li*, 1999], nonparametric regression based on splines, generalized additive models [*Vrac et al.*, 2007b; *Salameh et al.*, 2009], and neural networks [e.g., *Snell et al.*, 2000; *Cannon and Whitfield*, 2002; *Haylock et al.*, 2006; *Huth et al.*, 2008; *Ghosh and Mujumdar*, 2008].

[4] Stochastic weather generators (WGs) provide a way to simulate meteorological variables such as precipitation or temperature on the basis of probability density function (pdf) models [e.g., *Semenov and Barrow*, 1997; *Semenov et al.*, 1998; *Wilks*, 1999]. WGs are calibrated so that the simulated observations reproduce the statistical properties of the corresponding local observations. In a downscaling framework, WGs simulate a local variable Y given large-scale information x by building a model for the conditional distribution $Y|\mathbf{X} = x$. To achieve this goal, the parameters of the density model can be seen as functions of some appropriate large-scale information such as American weather regimes [e.g., *Vrac et al.*, 2007a], the North Atlantic Oscillation index [e.g., *Yang et al.*, 2005], or other large-scale climate variables (see *Wilks and Wilby* [1999] for a review). Thus, changes in large-scale variables are transferred into the local-scale density parameters so that the simulated observations evolve accordingly [*Vrac and Naveau*, 2007].

[5] The last set of methods, the weather typing methods, seeks to cluster and classify large-scale atmospheric circulation situations into recurrent weather patterns and assumes that each weather pattern gives rise to similar local-scale meteorological conditions or distributions [e.g., *Huth*, 2001; *Vrac et al.*, 2007c]. Weather typing can serve as a preprocessing step before building transfer functions [e.g., *Huth et al.*, 2008] or weather generators [e.g., *Schnur and Lettenmaier*, 1998; *Vrac et al.*, 2007a; *Vrac and Naveau*, 2007].

[6] Transfer function methods based on neural networks were applied successfully to downscaling [e.g., *Snell et al.*, 2000; *Cannon and Whitfield*, 2002]. They are able to reproduce the nonlinear relationship between large-scale atmospheric data and precipitation [*Hewitson and Crane*, 1996]. However, as regression algorithms, neural networks are limited in the following respects: they underestimate extreme events [*Haylock et al.*, 2006], they provide only pointwise prediction (i.e., no confidence interval or other measure of uncertainty is provided since twice the same input will produce twice the same output), and they cannot account for trends in the variability. In order to enable neural networks to provide probabilistic information, *Williams* [1998] has proposed the following model for rainfall data. The occurrence and the intensity processes are modeled jointly by means of a mixture with a discrete and a continuous component. The discrete component relates to the occurrence process, and the continuous component, which is taken to be a Gamma density, relates to the rainfall intensity process. The parameters of this two-component mixture depend on predictor variables by the functions computed by a neural network. The error function of regression neural networks is the mean-square error, which leads to the estimation of the conditional expectation. In Williams' model, the error function is the conditional log-likelihood of the two-component mixture. In this way, the neural network provides information on the whole distribution of rainfall, not only the conditional expectation. This kind of conditional density model based on a neural network can be seen as a continuous extension of quantile regressions [e.g., *Cannon*, 2011; *Friederichs and Hense*, 2007]. *Williams* [1998] used lagged observations of precipitation as predictor variables to model trend and seasonality of rainfall. Recently, *Haylock et al.* [2006] and *Cawley et al.* [2007] proposed Williams' model as a way for neural networks to model predictive uncertainty in a downscaling application. In the latter, large-scale atmospheric variables are taken as predictor variables for the two-component mixture parameters. Such a model belongs to the class of stochastic WGs since it provides a conditional density model of the local variable given large-scale information.

[7] In this paper, we extend Williams' model by considering a mixture of distributions rather than a single Gamma distribution to model rainfall intensity. Mixtures are flexible nonparametric density estimators that can account for asymmetric and multimodal distributions [e.g., *Priebe*, 1994; *McLachlan and Peel*, 2000]. Our proposed downscaling model is thus a conditional mixture model (CMM) in which one of the components is discrete to model rainfall occurrence. Mixture parameters are estimated by a single layer feed-forward neural network given predictor variables. We evaluate the performances at modeling rainfall intensity of three CMMs that differ in the type of continuous densities (Gaussian, log-normal, or hybrid Pareto) they use as mixture components. We compare CMMs with the two-component conditional mixture from *Williams* [1998] that we use as a benchmark model. We compare these four stochastic models at downscaling precipitation at three rain gauge stations in the French Mediterranean area. The distribution of precipitation is allowed to evolve according to the large-scale atmospheric information in all four stochastic downscaling models. The difference resides in the density model chosen for rainfall intensity. In CMMs, given the state of the atmosphere, described by the predictor variables, precipitation can be in one of several states or regimes, which can be thought of as representing the smaller-scale processes. These regimes are represented by the mixture components.

[8] This paper is structured as follows. Section 2 describes in detail the conditional mixture models. Section 3 presents the precipitation and large-scale data of our downscaling

application together with the preprocessing, training, and model selection steps. Results in terms of log-likelihood and analyses of conditional quantiles and climatological characteristics of the downscaling application are given in section 4. Section 5 provides discussions and conclusions.

2. Statistical Downscaling Models

[9] We adopt the following notation. Let Y be a random variable representing the precipitation process at a given station and let $\mathbf{X}$ be a vector of random variables representing the predictors that include large-scale atmospheric information. Lowercase letters y and x refer to values taken by the corresponding random variables.

2.1. Conditional Mixture Models

[10] Following *Williams* [1998], we consider a mixture with a discrete component to model jointly the occurrence and intensity processes of precipitation:

$$\phi(y;\psi) = (1-\alpha)\delta(y) + \alpha\phi_0(y;\psi_0), \tag{1}$$

where α is the probability of rain occurrence, δ is the Dirac function, which is such that $\int_{-\infty}^{\infty} f(z)\delta(z-a)\mathrm{d}z = f(a)$ and $\delta(z-a) = 0$ for $z \neq a$, and $\phi_0(\cdot;\psi_0)$ is the density model with parameter ψ_0 for rainfall intensity. Therefore, $(1-\alpha)\delta(y)$ handles the "no rain" events while $\alpha\phi_0(\cdot;\psi_0)$ handles the case of positive rainfall. The parameter vector of the mixture in equation (1) is thus $\psi = (\alpha, \psi_0)$. In the work by *Williams* [1998], $\phi_0(\cdot;\psi_0)$ is the Gamma density. We propose to use mixture models instead:

$$\phi_0(y;\psi_0) = \sum_{j=1}^{m} \pi_j f\left(y;\theta_j\right), \tag{2}$$

where $f\left(\cdot;\theta_j\right)$ is a density with parameter vector θ_j, π_j is the weight of component j, and ψ_0 is the vector that concatenates all the mixture parameters $(\pi_1,\ldots,\pi_m,\theta_1,\ldots,\theta_m)$.

[11] We can take into account the dependence of the distribution of precipitation on predictor variables $\mathbf{x}$ by considering the parameters of the mixture of equation (1) as functions of $\mathbf{x}$: $\psi(\mathbf{x}) = [\alpha(\mathbf{x}), \pi_1(\mathbf{x}),\ldots,\pi_m(\mathbf{x}), \theta_1(\mathbf{x}),\ldots,\theta_m(\mathbf{x})]$. In the downscaling application, the predictors are large-scale atmospheric variables. In the conditional mixture model, precipitation can be thought of as being in one of several states or regimes given the state of the atmosphere, as represented by the predictor variables. These states are not directly observed. The hidden states could be seen as resulting from subscale processes that are not accounted for by the large-scale atmospheric variables. Each of the hidden states is modeled by a component of the mixture. The mixture weight $\pi_j(\mathbf{x})$ gives the probability of occurrence of state j, and $f(y;\theta_j(\mathbf{x}))$ is the density of intensity given state j. This view of CMMs bears similarities with nonhomogeneous hidden Markov models (NHMMs) [*Bellone et al.*, 2000]. For NHMMs, the hidden state represents a weather pattern and is assumed to follow a first-order Markov chain whose transition probability depends on the predictors. This is one main difference with CMMs, which capture the serial dependence through the predictor variables exclusively. In NHMMs, the distribution of rainfall in a given hidden state (the so-called emission density) is similar to the component density in the CMMs.

[12] Some authors [e.g., *Hewitson and Crane*, 1996] have shown that the relationship between large-scale atmospheric variables and precipitation is nonlinear. To take this into account, a convenient way to implement the functions in $\psi(\mathbf{x})$ in the conditional mixture is by means of a one-layer feed-forward neural network [*Bishop*, 1995]. Such neural networks are flexible nonparametric models that can, in principle, approximate any continuous function [see *Hornik*, 1991]. We implemented the feed-forward neural network in a standard way but added an extra linear connection between the input variables and the neural network outputs so that the linear model is a special case of the neural network corresponding to zero hidden units. Let H be the number of hidden units. Each hidden unit h, with $h = 1,\ldots,H$, computes a linear combination of the predictors x_i, which is then nonlinearly transformed by means of the hyperbolic tangent (tanh):

$$z_h = \tanh\left(\sum_{i=1}^{d} \upsilon_{h,i}x_i + \upsilon_{h,0}\right), \tag{3}$$

where $\upsilon_{h,i}$ are the input layer weights linking the predictors to the hidden units. Then, a linear combination of the hidden unit activations z_h plus a linear combination of the predictors (the extra linear connection mentioned above) is transformed by a function g in order to ensure range constraint such as positivity:

$$\psi_j = g\left(\sum_{h=1}^{H} w_{j,h}z_h + \sum_{i=1}^{d} \tilde{\upsilon}_{j,i}x_i + w_{j,0}\right), \tag{4}$$

where $w_{j,h}$ are the hidden unit weights which compute the non-linear part, $\tilde{\upsilon}_{j,i}$ are the linear weights of the extra linear connection and g is chosen according to the mixture parameter ψ_j (see *Carreau and Bengio* [2009b] for more details on the function g). Let ω represent the neural network weights, i.e., $\upsilon_{h,i}$ in equation (3) and $w_{j,h}$ and $\tilde{\upsilon}_{h,i}$ in equation (4). Then the conditional mixture can be written as

$$\phi_\omega(y|x) = \phi(y;\psi_w(\mathbf{x})), \tag{5}$$

where $\phi(\cdot;\psi)$ is defined in equation (1) and $\psi_\omega(\mathbf{x})$ emphasizes that the mixture parameters depend on the neural network weights ω. Those weights are calibrated by minimizing the negative log-likelihood of the conditional mixture over the training set. The optimization is done with a conjugate gradient descent algorithm, and the gradient is computed with the back-propagation algorithm [*Rumelhart et al.*, 1986]. To avoid local minima, the optimization is restarted several times from different initial values of the neural network weights, and the weights leading to the lowest training error are kept. Depending on the type of CMMs and on the data set, three to five restarts seem to be enough. This procedure helps to stabilize the optimization and hence the performance of each model. The complexity level of the conditional mixture, that is, its degree of adaptiveness, is controlled by both the number of hidden units of the neural network and the number of components of the

mixture. Those are called hyperparameters. The hyperparameters have to be carefully selected in order to trade off bias (the model misfit with respect to the data) and variance (also called overfitting or learning by heart). This can be done by selecting the number of hidden units and components via the so-called cross-validation method, which will be described in section 3.2.

2.2. Three Families of CMMs

[13] We evaluate three conditional mixture models, all with a discrete component, which differ in the type of mixture components (i.e., $f(\cdot;\theta_j)$ in equation (2)) and compare them with the two-component conditional mixture from *Williams* [1998]. We took Gaussian, log-normal, and hybrid Pareto as mixture components. Since the intensity of rain is strictly positive, we ensure that the Gaussian and the hybrid Pareto mixtures have only positive density on the positive axis by truncation:

$$\phi_0(y;\psi_0) = \begin{cases} \tilde{\phi}_0(y;\psi_0)/\left[1-\tilde{\Phi}_0(0;\psi_0)\right] & \text{if } y<0 \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where $\tilde{\phi}_0(\cdot;\psi_0)$ is either the Gaussian or the hybrid Pareto mixture and $\tilde{\Phi}_0(\cdot;\psi_0)$ is the corresponding cumulative distribution function. The conditional Gaussian mixture was initially presented by *Bishop* [1994]. It combines the approximation capabilities of neural networks and mixture models and can, in principle, represent arbitrary conditional distributions. This motivates the use of conditional Gaussian mixtures to model rainfall intensity. However, when the data set is small and the distribution is heavy tailed, Gaussian mixtures might underestimate the upper tail of the distribution [*Carreau and Bengio*, 2009a]. Since precipitation in the French Mediterranean area where the rain gauges are located is typically heavy tailed [*Delrieu et al.*, 2005], we also considered conditional mixtures with log-normal and hybrid Pareto [*Carreau and Bengio*, 2009a] components. The log-normal density is often employed to model positive intensities of precipitation [see, e.g., *Cho et al.*, 2004]. Indeed, the shape of the log-normal distribution, i.e., its asymmetry and its support on the positive axis, is more suited to model precipitation data. Also, log-normal mixtures are often considered to model moderately heavy tailed data [*McNeil*, 1997; *Frigessi et al.*, 2002]. However, the log-normal might suffer from the same difficulty to model heavy-tailed distribution as the Gaussian. This is because, like the Gaussian, the log-normal tail eventually decreases exponentially fast, whereas the heavy-tailed distribution decreases polynomially fast [*Embrechts et al.*, 1997]. So we propose the hybrid Pareto as a mixture component to explicitly take extreme observations into account. The generalized Pareto distribution (GPD) has been put forward as a model that can approximate all kinds of tails (exponential, polynomial, and finite) [*Pickands*, 1975]. The GPD is designed to model only large observations. The hybrid Pareto provides a smooth extension of the GPD to the whole real axis and makes possible seamless inclusion of the GPD in a mixture while inheriting its tail approximation properties. Another way to include the GPD into a mixture model in a downscaling application was proposed by *Vrac and Naveau* [2007].

[14] The hybrid Pareto is made of a Gaussian stitched together with a generalized Pareto so as to ensure continuity of the density and of its derivative [*Carreau and Bengio*, 2009a]. The hybrid Pareto density is given by

$$h(y;\mu,\sigma,\xi) = \begin{cases} f(y;\mu,\sigma)/\gamma & \text{if } y \le u \\ g(y-u;\xi,\beta)/\gamma & \text{otherwise} \end{cases}, \quad (7)$$

where $f(y;\mu,\sigma)$ is the Gaussian density with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, $u \in \mathbb{R}$ is the junction point or threshold, $\gamma = 1 + \int_{-\infty}^{\alpha} f(y;\mu,\sigma)\mathrm{d}y$ is the normalization factor, and $g(y-u;\xi,\beta)$ is the generalized Pareto density with scale parameter $\beta > 0$ and tail index parameter $\xi \in \mathbb{R}$:

$$g(y-u;\xi,\beta) = \begin{cases} \frac{1}{\beta}\left(1+\frac{\xi}{\beta}(y-u)\right) & \text{if } \xi \neq 0 \\ \frac{1}{\beta}\exp\left(-\frac{y-u}{\beta}\right) & \text{if } \xi = 0 \end{cases}. \quad (8)$$

In addition to the location and scale parameters μ and σ of the Gaussian, the hybrid Pareto has a third parameter, the tail index ξ, which characterizes the heaviness of the tail of the distribution. Positive ξ indicates a polynomially decreasing tail, i.e., a heavy tail; the tail is called exponential or light when $\xi = 0$ and is called finite for negative ξ. Since precipitation in the French Mediterranean area where the rain gauges are located is typically heavy tailed [*Delrieu et al.*, 2005], we will focus on $\xi > 0$ for the hybrid Pareto tail index parameter. Because of the continuity constraints, the junction point u and the scale parameter of the GPD β are functions of ξ, μ, and σ. Note that, by construction, the hybrid Pareto inherits the tail approximation property of the generalized Pareto. The tail heaviness of a mixture of hybrid Pareto is driven by the component with the heaviest tail. The tail index that characterizes the mixture is then given by $\xi^* = \max_j \xi_j$, where ξ_j denotes the tail index parameter of component j. In the conditional case, the tail index depends on the predictor variables $\xi^*(x) = \max_j \xi_j(x)$. In our downscaling application, it means that the behavior of the tail of the distribution of the precipitation process is allowed to vary with regard to the large-scale atmospheric conditions. *Carreau and Bengio* [2009a, 2009b] have shown that in most cases, the hybrid Pareto mixture, conditional or not, is able to provide a decent estimator of the tail index of the data. It is more challenging in the case of conditional density estimation because of the introduction of large uncertainties in the tail index estimation [*Friederichs*, 2010]. For this reason, a penalty term to control the tail index parameter estimation within a hybrid Pareto mixture was proposed by *Carreau et al.* [2009].

2.3. Penalty Term for Tail Index Estimation

[15] Introducing a penalty term is similar to the approach taken by *Coles and Dixon* [1999]. They argued, by analogy with the probability weighted moment (PWM) estimator [*Hosking et al.*, 1985], that the performance of the maximum likelihood estimator (MLE) in small samples could be improved by imposing a restriction similar to $\xi < 1$ (which implies that the expectation is finite). This assumption leads to a reduced variance of the PWM estimator at the cost of a negative bias [*Coles and Dixon*, 1999]. *Coles*

and Dixon [1999] show that adding a penalty that enforces a similar prior assumption into the MLE leads to similar improvement of the MLE estimator.

[16] The penalty designed for the hybrid Pareto mixture is based on the following assumptions. One or a few components should have tail index parameters high enough to model the upper tail of precipitation data correctly. On the other hand, most components should have their tail index parameters close to zero and be dedicated to modeling the central part of the distribution. Previous studies [e.g., *Gardes and Girard*, 2010] lead us to assume that the conditional tail index $\xi^*(x)$ in the region where the studied rain gauges are located varies between 0.16 and 0.26. We thus suggest penalizing the log-likelihood of the hybrid Pareto mixture according to a density, examples of which are shown in Figure 1. This density has one mode at zero (for the majority of components with light tail indexes that are expected to model the central part of the distribution) and a second mode at 0.2 that covers the interval [0.16,0.26] (for the few components that are expected to model the upper tail of rainfall distribution.) The bimodal densities of Figure 1 are given by a two-component mixture made of an exponential and a Gaussian:

$$p(\xi;\eta,\rho) = \eta\exp\{-\eta x\}/2 + \exp\left\{-(\xi-0.2)^2/(2\rho^2)\right\}/\left(2\sqrt{2\pi\rho}\right), \quad (9)$$

where η and ρ are the parameters of the exponential and the standard deviation of the Gaussian, respectively. The exponential puts one mode at zero while the Gaussian is centered at 0.2. The estimation then consists in maximizing a penalized log-likelihood obtained by adding the logarithm of the bimodal density to the usual log-likelihood. In practice, for the conditional mixture, the neural network weights ω are found by minimizing

$$\mathcal{L}(\omega) = -\sum_{i=1}^{n}\log[\phi_\omega(y_i|\mathbf{x}_i)] - \frac{\lambda}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\log p\left(\xi_{i,j};\eta,\rho\right),$$

where n is the number of observations, m is the number of mixture components, $\phi_\omega(y_i|\mathbf{x}_i)$ is the conditional mixture defined in equation (5), evaluated at point i, $p\left(\xi_{i,j};\eta,\rho\right)$ is the bimodal density defined in equation (9) with $\xi_{i,j} = \xi_j(\mathbf{x}_i,\omega)$, and λ controls the trade-off between minimizing the negative log-likelihood and the penalty. The penalty term introduces extra hyperparameters, namely, λ, η, and ρ. These hyperparameters will be chosen by cross-validation; see section 3.2. As η and ρ influence the range of values taken by the tail index parameters of the mixture, we restrict these hyperparameters to vary as follows: $\eta \in \{60, 20\}$ and $\rho \in \{0.02, 0.03, 0.05\}$. This gives the possibility for the penalty to adapt to the rain gauge–specific tail behavior without letting the tail index parameters take unrealistic values.

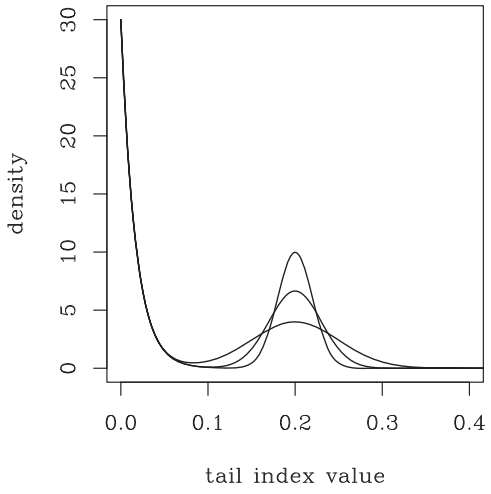


Figure 1. Three examples of bimodal density that reflect our prior assumptions regarding the tail index parameters of the hybrid Pareto mixture in the downscaling application. Most components are assumed to have tail index parameters values near zero (the first mode of the density) to model the central part of the distribution, and some tail index parameters take values near 0.2 (the second mode) to model the upper tail of the distribution.

3. Data Sets and Calibration of the Models

3.1. Large- and Local-Scale Data

[17] The local-scale data are precipitation from three rain gauge stations: Orange, Sète, and Le Massegros. These stations are located in the Cévennes-Vivarais region, which is part of the French Mediterranean area. Because of the Mediterranean influence and of the mountainous back country, the Cévennes-Vivarais region is well known for intense rain events, especially in the fall [*Delrieu et al.*, 2005]. For each rain gauge, we have daily rainfall measurements over 46 years (from 1 January 1959 to 31 December 2004) extracted from the European Climate Assessment & Dataset (ECA&D [*Klein Tank et al.*, 2002]). Precipitation values smaller than 1 mm are set to zero in order to discard values possibly resulting from measurement error.

[18] In this work, we did not seek the best set of predictors to drive the downscaling models. Our goal is to illustrate and compare the performances and advantages of the four downscaling models. For this, we selected a set of predictors with a decent predictive power to drive all four downscaling models, including the benchmark two-component model, so that the comparison among models is fair. Although many large-scale atmospheric variables are relevant to downscale precipitation, we only include the sea level pressure (SLP) field because it has been shown to be a good predictor of precipitation [e.g., *González-Rouco et al.*, 2000]. We selected a subset of the SLP field from the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis data [*Kalnay et al.*, 1996]. We decided to use a simple geometry (a rectangle of reanalysis grid cells centered on the rain gauges) and then proceeded by trial and error to select the size of the rectangle. We found a 6×6 grid to be a reasonable choice to provide large enough regional information in order to capture the relevant large-scale synoptic information.

[19] In addition to the SLP reanalysis data, we include as predictors three date variables representing the year, the

month, and the week of an observation. The year variable encodes the year as the difference with a reference year (here 1970). This allows us to model a potential trend. The month variable is built from the circular difference with the month of January: this difference goes from zero (January) to six (July) and decreases again to 1 (December). The same kind of computation is done to produce the week variable, but the circular difference operation is taken over 52 weeks. The circular differences are then smoothed in order to have continuous variables in the range [0,1]. These circular difference variables (see *Carreau et al.* [2009] for another application of the date variables) are an alternative to using sine and cosine [see *Williams*, 1998] in order to provide a way for the conditional mixtures to characterize seasonality in the distribution of precipitation.

[20] Seasonal signals might be present both in the SLP and the date variables. Therefore, although we are aware that a more classical application of principal component analysis (PCA) [*Jolliffe*, 1986] would separate the treatment of these two kind of variables, we apply PCA on all 39 initial variables (36 SLP plus 3 date variables). Our use of PCA serves merely to reduce dimensionality and remove redundancy among the predictors. The observed predictors are centered and scaled before the application of PCA in order to ensure that their values are in the same range. We extract the four principal components in order to keep 90% of the variance of the data. The four projected predictors are further standardized to have zero mean and unit standard deviation (this preprocessing is required to facilitate the training of neural networks).

3.2. Training and Hyperparameter Selection

[21] Training refers to the optimization of the parameters of the downscaling models (in all four cases, the neural network weights) with respect to an error function that measures the misfit between the model and a data set, called the training set. The error function is chosen according to the task at hand. Since we deal here with conditional density estimation, the error function is the conditional negative log-likelihood (eventually with a penalty term). Hyperparameters, such as the number of hidden units, control the complexity level of an algorithm, which often is proportional to the number of parameters. The higher the complexity level is, the better the fit of the model to the training set becomes. Therefore, hyperparameters cannot be selected by minimizing the error function on the training set. Theoretically, the optimal hyperparameters should minimize simultaneously bias (the model misfit) and variance (also called overfitting). In practice, since the underlying process is unknown, bias and variance have to be estimated. This can be achieved with the cross-validation procedure [*Bishop*, 1995], which computes an estimation of the validation error (the value of the error function on data that are not used for training). The validation error can be decomposed into a sum of bias plus variance and some noise (the residual part of the error that cannot be fit). Let's define a set $\mathcal{H}$ that contains possible values from which to choose the hyperparameters. We implemented a fivefold cross-validation as follows.

[22] 1. Split the training set into five subsets, called folds: $L_1, \ldots, L_5$.

[23] 2. Leave one of the folds, L_k, aside.

[24] 3. For each $h \in \mathcal{H}$, train the model (i.e., determine its parameters given $h \in \mathcal{H}$) on the four remaining folds.

[25] 4. Evaluate the error function $E_k(h)$ (the conditional negative log-likelihood) on the left-aside fold L_k.

[26] 5. Return to step 2 until all fives are left aside in turn.

[27] 6. Select the best hyperparameter: $h^* = \arg\min_h \sum_{k=1}^{5} E_k(h)$.

[28] For the benchmark Gamma conditional model, the only hyperparameter is the number of hidden units. For the conditional mixtures described in sections 2.1 and 2.2, there is a second hyperparameter, which is the number of mixture components. For the hybrid Pareto conditional mixture, there are three additional hyperparameters for the penalty term. The number of hidden units are chosen in the set $\{0,2,4,8\}$, where zero hidden units means that only a linear function is computed by the hidden layer of the neural network, and the number of mixture components is chosen among $\{1,2,4,8\}$. The penalty hyperparameters of the hybrid Pareto CMMs vary as follows: $\lambda \in \{0,1,10,20, 60,100\}$, $\eta \in \{60,120\}$, and $\rho \in \{0.02,0.03,0.05\}$. Therefore, in the cross-validation procedure, we choose among four hyperparameters for the benchmark model, 16 combinations for the Gaussian and log-normal CMMs, and 576 combinations for the hybrid Pareto CMM. For this last model, hyperparameter selection takes about 3 days of computation time on a single CPU.

[29] The 46 year data set is split into a training set of 25 years (from 1 January 1959 to 31 December 1983) and a test set of 21 years (from 1 January 1984 to 31 December 2004). The training set is first used to select the hyperparameters of each SDM with the fivefold cross-validation method. Once the hyperparameters are selected, each model is trained anew on the whole training set. The test set serves exclusively for comparison and evaluation of the SDMs.

4. Results and Analyses

4.1. Global Comparisons

[30] The hybrid Pareto conditional mixture is the most complex model and requires careful and longer training to address properly the issue of conditional tail index estimation. In order to determine if using such a model is justified by the data, we first compare the other three downscaling models in terms of average of differences in conditional log-likelihood with the hybrid Pareto CMM on the test set:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\log \phi_\omega^h(y_i|x_i) - \log \phi_\omega(y_i|x_i)\right], \tag{10}$$

where $\phi_\omega^h(y_i|x_i)$ is the hybrid Pareto CMM density and $\phi_\omega(y_i|x_i)$ is the density of one of the other three models. Positive values indicate that the hybrid Pareto CMM performs better. The evaluation of the performances (or differences in performance) on the test set provides a fair way to compare models even if the number of parameters might vary across models. The risk of overfitting resulting from too many parameters is implicitly taken into account because the performances are evaluated on new data, which did not serve for training or hyperparameter selection. Therefore, we do not need to introduce a penalty term for the number of parameters in the model, as is the case in popular fit criteria

Table 1. Hyperparameters Selected via Cross-Validation for the Three Rain Gauge Stations for Conditional Mixture Models by Type of Components[a]

	Hybrid Pareto	Gaussian	Log-normal	Benchmark
		Orange		
(h, m)	(2, 2)	(0, 4)	(2, 2)	$h = 2$
(λ, η, ρ)	(20, 120, 0.03)	–	–	–
		Sète		
(h, m)	(2, 2)	(0, 8)	(2, 2)	$h = 2$
(λ, η, ρ)	(20, 120, 0.05)	–	–	–
		Le Massegros		
(h, m)	(2, 4)	(4, 4)	(2, 2)	$h = 4$
(λ, η, ρ)	(10, 60, 0.03)	–	–	–

[a]The hyperparameters are the number of hidden units and of components selected (h, m) and penalty parameters (λ, η, ρ) for hybrid Pareto conditional mixtures. The only hyperparameter for the benchmark model is the number of hidden units h.

such as the Bayesian information criterion [*Schwarz*, 1978], which are measured on the training set.

[31] Table 1 presents the hyperparameters selected by fivefold cross-validation on the training set for the four downscaling models. Table 2 shows the average of differences in conditional log-likelihood on the test set along with standard errors for the three competing models (Gaussian CMM, log-normal CMM, and Gamma benchmark) on the three rain gauge stations. The cases where the hybrid Pareto CMM performed significantly better are in bold. We see that the hybrid Pareto outperforms the Gaussian CMM and the Gamma benchmark on all three stations. However, we cannot really distinguish the hybrid Pareto CMM from the log-normal CMM on the basis of this criterion. It is not so surprising that log-normal CMMs have a good performance because the data set is fairly large (which helps training) and the asymmetry and positive support of the log-normal are well suited to model rainfall data.

[32] We illustrate the forthcoming analyses on the Orange station only since the two other stations provide similar insights into the differences between SDMs. We computed the mixture parameters corresponding to the predictors on the test set and then randomly generated precipitation data according to the conditional mixture parameters. This gives us realizations of the precipitation process according to each SDM over the test set. We repeated this a thousand times. To visually assess how realistic a model is at reproducing the precipitation process, we looked at QQ-plots (on logarithmic scale) of the simulated values against the observations in the test set for values greater than 1 mm. This is illustrated in Figure 2 for the hybrid Pareto CMM (the other two CMMs are not shown because they are similar) and the benchmark model. Models that are in accordance with the data should be close to the diagonal line. We see that the benchmark model is less apt at modeling both the central part (overestimation) and the upper part (underestimation) of the distribution.

[33] For each of the 1000 generated time series, we computed the frequency of wet and dry spells of at least k days. These frequency counts were then normalized to obtain proportions. We took the 5% and 95% empirical quantiles over the 1000 replications for each spell length. We also computed the proportion of wet and dry spells of at least k days from the test data. The resulting wet and dry spell confidence intervals together with the spells computed from the observations are compared on the logarithmic scale in Figure 3 for the hybrid Pareto CMM and the benchmark model, with the other two CMMs giving almost identical results. From these wet and dry spell plots, we can conclude that the separate modeling of the occurrence process (for both CMMs and the benchmark model) by means of the discrete mixture component allows us to capture most of the serial dependence, although the shortest wet spell (2 and 3 days) probabilities are slightly underestimated. This dependence is taken into account implicitly by the neural network, which estimates rainfall probability given the predictor variables. Serial dependence could also be explicitly taken into account by including lagged observations in the predictor variables.

[34] As shown in this section, there is a significant gain (in terms of log-likelihood and through the analyses of the QQ-plots) from using a mixture model instead of a single Gamma distribution to model rainfall intensity. In the following analyses, we thus focus on the hybrid Pareto and log-normal CMMs, which are the most likely models for our downscaling application.

4.2. Seasonal Cycles

[35] In this section, we analyze climatologies on the Orange test set. For each day of the 21 years in the test set, we computed the 99% quantile $y_{0.99}(\mathbf{x})$ and the rain probability $\alpha(x)$ according to the SDMs. Then, for each of the 365 days of the generic year (or 366 considering leap years), we estimated empirical quantiles of levels 5%, 50%, and 95% from the 21 values of $y_{0.99}(\mathbf{x})$ and $\alpha(x)$ computed from the models. In order to compare the SDMs with the observations, we also computed the empirical 99% quantile from the test data for each day of the 365/366 days of the year and the proportion of rainy days. For CMMs, conditional quantiles are obtained by solving numerically the following equation for y_p: $\Phi_\omega(y_p|\mathbf{x}) = p$, where $p \in [0,1]$ is the quantile level (i.e., a probability) and $\Phi_\omega(y_p|\mathbf{x})$ is the distribution function associated with the conditional mixture in equation (5). Note that y_p is equal to zero for all $p < 1 - \alpha(x)$, the

Table 2. Average Differences in Log-Likelihood on the Test Set for the Three Rain Gauge Stations Between the Hybrid Pareto CMM and the Other Models[a]

	Gaussian	Log-normal	Benchmark
Orange	**0.02146 (0.003139)**	0.0022512 (0.001910)	**0.02275 (0.002866)**
Sète	**0.01595 (0.003034)**	−0.003530 (0.001647)	**0.01847 (0.002690)**
Le Massegros	**0.01948 (0.006671)**	−0.004606 (0.002121)	**0.02068 (0.003005)**

[a]Standard errors are given in parentheses. Positive numbers indicate that the hybrid Pareto CMM performed better. Significant differences are in bold.

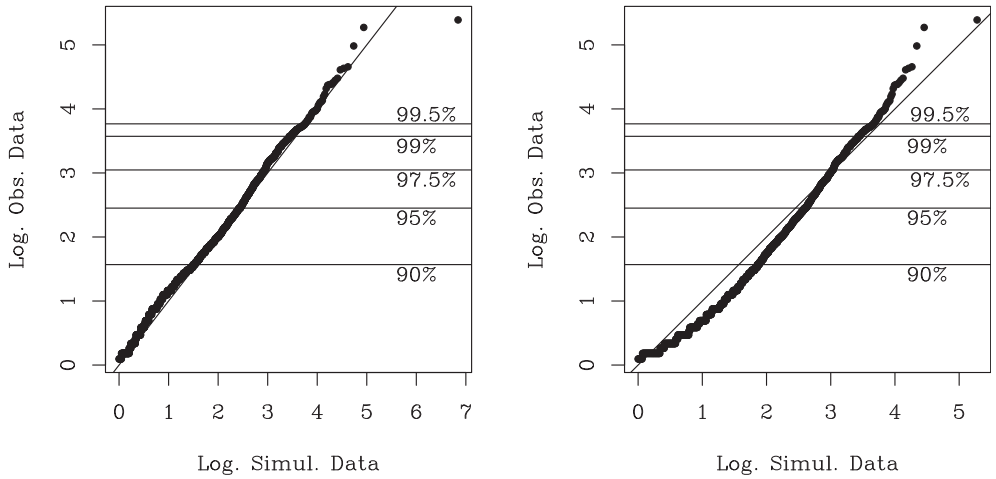


Figure 2. QQ-plots on a logarithmic scale of the simulated precipitation versus observations >1 mm on the Orange test set for (left) the hybrid Pareto conditional mixture and (right) the benchmark model. The horizontal lines are the empirical unconditional quantiles from observations of the test set.

no-rain probability, since in that case y_p falls in the discrete part of the distribution.

[36] Figure 4 shows the daily seasonal cycle of the rain probability and the daily seasonal cycle of the 99% conditional quantile for the Orange test data. The black line is the climatology computed from the observations in the test set: in Figure 4 (left), this is the proportion of rainy days, whereas in Figure 4 (right) it is the empirical 99% quantile. The gray band is the 90% confidence interval based on the empirical quantiles of 5% and 95% levels computed from the hybrid Pareto CMM, and the white line corresponds to the empirical median. In Figure 4 (left), the empirical quantiles are estimated from the 21 modeled conditional probabilities of rain $\alpha(x)$ per day, and in Figure 4 (right) they are estimated on the 21 modeled conditional quantiles of the 99% level, $y_{0.99}(\mathbf{x})$, per day. It is not meaningful to look at more central conditional quantiles because of the presence of the discrete component, which makes these conditional quantiles hard to interpret. From Figure 4 (left), we can identify two seasonal modes, around March (month 3) and October (month 10), which translates into higher probabilities of rain around these two months, while summer (i.e., around July) presents smaller probabilities of rain. This is in agreement with the observations over the test set, showing the same features. Regarding the 99% quantile in Figure 4 (right), the picture is a bit less clear, but we can nevertheless identify two modes in the median (white line) just after March and around fall (months 9 and 10). Larger rainfall amounts are thus expected in spring and fall. The climatologies are based on empirical quantiles computed on 21 values (because of the 21 years in the test set). Few positive observations occurred for a given day out of the 21, and empirical quantiles are thus not very stable (especially at the 99% level). This explains the spiky features in Figure 4. The seasonal cycles for the other three SDMs are very similar and are thus not shown.

[37] In order to get some understanding of the way conditional mixtures approximate the conditional distribution of precipitation intensity, we now look at the daily seasonal cycle of the mixture parameters (the continuous part of the mixture; see equation (1)) on the test set. The mixture parameters are deterministic functions of the predictor variables and therefore vary with the values taken by these. Just as before, for each of the 365/366 days of the year, empirical quantiles of the 5%, 50%, and 95% levels are estimated from the 21 values of the mixture parameters on the test set. Figure 5 depicts the climatology of the hybrid Pareto CMM mixture weight, tail index, location, and scale parameters over the Orange test set. For this station, the cross-validation procedure selected two components in the mixture. The white and the black lines represent the empirical median of the mixture parameter values for each component. The gray bands are the 90% confidence intervals based on the empirical quantiles of the 5% and 95% levels. The seasonal cycle of the mixture weights (or priors) in Figure 5 (top left) shows the predominance of one or the other component in the density. For the hybrid Pareto CMM, we see that in July (month 7), the component associated with the black line tends to dominate, although not significantly, as we see that the confidence intervals overlap. The seasonal cycle of the tail index parameters is shown in Figure 5 (top right). We observe a slight decrease of the tail index parameter from both components in the summer, but globally, they are around the value 0.2, which is in agreement with the penalty of the MLE (see equation (9)). For the Orange data, the penalty hyperparameters η and ρ selected via cross-validation imply that the tail index parameters have just one mode centered on 0.2. For the other two data sets, Sète and Le Massegros, the penalty hyperparameters selected entail that the tail index parameters have two modes centered on zero and on 0.2 with different heights. These results are not shown. Figure 5 (bottom) depicts the cycles of the location and scale parameters of the hybrid Pareto components. There is a strong seasonal signal with a clear bump of the white line component for both location and scale parameters around summertime,

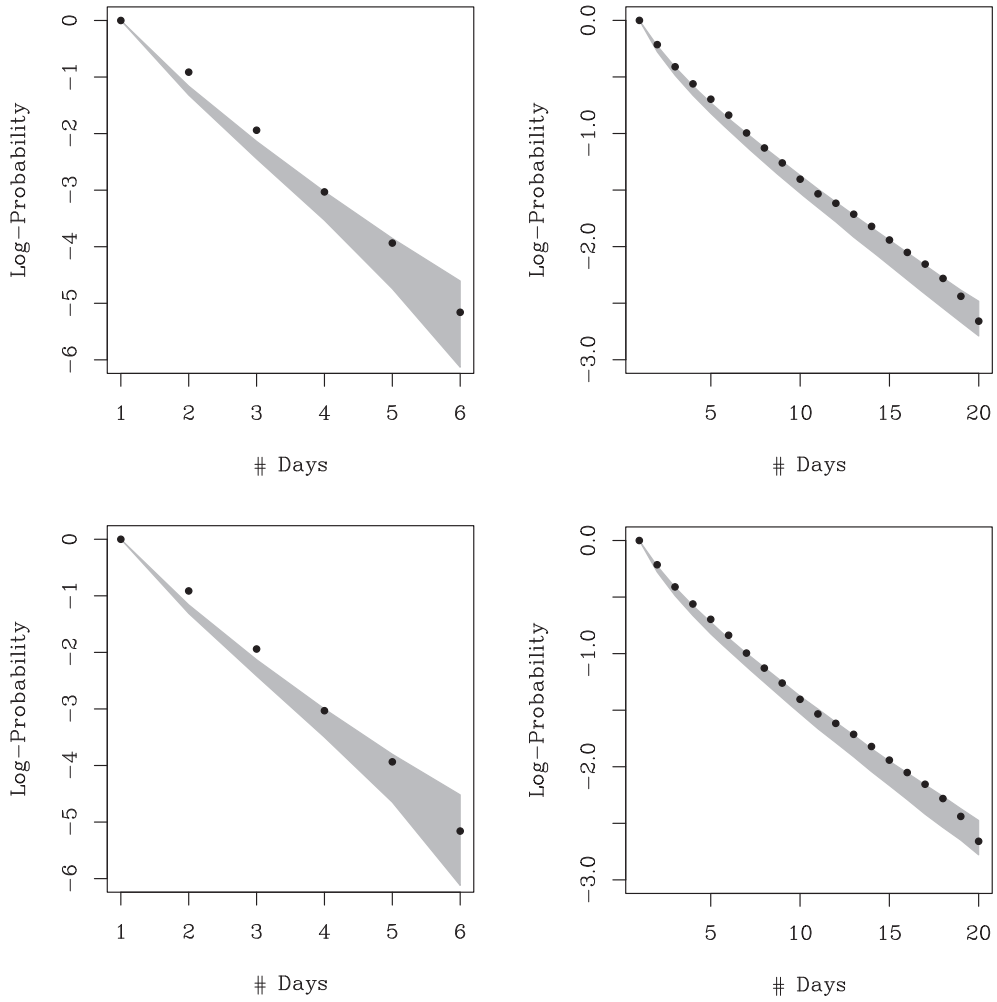


Figure 3. Logarithm of the proportion of (left) wet spells and (right) dry spells of at least k days for the Orange test data (dots) together with 90% empirical confidence interval (gray band) from (top) the hybrid Pareto CMM and (bottom) the benchmark model. The predictor variables allow implicit modeling of serial dependence of rainfall occurrence and provide similar spell probabilities in all four downscaling models.

thereby providing a larger probability of high-intensity rainfall. Keeping in mind that this continuous mixture density is conditional on the fact that it rains ($Y > 0$), it means that when it rains, the rainfall intensity is potentially high. However, we saw in Figure 4 (left) that the probability of rain reaches a low in the middle of summer. This reflects common knowledge about rainfall in this area of France: summer rain is rare but intense.

[38] As a means of comparing the log-normal versus the hybrid Pareto CMMs, we also include the climatologies of the log-normal conditional mixture parameters on the Orange test data; see Figure 6. In this case as well, two components were selected via cross-validation. From the climatology of the mixture weights in Figure 6 (top), we observe a pattern similar to the one of the hybrid Pareto CMM: one of the components dominates in the summer. In this case, the difference is more accentuated. The log-normal distribution has two parameters, μ and σ^2, which are directly linked to the Gaussian parameters. However, unlike the Gaussian, they do not represent the expectation and the variance of the log-normal. So instead of the μ and σ^2, we have computed the climatologies of the expectation and the variance of the log-normal components, which are given by $e^{\mu+\sigma^2/2}$ and $e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$, respectively. These are plotted in Figure 6 (middle and bottom). In summer, we see the same bump as in Figure 5 (bottom), albeit less pronounced for the expectations (Figure 6, middle). However, the values taken by the variance (Figure 6, bottom) are much higher: the full 90% confidence interval takes values up to about 600, which makes for a standard deviation up to about 25. This means that the tail is not heavy enough to

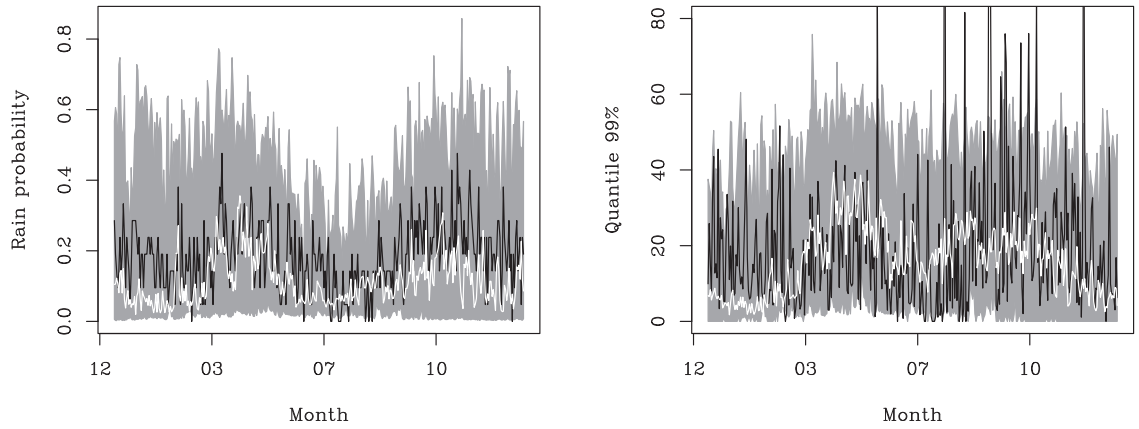


Figure 4. Daily seasonal cycles of (left) the rain occurrence probability and (right) the 99% quantile from the observations (black line) together with an empirical 90% confidence interval (gray band) and median (white line) from the hybrid Pareto conditional mixture for the Orange station test data. Peaks in the seasonal cycle of the occurrence process are visible in spring and fall.

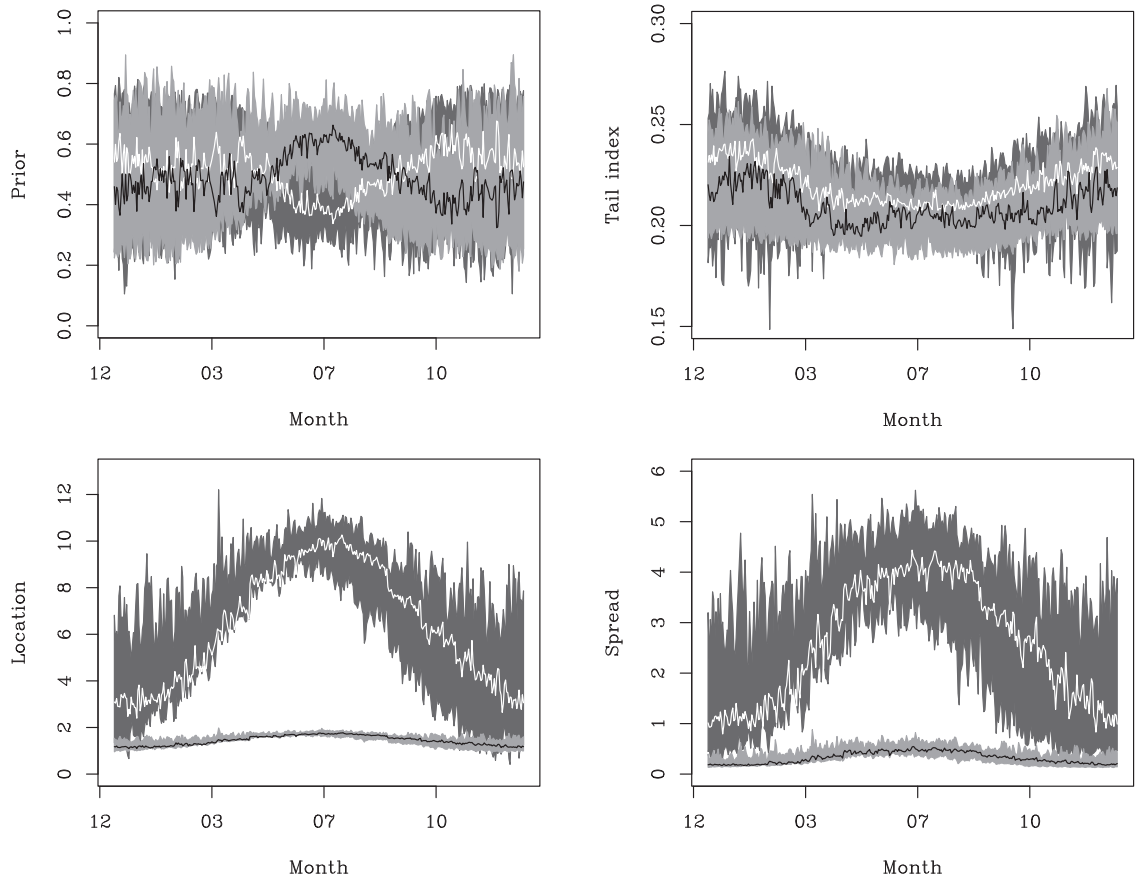


Figure 5. Daily seasonal cycles of the hybrid Pareto conditional mixture parameters (top left to bottom right: mixture weights π_j, tail index parameters ξ_j, location parameters μ_j, and scale parameters σ_j) together with an empirical 90% confidence interval. The mixture has two components whose parameters are represented by the black and white lines.

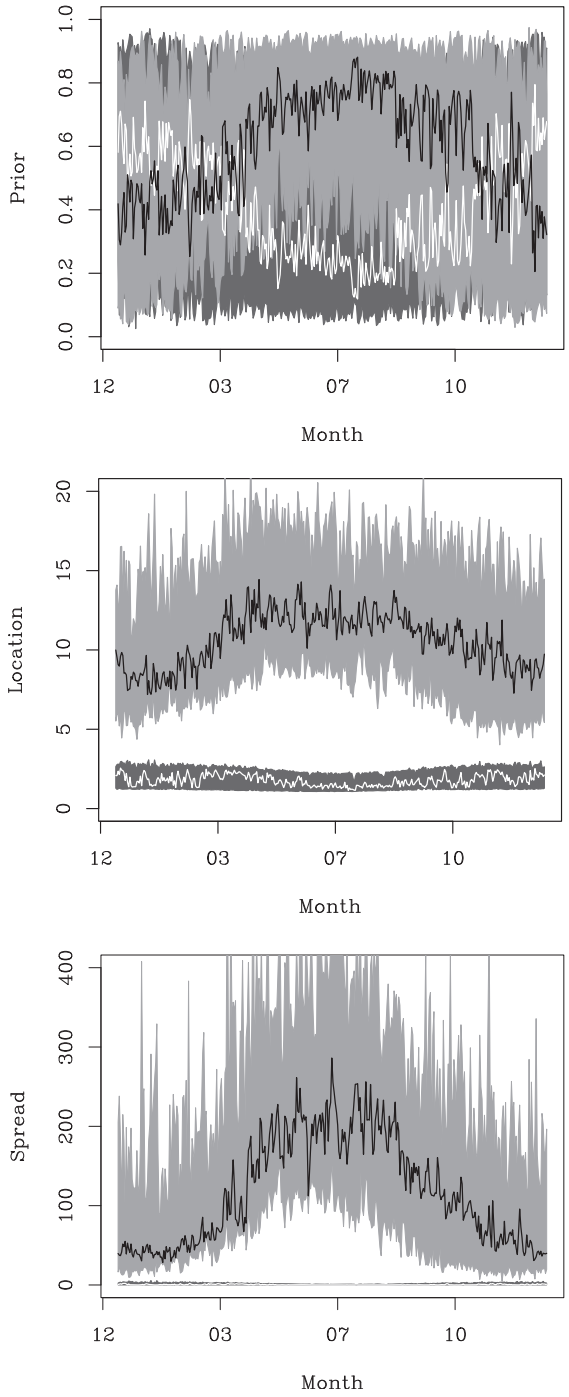


Figure 6. Daily seasonal cycles of the log-normal conditional (top) mixture weights π_j, (middle) component expectations $e^{\mu_j+\sigma_j^2/2}$, and (bottom) component variances $e^{2\mu_j+\sigma_j^2}\left(e^{\sigma_j^2}-1\right)$ together with an empirical 90% confidence interval. The mixture has two components whose parameters are represented by the black and white lines.

correctly represent rainfall. To compensate, the log-normal conditional mixture needs to inflate the variance of one of the components in order to take properly extreme rainfalls into account. We also note that the component with large location and scale in the summer has the largest prior. Therefore, the log-normal CMM is not very realistic. The hybrid Pareto CMM, because of the tail index parameters, does not need to increase the variance. However, the fine-tuning of the penalty term of the MLE (see equation (9)) requires some extra care and greatly influences the final model.

4.3. Analyses of Conditional Events

[39] This section concerns analyses of two specific rain events at the Orange station from the test set: the wet spell with the highest volume of rain (322 mm on 8–9 September 2002) and the longest wet spell (9 days, from 24 April 1993 to 2 May 1993). Our goal is to check whether the SDMs' continuous part can capture rainfall intensity during particularly severe events that might be challenging to reproduce. Results are shown for the hybrid Pareto CMM only. In Figure 7, modeled conditional quantiles of the 95%, 99%, and 99.9% levels, i.e., $y_{0.95}(\mathbf{x})$, $y_{0.99}(\mathbf{x})$, and $y_{0.999}(\mathbf{x})$, are shown in light, medium, and dark gray, respectively, for the hybrid Pareto CMM for these two rain events. The black vertical lines represent the observed precipitation on each day. We can see how the hybrid Pareto CMM adapts to the atmospheric conditions that yield a sequence of dry days or little rain, then a high or long volume of precipitation, and then back to a drier period. If the model is right, we expect the observed precipitation to be below the quantile level $p \times 100\%$ of the time. In other words, we expect the observed precipitation to exceed the modeled conditional 99.9% quantile on average 0.01%, which means, given the test set size, about one observation. The largest observation in the training set for the Orange station is 137 mm, whereas the largest observation in the test set is 220 mm. The latter appears at the date 09/08 in Figure 7 (left) and corresponds to a quantile level of 99.99% according to the hybrid Pareto CMM. On average, the tested SDMs are pretty accurate with regard to conditional quantile modeling; see Table 3.

[40] Another way to look at the evolution of the model through a rain event is by looking directly at the conditional densities, which are associated with different atmospheric conditions, that is, for different predictors. More precisely, we depicted the conditional densities of Y given the predictors x and given that $Y > 0$. For the conditional mixtures, it corresponds to the continuous part of the mixture in equation (1): $\alpha(\mathbf{x})\phi_0(y;\psi_\omega(\mathbf{x}))$. These conditional densities are illustrated in Figure 8 for the hybrid Pareto CMM on the wet spell with the highest volume of rain in the Orange test data. This is the same rain event as in Figure 7 for the conditional confidence intervals. Figure 8 (left) shows the central part of the conditional densities, while Figure 8 (right) represents the upper tails in logarithmic scale. Each curve corresponds to a different day, which is connected in the legend with the amount of rain observed on that day. The days in the legend are presented in chronological order (from top to bottom). We see from Figure 8 (left) that the density of the rain intensity is very low, almost flat, for days where no precipitation occurred

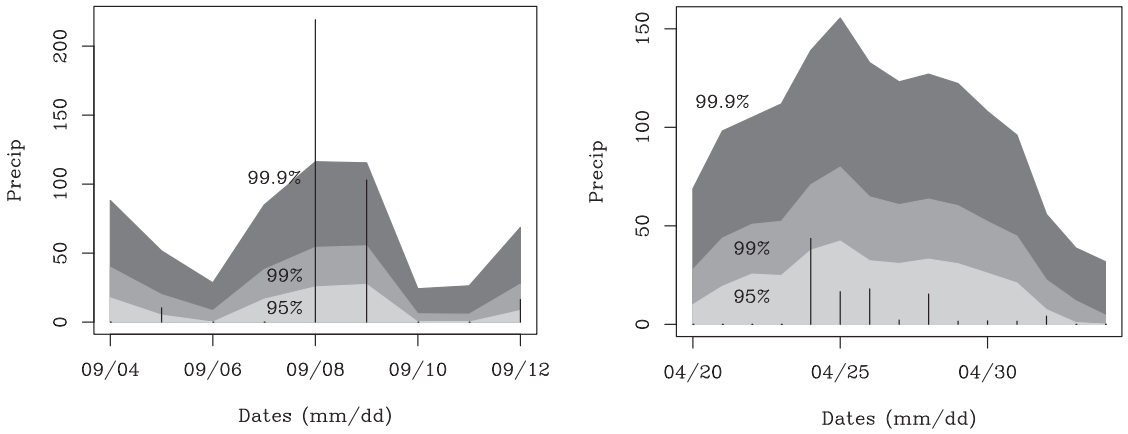


Figure 7. Hybrid Pareto conditional mixture: conditional quantiles $y_{0.95}(\mathbf{x})$ (light gray), $y_{0.99}(\mathbf{x})$ (medium gray), and $y_{0.999}(\mathbf{x})$ (dark gray) on (left) the wet spell with the highest volume of rain and (right) the longest wet spell from the Orange test data. Vertical lines represent the observed precipitation. The largest observed rainfall, 220 mm on 8 September (Figure 7, left), corresponds to a quantile level of 99.99% for the hybrid Pareto CMM.

(curves in warm hues). The corresponding tails in Figure 8 (right) decrease rapidly. The yellow curve, which is the last day with zero precipitation preceding the wet spell, is an exception; this density curve is similar to the density curves of the intense wet days (in green hues). For the days of heavy rains, the upper tail of the conditional density is heavier. We performed similar analyses on the other data sets and with the other three SDMs. There are differences (not shown) caused by the choice of the statistical model for rainfall intensity and reflecting the underlying assumptions. For instance, during an intense rain event, the upper tails of the Gaussian mixture density curves decrease much more rapidly than the upper tails of the hybrid Pareto CMM, indicating a lesser risk of extreme rainfalls for the Gaussian CMM. In general, from these analyses, we see that the conditional mixture model is very responsive to a change in atmospheric conditions. The mixture parameters change to adjust the shape of the distribution of precipitation in a consistent way (heavier upper tail is associated with extreme rainfalls).

5. Conclusions and Discussion

[41] The focus of this paper is on conditional mixture models (CMMs), which to our knowledge, are used for the first time in a downscaling context and open interesting ways to study the interactions between large- and small-scale climate variables. CMMs are flexible stochastic downscaling models: a discrete component in the mixture serves to simulate the occurrence process, whereas the continuous mixture part simulates the intensity process. The continuous mixture is made of either one of the following types of component: Gaussian, log-normal, or hybrid Pareto. The mixture parameters, i.e., mixture weights and both discrete and continuous components parameters, are functions of predictor variables (including large-scale information) and are computed by means of neural networks. CMMs extend the two-component mixture proposed initially by *Williams* [1998], which has a discrete component like CMMs to model the occurrence process but relies on a single density, the Gamma, for the rainfall intensity process. This simpler model, which was used in a downscaling context [*Haylock et al.*, 2006; *Cawley et al.*, 2007], acted as a benchmark model.

Table 3. Percentage of Observations Below Estimated Quantiles of Various Probability Levels (p = 0.05, 0.5, 0.9, 0.0975, 0.99, 0.999) for the Hybrid Pareto, Gaussian, and Log-normal CMMs and the Benchmark Model for the Three Rain Gauge Stations on the Test Set[a]

	Hybrid Pareto	Gaussian	Log-normal	Benchmark
		Orange		
$p = 0.05$	–	0	–	–
$p = 0.5$	44.1	50.9	45	46.3
$p = 0.9$	88.6	88.8	89.6	89.9
$p = 0.95$	94.2	94.7	94.4	94.9
$p = 0.975$	96.4	97.2	96.7	96.8
$p = 0.99$	98.5	98.9	98.5	98.5
$p = 0.999$	99.7	99.8	99.8	99.6
		Sète		
$p = 0.05$	–	–	–	–
$p = 0.5$	47.8	51.4	48.9	47.4
$p = 0.9$	89.6	89.7	90.6	91.7
$p = 0.95$	94.3	95.1	94.6	95.1
$p = 0.975$	97.4	97.7	97.5	97.5
$p = 0.99$	98.8	99.1	98.9	98.9
$p = 0.999$	99.7	100	99.9	99.7
		Le Massegros		
$p = 0.05$	8.3	16.7	15.6	9.7
$p = 0.5$	47.6	48.7	48.4	50.4
$p = 0.9$	89.1	89.2	89.0	89.8
$p = 0.95$	95.0	95.0	95.2	95.1
$p = 0.975$	97.4	97.5	97.6	97.2
$p = 0.99$	99.0	98.9	99.0	98.5
$p = 0.999$	99.9	99.9	99.9	99.8

[a]Missing values are because the estimated quantiles fall in the discrete part of the distribution. If the model is correct, the percentage of observations should be close to $p \times 100$.

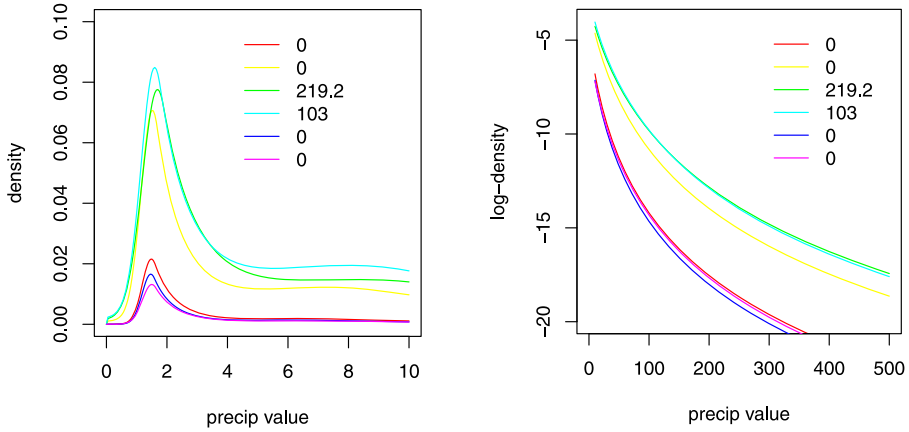


Figure 8. (left) Central part and (right) upper tail in logarithmic scale of the conditional densities $\alpha(\mathbf{x})\phi_0(y;\psi_\omega(\mathbf{x}))$ for the hybrid Pareto CMM day by day for a period comprising the wet spell with the highest volume of rain in the test Orange data (same period as in Figure 7, left). Each daily density is represented with a different color, which is represented in the legend in chronological order, from top to bottom, with the amount of rainfall observed on that day.

[42] We draw the following conclusions from our analyses of the three stations in the French Mediterranean area. First, conditional mixture models have a clear advantage over the benchmark model in terms of flexibility to represent both the central and the extremal part of rainfall intensity distribution. Second, modeling the occurrence process separately with a discrete component allows the implicit introduction of serial dependence through the predictor variables and hence the reproduction of wet and dry spell sequences. Finally, the choice of component in CMMs depends on the data. In our case, Gaussian components are not well suited. Log-normal CMMs offer a good performance and are more straightforward to implement than hybrid Pareto CMMs. However, the assumption of heavy tails of the hybrid Pareto CMM seems more realistic for the precipitation data considered in this work.

[43] The downscaling models considered in this work are all weather generators: they downscale the whole conditional distribution of precipitation from which we can answer all kind of questions. (1) Conditional quantiles can be computed, and from these, climatologies can be examined and confidence intervals can be constructed. (2) Conditional densities give insight into the influence of the atmospheric information on the distribution of precipitation. (3) It is easy to simulate rainfall and check whether the features of observed precipitation, such as wet and dry spells, are well captured by the models.

[44] Although conditional mixtures are rather complex models, some understanding of the modeling mechanisms can be gained by looking at the climatologies of the mixture parameters as functions of the covariates. We believe that the multiple benefits from the conditional mixtures compensate for the extra work of implementation. Note that a package in the R language [*R Development Core Team*, 2010] named CondMixt has been developed for this study and should be made available in the near future.

[45] This study has multiple perspectives and future works. For example, the choice of appropriate predictors should be made with care and requires further analyses. In this paper, our goal was to illustrate and compare the performances and advantages of the proposed downscaling models. We provided the downscaling models with a decent set of predictors without looking for the best set of predictors. The results from the hyperparameter selection for the mixtures show that single-component models are inadequate. This could indicate that more than one component is required, but it could also be due to inappropriate selection of predictors. Hence, although the comparisons between models are fair, a more complete set of predictors describing better the physical processes at play could yield more accurate statistical properties and simulations for all models.

[46] Another interesting perspective would be to evaluate climate change in precipitation according to these downscaling models. For this, we need to validate the use of GCM outputs as predictors in CMMs trained on reanalysis data. Indeed, this is essential to assess the reliability of the "couple" (GCM and CMM) and to give confidence in their present projections before applying CMMs to downscale distributions of rainfall under various potential future greenhouse gas emission scenarios. This would help us to evaluate the impact of climate change on very important features of rainfall, such as (interannual or statistical) variability, seasonality, or extremes.

[47] Finally, a challenging extension of this downscaling approach would be to take into account spatial dependencies between different rain gauges. This would allow joint modeling of precipitation at multiple sites with multivariate CMMs. Possible approaches to model the spatial dependence structure of precipitation include the use of copulas [*Nelsen*, 2006] and the approach suggested by *Cannon* [2008] to change the error function in order to encourage the model to match the observed covariance matrix. The resulting coherent spatial simulations should preserve observed dependencies and would provide tools to understand their temporal evolutions during a control time period,

past or future climate, completing the information brought by CMMs to study the many facets of climate changes.

[48] **Acknowledgments.** The authors thank the AssimileX and ACQWA project teams. M. Vrac has been partly funded by the GIS-REGYNA project. J. Carreau has been partly funded by the Fonds québecois de la recherche sur la nature et les technologies (FQRNT).

References

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

Bishop, C. M. (1994), Mixture density networks, technical report, Aston Univ., Birmingham, U. K.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Clarendon, Oxford, U. K.

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

Cannon, A. J. (2008), Probabilistic multisite precipitation downscaling by an expanded Bernouilli-Gamma density network, *J. Hydrometeorol.*, *9*, 1284–1300.

Cannon, A. J. (2011), Quantile regression neural networks: Implementation in R and application to precipitation downscaling, *Comput. Geosci.*, *37*(9), 1277–1284.

Cannon, A. J., and P. H. Whitfield (2002), Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models, *J. Hydrol.*, *259*, 136–151.

Carreau, J., and Y. Bengio (2009a), A hybrid Pareto model for asymmetric fat-tailed data: The univariate case, *Extremes*, *12*, 53–76.

Carreau, J., and Y. Bengio (2009b), A hybrid Pareto mixture for conditional asymmetric fat-tailed distributions, *IEEE Trans. Neural Networks*, *20*(7), 1087–1101.

Carreau, J., P. Naveau, and E. Sauquet (2009), A statistical rainfall-runoff mixture model with heavy-tailed components, *Water Resour. Res.*, *45*, W10437, doi:10.1029/2009WR007880.

Cawley, G. C., G. J. Janacek, M. R. Haylock, and S. R. Dorling (2007), Predictive uncertainty in environmental modelling, *Neural Networks*, *20*, 537–549.

Cho, H.-K., K. P. Bowman, and G. R. North (2004), A comparison of Gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission, *J. Appl. Meteorol.*, *43*, 1586–1597.

Coles, S. G., and M. J. Dixon (1999), Likelihood-based inference for extreme value models, *Extremes*, *2*, 5–23.

Delrieu, G., et al. (2005), The catastrophic flash-flood event of 8–9 September 2002 in the Gard region, France: A first case study for the Cévennes Vivarais Mediterranean hydrometeorological observatory, *J. Hydrometeorol.*, *6*(1), 34–52.

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997), *Modelling Extremal Events*, Springer, New York.

Friederichs, P. (2010), Statistical downscaling of extreme precipitation events using extreme value theory, *Extremes*, *13*, 109–132.

Friederichs, P., and A. Hense (2007), Statistical downscaling of extreme precipitation events using censored quantile regression, *Mon. Weather Rev.*, *135*, 2365–2378.

Frigessi, A., O. Haug, and H. Rue (2002), A dynamic mixture model for unsupervised tail estimation without threshold selection, *Extremes*, *5*, 219–235.

Gardes, L., and S. Girard (2010), Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels, *Extremes*, *13*, 177–204.

Ghosh, S., and P. P. Mujumdar (2008), Statistical downscaling of GCM simulations to streamflow using relevance vector machine, *Adv. Water Resour.*, *31*(1), 132–146.

Gladstone, R., et al. (2005), Mid-Holocene NAO: A PMIP2 model intercomparison, *Geophys. Res. Lett.*, *32*, L16707, doi:10.1029/2005GL023596.

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

Goubanova, K., V. Echevin, B. Dewitte, F. Codron, K. Takahashi, P. Terray, and 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, *Clim. Dyn.*, *36*(7–8), 1365–1378.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby, and C. M. Goodess (2006), Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios, *Int. J. Climatol.*, *26*, 1397–1415.

Hewitson, B. (1994), Regional climates in the GISS general circulation model: Surface air temperature, *J. Clim.*, *7*(2), 283–303.

Hewitson, B. C., and R. G. Crane (1996), Climate downscaling: Techniques and application, *Clim. Res.*, *7*, 85–95.

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

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985), Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, *Technometrics*, *27*, 251–261.

Huth, R. (2001), Disaggregating climatic trends by classification of circulation patterns, *J. Climatol.*, *21*, 135–153.

Huth, R. (2002), Statistical downscaling of daily temperature in central Europe, *J. Clim.*, *15*, 1731–1742.

Huth, R., S. Kliegrová, and L. Metelka (2008), Non-linearity in statistical downscaling: Does it bring an improvement for daily temperature in Europe?, *J. Climatol.*, *28*, 465–477.

Intergovernmental Panel on Climate Change (2007a), *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., Cambridge Univ. Press, Cambridge, U. K.

Intergovernmental Panel on Climate Change (2007b), *Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by M. L. Parry et al., Cambridge Univ. Press, Cambridge, U. K.

Jolliffe, I. T. (1986), *Principal Component Analysis*, Springer, New York.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*(3), 370–471.

Klein Tank, A. M. G., et al. (2002), Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment, *Int. J. Climatol.*, *22*, 1441–1453.

Maraun, D., et al. (2010), Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, *Rev. Geophys.*, *48*, RG3003, doi:10.1029/2009RG000314.

McLachlan, G., and D. Peel (2000), *Finite Mixture Model*, John Wiley, New York.

McNeil, A. J. (1997), Estimating the tails of loss severity distributions using extreme value theory, *Astin Bull.*, *27*, 117–137.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer, New York.

Pickands, J. (1975), Statistical inference using extreme order statistics, *Ann. Stat.*, *3*, 119–131.

Priebe, C. E. (1994), Adaptive mixtures, *J. Am. Stat. Assoc.*, *89*, 796–806.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986), Learning internal representations by error propagation, in *Parallel Distributed Processing: Explorations in the Macrostructure of Cognition*, vol. 1, pp. 318–362, MIT Press, Cambridge, Mass.

Sailor, D. J., and X. Li (1999), A semi-empirical downscaling approach for predicting regional temperature impacts associated with climatic change, *J. Clim.*, *12*, 103–114.

Salameh, T., P. Drobinski, M. Vrac, and P. Naveau (2009), Statistical downscaling of near surface wind field over complex terrain in southern France, *Meteorol. Atmos. Phys.*, *103*, 253–265.

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

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*, 461–464.

Semenov, M. A. (2007), Development of high-resolution UKCIP02-based climate change scenarios in the UK, *Agric. For. Meteorol.*, *144*, 127–138.

Semenov, M. A., and E. M. Barrow (1997), Use of a stochastic weather generator in the development of climate change scenarios, *Clim. Res.*, *35*, 397–414.

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

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

Vrac, M., and P. Naveau (2007), Stochastic downscaling of precipitation: From dry events to heavy rainfalls, *Water Resour. Res.*, *43*, W07402, doi:10.1029/2006WR005308.

Vrac, M., M. Stein, and K. Hayhoe (2007a), Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing, *Clim. Res.*, *34*, 169–184.

Vrac, M., P. Marbaix, D. Paillard, and P. Naveau (2007b), Non-linear statistical downscaling of present and LGM precipitation and temperatures over Europe, *Clim. Past*, *3*, 669–682.

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

Wigley, T. M. L., P. D. Jones, K. R. Briffa, and G. Smith (1990), Obtaining sub-grid scale information from coarse resolution general circulation model output, *J. Geophys. Res.*, *95*, 1943–1953, doi:10.1029/JD095iD02p01943.

Wilby, R. L., C. W. Dawson, and E. M Barrow (2002), SDSM—A decision support tool for the assessment of regional climate change impacts, *Environ. Modell. Software*, *17*(2), 145–157.

Wilks, D. S. (1999), Multisite downscaling of daily precipitation with a stochastic weather generator, *Clim. Res.*, *11*, 125–136.

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

Williams, M. P. (1998), Modelling seasonality and trends in daily rainfall data, *Adv. Neural Inf. Process. Syst.*, *10*, 985–991.

Yang, C., R. E. Chandler, and V. S. Isham (2005), Spatial-temporal rainfall simulation using generalized linear models, *Water Resour. Res.*, *41*, W11415, doi:10.1029/2004WR003739.

J. Carreau, HydroSciences Montpellier, UMR 5569, CNRS/IRD/UM1/UM2, Université de Montpellier 2, Case MSE, place Eugène Bataillon, F-34095 Montpellier CEDEX 5, France. (julie.carreau@univ-montp2.fr)

M. Vrac, Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CNRS/CEA/UVSQ, Orme des Merisiers, F-91191 Gif-sur-Yvette, France. (mathieu.vrac@lsce.ipsl.fr)

A.3 Caractérisation spatiale des précipitations intenses

A.3.1 Approche régionale basée sur la loi de Pareto généralisée

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2017WR020758

Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation

J. Carreau[1], P. Naveau[2], and L. Neppel[1]

[1]HSM, CNRS/IRD/UM, Université de Montpellier, Montpellier, France, [2]LSCE, IPSL-CNRS, Orme des Merisiers, Gif-sur-Yvette, France

Key Points:
- Regional peaks-over-threshold formalized as a conditional mixture model
- Inference strategy based on probability weighted moments and nonparametric estimators
- Selection of the number of subregions with a cross-validation procedure

Supporting Information:
- Supporting Information S1

Correspondence to:
J. Carreau,
julie.carreau@ird.fr

Citation:
Carreau, J., P. Naveau, and L. Neppel (2017), Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation, *Water Resour. Res.*, *53*, 4407–4426, doi:10.1002/2017WR020758.

Received 17 MAR 2017
Accepted 11 MAY 2017
Accepted article online 17 MAY 2017
Published online 31 MAY 2017

Abstract The French Mediterranean is subject to intense precipitation events occurring mostly in autumn. These can potentially cause flash floods, the main natural danger in the area. The distribution of these events follows specific spatial patterns, i.e., some sites are more likely to be affected than others. The peaks-over-threshold approach consists in modeling extremes, such as heavy precipitation, by the generalized Pareto (GP) distribution. The shape parameter of the GP controls the probability of extreme events and can be related to the hazard level of a given site. When interpolating across a region, the shape parameter should reproduce the observed spatial patterns of the probability of heavy precipitation. However, the shape parameter estimators have high uncertainty which might hide the underlying spatial variability. As a compromise, we choose to let the shape parameter vary in a moderate fashion. More precisely, we assume that the region of interest can be partitioned into subregions with constant hazard level. We formalize the model as a conditional mixture of GP distributions. We develop a two-step inference strategy based on probability weighted moments and put forward a cross-validation procedure to select the number of subregions. A synthetic data study reveals that the inference strategy is consistent and not very sensitive to the selected number of subregions. An application on daily precipitation data from the French Mediterranean shows that the conditional mixture of GPs outperforms two interpolation approaches (with constant or smoothly varying shape parameter).

1. Introduction

In the French Mediterranean, heavy precipitation events occurring mainly in autumn, often called *Cevenol events* tend to gather in very specific areas. Factors explaining the spatial distribution of these events are the presence of mountains and the trajectories usually taken by contrasting air masses, humid, and warm coming from the Mediterranean sea and cold from the North. Heavy precipitation might trigger flash floods, the main natural danger in the Mediterranean area, that can potentially cause fatalities and important material damage [*Delrieu et al*., 2005; *Borga et al*., 2011; *Braud et al*., 2014].

Extreme value theory [*Coles*, 2001] provides a sound asymptotic framework commonly used to model the distribution of extremes such as heavy precipitation. The Extremal-types theorem [*Fisher and Tippett*, 1928; *Gnedenko*, 1943] states that the behavior of the upper tail of the distribution, which governs the probability of extreme events, can be of three distinct types. Two classical strategies for statistical inference are as follows. In the block maxima approach, the generalized Extreme Value (GEV) distribution is fitted to maxima over sufficiently large blocks of observations, often taken as years. On the other hand, the peaks-over-threshold approach consists in approximating the distribution of the excesses over a large enough threshold by the generalized Pareto (GP) distribution [*Balkema and de Haan*, 1974; *Pickands*, 1975]. Both the GEV and the GP distribution have a shape parameter which determines the type of extremal behavior: Fréchet (heavy tail), Gumbel (exponential or light tail), or Weibull (finite bounded tail) type depending on the range of values of the shape parameter (positive, zero or negative, respectively).

Although risk assessment is conventionally performed by hydrologists based on return levels such as the 100 year return level, the probability of extreme events and therefore hazard levels are strongly influenced by the shape parameter. Indeed, return levels of period T years (the amount which is expected to be exceeded on average once in T years) can be expressed as quantiles of the GEV or the GP distribution. In both cases, as T gets large, the shape parameter becomes the determining factor of the value of return

levels. Therefore, to assess hazard levels, considering return levels with a long period is strongly linked to the value of the shape parameter.

Several regression approaches have been developed to interpolate the distribution of extremes in a region. As the estimation of the shape parameter is known to be difficult, it is often assumed to be constant in space, thereby assuming a constant hazard level in the region [*Sang and Gelfand*, 2009; *Renard*, 2011; *Naveau et al*., 2014]. However, the shape parameter has also been modeled as functions of covariates with various degrees of freedom [*Blanchet and Lehning*, 2010; *Carreau and Girard*, 2011]. In *Cooley et al*. [2007], hierarchical Bayesian models of increasing complexity for the shape parameter are compared: a single value for the entire region (constant shape parameter in space), two values for two predefined regions (the selected model) and as functions of covariates with a Gaussian process.

Regional frequency analysis can also be applied as an interpolation approach of the distribution of extremes [*Hosking and Wallis*, 2005; *Carreau et al*., 2013]. It relies on the concept of homogeneous regions that can be defined either contiguously in space or as neighborhoods around target sites [*Burn*, 1990]. Extreme observations (annual maxima or excesses above a high threshold) from all the sites in a given homogeneous region are scaled by a site-specific factor and pooled together to estimate the so-called regional distribution. Both with the block maxima or the peaks-over-threshold strategies, this implies that homogeneous regions have constant shape parameter and thus the corresponding hazard levels can be either piecewise constant (in the case of contiguous regions) or smoothly varying (with the neighborhood approach also called *region of influence*).

In this work, we propose a regional peaks-over-threshold model to interpolate the distribution of extremes, defined as excesses above a high threshold, for regions in which the hazard level is assumed to vary moderately. More precisely, we assume that the hazard level is piecewise constant and that the region can be partitioned into subregions with constant shape parameter. The scale parameter is assumed to vary smoothly spatially as a function of covariates. The regional peaks-over-threshold model can be formalized as a conditional mixture of GP distributions, see section 2.2. We develop in section 2.2.1 a two-step inference and interpolation procedure to estimate the partition and the GP parameters, analogously to the Expectation-Maximization (EM) algorithm [*Dempster et al*., 1977]. The inference is fast since it relies on probability weighted moment estimators [*Diebolt et al*., 2007] and could be used to initialize EM or any other inference strategy involving optimization. The estimation of the GP parameters in a given subregion is a reformulation, with simplified expressions (requiring two instead of three probability weighted moments), of the inference strategy proposed in *Naveau et al*. [2014] (see Appendix A for details).

The number of subregions, i.e., the size of the partition or equivalently the number of components in the mixture, is selected with a cross-validation procedure, see section 2.2.2. This is a standard way to assess out-of-sample performance by creating test sets, i.e., data not used for estimation, which makes an efficient use of the available data [see *Bishop*, 2011, section 1.3]. By comparing several partition sizes with out-of-sample performance, the cross-validation procedure can find a balance between sufficient adaptability of the model obtained with enough subregions and too much variability as a result of too many shape parameters to infer. This is a typical bias-variance trade-off which is often addressed in regional frequency analysis with statistical tests to assess homogeneity [*Viglione et al*., 2007].

The proposed regional peaks-over-threshold model can be cast into regional frequency analysis combined with peaks-over-threshold [*Madsen and Rosbjerg*, 1997a; *Roth et al*., 2012; *Evin et al*., 2016]. However, none of these studies has considered the case of a partition into subregions with constant shape parameter (i.e., contiguous homogeneous regions). In addition, the procedure developed in this work to identify the partition takes an entirely different angle. In particular, the partition is formed based on a probability weighted moment that depends only on the shape parameter rather than from physiographic variables such as geographical and climatological characteristics [*Hosking and Wallis*, 2005]. Besides, although the inference strategy is completely different, the proposed conditional mixture of GPs can also be thought of as an extension of the two subregion model proposed in *Cooley et al*. [2007] in which the number of subregions is not fixed a priori.

The performance of the proposed regional peaks-over-threshold model is evaluated on 18 different synthetic data sets in section 3 in terms of its ability to select the appropriate number of hazard subregions and to estimate the GP parameters when the generative model is known. In section 4, the proposed model

is compared on daily precipitation from the French Mediterranean to two other interpolation approaches with different assumptions on the variability of the shape parameter (either constant on the entire region or smoothly varying). To this end, the performance of each interpolation approach is evaluated via cross validation. In addition, thanks to spatial block bootstrap, 95% confidence bands of return level curves are obtained for each approach at two stations that were kept aside for validation.

2. Regional Peaks-Over-Threshold Approach

2.1. Peaks-Over-Threshold Approach

The peaks-over-threshold approach is based on a theorem which states that, under mild assumptions, the generalized Pareto (GP) distribution can be used as an approximation to the upper tail of the distribution of most random variables [*Pickands*, 1975]. In other words, given a high enough threshold u suitably chosen, the GP distribution approximates the distribution of the excesses over u. Let $Y \sim G(\sigma, \xi)$ be a random variable representing the excesses that follows a GP distribution with scale parameter $\sigma > 0$ and shape parameter $\xi \in \mathbb{R}$. The survival function of Y is provided in equation (1) and obey the following domain restrictions: $y \geq 0$ when $\xi \geq 0$ and $y \in [0, -\sigma/\xi)$ when $\xi < 0$.

$$\mathbb{P}(Y > y) = \bar{G}(y; \sigma, \xi) = 1 - G(y; \sigma, \xi) = \begin{cases} \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ \exp\left(-\frac{y}{\sigma}\right) & \text{if } \xi = 0. \end{cases} \tag{1}$$

The shape parameter describes the upper tail behavior and can be thought of as a way to associate a hazard level to Y. Indeed, the larger the shape parameter is, the higher the probability of extreme events. If $\xi > 0$, Y is said to have a heavy or Pareto-type upper tail and $\mathbb{P}(Y > y) \approx 1/y^{1/\xi}$ for $y \uparrow \infty$. The upper tail is said to be light or exponential when $\xi = 0$ and $\mathbb{P}(Y > y) \approx 1/\exp(-y)$ for $y \uparrow \infty$. The upper tail is bounded for $\xi < 0$.

2.1.1. Return Levels

High quantiles associated to long return periods such as 100 years are often used by practitioners for risk assessment. Let $I(T)$ be the quantile, also termed return level, with a return period of T years, i.e., $I(T)$ is the level that is exceeded on average once every T years. Thanks to the GP tail approximation, $I(T)$ can be estimated as a quantile of the GP distribution as follows:

$$I(T) = \begin{cases} u + \frac{\sigma}{\xi}\left((T N_{exc})^{\xi} - 1\right) & \text{if } \xi \neq 0 \\ u + \sigma \log(T N_{exc}) & \text{if } \xi = 0 \end{cases} \tag{2}$$

provided that $I(T)$ is greater than the threshold u and where $N_{exc} = 365.25\,\zeta_u$ is the average number of excesses per year with ζ_u the probability of exceeding the threshold u.

Underestimation of the shape parameter leads to underestimation of return levels with greater discrepancies for longer return periods. Indeed, it can be seen from equation (2) that the larger T is, the greater the influence of the value of the shape parameter on the return level $I(T)$.

2.1.2. Probability Weighted Moment Estimators

To estimate the GP parameters, we rely on a method based on probability weighted moments (PWM). For $r \geq 0$, the PWMs of $Y \sim G(\sigma, \xi)$ are given by [*Diebolt et al.*, 2007]:

$$\mathbb{E}[Y \bar{G}(Y; \sigma, \xi)^r] = \frac{\sigma}{(1+r)(1+r-\xi)}. \tag{3}$$

Provided that $\xi < 1$, let $\mu = \mathbb{E}[Y] = \sigma/(1-\xi)$ be the first probability weighted moment of Y (plug $r = 0$ in equation (3)). Let us define $Z = Y/\mu$. We have that $Z \sim G(1-\xi, \xi)$ since, by making use of equation (1):

$$\mathbb{P}(Z > z) = \mathbb{P}(Y > \mu z) = \mathbb{P}\left(Y > \frac{\sigma z}{1-\xi}\right) = \bar{G}(1-\xi, \xi). \tag{4}$$

Note that the first PWM of Z is equal to one, i.e., $\mathbb{E}[Z] = 1$. Let ν denote the second PWM of Z:

$$\nu = \frac{1-\xi}{4-2\xi}. \tag{5}$$

We express the GP parameters ξ and σ as functions of the two aforementioned PWMs, ν and μ:

$$\xi=\frac{1-4\nu}{1-2\nu} \quad \text{and} \quad \sigma=\mu(1-\xi). \tag{6}$$

Sample estimates of PWMs can be computed using U-statistics [*Furrer and Naveau*, 2007].

2.2. Conditional Mixture of GPs

We assume that the region of interest can be partitioned into N_{reg} subregions with different hazard levels. In terms of GP distribution, this translates into each subregion having distinct but constant shape parameters. The scale parameter is assumed to vary smoothly as a function of the covariates.

For a given site, let Y be the random variable representing the excesses above a high enough threshold and let $\boldsymbol{x}$ be a vector of covariates associated to the site. In addition, let C be a discrete random variable taking values in $\{1,\ldots,N_{reg}\}$ representing the label of the subregion to which the site belongs. Then Y can be thought of as following a conditional mixture whose distribution is given by:

$$\mathbb{P}(Y\leq y|\boldsymbol{x})=\sum_{j=1}^{N_{reg}}\mathbb{P}(C=j|\boldsymbol{x})\mathbb{P}(Y\leq y|C=j,\boldsymbol{x}), \tag{7}$$

where $\mathbb{P}(C=j|\boldsymbol{x})$ is the probability of belonging to the j^{th} subregion given the covariates $\boldsymbol{x}$ and $\mathbb{P}(Y|C=j,\boldsymbol{x})$ is the conditional distribution of Y given that it belongs to the j^{th} subregion.

Thanks to the GP tail approximation and the assumptions on the GP parameters, we have that:

$$\mathbb{P}(Y\leq y|C=j,\boldsymbol{x})=G(y;\sigma(\boldsymbol{x}),\xi_j), \tag{8}$$

where G is the distribution function given in equation (1) and with the scale parameter $\sigma(\cdot)$ a smooth function of $\boldsymbol{x}$ and ξ_j the shape parameter of the j^{th} subregion.

2.2.1. Inference and Interpolation

We develop a two-step inference strategy adapted to the fact that the subregion label is not observed, i.e., C is a hidden (or latent) variable. Both steps, called E and M steps by analogy with the Expectation-Maximization (EM) algorithm [*Dempster et al.*, 1977], relies on PWM estimators (section 2.1.2). The first step or E-step aims to assign hazard levels to the sites by estimating C, i.e., it aims to estimate the partition of the subregions. In the second step or M-step, $\mathbb{P}(C=j|\boldsymbol{x})$ along with the GP scale parameter $\sigma(\cdot)$ and the shape parameters $\{\xi_1,\ldots,\xi_{N_{reg}}\}$ are estimated.

Let M be the number of gauged sites in the region of interest. For a given gauged site i, let $\{y_{i1},\ldots,y_{in_i}\}$ be the n_i observed excesses and let $\boldsymbol{x}_i$ be the corresponding vector of covariates. In addition, let $C_i\in\{1,\ldots,N_{reg}\}$ be the unobserved subregion label.

2.2.1.1. E-Step: Partitioning into Hazard Subregions

Let $Y_i|C_i,\boldsymbol{x}_i\sim G(\sigma(\boldsymbol{x}_i),\xi_{C_i})$. Similar expressions as in section 2.1.2 can be developed with σ replaced by $\sigma(\boldsymbol{x}_i)$ and ξ by ξ_{C_i}. In particular, given C_i, equation (4) yields:

$$Z_i=\frac{Y_i|C_i,\boldsymbol{x}_i}{\mu(\boldsymbol{x}_i)}\sim G(1-\xi_{C_i},\xi_{C_i}), \tag{9}$$

where the conditional average of the excesses is given by $\mu(\boldsymbol{x}_i)=\sigma(\boldsymbol{x}_i)/(1-\xi_{C_i})$.

To estimate $\mu(\cdot)$, we employ kernel regression, a nonparametric approach [*Nadaraya*, 1964; *Watson*, 1964]. The implementation details are provided in section 2.2.3. Let $\hat{\mu}(\cdot)$ be the kernel regression estimator of $\mu(\cdot)$. The estimated scaled excesses are then given by $\hat{z}_{ik}=y_{ik}/\hat{\mu}(\boldsymbol{x}_i)$ for $1\leq i\leq M$ and $1\leq k\leq n_i$.

For any two sites $1\leq i,j\leq M$, let ν_i and ν_j be the second PWMs of Z_i and Z_j, respectively (see equation (5)). Since ν_i and ν_j only depend on ξ_{C_i} and ξ_{C_j} respectively, we have that

$$\nu_i=\nu_j \iff C_i=C_j. \tag{10}$$

The partitioning into hazard subregions can be thought of as an unsupervised classification problem. Let $\hat{\nu}_i$ be estimators computed using U-statistics [*Furrer and Naveau*, 2007] of the second PWM of Z_i, $1\leq i\leq M$ based on $\{\hat{z}_{i1},\ldots,\hat{z}_{in_i}\}$, the sample of estimated scaled excesses at site i. We resort to K-Means [*Ripley*,

1996] applied to $\hat{\nu}_i$ to obtain a partition of the hazard subregions. Any other statistic of Z_i, such as higher PWMs, could be considered to form the partition.

2.2.1.2. M-Step: Parameter Estimation

The estimation of the conditional probability of belonging to a subregion follows from the partitioning obtained in the *E-Step*. Since K-Means yields *hard* assignment rules, i.e., a site is or is not in a subregion, we have that $\hat{\mathbb{P}}(C=j|\boldsymbol{x}_i)=1_{\{\hat{C}_i==j\}}$. In the conditional mixture setup, the expressions in equation (6) to estimate the GP parameters become:

$$\xi_j=\frac{1-4\nu_j}{1-2\nu_j}\,1\le j\le N_{reg} \tag{11}$$

$$\sigma(\boldsymbol{x}_i)=\mu(\boldsymbol{x}_i)(1-\xi_j) \text{ if } C_i=j. \tag{12}$$

For a given subregion j, the scaled excesses from all the sites in the subregion can be pooled together to estimate ν_j. Let $\hat{C}_i$ be the estimated subregion labels obtained at the *E-Step*, i.e., the cluster labels determined by K-Means. For each $1\le j\le N_{reg}$, ν_j is estimated from the pooled sample $\{\hat{z}_{ik}, 1\le k\le n_i : \hat{C}_i=j\}$, i.e., the set of estimated scaled excesses from all the sites that are assigned to subregion j. The estimated shape parameter $\hat{\xi}_j$ for each subregion is then obtained through equation (11) and the estimated scale parameter $\hat{\sigma}(\boldsymbol{x}_i)$ is provided by equation (12). See Appendix A for details of the related inference strategy developed in *Naveau et al.* [2014].

2.2.1.3. Interpolation to Ungauged Sites

Let i^* be a target site, poorly gauged or ungauged where we wish to estimate the parameters of the GP distribution according to the model of equation (7). The first step (or *E-Step*) is to assign a hazard level to i^*, i.e., to estimate the subregion label $C_{i^*}\in\{1,\ldots,N_{reg}\}$. This is a supervised classification problem for which we use the k-nearest neighbor rule, a nonparametric classifier [*Ripley*, 1996], based on the covariates $\boldsymbol{x}_{i^*}$ and $\boldsymbol{x}_i$, $1\le i\le M$. See section 2.2.3 for the implementation details.

In the *M-Step*, as the k-nearest neighbor rule also yields *hard* assignment rules, $\hat{\mathbb{P}}(C=j|\boldsymbol{x}_{i^*})=1_{\{\hat{C}_{i^*}==j\}}$. The estimated shape parameter at i^* is given by $\hat{\xi}_{\hat{C}_{i^*}}$. Then $\hat{\sigma}(\boldsymbol{x}_{i^*})$ is obtained via equation (12) in which the estimated $\hat{\mu}(\cdot)$ is applied to the covariates $\boldsymbol{x}_{i^*}$.

2.2.2. Selection of the Number of Hazard Subregions

The number of hazard subregions is the number of mixture components and it controls the complexity level of the conditional mixture of equation (7). Indeed, the number of components is equal to the number of free parameters in the mixture which corresponds to the shape parameters of the GP for each subregion. Therefore, increasing the number of subregions provides the conditional mixture with a greater ability to adapt to the variability of the hazard level in the region. On the other hand, added adaptability translates into higher variance of the conditional mixture since more parameters must be estimated from the data. This is particularly an issue in our setup since the shape parameter is notoriously difficult to estimate and is subject to considerable uncertainty [*Coles*, 2001].

We address this bias-variance trade-off (sufficient adaptability while keeping variance under control) by selecting the number of subregions such that the conditional mixture yields the highest performance on out-of-sample data. In practice, we implement cross validation as follows. A small set of sites is held out to assess performance while the remainder of the sites are used for parameter estimation. The procedure is repeated for all possible choices for the held-out set of sites and the performance is computed from all the held-out sets of sites [see *Bishop*, 2011, section 1.3].

The performance is measured in terms of loss functions which provide a measure of how poorly models perform. Thus, the best model is the one yielding the smallest lost. In order to stress the contribution of the shape parameter, the lost functions are computed on the estimated scaled excesses $\hat{z}_{ik}$. Due to the *hard* assignment rules used to form the partition, the fitted model for the scaled variable at a given held-out site i reduces to:

$$\hat{\mathbb{P}}(Z_i\le z|\boldsymbol{x}_i)=\hat{\mathbb{P}}(Y_i\le\hat{\mu}(\boldsymbol{x}_i)z|\boldsymbol{x}_i)=\sum_{j=1}^{N_{reg}}1_{\{\hat{C}_i==j\}}G(z;1-\hat{\xi}_j,\hat{\xi}_j)=G(z;1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i}). \tag{13}$$

We consider three lost functions to evaluate the fitted conditional mixture. The first loss function is the negative log-likelihood of the estimated scaled excesses which is given by:

$$-\sum_{i=1}^{M}\sum_{k=1}^{n_i}\ln g(\hat{z}_{ik};1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i}), \tag{14}$$

where g is the density function of the GP distribution.

The second loss function is the sum-of-squares quantile error:

$$\sum_{i=1}^{M}\sum_{k=1}^{n_i}\left(\hat{z}_{i(k)}-G^{-1}((k-0.5)/n_i;1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i})\right)^2 \tag{15}$$

where $\hat{z}_{i(1)}\leq\ldots\leq\hat{z}_{i(n_i)}$ are the ordered estimated scaled excesses and G^{-1} is the quantile function of the GP distribution computed on empirical frequencies.

The third loss function is the Anderson-Darling statistic that is "sensitive to discrepancies at the tails of the distribution" [*Anderson and Darling*, 1954]:

$$\sum_{i=1}^{M}\left\{-n_i-\frac{1}{n_i}\sum_{k=1}^{n_i}(2k-1)\left[\ln G(\hat{z}_{i(k)};1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i})+\ln\left(1-G(\hat{z}_{i(n_i-k+1)};1-\hat{\xi}_{\hat{C}_i},\hat{\xi}_{\hat{C}_i})\right)\right]\right\}. \tag{16}$$

2.2.3. Implementation Details for Fixed N_{reg}

The algorithm for the regional peaks-over-threshold approach is described in details in Algorithm 1 and is implemented in a package available upon request in the R environment [*R Core Team*, 2016]. Each step of the algorithm is illustrated on synthetic data generated according to a model that satisfies the assumptions of the approach. More precisely, let $x\in[0,1000]$ be a one-dimensional covariate, then:

$$Y|x\sim G(y;\sin(2\pi x/250)+\exp(x/500),\xi_C). \tag{17}$$

For the illustration, we use the sample in Figure 1a with four hazard subregions, i.e., $C\in\{1,2,3,4\}$ such that:

$$\begin{aligned}\xi_1&=0.3 \quad \text{if} \quad x\leq 250\\ \xi_2&=0.2 \quad \text{if} \quad 250<x\leq 500\\ \xi_3&=0.1 \quad \text{if} \quad 500<x\leq 750\\ \xi_4&=0 \quad \text{if} \quad x>750.\end{aligned} \tag{18}$$

The shape parameter values are chosen so as to span approximately the range of estimated values for the precipitation data, see section 4. The scale parameter is taken as a combination of periodic and exponential signal. Therefore, the conditional mean also follows a combination of periodic and exponential signal but with discontinuities at the borders of the subregions since $\mu(x_i)=\sigma(x_i)/(1-\xi_{C_i})$. We can thus assess the ability of kernel regression at estimating a nonlinear function with a small number of discontinuities.

To perform the so-called *E-Step* of the inference scheme in section 2.2.1, the conditional mean $\mu(\cdot)$ must first be estimated (see Figure 1b) and then the estimated scaled excesses can be computed $\hat{z}_{ik}=y_{ik}/\hat{\mu}(\boldsymbol{x}_i)$, $1\leq k\leq n_i$ and $1\leq i\leq M$ (see Figure 2a).

Let $K_h(\cdot)$ be a kernel function that can be thought of as a symmetric density function for which the so-called *bandwidth* h, which controls the amount of smoothing in the regression, acts as a scale parameter. The kernel regression estimator of $\mu(\boldsymbol{x})$ is given by:

$$\hat{\mu}(\boldsymbol{x})=\frac{1}{\sum_i K_h(\boldsymbol{x}-\boldsymbol{x}_i)}\sum_{i=1}^{M}\left(\frac{1}{n_i}\sum_{k=1}^{n_i}y_{ik}\right)K_h(\boldsymbol{x}-\boldsymbol{x}_i). \tag{19}$$

We rely on the `np` package of `R` [*Hayfield and Racine*, 2008] that implements kernel regression with various types of kernels and several automated bandwidth selection methods. We employed the Epanechnikov kernel which is optimal in the sense that it minimizes the asymptotic mean integrated square error [*Epanechnikov*, 1969; *Abadir and Lawford*, 2004]. Bandwidth selection is performed with cross validation [*Li and Racine*, 2004].

For each site $1\leq i\leq M$, v_i, is estimated from $\{\hat{z}_{ik},1\leq k\leq n_i\}$ with U-statistics [*Furrer and Naveau*, 2007] and then smoothed with kernel regression similarly as for $\mu(\cdot)$ to reduce the sampling variability. K-Means

Algorithm 1: Regional peaks-over-threshold model with fixed N_{reg}

input : N_{reg} the number of subregions ;

$\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{in_i}\}$ observed excesses, $\boldsymbol{x}_i$ vector of covariates for each site $1 \leq i \leq M$;

$\boldsymbol{x}_{i^*}$ $1 \leq i^* \leq M^*$ for ungauged sites (optional)

output $\{\hat{\xi}_1, \ldots, \hat{\xi}_{N_{reg}}\}$ shape parameter estimates for each subregion ;

:

$\hat{\sigma}(\boldsymbol{x}_i)$, $\hat{C}_i$ $1 \leq i \leq M$ the scale parameter and the subregion label estimates for each site ;

$\hat{C}_{i^*}$ and $\hat{\sigma}(\boldsymbol{x}_{i^*})$ for $1 \leq i^* \leq M^*$

1 Estimate $\mu(\cdot)$ by regressing $\hat{\mu}_i = 1/n_i \sum_{k=1}^{n_i} y_{ik}$ over $\boldsymbol{x}_i$;

2 Compute the scaled excesses $\hat{z}_{ik} = y_{ik}/\hat{\mu}(\boldsymbol{x}_i)$ for $1 \leq k \leq n_i$;

3 **if** $N_{reg} == 1$ **then** // `single subregion case`

4 Assign all the sites to a single subregion $\hat{C}_i = 1 \; \forall i$ and $\hat{C}_{i^*} = 1 \; \forall i^*$;

5 **else**//`partitioning into` N_{reg} `subregions`

6 Estimate v_i by regressing the second PWM sample estimate of Z_i over $\boldsymbol{x}_i$;

7 Estimate C_i i.e., assign each site i to a subregion j, $1 \leq j \leq N_{reg}$ by clustering $\hat{v}_i$;

8 Estimate C_{i^*} i.e assign each i^* to a subregion j with a classifier based on $\boldsymbol{x}_{i^*}$ and $\boldsymbol{x}_i$;

9 **end**

10 **for** $j \leftarrow 1$ **to** N_{reg} **do**

11 Estimate v_j from all $\hat{z}_{ik}$, $1 \leq k \leq n_i$, such that $\hat{C}_i = j$;

12 Estimate ξ_j, the shape parameter of subregion j as $\hat{\xi}_j = (1 - 4\hat{v}_j)/(1 - 2\hat{v}_j)$;

13 Estimate $\sigma(\boldsymbol{x}_i)$ thanks to $\hat{\sigma}(\boldsymbol{x}_i) = \hat{\mu}(\boldsymbol{x}_i)(1 - \hat{\xi}_{\hat{C}_i})$;

14 Estimate $\sigma(\boldsymbol{x}_{i^*})$ similarly ;

15 **end**

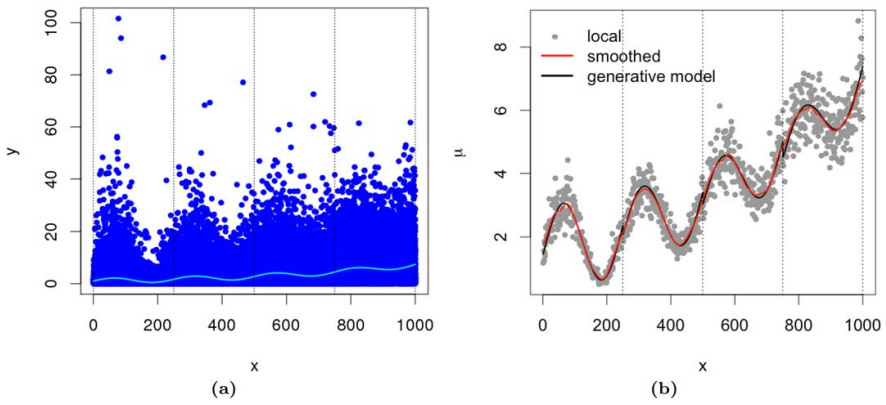


Figure 1. (a) Synthetic data set made of $n_i = 100$ random GP values Y with $x = i$ and $1 \leq i \leq 1000$ from the generative model in equations (17) and (18) whose scale parameter is represented by the cyan curve. (b) At each x, the sample average of the excesses is computed (gray dots) and then smoothed with kernel regression to obtain $\hat{\mu}(x_i)$ (red curve). The black curve represents the conditional mean of the generative model $\mu(x) = \sigma(x)/(1 - \xi_C)$. See line 1 in Algorithm 1. In both figures, the hazard subregions of the generative model are defined by the vertical bands.

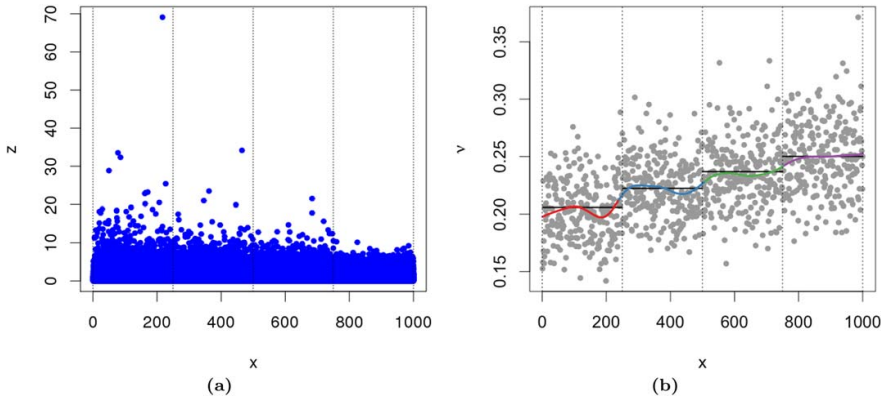


Figure 2. (a) Estimated scaled excesses $\hat{z}_{ik}=y_{ik}/\hat{\mu}(x_i)$, $1 \le k \le n_i$ and $1 \le i \le M$ of the data set in Figure 1a with $\hat{\mu}(x_i)$ as the red curve in Figure 1b. See line 2 in Algorithm 1. (b) The colored curves represent $\hat{v}_i$, the second PWM of Z_i estimate, obtained by smoothing with kernel regression the estimates computed from $\{\hat{z}_{ik}, 1 \le k \le n_i\}$ at each site (gray dots). Hazard subregions, identified by colors, are determined by applying K-Means to $\hat{v}_i$ with $N_{reg}=4$ clusters. The horizontal black lines represent the v values of the generative model. See lines 3–9 in Algorithm 1. In both figures, the hazard subregions of the generative model are defined by the vertical bands.

(implementation of the stats package in R) is then applied to the smoothed $\hat{v}_i$ estimate in order to partition the sites into subregions corresponding to clusters (see Figure 2b where the number of clusters is equal to the number of subregions of the generative model in equation (18)).

To ensure that K-Means always converges to the same partition, we set initial cluster centers to N_{reg} empirical quantiles of $\hat{v}_i$, $1 \le i \le M$, with probabilities that spread regularly the $[0,1]$ interval. The sites are iteratively assigned to the cluster whose center is closer in terms of $\hat{v}_i$ and then the cluster centers are updated as the averages of the $\hat{v}_i$ of the sites belonging to each cluster.

The k-nearest neighbor rule assigns a subregion label to ungauged sites based on the partition established by K-Means. The k nearest neighbors of an ungauged site i^* are determined from the Euclidean distances $d(\boldsymbol{x}_{i^*}, \boldsymbol{x}_i)$ for all $i \in \{1,\ldots,M\}$. The subregion C_{i^*} is the result of a majority vote among the k nearest neighbors. We set the number of neighbor to $k=5$, although it could be optimized just like the bandwidth of kernel regression.

In the so-called *M-Step* in section 2.2.1, v_j is estimated for each subregion $1 \le j \le N_{reg}$ with the pooled sample $\{\hat{z}_{ik}, 1 \le k \le n_i : \hat{C}_i=j\}$. The GP parameters are then estimated thanks to equations (11) and (12). The parameters of the generative model are shown as black curves in Figure 3 while the gray bands represent 95% confidence intervals obtained with parametric bootstrap (1000 replications).

3. Synthetic Data Study

We simulate synthetic data sets that vary in terms of generative models, number of sites, and sample sizes with the aim to evaluate the regional peaks-over-threshold model.

The two generative models considered share the functional form of equation (17) for the scale parameter but differ in terms of partitions into hazard subregions. The first partition has two subregions with very distinct hazard level values:

$$\begin{aligned} \xi_1 &= 0.3 \quad \text{if} \quad x \le 500 \\ \xi_2 &= 0 \quad \text{if} \quad x > 500. \end{aligned} \tag{20}$$

The second partition has four subregions with closer hazard level values as given in equation (18).

We only include differences in partitions in the generative models in order to assess the main contributions of the regional peaks-over-threshold model, namely the *E-Step* to determine the partition and the cross-validation procedure to select the number of subregions. Note that the estimation of the partition

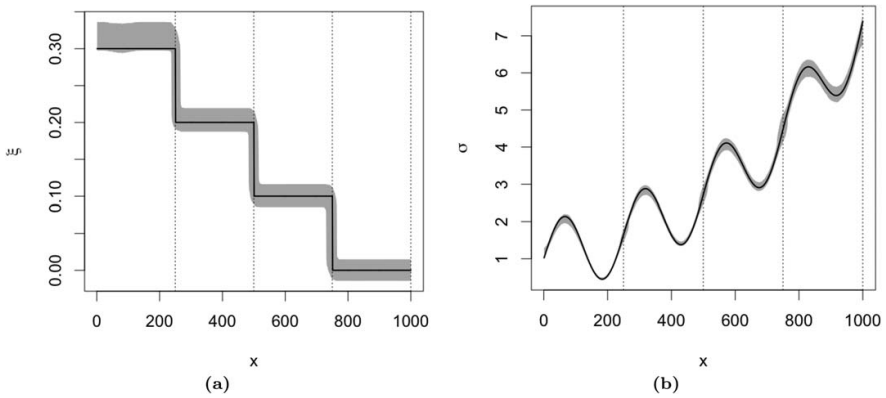


Figure 3. 95% confidence bands in gray obtained with parametric bootstrap (1000 replications) for the GP parameter estimates (*M-Step* in section 2.2.1). The parameters of the generative model are shown as black curves. (a) Shape parameter estimates (equation (11) with v_j estimated from the pooled sample $\{\tilde{z}_{ik}, 1 \leq k \leq n_i : \hat{C}_i{=}j\}$). (b) Scale parameter estimates (equation (12)). See lines 12 to 15 of Algorithm 1. In both figures, the hazard subregions of the generative model are defined by the vertical bands.

influences both the shape and scale parameter estimates (see equations (11) and (12)). On the other hand, as mentioned in section 2.2.3, the functional form for $\sigma(\cdot)$ is reasonably challenging for kernel regression.

There are, in all, 18 different types of synthetic data sets with either two or four subregions, made of M sites for which x is sampled randomly in $[0, 1000]$ with $M = 100$, 200, or 400 and the sample size is either 25, 50, or 100 for each site. Each synthetic data set is replicated 1000 times to assess uncertainty.

3.1. Selection of the Number of Hazard Subregions

We select the number of hazard subregions as the partition size that yields the highest out-of-sample performance, as described in section 2.2.2. We consider partitions with either 1, 2, 4, or 8 hazard subregions (a classical geometric progression). This choice of partition sizes is made to keep computation time tractable.

The cross-validation procedure is implemented as follows. The size of the held-out sets is established so that, whichever synthetic data set, there are 100 held-out sets. More precisely, with $M{=}100, 200, 400$, the held-out sets contain 1, 2, and 4 sites, respectively.

For each of the three loss functions (see equations (14–16)), we computed the percentage of times, out of 1000 replications, each number of subregions is selected. The results for the negative log-likelihood loss function are presented in Figure 4 for each of the 18 types of synthetic data sets. For the sum-of-squares quantile error and the Anderson-Darling statistic, the results are provided in the supporting information.

3.2. Estimation of the GP Parameters

Once the number of hazard subregions is selected via cross validation, the conditional mixture of GPs is estimated anew on the whole data set. The GP parameters are then interpolated to fictitious "ungauged" sites, i.e., on a test set. The test set is made of 1001 sites for which x takes integer values $\{0, 1, \ldots, 1000\}$. Thanks to parametric bootstrap (1000 replicates), 95% confidence bands are computed.

Figures 5 and 6 presents, for the two-region and four-region generative models, respectively, the 95% confidence bands obtained for the smallest data set (100 sites and a sample size of 25) and the largest data set (400 sites and a sample size of 100), when the negative log-likelihood loss function is employed to select the partition size. The results for the other two loss functions are provided in the supporting information.

4. Daily Precipitation Data Application

We focus on the area in the French Mediterranean shown in Figure 7a. The Rhône river valley (in shades of dark green in Figure 7a) runs from north to south and encompasses the cities of Valence and Montpellier.

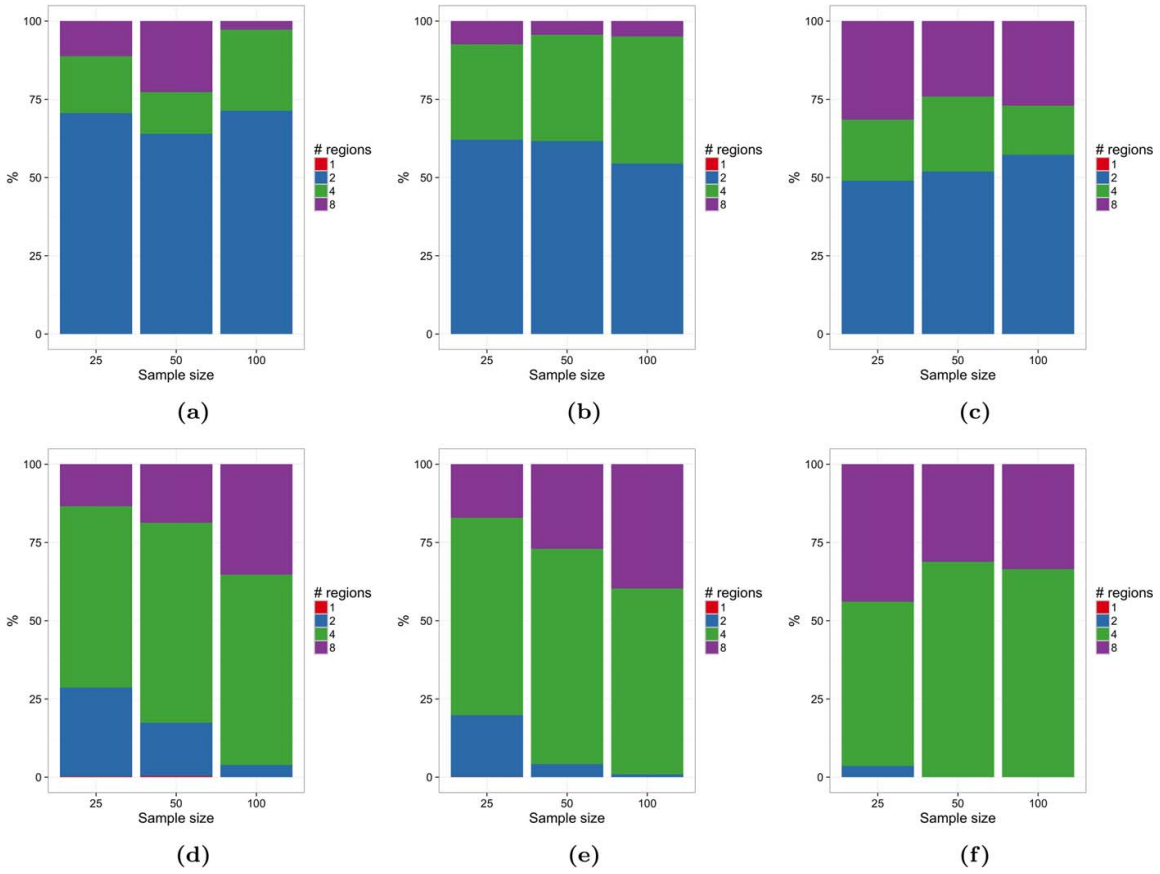


Figure 4. Selection of each partition size (# of subregions) in % over 1000 replications for (a)–(c) the two subregion generative model and (d)–(f) the four subregion generative model. In Figures 4a and 4d, the number of sites is $M = 100$, in Figures 4b and 4e $M = 200$ and in Figures 4c and 4f $M = 400$. The negative log-likelihood loss function is used to assess performance, see section 2.2.2.

On the left bank of the river, sits the prealps (highest point about 2700 m) while on the right bank, sits the Cevennes mountain range (highest point about 1700 m). The latter is well-known for the intense rainfall events called *épisodes cévenols* that occur mainly in autumn [*Delrieu et al*., 2005; *Braud et al*., 2014].

We selected 332 stations of the Météo-France network, the French weather service, depicted in Figure 7b, that belong to the French Mediterranean area shown in Figure 7a. Daily precipitation measurements are collected over the period 1 January 1958 to 31 December 2014 (57 years). In Figure 7b, the size of the plotting symbol is proportional to the length of the observation period available (from 10 to 57 years). The color indicates the percentage of missing values over the observation period (from 0% in light orange to 10% in dark red). Two contour level curves, 400 m and 800 m, of the digital elevation map in dark and light shades of gray recall the orography of the area.

A regular grid (approximately 500 m) is set up to cover the region where the stations lie. Interpolation of the distribution of heavy precipitation is carried out onto the grid. The vector of covariates $\boldsymbol{x}$ is taken as the x and y coordinates (extended Lambert II projections of latitude and longitude).

For each station, the threshold that defines the excesses for the application of the peaks-over-threshold approach (section 2.1) is set to the 98% quantile of the precipitation intensities (the observations greater than 0.1 mm which is the sensitivity of daily rain gauges). This is in line with conventional choices of threshold, see for instance *Roth et al*. [2012]. The resulting overall number of excesses per station ranges from 20

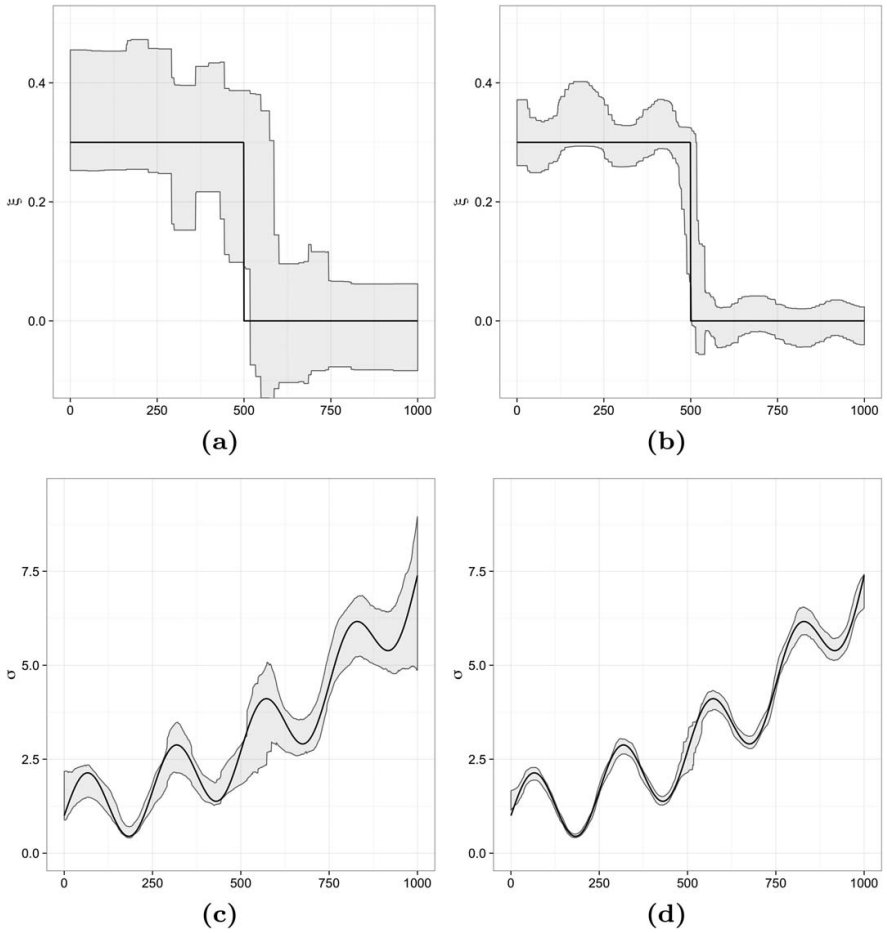


Figure 5. 95% confidence bands (gray) obtained by interpolating on the test set ($x \in \{0, 1, \ldots, 1000\}$) with 1000 replicates (a) and (b) the shape and (c) and (d) the scale parameters of the GP distribution of the two-region generative model (black curves). (a) and (c) Results for the smallest data set (100 sites and a sample size of 25) and (b) and (d) for the largest data set (400 sites and a sample size of 100). The partition size is selected with the negative log-likelihood loss function, see section 2.2.2.

to 191. The threshold and the average number of excesses per year are estimated at each station and then interpolated onto the regular grid with kernel regression (see section 2.2.3 and equation (19)). Figure 8 shows the interpolation results.

4.1. Interpolation Approaches

We compare three approaches to interpolate the distribution of heavy precipitation, i.e., the distribution of the excesses above a high threshold. Each approach differs in the way the shape parameter is assumed to vary in the region.

The first approach considered is the regional peaks-over-threshold model with a single region, i.e., the number of subregions is fixed to one. Therefore, the shape parameter is assumed to be constant (this is equivalent to the *Naveau et al*. [2014] approach, see Appendix A). The shape parameter estimate obtained by pooling the 332 stations of Figure 7b together in equation (11) is $\hat{\xi}_1 = 0.11$.

In the second approach, we let the number of subregions be determined by the data as described in section 2.2.2. For the precipitation data, cross validation is implemented with held-out sets containing four sites

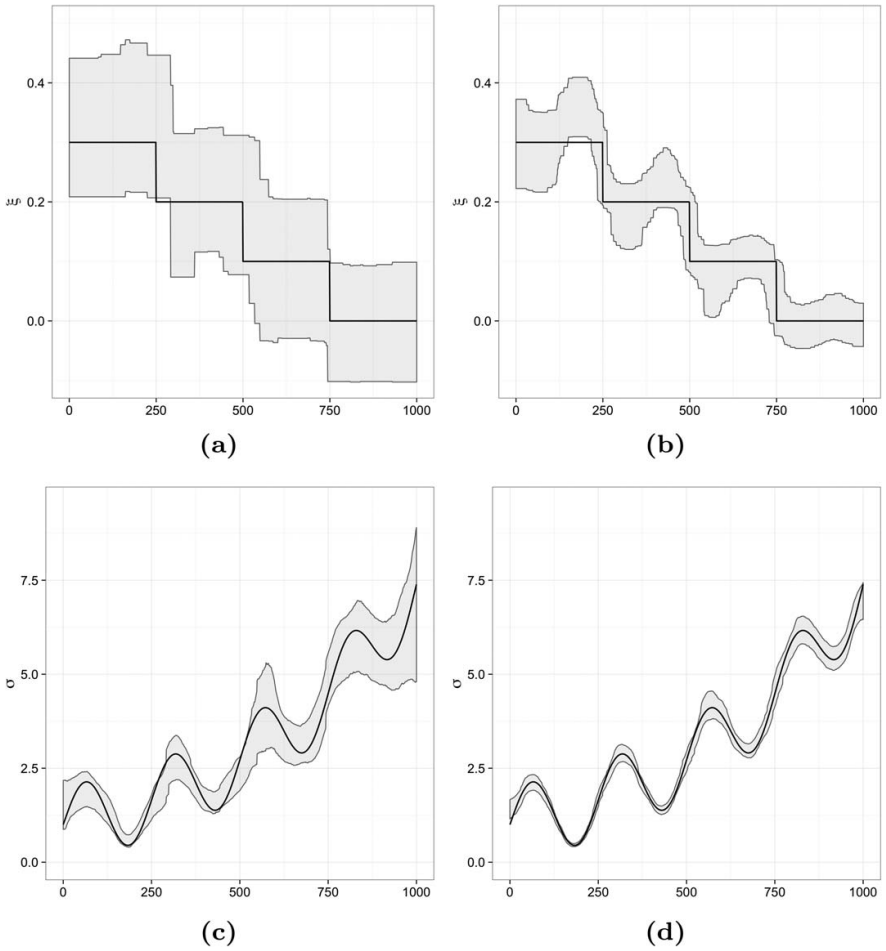


Figure 6. 95% confidence bands (gray) obtained by interpolating on the test set ($x \in \{0, 1, \ldots, 1000\}$) with 1000 replicates (a) and (b) the shape and (c) and (d) the scale parameters of the GP distribution of the four-region generative model (black curves). (a) and (c) Results for the smallest data set (100 sites and a sample size of 25) and (b) and (d) for the largest data set (400 sites and a sample size of 100). The partition size is selected with the negative log-likelihood loss function, see section 2.2.2.

taken at random (there are 83 such sets). A partition of size four yielded the highest performance according to the three loss functions (equations (14–16)). In Figure 9a, we see that the shape parameter estimate varies in a piecewise constant manner across the region. Each subregion has a different hazard level related to the values of the estimated shape parameter, namely $\hat{\xi}_1 = 0.01$, $\hat{\xi}_2 = 0.07$, $\hat{\xi}_3 = 0.12$, and $\hat{\xi}_4 = 0.29$. The difference between the interpolated scale parameter of the regional peaks-over-threshold model with a single region versus the four subregion partition is presented in Figure 10a.

In the third approach, the shape and scale parameters are estimated at each station thanks to equation (6) and then interpolated with kernel regression (see section 2.2.3 and equation (19)). In this approach, the shape parameter can vary almost freely. In Figure 9b, its estimated values range mainly from -0.05 to 0.3 with less than 5% of its values below -0.05 (in dark gray in Figure 9b with a minimum of -0.24). Figure 10b shows the difference between the interpolated scale parameter of this approach and the one of the regional peaks-over-threshold model with four subregions. Less than 1% of the differences are greater than 7 mm in magnitude (in dark gray in Figure 10b).

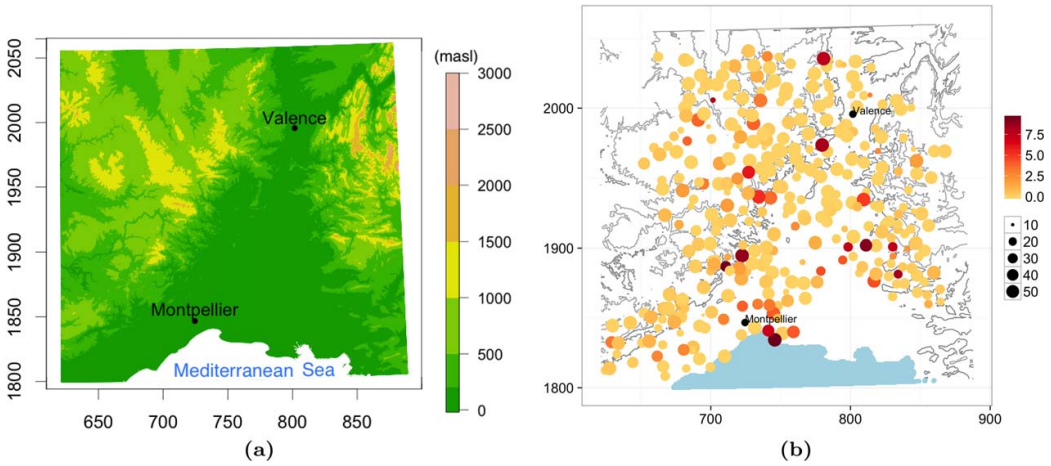


Figure 7. (a) Digital elevation map of the area of interest in the French Mediterranean. (b) 332 rain gauge stations of the Météo-France network (French weather service) covering the period 1 January 1958 to 31 December 2014 (57 years). The size of the symbol is proportional to the length of the observation period (10–57 years) and the color shade (light orange to dark red) indicates the percentage of missing values (0–10%).

4.2. Out-of-Sample Performance Comparison

Cross validation, with held-out sets of size four taken at random, is employed to evaluate the out-of-sample performance of each of the three interpolation approaches described previously. Note that cross validation was also applied to determine the number of hazard subregions of the second interpolation approach. This can be thought of as a form of double layer of cross validation since thanks to the randomization of the sites, the held-out sets are different in each layer [*Arlot and Celisse*, 2010].

With the GP parameter interpolated to the held-out sets, we computed the three loss functions from equations (14–16) on the excesses y_{ik} in order to account for the estimation of *both* the shape and the scale parameters. In addition, the loss functions are computed relatively to the regional peaks-over-threshold model with four hazard subregions. For instance, the negative log-likelihood loss function becomes:

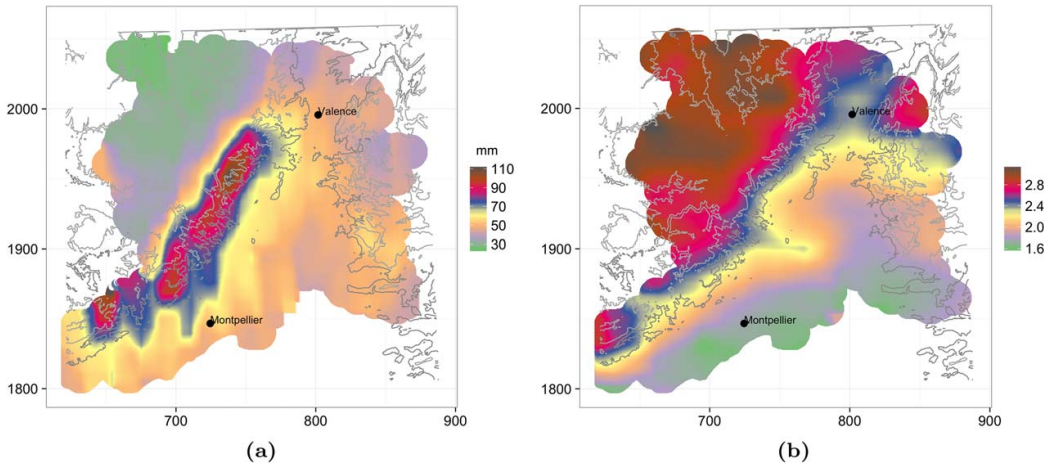


Figure 8. French Mediterranean precipitation data: (a) Threshold defined as the 98% quantile of precipitation intensities (b) Average number of excesses above the threshold per year. Interpolation onto the grid is performed with kernel regression.

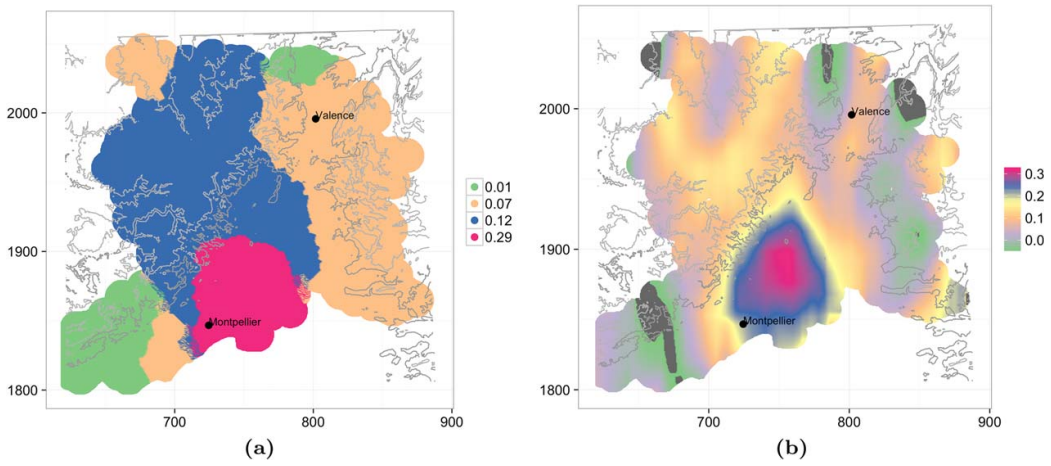


Figure 9. Interpolated shape parameter of the GP distribution onto the grid with (a) the regional peaks-over-threshold model with four subregions and (b) kernel regression applied to at-site shape parameter estimates. Less than 5% of the values are below −0.05 and are shown in dark gray. The shape parameter estimate of the regional peaks-over-threshold model with a single region is $\hat{\xi}_1 = 0.11$.

$$-\sum_{i=1}^{M}\sum_{k=1}^{n_i} \ln g(y_{ik}; \hat{\sigma}_i^O, \hat{\xi}_{\hat{C}_i}^O) + \sum_{i=1}^{M}\sum_{k=1}^{n_i} \ln g(y_{ik}; \hat{\sigma}_i^R, \hat{\xi}_{\hat{C}_i}^R), \tag{21}$$

where g is the density function of the GP distribution, $\hat{\sigma}_i^R$ and $\hat{\xi}_{\hat{C}_i}^R$ are the parameter estimates of the regional model with four hazard subregions while $\hat{\sigma}_i^O$ and $\hat{\xi}_{\hat{C}_i}^O$ are the estimates of one of the other interpolation approaches. Positive values of the relative log-likelihood from equation (21) indicate that the regional model with four hazard subregions outperforms the other interpolation approach. The two other relative loss functions based on the sum-of-squares quantile error in equation (15) and on the Anderson-Darling statistic in equation (16) are built similarly and can be interpreted in the same way.

Figure 11 shows the relative negative log-likelihood loss from equation (21) at each station for the regional model with a single region and the kernel regression interpolation of at-site estimates. At one station

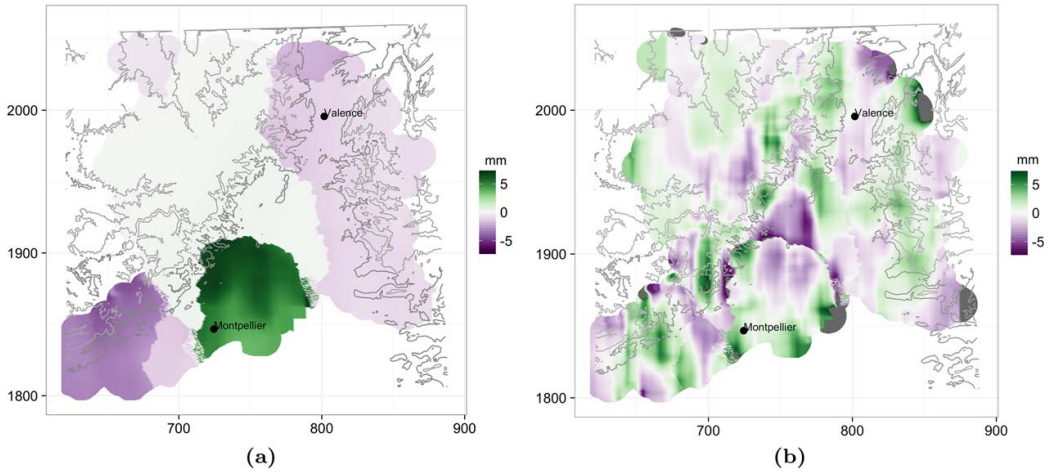


Figure 10. Differences in interpolated scale parameter of the GP distribution onto the grid: (a) The regional peaks-over-threshold model with a single region or (b) The kernel regression interpolation of the at-site estimates minus the interpolated scale parameter of the regional model with four subregions. In Figure 10b, less than 1% of the differences are greater than 7 mm in magnitude and are shown in dark gray.

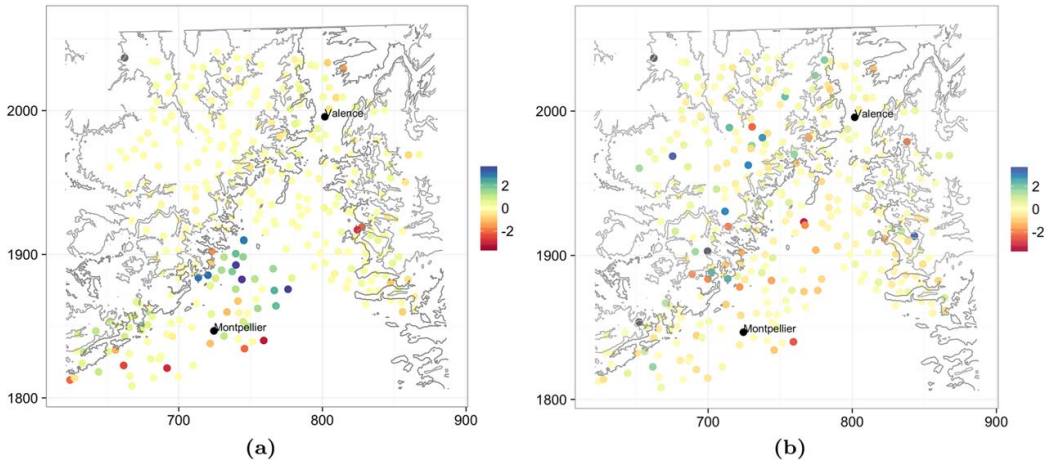


Figure 11. Negative log-likelihood relative to the regional model with four subregions. (a) For the regional model with a single region and (b) for the kernel regression interpolation of the at-site estimates. Positive values (blue shades) indicate that the regional model with four subregions outperforms the other interpolation approach in terms of log-likelihood.

(shown in gray in the north-west of both figures), kernel regression interpolation failed (for μ in the regional peaks-over-threshold model and for σ and ξ in the interpolation of the at-site estimates). In Figure 11b, two stations have values beyond the color scale (–4.5 and 6.9) (also shown in gray). The figures for the other relative loss functions have similar spatial patterns and are provided in the supporting information. The spatial average with standard errors in parentheses of the three relative loss functions is provided in Table 1.

Last, we performed the inference on 330 stations, i.e., with two stations (the closest ones to the cities of Montpellier and Valence, see Figure 7b) kept aside for validation. From the interpolated GP parameters from each of the three interpolation approaches, we computed return level curves (see equation (2)). To obtain the 95% confidence bands shown in Figure 12, we implemented nonparametric spatial block bootstrap as follows. Blocks of 3 days are randomly sampled from the original observations for all the stations simultaneously to preserve both the temporal and spatial dependence present in the observations. The size of the block was determined from the maximum number of consecutive excesses in the precipitation data. Other propositions to take into account the spatial dependence in the uncertainty estimation can be found in *Madsen and Rosbjerg* [1997b]; *Madsen et al.* [2002]; and *Van de Vyver* [2012].

5. Discussion

5.1. Synthetic Data Study

The synthetic data study in section 3 serves two purposes. First it allows to evaluate the procedure to select the number of hazard subregions from section 2.2.2 on data whose generative model is known.

For both the two subregion and the four subregion generative models (see equation (20) and equation (18), respectively), the number of subregions of the generative model is selected approximately in 50% of the bootstrap replicates when the negative log-likelihood loss function is used to assess performance see Figure 4. The picture is more mixed for the other two loss functions, see the supporting information. This percentage does not seem to follow a pattern, in particular it does not clearly increase with the number of sites or the sample

Table 1. Average (Standard Error) Loss Functions Relative to the Regional Model With Four Subregions[a]

	Neg Log-Like	Sum-of-Squares	Anderson-Darling
Single region	**0.1160 (0.04148)**	**784.5 (336.8)**	**0.06201 (0.02805)**
Kernel regression	0.03704 (0.05053)	**641.7 (286.6)**	0.009116 (0.04415)

[a]Positive values means that the regional model with four subregions outperforms the other interpolation approach, results in boldfaced are approximately significant at 95%.

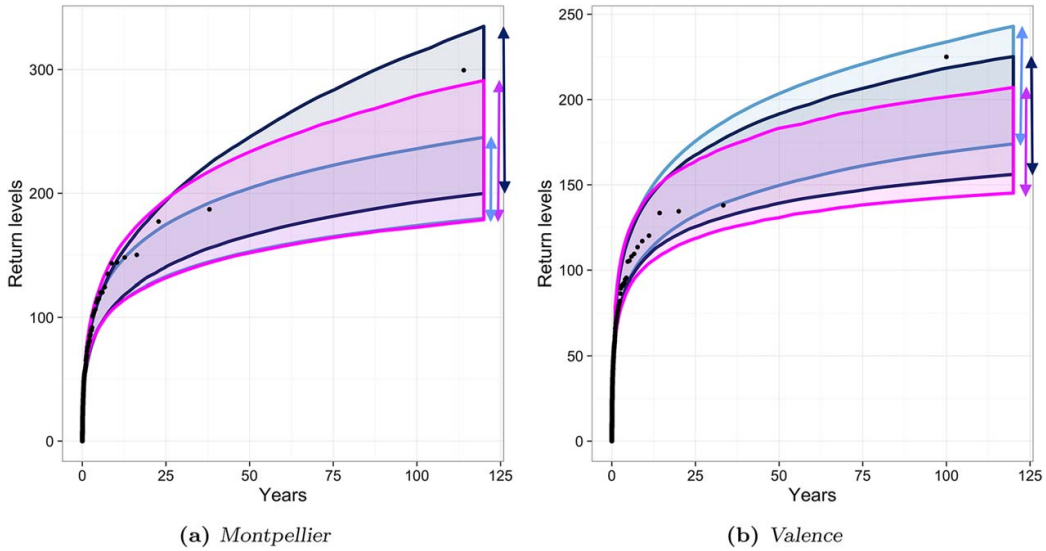


Figure 12. Return level curves from the regional peaks-over-threshold model with a single region (light blue), with four hazard subregions (dark blue) and the kernel regression interpolation of at-site estimates (magenta). The points represent the empirical return levels.

size as we might expect. There is a tendency to select more rather than less subregions than present in the generative model, especially as the number of sites and the sample size get larger. Indeed, the partition with a single region (in red in the figures) is rarely selected.

The focus of the procedure is only on the selection of the number of subregions, not on their identification. As the partitioning or *E-Step* is fairly stable, i.e., a partition of a given size will always define very comparable subregions, the selection of the number of subregions also determines their shape. On the other hand, the partition is built so that within each subregion, the hazard level of the sites is as similar as possible. Thus, if within some subregions the hazard level is too variable, the only possibility to improve the fit of the model is to increase the number of subregions, not to change their shape. In contrast, in regional frequency analysis, subregions are conventionally identified as geographically coherent areas with similar climatic and physical features. Such subregions are often found to be heterogeneous in terms of extremal behavior. Statistical tests are used to assess homogeneity and the subregions are reshaped until they pass the tests [*Madsen and Rosbjerg*, 1997b; *Castellarin et al.*, 2001].

The second purpose of the synthetic data study is to evaluate the inference and interpolation strategy of the regional peaks-over-threshold model, see section 2.2.1. The 95% confidence bands for the shape and scale parameter become narrower as the number of sites and the sample size increases, see Figures 5 and 6 and the supporting information. The width of the 95% confidence bands of the scale parameter is comparable for both generative models. In contrast, with the four subregion generative model, the confidence bands of the shape parameter tend to overlap more between subregions than with the two subregion generative model.

5.2. Daily Precipitation Data Application

The threshold that defines the excesses (the 98% quantile of the precipitation intensities) takes its larger values, about 100 mm, along the Cevennes mountain range, see Figure 8a. As the Cevennes acts as a barrier, the threshold values drop quickly behind it. The spatial pattern of the average number of excesses per year is noticeably different, see Figure 8b. In section 2.1.1, it is defined as $N_{exc}=365.25\,\zeta_u=365.25\,\mathbb{P}(P>u|\boldsymbol{x})=365.25\,\mathbb{P}(P>u|P>0,\boldsymbol{x})\,\mathbb{P}(P>0|\boldsymbol{x})$, where P is the random variable representing daily precipitation (not the excesses, that are represented by Y). By construction, $\mathbb{P}(P>u|P>0,\boldsymbol{x})=0.02$. Therefore, the north-south pattern of N_{exc} is driven by $\mathbb{P}(P>0|\boldsymbol{x})$, the probability of occurrence of positive precipitation intensities.

The shape parameter estimate of the regional peaks-over-threshold model with a single region $\hat{\xi}_1 = 0.11$ is very close to the average of the estimates of the four subregions model which is of 0.12, see Figure 9a. Similarly, the shape parameter estimates of the model with four subregions can be seen as approximately the average of the estimates obtained with kernel regression within each subregion, see Figure 9. Most of the values (84%) in Figure 9b are in the $[0, 0.3]$ interval. Thus, the range of values of the shape parameter estimates is globally the same for the four subregions model and the interpolation with kernel regression. Also, in both models, the higher hazard subregion is located in the south (in pink) with shape parameter estimates close to 0.3. The region starts at the coast, goes up to the foothills of the Cevennes mountain range and is consistent with expert knowledge [*Delrieu et al*., 2005; *Braud et al*., 2014].

The borders of the subregions are relatively smooth in Figure 9a except in a few places such as near the southern east and west borders of the high hazard subregion. The amount of smoothness is controlled by k, the number of neighbors in the k-nearest neighbor rule, which was set to 5. This choice might have an impact on the cross-validation results (held-out sites might be classified into different hazard subregions with other values of k). We leave the selection of the number of neighbors for further studies.

The partition obtained with the regional peaks-over-threshold model with data-driven number of subregions has two subregions with relatively close shape parameter estimates in Figure 9a, $\hat{\xi}_1 = 0.01$ and $\hat{\xi}_2 = 0.07$ in green and orange, respectively. It seems likely that three instead of four hazard subregions would be sufficient. However, we considered only partitions of size 1, 2, 4, and 8. To assess the validity of a partition of size 3, we would have to tune the procedure of selection of the number of subregions to refine the search over more precise partition sizes. Given the results in the synthetic data study, it seems, on one hand, that determining a very precise partition size is not an easy task but on the other hand, a slightly higher number of subregions should not harm much the GP parameter estimates.

The differences between the scale parameter estimates from the three interpolation approaches in Figure 10 are generally small compared to the threshold values in Figure 8a. Indeed, 93% of the differences between the single region and the four subregion model are less than 5 mm in magnitude (see Figure 10a). For the kernel regression interpolation of at-site estimates, this percentage goes up to 97%, see Figure 10b. This is consistent with the findings in the synthetic data study that the scale parameter estimates is not very sensitive to the shape parameter estimate.

The spatial pattern of the cross-validation results for the negative log-likelihood loss function in Figure 11a shows that the single region model is outperformed by the four subregion model in the higher hazard area (shown in pink in Figure 9a). Similar patterns can be seen for the other two loss functions in the supporting information. In contrast, there is no clear spatial pattern for the performance of the kernel regression interpolation of the at-site estimates relative to the four subregion model, see Figure 11b. These conclusions are supported by Figure 12. In Figure 12a, for the city of Montpellier which sits in the higher hazard subregion, the 95% confidence bands of the return level curve of the single region model underestimate the empirical return levels. In Figure 12b, no similar underestimation occurs at the city of Valence which belongs to a low hazard region with shape parameter about 0.07, see Figure 9a. The 95% confidence bands of the return level curve of the kernel regression interpolation are somewhat lower than the empirical return levels and in particular, the larger empirical return level is outside the confidence bands in both cases. The spatial averages of the relative loss functions (with standard errors in parentheses) in Table 1 are all positive and most of them are significant at 95%, indicating that the four subregion model performed better relatively to the other two interpolation approaches.

6. Conclusions

In some regions such as the French Mediterranean, the distribution of heavy precipitation is known to follow specific spatial patterns. Since the shape parameter of the GEV or the GP distribution governs the behavior of the upper tail of the distribution where the extremes lie, it can be related directly to the hazard level. Therefore, when interpolating the distribution of extremes in such a region, we would be tempted to take into account the variability of the hazard level and to let the shape parameter vary in space. However, the shape parameter is difficult to estimate due to the high uncertainty of its estimators and for this reason, it is often assumed to be constant. In this work, we chose a middle ground and assume a moderately varying hazard level. More precisely, it is allowed to vary in a piecewise constant fashion.

The main contributions of this paper are the following. We formulated the regional peaks-over-threshold model as a conditional mixture model of GP distributions. Each component has a smoothly varying scale parameter together with a constant shape parameter and can be thought of as affecting a subregion with constant hazard level. We developed a two-step inference and interpolation strategy that is similar in spirits to the Expectation-Maximization algorithm. In the first step, the partition into subregions with constant hazard level is estimated i.e., each site is assigned to a mixture component. The second step consists in estimating the GP parameter for each subregion while the probability of belonging to each subregion follows from the partitioning of the first step. The two-step strategy relies on two probability weighted moments of the GP. The computations are fast and could be used to initialize maximum likelihood estimators. We proposed a classical cross-validation procedure with three different loss functions to assess performance in order to select the number of subregions.

Thanks to 18 synthetic data sets, we assess how the performance of the regional peaks-over-threshold model is affected by the number of subregions, the variability of the hazard level, the number of sites, and the sample size at each site. To this end, we designed two generative models with different types of partitions: one with two subregions with very different hazard levels and another one with four subregions and more similar hazard levels. We were able to evaluate the following two aspects: the ability of the cross-validation procedure to retrieve the number of subregions of the generative model and the performance of the GP estimators when the number of subregions is selected with cross validation and therefore prone to error.

The synthetic data study indicates that the proposed procedure to select the number of subregions is only partially successful at identifying the number of subregions of the generative model and tend to overestimate it, in particular for larger data sets. Note that conventional homogeneity test in regional frequency analysis also suffer from a lack of power [*Viglione et al*., 2007]. On the other hand, although the selected number of subregions does not tend to the generative model's, the accuracy of the GP parameter estimates is improved with larger data sets. As expected, hazard levels are more clearly determined (i.e., the confidence intervals of the shape parameter estimates overlap less) for synthetic data from the generative model with the two very distinct hazard subregions than with the four subregions with closer hazard levels.

In the French Mediterranean daily precipitation data application, the cross-validation procedure selected four subregions for the regional peaks-over-threshold model according to all three loss functions. We conducted a comparison with two other interpolation approaches. The first approach assumes that the shape parameter is constant in the region while the second approach let it varies smoothly as a function of covariates. The comparison of the out-of-sample performance of the three interpolation approaches leads us to conclude that the assumption of a constant shape parameter is not appropriate for this region. On the other hand, the gain in performance of the four subregion model relative to the smooth interpolation of the at-site estimates is not so clear. As shown in *Carreau et al*. [2013] for the block maxima approach, it is likely that the gain in performance of the regional model would be greater in applications in which the spatial variability of the variable of interest is high compared to the spatial density of the sites such as in sparsely gauged networks, in more arid climates or when considering subdaily precipitation. Potential advantages of the regional model are the partitioning into hazard subregions which could be useful for decision makers and shape parameter estimates that are less likely to take negative values thanks to the pooled sample.

One interesting perspective for this work is to exploit the formulation of the regional peaks-over-threshold model as a mixture of GPs. First, by making the assignment rules to make the partition *soft*, i.e., probabilistic, smooth transitions between subregions could be obtained. More precisely, the probability of belonging to a subregion would vary smoothly, being closer to 1 in the center and gradually decreasing when getting near the borders. Second, as mentioned previously, the mixture parameters could be estimated with an EM algorithm or by maximizing directly the log-likelihood starting from the parameter values provided by the two-step inference strategy developed in this work. Another perspective is to improve the procedure to select the number of subregions. In particular, we could work out a penalty term in order to limit the number of subregions when the differences in shape parameter estimates are low relative to their uncertainty. Last, the regional peaks-over-threshold model could be applied to a sparsely gauged network to better assess the gain in performance of the regional peaks-over-threshold model compared to a direct interpolation of the at-site GP parameter estimates. Most likely, x and y coordinates would not be informative enough and, in addition to altitude, covariates related to the orography could be of interest [*Benichou and*

Le Breton, 1987]. Besides, nonparametric methods, such as the k-nearest neighbor rule and kernel regression, might not perform so well in sparser networks and it might be more appropriate to seek parsimonious parametric models.

Appendix A: Alternative Inference Strategy for a Given Subregion

We present the inference strategy developed in *Naveau et al*. [2014] to estimate the GP parameters in a region with constant shape parameter and smoothly varying scale parameter. It is related to the so-called *M-Step* of the inference strategy in section 2.2.1.

It relies on a different scaled variable $Z \sim G(1, \xi)$. Let α denotes the ratio of the second and third PWMs of Z (replace $r = 1$, 2 and $\sigma(\boldsymbol{x})=1$ in equation (3)), i.e.,

$$\alpha = \frac{\mathbb{E}[Z\bar{G}(Z;1,\xi)]}{\mathbb{E}[Z\bar{G}(Z;1,\xi)^2]}. \tag{A1}$$

As before, let $\mathbb{E}[Y|\boldsymbol{x}]=\mu(\boldsymbol{x})=\sigma(\boldsymbol{x})/(1-\xi)$ be the first PWM of Y given $\boldsymbol{x}$. *Naveau et al*. [2014] expressed ξ and $\sigma(\boldsymbol{x})$ as functions of the aforementioned three PWMs:

$$\xi = \frac{9-4\alpha}{3-2\alpha} \quad \text{and} \quad \sigma(\boldsymbol{x})=\mu(\boldsymbol{x})(1-\xi). \tag{A2}$$

To infer ξ and $\sigma(\boldsymbol{x})$, estimates of α and $\mu(\boldsymbol{x})$ are required. Since $Z=Y(\boldsymbol{x})/\sigma(\boldsymbol{x})$ is not observable, *Naveau et al*. [2014] circumvent this issue by relying on the probability weighted moment of $Y(\boldsymbol{x})/\mu(\boldsymbol{x})$ to estimate α. The latter remains unchanged since it is a ratio and $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ differs by the factor $1-\xi$.

Acknowledgments
This work was partially supported by the Agence Nationale de la Recherche (French Research Agency) through its Blanc program with the projects Floodscale and DADA and through its JCJC program with the project Starmip. Part of this was has also been supported by LEFE-INSU-Multirisk, LEFE-INSU-Cerise, AMERISKA, and A2C2 projects. We are grateful to Météo-France for the precipitation data available at `https://publitheque.meteo.fr` and to Juliette Blanchet (IGE) for the preprocessing.

References

Abadir, K. M., and S. Lawford (2004), Optimal asymmetric kernels, *Econ. Lett.*, *83*(1), 61–68.
Anderson, T. W., and D. A. Darling (1954), A test of goodness of fit, *J. Am. Stat. Assoc.*, *49*(268), 765–769.
Arlot, S., and A. Celisse (2010), A survey of cross-validation procedures for model selection, *Stat. Surv.*, *4*, 40–79.
Balkema, A. A., and L. de Haan (1974), Residual life time at great age, *Ann. Probab.*, *2*(5), 792–804.
Bénichou, P., and O. Le Breton (1987), Prise en compte de la topographie pour la cartographie des champs pluviométriques statistiques, La Météorologie, 7e série, no. 19.
Bishop, C. M. (2011), *Pattern Recognition and Machine Learning, Information Science and Statistics*, Springer, New York.
Blanchet, J., and M. Lehning (2010), Mapping snow depth return levels: Smooth spatial modeling versus station interpolation, *Hydrol. Earth Syst. Sci.*, *14*(12), 2527–2544.
Borga, M., E. Anagnostou, G. Blöschl, and J.-D. Creutin (2011), Flash flood forecasting, warning and risk management: The HYDRATE project, *Environ. Sci. Policy*, *14*(7), 834–844, doi:10.1016/j.envsci.2011.05.017, adapting to Climate Change: Reducing Water-related Risks in Europe.
Braud, I et al. (2014), Multi-scale hydrometeorological observation and modelling for flash flood understanding, *Hydrol. Earth Syst. Sci.*, *18*(9), 3733–3761, doi:10.5194/hess-18-3733-2014.
Burn, D. (1990), Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resour. Res.*, *26*(10), 2257–2265.
Carreau, J., and S. Girard (2011), Spatial extreme quantile estimation using a weighted log-likelihood approach, *J. Soc. Française Stat.*, *152*(3), 66–82.
Carreau, J., L. Neppel, P. Arnaud, and P. Cantet (2013), Extreme rainfall analysis at ungauged sites in the South of France: Comparison of three approaches, *J. Soc. Française Stat.*, *154*(2), 119–138.
Castellarin, A., D. Burn, and A. Brath (2001), Assessing the effectiveness of hydrological similarity measures for flood frequency analysis, *J. Hydrol.*, *241*(3), 270–285.
Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, *Springer Series in Statistics*, Springer, London, U. K.
Cooley, D., D. Nychka, and P. Naveau (2007), Bayesian spatial modeling of extreme precipitation return levels, *J. Am. Stat. Assoc.*, *102*(479), 824–840.
Delrieu, G., et al. (2005), The catastrophic flash-flood event of 8–9 September 2002 in the Gard region, France: A first case study for the Cévennes-Vivarais Mediterranean Hydrometeorological Observatory, *J. Hydrometeorol.*, *6*(1), 34–52.
Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Stat. Soc. Ser. B*, *39*(1), 1–38.
Diebolt, J., A. Guillou, and I. Rached (2007), Approximation of the distribution of excesses through a generalized probability-weighted moments method, *J. Stat. Plann. Inference*, *137*(3), 841–857.
Epanechnikov, V. A. (1969), Non-parametric estimation of a multivariate probability density, *Theor. Probab. Appl.*, *14*(1), 153–158.
Evin, G., J. Blanchet, E. Paquet, F. Garavaglia, and D. Penot (2016), A regional model for extreme rainfall based on weather patterns subsampling, *J. Hydrol.*, *541*, 1185–1198.
Fisher, R., and L. H. C. Tippett (1928), Limiting forms of the frequency distribution of the largest and smallest member of a sample, in Cambridge Philosophical Society, vol. 24, pp. 180–190, Cambridge Univ. Press, Cambridge, U. K.
Furrer, R., and P. Naveau (2007), Probability weighted moments properties for small samples, *Stat. Probab. Lett.*, *77*(2), 190–195.
Gnedenko, B. (1943), Sur la distribution limite du terme maximum d'une serie aleatoire, *Ann. Math.*, *44*, 423–453.
Hayfield, T., and J. Racine (2008), Nonparametric econometrics: The np package, *J. Stat. Software*, *27*(5), 1–32.

Hosking, J. R. M., and J. R. Wallis (2005), *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge Univ. Press, Cambridge, U. K.
Li, Q., and J. Racine (2004), Cross-validated local linear nonparametric regression, *Stat. Sin.*, *14*(2), 485–512.
Madsen, H., and D. Rosbjerg (1997a), The partial duration series method in regional index-flood modeling, *Water Resour. Res.*, *33*(4), 737–746.
Madsen, H., and D. Rosbjerg (1997b), Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling, *Water Resour. Res.*, *33*(4), 771–781.
Madsen, H., P. S. Mikkelsen, D. Rosbjerg, and P. Harremoës (2002), Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series statistics, *Water Resour. Res.*, *38*(11), 1239, *doi*:10.1029/2001WR001125.
Nadaraya, E. A. (1964), On estimating regression, *Theor. Probab. Appl.*, *9*(1), 141–142.
Naveau, P., A. Toreti, I. Smith, and E. Xoplaki (2014), A fast nonparametric spatiotemporal regression scheme for generalized Pareto distributed heavy precipitation, *Water Resour. Res.*, *50*, 4011–4017, doi:10.1002/2014WR015431.
Pickands, J. (1975), Statistical inference using extreme order statistics, *Ann. Stat.*, *3*, 119–131.
R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna, Austria.
Renard, B. (2011), A Bayesian hierarchical approach to regional frequency analysis, *Water Resour. Res.*, *47*, W11513, doi:10.1029/2010WR010089.
Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge Univ. Press.
Roth, M., T. Buishand, G. Jongbloed, A. Klein Tank, and J. van Zanten (2012), A regional peaks-over-threshold model in a nonstationary climate, *Water Resour. Res.*, *48*, W11533, doi:10.1029/2012WR012214.
Sang, H., and A. E. Gelfand (2009), Hierarchical modeling for extreme values observed over space and time, *Environ. Ecol. Stat.*, *16*(3), 407–426.
Van de Vyver, H. (2012), Spatial regression models for extreme precipitation in Belgium, *Water Resour. Res.*, *48*, W09549, doi:10.1029/2011WR011707.
Viglione, A., F. Laio, and P. Claps (2007), A comparison of homogeneity tests for regional frequency analysis, *Water Resour. Res.*, *43*, W03428, doi:10.1029/2006WR005095.
Watson, G. S. (1964), Smooth regression analysis, *Sankhyā Ser. A*, *26*, 359–372.

A.3.2 Étude comparative des choix de modèles de densité multivariée

Stoch Environ Res Risk Assess
DOI 10.1007/s00477-015-1166-6

ORIGINAL PAPER

Multivariate density model comparison for multi-site flood-risk rainfall in the French Mediterranean area

Julie Carreau[1] · **Christophe Bouvier**[1]

Abstract The French Mediterranean area is subject to intense rainfall events which might cause flash floods, the main natural hazard in the area. Flood-risk rainfall is defined as rainfall with a high spatial average and encompasses rainfall which might lead to flash floods. We aim to compare eight multivariate density models for multi-site flood-risk rainfall. In particular, an accurate characterization of the spatial variability of flood-risk rainfall is crucial to help understand flash flood processes. Daily data from eight rain gauge stations at the Gardon at Anduze, a small Mediterranean catchment, are used in this work. Each multivariate density model is made of a combination of a marginal model and a dependence structure. Two marginal models are considered: the Gamma distribution (parametric) and the Log-Normal mixture (non-parametric). Four dependence structures are included in the comparison: Gaussian, Student t, Skew Normal and Skew t in increasing order of complexity. They possess a representative set of theoretical properties (symmetry/asymmetry and asymptotic dependence/independence). The multivariate models are compared in terms of three types of criteria: (1) separate evaluation of the goodness-of-fit of the margins and of the dependence structures, (2) model selection with a leave-one-out evaluation of the Anderson-Darling and Cramer-Von Mises statistics and (3) comparison in terms of two hydrologically interpretable quantities (return periods of the spatial average and conditional probabilities of exceedances). The key outcome of the comparison is that the Skew Normal with the Log-Normal mixture margins outperform significantly the other models. The asymmetry introduced by the Skew Normal is an added-value with respect to the Gaussian. Therefore, the Gaussian dependence structure, although widely used in the literature, is not recommended for the data in this study. In contrast, the asymptotically dependent models did not provide a significant improvement over the asymptotically independent ones.

Keywords Intense rainfall events · Strong spatial variability · Small Mediterranean catchments · Elliptical and skew multivariate distributions · Asymptotic dependence/independence

1 Introduction

The French Mediterranean area is subject to intense rainfall events occurring mainly in the fall. They can be triggered by a combination of three factors: the moisture generated by the Mediterranean Sea, upper-level cold troughs coming from the North and the complex orography in the region (the Alps, the Pyrenees and the Massif Central Mountains in the South of France) (Delrieu et al. 2005). Such heavy rainfall might cause flash floods that can be defined as a sudden rise of the water level (in a few hours or less) together with a significant peak discharge (Braud et al. 2014). Flash floods can potentially cause fatalities and important material damage and are known as the main natural hazard in the Mediterranean area (Borga et al. 2011). We refer to rainfall which might lead to flash floods as *flood-risk rainfall*.

A key feature of flood-risk rainfall is its strong spatial variability at high temporal and spatial resolutions.

✉ Julie Carreau
Julie.Carreau@univ-montp2.fr

[1] HydroSciences Montpellier, UMR 5569, CNRS/IRD/UM, Université de Montpellier, Case 17, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

Published online: 07 October 2015

Indeed, Gaume et al. (2009) stressed that albeit flash floods are generally associated with localized intense rainfall that lasts a few hours, they can also be generated by long lasting rainfall with moderate intensities that affects the whole catchment. In the French Mediterranean region, streamflow simulation accuracy and dynamics can be significantly enhanced when exploiting information from rainfall at higher spatial resolution (Lobligeois et al. 2014; Patil et al. 2014; Braud et al. 2014). Therefore, analyses to characterize the spatial variability of flood-risk rainfall will contribute to the understanding of flash flood processes.

We define flood-risk rainfall as rainfall with a high spatial average. More precisely, the spatial average is *high* when it is above a threshold that should be set according to the catchment at hand. This definition encompasses both intense localized events and moderate widespread events, in accordance with expert knowledge. It is straightforward to cast flood-risk rainfall modeling with such a definition into a multivariate or spatial process extreme-value theory (EVT) framework (Coles 2001; Beirlant et al. 2006). In the peaks-over-threshold approach of EVT, models are developed for multivariate or spatial extremes defined as events which are large according to a given norm. With the $\mathcal{L}_1$-norm, this corresponds exactly to the definition of flood-risk rainfall (see Sabourin and Naveau 2014 who proposed a non-parametric multivariate model in this framework). However, the application of these models to flood-risk rainfall raises a number of technical questions (for example, an extreme in the hydrological sense might not be an extreme in the statistical sense).

An alternative approach to analyze and characterize flood-risk rainfall is by means of stochastic rainfall generators (or more generally weather generators, see Ailliot et al. 2015). They can simulate long series of observations from which observations corresponding to flood-risk rainfall (high spatial average) can be extracted and studied. Stochastic generators are complex statistical models which must handle rainfall intermittency (the determination of rainy and dry areas) and rainfall inhomogeneity (the presence of different types of rainfall such as convective and stratiform and of seasonal or diurnal cycles). Intermittency can be addressed either by including an atom at zero in the transformation of the marginal distribution (Bouvier et al. 2003; Vischel et al. 2009; Baxevani and Lennartsson 2015) or by applying an indicator function (Barancourt et al. 1992; Wilks 1998; Hughes et al. 1999; Kleiber et al. 2012; Leblois and Creutin 2013). Inhomogeneity can be incorporated by means of rainfall or weather patterns (Bellone et al. 2000; Thompson et al. 2007; Garavaglia et al. 2010) or by introducing covariates in the distribution parameters (Chandler and Wheater 2002; Kleiber et al. 2012; Baxevani and Lennartsson 2015).

The spatial dependence structure of rainfall is an essential building block of multi-site stochastic generators. Many rely on the meta-Gaussian distributions, i.e. the Gaussian dependence structure combined with a transformation of the marginal distributions (Lebel and Laborde 1988; Wilks 1998; Guillot and Lebel 1999; Bouvier et al. 2003; Vischel et al. 2009; Kleiber et al. 2012; Leblois and Creutin 2013; Serinaldi and Kilsby 2014; Baxevani and Lennartsson 2015). Other multivariate distributions have been employed, with possibly a transformation of the marginals, to infer the spatial structure in stochastic generators but none, as far as we are aware, in a truly multi-site framework (see Flecher et al. 2010 for a single-site multivariable weather generator based on the multivariate Skew Normal distribution and Vrac et al. 2007 for a two-site rainfall generator merging a bivariate Gamma with a bivariate model from EVT). The dependence structure can also be modeled with copulas (Genest and Favre 2007). Besides the Gaussian copula which is equivalent to the meta-Gaussian distribution, several copula families exist such as the Student t, the Archimedean or Extreme Value families but not many are available in dimension greater than two. For instance, Schoelzel and Friederichs (2008) performed modeling of rainfall at two sites with a bivariate Gumbel copula, Bárdossy and Pegram (2009) proposed an asymmetric copula to model rainfall at 32 sites and Serinaldi (2009) proposed a copula-based mixed model for bivariate rainfall.

Extreme rainfall in the French Mediterranean area has been widely studied. To our knowledge, most of the time, a univariate viewpoint is adopted with the block maxima approach of EVT where extreme events are taken as maxima over a period of time such as the year or the month, see Gardes and Girard (2010), Ceresetti et al. (2012) and Carreau et al. (2013) for instance. In contrast, spatial dependence is taken into account in Lebel and Laborde (1988) who proposed a geostatistical approach to model monthly areal rainfall maxima and in Neppel et al. (2011) who developed a multivariate regional test in which the spatial dependence structure is modeled with the Student t copula.

Comparison studies of spatial or multivariate models for extremes were conducted in other application domains or in other study areas. The meta-Gaussian distribution is often chosen because it is easy to implement even in high dimension. However, the Gaussian has very specific dependence properties which should be validated. In finance, impacts on risk measures of the choice of the Gaussian dependence structure compared to other choices were studied in Embrechts et al. (2002) and Poon et al. (2004). In particular, Embrechts et al. (2002) emphasized the potential under-estimation of the probability of joint extreme events when employing the Gaussian copula. In

hydrology, Berg and Aas (2009) modeled daily rainfall at four sites with five types of combined Archimedean copulas and compared the goodness-of-fit with the Cramer-Von-Mises statistics. In Dupuis and Tawn (2001), the effects of mis-specification of the dependence structure on bivariate extreme-value problems were studied on synthetic data while Dupuis (2007) showed the effect of model mis-specification on bivariate hydrometric data sets. More recently, Blanchet and Davison (2011) and Thibaud et al. (2013) (see also references therein) performed model selection of spatial processes for extremes of snow and rainfall, respectively, in Switzerland. These studies show that the choice of the spatial dependence structure for extremes must be made with great care and that the meta-Gaussian distribution can fit very poorly.

In this work, we aim to analyze and compare multivariate density models for multi-site flood-risk rainfall. In particular, we seek to evaluate whether the spatial dependence properties of the models can reproduce the spatial variability of flood-risk rainfall. The study area is a small representative French Mediterranean catchment, the Gardon at Anduze, that is vulnerable to devastating floods (Delrieu et al. 2005). Flood-risk rainfall can be thought of as a type of rainfall (or rainfall pattern) and the strategy that we adopt in this work can thus be seen as focussing on a single rainfall type within a stochastic generator. The flood-risk rainfall type is the most important feature multi-site stochastic generators should be able to reproduce when applied in small Mediterranean catchments. In addition, the adopted strategy is likely to reduce the need to deal with rainfall intermittency and inhomogeneity, as is the case for the Gardon at Anduze catchment. This work is intended as a preliminary study before developing a spatial stochastic rainfall generator adapted for flood-risk rainfall in the Mediterranean area.

The paper is structured as follows. Daily flood-risk rainfall data at eight rain gauge stations in the Gardon at Anduze catchment together with pairwise exploratory analyses are presented in Sect. 2. Although the response time of the catchment is in the order of the hours, we make do with analyses at the daily time-step because a longer and more complete data base is available. We assume that the dependence structure of flood-risk rainfall at the daily time-step provides relevant information on the flash flood processes, even when they occur at the sub-daily scale. Section 3 is dedicated to the description of the eight multivariate density models included in the comparison. Each model consists of marginal distributions, which describe the univariate behavior of daily flood-risk rainfall at each site, combined with a spatial dependence structure, which captures the site-to-site variability at a given day. A parametric marginal model, the Gamma distribution, and a non-parametric marginal model, the mixture of Log-Normal distributions, are described in Sect. 3.1. Four dependence structures, the Gaussian, the Student t, the Skew Normal and the Skew t, in increasing order of complexity, are presented in Sect. 3.2. They possess a representative set of theoretical properties (symmetry/asymmetry and asymptotic dependence/independence for the extremes) and they can be fitted in dimension 8 with available R libraries (Kojadinovic and Yan 2010; Azzalini 2015). Section 4 contains the comparative results in terms of three types of criteria. We first examine separately the goodness-of-fit of the marginal models and of the dependence structures in Sect. 4.1. Second, the best model is selected by performing leave-one-out validation (also called jackknife) with two goodness-of-fit statistics, Cramer-Von Mises and Anderson-Darling, in Sect. 4.2. Third, we look at hydrologically interpretable quantities (return periods of observed spatial average and conditional probability of exceedances, see Thibaud et al. 2013 for similar criteria) which involve the whole multivariate models in Sect. 4.3. We discuss the results and conclude in Sect. 5.

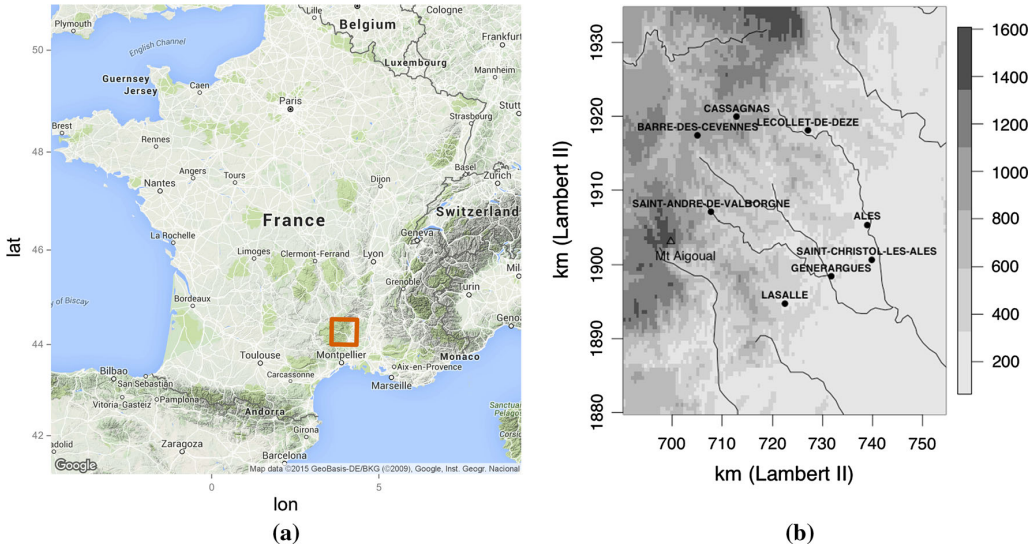
2 Data and exploratory analyses

The catchment of the Gardon at Anduze is a small catchment of about 545 km^2 located in the Cevennes mountain range, in the South of France, see Fig. 1a. It is subject to the Mediterranean climate that, in combination with its sharp orography, can trigger heavy precipitation events especially in the fall season (Ducrocq et al. 2008). Daily rainfall observations are collected over a 43 year period, from 01/01/1958 to 12/31/2000, at eight stations scattered around the catchment, as shown in Fig. 1b and Table 1. Horizontal distance for pairs of stations varies between 5 and 40 km. Elevation ranges from 135 m, in the valleys, to 930 m near the crest of the mountain range.

Based on expert knowledge on the Gardon catchment (Bouvier et al. 2007), we established that rainfall with a spatial average above 50 mm can potentially provoke flooding. In this work, flood-risk rainfall is thus defined as rainfall at the eight rain-gauges provided that the spatial average is above the threshold of 50 mm. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_8)$ be the vector of random variables of the rainfall intensities at each of the eight rain-gauges. Then, flood-risk rainfall corresponds to the following set:

$$\mathcal{F} = \left\{ \boldsymbol{X} \in \mathbb{R}_+^8 \quad | \quad \frac{1}{8}\sum_{s=1}^{8} X_s > 50 \right\}. \tag{1}$$

Out of the 15,706 days of the 43 year observation period, 265 are such that the spatial average is above 50 mm. This is less than 2 % of all the observations and about 5.5 % of the days where it rains (defined as days for which at least



(a) (b)

Fig. 1 Country- and local-scale view of the catchment of the Gardon at Anduze in the South of France. **a** Map of metropolitan France; the catchment of interest is located within the orange rectangle depicted in the South. **b** The eight daily rain-gauge stations used in our study against a digital elevation map (m) revealing the sharp orography in the Cevennes mountain range

Table 1 X and Y Lambert II extended coordinates of the eight daily rain-gauge stations and elevation (Z). See the map in Fig. 1b

Station name	X (km)	Y (km)	Z (m)
BARRE-DES-CEVENNES	705	1917	930
CASSAGNAS	713	1920	800
LECOLLET-DE-DEZE	727	1918	348
ALES	739	1905	135
GENERARGUES	732	1898	138
LASALLE	722	1895	278
SAINT-ANDRE-DE-VALBORGNE	708	1907	450
SAINT-CHRISTOL-LES-ALES	740	1901	138

one station receives more than 1 mm). Among the $8 \times 265 = 2120$ observations, only four are zero. Hence, in order to keep the statistical models as parsimonious as possible, we chose not to model rainfall intermittency. Instead, we assume that rainfall intensities on flood-risk days are strictly positive at every station and these four zero observations are lifted to 0.2 mm, the rain-gauge measurement error.

Although preliminary analyses detect some order 1 auto-correlation when flood-risk rainfall happens on two consecutive days, i.e. $Cor(\boldsymbol{X}_t, \boldsymbol{X}_{t+1})$, we model $\boldsymbol{X}_t$ as independent. As a result, the confidence intervals presented in the analyses might be narrower than if temporal dependence was taken into account, since the effective number of observations might be somewhat reduced. Consecutive flood-risk rainfall days occur 43 times and the largest magnitude of the auto-correlation is about 0.3 so we expect that the independence assumption does not have very significant impacts.

We further assume that flood-risk rainfall is identically distributed (homogeneity assumption). Flood-risk rainfall happens mainly during the fall season but there are occurrences throughout the year. It is likely that both convective and stratiform types of rainfall are included in our definition of flood-risk rainfall. In order to distinguish between the two of them, additional information, unavailable to us, such as the prevalent atmospheric circulation or sub-daily rainfall intensities, is needed. Since this information is rarely available, a classic way to attempt to ensure homogeneity is to perform separate modeling for each season or each month. The resulting model is a mixture with one component per season or per month, see Garavaglia et al. (2010) for example. For the catchment considered in this work, homogeneity would not be guaranteed by seasonal or monthly modeling as both convective and non-convective processes, known to yield heavy rainfall, can occur during the same season or the same month (Delrieu et al. 2005). Instead, we take a statistical approach to address the homogeneity assumption. We allow for mixture of distributions for the marginal models (Sect. 3.1) and for the dependence structures (Sect. 3.2.3). The adequate number of components is selected with the Bayesian Information Criterion (BIC) (Schwarz 1978).

2.1 Pairwise dependence

Pairwise dependence is first evaluated with the estimation of Kendall's τ coefficients. Kendall's τ is a measure of dependence based on the difference between the probability of concordant and discordant pairs. Let X_i and X_j be two random variables representing flood-risk rainfall at station i and station j respectively. Then the τ coefficient for these two stations is given by:

$$\begin{aligned}\tau(X_i, X_j) &= P\big((X_i - \tilde{X}_i)(X_j - \tilde{X}_j) > 0\big) \\ &\quad - P\big((X_i - \tilde{X}_i)(X_j - \tilde{X}_j) < 0\big) \\ &= \mathbb{E}[sign(X_i - \tilde{X}_i)(X_j - \tilde{X}_j)]\end{aligned} \tag{2}$$

where $(\tilde{X}_i, \tilde{X}_j)$ is an independent copy of the pair (X_i, X_j) with identical distribution. Kendall's τ is invariant to strictly increasing transformations (Joe 1997) and thus, it is unaffected by marginal transformations. It takes values in the interval $[-1, 1]$, where $\tau = -1$ or $\tau = 1$ means perfect negative or positive dependence and $\tau = 0$ means independence.

Kendall's τ coefficients, significantly different from zero at the 5 % level, are presented in the upper triangular part of Fig. 2. Pairwise scatterplots are shown in the lower triangular part of Fig. 2 together with a smooth regression line obtained from local regression (Cleveland 1981) to help detect dependencies. Figure 2 reads as follows. Each station name and histogram appears on the diagonal. Row i and column i concerns the ith station. At the intersection of column i and row j, with $j > i$, there is the scatterplot of the pair of stations i (x-axis) and j (y-axis). Conversely, at the intersection of column j and row i, the corresponding Kendall's τ coefficient is written when significant at the 5 % level. The axes shown can be associated to the lower triangular scatterplots or to the histograms on the diagonal.

As is well recognized in the literature (Serinaldi and Kilsby 2014), Kendall's τ depends on the distance between the stations. For the flood-risk rainfall data, the τ estimates appear to be linearly decreasing with the horizontal distance, as shown in Fig. 3 with a regression line. Kendall's τ for pairs of stations which are less than about 12 km apart ranges from 0.4 to 0.7 and then decreases to values close to zero or negative for stations which are more than 25 km apart.

We consider a second measure of pairwise dependence, the χ coefficient, which measures extremal dependence

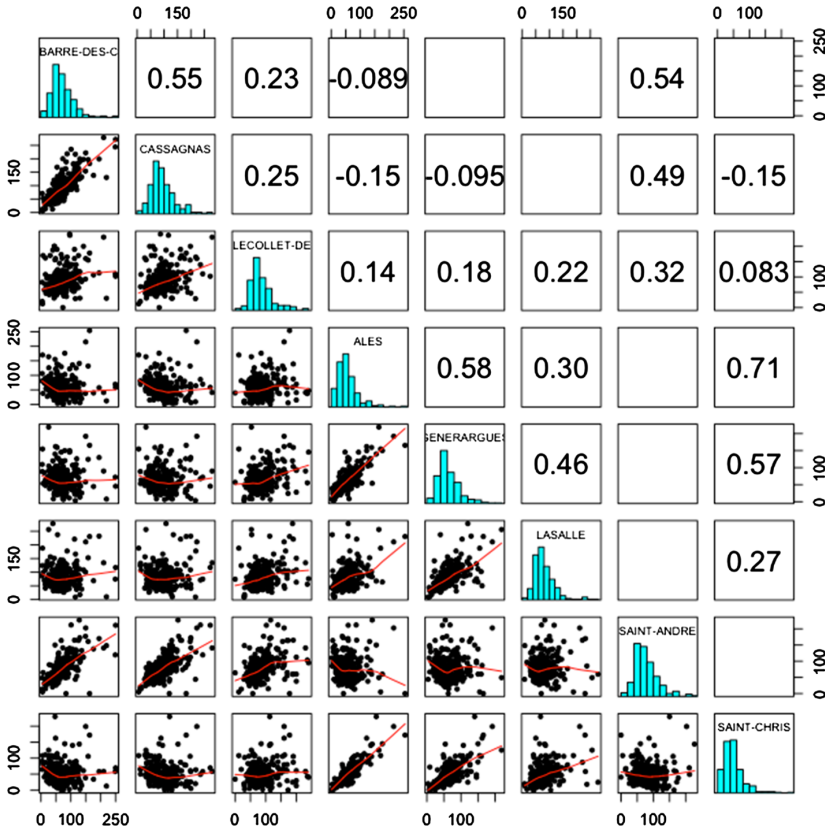


Fig. 2 Evaluation of pairwise dependence for the eight rain gauge stations. Diagonal: station names and histogram of flood-risk rainfall. *Lower triangle*: scatterplot of the flood-risk rainfall (*black dots*) with a locally smoothed regression "lowess" line. *Upper triangle*: Kendall's τ coefficient significant at the 5 % level

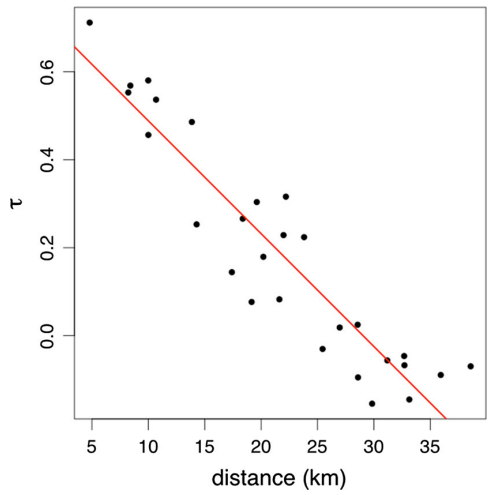


Fig. 3 Plot of the estimated Kendall's τ coefficients with respect to horizontal distances together with a regression line

(Coles et al. 1999). For a pair of variables (X_i, X_j), it is defined as:

$$\chi(X_i, X_j) = \lim_{u \to 1} P(F_{X_j}(X_j) > u | F_{X_i}(X_i) > u), \tag{3}$$

where F_{X_i} and F_{X_j} are the distribution functions of X_i and X_j respectively. Loosely stated, χ is the probability of one variable being extreme given that the other is extreme. Since χ is the limit of a conditional probability, it takes values in [0, 1]; when $\chi = 0$, X_i and X_j are said to be asymptotically independent whereas when $\chi > 0$, they are asymptotically dependent (perfect dependence is achieved when $\chi = 1$). In practice, χ is estimated for a fixed threshold u, taken as high as possible.

A useful tool to assess whether a pair of variables X_i and X_j is asymptotically dependent or independent is the so-called χ-plot (Coles et al. 1999). In such a plot, $\chi(X_i, X_j)$ is estimated and plotted against increasing thresholds u expressed as quantiles of level $q \in [0, 1]$. Confidence intervals can be constructed with the delta method but are not very reliable near $q = 0$ or $q = 1$.

In the flood-risk rainfall data, both asymptotically dependent and independent pairs of stations appear to be present. In Fig. 4, the χ-plots for two representative pairs of stations are shown together with a 95 % confidence interval (`R` package `evd` Stephenson 2002). The χ-plot of the pair of nearby stations, Barre-des-Cevennes and Cassagnas, in Fig. 4a, is rather stable around the value 0.6, regardless of the threshold, and thus indicates asymptotic dependence. In contrast, for the pair of distant stations, Barre-des-Cevennes and Generargues, in Fig. 4b, the χ estimates increase from negative values (which are caused by the estimator employed, see Coles et al. 1999) to values near zero, indicative of asymptotic independence.

As discussed in Serinaldi et al. (2014), χ estimators are strongly positively related to Kendall's τ coefficients. In Fig. 5, the χ estimates for the flood-risk rainfall data with a threshold u set to the 95 % quantile (`R` package `extRemes` from Gilleland and Katz 2011) are plotted with respect to the τ estimates. When the τ estimates are positive, the scatter plot is quite well aligned with the $y = x$ line.

3 Multivariate density models

3.1 Marginal distributions

The first marginal model considered is the Gamma distribution which has often been used to model daily precipitation (Chandler and Wheater 2002; Flecher et al. 2010; Kleiber et al. 2012). The Gamma density is given by:

$$f^{\text{Gam}}(x; k, \eta) = \frac{x^{k-1} e^{-x/\eta}}{\eta^k \Gamma(k)} \quad x > 0, \tag{4}$$

where k and η are the shape and scale parameters respectively with $k, \eta > 0$ and $\Gamma(k)$ is the Gamma function evaluated at k. Estimates of k and η are obtained by the maximum likelihood estimation method (`R` package `MASS` from Venables and Ripley 2002).

We consider as a second marginal model a mixture of Log-Normal distributions although some authors recommend to model rainfall with a hybrid distribution. Such a hybrid distribution combines a parametric (Carreau and Bengio 2009; Li et al. 2012) or non-parametric model (Lennartsson et al. 2008) for the bulk of the distribution with the Generalized Pareto distribution (GPD) in the upper tail. Univariate extreme value theory (EVT) provides an asymptotic justification for the GPD to be an appropriate model for the distribution of values exceeding a suitably chosen high threshold (Pickands 1975). An advantage of such a hybrid distribution is its ability to adapt to any type of upper tail behavior be it finite, exponential (light-tail) or power-law (heavy-tail).

The motivation for the choice of the Log-Normal mixture instead of the hybrid distribution is twofold. First, preliminary analyses based on fitting the GPD revealed that the marginal distributions of the flood-risk rainfall data appear to be light-tailed. The Gamma is light-tailed but, because it has only two parameters, might lack the flexibility to model both the bulk of the rainfall distribution and its upper tail. Second, the Log-Normal mixture is straightforward to fit, has shown to be a good model for rainfall in Southen France (Carreau and Vrac 2011) and

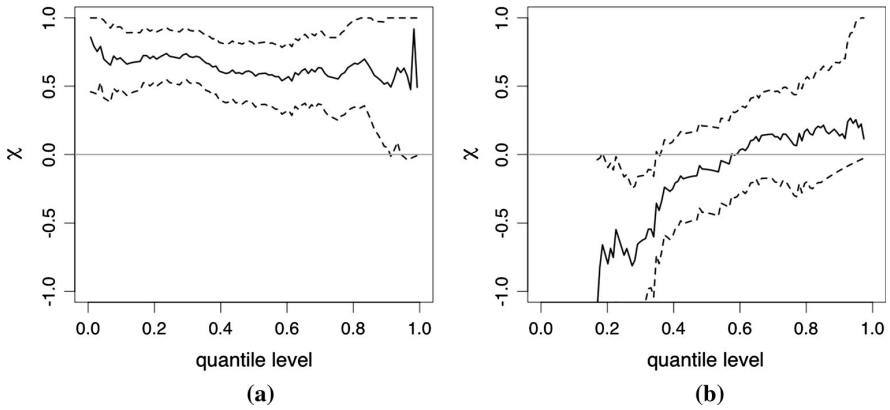


Fig. 4 Two representative χ-plots: χ estimates with respect to u with 95 % confidence intervals (*black dashed lines*). **a** Two nearby stations Barre-des-Cevennes and Cassagnas. The χ-plot indicates asymptotic dependence. **b** Two distant stations Barre-des-Cevennes and Generargues. The χ-plot indicates asymptotic independence

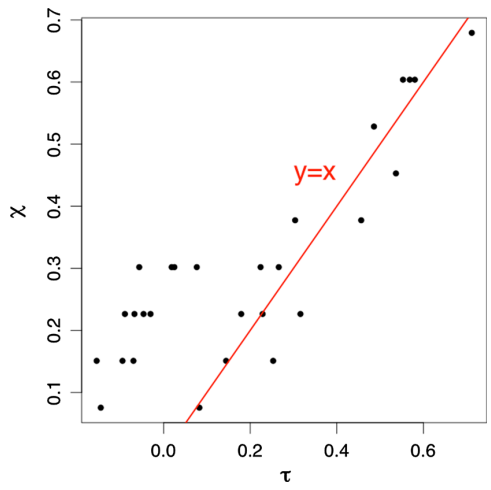


Fig. 5 Relationship between estimated χ coefficients and estimated Kendall's τ

can take into account the presence of more than one sub-population of rainfall, such as convective and stratiform, if needed.

The number of mixture components must be chosen carefully according to the data set. Indeed, a mixture of distributions is a non-parametric model which mean that the complexity, that is the number of free parameters driven by the number of mixture components, can increase as the data set gets larger (Carreau and Bengio 2009). For all eight stations, two components in the Log-Normal mixture were selected with the BIC. We used the `R` package from Frayler and Raftery (1999) for Gaussian mixtures on log-transformed data. From a statistical viewpoint, the population of flood-risk rainfall at each station is adequately modeled with a two-component Log-Normal mixture. The marginal model, see Eq. (5), has thus 5 parameters $\boldsymbol{\psi} = (\lambda, \tilde{\mu}_1, \tilde{\sigma}_1, \tilde{\mu}_2, \tilde{\sigma}_2)$ where $\lambda \in [0,1]$ is the mixture proportion, $\tilde{\mu}_i \in \mathbb{R}$ and $\tilde{\sigma}_i > 0$, $i = 1, 2$, are the location and scale parameters of the ith Log-Normal component.

$$f^{\text{lnmix}}(x;\boldsymbol{\psi}) = \lambda \frac{1}{2\sqrt{\tilde{\sigma}_1 \pi x}} \exp\left\{-(\ln x - \tilde{\mu}_1)^2/(2\tilde{\sigma}_1)\right\} + (1-\lambda)\frac{1}{2\sqrt{\tilde{\sigma}_2 \pi x}} \times \exp\left\{-(\ln x - \tilde{\mu}_2)^2/(2\tilde{\sigma}_2)\right\} \quad x > 0 \tag{5}$$

3.2 Spatial dependence structures

3.2.1 Gaussian and student t copulas

The first two dependence structures included in the comparison are the Gaussian and Student t copulas which belong to the elliptical family (`R` package `copula` from Kojadinovic and Yan 2010). As a widespread model among practitioners, the Gaussian copula, that represents the class of meta-Gaussian models, is taken as the benchmark model. The Student t copula has an additional parameter, the degree of freedom ν, which provides greater modeling flexibility in terms of tail dependence and encompasses the Gaussian copula as a limiting case, when $\nu \to \infty$.

As mentioned in Genest and Favre (2007), the main advantage of the copula approach is that a valid multivariate model can be built by selecting a dependence structure represented by the copula and then selecting independently the marginal distributions. Let $\boldsymbol{X} \in \mathcal{F}$ be as in Eq. (1), the random vector of flood-risk rainfall intensities, let $F_{X_j}, j = 1, \ldots, 8$ be its marginal distributions and $F_{\boldsymbol{X}}$ be its joint distribution function. Then, by Sklar's theorem (Sklar 1959), the associated copula, assuming $\boldsymbol{X}$ is continuous, is a function $C_{\boldsymbol{\theta}} : [0,1]^8 \to [0,1]$, with parameter vector $\boldsymbol{\theta}$, such that:

$$F_{\boldsymbol{X}}(x_1,\ldots,x_8) = C_{\boldsymbol{\theta}}(F_{X_1}(x_1),\ldots,F_{X_8}(x_8)) \quad (x_1,\ldots,x_8) \in \mathbb{R}_+^8. \tag{6}$$

There are no closed-form expressions for the Gaussian and Student t copulas as expressions for $C_{\boldsymbol{\theta}}$ can be obtained by means of the formula:

$$C_{\boldsymbol{\theta}}(u_1,\ldots,u_8) = F_{\boldsymbol{X}}(F_{X_1}^{-1}(u_1),\ldots,F_{X_8}^{-1}(u_8)) \quad (u_1,\ldots,u_8) \in [0,1]^8 \tag{7}$$

where $F_{X_j}^{-1}$ denotes the quantile function of the margins that do not have closed-form expressions for the Gaussian and Student t distributions.

The Gaussian and Student t copula parameters stem from the parameters of their associated standardized multivariate distribution functions. This is because copulas, by definition, are invariant under a standardization of the marginal distributions. The expressions of the standardized densities are given in Eq. (8) for the Gaussian and Eq. (9) for the Student t with $\boldsymbol{x} \in \mathbb{R}^d$. In the Gaussian case, the parameter vector $\boldsymbol{\theta}$ contains the free parameters of the $d \times d$ correlation matrix P which must be symmetric and positive definite. In the Student t case, $\boldsymbol{\theta}$ contains, in addition to P, the degree of freedom parameter $\nu > 0$.

$$f^{\text{Gauss}}(\boldsymbol{x};P) = \frac{1}{\sqrt{(2\pi)^d|P|}}\exp\left\{-\frac{1}{2}\boldsymbol{x}^T P^{-1}\boldsymbol{x}\right\} \tag{8}$$

$$f^{\text{Stu}}(\boldsymbol{x};P,\nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\nu\pi)^d|P|}}\left\{1+\frac{\boldsymbol{x}^T P^{-1}\boldsymbol{x}}{\nu}\right\}^{-\frac{\nu+d}{2}} \tag{9}$$

For the Gaussian and Student t copulas, Kendall's τ takes the same form, see Eq. (10). It is positively related to the correlation parameter ρ_{ij} in the matrix P associated to the pair (X_i, X_j) (Demarta and McNeil 2005). In regard to the variety of strengths of empirical Kendall's τ coefficients in the flood-risk rainfall data (see Fig. 2), we chose not to impose any specific parametric form on the correlation matrix P so that it has as much flexibility as needed.

$$\tau(X_i,X_j) = 4\int_0^1\int_0^1 C(u_1,u_2)dC(u_1,u_2) - 1 = \frac{2}{\pi}\arcsin\rho_{ij} \tag{10}$$

In contrast, the extremal behavior of the Gaussian and Student t copulas differ. This can be analyzed through the χ-coefficient of extremal dependence, defined in Eq. (3), that can be expressed in terms of the copula function, see Eq. (11). The Gaussian is asymptotically independent with $\chi = 0$, provided that $|\rho_{ij}| < 1$ although this behavior might come into play only for very extreme values if ρ_{ij} is close enough to 1 (Coles et al. 1999). In contrast, the Student t is asymptotically dependent with positive χ as long as $\rho_{ij} > -1$ and $\nu < \infty$ (Demarta and McNeil 2005). The smaller ν is, the larger χ becomes.

$$\chi(X_i,X_j) = \lim_{u\uparrow 1}\frac{1-2u+C(u,u)}{1-u} \quad \begin{cases} = 0 & \text{Gaussian with } |\rho_{ij}|<1 \\ > 0 & \text{Student t with } \rho_{ij} > -1 \text{and } \nu < \infty \end{cases} \tag{11}$$

The copula parameters are estimated by the method of maximum pseudo-likelihood (MPL) that relies on the ranks of the observations (Genest and Favre 2007). This way, the estimation of the dependence structure is completely independent from the estimation of the marginal distributions. Provided that the density $c_{\boldsymbol{\theta}}$ associated to the copula exists, MPL involves the maximization of a log-likelihood of the form:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\log\left\{c_{\boldsymbol{\theta}}\left(\hat{F}_{X_1}(x_1),\ldots,\hat{F}_{X_8}(x_8)\right)\right\}, \tag{12}$$

where $\hat{F}_{X_s}(x_s)$, $s = 1,\ldots,8$, are the marginal empirical distribution functions which depend essentially on the ranks.

The densities of the Gaussian and Student t copulas in the bivariate case are illustrated in Fig. 6a, b respectively. For both copulas, $\rho = 0.5$ and thus Kendall's τ is equal to 0.33, according to Eq. (10). The degree of freedom parameter is $\nu = 4$ which, together with $\rho = 0.5$, yields a coefficient of extremal dependence of $\chi = 0.25$ for the Student t copula (Demarta and McNeil 2005). The asymptotic dependence of the Student t copula is associated with a higher density of joint extremes, as can be seen by comparing the Gaussian and the Student t copulas in Fig. 6.

3.2.2 Skew normal and Skew t

The last two dependence structures included in the comparison are the multivariate Skew Normal and Skew t distributions (`R` package `sn` Azzalini 2015). They can be thought of as asymmetric extensions of their generating distribution (Gaussian for the Skew Normal and Student t for the Skew t). The Skew distributions have an additional vector of d parameters, $\boldsymbol{\alpha}$, which act as skewness parameters. The Gaussian and Student t distributions appear as special cases when $\boldsymbol{\alpha} = 0$.

The density of the Skew Normal (resp. Skew t) are given in Eq. (13) (resp. Eq. (14)) where $\boldsymbol{x} \in \mathbb{R}^d$ (Azzalini and Capitanio 2003). For both the Skew Normal and the Skew t, $\boldsymbol{\alpha} \in \mathbb{R}^d$ projects $\boldsymbol{x}$ onto a line and P is a $d \times d$ correlation (or dispersion) matrix that is symmetric and positive definite. The Skew t has, in addition, the degree of freedom parameter, $\nu > 0$, inherited from the Student t.

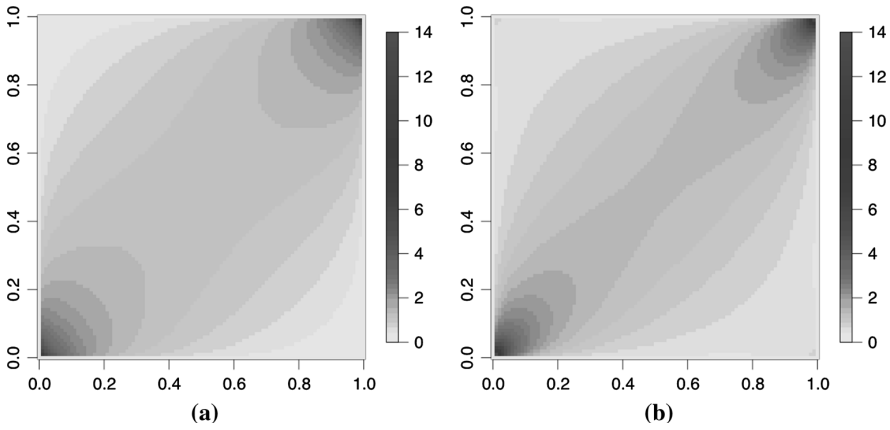


Fig. 6 Bivariate elliptical copula densities. **a** Gaussian copula with $\rho = 0.5$. **b** Student t copula with $\rho = 0.5$ and $\nu = 4$

The densities of the Skew distributions from Eqs. (13–14) are obtained by multiplying the multivariate Gaussian density $f^{\text{Gauss}}(\boldsymbol{x}; P)$ from Eq. (8) or the multivariate Student t density $f^{\text{Stu}}(\boldsymbol{x}; P, \nu)$ from Eq. (9) by a skewing factor which is based on the univariate standard Gaussian distribution function $F^{\text{Gauss}}(\cdot)$ or the univariate standard Student t distribution function with $\nu + d$ degree of freedom $F^{\text{Stu}}(\cdot; \nu + d)$.

$$f^{\text{SN}}(\boldsymbol{x}; P, \boldsymbol{\alpha}) = 2\, f^{\text{Gauss}}(\boldsymbol{x}; P)\; F^{\text{Gauss}}(\boldsymbol{\alpha}^T \boldsymbol{x}) \tag{13}$$

$$f^{\text{St}}(\boldsymbol{x}; P, \nu, \boldsymbol{\alpha}) = 2\, f^{\text{Stu}}(\boldsymbol{x}; P, \nu)\; F^{\text{Stu}}\left(\boldsymbol{\alpha}^T \boldsymbol{x} \left\{ \frac{\nu + d}{\boldsymbol{x}^T P^{-1} \boldsymbol{x} + \nu} \right\}^{1/2}; \nu + d \right) \tag{14}$$

Since the Skew distributions include their generating distributions as special cases, we expect that they might share their pairwise dependence properties. In terms of Kendall's τ, to our knowledge, no closed-form expressions were derived for the Skew distributions. In terms of χ-coefficient, Bortot (2010) has shown that the Skew Normal is asymptotically independent ($\chi = 0$) and the Skew t is asymptotically dependent ($\chi > 0$) as it is the case for the symmetric distributions of Eq. (11). However, Bortot (2010) argued that the Skew distributions have greater flexibility and can adapt to a larger variety of extremal dependence strengths than their generating distributions.

In order to separate the inference of the margins from the inference of the spatial structure, we adapted the margin transformation proposed in (Flecher et al. 2010):

$$H^{-1}(\hat{F}_{X_i}(X_i)) \tag{15}$$

where $\hat{F}_{X_s}(x_s)$, $s = 1, \ldots, 8$, are the marginal empirical distribution functions and H^{-1} is the quantile function of a suitable univariate distribution. When the spatial structure is the multivariate Skew Normal (resp. Skew t), the univariate standard Normal (resp. standard Student t with fixed degree of freedom parameter) is used in the transformation of Eq. (15). We did not employed copulas for skew distributions because they were not available yet in `R` packages although theoretical developments are underway (Kollo et al. 2013).

The parameters of the Skew distributions are estimated by maximizing the following log-likelihood:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left\{ f^{\text{Skew}}\left(H^{-1}(\hat{F}_{X_1}(X_1)), \ldots, H^{-1}(\hat{F}_{X_8}(X_8)); P, \boldsymbol{\alpha}\right)\right\}, \tag{16}$$

where f^{Skew} is the density of either the Skew Normal from Eq. (13) or the Skew t from Eq. (14) with degree of freedom fixed to the value estimated for the Student t copula. Therefore, the parameter vector $\boldsymbol{\theta}$ contains the free parameter of the 8×8 correlation matrix P and the skewness parameters $\boldsymbol{\alpha}$. In practice, the `sn` package (Azzalini 2015) did not allow to fix the location parameters to zero and the scale parameters to 1 in the estimation as would be required by the margin transformation. In order to stay as close as possible to these parameter values, they were used as starting parameter values for the optimization. The optimized parameter values did not wander too far from the starting values. The Skew t, the most complex model, was difficult to fit. Sensible starting values, taken from the fitted Skew Normal, were provided to the optimizer to help the estimation of the parameters.

Two types of departure from symmetry are illustrated in Fig. 7 in terms of bivariate copula density for the Skew Normal (left column) and the Skew t (right column). Copula densities are computed by deriving the expression in Eq. (6) with respect to x_i, $i = 1, \ldots, 8$ (see Kollo et al. 2013). In all four cases, $\rho = 0.5$ and $\nu = 4$ for the Skew t so that these copula densities can be compared to their symmetric counterparts in Fig. 6. In the top row, the skewness parameter is $\alpha = (-1, 1)$ and produces an asymmetry with respect to the line $y = x$. In this case, the x-axis variable most often takes higher values than the y-

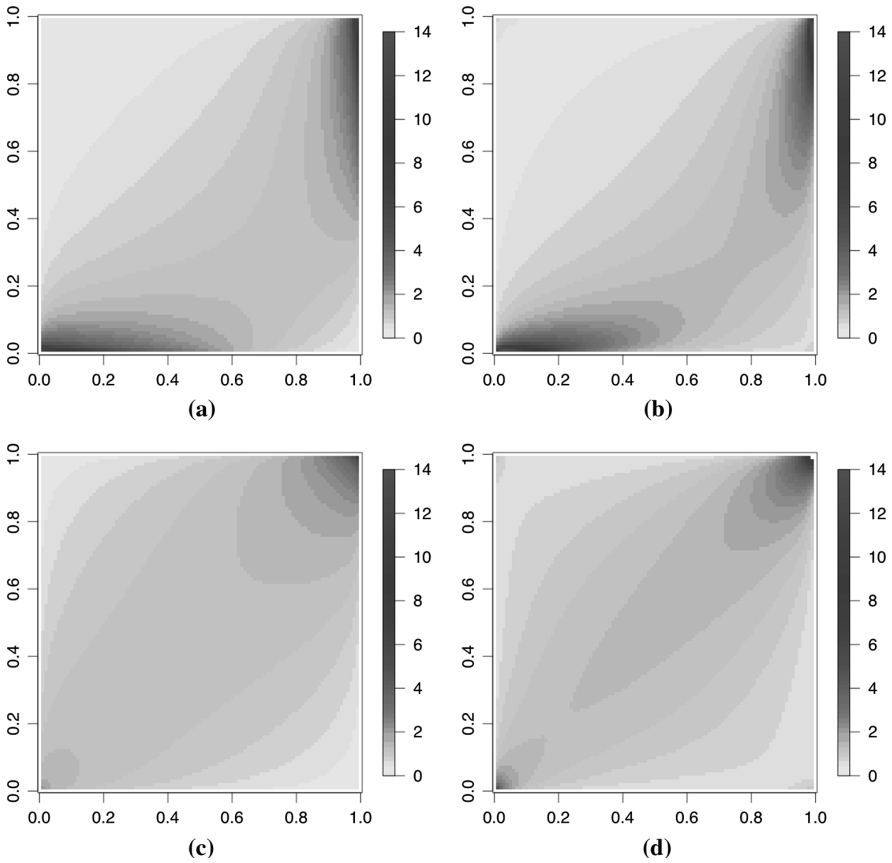


Fig. 7 Bivariate skew copula densities. **a** Skew Normal copula with $\rho = 0.5$ and $\alpha = (-1, 1)$. **b** Skew t copula with $\rho = 0.5$, $\nu = 4$ and $\alpha = (-1, 1)$. **c** Skew Normal copula with $\rho = 0.5$ and $\alpha = (0.5, 0.5)$. **d** Skew t copula with $\rho = 0.5$, $\nu = 4$ and $\alpha = (0.5, 0.5)$

axis variable. In the rainfall application, this translates into one station generally hitting higher quantile values of its marginal distribution with respect to another station. In the bottom row, $\alpha = (0.5, 0.5)$ yields an asymmetry with respect to the line $y = 1 - x$. This results in lower dependence at the smaller values than at the larger values. This can be related to the fact that low rainfall intensities tend to be scattered and intermittent and thus often display poor spatial dependence whereas high rainfall intensities tend to be more dependent (Bárdossy and Pegram 2009).

3.2.3 Multivariate mixture

In order to account for the possible presence of more than one sub-population of rainfall, we tested whether a multivariate mixture with more than one component was required. The margins of the flood-risk rainfall data were transformed to standard Gaussian with the empirical marginal distribution functions, see Eq. (15), and a multivariate Gaussian mixture was fitted to the transformed data. Then, the BIC was used to select the appropriate number of components (Frayler and Raftery 1999).

According to the BIC, a single Gaussian component is needed to model the dependence structure. We expect that only one component would be selected as well when considering a mixture with the other models (Student t, Skew Normal and Skew t). Indeed, these models include the Gaussian as a special case and have a larger number of parameters. For the BIC to select more than one component, the increase in goodness-of-fit versus the increase in complexity (number of parameters) would have to be very significant. The test provide sufficient grounds to keep a single dependence structure model and not to consider further multivariate mixture modeling.

4 Comparative results

4.1 Statistical inference

First, we seek to evaluate independently how good the marginal and dependence structure models are at fitting the flood-risk rainfall data.

4.1.1 Margin fit

The fit of the two marginal distributions considered is evaluated by means of quantile-quantile plots (qq-plots) as shown in Fig. 8 for the Gamma distribution and in Fig. 9 for the 2-component Log-Normal mixture. In all qq-plots, the empirical quantiles are represented on the x-axis and the theoretical quantiles from the marginal distributions on the y-axis. Confidence intervals at 95 % for the theoretical quantiles are computed with 1000 parametric bootstrap replications. To ease comparison across qq-plots, the first diagonal is drawn on the interval [0, 300]. For a given station, the marginal model is considered to fit well if the first diagonal is within the confidence interval most of the time.

The Gamma distribution fails at representing the upper tail and thus the largest observations of flood-risk rainfall at at least four stations (Barre-des-Cevennes, Ales, Lasalle and Saint-Christol-les-Ales). In contrast, the 2-component Log-Normal mixture yields a better fit at the expense of wider confidence intervals that are most likely due to the higher number of parameters (5 parameters as compared to 2 for the Gamma distribution).

4.1.2 Dependence structure fit

There is no straightforward way to visually assess the fit of a dependence structure, especially in high dimension. We make do with comparisons in terms of pairwise dependence. First, we evaluate whether the models are able to reproduce the empirical Kendall's τ for all pairs of stations. We dropped the evaluation in terms of the extremal χ coefficients as we have seen that the χ estimators are positively related to the τ coefficient estimators, see Fig. 5. Second, we look at bivariate densities for two representative pairs of stations.

All four dependence structures give theoretical Kendall's τ coefficients that are quite close to the empirical estimates, as can be seen in Fig. 10. For the Gaussian and Student t copulas, the theoretical τ coefficients are computed thanks to the relationship with the correlation coefficients in Eq. (10). For the Skew distributions, the theoretical τ coefficients are estimated by computing the empirical τ estimates on random samples of size 5000 from the fitted distributions.

In order to gain more insight into the models, we also look at the fitted bivariate copula densities for the same two pairs of stations as chosen for illustration for the χ-plots in Fig. 4: a nearby pair, Barre-des-Cevennes and Cassagnas, in Fig. 11 and a distant pair, Barre-des-Cevennes and Generargues, in Fig. 12.

The fitted bivariate copula densities are estimated with bivariate hexagonal histograms (`R` package `fMultivar` provided by Rmetrics https://www.rmetrics.org/) on random samples of size 10^6 from the copulas associated to

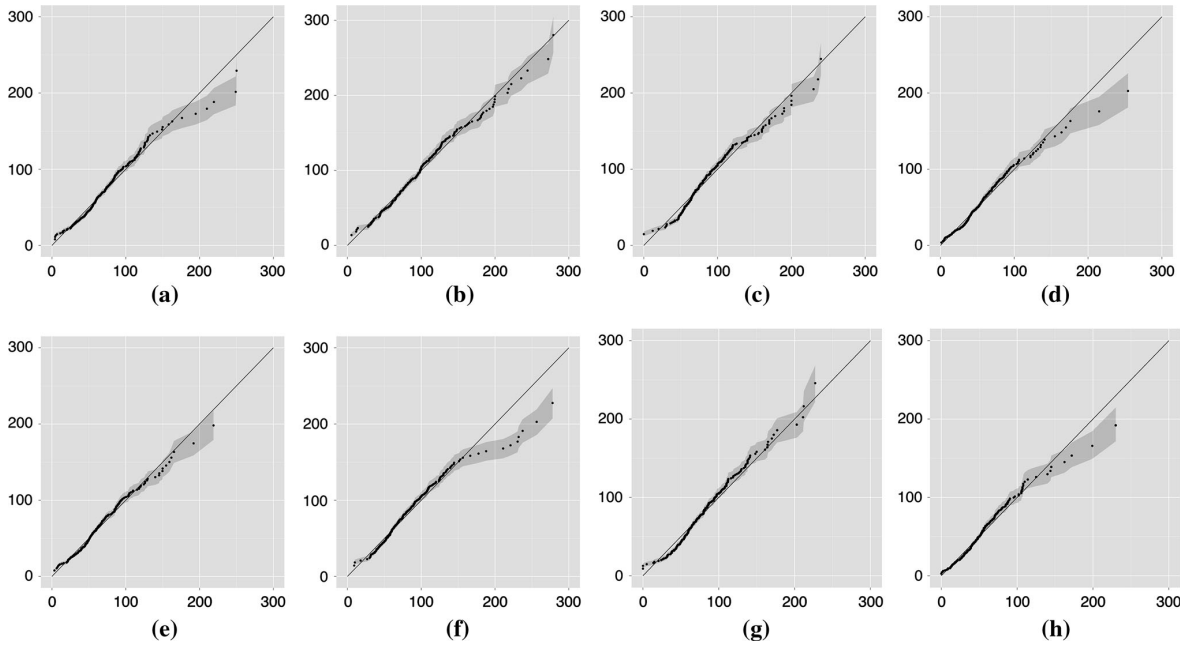


Fig. 8 Quantile–quantile plots of the Gamma distribution at each station with parametric bootstrap 95 % confidence interval. Empirical quantiles (theoretical quantiles) are on the x-axis (y-axis).The range of the first diagonal covers [0,300] on both axes in all plots. **a** Barre-des-Cevennes. **b** Cassagnas. **c** Lecollet-de-Deze. **d** Ales. **e** Generargues. **f** Lasalle. **g** Saint-Andre-de-Val. **h** Saint-Christol-les-A

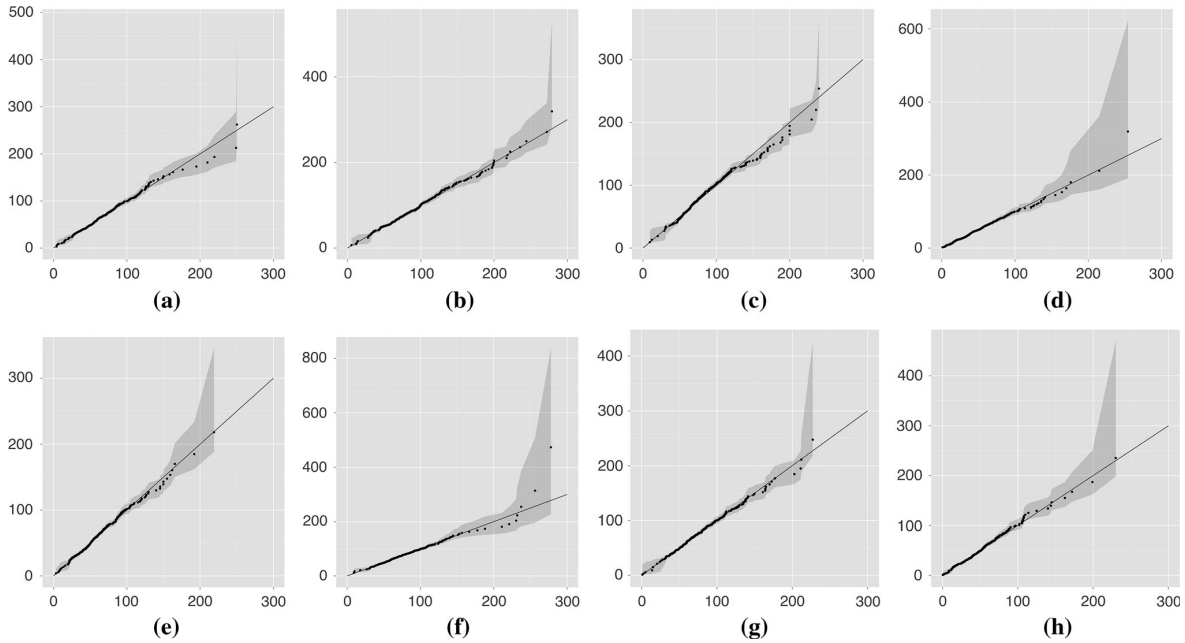


Fig. 9 Quantile–quantile plots of the 2-component Log-Normal mixture distribution at each station with parametric bootstrap 95 % confidence interval. Empirical quantiles (theoretical quantiles) are on the x-axis (y-axis).The range of the first diagonal covers [0,300] on both axes in all plots. **a** Barre-des-Cevennes. **b** Cassagnas. **c** Lecollet-de-Deze. **d** Ales. **e** Generargues. **f** Lasalle. **g** Saint-Andre-de-Val. **h** Saint-Christol-les-A

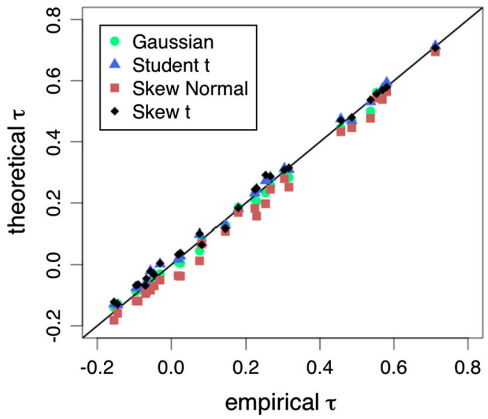


Fig. 10 Theoretical Kendall's τ coefficients of the fitted four spatial dependence structures with respect to the empirical Kendall's τ coefficients for all pairs of stations

each fitted model. We made this choice because the bivariate margins of the Skew distributions are not easy to deduce (Azzalini 2013). The darker the histogram bin is, the higher the density is estimated. The same scale of grey is used for each pair of stations. The dots represent the observations.

Figures 11, 12 are organized as follows. The asymptotically independent dependence structures are in the left column (Gaussian and Skew Normal) and the asymptotically dependent ones are in the right column (Student t and Skew t). The symmetric dependence structures are in the top row (Gaussian and Student t) while the asymmetric ones (Skew Normal and Skew t) are in the bottom row.

Although the theoretical Kendall's τ coefficients are very similar to the empirical ones (0.55 for Barre-des-Cevennes and Cassagnas and -0.067 for Barre-des-Cevennes and Generargues), there might be important differences in the bivariate densities such as those observed for the distant pair, Barre-des-Cevennes and Generargues in Fig. 12. Therefore, assessing whether the empirical Kendall's τ coefficients are reproduced is clearly not enough to determine which model is the most appropriate.

4.2 Leave-one-out model selection

Second, model selection is achieved by performing an automatic quantitative evaluation of the fit of the multivariate density models based on leave-one-out validation (sometimes also called jackknife). With such a validation

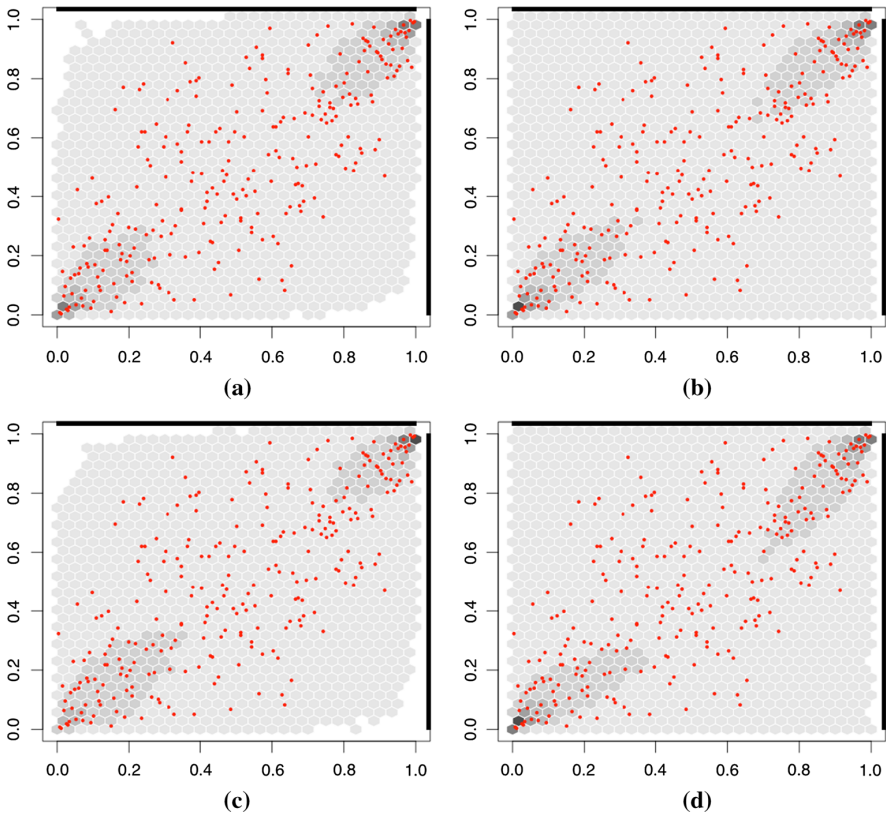


Fig. 11 Fitted bivariate copula densities for the nearby pair of stations (Barre-des-Cevennes and Cassagnas) as estimated by bivariate hexagonal histograms on random samples of size 10^6 from the copulas associated with each fitted model. The darker the histogram bin is, the higher the density is estimated. The *dots* represent the observations. The empirical Kendall's τ is 0.55. **a** Gaussian copula. **b** Student t copula. **c** Skew Normal. **d** Skew t

scheme, each observation is left aside in turn and the models are fitted on the $n-1$ observations. Performance measures are then computed on the observation that was left aside. Since the performance is evaluated out-of-sample, the comparison is fair between models even when they have different numbers of parameters (see Chapter 2.7 in Ripley 1996).

We used the Cramer-Von Mises and Anderson-Darling goodness-of-fit statistics as performance measures. These goodness-of-fit statistics can be seen as distances between the empirical distribution function and the theoretical distribution function $F_{\boldsymbol{\phi}}$ of a given multivariate density model, with $\boldsymbol{\phi}$ including margin and dependence parameters. The Cramer-Von Mises statistic is simply defined as the square distance between the two distribution functions while in the Anderson-Darling statistic, weights are introduced to emphasize an accurate representation of extreme values (Genest et al. 2013).

In the first step of the leave-one-out scheme, for a given $1 \leq k \leq n$, we compute $\hat{F}_{k:n-1}$, the empirical distribution function, and $\hat{\boldsymbol{\phi}}_{k:n-1}$, the parameter estimates of the theoretical distribution function, on sets of the form:

$$\mathcal{F}_{k:n-1} = \left\{ \boldsymbol{X}_j = (X_1^j, X_2^j, \ldots, X_8^j) \mid j \in \{1, \ldots n\} \setminus \{k\} \wedge \frac{1}{8} \sum_{s=1}^{8} X_s^j > 50 \right\}. \tag{17}$$

In the second step of the leave-one-out scheme, the goodness-of-fit statistics are evaluated on the left-out observation $\boldsymbol{X}_k$ with the distribution functions fitted on $\mathcal{F}_{k:n-1}$. The expressions for the Cramer-Von Mises and the Anderson-Darling statistics are given in Eqs. (18) and (19) respectively.

$$CvM_k = \left\{ \hat{F}_{k:n-1}(\boldsymbol{X}_k) - F_{\hat{\boldsymbol{\phi}}_{k:n-1}}(\boldsymbol{X}_k) \right\}^2 \tag{18}$$

$$AD_k = \frac{\left\{ \hat{F}_{k:n-1}(\boldsymbol{X}_k) - F_{\hat{\boldsymbol{\phi}}_{k:n-1}}(\boldsymbol{X}_k) \right\}^2}{F_{\hat{\boldsymbol{\phi}}_{k:n-1}}(\boldsymbol{X}_k)(1 - F_{\hat{\boldsymbol{\phi}}_{k:n-1}}(\boldsymbol{X}_k)) + 0.05} \tag{19}$$

Figure 13 shows the averages over $1 \leq k \leq n$ and confidence intervals at 95 % from standard errors for the two statistics of Eqs. (18)–(19) for the eight multivariate density models. In the model acronyms (to the left of Fig. 13), GC and TC stand for Gaussian and Student t copula and SN and ST for Skew Normal and Skew t. The marginal models are indicated by Gam for Gamma (in blue) and

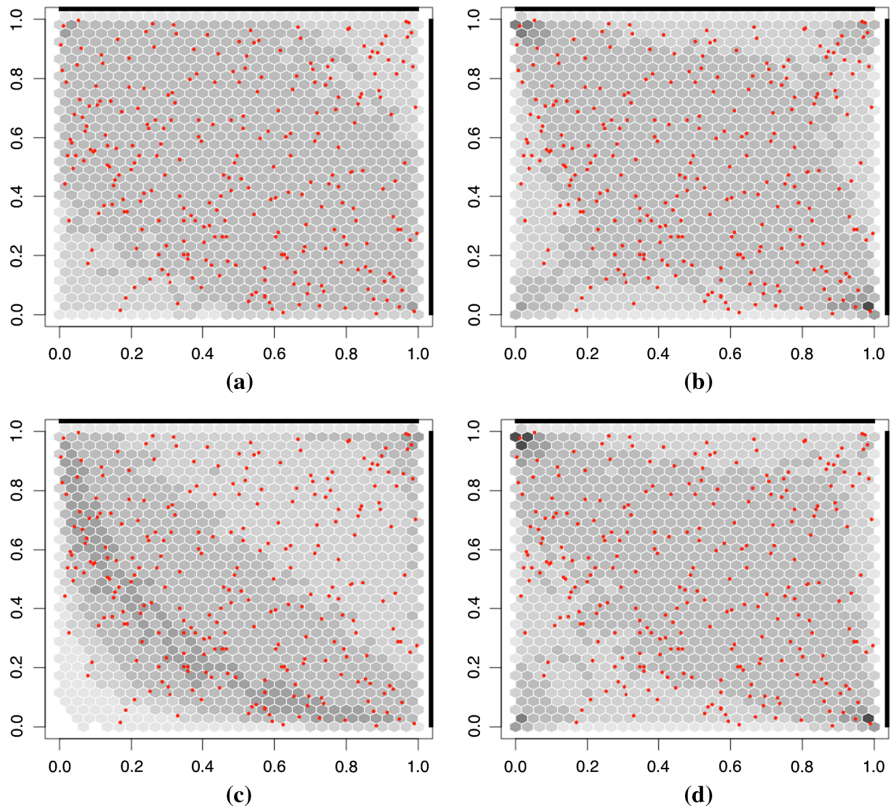


Fig. 12 Fitted bivariate copula densities for the distant pair of stations (Barre-des-Cevennes and Generargues) as estimated by bivariate hexagonal histograms on random samples of size 10^6 from the copulas associated with each fitted model. The darker the histogram bin is, the higher the density is estimated. The *dots* represent the observations. The empirical Kendall's τ is -0.067. **a** Gaussian copula. **b** Student t copula. **c** Skew Normal. **d** Skew t

LNormMix for the 2-component Log-Normal mixture (in black). The x-axis has a logarithmic scale to enhance differences between models. Smaller value of the statistics means a better performance.

The multivariate model with the Skew Normal dependence structure and 2-component Log-Normal margins (SN-LNorMix) outperforms the other seven models in terms of both goodness-of-fit statistics. In all cases but one, the performance of the multivariate models, in terms of both goodness-of-fit statistics, is improved when 2-component Log-Normal mixture margins are used instead of Gamma margins. The exception concerns the models with Skew t dependence structure that have similar performance with both types of margins. When Gamma margins are employed, all four dependence structures yield multivariate models with comparable performance. Only the Skew Normal displays a significantly better fit and only in terms of the Anderson-Darling statistic. The asymptotically dependent models (TC, ST) are not performing better than their asymptotically independent counterparts (GC, SN).

4.3 Hydrological criteria

Last, we propose to obtain complementary insight into the multivariate models by means of two hydrologically meaningful quantities: the return periods of the spatial average of flood-risk rainfall (Sect. 4.3.1) and the conditional probability that at one station, rainfall exceeds a high level given that a high level is exceeded at another station (Sect. 4.3.2).

4.3.1 Spatial average return periods

The distribution of the spatial average $\overline{X} = X_1 + \cdots + X_8/8$ and consequently of the return periods of the spatial average, involves both the margins and the dependence structure of the multivariate density models.

The empirical return periods of the observed spatial averages are computed as follows. Let $\bar{x}_{(k)} = \sum_{j=1}^{8} x_j^{(k)}/8$, be the kth largest observed spatial average, $k = 1, \ldots, 265$, and let $n_{\bar{x}} = 265/43 \approx 6.16$ be the average number of

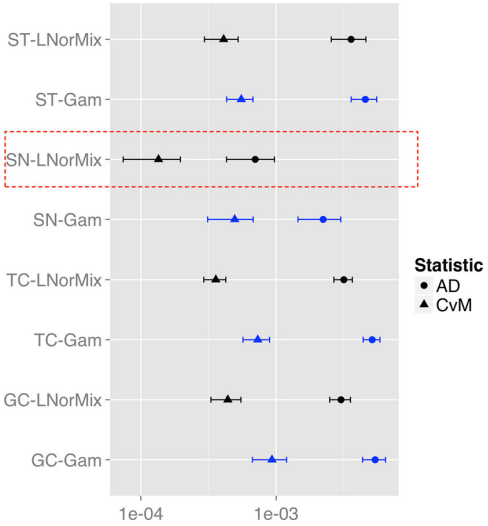


Fig. 13 Model selection based on the leave-one-out evaluation of the Cramer-Von Mises (CvM) and the Anderson-Darling (AD) goodness-of-fit statistics: average value and 95 % confidence interval are shown on a logarithmic scale. In blue, models with Gamma margins (Gam), and in black, with 2-component Log-Normal mixture margins (LNorMix). GC and TC stand for Gaussian and Student t copulas and SN and ST for Skew Normal and Skew t. The model SN-LNorMix outperforms significantly the other models

observations per year with a spatial average greater than 50 mm. Then, the empirical return period of $\bar{x}_{(k)}$ is estimated by:

$$\hat{T}_k = \frac{1}{P(\overline{X} > \bar{x}_{(k)} | \overline{X} > 50) \times n_{\bar{x}}} \tag{20}$$

where $P(\overline{X} > \bar{x}_{(k)} | \overline{X} > 50) \approx (265.5 - k)/265$ are the empirical Hazen frequencies. For example, the return period of the smallest observed spatial average (50 mm) is estimated to $\hat{T}_1 = 0.163$ year or approximately 60 days whereas the return period of the largest observed spatial average (194 mm) is estimated to $\hat{T}_{265} = 86$ years.

The theoretical return periods T_k, that is as predicted by the fitted models, of the observed spatial averages are estimated by bootstrap resampling as computing exact return periods from the multivariate models would be very involved. An 8-dimensional sample of size 10^6 was drawn from each of the eight multivariate density models ensuring that the spatial average is always greater than 50. The theoretical return periods are then estimated with Eq. (20) in which $P(\overline{X} > \bar{x}_{(k)} | \overline{X} > 50)$ is now approximated by the proportion of exceedances of $\bar{x}_{(k)}$ in the bootstrap sample of 10^6 simulated spatial averages.

Confidence intervals at 95 % are obtained for the empirical and theoretical return periods as follows. For the empirical estimates $\hat{T}_k$, since the size of the observed sample is small, bootstrap resampling is employed. To this end, 10,000 random samples of size 265 were drawn with replacement from the set of 265 8-dimensional observed rainfall so as to preserve spatial dependence. For the theoretical estimates T_k, as the sample size is large, 95 % confidence intervals can be computed from standard errors. This is done in two steps. First, standard errors for the sample proportion of exceedances of $\bar{x}_{(k)}$ are estimated as the standard deviation of the sample proportion divided by the square-root of the sample size (1000 in this case). The confidence intervals deduced for the sample proportion of exceedances are translated into confidence intervals for the return periods via Eq. (20).

Empirical and theoretical return periods in logarithmic scale are plotted against the observed spatial averages in Fig. 14. Each of the four panels is dedicated to one dependence structure in which the empirical estimates with their 95 % confidence intervals are shown as red dots surrounded by a grey band and the theoretical estimates appear as curves with confidence intervals pictured as tiny vertical bars (the model with Gamma margins is in blue and the model with 2-component Log-Normal mixture margins is in a black). The tiny vertical bars are hardly visible, they appear as small dots along the curves, this is because the confidence intervals for the theoretical estimates are very narrow. As in Figs. 11, 12, the effect of allowing for skewness can be assessed by comparing the top row with the bottom row and of allowing for asymptotic dependence by comparing left and right panels.

The Skew Normal dependence structure stands out as it is the only one which, with both types of margins, is able to reproduce accurately the return periods of the smallest return levels (from 50 to 90 mm approximately). These are under-estimated by the other three dependence structures, which means that the models see these levels as more frequent than they should.

For the largest spatial averages (beyond 125 mm), the confidence intervals of the empirical estimates are very wide and contain, most of the time, the estimates of all eight models. These spatial averages are rare events (return periods greater than 30 years) with respect to the length of the data set (43 years). For the four dependence structures, the model with 2-component Log-Normal mixture margins yields lower return periods while the model with Gamma margins provides higher return periods and thus assigns smaller probabilities to the largest observed spatial averages.

4.3.2 Conditional probability of exceedance

Conditional probabilities can be deduced from the multivariate density models using their lower dimensional

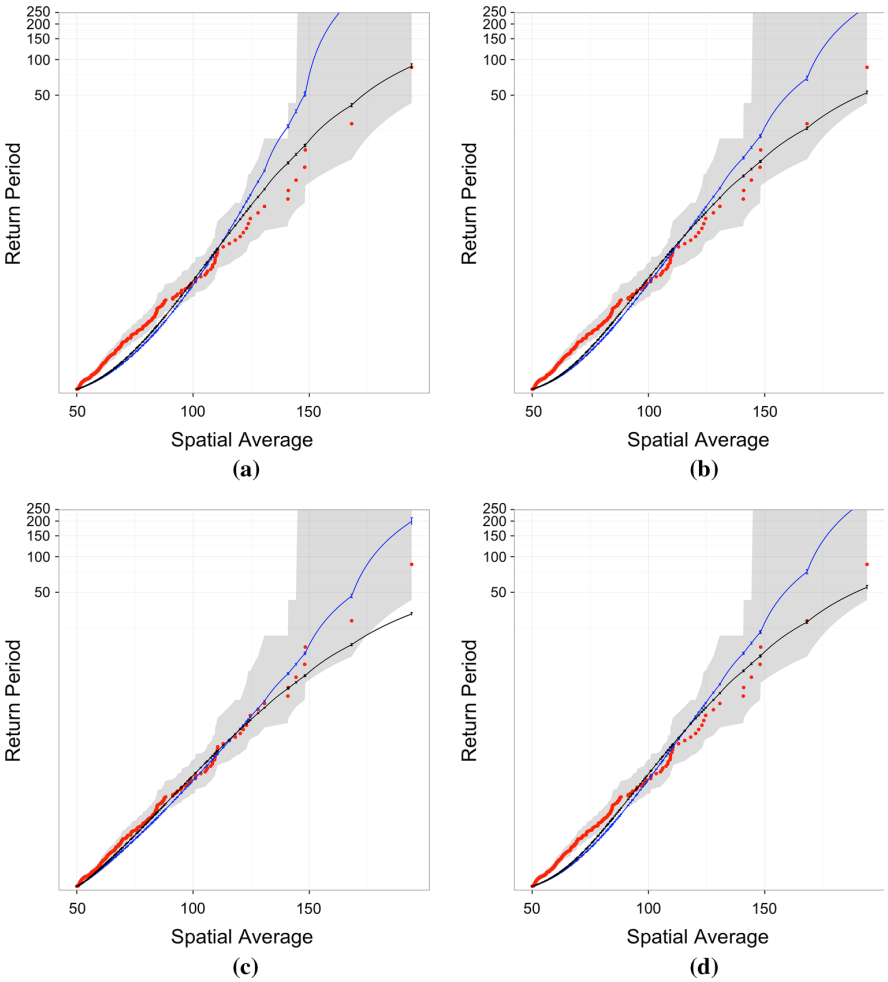


Fig. 14 Empirical (*red dots*) and theoretical return periods (*blue curve* for the Gamma margins and *black curve* for the 2-component Log-Normal mixture margins) in logarithmic scale of the observed spatial average are plotted against the observed spatial averages. The 95 % confidence band of the empirical estimates are shown in *grey* while those of the theoretical estimates are the tiny *vertical bars* along the *blue* and *black curves*. **a** Gaussian copula. **b** Student t copula. **c** Skew Normal. **d** Skew t

margins. We consider the conditional probability that rainfall at one station exceeds the at-site T-year return level given that it has exceeded the at-site T-year return level at another station in the catchment. For two stations i and j, this can be expressed as:

$$P(X_j > R_j(T)|X_i > R_i(T), \overline{X} > 50) = \frac{P(X_j > R_j(T), X_i > R_i(T)|\overline{X} > 50)}{P(X_i > R_i(T)|\overline{X} > 50)} \quad (21)$$

where $R_i(T)$ and $R_j(T)$ are the at-site T-year return levels for stations i and j that satisfy $P(X_i > R_i(T)) = P(X_j > R_j(T)) = 1/T$.

The at-site return levels are estimated individually by fitting the Generalized Pareto distribution (GPD) above a suitably high threshold (Coles 2001). For station j, the upper tail is approximated by the GPD, for all x above the threshold u_j, as follows:

$$P(X_j > x|X_j > u_j) = \left[1 + \xi_j \frac{(x-u_j)}{\beta_j}\right]_+^{-1/\xi_j} \Leftrightarrow$$
$$P(X_j > x) = P(X_j > u_j)\left[1 + \xi_j \frac{(x-u_j)}{\beta_j}\right]_+^{-1/\xi_j} \quad (22)$$

where β_j and ξ_j are the scale and shape parameters respectively of the GPD and the $+$ indicates the positive fraction, that is $z_+ = \max(z, 0)\ \forall z \in \mathbb{R}$. The threshold u_j is set to the 95 % empirical quantile of the rainfall intensities greater than 1 mm. The T-year return level at station j can be derived from Eq. (22) as:

$$R_j(T) = u_j + \frac{\beta_j}{\xi_j}\left[(T\ 365.25\ P(X_j > u_j))^{\xi_j} - 1\right], \quad (23)$$

where $P(X_j > u_j)$ is taken as the sample proportion of threshold exceedances.

Table 2 Estimated at-site T-year return levels $R_j(T)$ in mm, see Eq. (23) for three stations forming the pairs whose conditional probability estimates (Eq. (21)) are shown in Fig. 15

Station name	$R(2)$	$R(5)$	$R(10)$	$R(20)$
BARRE-DES-CEVENNES	124	157	184	214
CASSAGNAS	160	197	226	255
GENERARGUES	110	137	159	182

The at-site return level estimates for $T = 2$, 5, 10 and 20 years at three stations that form the two pairs of representative stations examined before (see Sect. 4.1.2) are shown in Table 2. The first two stations, Barre-des-Cevennes and Cassagnas, are close spatially and sit near the mountain crest while the third station named Generargues lies in the valley and has lower return levels.

The empirical and theoretical (as predicted by each fitted model) conditional probabilities of Eq. (21) are estimated as the sample proportion of the conditional exceedances. In other words, among the observations in the sample for which $R_i(T)$ is exceeded at station i, we computed the proportion for which $R_j(T)$ is also exceeded at station j. For the theoretical conditional probabilities, a sample of size 10^6 was simulated by each fitted model such that the spatial average is greater then 50. As already mentioned, we resorted to simulation to estimate the theoretical conditional probabilities because the lower dimensional margins of the Skew distributions are not easy to deduce (Azzalini 2013).

The 95 % confidence intervals are estimated in a similar way as those for the return periods of the spatial average in Sect. 4.3.1. For the empirical estimates, as the sample size is small, 95 % confidence intervals are obtained by bootstrap resampling (10,000 random samples of size 265 were drawn with replacement from the set of 265 8-dimensional observed rainfall). For the theoretical estimates, as the sample size is large, the confidence intervals were computed with the standard errors (the standard deviation of the empirical proportion divided by the square-root of the sample size which is the number of exceedances of X_i).

The estimated conditional probabilities for the eight multivariate density models are shown in Fig. 15. The pair of nearby stations Barre-des-Cevennes and Cassagnas appear in the top row and the distant pair, Barre-des-Cevennes and Generargues, in the bottom row. In both cases, Barre-des-Cevennes is the conditioning station (X_i in Eq. (21)). The left column compares the models with Gamma margins and the right column the models with 2-component Log-Normal mixture margins. Therefore, each panel depicts the results for a given pair of stations and for a given marginal model and has thus four curves with vertical error bars for each of the spatial dependence structure. The same acronyms as before are used in the legend: GC and TC for Gaussian and Student t copula and SN and ST for Skew Normal and Skew t. The empirical conditional probabilities are represented with black dots surrounded by their 95 % grey confidence band.

Given the small sample size, the empirical estimates of the conditional probabilities are unreliable. For the nearby pair, the empirical estimates provide no information for $T \geq 10$ since the 95 % confidence intervals reach the bounds [0,1], see Fig. 15a, b where the y-axis is truncated. Conversely, for the distant pair, the confidence intervals collapse to 0, also for $T \geq 10$, see Fig. 15c, d. Indeed, the numbers of exceedances of the conditioning station Barre-des-Cevennes are very low: 21, 8, 5, and 4 exceedances for T = 2, 5, 10 and 20 respectively.

As expected, since this is a monotone transformation, the choice of margins (left column versus right column in Fig. 15) does not affect the ordering of the curves or their global features (rising, declining or stabilizing). However, for the nearby pair, the estimated conditional probabilities are clearly higher with the 2-component Log-Normal mixture margins, see Fig. 15b.

Unsurprisingly, the asymptotically dependent models, t copula (TC) and Student t (ST), yield generally the highest conditional probability estimates. This is especially true for the longer return periods and for the nearby pair, see Fig. 15a, b. In contrast, the Skew Normal model, which is asymptotically independent, provides estimates that are nearly comparable to the asymptotically dependent model estimates for the distant pair, see Fig. 15c, d.

Among the two asymptotically independent models, the Skew Normal model gives higher conditional probability estimates than the Gaussian model. For the nearby pair, the empirical Kendall's τ estimate is of 0.55, and thanks to this strong correlation the Gaussian model is able to predict quite high conditional probabilities as the asymptotic independence property comes into play for much longer return periods. Conversely, for the distant pair that has an empirical Kendall's τ estimated at -0.067, the Gaussian is almost independent and predicts a conditional probability decreasing very quickly to zero.

5 Discussion and conclusion

We conducted a comparative study of eight multivariate density models (marginal model combined with a dependence structure) for flood-risk rainfall, i.e. rainfall susceptible of causing flash floods in small Mediterranean catchments. The characterization of flood-risk rainfall and in particular, of its spatial variability, is crucial to improve flash-flood understanding. The study area is the Gardon at Anduze, a representative small Mediterranean catchment of

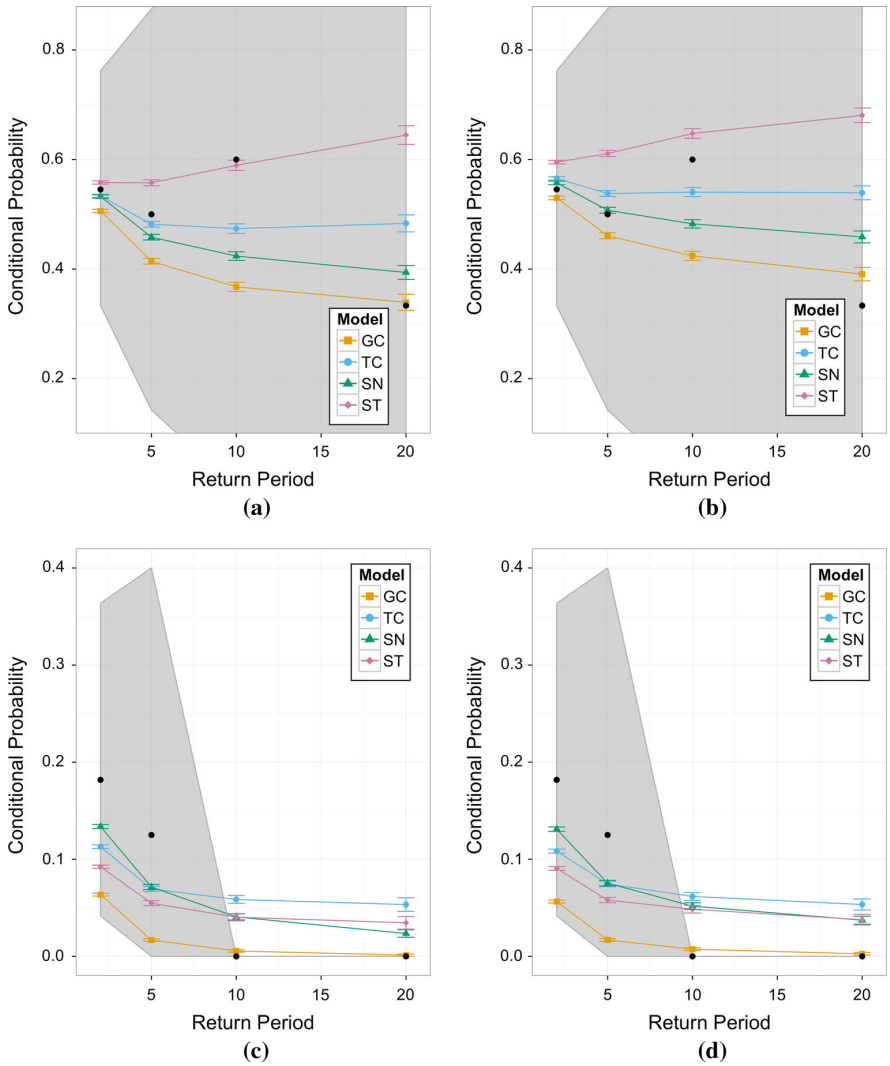


Fig. 15 Empirical (*black dots*) and theoretical (*curves*) conditional probabilities of exceedances of high thresholds expressed as at-site return levels $R_i(T)$, see Eq. (21) and Table 2. The x-axis represents the associated return period T. The 95 % confidence band of the empirical estimates is shown in *grey* while those of the theoretical estimates are the *colored vertical bars*. In the left (right) column, the models have Gamma (2-component Log-Normal mixture) margins. GC and TC stand for Gaussian and Student t copulas and SN and ST for Skew Normal and Skew t. **a** Nearby pair Cassagnas |Barre-des-Cevennes: Gamma margins. **b** Nearby pair Cassagnas |Barre-des-Cevennes: 2-component Log-Normal mixture margins. **c** Distant pair Generargues |Barre-des-Cevennes: Gamma margins. **d** Distant pair Generargues |Barre-des-Cevennes: 2-component Log-Normal mixture margins

about 545 km^2. Flood-risk rainfall is defined as rainfall with a high spatial average. For the Gardon at Anduze catchment, spatial average is considered high when it is greater than 50 mm. We used data from eight rain gauge stations at the daily time-step. The pairwise exploratory analysis revealed that the bivariate dependence varies widely from strong for nearby pairs of stations (5 km apart) to weak or zero for distant pairs (40 km apart). This confirms the strong spatial variability of flood-risk rainfall over the catchment.

Two marginal models were considered: the Gamma distribution and a mixture of Log-Normal distributions. The Gamma is a parametric model with 2 parameters that was often used to model the univariate distribution of rainfall. The Log-Normal mixture is a non-parametric model whose complexity, i.e. the number of parameters,

can increase with the size of the data set. For all eight stations, two mixture components were selected based on the BIC. As a result, the mixture has 5 parameters, for this data set. Both marginal models are light-tailed, i.e. exponential decay of the upper tail. However, the 2-component Log-Normal mixture has considerably more flexibility due to its larger number of parameters. Such a mixture can adapt, in principle, to more complex distributions caused by the presence of several sub-populations of rainfall.

Four dependence structures with different theoretical properties were included in the comparison: the Gaussian, the Student t, the Skew Normal and the Skew t. The Gaussian is symmetric and asymptotically independent (except when $|\rho| = 1$). Its parameters are the free parameters of the 8×8 correlation matrix (constrained to be symmetric and positive definite), where 8 is the number of rain gauge stations. The Student t is symmetric and asymptotically dependent (except when $\rho = -1$). The asymptotic dependence, loosely speaking, characterizes the fact that extremes, i.e. asymptotically high values, tend to occur simultaneously at different sites. The Student t has one additional parameter ν, compared to the Gaussian, called the degree-of-freedom. This parameter controls the behavior of the tails: the smaller it gets, the greater the asymptotic dependence becomes. The Skew Normal introduces asymmetry in the Gaussian. It has, in addition to the correlation matrix, a vector of skewness parameter $\boldsymbol{\alpha} \in \mathbb{R}^8$ which define the orientation of the asymmetry, see Fig. 7. The Skew Normal, like its generating distribution, is asymptotically independent but has more flexibility thanks to its eight extra parameters. The Skew t, an asymmetric version of the Student t, combines the properties of the Student t and Skew Normal: it is asymptotically dependent and more flexible than its generating distribution.

The models were included in the comparison either because they were widely used in the literature for similar applications or because they are variants of these models with different theoretical properties (non-parametric, asymmetric, asymptotically dependent). All models are relatively easy to implement thanks to `R` libraries mentioned throughout the text. The Gaussian with Gamma margins is the most parsimonious model while the Skew t with 2-component Log-Normal margins is the most complex (12 additional parameters). Moreover, we gained reasonable confidence that no multivariate mixture modeling was needed by testing for the number of components in a multivariate Gaussian mixture.

Three types of criteria were taken into account in the comparison of the multivariate density models. First in terms of *statistical inference*, we sought to evaluate if the marginal and dependence structure models independently provided a reasonable fit. As can be seen from the quantile-quantile plots in Fig. 9, the 2-component Log-Normal mixture, thanks to its greater flexibility, is able to fit all eight stations. Greater flexibility comes with greater variance as indicated by the large confidence intervals for the upper tail of the distribution. In contrast, the Gamma lacks some flexibility as it under-estimates the upper tail of the distribution for four stations, see Fig. 8. Although the four dependence structures all reproduce well the empirical Kendall's τ (see Fig. 10), they might have important differences in the fitted bivariate densities. For instance, the asymmetry of the Skew Normal appears very clearly for the pair of stations Barre-des-Cevennes/Generargues, see Fig. 12. Moreover, the effect of the asymptotic dependence can be seen for the Student t and the Skew t that have greater density in the left-top and bottom-right corners.

Second, *model selection* was achieved based on the evaluation of the Cramer-Von Mises and the Anderson-Darling statistics with a leave-one-out scheme. With such a scheme, an over-parametrized model is penalized as it will tend to fit too well the calibration sets $\mathcal{F}_{k:n-1}$, see Eq. (17), and perform poorly on the left-out observations. Therefore, the leave-one-out evaluation allows a trade-off between goodness-of-fit and complexity. In regard of this quantitative evaluation, the Skew Normal with 2-component Log-Normal mixture margins outperforms significantly the other seven models, see Fig. 13.

Third, to obtain complementary insight into the models, they were compared in terms of two *hydrologically interpretable quantities*: the return periods of the observed spatial averages and the conditional probability of exceedances of at-site return levels for two representative pairs of stations. In both cases, it is not possible to select a model based on comparisons with the empirical estimates because of the high uncertainty of these rare events. However, inter-model comparisons emphasize some differences between the dependence structures. In particular, the Skew Normal is the only dependence structure providing consistent return periods for the smaller spatial averages, see Fig. 14. In addition, despite being also asymptotically independent, the Skew Normal provides higher conditional probabilities and therefore reveals stronger dependence than the Gaussian, see Fig. 15. For the distant pair of stations, the Skew Normal is almost comparable to the asymptotically dependent models. The Gaussian yields the lowest conditional probabilities and thus is the model with the weakest spatial dependence. In contrast, the Skew t can display very strong spatial dependence, especially for the nearby pair of stations, see Fig. 14.

In conclusion, for the Gardon at Anduze catchment, the Skew Normal with 2-component Log-Normal mixture margins achieved the best fit. The increase in complexity of the mixture model for the margins with respect to the

Gamma is compensated by a significant increase in goodness-of-fit. Similarly, the asymmetry introduced by the Skew Normal is an added-value with respect to the Gaussian. In contrast, the asymptotically dependent models did not improve the fit over the asymptotically independent ones. As mentioned in Serinaldi et al. (2014), asymptotic dependence is very difficult to detect when the time series is short, as it is the case in the present work. The Gaussian, which is the benchmark model in this comparison, is not recommended for the data at hand. Even when considering the more complex 2-component Log-Normal mixture model for the margins, its performance remains significantly lower than the Skew Normal. Moreover, preliminary testing lead us to conclude that considering a multivariate mixture of Gaussians, instead of a single Gaussian, would not improve the fit.

The contributions of this work are as follows.

1. The strategy that we adopted to focus on flood-risk rainfall, the type of rainfall associated to flash-floods, allows us to tackle the most important feature multi-site stochastic generators should be able to reproduce when applied to small Mediterranean catchments. This strategy circumvents the need to build a complex stochastic model that must account for rainfall intermittency and inhomogeneity. Homogeneity is dealt with a statistical approach, namely the selection of the number of components in mixture models based on the BIC, rather than by fixing the number of components based on the seasons or the months.
2. We compared multivariate density models of increasing complexity with a different combinations of theoretical properties thanks to the decomposition into marginal and dependence structure models. We were able to determine which properties are most relevant for the data at hand. Multivariate EVT models were not included in the comparison because high dimensional models that could be easily implemented are too simplistic (e.g. Gumbel).
3. We proposed three types of criteria that serve different purposes: (i) *statistical inference* is meant to asses basic model goodness-of-fit, (ii) *model selection* serves to identify the best model and (iii) *hydrological interpretable quantities* helps to gain deeper understanding into the models that could be relevant for hydrological applications.

The perspectives for this work are centered around the development of a spatial stochastic rainfall generator adapted for flood-risk rainfall. A first step would be to study flood-risk rainfall at the hourly time-step. This would be more consistent with the response time of small Mediterranean catchments. As the daily data sets are longer and more complete than hourly data sets, a possible alternative is to rely on temporal disaggregation, such as in Allard and Bourotte (2014). A second step would be to go from a multivariate to a spatial process framework. The multivariate Gaussian extends naturally to the Gaussian process. However, no straightforward extensions to a continuous random process seem available at the moment for the Skew Normal (see Zareifard and Khaledi 2013 and references therein). A possible solution is to rely on the spatial vine copula construction proposed by Gräler (2014). Another issue is to perform the so-called *regionalization* for the margin parameters, i.e. spatial interpolation, in order to define the margins of a continuous process at every point in space. Finally, it would be interesting to evaluate, in our application, some recent flexible models from multivariate EVT such as those proposed in Salvadori and De Michele (2010) or Bacro et al. (2015) for spatial processes.

Acknowledgments This work has been has been partly supported by the StaRMIP project and the FloodScale project, funded by the French National Research Agency (ANR). The FloodScale project contributes to the HyMeX program and benefits from funding by the MISTRALS/HyMeX program (http://www.mistrals-home.org). The rainfall data are provided by the OHM-CV, an observation service certified in 2006 and funded by the Institut National des Sciences de l'Univers/Surface et Interfaces Continentales. We are thankful to all the R-package developers that we mentioned throughout the paper. We thank F. Serinaldi and two anonymous reviewers for their valuable comments, which greatly helped improve the quality of the paper.

References

Ailliot P, Allard D, Monbet V, Naveau P (2015) Stochastic weather generators: an overview of weather type models. J de la Société Française de Stat 156(1):101–113

Allard D, Bourotte M (2014) Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process. Stoch Environ Res Risk Assess 29(2):453–462

Azzalini A (2013) The skew-normal and related families, vol 3. Cambridge University Press, Cambridge

Azzalini A (2015) The R package sn: The skew-normal and skew-t distributions (version 1.2-3). Università di Padova, Italia. http://azzalini.stat.unipd.it/SN

Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J R Stat Soc 65(2):367–389

Bacro JN, Gaetan C, Toulemonde G (2015) A flexible dependence model for spatial extremes. (**in revision**)

Barancourt C, Creutin JD, Rivoirard J (1992) A method for delineating and estimating rainfall fields. Water Resour Res 28(4):1133–1144

Bárdossy A, Pegram GGS (2009) Copula based multisite model for daily precipitation simulation. Hydrol Earth Syst Sci 13(12):2299–2314

Baxevani A, Lennartsson J (2015) A spatiotemporal precipitation generator based on a censored latent Gaussian field. Water Resour Res

Beirlant J, Goegebeur Y, Segers J, Teugels J (2006) Statistics of extremes: theory and applications. Wiley, New York

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

Berg D, Aas K (2009) Models for construction of multivariate dependence—a comparison study. Eur J Financ 15(7):639–659

Blanchet J, Davison AC (2011) Spatial modeling of extreme snow depth. Ann Appl Stat 5:1699–1725

Borga M, Anagnostou EN, Blöschl G, Creutin JD (2011) Flash flood forecasting, warning and risk management: the HYDRATE project. Environ Sci Policy 14 (7):834–844. doi:10.1016/j.envsci.2011.05.017. ISSN 1462-9011. http://www.sciencedirect.com/science/article/pii/S1462901111000943. Adapting to Climate Change: Reducing Water-related Risks in Europe

Bortot P (2010) Tail dependence in bivariate skew-normal and skew-t distributions. Available online: www2.stat.unibo.it/bortot/ricerca/paper-sn-2.pdf

Bouvier C, Cisneros L, Dominguez R, Laborde J-P, Lebel T (2003) Generating rainfall fields using principal components (PC) decomposition of the covariance matrix: a case study in Mexico city. J Hydrol 278(1):107–120

Bouvier C, Ayral PA, Brunet P, Crespy A, Marchandise A, Martin C (2007) Recent advances in rainfall-runoff modelling: extrapolation to extreme floods in southern France. In: First international workshop on hydrological extremes. Observing and modelling exceptional floods and rainfalls, pp 229–238, Cosenza. FRIEND-AMHY

Braud I, Ayral PA, Bouvier C, Branger F, Delrieu G, Le JC, Nord G, Vandervaere JP, Anquetin S, Adamovic M, Andrieu J, Batiot C, Boudevillain B, Brunet P, Carreau J, Confoland A, Didon-Lescot JF, Domergue JM, Douvinet J, Dramais G, Freydier R, Gérard S, Huza J, Leblois E, Le OB, Le RB, Marchand P, Martin P, Nottale L, Patris N, Renard B, Seidel JL, Taupin JD, Vannier O, Vincendon B, Wijbrans A (2014) Multi-scale hydrometeorological observation and modelling for flash flood understanding. Hydrol Earth Syst Sci 18 (9): 3733–3761. doi:10.5194/hess-18-3733-2014. http://www.hydrol-earth-syst-sci.net/18/3733/2014/

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

Carreau J, Vrac M (2011) Stochastic downscaling of precipitation with neural network conditional mixture models. Water Resour Res 47(10)

Carreau J, Neppel L, Arnaud P, Cantet P (2013) Extreme rainfall analysis at ungauged sites in the South of France: comparison of three approaches. J de la Société Française de Stat 154(2):119–138

Ceresetti D, Ursu E, Carreau J, Anquetin S, Creutin J-D, Gardes L, Girard S, Molinie G (2012) Evaluation of classical spatial-analysis schemes of extreme rainfall. Nat Hazards Earth Syst Sci 12:3229–3240

Chandler RE, Wheater HS (2002) Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland. Water Resour Res 38(10):10

Cleveland WS (1981) LOWESS : a program for smoothing scatterplots by robust locally weighted regression. Am Stat 35:54

Coles S (2001) An introduction to statistical modeling of extreme values., Springer series in statisticsSpringer, New York

Coles S, Heffernan J, Tawn J (1999) Dependence measures for extreme value analyses. Extremes 2(4):339–365

Delrieu G, Nicol J, Yates E, Kirstetter P-E, Creutin J-D, Anquetin S, Obled C, Saulnier G-M, Ducrocq V, Gaume E, Payrastre O, Andrieu H, Ayral P-A, Bouvier C, Neppel L, Livet M, Lang M, du Châtelet JP, Walpersdorf A, Wobrock W (2005) The catastrophic flash-flood event of 8–9 september 2002 in the Gard region, France: a first case study for the Cévennes-Vivarais Mediterranean Hydrometeorological Observatory. J Hydrometeorol 6(1):34–52

Demarta S, McNeil AJ (2005) The t copula and related copulas. Int Stat Rev 73(1):111–129

Ducrocq V, Nuissier O, Ricard D, Lebeaupin C, Thouvenin T (2008) A numerical study of three catastrophic precipitating events over southern France. II: mesoscale triggering and stationarity factors. Q J R Meteorol Soc 134(630):131–145

Dupuis DJ, Tawn JA (2001) Effects of mis-specification in bivariate extreme value problems. Extremes 4(4):315–330

Dupuis DJ (2007) Using copulas in hydrology: benefits, cautions, and issues. J Hydrol Eng 12(4):381–393

Embrechts P, McNeil A, Straumann D (2002) Correlation and dependence in risk management: properties and pitfalls. Risk Manag, pp 176–223

Flecher C, Naveau P, Allard D, Brisson N (2010) A stochastic daily weather generator for skewed data. Water Resour Res 46(7)

Frayler C, Raftery AE (1999) MCLUST: software for model-based cluster analysis. J Classif 16:297–306

Garavaglia F, Gailhard J, Paquet E, Lang M, Garçon R, Bernardara P (2010) Introducing a rainfall compound distribution model based on weather patterns sub-sampling. Hydrol Earth Syst Sci Discuss 14:951

Gardes L, Girard S (2010) Conditional extremes from heavy-tailed distributions: an application to the estimation of extreme rainfall return levels. Extremes 13(2):177–204

Gaume E, Bain V, Bernardara P, Newinger O, Barbuc M, Bateman A, Blaškovičová L, Blöschl G, Borga M, Dumitrescu A et al (2009) A compilation of data on European flash floods. J Hydrol 367(1):70–78

Genest C, Favre A-C (2007) Everything you always wanted to know about copula modeling but were afraid to ask. J Hydrol Eng 12(4):347–368

Genest C, Huang W, Dufour J-M (2013) A regularized goodness-of-fit test for copulas. J de la Société Française de Statistique & revue de statistique appliquée 154(1):64–77

Gilleland E, Katz RW (2011) New software to analyze how extremes change over time. EOS, Trans Am Geophys Union 92(2):13–14

Gräler B (2014) Modelling skewed spatial random fields through the spatial vine copula. Spat Stat 10:87–102

Guillot G, Lebel T (1999) Approximation of Sahelian rainfall fields with meta-gaussian random functions. Stoch Environ Res Risk Assess 13(1–2):113–130

Hughes JP, Guttorp P, Charles SP (1999) A non-homogeneous hidden markov model for precipitation occurrence. Appl Stat 48:15–30

Joe H (1997) Multivariate models and multivariate dependence concepts, vol 73. CRC Press, Boca Raton

Kleiber W, Katz RW, Rajagopalan B (2012) Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. Water Resour Res 48(1):1

Kojadinovic I, Yan J (2010) Modeling multivariate distributions with continuous margins using the copula R package. J Stat Softw 34 (9): 1–20. http://www.jstatsoft.org/v34/i09/

Kollo T, Selart A, Visk H (2013) From multivariate skewed distributions to copulas. In: Combinatorial matrix theory and generalized inverses of matrices. Springer, New York, pp 63–72

Lebel T, Laborde JP (1988) A geostatistical approach for areal rainfall statistics assessment. Stoch Hydrol Hydraul 2(4):245–261

Leblois E, Creutin JD (2013) Space-time simulation of intermittent rainfall with prescribed advection field: adaptation of the turning

band method. Water Resour Res, pp n/a–n/a. ISSN 1944-7973. doi:10.1002/wrcr.20190
Lennartsson J, Baxevani A, Chen D (2008) Modelling precipitation in Sweden using multiple step markov chains and a composite model. J Hydrol 363(1):42–59
Li C, Singh VP, Mishra AK (2012) Simulation of the entire range of daily precipitation using a hybrid probability distribution. Water Resour Res 48(3)
Lobligeois F, Andréassian V, Perrin C, Tabary P, Loumagne C (2014) When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. Hydrol Earth Syst Sci 18 (2): 575–594. doi:10.5194/hess-18-575-2014. https://hal.archives-ouvertes.fr/hal-00952657
Neppel L, Pujol N, Sabatier R (2011) A multivariate regional test for detection of trends in extreme rainfall: the case of extreme daily rainfall in the French Mediterranean area. Adv Geosci 26(26):145–148
Patil SD, Wigington Jr. PJ, Leibowitz SG, Sproles EA, Comeleo RL (2014) How does spatial variability of climate affect catchment streamflow predictions? J Hydrol 517(0): 135–145. ISSN 0022-1694. doi:10.1016/j.jhydrol.2014.05.017. http://www.sciencedirect.com/science/article/pii/S0022169414003710
Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat 3:119–131
Poon S-H, Rockinger M, Tawn J (2004) Extreme value dependence in financial markets: diagnostics, models, and financial implications. Rev Financ Stud 17(2):581–610
Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
Sabourin A, Naveau P (2014) Bayesian Dirichlet mixture model for multivariate extremes: a re-parametrization. Comput Stat Data Anal 71:542–567
Salvadori G, De Michele C (2010) Multivariate multiparameter extreme value models and return periods: a copula approach. Water Resour Res 46(10)
Schoelzel C, Friederichs P (2008) Multivariate non-normally distributed random variables in climate research-introduction to the copula approach. Nonlinear Process Geophys 15(5):761–772
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Serinaldi F (2009) Copula-based mixed models for bivariate rainfall data: an empirical study in regression perspective. Stoch Environ Res Risk Assess 23(5):677–693
Serinaldi F, Kilsby CG (2014) Simulating daily rainfall fields over large areas for collective risk estimation. J Hydrol 512:285–302
Serinaldi F, Bárdossy A, Kilsby CG (2014) Upper tail dependence in rainfall extremes: would we know it if we saw it? Stoch Environ Res Risk Assess, pp 1–23
Sklar M (1959) Fonctions de répartition à n dimensions et leurs marges. Université Paris 8
Stephenson AG (2002) Evd: extreme value distributions. R News 2 (2): 0, June 2002. http://CRAN.R-project.org/doc/Rnews/
Thibaud E, Mutzner R, Davison AC (2013) Threshold modeling of extreme spatial rainfall. Water Resour Res 49(8):4633–4644
Thompson CS, Thomson PJ, Zheng X (2007) Fitting a multisite daily rainfall model to New Zealand data. J Hydrol 340:25–39
Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0
Vischel T, Lebel T, Massuel S, Cappelaere B (2009) Conditional simulation schemes of rain fields and their application to rainfall-runoff modeling studies in the Sahel. J Hydrol 375:273–286
Vrac M, Naveau P, Drobinski P (2007) Modeling pairwise dependencies in precipitation intensities. Nonlinear Process Geophys 14(6):789–797
Wilks DS (1998) Multisite generalization of a daily stochastic precipitation generation model. J Hydrol 210(1):178–191
Zareifard H, Khaledi MJ (2013) Non-Gaussian modeling of spatial data using scale mixing of a unified skew Gaussian process. J Multivar Anal 114:16–28