Centre Eau Terre Environnement

Génération stochastique de séries temporelles de pluie pour l'évaluation des risques

Antoine Chapon

Doctorat en Sciences de l'Eau

Examinateur externe 1 : Félix Francés

Examinateur externe 2 : Thomas Opitz

Prédisent du jury : André St-Hilaire

Directeur de thèse : Taha B. M. J. Ouarda

date de dépôt : 3 septembre 2025

Je tiens à remercier Taha pour son encadrement, pour m'avoir fait confiance et laisser le temps d'essayer de faire des choses originales. En dehors de la recherche, je le remercie aussi pour ses qualités humaines. Je remercie Nathalie pour son encadrement à distance depuis la France, et pour toujours avoir été disponible et encourageante. Je remercie également les autres personnes de l'ASNR et de l'INRS que j'ai rencontré, pour leur bienveillance et pour avoir rendu ces deux endroits très agréables.

Merci aux membres du jury, André St-Hilaire, Félix Francés et Thomas Opitz, pour avoir accepté d'évaluer ma thèse.

J'ai eu la chance de rencontrer des personnes formidables pendant mes années à Québec. Dans une petite liste non-exhaustive, je tiens à remercier en particulier Enzo, Bárbara, Freddy, Pia, Lorenzo, Andrea, Victor, Hamza, Simon, Savannah, Pupu et Maÿ.

Enfin, je remercie ma famille pour leurs encouragements et leur soutien. Je remercie ma mère pour avoir été la première personne à me faire confiance, sans qui je n'aurais pas essayé de faire de la recherche.

Résumé

Les pluies extrêmes sont la cause principale des inondations. Il est donc nécessaire de modéliser ce phénomène afin d'en estimer le risque. Comme il s'agit d'évènements rare, cette modélisation passe par des méthodes statistiques, avec notamment la théorie des valeurs extrêmes.

Les méthodes et classes de modèles appliqués aux pluies extrêmes sont nombreuses. Cette thèse s'intéresse aux générateurs stochastiques temporelles, qui permettent de reproduire les patrons statistiques d'une chronique de pluie observée dans des simulations. Toutes les valeurs de pluies sont considérées par ces modèles, allant des valeurs nulles sans pluie aux valeurs extrêmes. Ces modèles doivent notamment capturer le comportement asymptotique des valeurs les plus rares, afin d'être adapté à l'estimation du risque.

Deux générateurs stochastiques de séries temporelles de pluie sont développés. Leur point commun est l'utilisation d'une distribution Pareto généralisée étendue (EGP) modélisant toutes les valeurs non-nulles de pluie. Cette distribution est conditionnée aux valeurs passées afin de reproduire la dépendance temporelle des valeurs non-nulles.

Un élément séparé de la distribution EGP modélise l'intermittence des valeurs nulles et non-nulles de pluie. Le premier générateur stochastique est paramétrique, avec une EGP définie par distributions tronquées et une copule temporelle pour la dépendance aux valeurs passées. Ce premier générateur stochastique est appliqué à une série chronologique de pluie horaire. Un second modèle utilise l'apprentissage profond, avec une distribution construite par réseau de neurones afin de définir une EGP conditionnelle aux valeurs passées. Ce second modèle est appliqué à des données de pluie journalière. Ces deux modèles reproduisent dans des simulations la plupart des patrons statistiques important pour l'estimation du risque. Ces patrons sont la distribution marginale, la dépendance temporelle, l'intermittence, la variabilité annuelle et le comportement asymptotique des valeurs hautes.

Un troisième modèle Bayésien concerne l'imputation de valeurs manquantes dans une série temporelle, via une approche de régionalisation, pour des données au pas de temps irrégulier correspondant au principal cycle de marée (environ 12 heures). Les développements de ce modèle sont partiellement réutilisés dans le générateur stochastique paramétrique.

Mots-clés : pluie ; générateur stochastique ; séries temporelles ; valeurs extrêmes

Abstract

Extreme rainfalls are the main cause of floods. It is therefore necessary to model this phenomenon, so as to estimate the associated risks. Since rare events are concerned, this modelling uses statistical methods, in particular the extreme value theory.

The methods and classes of models applied to extreme rainfall are numerous. This thesis focuses on temporal stochastic generators, which reproduce the statistical patterns of an observed rainfall time series in simulations. All the rainfall values are considered by these models, from the dry time steps to the upper extreme values.

Two stochastic generators for rainfall time series are developed. Their common point is the usage of an extended generalized Pareto (EGP) distribution for all the non-zero rainfall values. This distribution is conditioned on past values, to reproduce the temporal dependence of non-zero values. An element separated from the EGP models the intermittency of zero and non-zero values. The first stochastic generator is parametric, with an EGP defined by truncated distribution and a temporal copula modelling the dependence to past values. This first stochastic generator is applied to a time series of hourly rainfall. The second model uses deep learning, with a distribution built from a neural network to define an EGP conditioned on past values. This second model is applied to daily rainfall data. These two models reproduce in simulations most statistical patterns relevant to risk estimation. These patterns are the marginal distribution, the temporal dependence, the intermittency, the yearly variability, and the asymptotic behaviour of upper values.

A third Bayesian model concerns the imputation of missing values in a time series, via a regionalization approach, for data with an irregular time step, corresponding to the main tidal cycle (approximately 12 hours). Developments of this model are partially reused in the parametric stochastic generator.

Keywords : rainfall ; stochastic generator ; time series ; extreme values

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction générale

1.1 Contexte

Le sujet de thèse porte sue la modélisation statistique des précipitations, incluant les précipitations extrêmes. La motivation pour cette modélisation est le fait que les pluies extrêmes sont la cause principale du risque d'inondation. Le risque est défini comme la combinaison d'un aléa et d'une vulnérabilité (Cardona et al. 2012). Dans le cas des inondations, l'aléa principal est les pluies extrêmes, mais peuvent se combiner à d'autres aléas tels que par exemple les remontées de nappes, les surcotes marines, ou les crues de rivières. Une vulnérabilité représente des conséquences négatives pour des personnes ou biens lorsqu'un certain aléa arrive, par exemple l'inondation d'une zone habitée lorsqu'un certain niveau d'eau est atteint. Il n'y a donc pas de risque s'il n'existe pas de vulnérabilité. Le risque dans son entièreté dépasse le cadre de la thèse, qui ne concerne que la modélisation de l'aléa, mais il justifie la modélisation des pluies extrêmes.

La thèse est co-financée et co-encadrée par l'Institut national de la recherche scientifique (INRS) au Québec et l'Autorité de sûreté nucléaire et de radioprotection (ASNR) en France. L'INRS a une expertise dans la modélisation statistique des extrêmes hydroclimatiques et a développé des approches innovantes dans ce domaine. L'ASNR travaille aussi sur les extrêmes environnementaux et s'intéresse en particulier à l'estimation des risques auxquels sont soumises les centrales nucléaires françaises. Dans le cas des pluies, son rôle est de s'assurer que les centrales soient suffisamment protégées pour faire face à des évènements rares, ayant par exemple une chance sur cent ou mille d'occurrence annuelle. Les modèles développés dans la thèse sont donc appliqués à des données en

France.

La théorie des valeurs extrêmes donne un cadre statistique permettant l'extrapolation des probabilités des valeurs hautes d'une variable à des niveaux non-observés. Cette branche des statistiques est donc utilisée dans le cadre de la thèse. Les distributions des valeurs extrêmes sont au cœur de cette théorie, avec la distribution d'extremum généralisée (GEV) considérant les extrêmes comme les plus hautes valeurs par long blocs de temps (maxima par blocs), et la distribution Pareto généralisée (GP) considérant les extrêmes comme les valeurs indépendantes dépassant un seuil haut. Ces deux distribution ont de nombreuses extensions, par exemple pour une meilleur utilisation des données, pour la prise en compte de la non-stationnarité, ou pour une modélisation multivariée (Coles 2001). La pluie est une variable pour laquelle la durée des évènements extrêmes contribue au risque, en plus de l'intensité, ce qui a amené au développement du modèle intensité-durée-fréquence (IDF) (Koutsoyiannis et al. 1998), qui est une extension de la distribution GEV.

Du point de vue du contenu en information des données, un modèle statistique capable d'utiliser toutes les observations est préférable, car il permet de chercher les patrons statistique d'intérêt dans toute l'information disponible. Dans le cas particulier des extrêmes, les patrons d'intérêt sont a priori contenus dans une fraction des observations, ce qui justifie les modèles GEV et GP n'utilisant que les valeurs définies comme extrêmes. Cependant, malgré le fait que l'information des valeurs extrêmes est beaucoup plus importante que celle des valeurs non-extrêmes, ces deux modèles classiques forcent l'hypothèse qu'il n'y a pas d'information d'intérêt dans la majorité des observations, et qu'il est possible de discerner quelles observations contiennent l'information utile (i.e. les maxima par blocs ou les dépassements d'un seuil). L'approche par blocs ne concerne nécessairement qu'un sous-ensemble des données, tandis que l'approche par seuil peut être étendue afin d'intégrer les valeurs non-extrêmes en plus des extrêmes.

La problématique de la thèse était initialement de proposer un nouveau modèle pour les extrêmes de pluie, sans a priori sur les méthodes utilisées. Une possibilité aurait été d'utiliser le modèle IDF comme base, mais cela n'aurait pas permis d'utiliser toute l'information disponible dans les données de pluie. Le modèle GP peut être étendu aux observations non-extrêmes par l'approche Pareto généralisée étendue (EGP) (Papastathopoulos et Tawn 2013), jusqu'à intégrer toutes les observations dans le cas d'une variable positive comme la pluie, ce qui revient à ne plus avoir de seuil (Naveau et al. 2016). De plus, la dépendance temporelle peut être modélisée afin de retirer l'hypothèse d'indépendance des dépassements du seuil. Avec ces deux extensions du modèle GP, toute la série temporelle des observations est modélisée, et il est possible de simuler de nouvelles séries

temporelles, ce qui fait que le modèle est un générateur stochastique (GS). Donc l'utilisation d'une distribution EGP dans un GS peut être vue comme une extension de l'approche par seuil. Pour qu'un tel GS soit adapté aux valeurs extrêmes, et puisse être considéré comme une alternative au modèle IDF, la dépendance temporelle des extrêmes doit être modélisée, afin que les simulation reproduisent la relation entre intensité et durée des évènements les plus rares. Avec ces considérations, il a été décidé tôt dans la thèse de travailler sur les GSs. La problématique est donc de proposer un GS pour une série temporelle de pluie modélisant toutes les observations, mais avec une distribution des valeurs extrêmes.

Le reste de l'introduction générale présente dans la section 1.2 une revue de littérature sur la modélisation des extrêmes de pluie, les copules, les GSs pour la pluie, la modélisation de l'intermittence de la pluie, les distributions par apprentissage profond, et le lien entre les modèles prédictifs et les générateurs stochastiques. La section 1.3 présente la structure de la thèse et un résumé des articles.

1.2 Revue de littérature

1.2.1 Estimation des risques

L'estimation du risque est le fait de mettre une probabilité sur un aléa rare auquel est soumis une vulnérabilité. Comme cette estimation repose sur une distribution des valeurs extrêmes, le risque est exprimé par des quantiles de cette distribution.

Dans le cas non-stationnaire, le quantile d'une probabilité donnée change dans le temps, en fonction des covariables temporelles. Parey et al. (2010) ont défini des niveaux de retour pour des modèles variant en fonction du temps. Dans le cas de modèle non-stationnaire variant en fonction d'une covariable stochastique, tel qu'un indice climatique, la probabilité de non-dépassement dépend de la valeur de la covariable (Ouarda et al. 2019), mais pour une application d'ingénierie il faudrait alors choisir une valeur de la covariable.

Dans le cas multivarié, différentes définitions d'un quantile sont possible, avec un risque estimé très variable d'une définition à l'autre. Il est alors important de choisir la définition adaptée aux variables modélisée (Brunner et al. 2016).

D'après Serinaldi (2015), il est préférable d'estimer le risque à partir de la probabilité de dépassement d'une variable de conception sur une période donnée, par exemple la probabilité qu'un certain niveau d'eau soit atteint au cours de la période d'utilisation d'un ouvrage de protection, plutôt que par un quantile. Estimer le risque par simulation règle les différents problèmes liés aux quantiles, à

la fois dans le cas de la non-stationnarité temporelle et des modèles multivariés. La simulation permet aussi de prendre en compte l'incertitude, en estimant le risque à partir de nombreuses répliques des simulations.

La définition du risque par simulation est compatible avec le concept de variable de conception, car le risque peut être exprimé de manière probabiliste comme le ratio de simulations conduisant à un échec du système, comme par exemple le dépassement d'un seuil. De plus, les simulations d'une variable telle que la pluie peuvent être utilisées en entrée d'un modèle la transformant en débit, pour l'évaluation du risque d'inondation (Toulemonde et al. 2020).

1.2.2 Distributions des valeurs extrêmes de pluie

Les valeurs extrêmes de pluie peuvent être modélisées avec l'approche par bloc maxima via la distribution GEV, ou par l'approche par dépassement de seuil avec la distribution GP.

La distribution GEV a été étendue par Koutsoyiannis et al. (1998) pour faire dépendre la probabilité et l'intensité d'un évènement à sa durée. Le modèle d-GEV est donné par

$$G(z, d|\tilde{\mu}, \sigma_0, \xi, \theta, \eta) = \exp\left[-\left\{1 + \xi\left(\frac{z}{\sigma_0(d+\theta)^{-\eta}} - \tilde{\mu}\right)\right\}^{-1/\xi}\right], \tag{1.1}$$

où z sont les maxima des blocs de temps de durée d, $\tilde{\mu} > 0$, $\sigma_0 > 0$, $\xi \neq 0$, $\theta \geq 0$ et $\eta \in (0, 1]$ (Ulrich et al. 2020). Les quantiles du modèle sont données par des courbes IDF, faisant correspondre la probabilité et l'intensité d'un évènement à sa durée. Des formes non-stationnaires de ce modèles ont été développées. Sarhadi et Soulis (2017) ont développé un modèle d-GEV dont les paramètres de localisation et d'échelle dépendent du temps, avec par exemple $\mu(t) = \mu_0 + \mu_1 t$, où t est un vecteur représentant le temps. Ulrich et al. (2020) ont défini une distribution d-GEV dont les paramètres varient en fonction de covariables spatiales, via des modèles linéaires généralisés de la forme

$$\theta = s\left(\theta_0 + \sum_{\forall i} \alpha_i x_i\right), \tag{1.2}$$

où le paramètre θ dépend des covariables x_i en fonction des coefficients α_i, de l'ordonnée à l'origine θ_0 et d'une fonction de lien $s(\cdot)$.

Les valeurs extrêmes définies comme les dépassements d'un seuil haut sont modélisées par la distri-

bution GP, dont la fonction de probabilité cumulée est

$$G(y) = 1 - \left(1 + \frac{\xi\, y}{\sigma}\right)^{-1/\xi}. \tag{1.3}$$

Le seuil peut être abaissé à un niveau sub-extremal en étendant la distribution par une transformation intégrale de probabilité (probability integral transform) (PIT), définissant ainsi une distribution EGP (Papastathopoulos et Tawn 2013). La fonction de probabilité cumulée d'une distribution EGP est donnée par

$$W(y) = F(G(y)), \tag{1.4}$$

où $F(\cdot)$ est la PIT au support $[0, 1]$. La figure 1.1 présente un exemple de la densité d'une distribution EGP, construite à partir d'une PIT et de la GP. Cette figure montre le cas classique d'une EGP pour la pluie, où la PIT va augmenter la densité pour les plus faibles valeurs.

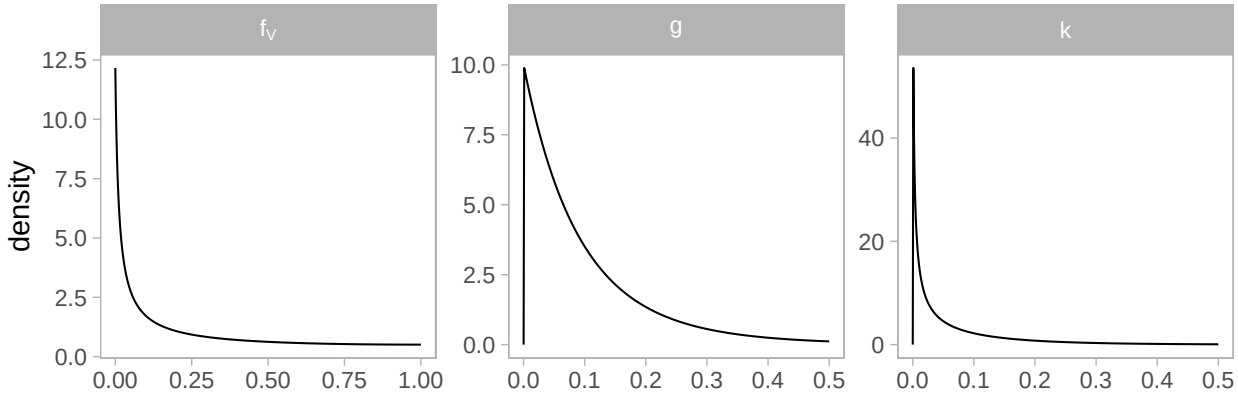


Figure 1.1 – Densité d'une PIT au support $[0, 1]$ (gauche), densité de la distribution GP (centre), densité de la distribution EGP résultant de la PIT et de la GP (droite). Ici la forme de la PIT serait adaptée à des valeurs de pluie, augmentant la densité pour les plus faibles valeurs (vers 0) afin d'abaisser le seuil, potentiellement jusqu'à 0.

Dans le cas de la pluie qui est une variable positive, abaisser le seuil de l'EGP à 0 donne une distribution pour toutes les valeurs non-nulles, allant des valeurs basses à extrêmes (Naveau et al. 2016). La fonction la plus utilisée comme PIT est la fonction puissance, donnant la distribution $W(y) = G(y)^{\kappa}$ avec $\kappa > 0$. Gamet et Jalbert (2022) ont défini des distribution EGP pour la pluie à partir de distribution tronquées. Tencaliec et al. (2020) ont défini une distribution EGP semi-paramétrique pour la pluie, en approximant la PIT par des polynômes de Bernstein.

La différence entre les deux approches par d-GEV et par EGP est que la première prend en compte la durée des évènements mais ne modélise qu'un sous-ensemble des données via les maxima par bloc, tandis que la seconde approche ne modélise pas la durée mais permet d'obtenir une distribution pour toutes les valeurs non-nulles de pluie.

Ces distributions permettent d'extrapoler les valeurs de pluie à des niveaux non-observés. D'après ces modèles, il n'y a pas de valeur maximale que peut prendre la pluie quand $\xi > 0$. Le paramètre de forme ξ déterminant la forme de la queue de la distribution est positif pour la pluie (Serinaldi et Kilsby 2014), ce qui correspond à une queue "lourde", donnant de plus fortes probabilités aux valeurs les plus hautes qu'une distribution à queue exponentielle.

1.2.3 Structure de dépendance par copules

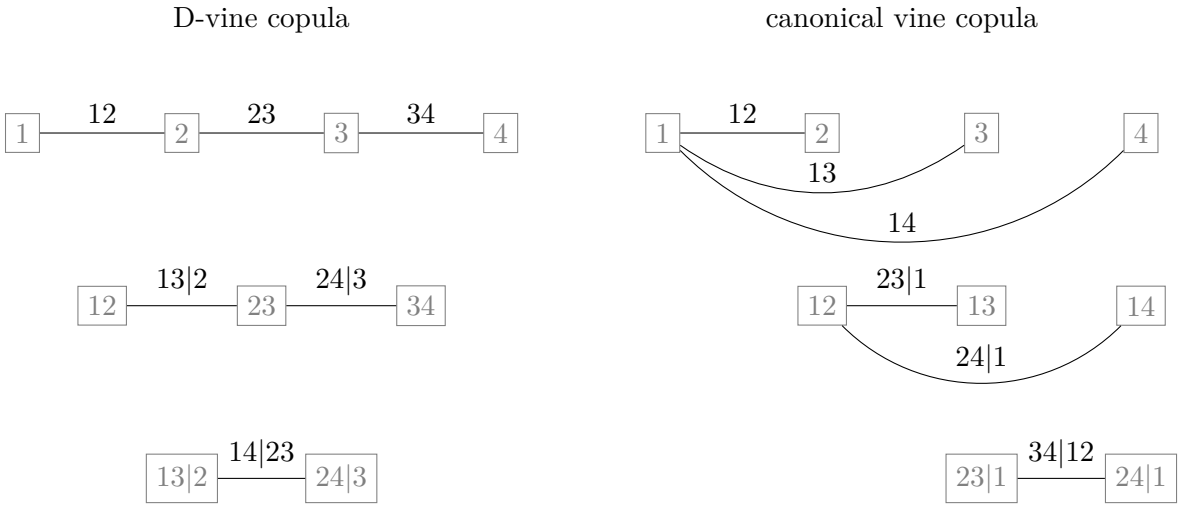


Figure 1.2 – Structures d'une copule de type D-vine (gauche) et d'une copule en vigne canonique (droite) pour quatre dimensions. Les traits reliant les boites représentent les copules bivariées composant la copule en vigne. Le nom de chaque copule bivariée indique quelles deux dimensions sont reliées et à quelles dimensions la copule bivariée est éventuellement conditionnée.

Les copules sont des distribution multivariées permettant de séparer les distributions marginales et la structure de dépendance (Sklar 1959). La fonction de probabilité cumulée d'une distribution en n dimensions peut être formulée par

$$F(\boldsymbol{x}) = C\{F_1(x_1), \dots F_n(x_n)\}, \tag{1.5}$$

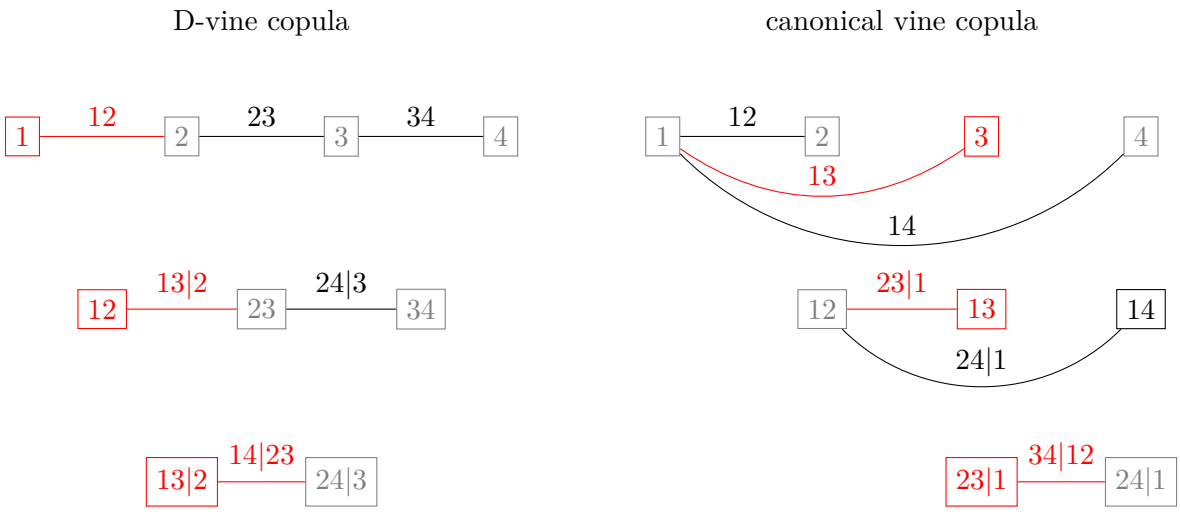


FIGURE 1.3 – Exemple de sous-ensemble des vine copulas de la figure 1.2. Les éléments en rouge correspondent aux dimensions retirées. Pour une copule de type D-vine (gauche), le dimensions peuvent être retirées à partir des bords de la structure (ici la première dimensions. Pour une copule en vigne canonique (droite), les dernières dimensions peuvent être retirées (ici l'avant dernière troisième dimension).

où F_1 est la distribution marginale de la première dimension et C est la fonction de probabilité cumulée de la copule au support $[0,1]^n$ et aux marges uniformes. Le copules sont beaucoup utilisées pour modéliser des distributions multivariées en hydrologie (TOOTOONCHI et al. 2022). Leur intérêt vient de la flexibilité dans la modélisation de la structure de dépendance. Dans le cas des valeurs extrêmes, les copules permettent de modéliser l'éventuelle dépendance ou indépendance asymptotique entre les dimensions, quelque soit les distributions marginales.

Dans le cas d'une distribution en deux dimensions, la dépendance asymptotique (ou dépendance de queues) correspond à la dépendance entre les valeurs extrêmes basses et hautes des deux variables (EMBRECHTS et al. 2003). Ce concept s'étend au delà de deux dimensions, et est directement une propriété des copules. Des variables peuvent être dépendantes pour le cœur de la distribution mais indépendantes dans les queues, comme c'est le cas pour la distribution normale multivariée. Il y a alors indépendance asymptotique, car plus les valeurs des différentes dimensions sont extrêmes et moins elles sont dépendantes. À l'inverse, il y a dépendance asymptotique si les valeurs extrêmes des différentes dimensions sont dépendantes, comme c'est le cas par exemple pour la distribution t de Student multivariée.

De nombreuses familles de copules en deux dimensions existent (Joe 2015), mais le choix devient plus limité à partir de trois dimensions. La plupart des copules paramétriques n'ont pas d'extension au delà de deux dimensions, et celles généralisables à n dimensions perdent en flexibilité. Une copule normale en trois ou plus dimensions force l'indépendance asymptotique entre toutes les dimensions. À l'inverse, une copule t de Student force la dépendance asymptotique. Ces deux copules généralisables à n dimensions ne permettent donc pas de modéliser une distribution pour laquelle certaines dimensions seraient asymptotiquement indépendantes et d'autres dépendantes.

Pour retrouver la flexibilité des copules bivariées en trois dimensions ou plus, Bedford et Cooke (2002) ont défini les vine copulas comme des copules en n dimensions, composées de $n(n-1)/2$ copules bivariées. La densité d'une la copule en vigne en trois dimensions est donnée par

$$c_{123}(x_1, x_2, x_3) = c_{12}\{F(x_1), F(x_2)\}\; c_{23}\{F(x_2), F(x_3)\}\; c_{13|2}\{F(x_1|x_2), F(x_3|x_2)\}, \tag{1.6}$$

où c_{12} et c_{23} sont des copules bivariées inconditionnelles tandis que $c_{13|2}$ et une copule bivariée entre les dimensions 1 et 3, conditionnée à la dimension 2. Les distributions marginales conditionnels apparaissant dans la copule en vigne sont calculées en dérivant ses copules bivariées, avec

$$F(x_1|x_2) = \frac{\partial\, C_{12}\{F(x_1), F(x_2)\}}{\partial\, F(x_2)}. \tag{1.7}$$

À partir de quatre dimensions, différentes structures de copule en vigne sont possibles, en fonction des dimensions reliées par des copules bivariées conditionnées à d'autres dimensions ou inconditionnelles. Les deux structures les plus régulières sont les copule de type D-vine et des canonical vine copulas (Aas et al. 2009). La figure 1.2 présente les graphs de ces deux structures pour quatre dimensions. Dans les deux cas, la copule en vigne est organisées en trois "arbres", ici représentés verticalement par les rangées. Les boites du premier arbre en haut représentent les quatre dimensions, mais les boites des arbres suivantes ne sont plus qu'une aide pour écrire la structure de la copule. Les traits reliant les boites représentent les copules bivariées composant la copule en vigne, avec des copules conditionnelles à partir du deuxième arbre. La différence entre les deux structures concerne quelles dimensions sont reliées directement par une copule bivariée dans le premier arbre, et quelles dimensions conditionnent les copules des arbres suivants. Dans le cas de la copule de type D-vine, chaque dimension est reliée à deux autres, ou à une seule pour la première et dernière dimensions aux "bords" de la structure (dimensions 1 et 4 sur la figure 1.2). Dans le cas de la copule en vigne canonique, la première dimension est reliée à toute les autres dans le premier arbre, et

chaque copule des arbres suivants sera conditionnée par la première dimension.

Les vine copulas sont des modèles emboités (nested models), dont l'unité de base sont les copules en deux dimensions. Cette propriété est exploitée par Hasler et al. (2018) dans un modèle d'imputation de données manquantes basée sur une copule de type D-vine à dimension variable. Dépendamment de la structure d'une copule en vigne, certaines de ses dimensions peuvent être retirées tout en conservant une structure valide pour les dimensions restantes. La condition pour que la structure reste valide est que les copules en deux dimensions restantes ne soient pas conditionnées aux dimensions retirées. Dans le cas d'une copule de type D-vine, les premières et dernières dimensions aux bords de la structure peuvent être retirées, et les dimensions restantes constituent en copule de type D-vine valide. Hasler et al. (2018) utilise cette propriété des copule de type D-vine en plaçant les dimensions ayant des valeurs manquantes au bord de la structure. Dans le cas d'une copule en vigne canonique, les dimensions peuvent être retirées à partir de la fin de la structure. La figure 1.3 montre un exemple des vine copulas de la figure 1.2 avec des dimensions retirées. Dans cet exemple, les vine copulas deviennent en trois dimensions, ce qui fait disparaitre les structure particulières des D-vine ou canonical copulas, n'apparaissant qu'à partir de quatre dimensions.

Min et Czado (2011) ont développé un modèle Bayésien de sauts réversibles modifiant les familles paramétriques de copules bivariées pendant l'inférence du modèle. Cela revient à considérer la copule en vigne comme un métamodèle regroupant différentes formes possible. l'algorithme de sauts réversibles échantillonne la distribution a posteriori du métamodèle, donnant la probabilité de chacune de ses versions. L'application de Min et Czado (2011) considère des copules bivariées t de Student et indépendantes, ce qui revient à considérer l'indépendance entre les dimensions. Cette indépendance est conditionnelle pour les copules bivariées conditionnées à d'autres dimensions.

La structure hiérarchique des vine copulas fait que les copules bivariées sont conditionnées à plus ou moins d'autres dimensions. Plus une copule bivariée est conditionnée, et moins elle a d'importance dans la représentation de la dépendance totale par le modèle. Brechmann et al. (2012) ont défini un algorithme permettant de tronquer une copule en vigne à partir d'un certain niveau hiérarchique, basé sur un test statistique. Cela permet de simplifier le modèle tout en conservant la majorité de la dépendance.

Dissmann et al. (2013) ont développé un algorithme sélectionnant automatiquement la forme de la copule en vigne et les familles paramétriques de copules bivariées. Cela permet de considérer n'importe quelles structures valides, en dehors des structures régulières des D-vine et copule en vigne canonique.

Les copules sont couramment utilisées en hydrologie statistique (Genest et Favre 2007 ; Tootoonchi et al. 2022). Plus spécifiquement, des vines copulas ont été utilisées par Vernieuwe et al. (2015) pour modéliser la dépendance entre différentes caractéristiques des orages (volume de pluie, durée de l'orage et durée de la période sèche après l'orage), et par Ahn (2021) pour modéliser la dépendance spatiale entre différentes stations mesurant des débits de rivières.

1.2.4 Générateurs stochastiques pour des variables hydrologiques

Un GS est un modèle reproduisant des patrons statistiques d'observations dans des simulations. Le nom de "générateur de climat" (weather generator) est aussi utilisé pour désigner les GS. Cela regroupe des types de modèles très différents, permettant de simuler des séries temporelle et/ou des champs de précipitations. La seule condition pour qu'une modèle soit qualifié de GS semble être que son utilisation repose sur de la génération de valeurs aléatoires, mais sans autres contraintes. Cette définition très (trop ?) large regroupe des modèles ayant des différences fondamentales dans leur méthodes et leurs objectifs. Un distinction importante doit être faite entre les GSs générant des valeurs à partir d'une distribution statistique, paramétrique ou non, et ceux n'utilisant pas de distribution. La second option n'est a priori pas adaptée à l'estimation du risque, car elle ne permet pas d'extrapoler les valeurs haute ou d'estimer la probabilité d'un évènement en accord avec la théorie des valeurs extrêmes. Même sans considérer les extrêmes, un générateur stochastique sans distribution statistique ne peut pas faire correspondre la variabilité de ses simulations à la variabilité statistique du phénomène modélisé. Cependant les méthodes de ces modèles peuvent avoir d'autres applications que les GSs, donc leur intérêt n'est pas remis en cause.

Les GSs peuvent être spatiaux et/ou temporels. La plupart des GSs présentés sont appliqués à la pluie, mais certains concernent d'autres variables hydrologiques.

La méthode des analogues est une des approches utilisées pour définir des GS sans distributions. Les analogues sont définis comme des champs de pressions atmosphériques observés dans une région donnée et se ressemblant le plus, par rapport à une métrique de distance (Yiou 2014). Pour un jour donné, ses analogues seront les K autres jours ayant les conditions atmosphériques les plus proches. Un GS est développé par Yiou (2014) à partir des analogues en sélectionnant les observations d'un jour $j = 0$ aléatoirement, puis en sélectionnant des analogues pour simuler les valeurs des jours $j = 1, \ldots, j = n$. La probabilité de sélectionner un des K analogues pour un jour est donnée par

$$p^k = \beta \left(\alpha_1 + \sum_{k=1}^{K} (c^k + 1) \exp\left(-\alpha_2 \, \delta(j_i, j_{i+1}^k) \right) \right), \tag{1.8}$$

où β est un facteur de normalisation pour que la somme des p^k soit 1, α_1 et α_2 sont des paramètres sélectionnés pour obtenir plus ou moins de persistance et saisonnalité dans le résultat, c^k est une mesure de la corrélation spatiale avec l'analogue k, et $\delta(j_i, j_{i+1}^k)$ est la distance entre le jour j_i et son analogue j_{i+1}^k. Cette méthode donne une probabilité pour chaque analogue d'être sélectionné lors de la simulation, mais ce n'est pas une distribution. Krouma et al. (2024) ont combiné cette méthode à un modèle dynamique pour obtenir un modèle prédictif. La méthode des analogues ne semble pas adaptée à un GS pour l'estimation des risques, car elle et basée sur du ré-échantillonnage des observations, avec une génération aléatoire qui n'utilise pas de distribution.

La désagrégation est une méthode permettant de simuler une série chronologique à une certaine fréquence d'observation, à partir d'une autre times séries à une plus basse fréquence, pour par exemple simuler des données à 5 minutes à partir de données horaires. Cette méthode est utilisée dans plusieurs GSs pour la pluie (Evin et al. 2018 ; Vorobevskii et al. 2024). Dans le GS d'Evin et al. (2018), des valeurs de cumuls de pluie sur 3 jours sont simulées, puis sont désagrégés en utilisant une méthode de ré-échantillonnage par plus proche voisin des observations sur 3 jours. Pour chaque cumul de 3 jours simulés, une séquence de 3 jours des K plus proche voisin est sélectionné avec la probabilité

$$p^k = \frac{1/K}{\sum_{i=1}^{k} 1/i}, \tag{1.9}$$

qui donne des probabilités décroissantes, les K voisins étant ordonnées par similarité décroissantes (Wójcik et Buishand 2003). Cette méthode de désagrégation n'est pas de la génération aléatoire à partir d'une distribution, pourtant le reste du modèle d'Evin et al. (2018) utilise des distributions, avec un processus de Markov pour l'intermittence de la pluie, et une distribution EGP pour les cumuls de pluie sur 3 jours. La désagrégation sur 3 jours est utilisée dans ce modèle pour reproduire les patrons de dépendance à 3 jours. Il s'agit donc d'une approche hybride, utilisant à la fois des distribution, et du ré-échantillonnage, comme dans la méthode des analogues. Le GS de Vorobevskii et al. (2024) simule par ré-échantillonnage des observations, qui sont ensuite désagrégées par une copule empirique en deux dimensions. Le processus est itératif, jusqu'à ce que le cumul des valeurs désagrégées ait une différence inférieure à un seuil pré-déterminé par rapport aux observations ré-échantillonnées.

Lee et al. (2020) ont défini un GS pour des séries temporelles à partir d'un modèle ayant une partie déterministe et une partie stochastique générant du bruit Gaussien, de la forme

$$y_t = f(x_t) + \epsilon, \tag{1.10}$$

où $f(\cdot)$ est une couche long short-term memory (LSTM). La seule partie stochastique du modèle est la génération d'un bruit Gaussien. Cette définition de ce qu'est un GS semble limiter l'applicabilité du modèle, car le bruit généré ne correspond à aucune propriété statistique de la variable modélisée.

Les méthodes basées sur le ré-échantillonnage des observations, par analogues ou désagrégation, ne semblent pas adaptées à l'estimation du risque, car elle ne peuvent par définition pas extrapoler en dehors des observations. En plus de ne pas pouvoir extrapoler, ces méthodes ne génèrent pas des valeurs aléatoires à partir d'une distribution, donc il n'y a aucune garantie que la variabilité des simulations corresponde à la variabilité réelle du phénomène. Cependant, l'intérêt premier de la désagrégation de séries temporelles est à faire correspondre des fréquences d"observations différentes, donc cela pourrait être utilisé sans ré-échantillonnage et uniquement avec des distributions.

Plusieurs GSs pour des séries temporelles de pluie utilisent une distribution EGP pour modéliser les valeurs non-nulles de pluie, et d'autres distributions pour modéliser l'intermittence de la pluie et/ou la dépendance spatiale (Evin et al. 2018 ; Ahn 2020 ; Li et al. 2021) [1]. Le GS d'Evin et al. (2018) modélise l'intermittence de la pluie journalière par un processus de Markov d'order 4, et le cumul de pluie des valeurs non-nulles avec une distribution EGP. La dépendance spatiale entre plusieurs stations est modélisée par un processus Gaussien. Le GS d'Ahn (2020) fonctionne à deux échelles, en modélisant des séries temporelles annuelles avec un processus auto-régressif, et des séries temporelles à une échelle journalière avec une distribution EGP. Cette approche décompose la simulation de la variabilité à basse fréquence temporelle, et celle à plus haute fréquence. Li et al. (2021) ont développé un processus ponctuel temporel dont les points correspondent au dépassement d'un seuil, et les valeurs de ces dépassements sont modélisés par une distribution EGP. Le processus ponctuel utilisé est un processus de Hawkes discret, qui fait dépendre la probabilité d'un évènement à l'historique des évènements passées, permettant ainsi de modéliser des grappes (clusters) temporels de points (Hawkes 2018). Dans ces trois exemples, la distribution EGP utilise la fonction puissance comme PIT. Notez que l'EGP n'est techniquement pas une distribution, mais une approche regroupant une infinité de distributions, qui différent par la manière de définir la PIT.

Beneyto et al. (2023) ont exploré l'incertitude d'un GS en fonction de l'information disponible, qui est notamment limitée dans les régions arides. Ils ont pour cela utilisé le modèle d'Evin et al. (2018), en y ajoutant de l'information régionale extrême afin de réduire les incertitudes.

Les processus max-stables sont une extension de la modélisation des extrêmes par bloc maxima au

1. L'article d'Evin et al. (2018) présente différentes versions du GSs, dont une utilisant uniquement des distributions, et une autre utilisant le ré-échantillonnage et la désagrégation en plus des distributions.

domaine spatiale, qui modélise des maxima d'un champ de valeurs (Toulemonde et al. 2020). Cette méthode est utilisé pour des générateurs stochastiques spatiaux de valeurs extrêmes. L'approche max-stable peut être combinée avec le modèle IDF afin de prendre en compte la durée des évènements (Stephenson et al. 2016 ; Jurado et al. 2020). Davis et al. (2013) ont étendu la méthode max-stable à la modélisation spatio-temporelle. Huser et al. (2024) discutent l'inadéquation des modèles max-stables pour la plupart des applications en environnement. Ils encouragent plutôt l'utilisation de méthodes modélisant l'intégralité des données, et non seulement les extrêmes, et notamment d'explorer les possibilités offerte par les méthodes d'apprentissage profond.

Les réseaux antagonistes génératifs (generative adversarial networks, GANs) sont une architecture d'apprentissage profond consistant en deux réseaux de neurones : un premier réseau apprenant à générer des données similaires aux observations, et un second réseau estimant si des données qui lui sont présentées sont des observations réelles ou une simulation issue du premier réseau. L'entrainement de ces deux réseaux conjointement permet au premier réseau d'apprendre à générer des données statistiquement similaire aux observations, donc permet d'apprendre la distribution des observations. Scher et Pessenteiner (2021) utilisent un GAN pour définir un modèle de désagrégation de la pluie, permettant de simuler des champs de précipitation. Cependant ces auteurs pointent plusieurs limitations des GANs pour cette application, notamment le fait que ces modèles assument la stationnarité temporelle des données et que les covariables sont compliquées à intégrer au modèle. Cela ne permet donc pas de facilement développer une version de GAN tenant compte de la variabilité climatique. Ji et al. (2024) ont développé un GAN pour l'estimation du risque d'inondation. Ce modèle simule des données spatio-temporelles, au pas de temps horaire pour 14 stations. Les pluies simulées par le GAN sont ensuite en entrée d'un modèle pluie-débit. Les auteurs évaluent notamment les performances de leur GAN pour les valeurs extrêmes, en évaluant sur les simulations transformées en débit. Le modèle sous-estime les valeurs extrêmes, comparé à une distribution de Gumbel.

1.2.5 Intermittence de la pluie

Les GSs temporels doivent modéliser l'intermittence des séries temporelles de pluie. Plus la fréquence d'observation est élevée, et plus une série temporelle de pluie aura une part importante de zéros, correspondant à un cumul de pluie nul pour l'intervalle de mesure. La mesure de la pluie est discrétisée dans le temps, avec des observations correspondant à des cumuls, par exemple horaires ou pour quelques minutes. Cela fait qu'un évènement de pluie plus long que la fréquence de mesure est observé comme plusieurs valeurs non-nulles successives. De manière similaire, une période sèche

correspond à plusieurs valeurs nulles successives dans les observations. Il y a donc une persistance temporelle à la fois pour les valeurs nulles et pour les valeurs non-nulles. La plupart des modèles ont un composant spécifique modélisant l'intermittence comme une variable binaire, et d'autres composants modélisant les valeurs non-nulles de pluie ou la dépendance spatiale.

Wilks (1998) a développé un GS où une chaine de Markov d'ordre 1 modélise l'intermittence pour des données journalières, ce qui signifie que la probabilité de pluie ou de valeur nulle dépend uniquement du pas de temps précédant. Evin et al. (2018) construisent leur GS en s'inspirant de Wilks (1998), mais utilisent un chaine de Markov d'ordre 4, aussi pour des données journalières. Leurs résultats montrent qu'une chaine plus longue donne de meilleures performance, en particulier pour la modélisation des périodes sèches. Pour des données à plus hautes fréquences d'observation que journalières, une chaine de Markov considérant encore plus de valeurs passées que 4 serait a priori nécessaire.

Le GS de Li et al. (2021) ne modélise pas la pluie, mais une autre variable hydrologique pour laquelle il y a alternance de périodes de valeurs nulles et de valeurs positives. Plutôt que d'utiliser une chaine de Markov, Li et al. (2021) utilisent un processus de Hawkes discret, donnant une probabilité de valeur non-nulle en fonction des valeurs passées. Un processus de Hawkes est un processus ponctuel pour lequel la fonction d'intensité dépend de toutes les valeurs passées, ce qui lui permet de modéliser des grappes d'évènements (Laub et al. 2015 ; Hawkes 2018). Dans le cadre discret des séries temporelles, ces grappes sont des valeurs consécutives de valeurs non-nulles. Les périodes de valeurs nulles sont aussi des grappes, qui sont indirectement modélisés par le processus de Hawkes quand l'intensité reste basse. Comme l'intensité du processus de Hawkes dépend de tout l'historique, il n'est pas nécessaire de choisir le nombre de valeurs passées à considérer, comme dans le cas d'une chaine de Markov.

Bárdossy et Pegram (2009) modélisent la dépendance spatiale entre plusieurs stations et l'intermittence de la pluie via des copules empiriques. Ces copules modélisent à la fois les valeurs nulles et non-nulles.

Papalexiou (2022) modélise séparément le processus binaire de l'intermittence par une distribution discrète de Bernoulli, et les valeurs non-nulles à partir d'un processus Gaussien, qui est ensuite transformé par une distribution marginale. Pendant la simulation, une série temporelle du processus binaire est générée, ainsi qu'une autre série temporelle du processus Gaussien. La série temporelle de pluie est obtenue en multipliant les séries des deux processus binaire et Gaussien.

Papalexiou et al. (2023) utilisent une distribution uniforme mixte pour modéliser l'intermittence,

qui est une distribution uniforme avec une probabilité de masse p_0 à 0, donnant la probabilité de valeur nulle. La densité de cette distribution uniforme mixte est

$$f(u) = \begin{cases} p_0 & \text{if } u = 0 \\ (1 - p_0) & \text{if } 0 < u \geq 1. \end{cases} \tag{1.11}$$

1.2.6 Distributions par réseaux de neurones

L'apprentissage profond permet de modéliser des patrons complexes et non-linéaires, ce qui en fait une approche intéressante pour développer des GSs. Cette section présente deux méthodes pour définir des distributions par apprentissage profond, via diffusion et les réseaux monotones. Le GAN est une troisième méthode par apprentissage profond, qui a été présentée dans la section 1.2.4.

La diffusion est une architecture d'apprentissage profond pour la génération stochastique (Song et al. 2020). Le principe à la base de la diffusion est que des données auxquelles sont ajoutées suffisamment de bruit deviennent indistinguables d'un bruit brut. Pendant l'entrainement du modèle, plus ou moins de bruit est ajouté aux données, allant d'un niveau très faible jusqu'à un tel niveau que les données deviennent indistinguables, et le réseau de neurones est entrainé à débruiter ces données altérées. Pendant la génération à partir du modèle entrainé, du bruit brut est généré, puis est "débruité" jusqu'à obtenir un échantillon de la distribution des données. Ce processus de débruitage est discrétisé pendant la génération avec une chaîne de Markov d'ordre 1. Les différents niveaux de cette chaîne correspondent à la variance du bruit que le modèle retire successivement, jusqu'à ce qu'il ne reste quasiment aucun bruit et que la valeur générée corresponde suffisamment à un échantillon de la distribution. Le modèle peut être conditionné à des covariables. La génération est stochastique car la valeur initiale de bruit brut est générée aléatoirement, mais le débruitage est déterministe dans la plupart des modèles de diffusion. Karras et al. (2022) ont développé des modèles de diffusion stochastique, où du bruit est rajouté à chaque étape de la chaîne de Markov pendant le débruitage, ce qui améliore le résultat final. La diffusion peut être utilisé pour la génération stochastique de variables environnementales spatiale et/ou temporelle. Dans le cas des extrêmes, la diffusion seule ne peut pas extrapoler en accord avec la théorie des valeurs extrêmes, donc devrait être combinée à une distribution des valeurs extrêmes. Néanmoins, la diffusion est une des techniques les plus avancées en génération stochastique par apprentissage profond, et permet de modéliser une distribution, donc pourrait servir dans un générateur stochastique.

Un réseau de neurones peut aussi représenter une fonction de probabilité cumulée sous condition que

sa valeur en sortie évolue de façon monotone avec ses entrées. Ainsi, la dérivée de la valeur en sortie respectivement aux entrées est strictement positive, et représente donc une pdf. Deux approches sont utilisées pour définir une telle distribution neurale, avec la monotonicité "faible" ou "forte" (Sartor et al. 2025). Un réseau de neurones faiblement monotone est une architecture classique, telle que des couches densément connectées, avec des termes dans la fonction objectif pénalisant le réseau quand il ne se comporte pas comme une fonction de probabilité cumulée. Zeng et Wang (2022) utilisent un réseau de neurones avec monotonicité faible pour définir une copule neurale, dont la fonction objectif comprend la vraisemblance et des termes pénalisant la densité négative, la différence entre l'intégrale de la densité et 1, et les probabilités incompatibles avec la définition d'une copule aux limites du support. Une distribution neurale à monotonicité faible à l'avantage de ne pas nécessité d'architecture particulière, car toute la définition du modèle passe par la fonction objectif. Cependant, cette définition d'une distribution neurale ne garanti pas que la distribution soit valide, et il est possible par exemple que la densité ne soit pas strictement positive partout. À l'inverse, dans le cas de la monotonicité forte, l'architecture du réseau est modifiée pour garantir la monotonicité. Soit $\sigma(Wx + b)$ une couche neurale. Une façon pour qu'elle soit monotone est de contraindre la matrice des poids W a être positive (Chilinski et Silva 2018). Sartor et al. (2025) montrent qu'une couche monotone peut aussi être définie par

$$f(x) = W^+\sigma(x) + W^-\sigma(-x) + b, \tag{1.12}$$

où W^+ et W^- sont les parties positives et négatives des poids, respectivement, et $\sigma(\cdot)$ est une fonction d'activation monotone saturant à droite (comme par exemple les fonctions ReLU et softplus, voir table 4.2). Une formulation alternative à celle de l'équation 1.12 est

$$f(x) = \sigma(W^+x + b) - \sigma(W^-x + b). \tag{1.13}$$

Dans le cas d'une distribution multivariée, la dérivée d'ordre n respectivement à toutes les dimensions doit être positive, ce qui contraint l'utilisation de fonctions d'activation convexes, comme par exemple la fonction softplus (Chilinski et Silva 2018).

1.2.7 Modèles prédictifs et générateurs stochastiques

Un modèle prédictif peut être purement déterministe, ou avoir une partie stochastique pour la prise en compte des incertitudes. À l'inverse, un GS peut être purement stochastique, ou peut avoir une part de déterminisme via des covariables déterministes dépendantes du temps. Il est donc possible

d'avoir des GS et des modèles prédictifs utilisant des méthodes très similaires, et la différence entre les deux types de modèles ne devient plus qu'une question d'échelle temporelle. Le GS va simuler des valeurs pour une longue période tandis que le modèle prédictif ne simule que pour un horizon plus court, conditionnellement à des observations. Une autre différence et que le modèle prédictif va être initialisé par des observations, mais cela peut aussi être le cas pour un modèle stochastique. La part stochastique d'un modèle prédictif prenant en compte l'incertitude doit idéalement être une distribution statistique, plutôt qu'une manière d'obtenir de la variabilité qui ne corresponde pas à la variabilité statistique du phénomène, comme par exemple en modifiant l'état initial d'un modèle déterministe. À l'inverse, un GS peut avoir des parties déterministes de son architecture qui ne soit pas des distributions statistiques, du moment que la partie stochastique soit une distribution. Les modèles prédictifs existants et les avancées récentes dans ce domaine sont donc d'intérêt pour le développement des GSs, notamment compte-tenu du succès des modèles prédictifs par apprentissage profond.

Le modèle MetNet de Sønderby et al. (2020) prédit la pluie jusqu'à 8 heures dans le future avec une résolution spatiale de 1 km^2 et temporelle de 2 min. MetNet donne une prédiction probabiliste via une distribution discrétisée obtenue par un réseau de neurones. L'architecture du modèle repose sur des convolutions spatiales, des couches convolution-LSTM (long short-term memory) temporelles et des blocs de self-attention. Il s'agit du premier modèle par apprentissage profond qui améliore les prédictions par rapports aux modèles physiques. Ces derniers modélisent l'incertitude de la prédiction via différentes conditions initiales, tandis qu'un modèle statistique comme MetNet modélise directement la distribution, ce qui est préférable. Une distribution discrétisée comme celle utilisée dans MetNet n'est pas adaptée pour les valeurs extrêmes, car cette distribution a un nombre pré-déterminé d'intervalles, avec toutes les valeurs les plus extrêmes regroupés dans l'intervalle maximum. Cependant le reste de l'architecture du modèle pourrait inspirer le développement d'un GS.

Espeholt et al. (2022) ont développé le modèle prédictif MetNet-2 pour la précipitation. Ce modèle rajoute des couches de convolutions dilatées spatiales par rapport au premier MetNet. Les convolutions dilatées utilisent des filtres de taille constante entre les différentes couches, mais en sautant des observations en fonction du niveau de dilatation. Cela permet d'augmenter la quantité d'observations vues par la convolution tout en gardant un nombre de paramètres constant entre les couches. Pour plus de détails sur les convolutions dilatées, voir l'article du chapitre 4 qui les utilise. Les prédictions de MetNet-2 sont aussi probabilistes, via une distribution discrétisées.

Rasul et al. (2021) ont développé un modèle prédictif probabiliste par diffusion conditionnelle. L'architecture repose sur des couches récurrentes par lesquelles passe la diffusion. Le modèle GenCast de Price et al. (2023) utilise aussi de la diffusion, avec une résolution de 12 heures et 0.25°.

1.3 Structure de la thèse et résumé des articles

La thèse est composée de trois articles, dont deux déjà publiés, portant sur trois modèles stochastiques aux architectures différentes. Le premier modèle est une distribution multivariée spatiale pour l'imputation de données manquantes, dont le composant principal est une copule en vigne à dimension variable. L'intérêt de cette copule d'un point de vue des valeurs extrêmes est la prise en compte de différents types de dépendance asymptotique. Le second modèle est un GS paramétrique pour la pluie combinant une distribution EGP, un processus de Hawkes, et une copule en vigne temporelle. Les développements du premier articles ont partiellement été réutilisés pour ce GS car la copule en vigne est de nouveau à dimension variable, mais avec une application différente. Dans le premier article la copule en vigne spatiale change de dimension selon les stations observées ou manquantes, tandis que dans le GS du second article la copule en vigne temporelle adapte sa dimension selon l'intermittence des valeurs de pluie horaire. Le troisième article est un second GS dont l'objectif est similaire au GS paramétrique, mais où tous les composants du modèle autre que la distribution GP sont remplacés par des réseaux de neurones. La contribution principale de cet article est la définition d'un EGP neural conditionnelle, permettant de simuler des séries temporelles de pluie à une station.

Les points communs reliant ces trois articles sont de développer des modèles utilisant le plus possible d'observations afin d'utiliser au mieux l'information disponible, et de modéliser la dépendance asymptotique entre les dimensions spatiales ou temporelles. Comme indiqué en section 1.1, utiliser toute les observations disponibles et sans les transformer permet de maximiser l'information disponible par le modèle, donc revient à ne pas poser d'hypothèses préalables. À l'inverse, un modèle classique des valeurs extrêmes avec les distribution GEV ou GP pose l'hypothèse forte de séparer les observations entre celles ayant a priori les patrons d'intérêt et celles que le modèle ne va pas regarder. Le second point concernant la dépendance asymptotique est crucial du point de vue des extrêmes, notamment pour la pluie dont les évènements extrêmes dépendent de la durée. Les trois modèles utilisent des distributions multivariées et/ou conditionnelles afin de modéliser cette dépendance.

1.3.1 Imputation par copule en vigne à dimension variable

Titre de l'article : Imputation of missing values in environmental time series by D-vine copulas

Objectif : Développer une méthode d'imputation pour une série temporelle à une station en utilisant l'information d'autres stations voisines, avec un focus particulier sur les valeurs extrêmes.

La problématique de cet article porte sur les valeurs extrêmes dans le cadre de l'estimation du risque d'inondation, mais il s'agit ici du risque de submersion marine via la surcote. La surcote marine est une élévation transitoire et localisée du niveau marin, causée par une dépression atmosphérique. En plus de causer des surcotes, les dépressions amènent des masses d'air humides et donc des pluies, ce qui peut causer des inondations multifactorielles sur le littoral. Plusieurs centrales nucléaires françaises sont situées sur le littoral ou dans des estuaires, ce qui explique l'intérêt de cette variable pour l'ASNR. La surcote n'est pas directement mesurée mais est calculée comme la différence entre le niveau marin observé et la marée maximale (Saint Criq et al. 2022), cette dernière étant modélisée via des harmoniques de marée.

Le niveau marin est mesuré à intervalles réguliers par plusieurs marégraphes sur le littoral, mais ces données sont incomplètes. Les observations peuvent être manquantes pendant des périodes plus ou moins longues, allant d'une seule observation à plusieurs années. Dans le cas des courtes périodes manquantes, la cause peut justement être des évènements climatiques extrêmes, ce qui introduit un biais d'observation cas les valeurs extrêmes ont plus de chance d'être manquantes. Les dépressions atmosphériques affectent des régions suffisamment étendues pour que les surcotes extrêmes entre marégraphes voisins soient dépendantes. Le modèle d'imputation proposé repose donc sur cette dépendance, en utilisant l'information des stations (i.e. marégraphes) voisines pour estimer les valeurs manquantes à une stations particulière.

La méthode d'imputation utilisée est de modéliser la distribution jointe entre la station à imputer et des stations voisines avec une copule en vigne spatiale, où chaque dimension correspond à une station. Comme ces stations voisines ont aussi des données manquantes, la méthode d'imputation de Hasler et al. (2018) reposant sur une copule en vigne à dimension variable est utilisée. En fonction de son architecture, certaines dimensions d'une copule en vigne peuvent être retirées, ce qui résulte en une copule en vigne de dimension réduite. Autrement dit, les vine copulas sont des modèles emboités. Tout comme une copule en vigne est un assemblage de copules bivariées, une copule en vigne de dimension réduite peut se trouver dans une plus grande copule en vigne. Pour qu'une dimension puisse être retirée d'un copule en vigne, il faut que les copules bivariées des

dimensions restantes n'y soient pas conditionnées. Hasler et al. (2018) ont exploité cette propriété des vine copulas pour un modèle d'imputation utilisant l'information d'autres variables ayant aussi des données manquantes. Le modèle d'imputation proposé permet d'imputer la surcote à une station en utilisant l'information de observations non-manquantes des stations voisines via une copule de type D-vine à dimension variable. Les distribution marginales des stations sont modélisées avec une distribution t de Student asymétrique généralisée (Kerman et McDonald 2013).

Le modèle permet d'imputer toute la série temporelle le surcote, mais il doit en particulier capturer la dépendance asymptotique des valeurs hautes afin d'être valable pour les extrêmes. Différentes copules paramétriques bivariées sont considérées dans la copule en vigne afin de prendre en compte l'éventuelle dépendance ou indépendance asymptotique pour chaque paire de dimension. En suivant la méthodologie de Min et Czado (2010) et Min et Czado (2011), la copule en vigne est estimée avec un algorithme Bayésien de Monte-Carlo par chaînes de Markov avec sauts réversibles (RJMCMC). En plus d'estimer les distributions a posteriori des paramètres de la copule en vigne, cet algorithme permet de changer les familles de copule bivariées pendant l'inférence Bayesienne, via des sauts réversibles (Green 1995). Ces modifications des copules bivariées reviennent à considérer la copule en vigne comme un métamodèle regroupant différents modèles selon la combinaison de familles de copules bivariées. La distribution a posteriori de ce meta-modèle est échantillonnée par l'algorithme RJMCMC. Cet algorithme permet donc une sélection des familles des copules bivariées automatique et simultanée avec l'ajustement de leur paramètres. La combinaison de familles de copules bivariées la plus visitée pendant l'échantillonnage Bayésien est le mode de la distribution a posteriori du métamodèle, qui est ensuite utilisé comme modèle d'imputation. La distribution a posteriori du métamodèle pourrait aussi être échantillonnée pendant l'imputation pour une meilleur prise en compte des incertitudes du modèle, mais cela aurait nécessité des développements supplémentaires et dépassait le cadre de l'article. L'incertitude des valeurs imputées est prise en compte en échantillonnant plusieurs valeurs pour chaque observation manquante.

Ce modèle d'imputation est appliqué pour une série temporelle de 46 ans au pas de temps d'environ 12 heures (le calcul de la surcote rend le pas de temps irrégulier), pour une station en Manche, le bras de mer séparant la France et l'Angleterre. Cette station à imputer et huit autres stations voisines en Manche sont modélisées par une vine copule en neuf dimensions. Les résultats avec validation croisée donnent un score d'efficacité de Nash–Sutcliffe de 0.857 pour l'ensemble des valeurs. La méthode était aussi capable d'estimer les valeurs extrêmes potentiellement non-observées, avec les incertitudes associées, ce qui la rend pertinente dans le cadre de l'estimation des risques.

1.3.2 EGP conditionnelle paramétrique pour la génération de pluie

Titre de l'article : Stochastic generator for rainfall with a Hawkes process marked by an extended generalized Pareto and a vine copula

Objectif : Développer un générateur stochastique paramétrique pour une série temporelle de pluie horaire adaptée aux valeurs extrêmes.

La problématique de cet article porte sur la simulation de série temporelles de pluie pour l'estimation du risque. Pour cela, les simulations doivent reproduire les comportements asymptotiques de la distribution marginal et de la dépendance temporelle. Si le modèle reproduit correctement l'autocorrélation mais ne tenait pas compte de l'éventuelle dépendance ou indépendance asymptotique temporelle, l'estimation du risque serait sous- ou sur-évalué.

L'architecture du GS proposé repose sur une distribution marginale EGP pour l'ensemble des valeurs non-nulles de pluie, une copule en vigne temporelle modélisant la dépendance temporelle des valeurs non-nulles, et un processus de Hawkes discret pour l'intermittence de la pluie.

En prenant inspiration de Gamet et Jalbert (2022), deux nouvelles distributions EGP sont définies à partir de distributions tronquées comme PITs. Des résultats indépendants du GS (i.e. en ajustant uniquement les EGPs aux valeurs non-nulles) montrent que les deux distribution proposées dans l'article offrent un meilleur ajustement sur des données horaires de pluie, comparée aux EGPs dont les PITs ont un unique paramètre utilisées par Naveau et al. (2016) et Gamet et Jalbert (2022). Ces nouvelles EGPs paramétriques pour la pluie ont donc de l'intérêt, indépendamment du GS, et pourraient être utilisées dans d'autres applications.

L'architecture du GS s'inspire du modèle de Li et al. (2021), où un processus ponctuel modélise l'intermittence d'une série temporelle et une distribution EGP modélise les valeurs non-nulles. Comme une valeur de pluie est associé à chaque point, il s'agit d'un processus ponctuel avec "marque", laquelle est ici une variable distribuée par l'EGP. Ce processus ponctuel est discret dans le temps, car les séries temporelles sont à intervalles réguliers. Un processus ponctuel est défini par sa fonction d'intensité, qui donne une mesure directe ou indirecte de la probabilité d'un point. Dans le cas de la variable modélisée par Li et al. (2021) comme dans le cas de la pluie, un évènement correspond à plusieurs points successifs dans les observations, par exemple un épisode de pluie d'un durée de trois heures sera enregistré comme trois observations successives dans une chronique de pluie horaire. Ce regroupement en grappes (clustering) temporel des observations est artificiel, il résulte juste de la discrétisation temporelle de la mesure. Cependant ce regroupement en grappes doit être modélisé

car il est présent dans les données. Le type de processus ponctuel utilisé est un processus de Hawkes (Hawkes 2018), où l'intensité dépend des points passés, ce qui permet de modéliser des grappes (à la différence d'un processus de Poisson qui modélise des points indépendants). Une formulation simple de l'intensité λ d'un processus de Hawkess discret est donnée par

$$\lambda(t) = \mu + \sum_{i|i<t} \exp\{-\gamma(t-i)\}, \tag{1.14}$$

où i correspond aux points passés par rapport à t, avec $1 \leq i < t \leq T$. La valeur t correspond à l'instant présent du point de vue du modèle, et T est la durée totale de la série temporelle. Dans cet exemple utilisant le noyau exponentiel, l'influence des points passés sur l'intensité diminue exponentiellement avec la distance à t.

Tout comme les évènements de surcotes marines modélisés dans le premier article, les évènements de pluie sont associés aux périodes de basses pression atmosphérique. Un même épisode de basse pression peut durer plusieurs jours, et peut entrainer plusieurs épisodes pluvieux. Il y a donc un second niveau de regroupement en grappes dans les séries temporelles de pluie horaires, qui cette fois n'est pas artificiel mais a un sens physique. Une fonction d'intensité telle que celle de l'équation 1.14 ne peut modéliser que le premier niveau de grappes, et ne pourra donc pas modéliser la dépendance temporelles des pluies en fonction des périodes de basses ou hautes pression atmosphériques. Prendre en compte ce second niveau de regroupement en grappes lié aux basses pressions est important pour l'analyse du risque car des épisodes pluvieux successif peuvent saturer la capacité d'infiltration du sol, et donc augmenter transitoirement le risque d'inondation. Un nouvelle fonction d'intensité à deux niveaux est développée pour cet article, afin de modéliser ces deux niveaux de regroupement en grappes. La fonction d'intensité du premier niveau, modélisant le regroupement en grappes artificiel lié à la discrétisation de la mesure, ressemble à l'équation 1.14, avec une influence positive des évènements passés sur l'intensité à temps t. La fonction d'intensité du second niveau est définie comme une inhibition de l'intensité du premier niveau, ce qui revient à modéliser les périodes de hautes pressions atmosphériques pour lesquelles la probabilité de pluie est moindre. Les simulations utilisant cette fonction d'intensité reproduisent les deux niveaux de regroupement en grappes présents dans les observations.

Un processus de Hawkes marqué par une EGP va reproduire la dépendance temporelles des occurrences de valeurs non-nulles de pluie (i.e. s'il pleut ou non, de manière binaire), mais pas la dépendance temporelle des valeurs en elles-mêmes. Pour introduire cette dépendance entre les marques, Li et al. (2021) ont un paramètre d'échelle σ de l'EGP variable dans le temps. Cela modélise une

partie de la dépendance, mais ne permet pas de modéliser toute la dépendance car il s'agit d'une distribution variable dans le temps, mais pas d'une distribution conditionnelle aux valeurs passées. Dit de manière imagée, la densité de l'EGP avec σ variable aura toujours son pic vers 0 comme dans le cas stationnaire, et ce pic ne peut pas être déplacé en fonction des valeurs passées. De plus, cette EGP non-stationnaire ne tient pas compte de la dépendance temporelle asymptotique entre des valeurs extrêmes successives. Afin de remédier à cela, le nouveau GS proposé rajoute une copule temporelle à l'EGP afin d'obtenir une distribution conditionnelle. Comme dans le cas du modèle d'imputation de surcote, l'éventuelle dépendance ou indépendance est prise en compte en testant différentes familles de copules paramétriques. Une copule en vigne canonique est utilisé car cette structure de copule en vigne est adaptée aux séries temporelle en ayant la première dimension correspondant à t et les dimensions suivantes aux lags successifs (voir figure 1.2). Comme la pluie horaire est intermittente, la dimension de la copule en vigne est variable en fonction des valeurs passées non-nulles.

Les simulations du modèles reproduisent la plupart des patrons statistiques des observations, avec la distribution marginale, l'intermittence et l'autocorrélation temporelle. Les courbes IDF sont utilisés comme diagnostic en l'appliquant aux observations et aux simulations. Les courbes obtenues dans les deux correspondent globalement, ce qui montre que cette architecture de GS est pertinente pour modéliser une variable dont le risque dépend de la durée des évènements.

1.3.3 EGP par réseaux de neurones pour la génération de pluie

Titre de l'article : A neural extended generalized Pareto for rainfall time series stochastic generator

Objectif : Développer un générateur stochastique par apprentissage profond pour une série temporelle de pluie horaire adaptée aux valeurs extrêmes.

La problématique de ce troisième article est la même que pour le second article. La différence entre les deux articles concerne les méthodes. Le premier GS est totalement paramétrique et ce second modèle remplace la plupart des éléments paramétriques par des réseaux de neurones. La seule partie paramétrique restant est la distribution GP, afin que le modèle puisse extrapoler les valeurs hautes en accord avec la théorie des valeurs extrêmes.

L'architecture de ce second GS repose encore sur une distribution EGP, mais dont la PIT est modélisée par un réseaux de neurones. Une autre différence importante avec le modèle paramétrique est que son EGP est conditionnée par une copule temporelle, tandis que l'EGP neural est conditionnée directement via la PIT.

Pour qu'un GS de pluie par apprentissage profond soit utilisable pour l'estimation du risque, il faut que son architecture garantisse qu'il converge à reproduire les patrons asymptotiques de la distribution marginale et de la dépendance temporelle. La fonction objectif du modèle doit donc être la vraisemblance d'une distribution conditionnelle. Différentes modifications d'un réseau de neurone permettent de garantir que son résultat soit monotonic par rapport aux variables en entrées. La méthode communément utilisée pour obtenir ce réseau de neurone monotonic est de contraindre les poids à être positifs. L'équation d'une couche monotone avec cette contrainte est donnée par

$$\sigma(W^{+}x + b), \tag{1.15}$$

où x sont les données en entrée de la couche, W^{+} est la matrice des poids contraints à être positifs, $\sigma(\cdot)$ est une fonction d'activation et b est le vecteur de biais. Chilinski et Silva (2018) utilisent des couches monotones comme dans l'équation 1.15 afin de modéliser une fonction de probabilité cumulée. Cette fonction de probabilité cumulée est dérivée par rapport à une variable en entrée afin d'obtenir la pdf de cette variable, permettant d'entrainer le modèle par maximisation de la vraisemblance. La dérivée de la fonction de probabilité cumulée est obtenue simplement par la différenciation automatique d'un outil d'apprentissage profond. Pour définir l'EGP neurale, la probabilité $G(y)$ des observations par rapport à la distribution GP sont en entrées d'un réseau de couches monotones. La dérivée de ces couches par rapport à $G(y)$ donne la densité d'une PIT inconditionnelle. La dérivée second de cette densité par rapport à des covariables temporelles donne la densité d'une PIT conditionnelle. Les covariables en question sont des valeurs passées par rapport à t ou des représentations de ces valeurs passées obtenues par d'autres réseaux de neurones. Comme la dépendance temporelle entre ces covariables n'a pas besoin d'être modélisée, utiliser seulement la dérivée second de la fonction de probabilité cumulée de l'EGP est suffisant.

Pendant la simulation, plusieurs valeurs uniformes sur $[0, 1]$ sont générées comme valeurs proposées pour $G(y)$ pour la pluie à temps t. La densité conditionnelle de la PIT est calculée pour chacune de ces valeurs et une valeur est sélectionné aléatoirement avec une probabilité proportionnelle à la densité conditionnelle. La fonction de quantile de la distribution GP donne ensuite la valeurs de pluie simulée.

Les patrons des valeurs passées sont modélisées par des couches de convolution dilatées temporelles. Les convolution dilatées ont des filtres de taille constant, mais sautant des valeurs selon le facteur de dilatation. Cela permet à ces couches de représenter les patrons sur des gros jeux de données, ce qui est adapté dans le cas de haute fréquence d'observations temporelle ou spatiale. Espeholt

et al. (2022) utilisent ce type de couches afin de représenter des patrons spatiaux à grande emprise spatiale dans un modèle prédictif pour la précipitation. Dans le cas du GS, les résultats de ces couches de convolution conditionnent l'EGP et sont en entrées des deux autres éléments du modèle variant dans le temps : le paramètre de forme σ de la distribution GP et la probabilité de valeur sans pluie.

Un élément séparé de l'EGP neurale modélise l'intermittence de la pluie, via la probabilité de valeur sans pluie. Cet élément ressemble au processus de Hawkes discret du modèle paramétrique, mais n'en est pas un au sens strict. Comme l'historique est représenté par des convolutions temporelles, il ne couvre pas toutes les valeurs passées, mais seulement celles couvertes par la convolution. Cet élément du modèle pourrait être appelé processus de Hawkes s'il dépendait de couches récurrentes ou d'attention couvrant toutes les valeurs passées, mais dans les faits il ne diffère pas significativement d'un processus de Hawkes neural discret car les couches de convolutions dilatées peuvent remonter très loin dans le temps.

L'évaluation du modèle repose sur la comparaison des observations et des simulations, comme dans le cas du GS paramétrique. L'EGP neurale reproduit la plupart des patrons statistiques, incluant la distribution marginale, la variabilité annuelle, la dépendance temporelle de l'ensemble des valeurs ainsi que la dépendance asymptotique des valeurs extrêmes successives. Le modèle ne reproduit que partiellement la distribution des périodes sans pluie et le patron IDF, donc il est pour l'instant biaisé. Cependant, l'article montre l'intérêt d'un réseau neural monotonique convexe pour définir un GS.

Chapitre 2

Imputation de séries temporelles (article 1)

Ce premier article de la thèse propose un modèle d'imputation de séries temporelles adapté aux valeurs extrêmes, pour une station utilisant l'information d'autres stations voisines mesurant la même variable. L'application du modèle ne concerne pas la pluie mais la surcote marine, car ce travail était dans le prolongement de mon stage de Master 2 effectué à l'ASNR, qui co-finance la thèse. Aucun élément de la méthodologie n'est spécifique aux données de surcote, et le modèle pourrait être appliqué à d'autres variables environnementales. Cependant l'éventuelle intermittence de la variable n'est pas modélisée, donc la méthode ne pourrait pas être applicable à la pluie sans développement additionnel. Cet article a tout de même de l'intérêt du point de vue de la thèse car il utilise un modèle Bayesien pour la prise en compte des incertitudes et une copule en vigne pour construire une distribution multivariée. La méthode concernant la copule en vigne à dimension variable est réutilisée dans le second article.

Si je devais re-travailler sur la problématique de l'article aujourd'hui, je ne referais pas un modèle similaire. Imputer les valeurs manquantes revient à essayer de créer de l'information sans l'avoir observée, ce qui est impossible. Il me semble préférable d'avoir un modèle capable d'utiliser à la fois observations et valeurs manquantes, plutôt qu'une modélisation en deux étapes avec un premier modèle d'imputation puis un second modèle considérant les données imputées comme des vraies observations. De plus, la surcote marine n'est pas observée, mais est une variable transformée obtenue à partir des observations de niveau marin et de la modélisation de la marée. Il aurait fallu

imputer les données manquantes de niveau marin plutôt que celles de surcote.

L'algorithme RJMCMC présenté dans l'article a été codé entièrement en R, ce qui le rend particulièrement lent. D'un point de vue technique, essayer de tout coder moi-même pour un modèle complexe était trop chronophage, avec comme résultat un modèle difficilement réutilisable.

Article 1

Imputation de données manquantes dans les séries temporelles environnementales par D-vine copula

auteurs : A. Chapon, T. B. M. J. Ouarda, Y. Hamdi

journal : Weather and Climate Extremes

soumis le 18/01/2023, accepté le 12/06/2023, publié le 28/06/2023

DOI : 10.1016/j.wace.2023.100591

Contributions : AC a conceptualisé le modèle, codé le modèle, produit les résultats, écrit la première version de l'article et la version révisée. TBMJO a supervisé le projet et révisé le différentes versions de l'article. YH a fourni les données.

Abstract

Missing values in environmental time series are common and must be imputed before carrying out an analysis requiring complete data. We propose an imputation method for the time series of a target station using information of neighboring stations measuring the same variable. The method allows these neighboring stations to have missing values themselves. The multivariate dataset comprising the time series of the target station and its neighboring stations is jointly modeled by a vine copula and parametric margins. Multiple imputation takes into account the uncertainty of missing data by generating several plausible values for each missing value in the time series of the target station. This is done in a Bayesian framework by sampling the posterior distribution of a missing value, which is conditional on the observed stations for the date. The method is suitable for extremes because the vine copula can model the eventual tail dependence between stations. The application to a skew surge time series is presented, with cross-validated results and a focus on the performance for the upper extremes.

2.1 Introduction

The presence of missing values in time series is a recurring issue in environmental sciences and many other disciplines. Data may be missing for various reasons, such as measurement device failure or measurement error (Kalteh et Hjorth 2009). Many common analyses applied to environmental variables, such as spectral analysis or extreme value analysis, require complete time series (Gao et al. 2018). The analysis can be performed on a subset of the dataset for which there are no missing values, but this can severely limit the length of the time series and thus negatively impact the results. Values can also be missing in a systematic way and introduce bias in an analysis. As an example, the probability of missingness can be higher during extreme events due to measurement device failure, which would artificially reduce the frequency of extremes in the recorded time series. Therefore, a prior imputation of the missing data is often necessary. Even for an analysis that can accommodate missing values, imputing rather than ignoring them can improve results by increasing the length of the time series and reducing the potential bias caused by the missingness mechanism.

Our interest is in the imputation of the time series at a given station (referred to as the target station thereafter), using the information of other stations (the neighboring stations) measuring the same variable in an homogeneous region. The imputation method must allow neighboring stations to have missing values, as they are also subject to missingness. This homogeneous region is defined according to the objective of the subsequent analysis of the imputed dataset. Hamdi et al. (2019)

and Andreevsky et al. (2020) defined this region based on the ratio of common extreme events between the target station and each neighboring station, which is adapted if the interest is in the extremes. Since the analysis of extreme values is of particular interest in environmental sciences, e.g., for risk assessment, the imputation must retain good performance for the tails of the distribution in addition to its bulk, and must take into account the uncertainty associated with missing data (Serinaldi et Kilsby 2015).

The imputation methods available for environmental sciences have been extensively reviewed (Kalteh et Hjorth 2009 ; Ben Aissia et al. 2017 ; Gao et al. 2018 ; Hamzah et al. 2020). The time series of the target station and its neighboring stations constitute a multivariate dataset. Joint modeling of this dataset with a parametric distribution is a suitable approach if the extremes are of interest, as the tails of the distribution are explicitly modeled. Furthermore, the uncertainty can be accounted for with multiple imputation by repeated sampling from the multivariate distribution. Multiple imputation involves replacing missing values with several plausible values of what could have been observed (Little et al. 2014). In a Bayesian framework, multiple imputation amounts to sampling several values from the posterior distribution of a missing value.

Copulas are a popular option for constructing multivariate distributions in many fields, including environmental sciences (Tootoonchi et al. 2022). A copula is a multivariate distribution with uniform margins on $[0, 1]$ which models a dependence structure (Grösser et Okhrin 2022). For a d-dimensional random vector $U \in [0, 1]^d$, a copula C is defined by :

$$C(u_1, \ldots, u_d) = \Pr(U_1 \leq u_1, \ldots U_d \leq u_d). \tag{2.1}$$

Each dimension of a dataset can be transformed to be uniform over $(0, 1)$ with its probability integral transformation (Yan 2007). A d-dimensional joint distribution F with margins $F_1, \ldots, F_d$ has a unique copula C, such that :

$$F(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_d(x_d)\}. \tag{2.2}$$

Thus, copulas allow separate modeling of the margins and the dependence between dimensions. Many parametric families of two-dimensional copulas (referred to as pair-copulas thereafter) exist to model a variety of dependence structures, but the choice becomes much more restricted for copulas in higher dimensions (Aas et al. 2009). Furthermore, the existing families of copulas usable in three or more dimensions, such as the Gaussian or Student t copulas, can impose a too restrictive dependence structure for a given dataset.

Copulas have already been applied to multiple imputation. Hollenbach et al. (2014) used Gaussian copulas in two or more dimensions, but this is not appropriate for the extremes because the dimensions are asymptotically independent in the Gaussian copula (i.e., it cannot model the eventual tail dependence). Di Lascio et al. (2015) considered the Gaussian and t copulas in high dimension, or several families of pair-copulas. The high-dimensional t copula has the inverse drawback compared to the Gaussian copula, as it forces a positive tail dependence. The imputation with a single pair-copula (i.e., not several pair-copulas organized as a vine copula) is restricted to two-dimensional datasets. Vine copulas solve both limitations of copulas by constructing high-dimensional dependence structures as assemblages of pair-copulas, taking advantage of the rich diversity of pair-copula families and allowing the dependence structure (including the tail dependence) to vary for each pair of dimensions. Ahn (2021) applied a D-vine copula to estimate streamflow at a partially gauged site conditionally on observations at neighboring stations, with the latter having complete records in their time series. Ahn (2021) compared the performance of the D-vine with six other imputation methods, including geostatistical methods with inverse distance weighting and Kriging, and found the D-vine to outperform them. Hasler et al. (2018) developed the imputation of a multivariate dataset with a D-vine copula for the case where each dimension may have missing values. The approach of Hasler et al. (2018) was restricted to monotone missingness patterns (i.e., only the lower and/or upper dimensions of the ordered dimensions of the D-vine are missing), and assumed the data to be missing completely at random (MCAR, meaning that the probability of missingness is independent on the actual data value). Hasler et al. (2018) compared the performance of their method with five alternatives, different from the six alternatives tested by Ahn (2021), and also found the D-vine to outperform them. In particular, the D-vine outperforms alternatives when the extremes are considered because it accounts for the eventual tail dependence. Jane et al. (2016) applied copulas to extend wave height time series beyond their measurement period using regional information. This can be achieved by considering the dates outside of the observation period as missing values, which does not require additional development of the multiple imputation model. Valle et Kaplan (2019) applied a Gaussian copula for a counterfactual analysis of a dataset where every dimension can have missing values, which shows that the ability of copulas to handle missing values has applications beyond imputation.

Since our objective is to have a multiple imputation method suitable for the extremes, we followed the methodologies of Ahn (2021) and Hasler et al. (2018) by using a D-vine copula to model the joint distribution of the target station and its neighboring stations. The selection of the family for each pair-copula of the D-vine is critical to account for the eventual tail dependence, or the tail

independence, so the parametric families are selected with a Bayesian framework (Min et Czado 2011). The generation of plausible values for multiple imputation is also performed in a Bayesian framework to account for uncertainty through credible intervals.

This paper is organized as follows. Section 2.2 presents the methodology. An application to a skew surge station located on the French Atlantic coast is presented in Section 2.3. The methodology and results are discussed in Section 2.4, along with a conclusion.

2.2 Methods

The time series at the target station and its neighbors constitute a multivariate dataset. The margins of this dataset are modeled with univariate parametric distributions. The marginal nonexceedance probabilities of the observations at each station are obtained from their respective margins. The dependence structure of the multivariate probabilities are then modeled with a vine copula, each station corresponding to a dimension of this multivariate distribution. For a given date, values are sampled from the vine copula for the missing dimensions, conditionally on the observed one. The multiple imputation accounts for the uncertainty of the values generated (Little et al. 2014). Finally the multiple imputed probabilities are transformed back to quantiles.

2.2.1 Marginal distribution

Joint modeling with copulas is often performed in a semiparametric way through the pseudo-likelihood method, which uses the rank of observations obtained from the empirical distribution of the margins (Genest et Favre 2007). The pseudo-likelihood method is not suitable when the extremes are of interest, because the empirical distribution is not precise enough in the tails where there are few observations, so a parametric distribution for the margins is required.

For time series with support on $\mathbb{R}$, the skewed generalized t distribution can offer a good fit in most cases and is used in this study. Its distribution is given by :

$$f_{SGT}(y|\mu,\sigma,\lambda,p,q) = p\left[2\sigma q^{\frac{1}{p}} B\left(\frac{1}{p},q\right)\right]^{-1}\left(1+\frac{1}{q}[1+\lambda\,\mathrm{sign}(y-\mu)]^{-p}\left|\frac{y-\mu}{\sigma}\right|^{p}\right)^{-(q+\frac{1}{p})}, \quad (2.3)$$

where μ is a location parameter, σ is a positive scale parameter, λ controls the skewness with $\lambda \in (-1,1)$, p is a positive parameter controlling the peakedness of the density, q is a positive parameter controlling the tails and $B(\cdot,\cdot)$ is the beta function (Kerman et McDonald 2013).

This distribution is implemented in the *sgt* R package (Davis 2015).

The adequacy of the margins is assessed with quantile-quantile plots. Note that different parametric distributions could be used to model the margins of the different dimensions of the joint distribution.

2.2.2 Vine copulas

A d-dimensional vine copula is composed of $n_c = d(d-1)/2$ pair-copulas. For the case of a 3-dimensional distribution, the density can be modeled with :

$$\begin{aligned} f(x_1, x_2, x_3) = & f(x_1)\, f(x_2)\, f(x_3) \\ & c_{12}\{F(x_1), F(x_2)\}\, c_{23}\{F(x_2), F(x_3)\} \\ & c_{13|2}\{F(x_1|x_2), F(x_3|x_2)\}, \end{aligned} \tag{2.4}$$

where c_{12} and c_{23} are the densities of the pair-copulas between the corresponding dimensions, and $c_{13|2}$ is the density of the pair-copula between dimensions 1 and 3, conditional on dimension 2. $F(x_1|x_2)$ and $F(x_3|x_2)$ are the marginal conditional distributions given by :

$$F(u|v) = \frac{\partial\, C_{uv}\{F(u), F(v)\}}{\partial\, F(v)}, \tag{2.5}$$

where C_{uv} is the distribution of a pair-copula (Aas et al. 2009). For vine copulas in dimension higher than three, (2.5) is used recursively to obtain the marginal distributions conditional on more than one dimension. A vine copula can be represented as a graph with nodes and edges, the latter corresponding to the pair-copulas (Figure 2.1). These nodes and edges are organized by trees, with the two dimensions and conditioning dimensions of a pair-copula in a given tree being specified by nodes on the lower tree. Thus, a tree contains the pair-copulas conditional on the same number of dimensions (with unconditional pair-copulas in the first tree). The D-vine is a special case of vine copulas which is fully specified by the order of the dimensions in its first tree (Figure 2.1). A joint density $f(x_1, \ldots, x_d)$ with a D-vine for the dependence between variables is given by :

$$\begin{aligned} f(x_1, \ldots, x_d) = & \prod_{m=1}^{d} f(x_m) \\ & \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,\ldots,i+j-1}\{F(x_i|x_{i+1}, \ldots, x_{i+j-1}), F(x_{i+j}|x_{i+1}, \ldots, x_{i+j-1})\}, \end{aligned} \tag{2.6}$$

where $c_{i,i+j|i+1,\ldots,i+j-1}$ is the pair-copula density between the dimensions x_i and x_{i+j} transformed to probabilities with their respective margins $F(x_i|x_{i+1},\ldots,x_{i+j-1})$ and $F(x_{i+j}|x_{i+1},\ldots,x_{i+j-1})$ (Hasler et al. 2018).

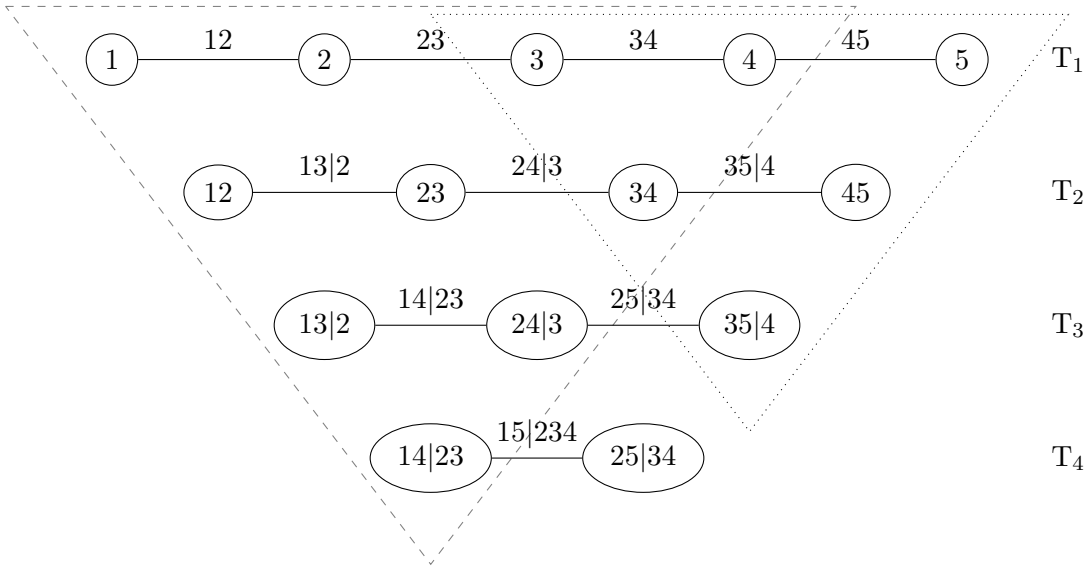


Figure 2.1 – 5-dimensional D-vine, with four trees from top to bottom (T_1 to T_4). The dimensions of the pair-copulas are indicated by the edges between the nodes, with unconditional pair-copulas in T_1 and conditional ones in the subsequent trees. As an example, the edge labeled 13|2 on the second tree represents the pair-copula between dimensions 1 and 3, conditional on dimension 2. The labels of the nodes show how the construction of a vine copula is systematic, with the two dimensions of the pair-copulas on a given tree and their set of conditioning dimensions depending on the previous tree. The dashed and dotted areas give examples of subsets of the D-vine to smaller ones with dimensions 1 to 4 including the six pair-copulas 12, 23, 34, 13|2, 24|3 and 14|23, and dimensions 3 to 5 including the three pair-copulas 34, 45 and 35|4.

2.2.3 D-vine with missing data

The construction of vine copulas by an assemblage of pair-copulas makes them nested models. Depending on its structure, a vine copula of a given dimension can be reduced to a smaller one of lesser dimension if the remaining pair-copulas are not conditional on the removed dimensions. In the case of the D-vine structure, this subsetting of the model results in a smaller D-vine. Figure 2.1 presents a 5-dimensional example, with the dashed and dotted lines delineating smaller D-vines obtained when the dimensions 1 and 2, or 5, respectively, are removed. Hasler et al. (2018) exploited this property of the D-vine to compute the likelihood corresponding to each date (i.e.,

to each multivariate observation). Assuming the data MCAR, the contribution of each date to the likelihood is the observed-data likelihood, given by :

$$f(x_{obs}) = \int f(x_{obs}, x_{mis})\, dx_{mis}, \tag{2.7}$$

where x_{obs} and x_{mis} are the observed and missing dimensions of this date, respectively. For the example of a 4-dimensional D-vine, the contribution to the likelihood of a date having the last fourth dimension (i.e., the rightmost) missing is :

$$f(x_1, x_2, x_3) = \int f(x_1, \ldots, x_4)\, dx_4,$$

which after integrating (2.6) results in the joint density of (2.4). This is applied recursively if the third dimension is missing along the fourth one. The same applies if one or several leftmost dimensions are missing. Note that dimensions could be missing on either side of the D-vine, but the remaining dimensions need to be continuously ordered (e.g., if only the third dimension of a four-dimensional D-vine is missing, a valid smaller D-vine cannot be obtained by removing this third dimension only).

The purpose of subsetting the full d-dimensional D-vine into smaller ones is to use more dates to compute the likelihood. In the case of Hasler et al. (2018), their dataset only had a monotone missingness pattern, so these subsets allowed every observation to contribute to the likelihood. In our case, the missingness patterns of some dates do not correspond to a valid D-vine subset, when the observed dimensions are not ordered continuously in the full D-vine, therefore these dates cannot contribute to the likelihood. Despite not being able to use every observation for the likelihood of the D-vine, this approach still uses much more information compared to only using dates without any missing value. Subsetting the full D-vine is also useful to reduce the computational requirement when sampling from the model for imputation, as will be presented in Section 2.2.5.

Let S be the set of missingness patterns corresponding to a subsettable D-vine, including the full D-vine. Let $x_{.,s}$ be the observations for which a valid subset $s \in S$ exists, and $m_s \subseteq 1, \ldots, d$ the observed dimensions of this subset. The likelihood of the D-vine computed with its subsets S is

given by :

$$L_S(\theta|x) = \prod_{s\in S} \left[\prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,\ldots,i+j-1}\{F(x_{i,s}|x_{i+1,s},\ldots,x_{i+j-1,s}), F(x_{i+j,s}|x_{i+1,s},\ldots,x_{i+j-1,s})\} \middle| i, i+j \in m_s \right], \tag{2.8}$$

where θ is the parameter vector of all the n_c pair-copulas of the D-vine. Note that compared to the density in (2.6), the likelihood in (2.8) only concerns the dependence structure modeled by the D-vine and does not include the margins.

The target station and its $d-1$ neighboring stations each correspond to a dimension of the d-dimensional D-vine. The dimensions (i.e., stations) in the D-vine are ordered by considering two criteria. The first criterion is to order the dimensions with the highest pairwise tail dependence next to each other, so that the corresponding pair-copulas are in the first trees of the D-vine. The first tree has unconditional pair-copulas and the lower trees are conditional on fewer dimensions than the higher trees (Figure 2.1). The pairwise tail dependence is estimated by fitting the t pair-copula on the observations of two stations transformed to probabilities with their respective margins. For the t copula, the lower and upper tail dependence λ is the same, given by :

$$\lambda = 2\, t_{\eta+1}\left(-\sqrt{\eta+1}\sqrt{\frac{1-\omega}{1+\omega}}\right), \tag{2.9}$$

where $t_{\eta+1}$ is the t distribution with $\eta+1$ degrees of freedom, ω is the parameter of the t pair-copula and η its degrees of freedom (Nagler et al. 2022). The estimate of λ given by (2.9) is always positive, therefore, the tail dependence is also tested with the method based on the Neyman–Pearson lemma described in Reiss et Thomas (2007). Since the time series in the application have serial correlation and that this test requires the data to be independent and identically distributed, it is applied to 100 resamples of 1 000 values for each pair of stations to break the serial correlation, and the mean of the 100 p-values is used.

The second criterion for this ordering is placing the dimensions with the highest ratio of missing values on the edges of the D-vine, to allow for more dates to be used in the likelihood computation with (2.8). Furthermore, less observations are unusable for the likelihood computation if the

dimensions closer to the edge of the D-vine have their missing values at the same date.

2.2.4 Selection and adjustment of pair-copulas by reversible jump Markov chain Monte Carlo (RJMCMC)

Once the dimensions are ordered, the selection of the pair-copula families in the D-vine and the adjustment of their parameters is done jointly via reversible jump Markov chain Monte Carlo (RJMCMC) (Green 1995). This RJMCMC algorithm applied to a D-vine is a modified version of the one developed by Min et Czado (2011). It alternates between *Stay* steps, during which the parameters of every pair-copula are updated with a Metropolis-Hastings move, and *Jump* steps, during which the family of one pair-copula changes with a reversible jump move. Only uniform priors are specified. The D-vine is fitted on the nonexceedance probabilities of the observations, which are obtained with the probability integral transformation of the skewed generalized t margins.

Six families of pair-copula are considered in the D-vine to model different types of dependence between its dimensions (Table 2.1). The independence pair-copula accounts for the independence between two dimensions of the D-vine, or for conditional independence in trees higher than the first one. The Gaussian copula covers cases when two dimensions are dependent for the bulk of the data but asymptotically independent (i.e., absence of tail dependence), and allows for a positive or negative correlation. Both the survival Clayton and Gumbel copulas have upper tail dependence. The BB1 and survival BB1 copulas are mixtures based on the Gumbel copula with dependence in both lower and upper tails. These pair-copulas are selected to allow a good fit of the D-vine for the upper extremes in particular, but if the lower extremes were of interest instead, the survival Clayton and Gumbel copulas could be swapped for the Clayton and survival Gumbel copulas, which have a positive lower tail dependence. If the Gaussian copula was not considered, all the families other than the independence copula would have a positive upper tail dependence, which could force the model to have asymptotic dependence between some dimensions.

More families could be considered but it is preferable to limit the set of families so that relevant pair-copulas are proposed more often during the *Jump* steps of the RJMCMC algorithm. The Student t copula is not considered because the evaluation of its likelihood takes much more time than for the six families mentioned previously (with the *VineCopula* R package). If the method is applied to a smaller dataset, the RJMCMC can be run for more iterations, so a larger set of pair-copula families can be tested. Likewise the t copula could be included for a small dataset, since the difference in computation time would then be negligible.

TABLE 2.1 – Parameter boundaries, lower and upper tail dependence of the six families of pair-copula (with 0 and + indicating the absence and positive tail dependence, respectively). The boundaries of the copula parameters and the code for each family in the *VineCopula* R package are also provided.

code	family	l_ρ	u_ρ	l_ν	u_ν	lower t.d.	upper t.d.
0	Independence					0	0
1	Gaussian	-1	1			0	0
13	survival Clayton	0	28			0	+
4	Gumbel	1	17			0	+
7	BB1	0	5	1	6	+	+
17	survival BB1	0	5	1	6	+	+

The D-vine is initialized with the selection method described in HASLER et al. (2018), where the pair-copula in each tree is selected recursively from the first to the last tree. The estimate of the parameter vector θ of the D-vine and the family of each pair-copula obtained by this initial selection are used as starting values in the RJMCMC. The algorithm of HASLER et al. (2018) is adapted to our setting by using only the observations with missingness patterns corresponding to valid D-vine subsets $s \in S$, as for the likelihood in (2.8).

Stay step : updating of pair-copula parameters

During the *Stay* step, the parameters of each pair-copula are updated sequentially. Let $i \in 1, \ldots, n_c$ be the index of the pair-copulas of the D-vine. Let θ be the parameter vector of the entire D-vine. Let k be the index of the pair-copula for which new parameter values are proposed, with $\theta_k = \{\rho_k\}$ or $\theta_k = \{\rho_k, \nu_k\}$ the parameter vector of this kth pair-copula, for a family with one or two parameters, respectively. We denote by $l_{\rho,i}$ and $u_{\rho,i}$ the lower and upper bounds, respectively, for the first parameter of the family of the ith pair-copula. We use the similar notation $l_{\nu,i}$ and $u_{\nu,i}$ for the eventual second parameter. These bounds are specific to each pair-copula family (Table 2.1).

Let *old* and *new* refer to the current and proposed states of the chains, respectively. A value θ_k^{new} is drawn from an adaptive proposal $N(\theta_k^{old}, \Sigma_k)$ in one or two dimensions, truncated to $(l_{\rho,k}, u_{\rho,k})$ and $(l_{\nu,k}, u_{\nu,k})$ for each dimension, respectively. The variance or covariance matrix of this proposal

is given by :

$$\Sigma_k = \begin{cases} (1-\beta)\, 2.4^2\, SamVar_k + \beta\, (u_{\rho,k} - l_{\rho,k})/1000, & \text{if } \theta_k = \{\rho_k\} \\ (1-\beta)\, 2.4^2/2\, SamVar_k + \beta\, \text{diag}(u_{\rho,k} - l_{\rho,k}, u_{\nu,k} - l_{\nu,k})/1000, & \text{if } \theta_k = \{\rho_k, \nu_k\} \end{cases} \tag{2.10}$$

where $SamVar_k$ is the sample variance or covariance matrix of the chains for the current family sampled so far for the kth pair-copula, and $\text{diag}(a, b)$ is a two-dimensional diagonal matrix with a and b on the main diagonal. This adaptive proposal is used when the one or two chains of θ_k contain at least 50 sampled values each, with $\beta = 0.01$. Up until this point, the proposal is nonadaptive with $\beta = 1$ (Roberts et Rosenthal 2009 ; Craiu et Rosenthal 2014).

The uniform prior of the D-vine is given by :

$$\pi(\theta) = \prod_{i=1}^{n_c} (u_{\rho,i} - l_{\rho,i})^{-1} (u_{\nu,i} - l_{\nu,i})^{-1}, \tag{2.11}$$

where $(u_{\nu,i} - l_{\nu,i})^{-1} = 1$ if the family of the ith pair-copula has only one parameter and $(u_{\rho,i} - l_{\rho,i})^{-1} = 1$ as well for the independence copula.

The new parameter vector θ^{new} of the entire D-vine is assembled with :

$$\theta_i^{new} = \begin{cases} \theta_i^{new}, & \text{if } i = k \\ \theta_i^{old}. & \text{if } i \neq k \end{cases} \tag{2.12}$$

The acceptance probability of this Metropolis-Hastings move is given by :

$$\alpha_{stay} = \min \left\{ 1, \frac{L_S(\theta^{new}|x)}{L_S(\theta^{old}|x)} \frac{\pi(\theta^{new})}{\pi(\theta^{old})} \frac{\phi(\theta_k^{old})}{\phi(\theta_k^{new})} \right\}, \tag{2.13}$$

where $L_S(\theta|x)$ is the likelihood of the D-vine computed with its subsets given by (2.8), and $\phi(\theta_k)$ is the density of the one or two dimensional truncated normal proposal (depending if the kth pair-copula has one or two parameters). The uniform priors in (2.13) cancel out since the families of pair-copulas are unmodified during the *Stay* step.

Note that the Gibbs sampler can not be used instead of the Metropolis-Hastings sampler because the full conditional distribution of the D-vine is not known (Min et Czado 2010).

***Jump* step : modification of pair-copula families**

During the *Jump* step, one pair-copula of the D-vine is uniformly selected to propose a modification of its family. Let n_f be the number of pair-copula families considered and f_k the family of the kth pair-copula selected. Each family other than the current one has a $1/(n_f - 1)$ probability of being proposed for a *Jump* step.

Once the pair-copula and the proposed family are selected, new parameter values θ_k^{new} are generated from a one or two dimensional normal distribution $N(\theta_k^*, \Sigma_k)$ truncated to $(l_{\rho,k}, u_{\rho,k})$ and $(l_{\nu,k}, u_{\nu,k})$, respectively. Σ_k is defined as (2.10) for each combination of pair-copula of the D-vine and family. If the family f_k has not been sampled at least 50 times for the kth pair-copula the proposal is nonadaptive, with $\beta = 1$.

θ_k^* is obtained by adjusting the proposed pair-copula to its marginal probabilities, which are given by
$F(x_i|x_{i+1}, \ldots, x_{i+j-1})$ and $F(x_{i+j}|x_{i+1}, \ldots, x_{i+j-1})$ for a copula between dimensions i and $i+j$ of the D-vine (as was mentioned in Section 2.2.3). The marginal probabilities are obtained recursively from conditional distributions given by (2.5). As an example, the marginal probabilities for the pair-copula $C_{14|23}$ on the third tree of the D-vine would be conditional on two dimensions (Figure 2.1). The marginal probabilities for the first dimension of this pair-copula are obtained with :

$$F(x_1|x_2, x_3) = \frac{\partial\, C_{13|2}\{F(x_1|x_2), F(x_3|x_2)\}}{\partial\, F(x_3|x_2)},$$

where $C_{13|2}$ is the pair-copula on the second tree between dimensions 1 and 3 and conditional on dimension 2, and $F(x_1|x_2)$ and $F(x_3|x_2)$ are marginal probabilities obtained from the copulas C_{12} and C_{23} on the first tree, respectively (Hasler et al. 2018). The proposed pair-copula is adjusted on its two marginal probabilities time series by maximum likelihood. The marginal probabilities are obtained with the complete observations of the concerned dimensions, e.g., for the $C_{14|23}$ pair-copula, the marginal probabilities are obtained from the observations without missing values in dimensions $1, \ldots, 4$. Min et Czado (2011) estimated the location parameter of the proposal by adjusting the entire D-vine with the modified pair-copula family by maximum likelihood. Instead, estimating θ_k^* only by adjusting the proposed pair-copula on its marginal probabilities does not required evaluating the likelihood of the entire D-vine, which is the costliest part of the algorithm. This modification of the algorithm makes a significant difference in computation time for applications with a large dataset.

As in the *Stay* step with (2.12), the full parameter vector θ^{new} is assembled with θ_k^{new}, whose family and parameters are modified, and the family and parameters of the pair-copulas other than the kth that remain unmodified. The Jacobian of this bijection is equal to 1.

The acceptance probability of the jump is given by :

$$\alpha_{jump} = \min\left\{1, \frac{L_S(\theta^{new}|x)}{L_S(\theta^{old}|x)} \frac{\pi(\theta^{new})}{\pi(\theta^{old})} \frac{g(f_k^{new} \to f_k^{old})}{g(f_k^{old} \to f_k^{new})} \frac{\phi(\theta_k^{old})}{\phi(\theta_k^{new})}\right\}, \tag{2.14}$$

where $g(a \to b)$ is the probability of proposing a jump from family a to b (which in our case is the same for each family, thus the corresponding ratio cancels out), and $\phi(\theta_k)$ is the density of the one or two dimensional truncated normal proposal. $\pi(\theta)$ is the prior defined in (2.11), which does not cancel out in (2.14), compared to (2.13) (Min et Czado 2011 ; Gruber et Czado 2018).

2.2.5 Sampling the D-vine conditionally on the observed dimensions

Multiple imputation is performed by sampling from the missing dimensions of the D-vine conditionally on the observed dimensions. The sampled values are nonexceedance probabilities, which are transformed to quantile with the margins.

The D-vine is subsetted before sampling values according to a date missingness pattern, similarly to the subsets for the likelihood with (2.8). The purpose of sampling from subsets of the D-vine instead of the full d-dimensional one is to reduce the computational requirement. For a given date, the continuously missing dimensions of neighboring stations from each end of the D-vine can be removed without losing any usable dependency from a pair-copula.

Formally, let $i \in 1, \ldots, d$ be the index of the ordered dimensions of a D-vine, with i_z the dimension of the target station to be imputed. For a given observation x_t (i.e., a d-dimensional vector corresponding to the date t), let :

$$a_{t,i} = \begin{cases} 1, & \text{if } x_{t,i} \text{ is observed or if } i = i_z \\ 0, & \text{otherwise} \end{cases} \tag{2.15}$$

and

$$b_{t,i} = \sum_{1}^{i} a_{t,i} \sum_{d}^{i} a_{t,i}. \tag{2.16}$$

The valid D-vine subset for x_t corresponds to the dimensions for which $b_{t,i} > 0$.

As an example, let us consider a 5-dimensional D-vine, as in Figure 2.1, where the third dimension would be the target station. If the dimensions 1, 2 and 4 were observed and the fifth dimension missing along the third, sampling only the third from a D-vine subsetted to dimensions $1, \ldots, 4$ (the dashed area in Figure 2.1) would be equivalent and faster compared to sampling jointly the third and fifth dimensions from the full D-vine, considering that only the third is of interest. However, if the second and fifth dimensions were missing, the valid subset would remain the same, with dimensions $1, \ldots, 4$, and sampling jointly from both the second and third dimensions would be required, because the pair-copulas between the first dimension and the third and fourth are conditional on the second one. Thus, the computational requirement for sampling an observation depends on its missingness pattern, and subsetting the D-vine reduces this requirement for most patterns.

For a given date t with a missing value for the target station, let y_t be the subset of x_t for which $b_{t,i} > 0$ (i.e., the subset of the observation vector corresponding to the dimension of the valid D-vine subset for this date). This subsetted observation vector y_t contains $m \geq 1$ missing values, with at least the missing value of the target station. Nonexceedance probabilities are sampled for these m dimensions of the D-vine subset (or full D-vine if $y_t = x_t$) with a Metropolis-Hastings algorithm.

The adaptive proposal is a m-dimensional normal distribution $N_m(y_{t,m}^{old}, \Sigma_m)$ truncated to the hypercube $[0,1]^m$, where $y_{t,m}$ corresponds to the m missing dimensions of y_t. The variance or covariance matrix Σ_m is defined similarly to (2.10), with :

$$\Sigma_m = (1-\beta)\, 2.4^2/m\, SamVar_m + \beta\, \mathrm{diag}_m(10^{-4}), \tag{2.17}$$

where $\mathrm{diag}_m(10^{-4})$ is a $m \times m$ diagonal matrix with 10^{-4} on the main diagonal and $SamVar_m$ is the sample variance or covariance matrix of the m chains sampled so far. $\beta = 0.01$ when at least 100 values have been sampled for each m dimension, otherwise $\beta = 1$. The acceptance probability is given by :

$$\alpha = \min\left\{1, \frac{L_s(\theta|y_t^{new})}{L_s(\theta|y_t^{old})} \frac{\pi(y_t^{new})}{\pi(y_t^{old})} \frac{\phi(y_{t,m}^{old})}{\phi(y_{t,m}^{new})}\right\}, \tag{2.18}$$

where y_t is the vector of observation assembled from the non missing values of y_t and the values $y_{t,m}^{old}$ or $y_{t,m}^{new}$ sampled for the previous or current iteration of the algorithm, respectively, $L_s(\theta|y_t)$ is the likelihood of the D-vine given by (2.8) for the subset $s \in S$ corresponding to y_t, and $\pi(y_t)$ is a uniform prior whose corresponding ratio cancels.

Convergence of the Markov chains is assessed following Gelman et al. (2015) with the $\widehat{R}$ test by running two or more times the m chains for a given date. The $\widehat{R}$ value indicates the potential scale

reduction of the posterior distribution if the chains of length n were ran further, declining to 1 as $n \to \infty$. As a rule of thumb, the convergence is considered satisfactory if $\widehat{R} < 1.1$. For simplicity, this convergence test is performed only for the chains corresponding to the target station.

2.2.6 Model validation

The performance of the imputation is assessed by k-fold cross-validation, with $k = 5$ (James et al. 2017). The k training and validation sets are obtained by repeating and nonoverlapping blocks of size b, using $b = 70$ because the autocorrelation of the daily skew surge approaches nonsignificative levels at this lag (not shown). The value of k is kept low to reduce computation time.

For each k model, the observations of the validation set at the target station are considered missing and are imputed by MCMC. For the sake of simplicity, the convergence of the chains were not assessed for the cross-validation. A highest posterior density credible interval is computed for each imputed date. These credible intervals are computed on the quantiles rather than their nonexceedance probabilities because the posterior distributions of the latter are skewed, with a negative (positive) skew for the upper (lower) extremes. The highest density intervals would be affected by the skewness of the probabilities (Hyndman 1996). This skewness disappears when the sampled probabilities are transformed back to quantile. The ratios of observations falling within their respective 90% credible interval, or below and above, indicate the validity of the imputation's uncertainty obtained by the D-vine. A perfect model would have 90% of observations inside their respective credible intervals. The highest density intervals are computed with the *hdrcde* R package (Hyndman et al. 2021).

The validity of the model is further assessed with the Nash–Sutcliffe efficiency (NSE) score, given by :

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{T}(X_t^{imp} - X_t^{obs})^2}{\sum_{t=1}^{T}(X_t^{obs} - \overline{X^{obs}})^2}, \tag{2.19}$$

where T is the total number of timesteps of the observed values, X_t^{imp} is the mean of the values generated by MCMC for the date t, X_t^{obs} is the observed value for the date t and $\overline{X^{obs}}$ is the mean of the observed values (Knoben et al. 2019). NSE = 1 indicates that the model perfectly reproduces the observations, while NSE = 0 indicates that it has the same explanatory power as the mean. The NSE is also computed through the k-fold cross-validation, with k NSE values.

2.3 Application

2.3.1 Data and case study

The imputation method is applied to a dataset of skew surge time series for nine tide gauges located along the French Atlantic coast (Figure 2.2). The skew surge is defined as the difference between the maximal observed sea level and the highest predicted astronomical high tide for a single tidal cycle, which may occur at different times in that cycle (Saint Criq et al. 2022), resulting in an approximately 12 hours and 25 minutes timestep. Extreme events are of interest for this variable as they can contribute to coastal flooding. Data availability varies by station, with measurements starting during the 1970s for most. The method is applied to a period of 46 years, from 1971 to 2016. This dataset is characterized by numerous missing values, ranging from 4.4% to 61.8% of the time series depending on the station (Figure 2.2). These missing values are sometimes isolated but at other times extend over long periods, spanning several years in the worst cases. This dataset was already used in the previous studies of Hamdi et al. (2019) and Andreevsky et al. (2020) (albeit with a different selection of neighbor stations).

The skew surge data is assumed to be MCAR. This simplifying assumption is made to follow the methodology of Hasler et al. (2018) and to avoid having to model the missingness mechanism. As a result, the probability of missingness is assumed independent of the value of the skew surge. For the skew surge this assumption may not always be valid, as an extreme sea level or tidal event could increase the chance of measurement device failure. Nonetheless, the methodology is tested with this assumption to assess the performance offered by the D-vine.

Among the tide gauge dataset spanning the French Atlantic coast, the target station of La Rochelle is chosen to test the model. This station has 42.4% missing values over the 46 years of the dataset (Figure 2.2). Eight other neighbor stations for La Rochelle are selected using the method presented in Hamdi et al. (2019) and Andreevsky et al. (2020). This method selects neighboring stations according to their ratio of common extreme events with the target station. The threshold for this ratio, above which a station is included as a neighboring station, is chosen to be sufficiently low so that there is enough regional information to impute each missing date from the target station. However, some stations with too few observations were not included in the D-vine to keep its dimensionality low enough for computational reasons, or because including these stations reduced the number of dates for which a valid subset can be obtained to compute the D-vine likelihood. The selection of these neighbors is not the focus of the present study, and another method of defining a homogeneous region from the standpoint of extremes could be used instead, e.g., a measure of the

pairwise tail dependence between the target station and each other station, a canonical correlation approach (Cavadias et al. 2001), or with the theory of complex networks (Han et al. 2020). Furthermore, the threshold for the metric defining the homogeneous region is chosen so as to have enough neighboring stations, but some stations are left out because of practical reasons related to fitting the D-vine with missing data. As a result, the set of neighboring stations used for imputation is not particularly sensitive to this metric, as long as it is adapted to the objective.

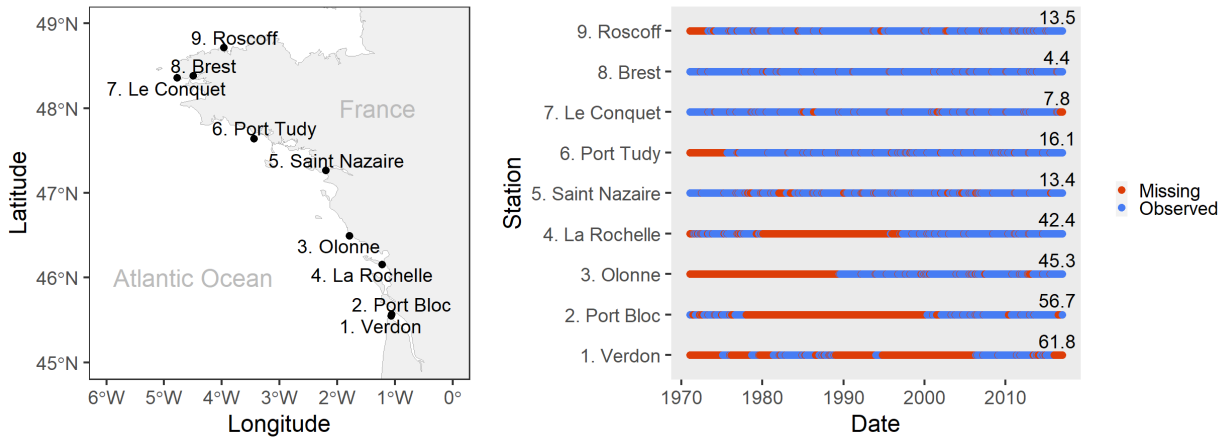


Figure 2.2 – Location of the target station (station number 4, La Rochelle) and its eight neighboring stations along the French Atlantic coast (left). Missing and observed dates for each station, with the percentage of missing values per time series indicated on the right side (right). The station numbers correspond to their order as dimensions of the D-vine copula.

2.3.2 Regional skew surge modeled by D-vine

Figure 2.3 presents the quantile-quantile plots of the skewed generalized t margins for the nine stations. These plots show that this distribution provides good fits for the skew surge time series, even for upper extremes which are of particular interest here. The largest observation is underestimated by the models for some stations, but this observation is much larger than the second largest observation (in particular for the target station of La Rochelle). The points deviate from the main diagonal in the lower tail for some stations (most visibly for Olonne, bottom left subplot), but this is not a great concern for the skew surges as the lower extremes are not of interest.

Figure 2.2 presents the observed and missing dates for each of the nine stations. The order of the stations from bottom to top corresponds to the order of the dimensions of the D-vine. Ordering the stations to maximize the pairwise tail dependence and Kendall's τ (Figure 2.4) in the first

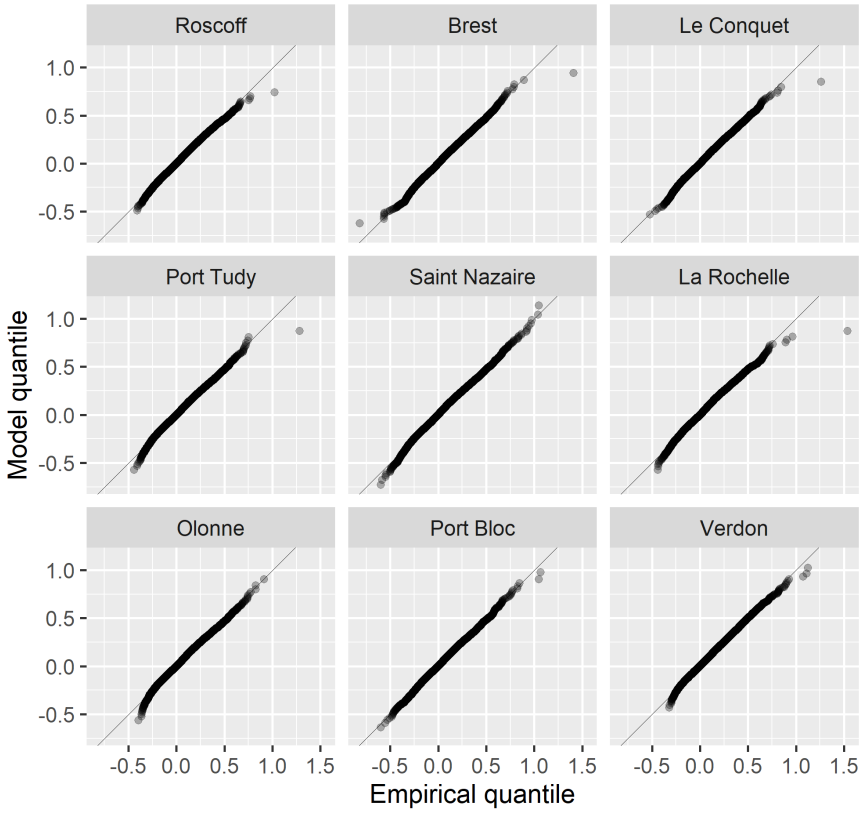


FIGURE 2.3 – Quantile-quantile plots of the skewed generalized t margins for the nine stations.

tree of the D-vine results in this order being in accordance with the spatial organization of the stations (i.e., the order of the stations along the coast, Figure 2.2). However, the periods of missing values of the stations need also to be considered so that a large part of the observations is usable when computing the likelihood of the D-vine with its subsets. Furthermore, it is preferable to allow most of the observations at the target station to be included in the likelihood computation, as this dimension of the D-vine is of particular interest.

In the case of La Rochelle (station number 4), the six stations to its North (stations 5 to 9) have few missing dates, except for Olonne (station 3) during the 1970s and 1980s, while the two stations to its South have long periods of missingness (stations 1 and 2). If the stations were ordered solely on the basis of relative spatial position, Olonne (station 3) should be placed between La Rochelle (station 4) and Saint Nazaire (station 5, Figure 2.2). But doing so would result in a D-vine that

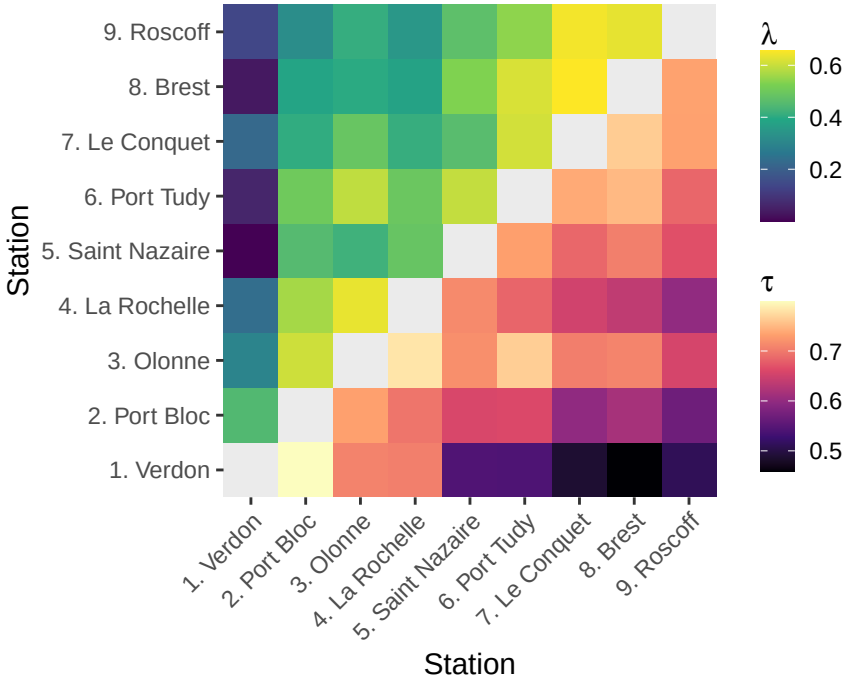


FIGURE 2.4 – Pairwise tail dependence λ and Kendall's τ between stations. The tail dependence is significant at the level $\alpha = 0.05$ for every pair of stations (p-values not shown).

could not be subsetted to include both the target station La Rochelle and the observed stations to its North when Olonne is missing. Instead, the Olonne station is placed to the other side of the D-vine relatively to the target station of La Rochelle (Figure 2.2).

Figure 2.4 shows the pairwise tail dependence between stations. In this figure the stations follow their ordering as dimensions of the D-vine. The highest tail dependence values are close to the main diagonal, indicating that these highest values correspond to pair of stations ordered next or close to each other in the D-vine. The stations ordered next to each other have corresponding pair-copulas on the first tree of the D-vine, which are not conditional on other stations (Figure 2.1). Similarly, the stations that are two orders apart correspond to a pair-copula on the second tree, solely conditional on one other station. Figure 2.4 also presents the pairwise Kendall's τ, for which the same conclusion can be drawn. Overall the highest values of τ are found close to the main diagonal, which indicates that the ordering of the stations is adequate.

The RJMCMC algorithm selecting and adjusting the pair-copulas is run for 5 000 iterations. The

subsets of the D-vine according to the missingness patterns allow 38.18% of the dates to be used in the D-vine likelihood computation (Equation 2.8). Figure 2.5 presents the trace plots of the pair-copulas first parameter. For each pair-copula of the D-vine, the family most sampled by the RJMCMC is the most probable among the six considered, and is kept for the final model used for the subsequent imputation. The first 1 000 sampled values for these most visited families are discarded as warm-up (lighter color in Figure 2.5). Table 2.2 gives the acceptance ratios of the RJMCMC for the pair-copula parameters, with most of them being inside the 20 to 80% range recommended by Min et Czado (2010). The mean acceptance ratio of 2.98% for the jump steps is satisfactory.

The final D-vine obtained by RJMCMC is presented in Table 2.3. The BB1 family is selected for seven out of eight unconditional pair-copulas in the first tree of the D-vine (bottom row of Table 2.3.c), with only the pair-copula between station 1 and 2 having no upper tail dependence. The BB1 family has a positive upper tail-dependence (Table 2.1). Similarly in the second tree (second row from the bottom of Table 2.3.c), the only pair-copula without upper tail dependence is between dimensions 1 and 3 (conditional on dimension 2). The two families without upper pair-dependence (coded 0 for the independence pair-copula and 1 for the Gaussian pair-copula, Table 2.1) are more often selected in higher trees, which are conditional on several dimensions (Figure 2.1). Having pair-copulas with a positive tail dependence in the lower trees of the D-vine is consistent with the pairwise tail dependence estimates of Figure 2.4 and shows that the final model obtained by RJMCMC is appropriate for imputation of upper extremes. For comparison, if the Gaussian family was the most selected in the lower trees of the D-vine, this would indicate that the regional information becomes less relevant for imputation as the values increase, making the model inappropriate for the extremes.

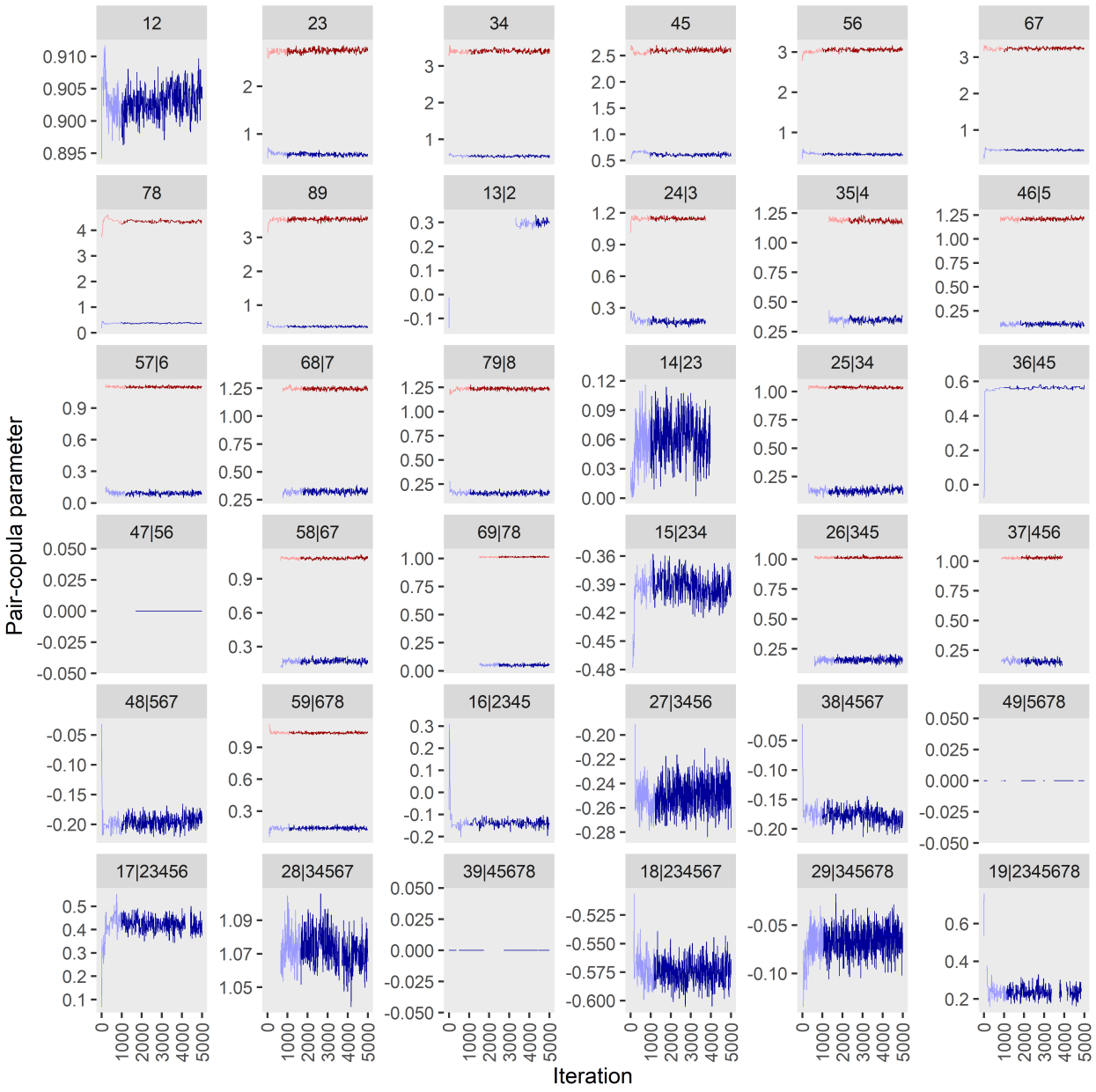


FIGURE 2.5 – Trace plot of the parameters for each pair-copula. The name of each subplot indicates the two dimensions of the pair-copula and its conditioning dimensions. For each subplot, only the trace of the pair-copula family in which the algorithm stayed the longest is displayed, which explains why not all traces start at the beginning and why some are discontinuous. The blue line corresponds to the first parameter ρ of the family and the red line corresponds to the eventual second parameter ν. The independence copula has no parameter but is nonetheless indicated by a constant value of 0. The lighter part of the traces indicate the first 1000 warm-up values for this family, which are discarded.

TABLE 2.2 – Acceptance ratios (in percentage) of the RJMCMC after the warm-up period, for the (a) first parameter of the pair-copulas, (b) second parameter and (c) the jumps between families. The reader is referred to Appendix 2.4 for explanations of the matrix representation of a vine copula used in this Table, with the actual matrix of the D-vine in Table 2.3.a. The dashes in the subtable (b) indicate that a two-parameter family has never been accepted for this pair-copula.

(a) parameter 1

26.2							
38.3	35.7						
47.3	34.1	25.9					
41.0	26.6	40.5	14.0				
23.1	24.9	33.4	33.5	25.7			
27.5	24.5	41.8	4.4	31.0	39.4		
23.4	25.1	26.2	26.8	23.1	22.4	17.7	
19.9	8.3	15.2	19.4	17.5	19.9	23.4	26.1

(b) parameter 2

13.1							
–	–						
–	21.9	28.7					
–	–	–	–				
23.1	–	33.7	35.7	–			
28.8	27.1	–	–	31.1	–		
23.3	28.8	26.0	28.0	23.2	22.3	20.2	
19.9	8.3	15.2	19.4	17.5	19.9	23.4	–

(c) jump

4.4							
0.7	1.2						
6.2	4.1	2.2					
19.8	0.0	2.5	0.0				
2.0	0.0	4.8	4.3	1.7			
8.0	2.1	4.1	0.0	5.7	1.8		
3.3	1.9	4.8	3.7	6.1	5.3	7.0	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 2.3 – D-vine for the skew surge with La Rochelle as target station (dimension 4), with (a) the vine copula matrix indicating the conditioning set of dimensions for each pair-copula, (b) their first parameter, (c) their families (see Table 2.1 for the code of each family) and (d) their second parameter. The reader is referred to Appendix 2.4 for explanations on the matrix representation of a vine copula used in this table. The values in the subtables (b, c, d) correspond to the dimensions of the subtable (a) ; for example the pair-copula between dimensions 1 and 9 has the family number 13, with a first parameter of 0.23 and no second parameter. The reader is referred to Figure 2.2 for the neighbor stations corresponding to the other dimensions. Note that the main diagonal is empty in the subtables other than (a), which is indicated by the dots. The dashes in the subtables (b) and (d) indicates that the pair-copula has no first and/or second parameter.

(a) vine copula matrix

9								
1	8							
2	1	7						
3	2	1	6					
4	3	2	1	5				
5	4	3	2	1	4			
6	5	4	3	2	1	3		
7	6	5	4	3	2	1	2	
8	7	6	5	4	3	2	1	1

(b) pair-copulas parameter 1

.								
0.23	.							
-0.07	-0.58	.						
–	1.07	0.42	.					
–	-0.18	-0.25	-0.14	.				
0.14	-0.20	0.15	0.16	-0.39	.			
0.05	0.17	–	0.56	0.12	0.06	.		
0.16	0.32	0.09	0.11	0.34	0.17	0.30	.	
0.37	0.38	0.45	0.46	0.60	0.52	0.57	0.90	.

(c) pair-copulas families

.								
13	.							
1	1	.						
0	4	13	.					
0	1	1	1	.				
17	1	7	7	1	.			
17	7	0	1	7	13	.		
17	7	7	7	17	17	1	.	
7	7	7	7	7	7	7	1	.

(d) pair-copulas parameter 2

.								
–	.							
–	–	.						
–	–	–	.					
–	–	–	–	.				
1.03	–	1.03	1.01	–	.			
1.01	1.08	–	–	1.03	–	.		
1.24	1.25	1.10	1.21	1.18	1.14	–	.	
3.54	4.35	3.24	3.05	2.60	3.42	2.74	–	.

2.3.3 Multiple imputation of the skew surge

The performance of the multiple imputation is evaluated by considering the observed values at the target station La Rochelle as missing. The method is first tested on a period of 100 time steps, from 2011-10-31 to 2011-12-09. This period is chosen because it has almost no missing values in the nine stations and includes the second highest observation at La Rochelle, which happened on 2011-12-16. Figure 2.6 presents the multiple imputation of the skew surge for four different combinations of neighboring stations considered to be missing in addition to the target station, in order to evaluate the performance of the method with respect to the information available in the region for a given date. Figure 2.6.a considers that only the target station is missing and the eight neighboring stations are observed. Figure 2.6.b considers the neighboring stations on either side of the target station (dimension 4) in the D-vine to also be missing (dimensions 3 and 5). Figure 2.6.c considers the stations corresponding to dimensions 2 to 6 to be missing. Lastly, Figure 2.6.d considers only the neighbors on either side of the target station in the D-vine to be observed (dimensions 3 and 5). For the four cases, the 100 time steps are imputed by sampling 25 000 values by MCMC for each (5 chains of 10 000 values, with the first half of each chain being discarded as warm-up). Figure 2.6 presents pseudo-histograms (color coded) of the posterior density for each date, the NSE and the mean of the variance of each date sample for the four combinations of observed and missing neighbors. The NSE is greater than 0.9 when all eight neighbors are observed but degrades as less regional information is available. Likewise, the mean variance and the spread of the posterior both increase as more stations are missing, indicating greater uncertainty. The order of the missing dimensions in the D-vine affect the performance. The case of Figure 2.6.b has only two neighbors missing, but these are ordered directly on each side of the target station, and the remaining six neighbors are observed. Figure 2.6.d presents an opposite case where only the two neighbors missing in Figure 2.6.b are observed and the remaining six are missing. Both the NSE and the mean variance show that the performance is better in the case of Figure 2.6.d rather than Figure 2.6.b, despite more regional information being available in the latter case. This comparison shows that the performance of the imputation is conditional on having the information for neighboring stations with high tail dependence and Kendall's τ with the target station (Figure 2.4). The $\widehat{R}$ convergence criterion is less than 1.1 for the 100 dates of the four cases (not shown). Note that the results of Figure 2.6 are not computed through cross-validation.

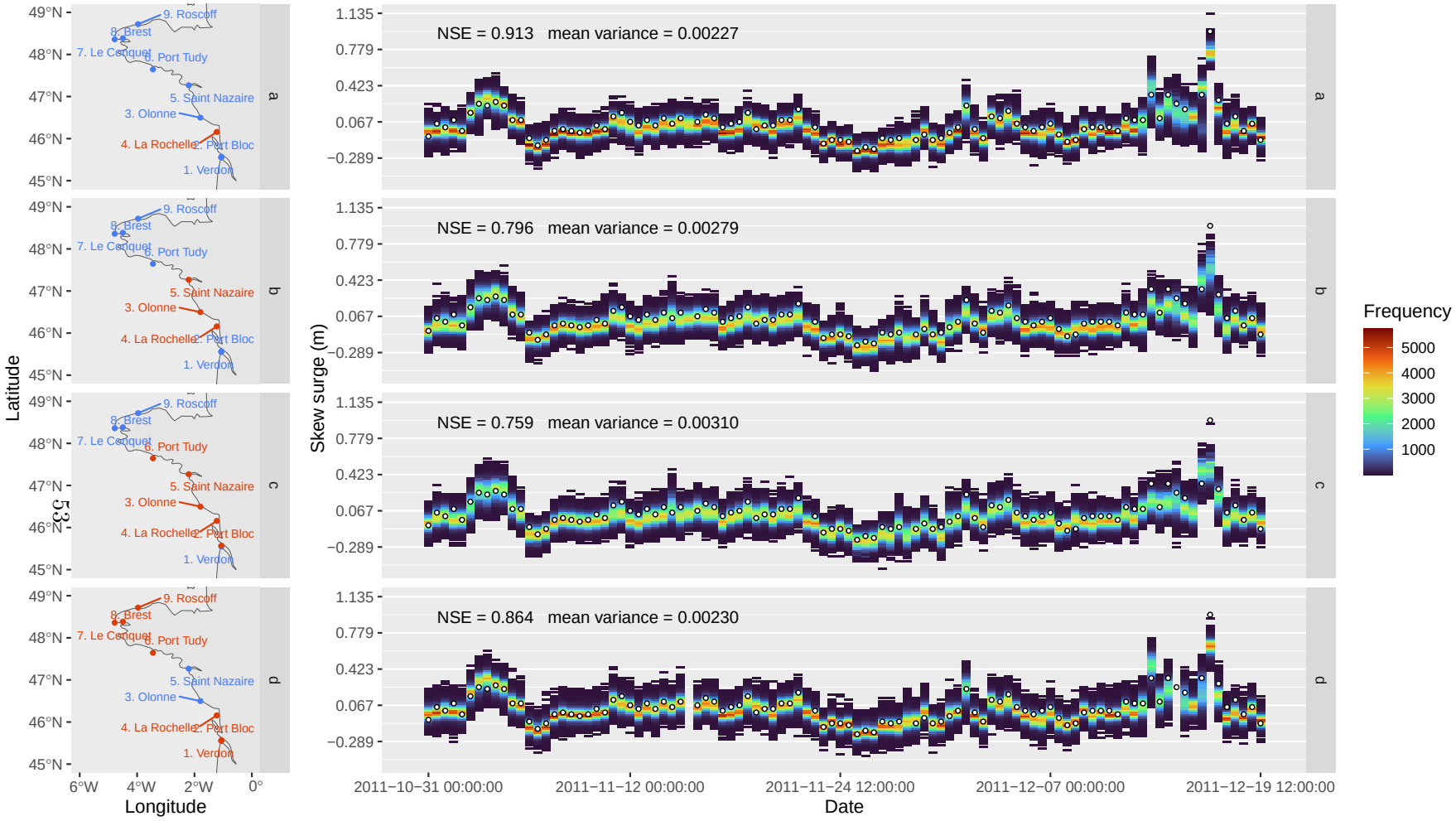


FIGURE 2.6 – Multiple imputation of the skew surge at La Rochelle for four missingness patterns (rows), from 2011-10-31 to 2011-12-09. The maps on the left column show which stations are considered observed (blue) and missing (red) for each test. The right column shows pseudo-histograms of the multiple imputation for each date (color coded, with 25 000 sampled values per date) and for each missingness pattern. The white dots indicate the observation value of each date. For each test, the NSE is computed using the mean value of each date sampled values. Likewise, the mean of the variance of each date sample is computed.

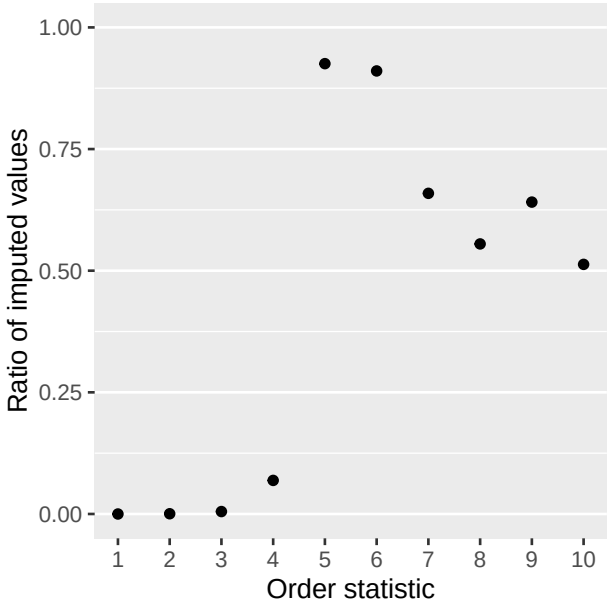


FIGURE 2.7 – Ratio of imputed values in the ten largest order statistics for 2 000 replicates of the completed skew surge time series at La Rochelle.

It can be assumed that the largest order statistics are among the 42.4% missing values at La Rochelle, which the multiple imputation should be able to recover. Figure 2.7 shows the ratio of imputed values among the ten largest order statistic for 2 000 completed time series (each missing date being imputed with two chains of length 2 000, the first halves being discarded as warm-up). The first to fourth largest values remain observations in almost all of the completed time series, but the fifth and sixth largest values are imputed in more than 90% of cases, and this ratio is above 50% for the seventh to tenth largest values. This demonstrates that the method is suitable to impute extremes as it consistently generates values of the largest order. Common extreme value analysis involves the block maxima approach in which the extremes are defined as the largest values in blocks of time (typically years), and the threshold-based approach in which extremes are defined as exceedances of a high threshold (COLES 2001). For both approaches, analyzing the multiple imputed time series would give a better result than restricting the analysis to the observations, as more extreme values would be included, resulting in better estimates and reduced uncertainties (but note that the uncertainty of the multiple imputation must be propagated to the uncertainty of the extreme value analysis).

For the k-fold cross-validation, the observations of the target stations are also treated as missing

and are imputed. The mean of the MCMC sample of each date is used to compare the imputed values to the observation and to compute the NSE through the cross-validation. Note that these mean values are not meant to be used for a subsequent analysis of the imputed time series, as doing so would not take into account the uncertainty of the imputed values. Figure 2.8 shows that there is an overall good correspondence between the observations and the means of the imputed values. The differences seem to increases for the upper extremes, but not to a large extent. The NSE values for the k-fold cross-validation are $\{0.852, 0.862, 0.851, 0.844, 0.876\}$, with a mean value of 0.857. This indicates a good performance of the model when considering the bulk of the data (since this criterion is not specific to extremes).

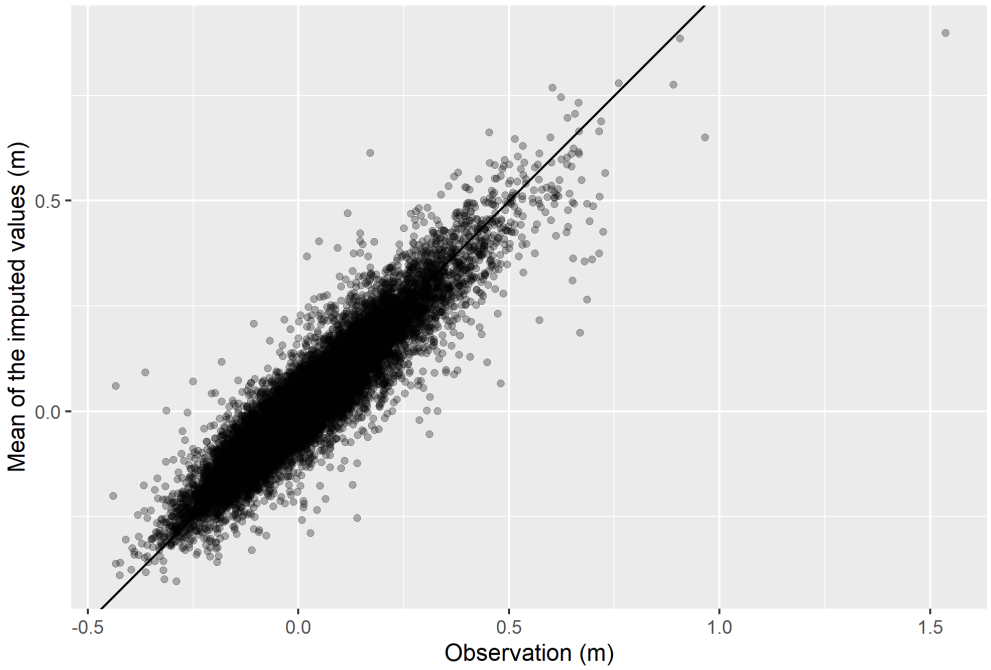


FIGURE 2.8 – Observations of the skew surge at La Rochelle compared to the means of the multiple imputed values. This imputation is done through the k-fold cross-validation.

The validity of the uncertainty obtained by the multiple imputation framework is also assessed through cross-validation. When all the observations are considered, 86.6% of them actually fall within their respective 90% credible intervals, with 6.3% and 7.1% of observations being below and above their interval bounds, respectively (Table 2.4). Therefore, when the whole distribution is considered, the uncertainty of the multiple imputation is only slightly underestimated, with credible

intervals that are a bit narrower than what they should be. When considering the observations above the quantile 0.9, the proportion falling inside the intervals is smaller with only 75.3% of values, indicating an underestimation of the uncertainty, and the ratio of observations above the upper bound of the credible intervals increases to 19.5%, which indicates an underestimation of the upper extremes. Although the lower extremes are not of interest for the skew surge, their imputation performance deteriorates in a manner comparable to that of the higher extremes.

TABLE 2.4 – Comparison of the credible intervals of sampled values from the cross-validated model with observations. The table indicates the percentage of observations falling inside, above and below their respective 90% credible intervals. These ratios are indicated for all the observations as well as for those below the 0.1 and above the 0.9 quantiles of the skewed generalized t margin, to assess the extent to which the model performance deteriorates for extremes.

	$<$ lower bound	inside 90% CI	$>$ upper bound
$<$ quantile 0.1	17.6	78.0	4.3
all observations	6.3	86.6	7.1
$>$ quantile 0.9	5.2	75.3	19.5

2.4 Discussion and conclusions

A method is developed to impute the missing values of a time series at a target station using the information of neighboring stations measuring the same variable, when these neighbors can themselves have missing values. The core of the proposed approach is to model the joint distribution of the time series of the target station and its neighbor stations by a D-vine copula. The uncertainty is accounted for by multiple imputation in a Bayesian framework.

The method is tested with the imputation of a skew surge time series at a station on the French Atlantic coast. The overall performance of the model is good, with a cross-validated NSE of 0.857. When the upper extremes are considered, the method consistently generates new values in the ten largest orders in each replicate of the imputed time series, which indicates that it is able to impute the missing extremes. The completed time series could subsequently be used for any analysis—including extreme value analysis—, with the uncertainty of the missing values being accounted for by the multiple imputation approach. However the cross-validated credible intervals reveal that the uncertainty of the imputed values is underestimated for the extremes. The performance of the model decreases for the upper extremes (which are of interest for the skew surge), but not to an extent suggesting that it is not suitable for extreme values.

The scope of this study is limited to the classical assumption of stationarity. Removing this assumption in future work would require nonstationary models for both the margins and the dependence structure. The latter could be achieved with a dynamic (i.e., nonstationary) vine copula, allowing the dependence between stations to vary in time according to covariates (Chebana et Ouarda 2021). These covariates could be climatic variables, such as atmospheric pressure or wind speed for skew surge, with the parameters of the pair-copulas depending on them. The fitting of the D-vine by RJMCMC could be expanded to include the selection of the covariates and the adjustment of their hyperparameters (El Adlouni et Ouarda 2009). A similar approach could be used for nonstationary models of the margins.

The imputation model could also be expanded by adding in the D-vine variables different from the one measured at the target station. These additional variables could be any that are sufficiently correlated with the time series to impute. If the imputation of the extremes is of interest, it would be preferable to use variables that are tail dependent with the one to be imputed. This could be useful in a situation where not enough neighbor stations measuring the same variable are available, or if they have too many missing values to impute every date of the target station. Adding these variables of a different nature would not require additional development of the present model, as long as they have the same timestep than the variable to impute.

Accounting for the autocorrelation of the time series and the eventual time lag of their correlation could further improve the imputation. Moreover these autocorrelation and time lag could themselves be dependent on covariates, such as storm related variables in the case of the skew surge.

The MCAR assumption may not be always valid in the case of the skew surge time series analyzed, as an extreme event can increase the chance of failure in measurement. Thus the missingness mechanism should be accounted for in the model.

Only the dates with a monotone missingness pattern are included in the computation of the D-vine likelihood (Equation 2.8), but more information could be used by allowing several D-vine subsets for the same date. For instance, a 6-dimensional D-vine with only the third dimensions missing could be subsetted to a pair-copula for dimensions $\{1, 2\}$ and a 3-dimensional D-vine for dimensions $\{4, 5, 6\}$. This would result in a likelihood value for each subset of a given date, which could be weighted to obtain a single likelihood value for the date.

As mentioned in the introduction, Ahn (2021) and Hasler et al. (2018) have both found that the imputation with a D-vine outperforms alternative methods, and particularly so for the extremes as a vine copula can model the eventual tail dependence between dimensions. Thus, although a proper

comparison of the performance of our method with other imputation methods was outside the scope of the article, we can assume that ours would at least offer a similar performance with alternatives for the bulk of the data and would outperform those that do not account for the eventual tail dependence for the extremes.

Funding : The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs Program, and the French Institute for Radioprotection and Nuclear Safety (IRSN) for funding this research. The authors would like also to extend their gratitude to Prof. Xuebin Zhang, the Editor, and two anonymous reviewers for their comments and suggestions, which significantly enhanced the quality of the paper.

Author contributions : Antoine Chapon : Conceptualization, Methodology, Software, Data Curation, Writing - Original Draft. Taha B. M. J. Ouarda : Writing - Review & Editing, Supervision, Project administration, Funding acquisition. Yasser Hamdi : Funding acquisition.

Appendix : Matrix representation of a vine copula

The matrix representation of a vine copula (Morales Napoles 2009 ; Dissmann et al. 2013) is used for the results of Tables 2.2 and 2.3. This representation is a $d \times d$ matrix indicating which dimensions correspond to each pair-copula of the vine, as well as their conditional dimensions for the trees other than the first one. Table 2.5.a gives an example for a 5-dimensional D-vine. The matrix is to be read by columns, with a number on the diagonal $\{d, \ldots, 1\}$ denoting a dimension of a pair-copula and another number below in the same column denoting the second dimension of this pair-copula. In this same column, any other number below both dimensions of the pair-copula indicates a conditioning dimension.

For instance, the pair-copula between dimensions 2 and 5 is indicated in the first column of Table 2.5.a (in bold), with every dimension below them both being the conditioning ones, here dimensions 3 and 4 (italicized). The 5-dimensional D-vine of Figure 2.1 also shows this 25|34 pair-copula in tree T_3. Table 2.5.b is the corresponding representation of a value for each pair-copula (which could be a pair-copula parameter value, the Kendall's τ of the pair-copula, etc.). The subscript notation of each value denotes the two dimensions of the pair-copula and its set of conditioning dimensions. The dots on the diagonal highlight the fact that it is empty, compared to Table 2.5.a. The value corresponding to the pair-copula between dimensions 2 and 5 taken as example is Table 2.5.a is indicated in bold.

Table 2.5 – Matrix representation of a 5-dimensional D-vine.

(a) vine copula matrix

5				
1	4			
2	1	3		
3	2	1	2	
4	3	2	1	1

(b) value x for each pair-copula

.				
$x_{15\|234}$	.			
$\boldsymbol{x_{25\|34}}$	$x_{14\|23}$	.		
$x_{35\|4}$	$x_{24\|3}$	$x_{13\|2}$	.	
x_{45}	x_{34}	x_{23}	x_{12}	.

Chapitre 3

Générateur stochastique paramétrique (article 2)

Ce second article concerne un GS paramétrique pour une série temporelle de pluie à une station. Ce premier des deux GSs proposés est totalement paramétrique, en associant une distribution EGP, un Hawkes process et une vine copula.

Lors du travail sur cet article, le plan pour la thèse était d'ensuite étendre le modèle à plusieurs stations, avec un modèle Bayésien hiérarchique spatial, en s'inspirant de la méthodologie de Johannesson et al. (2022).

Article 2

Générateur stochastique de pluie avec un Hawkes process marqué par une distribution extended generalized Pareto et une vine copula

auteurs : A. Chapon, T. B. M. J. Ouarda, N. Bertrand

journal : Environmental Modelling and Software

soumis le 11/11/2024, accepté le 22/04/2025, publié le 15/05/2025

DOI : 10.1016/j.envsoft.2025.106490

Contributions : AC a conceptualisé le modèle, codé le modèle, produit les résultats, écrit la première version de l'article et la version révisée. TBMJO et NB ont supervisé le projet et révisé le différentes versions de l'article.

Abstract

A stochastic generator for rainfall is built from a Hawkes process, which is modeling the occurrence and serial correlation of non-zero rainfall values. Hawkes processes are suited to model intermittent signals, which is the case of rainfall at a fine enough observation frequency. This Hawkes process has a two-scale intensity function accounting for two orders of clustering in rainfall time series. The rainfall amount of each non-zero value is modeled by an extended generalized Pareto (EGP) distribution with the whole range of rainfall as support, from low to extreme values. New parametric EGP forms adapted to high frequency rainfall time series are defined. The Hawkes process only models the serial correlation of occurrences but not that of the amounts. A conditional version of the EGP is hence developed by adding a copula, modeling the temporal dependence of rainfall amounts. A subsettable canonical vine copula models this dependency for multiple time lags, while accounting for the intermittency of non-zero rainfall values. An application to a 40-year time series of hourly rainfall in France is presented. Simulations from the model reproduce adequately the marginal distribution of rainfall, the temporal clustering of events, and the autocorrelation. The simulations are also able to reproduce the intensity-duration-frequency relation of the IDF extreme value model, showing that this stochastic generator is suitable for risk assessment of duration-dependent extremes.

3.1 Introduction

Rainfall modeling is important for flood risk assessment. Rainfall extremes are duration-dependent because the risk depends on both the amount and the duration of an event. Extreme value distributions and stochastic weather generators (WGs) are two types of statistical models suitable to model extreme rainfall. In recent years, WGs applied to rainfall have integrated extreme value distributions to properly account for its upper tail (Beneyto et al. 2023), which brings the two types of models closer. The opposite is also true, as expanding the classic extreme value model of the GP can end up transforming it into a WG.

The common extreme value framework for a duration-dependent variable is the IDF model (Langousis et Veneziano 2007 ; Ouarda et al. 2018), which considers the amount of rainfall (called intensity in this model) as a stochastic process conditional on both frequency and duration. The IDF model is an extension of the GEV distribution, which considers the maxima over long blocks of time (i.e., block maxima) as a stochastic process. The IDF links these block maxima to both frequency, as in the GEV, and duration. Therefore, the duration of events is not treated as a stochastic variable in

the IDF model. For non duration-dependent variables, the alternative to the GEV model is the GP distribution, which models independent extreme values above a high threshold. The advantage of the GP over the GEV is that the model can be inferred on more observations because the extremes are not defined as the single maxima in each long temporal window, but rather exceedances of a threshold, which improves the accuracy of the inferences (Coles 2001). However the two drawbacks of the GP is the difficulty to select an appropriate threshold, and the requirement that exceedances are independent, which often involves keeping only the maxima of each cluster of exceedances (Lang et al. 1999 ; Pan et al. 2023).

The threshold selection for the GP can be simplified by adding flexibility to the distribution with an EGP (Papastathopoulos et Tawn 2013). Selecting an appropriate threshold for an EGP becomes less critical than for the simpler GP because the added flexibility can adapt to different thresholds. The threshold can now be set at a sub-asymptotic level where both extremes and non-extreme values exceed it. For a variable with positive support, such as rainfall, the threshold can even be lowered down to 0, if the EGP is flexible enough, while retaining a valid extreme value distribution in the upper tail (Naveau et al. 2016). Therefore, the EGP corrects one of the two drawbacks of the GP.

The second drawback of the GP requiring independent observations can be corrected by modeling the serial correlation of threshold exceedances. Li et al. (2021) developed such model by considering the threshold exceedances as points in a Hawkes process. This class of point process models the autocorrelation of points by conditioning the intensity on previous points (Hawkes 2018). In other words, the Hawkes process models the temporal clustering of threshold exceedances, so every exceedance can now be kept and not only a subset of independent ones. A mark can be associated with each point, which is a value distributed by an EGP in the model of Li et al. (2021).

Therefore, for a variable with positive support such as rainfall, a Hawkes process marked by an EGP corrects the two drawbacks of the GP and allows the model to be fitted on every observation, while remaining an extreme value model. Since this marked Hawkes process models the whole time series of observations, it is also able to simulate new time series, and as such is a stochastic WG.

Risk assessment from extreme value distributions is commonly expressed as the quantile of a given probability, or the probability of a given quantile, with the corresponding uncertainties. However the risk is not caused directly by the rainfall. It is rather caused by its transformation into runoff, and ultimately by the water level reached at a vulnerable location. As such, a sensible way to express this risk is with respect to a design variable, which can be the maximum water level reached at a

vulnerable location without failure. A WG can simulate many time series, which can then be the input of a rainfall-runoff model, and a probabilistic risk assessment can be obtained by the ratio of simulations leading to a failure. Simulations from the fitted model can extrapolate the upper extremes beyond the range of observed values through the EGP.

This work presents a WG built from a Hawkes process marked by an EGP with the whole range of rainfall as support. Novelties of this model are the definition of new parametric EGPs for high frequency rainfall, a two-scale intensity Hawkes process intensity function to account for two orders of temporal clustering, and the addition of a subsettable canonical vine copula to the EGP to model the serial correlation of marks for several time lags. This WG is designed as a threshold-based equivalent to the block maxima based IDF model, as the GP is to the GEV. For this model to be a potential alternative to the IDF, it must be able to reproduce the relation between amount, duration and frequency of extreme events modeled by the IDF. To evaluate if this is the case, the IDF model is used as a diagnostic by applying it to the observations and a simulated time series of the same length.

Note that it is common to refer to the amount of rainfall for a given time interval as the rainfall intensity, for example in the IDF model, but to avoid any confusion the term "intensity" will be used to refer to point processes. Note also that the temporal point process discussed here is different from the point process representation of an extreme value distribution (Smith 2003), since the amount of rainfall of a given timestep is not a dimension of the point process, but is instead modeled by a mark.

Several WGs for rainfall or other hydrological variables have been developed in recent years. Evin et al. (2018) built a multisite WG for daily precipitation with an EGP for the marginal distribution and a Markov chain for the occurrence of non-zero values. Papalexiou et al. (2023) also developed a multisite parametric WG for daily precipitation, with a combination of multivariate Gaussian process for the spatial dependence and distributions on $[0, 1)$ having probability mass at 0 to model the intermittency of precipitation. Ahn (2020) developed a WG for precipitation and temperature by separating short term and long term variability. In their work the short term variability is modelled by a parametric distribution with an EGP and a spatial copula, while the long term variability is modelled by wavelet decomposition. Lee et al. (2020) used a deep learning model with recurrent layers in time to simulate monthly streamflow. Some of these models estimate the parameters separately for each month to account for the annual variability (Evin et al. 2018; Papalexiou et al. 2023).

The motivation behind the present work is to develop a new type of WG for hourly rainfall by combining EGP, Hawkes process and vine copula in a novel framework. Combining these three statistical methods permits to build a WG with an extreme value distribution that also accounts for the duration and intermittency of rainfall events. The Hawkes process modelling the intermittency could be removed from the model to build a WG for a non-intermittent variable, for example temperature. Non-stationarity of the rainfall distribution is left out of the scope of this work. The scope is also restricted to the WG and does not cover the rainfall-runoff model needed to obtain a flood risk assessment by simulation. These restrictions of the scope result that the proposed model is not operational, and would require additional developments to this end.

The model presented in this work is not specific to a particular location and should a priori be suitable for any location where the majority of precipitations are rainfall. The model is demonstrated with a 40 year time series of hourly rainfall in France, but longer or shorter time series could be used, granted that it is long enough for a proper estimation of the shape parameter of the EGP.

The presentation of the method is split into four parts. Section 3.2 presents new parametric forms for the EGP which are adapted to high frequency rainfall. Section 3.3 presents a Hawkes process with a novel intensity function for rainfall time series. Section 3.4 presents the addition of a vine copula to the EGP marked Hakwes process to account for the serial correlation of rainfall amount. Section 3.5 presents the inference method via likelihood and how to simulate from the model for risk assessment. An application to an hourly time series in presented in section 3.7. Section 3.8 concludes.

3.2 Distribution of hourly rainfall

3.2.1 Extended generalized Pareto

Extreme values above a sufficiently high threshold follow the GP distribution. The cdf of the GP is

$$G(m|\sigma,\xi) = \begin{cases} 1 - \left(1 + \dfrac{\xi\, m}{\sigma}\right)^{-1/\xi} & \text{for } \xi \neq 0, \\ 1 - \exp\left(-\dfrac{m}{\sigma}\right) & \text{for } \xi = 0, \end{cases} \tag{3.1}$$

where $\sigma > 0$ is the scale parameter and ξ is the shape parameter.

The EGP is a class of distributions where the probability integral transform of a distribution modifies and adds flexibility to the GP (Papastathopoulos et Tawn 2013). This added flexibility permits

to lower the threshold below the domain of the GP, to have a distribution for both extreme and non-extreme values.

The cdf of the EGP is

$$K(m|\sigma,\xi,\boldsymbol{\kappa}) = F_V\{G(m|\sigma,\xi)|\boldsymbol{\kappa}\}, \tag{3.2}$$

where F_V is the distribution used as probability integral transform and $\boldsymbol{\kappa}$ are the parameters of this transform. The corresponding pdf is

$$k(m) = f_V\{G(m)\}\, g(m). \tag{3.3}$$

For it to be a valid distribution, F_V must have $[0,1]$ as support and meet the condition

$$\lim_{v\to 1} f_V(v) = a, \tag{3.4}$$

where $a > 0$ is finite (Gamet et Jalbert 2022), so that the tail index ξ of the nested GP is unaffected by the transformation.

The most popular EGP uses the power function $F_V(v) = v^\kappa$ with $\kappa > 0$ as transform, thereafter named the EGP-power. Naveau et al. (2016) applied the EGP-power to model the whole range of hourly rainfall. The EGP-power is suited for rainfall, as it increases the density at the lower tail compared to the GP when $0 < \kappa < 1$. However, the one parameter power transform can lack flexibility for rainfall time series with high observation frequency. Gamet et Jalbert (2022) proposed to build EGPs with truncated distributions as F_V transform, and defined one suitable for rainfall.

Gamet et Jalbert (2022) also added the condition

$$\lim_{v\to 1} f'_V(v) = 0, \tag{3.5}$$

on the derivative of the transform's density, which ensures that the GP is asymptotically unmodified by F_V.

3.2.2 EGPs defined by truncated distributions

A truncated distribution $T(x)$ with support $[l, u]$, is defined from a distribution $F(x)$ by

$$T(x) = \frac{F[\max\{\min(x, u), l\}] - F(l)}{F(u) - F(l)}, \tag{3.6}$$

where l and u are the lower and upper boundaries of the truncation, respectively (Nadarajah et Kotz 2006). The corresponding density is given by

$$t(x) = \begin{cases} \dfrac{f(x)}{F(u) - F(l)}, & \text{if } l \leq x \leq u, \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

The EGP-beta of Gamet et Jalbert (2022) is defined with a truncated beta distribution. The pdf of its probability integral transform is given by

$$f_V(v|\boldsymbol{\kappa}, l) = \frac{f_{\mathcal{B}}\{(u - l)v + l\}\,(u - l)}{\mathcal{B}(u) - \mathcal{B}(l)}, \tag{3.8}$$

where $f_{\mathcal{B}}$ and $\mathcal{B}$ are the pdf and cdf of the beta distribution, respectively. The beta distribution has two parameters $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)$. Gamet et Jalbert (2022) parametrized the EGP-beta$_1$ with $\kappa_1 = \kappa_2$ and fixed truncation boundaries $l = 1/32$ and $u = 1/2$, resulting in a single parameter for the probability integral transform. The lower truncation value of $l = 1/32$ was empirically defined by Gamet et Jalbert (2022). When $\kappa_1 = \kappa_2$ the derivative $f'_{\mathcal{B}}(x)$ is always 0 at $x = 1/2$, so the condition of Eq. 3.5 is met when $u = 1/2$.

A more flexible EGP-beta$_3$ with three parameters can be defined by having l as a parameter, with $0 \leq l < u$, and allowing $\kappa_1 \neq \kappa_2$. The upper truncation is then obtained by

$$u = \frac{\kappa_1\,\kappa_2 - 1}{\kappa_2 - 2}, \tag{3.9}$$

with $f'_{\mathcal{B}}(u) = 0$.

Following the same principle as the EGP-beta$_3$, an EGP is defined with a truncated generalized beta of the first kind (McDonald 1984). The pdf of the generalized beta of the first kind is given by

$$f_{gb}(x|a, b, c) = c\,\mathcal{B}(a, b)^{-1}\,x^{a\,c-1}\,(1 - x^c)^{b-1}, \tag{3.10}$$

where $a > 0$, $b > 0$, and $c > 0$. The corresponding cdf is given by

$$F_{gb} = \mathcal{B}(x^c|a, b). \tag{3.11}$$

The probability integral transform of the EGP-genbeta is defined with $(a, b, c) = \boldsymbol{\kappa}$. Its pdf and cdf are given by

$$f_V(v|\boldsymbol{\kappa}, l) = \frac{f_{gb}\{(u-l)v + l\}\,(u-l)}{F_{gb}(u) - F_{gb}(l)} \tag{3.12}$$

and

$$F_V(v|\boldsymbol{\kappa}, l) = \frac{F_{gb}(v) - F_{gb}(l)}{F_{gb}(u) - F_{gb}(l)}, \tag{3.13}$$

respectively, where the upper truncation bound u is given by

$$u = \sqrt[\kappa_3]{\frac{\kappa_1\,\kappa_3 - 1}{\kappa_1\,\kappa_3 + \kappa_2\,\kappa_3 - \kappa_3 - 1}}, \tag{3.14}$$

with $f'_{gb}(u) = 0$. The EGP-genbeta has four parameters $(\kappa_1, \kappa_2, \kappa_3, l)$.

3.3 Rainfall occurrence and intermittency

In a rainfall time series, the non-zero timesteps can be seen as occurrences, and thus can be modeled by a temporal point process. Since rainfall is recorded at a discrete time interval, this point process is discrete as well. Most rainfall events will be observed as several consecutive non-zero values, and similarly most dry periods will span several timesteps. Therefore, from the perspective of the discrete point process, both occurrences and non-occurrences must be modeled as dependent and clustered in time.

A point process is characterized by its intensity function, which gives the expected number of points in an interval. In a discrete setting, this interval is the time series timestep and only one point can occur at each timestep, so the intensity gives the probability of occurrence. The simplest point process is the Poisson process which has a constant intensity, resulting in independent occurrences which would be inappropriate for rainfall. A point process suitable for rainfall must have an intensity function able to account for the dependence of occurrences.

This temporal point process only models the occurrence or non-occurrence of rainfall in a binary way. A mark, here modeled by an EGP, accounts for the amount of rainfall of each occurrence. In a general way, the mark of a point process carries extra dimensions that are not directly included in

the intensity function. In this work, the intensity of the point process has only a temporal dimension, and the rainfall amount is a second dimension whose stochastic behavior is not a dimension of the point process intensity. However the value of past marks can have an impact on the intensity of the point process. As an example, high amounts of rainfall could increase the short-term probability of occurrence.

3.3.1 Self-exciting Hawkes process for time series

Hawkes processes are a family of point processes able to account for temporal dependence with a variable intensity conditional on past occurrences (Hawkes 2018). The varying intensity $\lambda(t)$ of a discrete Hawkes process is given by

$$\lambda(t) = \mu + \sum_{i|i<t} f(t-i), \tag{3.15}$$

where μ is a base level intensity and the function $f(\cdot)$ controls the influence of past occurrences at times $i \in [1, t-1]$ on the intensity at time t. A common function for $f(\cdot)$ is the exponential kernel, which exponentially decreases the influence of past occurrences that are further in time from t, asymptotically reaching 0. The intensity function is then given by

$$\lambda(t) = \mu + \sum_{i|i<t} \gamma_1 \exp\{\gamma_2 - \gamma_2(t-i)\}, \tag{3.16}$$

where $\gamma_1 > 0$ controls how much past occurrences influence the intensity and $\gamma_2 > 0$ controls the decay rate of this influence. The exponential kernel results in a self-exciting Hawkes process, in which occurrences increase the short-term probability of future occurrences. In discrete time, this self-exciting point process models clusters of consecutive occurrences.

New occurrences are possible without the influence of past occurrences (or an almost null influence in the case of the exponential kernel) thanks to the base intensity μ. This component of the intensity which does not depend on the past is typically modeled as a constant but could also be itself varying, for example as a function of time (Hawkes 2018).

The rainfall amount of each occurrence is a mark modeled by an EGP. The influence of past marks on the intensity is added via a two-parameter impact function

$$\mathrm{M}(i) = \delta_1 \, m_i^{\delta_2}, \tag{3.17}$$

where m_i is the mark value at time i, $\delta_1 > 0$ and $\delta_2 > 0$.

If the linear influence of past occurrences directly yields the intensity, as in equations 3.15 and 3.16, the Hawkes process is said to be linear. Instead in a non-linear Hawkes process, the linear form gives a function of the intensity (Hawkes 2018). In this work, the generalized logistic type II distribution is used as link function to obtain a non-linear Hawkes process.

Combining the impact function and the link function, the intensity of the marked non-linear Hawkes process is given by

$$\lambda(t) = F_\omega \left[\mu + \sum_{i|i<t} \exp\{\gamma - \gamma(t-i)\} \, \mathrm{M}(i) \right]. \tag{3.18}$$

F_ω is the cdf of the generalized logistic type II, given by

$$F_\omega(x) = 1 - \frac{\exp\{-\omega_3(x-\omega_1)/\omega_2\}}{[1+\exp\{-(x-\omega_1)/\omega_2\}]^{\omega_3}}, \tag{3.19}$$

where ω_1 is a location parameter, $\omega_2 > 0$ is a scale parameter and $\omega_3 > 0$ is a shape parameter. The exponential kernel in Eq. 3.18 was simplified compared to Eq. 3.16 by removing the γ_1 parameter, since it becomes redundant with the extra flexibility of F_ω.

3.3.2 Two-scale intensity function

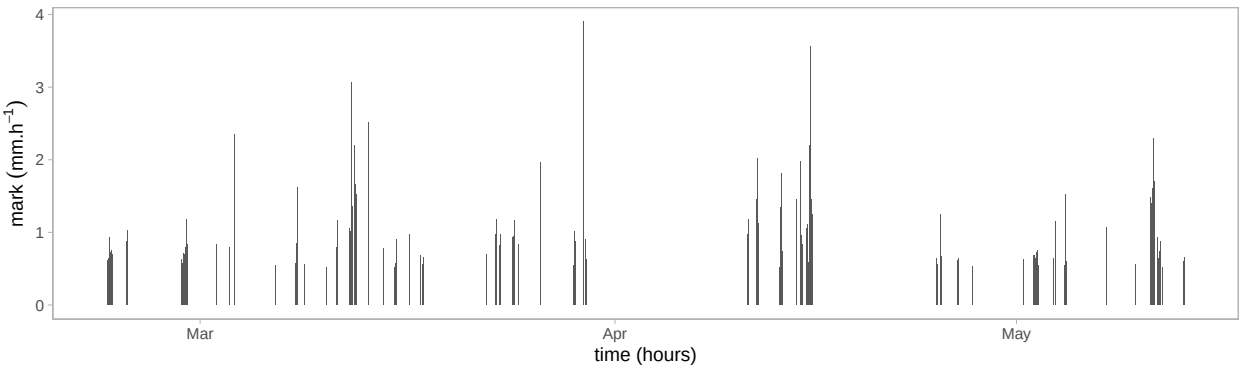


Figure 3.1 – 2000 hours of rainfall in France (49°N 0°E). Marks are here defined as the exceedances of the $u = 0.5$ mm.h^{-1} threshold.

Figure 3.1 presents 2 000 hours of rainfall in France (ERA5 total precipitation at 49°N 0°E). Some occurrences are isolated, but most are clustered in rainfall events spanning up to 10 hours in this

example. These multi-hour rainfall events are not randomly distributed in time but are also regrouped in second-order clusters separated by dryer periods, as can be clearly seen during the month of April in Fig. 3.1. The first-order clusters in the time series are an artifact due to how rainfall is discretely observed, but the second-order clusters are caused by the low atmospheric pressure events spanning a few days, during which several rainfall events can occur. Similarly the long dry periods separating the second-order clusters are caused by high atmospheric pressure events. A self-exciting Hawkes process with a single kernel, such as the one of Eq. 3.18, can only model the first-order of clustering in the time series. The arrival of these clusters would be controlled by the base intensity μ, resulting in them being distributed according to a Poisson process. Instead the arrivals of first-order clusters must also be clustered in time, to model the high and low pressure events.

To model this second-order clustering the base intensity μ is replaced by a second kernel which affects the probability of arrival of first-order clusters. This two-scale intensity function is given by

$$\lambda(t) = \mu_\beta \exp[\beta\{\mu_\alpha - \mu^*_\alpha(t)\}] + \sum_{i|i<t} f(t-i), \tag{3.20}$$

where $\mu_\beta > 0$, $\beta > 0$, and

$$\mu^*_\alpha(t) = \mu_\alpha \left[1 - \sum_{i|i<t} \exp\{\alpha - \alpha(t-i)\}\right], \tag{3.21}$$

where $\mu_\alpha > 0$ and $\alpha > 0$. The component $\mu^*_\alpha(t)$ is inhibited by occurrences, but overall the second kernel also acts a self-exciting part of the intensity.

Combined with the impact and the link functions, the intensity function is given by

$$\lambda(t) = F_\omega \left[\mu_\beta \exp[\beta\{\mu_\alpha - \mu^*_\alpha(t)\}] + \sum_{i|i<t} \exp\{\gamma - \gamma(t-i)\}\, \mathrm{M}(i)\right], \tag{3.22}$$

with ten parameters.

3.4 Serial correlation of rainfall amounts

The Hawkes process described in the previous section models the serial correlation of occurrences but does not model the marks serial correlation. Reproducing the serial correlation of marks is

necessary since rainfall extremes are duration-dependent. This is achieved by adding a copula to the EGP to obtain a distribution of the mark at time t conditional on past marks. Copulas are d-dimensional distributions with support on $[0,1]^d$ modelling the dependence structure between these dimensions, to the exclusion of their marginal distributions. To model the marks serial correlation, the dimensions of the copula are the EGP marginal probabilities at times $t-0,\ldots,t-d+1$.

3.4.1 Conditional EGP with copulas

In the context of the EGP, adding a copula can be seen as applying a second probability integral transform to the GP, on top of the first transform of the EGP. In the case of a pair-copula (i.e., two-dimensional copula), this is done through its conditional distribution, also called the h-function (Aas et al. 2009), given by

$$h(b|a) = F(b|a) = \frac{\partial\, C(a,b)}{\partial\, a}, \tag{3.23}$$

where C is the cdf of the copula. Note that despite the lowercase notation, the h-function is a cdf, here giving the probability of b given a, both having uniform margins.

The conditional cdf of the EGP is given by

$$\begin{aligned} W(m_t|m_{t-1}) &= h[F_V\{G(m_t)\}|F_V\{G(m_{t-1})\}] \\ &= h[K(m_t)|K(m_{t-1})], \end{aligned} \tag{3.24}$$

where m_t and m_{t-1} are the marks at time t and $t-1$, respectively. The corresponding pdf is given by

$$\begin{aligned} w(m_t|m_{t-1}) &= c[F_V\{G(m_t)\}, F_V\{G(m_{t-1})\}]\; f_V\{G(m_t)\}\; g(m_t) \\ &= c[K(m_t), K(m_{t-1})]\; k(m_t), \end{aligned} \tag{3.25}$$

where c is the pdf of the pair-copula.

Since rainfall time series are intermittent, not every mark at time t has a non-zero mark at $t-1$. To evaluate the likelihood of a Hawkes process marked with a copula, the pdf $w(m_t|m_{t-1})$ is used for marks with a previous non-zero mark, and $k(m_t)$ otherwise, the latter being the unconditional EGP. Likewise simulation is done with the quantile functions W^{-1} (with the inverse h-function) and K^{-1}.

3.4.2 Canonical vine copula for intermittent time series

The conditional EGP can also be constructed with a copula in more than two dimensions to directly model the serial correlation for longer lags. The structure of dependency between these dimensions could be too restricted with a single parametric copula, so a vine copula is used instead. Vine copulas are copulas in three or more dimensions built from an assemblage of pair-copulas (Aas et al. 2009). Any combination of parametric pair-copulas can be used for the $d(d-1)/2$ pair-copulas of a d-dimensional vine copula, which offers great flexibility. The structure of a vine copula is organized in so called trees, with unconditional pair-copulas in the first tree and conditional ones on subsequent trees (see Fig. 3.2). For vine copulas in four or more dimensions, different structures are possible depending on which dimensions are linked by unconditional or conditional pair-copulas.

The canonical vine copula structure has the particularity of having one dimension linked to all the other dimensions with unconditional pair-copulas (Fig. 3.2). By using the canonical structure to model marks serial correlation, the mark at time t can have an unconditioned pair-copula with marks at times $t-1, \dots, t-d+1$. The canonical vine copula is thus similar to a Markov process of order $d-1$.

Another property of canonical vine copulas that makes them suitable for time series is subsetability. Since the rainfall is intermittent, it is possible that not all the timesteps $t-1, \dots, t-d+1$ have a non-zero mark. In that case the full d-dimensional canonical vine copula cannot be used. However, as long as the first dimension has a non-zero mark (which will always be the case when the copula is used), the copula can be subsetted to a smaller canonical vine copula. Figure 3.3 gives the example of a subset of the canonical vine copula of Fig. 3.2 when one dimension is removed, in that case if there is a zero mark a time $t-2$. The remaining pair-copulas still constitute a valid canonical vine copula. These subsets are used both to compute the likelihood and to simulate. The vine copula reduces to a pair-copula if only one lag timestep has a non-zero mark, and is removed if all lags have zero marks.

Let $w_d(m_t)$ be the pdf of an EGP conditional on a canonical vine copula subsetted according to the non-zero $m_{t-1}, \dots, m_{t-d+1}$ marks. In other words, $w_d(m_t)$ can range from $k(m_t)$ if every $m_{t-1}, \dots, m_{t-d+1}$ marks are zero, up to an EGP conditional on the full d-dimensional canonical vine copula if every $d-1$ previous marks are non-zero, with every valid subsets of the canonical vine copula in between. In the case of the full vine copula, the conditional EGP pdf is given by

$$w_d(m_t|m_{t-1}, \dots, m_{t-d+1}) = c_d\{K(m_t), \dots, K(m_{t-d+1})\}\, k(m_t), \tag{3.26}$$

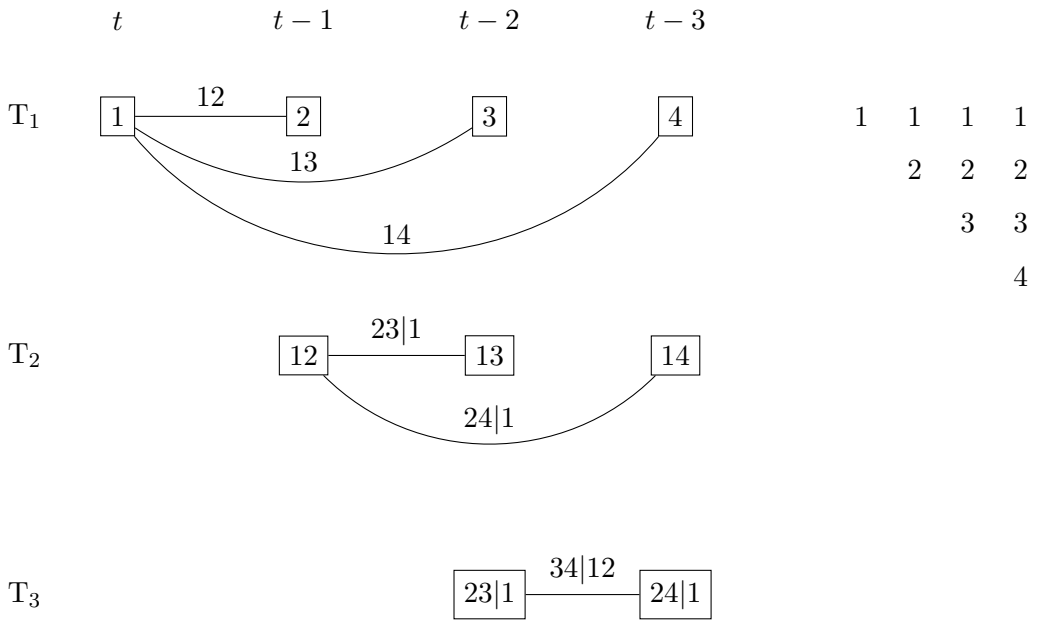


FIGURE 3.2 – Example of a four dimension canonical vine copula modeling the serial dependence up to lag three. The vine copula is organized in the trees T_1 to T_3 from top to bottom. The edges between the boxes represent the pair-copulas (for example the edge 23|1 is the pair-copula between dimensions 2 and 3, conditional on dimension 1). The unconditional pair-copulas are in T_1, all linked to the first dimension of the mark at time t. The matrix representation of the vine copula is displayed on the right.

where c_d is the pdf of the d-dimensional vine copula. The corresponding cdf is

$$W_d(m_t|m_{t-1},\dots,m_{t-d+1}) = C_{1|2,\dots,d}\{K(m_t)|K(m_{t-1}),\dots,K(m_{t-d+1})\}, \tag{3.27}$$

where $C_{1|2,\dots,d}$ is the conditional distribution of the first dimension of the d-dimensional canonical vine copula given the values of the other dimensions.

A convenient representation of vine copulas is a square matrix indicating the set of conditional dimensions for each pair-copula (MORALES NAPOLES 2009). The matrix representation of a 4-dimension canonical vine copula is

$$\begin{bmatrix} 1 & 1 & 1 & \mathit{1} \\ & 2 & 2 & \mathbf{2} \\ & & 3 & 3 \\ & & & \mathbf{4} \end{bmatrix}.$$

This matrix is read by column, with the number on the diagonal and another number in the same column indicating a pair-copula linking these two dimensions. Each other number above them in the same column indicates the conditioning dimensions of this pair-copula. For instance, the numbers 2 and 4 in bold in the matrix indicate a pair-copula, which is conditioned on the first dimension denoted in italic, but not on the third dimension (this is the 24|1 pair-copula in Fig. 3.2). When the triangular matrix is in the upper corner, the subsets of a canonical vine copula are obtained by removing the rows and columns of the removed dimensions (which can be any dimensions except for the first one). Figure 3.3 shows the matrix representation of the subsetted 4-dimensional canonical vine copula on the right, with the third row and column emptied. By removing these empty rows and columns, the triangular matrix of a smaller vine copula is obtained. The pair-copulas in the subset keep the same parametric form and parameter values as in the full vine copula.

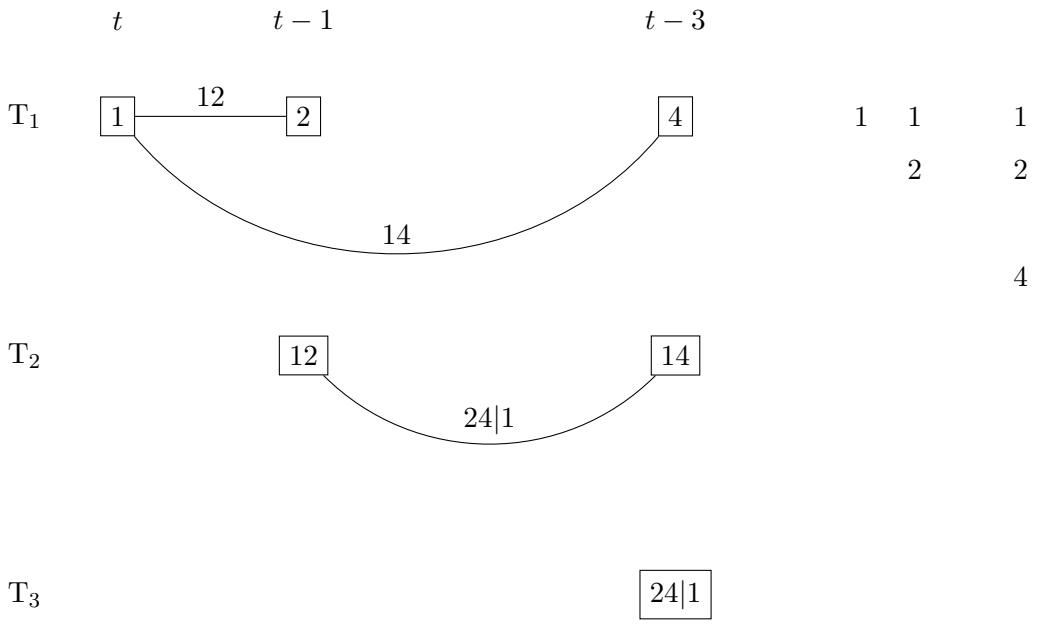


FIGURE 3.3 – Subset of the canonical vine copula of Fig. 3.2 when the third dimension is removed.

The parametric pair-copula considered in the vine copula must cover different types of dependency to ensure that the vine copula is flexible enough. This includes parametric pair-copulas with asymptotic tail dependence or asymptotic tail independence, and pair-copulas with negative dependence. The parametric pair-copulas considered in this application are the Gaussian, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7 and BB8 pair-copulas, which all have closed form h-functions (JOE 2015). Rotations by 90, 180 and 270° of the pair-copulas are also considered to further increase the

possibilities.

3.5 Model inference

3.5.1 Fitting by maximum likelihood

The model is fitted by maximum likelihood estimate. This could be done in one step in principle, but since the model has many parameters it becomes unpractical beyond a certain data size. For the application on 40 years of hourly data, the fitting was carried out in two steps. First the EGP parameters were estimated, then the Hawkes process and copula parameters. This two-steps fitting between marginal distributions and dependence structure is common in copula modelling (Grösser et Okhrin 2022). Here both the copula and the Hawkes process are part of the dependence structure.

In the case of rainfall, the model can in principle model the whole range with an EGP threshold at 0, provided that the EGP is sufficiently flexible. However rainfall time series are discretized due to instrument precision, with values rounded to 0.1 mm.h^{-1} for example, and sometimes censored below a certain threshold. This issue was already noted by Naveau et al. (2016) in the case of the EGP, but also causes issues for the copula, as the discretization can artificially increase the lower tail dependence. To circumvent this, a low threshold can be used instead of having it to 0. In that case, the marks are the threshold exceedances. This threshold must remain low enough to make no significant difference in the context of flood risk assessment.

Let $u \geq 0$ be the threshold and $\boldsymbol{x}$ a time series of T timesteps. The marks $\boldsymbol{m}$ are defined as

$$m(t) = \max(x_t - u, 0), \tag{3.28}$$

for $t \in 1, \ldots, T$.

The full model log-likelihood is given by

$$\log(\mathcal{L}) = \sum_{t|m_t>0} \log\{\lambda(t)\} + \log\{w_d(m_t)\} + \sum_{t|m_t=0} \log\{1 - \lambda(t)\}. \tag{3.29}$$

If the two-step optimization is used instead of Eq. 3.29, the log-likelihood is given by $\log(\mathcal{L}) = \log(\mathcal{L}_1) + \log(\mathcal{L}_2)$, with

$$\log(\mathcal{L}_1) = \sum_{t|m_t>0} \log\{k(m_t)\}, \tag{3.30}$$

to estimate $\hat{\boldsymbol{\kappa}}$ the parameters of K (including the lower truncation bound l if the EGP transform f_V in K is Eq. 3.8 or 3.12). The second term is then fitted conditionally on the estimates of the first term, with

$$\log(\mathcal{L}_2|\hat{\boldsymbol{\kappa}}) = \sum_{t|m_t>0} \log\{\lambda(t)\} + \log[c_d\{K(m_t|\hat{\boldsymbol{\kappa}}), \dots, K(m_{t-d+1}|\hat{\boldsymbol{\kappa}})\}\, k(m_t|\hat{\boldsymbol{\kappa}})] + \sum_{t|m_t=0} \log\{1-\lambda(t)\}, \tag{3.31}$$

where c_d is subsetted according to the $m_{t-1}, \dots, m_{t-d+1}$ non-zero previous marks.

Since the intensity of the Hawkes process at time t depends on all past observations and that their order in time matters, the evaluation of the likelihood cannot be paralleled in principle. However with long time series at high frequency, evaluating the likelihood on a few separate long blocks of the time series should not make any significant difference in the inference, since the length of these blocks remains much larger than the duration for which the second-order kernel in Eq. 3.21 is significantly above 0. The small error induced by the parallelism could be reduced by having the block slightly overlapping, so that the intensity and copula at the beginning of each block could still be computed on some previous observations, but without computing twice the likelihood of the overlapping observations. This correction has not been implemented because it would make an insignificant difference in this application, given the length of the time series.

3.5.2 Simulation

Risk assessment from the model is obtained through simulation with the algorithm 1, where W_d^{-1} is the quantile function of the conditional EGP, with the copula subsetted according to the non-zero previously simulated values in $m_{t-1}, \dots, m_{t-d+1}$. The first $d-1$ values could either be set to 0, or the copula can be subsetted to accommodate for the beginning of the simulation.

In the case of a two-dimensional copula, W^{-1} uses an inverse h-function. Instead if the EGP is combined with a canonical vine copula in W_d^{-1}, a random value must be generated from the conditional cdf of the vine copula $C_{1|2,\dots,d}$. The standard simulation algorithm for a canonical vine copula is recursive, with random values being generated from the first to the last dimension, conditionally on the values generated for the previous dimensions (Joe 2015). Therefore, it would only be possible to adapt this algorithm to obtain the conditional distribution of the last dimension $C_{d|1,\dots,d-1}$, but the values simulated for the first dimension would be uniformly distributed. Instead, a value is generated by rejection sampling in the first dimension of the pdf of the canonical vine copula, with the values for the other dimensions given by $K(m_{t-1}), \dots, K(m_{t-d+1})$.

Data : Length T of the time series to simulate.
Result : A time series of marks $\boldsymbol{m}$ of length T.
for $t \leftarrow 1$ **to** T **do**
  $p = \lambda(t)$
  $a \sim \mathcal{U}(0,1)$
  if $p > a$ **then**
    $b \sim \mathcal{U}(0,1)$
    $m_t = W_d^{-1}(b|m_{t-1}, \dots, m_{t-d+1})$
  else
    $m_t = 0$
  end
end

Algorithm 1 : Simulation from the discrete Hawkes process marked by a conditional EGP.

Unlike the likelihood evaluation, the algorithm 1 cannot be paralleled. However the risk assessment from the model would be obtained from several simulated time series, so different simulations could be ran in parallel.

3.6 Application

3.6.1 Dataset

An application of the model is presented with a time series of 40 years of hourly rainfall. This time series is the total precipitation of the ERA5 reanalysis (Hersbach et al. 2018), from 1980 to 2019, at the location 49°N 0°E. This location is in mainland France, with an oceanic climate in the Köppen classification. Rainfall in this climate type is frequent, including storms which are responsible for heavy rainfall events. Snowfall is rare in oceanic climates, and does not significantly contribute to flood risk as it would in other climates, so not accounting for this different type of precipitation in the model is not an issue for this application. The applicability of the model is therefore restricted to areas where the only type of precipitation contributing to flood risk is rainfall.

ERA5 total precipitations have numerical artifacts in very low values (Fig. 3.4), which is a different issue from the discretization of station data mentioned in section 3.5.1, but could similarly impair the inferences. Therefore, a low threshold of $u = 0.5$ mm.h^{-1} is used, which avoids the eventual spurious correlation in the lower tail of the copula and reduces computational requirement for both parameter estimation and simulation.

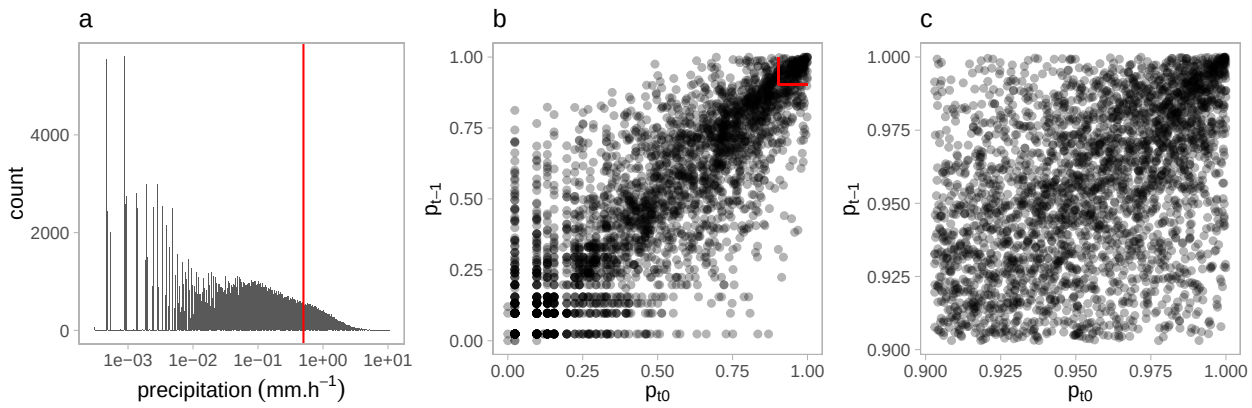


FIGURE 3.4 – Histogram (a) of the ERA5 hourly total precipitation time series. Empirical probabilities (b, c) of the precipitation plotted for 1 hour time lag. The red line indicates the $u = 0.5$ mm.h^{-1} threshold. The subplot c shows the region above the threshold in b. There are also values around 10^{-12} mm.h^{-1}, that are not displayed. Only 3000 values are displayed in b and c for readability.

Plotting the 40 years of precipitation along the day of the year reveals strong seasonal variability (Fig. 3.5), as it is expected for rainfall in most regions. This variability is also skewed, with the average values increasing faster during the first half of the year than they decrease during the second half (red curve on Fig. 3.5). In this case, upper values are higher during summer than winter. The seasonal effect is considered for the scale parameter σ, with a skewed sine wave. This season-dependent parameter is given by

$$\sigma(t) = \begin{cases} \sigma_0 + \eta\,\psi^{-1}\tanh\left\{\dfrac{\psi\,\sin(s_t-\phi)}{1-\psi\,\cos(s_t-\phi)}\right\} & \text{if } \psi \neq 0, \\ \sigma_0 + \eta\,\sin(s_t-\phi) & \text{if } \psi = 0, \end{cases} \tag{3.32}$$

where s_t in the time of the year mapped to $[0, 2\pi]$, $\sigma_0 > 0$, $\eta > 0$ is the amplitude of the seasonal effect, ψ controls the skewness of the sine wave, and ϕ is the phase.

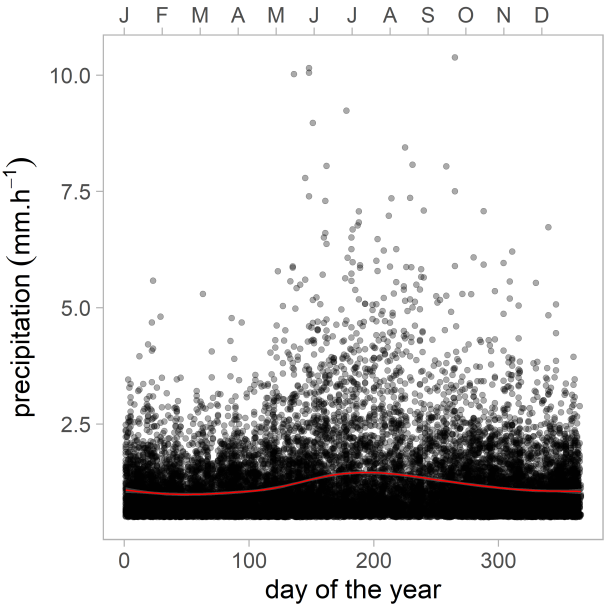


FIGURE 3.5 – 40 years of precipitation along the day of the year. The values smoothed along the day of the year (red curve) highlight the seasonal variability, which is also visible in the upper extremes.

3.6.2 EGP evaluation

The EGP-beta$_3$ with a seasonally varying scale parameter σ is strongly favored over a time-invariant model by the Akaike information criterion (AIC) and Bayesian information criterion (BIC) (not shown). Figure 3.6 presents the histogram of the marks (i.e., exceedances above u) along the EGP-beta$_3$ density for the minimum and maximum values of $\sigma(t)$. The difference between the two curves shows how a stationary model would underestimate the highest rainfall marks, compared to this seasonal model.

The shape parameter ξ (i.e., tail index) of an extreme value distribution is the most critical one, as it controls the heaviness of the upper tail. This parameter is generally positive for rainfall, which corresponds to a tail heavier than the exponentially tailed special case $\xi = 0$. SERINALDI et KILSBY (2014) found that the value of ξ for the GP should asymptotically be in the interval $(0.061, 0.097)$ with infinite sample size. In this work, the EGP-beta$_3$ is fitted on almost 17 000 exceedances above the $u = 0.5$ mm.h^{-1} (4.8 % of the 40 years of hourly time series), so the sample size is quite large from the perspective of the GP, even though only a fraction of these observations are extreme. The estimate of the tail index is $\hat{\xi} = 0.076$, which is in the range given by SERINALDI et KILSBY (2014).

The estimates of the other parameters are indicated in Table 3.2. If a constant σ is used instead of the seasonal $\sigma(t)$ of Eq. 3.32, the estimate becomes $\hat{\xi} = 0.148$, which could be an overestimation caused by ξ compensating for the lack of flexibility in σ. For $u = 0.5$ mm.h^{-1}, these results regarding the tail index are similar with the three other EGPs presented in section 3.2 : the EGP-power, EGP-beta$_1$ and EGP-genbeta.

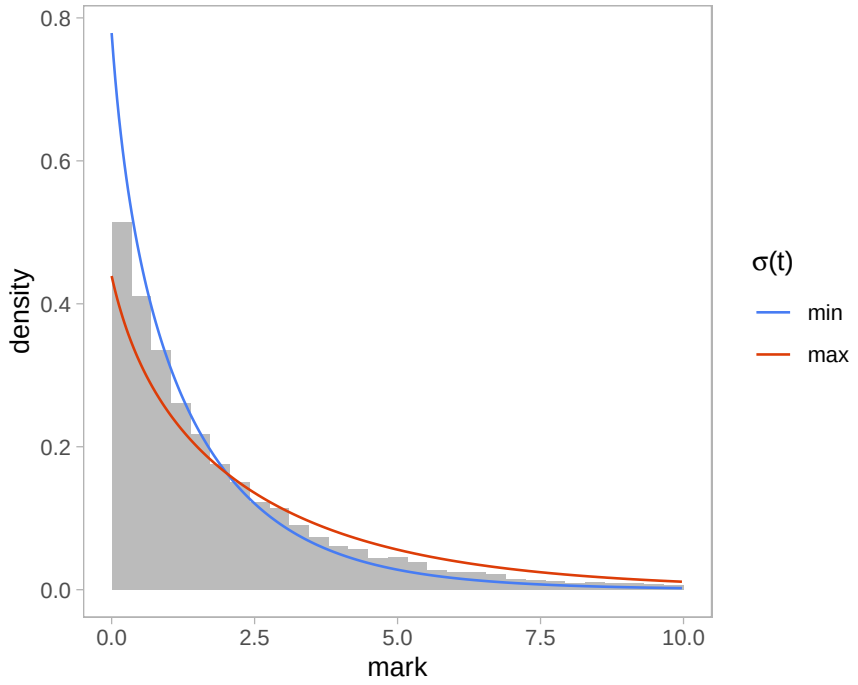


FIGURE 3.6 – Histogram of the marks and density of the EGP-beta$_3$ with a seasonally varying $\sigma(t)$. The EGP density is displayed for the minimum and maximum values of σ (which explains the apparent discrepancy with the histogram). The marks are standardized so are not in mm.h^{-1}, and the upper values are not displayed for readability.

The four EGPs are fitted with a lower threshold of $u = 0.05$ mm.h^{-1} to test if the extra flexibility of the EGP forms with more parameters is useful. In each case, the EGP is fitted both with a constant and seasonal parameter σ. Table 3.1 presents the estimated tail index $\hat{\xi}$ of each fitting. With this lower threshold, the lack of flexibility of the one-parameter EGP-power leads to a largely overestimated $\hat{\xi}$ of about 0.44, regardless of a constant or seasonally varying scale parameter σ. The EGP-beta$_1$ also lacks flexibility, with $\hat{\xi} = 0.178$ for both forms of σ. This EGP form is the only one with worse values of the information criteria AIC and BIC for the seasonal $\sigma(t)$. The fitting quality

improves significantly with the extra flexibility provided by the three- and four-parameter EGP-beta$_3$ and EGP-genbeta, respectively, with $\hat{\xi}$ estimates of about 0.09 falling in the $(0.061, 0.097)$ interval of Serinaldi et Kilsby (2014). These two EPGs also have lower values of AIC and BIC (note that Table 3.1 only reports the number of parameters of the probability transform of the EGP F_V, but the information criteria are computed according to the number of parameters of K, including the extra parameters when the seasonal $\sigma(t)$ is used). For this time series and a $u = 0.05$ mm.h^{-1} threshold, the EGP-genbeta does not significantly improve the fitting over the EGP-beta$_3$, but it can be assumed that the more flexible EGP-genbeta could be useful for rainfall time series at higher frequency than hourly. Interestingly, the results between the constant σ and seasonal $\sigma(t)$ EGPs are similar with this lower threshold, which indicates that the more complex seasonal form does not affect the estimated $\hat{\xi}$. The purpose of this comparison of the EGPs with a lower threshold is to demonstrate the potential usefulness of the more complex ones, but the EGP-beta$_3$ with $u = 0.5$ mm.h^{-1} discussed in the previous paragraph is used for the Hawkes process in the rest of the present work.

Table 3.1 – Inferences for four different EGP distributions with a lower threshold of $u = 0.05$ mm.h^{-1}.

F_V form	F_V parameters	seasonal σ	$\hat{\xi}$	AIC	BIC
power	1	no	0.438	143 072	143 104
		yes	0.448	142 945	143 004
EGP-beta$_1$	1	no	0.178	141 655	141 686
		yes	0.178	141 663	141 722
EGP-beta$_3$	3	no	0.096	141 514	141 564
		yes	0.109	141 418	141 495
genbeta	4	no	0.091	141 514	141 573
		yes	0.094	141 416	141 502

3.6.3 Hawkes process evaluation

Whether using the full likelihood of the Hawkes process in Eq. 3.29 or the two-step likelihood with equations 3.30 and 3.31, the parameters of the unconditional EGP K can be estimated separately with Eq. 3.31 for an exploratory analysis guiding the selection of the pair-copulas forms in the canonical vine copula. Software such as the R package *rvinecopulib* (Nagler et Vatter 2023) can automatically select the best pair-copula for a dataset among different parametric copulas, including their rotations. The probabilities of the non-zero marks are computed with the estimates

of the EGP with $K(m|\hat{\boldsymbol{\kappa}})$, then the best parametric form, including rotations, can be selected for each $d(d-1)/2$ pair-copulas of the canonical vine copula. As this automatic selection is only possible on the full d-dimensional vine copula in *rvinecopulib*, it is only done on the observations at time t having non-zero marks up to $t-d+1$. This restriction reduces the amount of data usable as d increases. This exploratory analysis is performed for $d=4$ to consider the lags up to $t-3$ in the vine copula. The unrotated BB8 pair-copula is selected for each pair-copula, among the nine forms and their rotations considered (see section 3.4.2 for the list of parametric pair-copulas tested). The BB8 pair-copula is asymmetric between its lower and upper tail, with both tails being asymptotically independent (Joe 2015).

After this exploratory analysis, the parameters of the Hawkes process and a 4-dimensional canonical vine copula for W_d are jointly estimated with the likelihood in Eq. 3.31. In this application $d=4$ because modeling up to lag $t-3$ gave better results than smaller vine copulas, while allowing stable parameter estimation. A principled size selection for the vine copula was not studied. Table 3.2 presents the estimates for the 30 parameters of the model, regrouped into its subcomponents.

Table 3.2 – Estimates of the 30 parameters of the model.

GP with seasonal $\sigma(t)$ (equations 3.1 and 3.32)				
σ_0	η	ψ	ϕ	ξ
2.073	0.575	0.382	2.112	0.076
EGP-beta$_3$ transform F_V (equation 3.8)				
κ_1	κ_2	l		
0.679	0.359	0.093		
Hawkes intensity (equations 3.20 and 3.21)				
μ_α	α	μ_β	β	γ
0.051	0.043	6.549	0.051	2.426
impact and link functions (equations 3.17 and 3.19)				
δ_1	δ_2	ω_1	ω_2	ω_3
1.983	0.273	6.524	0.005	0.002
canonical vine copula (all BB8 pair-copulas)				
$\boldsymbol{\theta}_{12}$	$\boldsymbol{\theta}_{13}$	$\boldsymbol{\theta}_{14}$	$\boldsymbol{\theta}_{23\|1}$	$\boldsymbol{\theta}_{24\|1}$	$\boldsymbol{\theta}_{34\|12}$
$\begin{bmatrix}2.587\\0.832\end{bmatrix}$	$\begin{bmatrix}6.277\\0.238\end{bmatrix}$	$\begin{bmatrix}5.246\\0.163\end{bmatrix}$	$\begin{bmatrix}4.204\\0.697\end{bmatrix}$	$\begin{bmatrix}6.700\\0.306\end{bmatrix}$	$\begin{bmatrix}2.683\\0.684\end{bmatrix}$

Figure 3.7 presents the second-order α and first-order γ kernels of the two-scale Hawkes intensity function (equations 3.20 and 3.21). These kernels behave as expected, since the second-order one

stays significantly above 0 for a much longer duration than the first-order one. The parametrization of 3.18 with γ appearing twice forces a value of 1 at lag 1, which simplifies the model fitting. A similar parametrization is used for the second-order kernel α (Eq. 3.21). The memory kernels are infinite, but in practice they drop to almost 0 at some point. To reduce the computational requirements, the memory kernels were censored for values below 10^{-4}. Forcing them to have a value of 1 at lag 1 also ensures that this censoring stays consistent for different parameter values (the censoring at a given threshold would be relatively higher or lower if the kernel had a value lower or higher, respectively, than 1 at lag 1).

Figure 3.8 presents 2 000 hours of simulated values (slightly less than three months). The top panel of Fig. 3.8 is the second-order clustering component μ_α^* of the intensity (Eq. 3.21). The middle panel is the full intensity λ controlling the probability of occurrence. The bottom panel is the marks (exceedances of the $u = 0.5$ mm.h^{-1} threshold). Occurrences inhibit the second-order component μ_α^*, which in turn increases the probability of first-order clusters. In other words, the inhibition of μ_α^* permits the succession of first-order clusters, consequently regrouping them in second-order cluster. These two orders of clustering are visible during the first half of the simulations. The second-order component stays at its maximum of $\mu_\alpha = 0.051$ for long periods in the second half of the simulation, during which the probability of occurrence without past influence reaches its lowest value with $\mu_\beta = 6.524$ (recall that this value is not a direct probability since the intensity in Eq. 3.22 is nonlinear). These periods of high μ_α^* correspond to dry spells in the marks (which are here only relatively dry since the threshold u is above 0). Note that a particularly dry period in the simulations is displayed here to illustrate the two-scale intensity.

Figure 3.9 presents the density of the three unconditional pair-copulas of the first tree T_1 of the 4-dimensional canonical vine copula (see Fig. 3.2). The correlation is higher between times t and $t-1$ modeled by the pair-copula c_{12}, than the subsequent lags $t-2$ and $t-3$ modeled by c_{13} and c_{14}, respectively.

Figure 3.10 presents the autocorrelation functions of the observations and 40 years of simulation. The autocorrelation is slightly underestimated at lag 1, then has a slight systematic overestimation from lag 4 and higher.

The IDF model is fitted on block maxima of both observations and simulations (40 years of hourly time series in both cases), to assess whether the Hawkes process reproduces the relation between amount of rainfall, duration and probability. Here the IDF model is only used as a diagnostic for the WG, and not as risk assessment. The parametrization of the IDF model is first selected based on

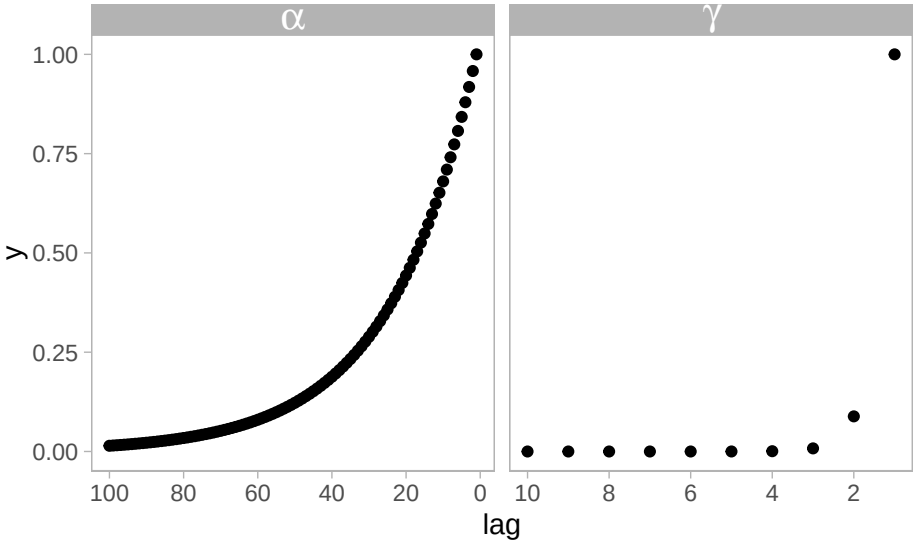


FIGURE 3.7 – Second-order α and first-order γ kernels of the two-scale Hawkes intensity. Note that these kernels are infinite, here only displayed for lags up to 100 and 10, respectively.

the best fit on the observations. As yearly block maxima are used, and for simplicity, the seasonal cycle is omitted. The parametrization of the IDF model is

$$G(z, d|\tilde{\mu}, \sigma_0, \xi, \theta, \eta) = \exp\left[-\left\{1 + \xi\left(\frac{z}{\sigma_0(d+\theta)^{-\eta}} - \tilde{\mu}\right)\right\}^{-1/\xi}\right], \tag{3.33}$$

where z are the block maxima for the duration d, $\tilde{\mu} > 0$, $\sigma_0 > 0$, $\xi \in \mathbb{R}$, $\xi \neq 0$, $\theta \geq 0$ and $\eta \in (0, 1]$. See ULRICH et al. (2020) for details on this IDF parametrization, which is implemented in the R package *IDF*. The same IDF parametrization is then fitted on the simulations from the Hawkes process. Since the vine copula in W_d (Eq. 3.27) is the primary component of the Hawkes process accounting for the duration of events, a simpler Hawkes process without copula was fitted and simulated from as comparison, with the marks modeled only by an unconditional EGP K (Eq. 3.3). Note that d in W_d refers to the vine copula dimension, while it refers to durations in the IDF model. The three IDF models are fitted for durations $d \in 1, \ldots, 120$ hours. Table 3.3 presents their estimated parameters. The tail index is estimated at $\hat{\xi} = -0.05$ for simulations from the model with vine copula, which is close to the estimate of -0.07 for the observations. However the tail index is lower at -0.16 for the simulations without vine copula. When IDF curves are plotted on a double log scale, the parameters θ and η (Eq. 3.33) control the curvature of the IDF curves for lower durations

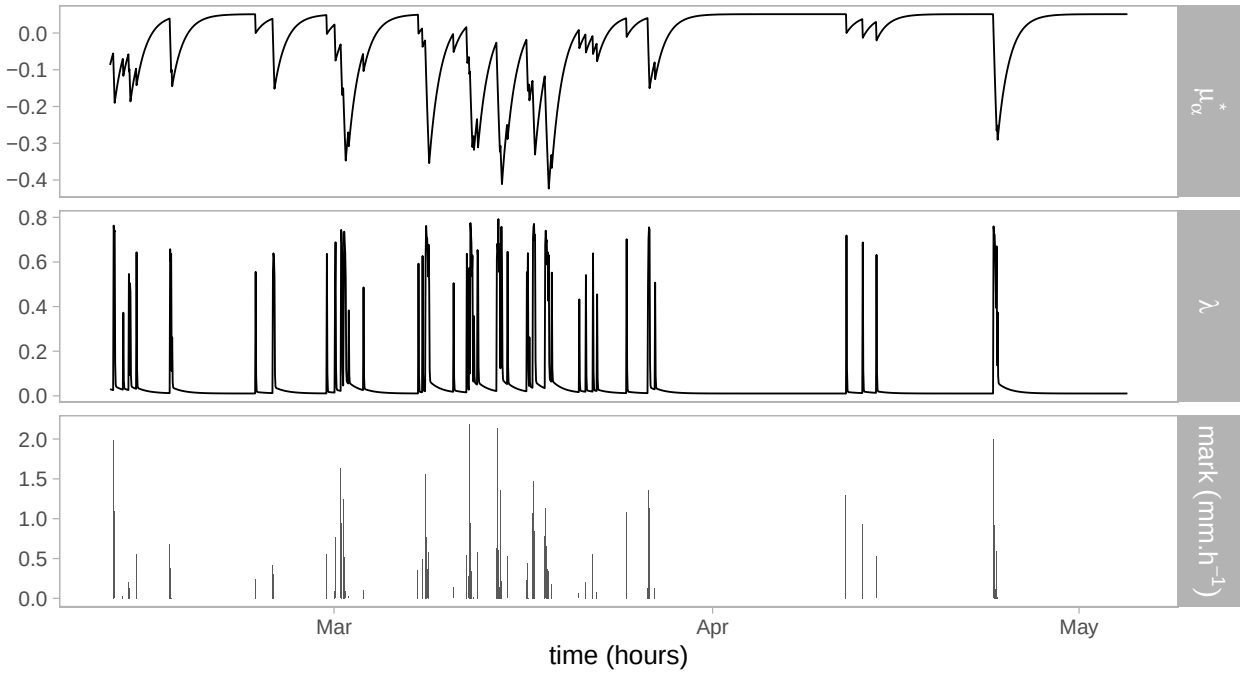


FIGURE 3.8 – Second-order clustering component μ_α^* (top panel) and full intensity λ (middle panel) of Eq. 3.22, and marks (bottom panel) for 2 000 hours of simulation.

and the slope for longer durations, respectively (ULRICH et al. 2020). For both IDF models from the simulations, $\hat{\theta} = 0$, whereas it is 1.08 for the observations. Likewise, both simulation models have an estimated $\hat{\eta}$ of about 0.6, while it is 0.8 for the observations.

Figure 4.11 presents the IDF curves for the yearly nonexceedance probabilities $\{0.9, 0.99, 0.999\}$. The curves from the simulations of the model without vine copula are systematically lower than the observations' curves, whereas the curves of the simulations from the full model with a vine copula are at the same level as the observations' curves, which is explained by the tail index ξ (Table 3.3). This indicates that the canonical vine copula in the conditional mark distribution W_d significantly improves the Hawkes process. If the autocorrelation of marks was not accounted for and only the autocorrelation of occurrences was modeled (i.e., in a Hawkes process with a mark distribution not conditional on past marks), the risk assessment obtained from the simulations would be underestimated. The observations' IDF curves display a curvature for the lower durations (here, under 10 hours), whereas there is none for both simulations, which is explained by the difference in the estimates of θ. Likewise observations' and simulations' curves have a different slope for

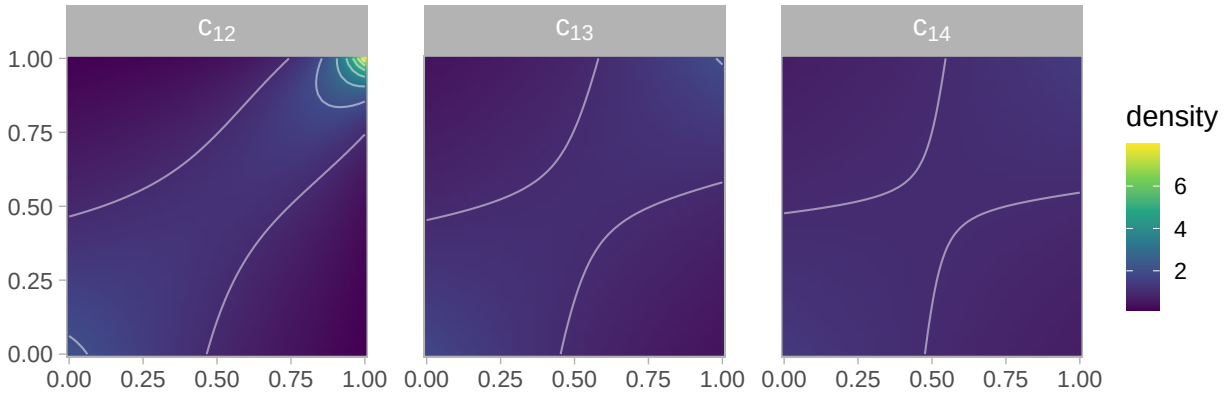


FIGURE 3.9 – Density of the unconditional pair-copulas in the first tree T_1 (top row of Fig. 3.2), modeling the dependence between the mark at t and previous marks at $t-1$ to $t-3$ from left to right, with the pair-copulas c_{12} to c_{14}, respectively. The three are unrotated BB8 pair-copulas.

higher durations, which is explained by the different estimates of η, and results in the probabilities being overestimated for longer durations in the simulations from the full Hawkes process. This overestimation at longer durations is probably linked to the overestimation in the autocorrelation seen in Fig. 3.10. These discrepancies between observations and simulations from the full model indicate that some patterns are not reproduced, and the current model is not reliable for longer durations.

TABLE 3.3 – Estimates of the IDF models for the observations, the simulations without vine copula and the simulations from the full Hawkes process with a 4-dimensional canonical vine copula. Note that θ is not a pair-copula parameter here, but a parameter of Eq. 3.33.

model	$\tilde{\mu}$	σ_0	ξ	θ	η
obs	3.87	2.67	-0.07	1.08	0.80
$\text{sim}_{\text{no cop}}$	3.82	1.88	-0.16	0.00	0.63
$\text{sim}_{\text{4d cvc}}$	5.12	1.00	-0.05	0.00	0.64

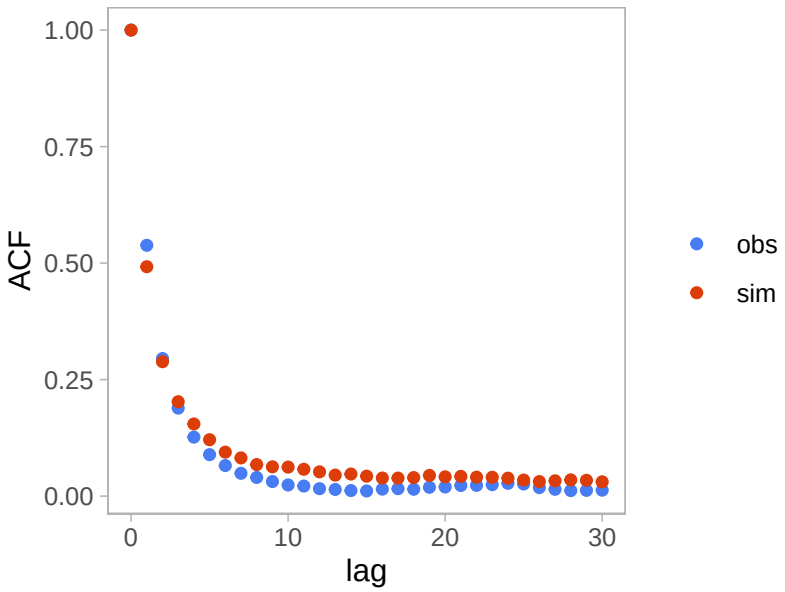


FIGURE 3.10 – Autocorrelation functions of the observations (exceedances of the threshold $u = 0.5$ mm.h^{-1}) and simulations (both 40 years of hourly rainfall).

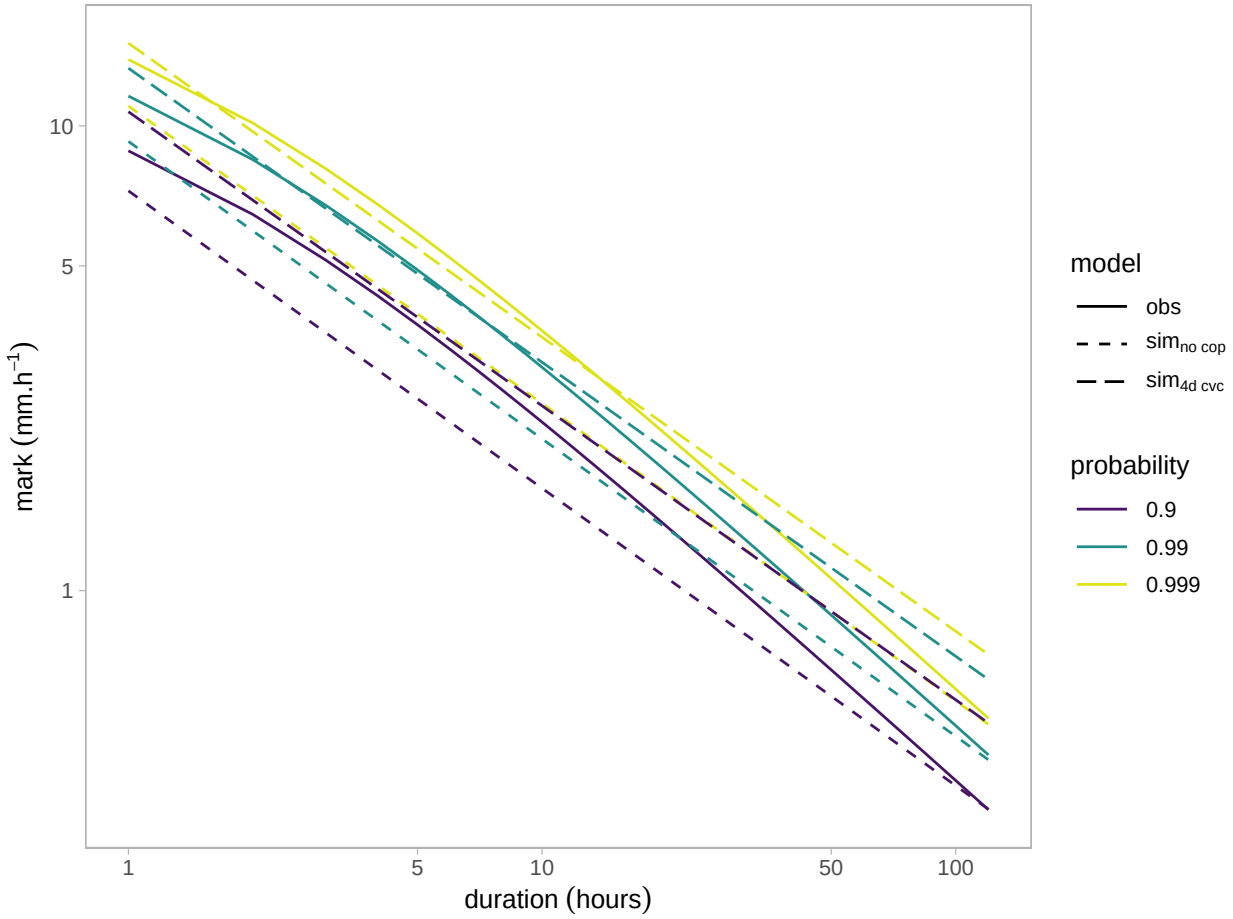


FIGURE 3.11 – IDF curves for the observations (obs, solid lines), simulations from the Hawkes process without vine copula ($\text{sim}_{\text{no cop}}$, short dash) and simulations from the full model with a 4-dimensional canonical vine copula ($\text{sim}_{\text{4d cvc}}$, long dash). The models are adjusted for durations of 1 to 120 hours. Both axes are in log.

3.7 Discussion and perspectives

The two major differences between the Hawkes process of the present work and the one of Li et al. (2021) are the addition of the vine copula to obtain a conditional distribution of the mark modeling their autocorrelation, and the use of a two-scale intensity function. Their Hawkes process was applied to daily river flow, with the sign of the river flow reversed, so that exceedances of the threshold (i.e., non-zero marks) corresponded to low flow values and hence to droughts (as opposed to this work where zero marks correspond to dry periods). In the model of Li et al. (2021), the scale parameter σ of the EGP for the marks is linearly dependent on the self-exciting part of the intensity function. This varying σ and the autocorrelation of occurrences will result in some autocorrelation of the whole time series in simulation, but not enough since the autocorrelation of exceedances (i.e., marks) is not explicitly modeled. In the dataset of Li et al. (2021), clusters of events almost only happen during the summer months, with only a few isolated non-zero values during winter. They accounted for this by having a partially deterministic intensity function, with the self-exciting component removed for winter months, during which the non-zero marks become Poisson distributed. It appears logical to have the yearly cycle (i.e., seasonal variability) as a deterministic component of the model, as it is the case for the scale parameter $\sigma(t)$ (Eq. 3.32) in the proposed model, but the resulting EGP and Hawkes intensity remained fully stochastic in the present work. In the model of Li et al. (2021), clusters of events can only happen from May to November, with this cutoff not being estimated through the likelihood but a priori selected. The two-scale intensity (equations 3.20 and 3.21) would not be appropriate for the river flow of Li et al. (2021), because the second-order of clustering in the dataset is the yearly cycle, which must be considered deterministic. However, the Hawkes process can remain fully stochastic if the yearly cycle is integrated in the model as a covariate of the intensity.

Only two clustering orders were considered in the proposed model, with the first-order being an artifact of the discrete time series and the second one being linked to the short-term meteorological variability. Higher orders of clustering are most probably present in the data. Precipitations in France are affected by the atmospheric pressure patterns of the North Atlantic Oscillation (NAO) (Hurrell et al. 2003), on a long-term time scale and large spatial scale. Massei et al. (2017) showed that taking into account large-scale and low frequency (i.e., lower than the yearly cycle) atmospheric pressure patterns improves the modeling of local-scale precipitation. For many regions, the El Niño-Southern Oscillation (ENSO) would be another climate pattern to consider for precipitation modeling (Sun et al. 2015 ; Ouarda et al. 2021). Such low frequency phenomenon could potentially affect the probability of rainfall, and thus could be incorporated in the Hawkes process

intensity. They could also not affect occurrences but affect the amount of rainfall via the marks distribution, or affect both occurrences and amounts. A WG using climate indices as covariate would require modeling them, then to either simulate or forecast the indices and subsequently conditionally simulate the rainfall. Lee et al. (2023) compared different forecasting methods for the ENSO and Pacific Decadal Oscillation indices.

The EGP-beta$_3$ and EGP-genbeta are defined to respect the condition of Eq. 3.5 proposed by Gamet et Jalbert (2022), but a more flexible version of these two EGPs could be defined by having the upper truncation boundary u as a parameter to estimate. This modification would result in the EGP only respecting the condition of Eq. 3.4. In the case of rainfall, the extra flexibility is required for the lower tail, and even more so as the threshold is lowered close to 0, or as the observation frequency increases, so for the application this extra flexibility for the upper tail was not required.

Despite the 4-dimensional canonical vine copula used in the mark distribution W_d, the autocorrelation function of the simulations did not fully reproduce the autocorrelation of the observations (Fig. 3.10). The choice of vine copula dimension and its parametric pair-copulas significantly affected the autocorrelation function of the simulations, so the lack of total adequacy of the vine copula must be the main reason for this discrepancy. Although the IDF diagnostics (Fig. 4.11) show that the vine copula improves the model, discrepancies between simulations and observations are also apparent here. A characteristic of rainfall data is the peak of density at the lower tail (Fig. 3.6), which could disproportionately affect the vine copula. This issue can be compounded by the artifact that can be present in the lower values of rainfall time series, either caused by the discretization of the observations or by numerical artifacts in reanalysis data (as in the ERA5 total precipitation). This could be corrected by a biased likelihood for the copula, giving more weight to the upper rainfall values.

A limitation of the current model is the a priori selection of the pair-copulas in the canonical vine copula, only justified by an exploratory analysis. Instead, the parametric pair-copulas could be selected jointly with the rest of the model by a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green 1995 ; El Adlouni et Ouarda 2009). This Bayesian inference algorithm swaps the parametric pair-copulas in the vine copula to give the posterior distributions of the parametric form and the parameters of each pair-copula (Min et Czado 2010 ; Gruber et Czado 2018 ; Chapon et al. 2023). Rotations of the pair-copulas could also be considered by a RJMCMC algorithm. Another a priori choice was the dimension of the canonical vine copula, which

could also be included in a RJMCMC algorithm. Such algorithm was considered to be outside the scope of the present work and was not incorporated in the proposed model.

The model was fitted in Stan through gradient descent optimization rather than Bayesian sampling to keep the computation time low. However, the current implementation can be inferred by both methods without modification of the code. The only modification needed would be to select parameter priors allowing stable inferences by the Hamiltonian Monte Carlo algorithm used in Stan. This would allow taking into account uncertainties of both the parameters and the process by simulating from the model different time series with different samples from the parameters' posteriors.

The subsets of the canonical vine copula were applied to an intermittent time series, but they could also be useful to handle missing values, for either intermittent or continuous time series. This would allow the use of all the available data to compute the likelihood, without a prior imputation step.

The first-order kernel γ drops from 1 at $t-1$ to 0.088 at $t-2$, then to almost 0 for further lags (Fig. 3.7). The second-order kernel α remains significantly above 0 for a much longer time. Therefore, a simplification of the model could be to have a first order Markov process instead of an infinite kernel for the first-order of clustering, in a hybrid of Hawkes and Markov processes. Taking only $t-1$ into account in this Markov process could allow a more complex form than a typical kernel used in a Hawkes process (exponential kernels in this work).

As mentioned in section 3.3.2, the first-order of clustering in rainfall time series, of several subsequent non-zero values, does not correspond to clusters in the real world. These clusters are an artifact due to the temporal discretization of rainfall observations. If the duration of rainfall events is omitted, each event can be considered as a point in time, which can be modeled as a point process. These events are not independent, so they must be modeled by a Hawkes process rather than a Poisson process. The duration of events would be accounted for in the mark, along with the amount, and a single kernel continuous Hawkes process would be adapted. In other words, the clustering scale which has physical meaning as a real cluster in the two-scale discrete Hawkes process is the second-order one. Therefore, a single kernel discrete Hawkes process for a rainfall time series only models the artificial clustering and misses the real world clustering (granted that no other part of the model accounts for it).

3.8 Conclusion

The Hawkes process presented in this article is both an extreme value model, extending the framework of the GP, and a stochastic WG able to simulate time series. The distinction between these two types of models is only a matter of perspective, as inferences from an extreme value model can be obtained through simulation, and the marginal distribution of a WG needs the extreme value theory to account for the heavy upper tail of rainfall. The application of the model to an hourly time series of 40 years shows that the Hawkes process reproduces the pattern between the amount and the duration of extreme events.

The three parts of the model are an EGP distribution for the non-zero values, a Hawkes process for the intermittency of rainfall and a copula for the temporal dependence. A new EGP parametrization for hourly rainfall is proposed. A two-scale intensity function for the Hawkes process models the two scales of clustering of non-zero values in hourly rainfall time series. The temporal dependence is modelled by a canonical vine copula, which can be subsetted for timesteps without rain.

Floods and droughts are the two types of extreme events at the two tails of the rainfall distribution. Even thought the focus was put on the upper extremes throughout the article, the Hawkes process simultaneously models both tails of the distribution. The low threshold of $u = 0.5$ mm.h^{-1} in the application was only to circumvent numerical artifacts in the ERA5 dataset, and most stations hourly rainfall time series would be censored below a value higher than 0.5 mm.h^{-1}, so the model is effectively applicable with a threshold of $u = 0$. Ward et al. (2020) advocate for considering droughts and floods together for risk reduction strategies, because of their many interactions. Jointly modeling them is the first step in this regard. The ability of a WG to handle both types of extremes is an advantage over the IDF model.

Code availability

The model is implemented in Stan (Carpenter et al. 2017), with the code available at https://github.com/antoinechapon/swg_param.

Data availability

The ERA5 hourly total precipitation dataset is available at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview.

Author contributions

AC : conceptualization, formal analysis, methodology, software, visualization, writing – original draft preparation, review & editing. TBMJO : funding acquisition, project administration, supervision, writing – review & editing. NB : funding acquisition, supervision, writing – review & editing.

Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs Program, and the French Institute for Radioprotection and Nuclear Safety (IRSN) for funding this research.

Chapitre 4

Générateur stochastique par apprentissage profond (article 3)

Ce troisième article propose un second générateur stochastique pour série temporelle à une station, mais en remplaçant la majorité des éléments paramétrique par des réseaux de neurones. L'apprentissage profond permet de reproduire des patrons non-linaires complexes et de traiter des gros jeux de données, mais il ne peut pas reproduire le comportement asymptotique des valeurs extrêmes. Autrement dit, un réseau de neurone ne peut pas trouver par lui-même la théorie des valeurs extrêmes (sauf preuve du contraire).

Comme le modèle paramétrique du chapitre précédant, le modèle par apprentissage profond repose sur une distribution EGP, mais le seul élément paramétrique restant est la distribution GP. La contrainte d'avoir au moins la distribution GP comme élément paramétrique du modèle, et le fait qu'il doit être un générateur stochastique, font que ce modèle ne pouvait pas avoir une architecture standard de réseau de neurone. En particulier, la fonction objectif du modèle doit garantir qu'il converge vers un générateur stochastique reproduisant les patrons extrêmes (i.e. comportement de la queue haute et dépendance temporelle des extrêmes).

Ce dernier modèle de la thèse a des biais, ce qui le rend inapplicable en l'état. Cependant c'est une base plus solide que le modèle paramétrique, qui pourrait être étendue dans différentes directions.

Article 3

Extended generalized Pareto neurale pour un générateur stochastique de séries temporelles de pluie

auteurs : A. Chapon, T. B. M. J. Ouarda, N. Bertrand

journal : TEMP

soumis le 01/09/2025 à Advances in Water Resources, en cours de révision

Contributions : AC a conceptualisé le modèle, codé le modèle, produit les résultats, écrit la première version de l'article et la version révisée. TBMJO et NB ont supervisé le projet et révisé le différentes versions de l'article.

Abstract

A stochastic generator (SG) for rainfall time series is built by combining a neural network and an extreme value distribution. This offers the ability of deep learning to model complex non-linear patterns while also allowing extrapolation of the upper tail in accordance with the extreme value theory. The neural EGP is trained via maximum likelihood. The temporal dependence is accounted for via convolution layers across time, but it is also explicitly targeted in the loss function with a conditional distribution. Rainfall intermittency is represented by a dedicated model component. Simulations from the SG reproduce several patterns of the observations, which are the marginal distribution, the temporal dependence, and the annual variability. The distribution of dry periods and the IDF pattern are reproduced with some bias. The neural EGP is demonstrated on a time series of daily rainfall from a station in France.

4.1 Introduction

SGs are statistical models aiming to reproduce the statistical patterns of observed time series and/or spatial fields in simulations. Risk assessment is one of their application, as the probability of an event to occur can be inferred from the simulations. The quality of a SG lies in its ability to reproduce many statistical patterns as accurately as possible. The main patterns to reproduce for rainfall risk assessment are the marginal distribution and the temporal dependence, which includes the duration of events. Our model focuses on rainfall time series at a single station, so spatial aspects will not be covered. Several recent SGs for rainfall used an distribution des valeurs extrêmes (EVD) for proper extrapolation of the upper extremes beyond the range of observations (Evin et al. 2018 ; Ahn 2020 ; Beneyto et al. 2023). These models used an EGP distribution (Papastathopoulos et Tawn 2013) for the whole range of rainfall, from low to extreme values, which is a class of distributions giving more flexibility to the EVD GP.

Since SGs aim to reproduce as many patterns of the observations as possible, they are necessarily complex models. These models do not aim to simplify the data to make it understandable, but rather to reproduce both broad and fine patterns, including the relationships between them. Deep learning is successful at modelling non-linear patterns, which makes it a good option for SGs. However, neural networks do not inherently extrapolate in accordance with extreme value theory, unless specifically designed or constrained to do so. Extending the GP to transform it into a distribution for the whole range of rainfall can be done in several ways, using either parametric (Naveau et al. 2016 ; Gamet et Jalbert 2022) or semi-parametric (Tencaliec et al. 2020) approaches. To our knowledge, it

has not been done with deep learning yet. Therefore, we introduce a novel model using a neural network to extend the GP. This neural EGP is the main component of a SG for rainfall time series.

The EGP class of distributions extends the GP with a PIT, which is itself a distribution. The EGP will be presented in more details in section 4.2.1. We build a neural EGP by modelling the PIT as a neural network. A distribution can be obtained from a neural network by having its output be monotonic with respect to the variable, so that the network's output represents a cumulative distribution function (cdf). Its derivative with respect to the variable yields the corresponding probability density function (pdf), which can be obtained by the automatic differentiation. The loss of the neural network is the likelihood obtained from the pdf (Chilinski et Silva 2018). There are two main ways to make a neural network monotonic so that it represents a distribution, with soft- and hard-monotonicity (Sartor et al. 2025). Soft-monotonicity corresponds to a standard neural network, such as a stack of fully-connected layers, where additional terms in the loss penalize the output for not being a proper cdf. For example, a loss term penalizes negative density values. See Zeng et Wang (2022) for an example of soft-monotonicity to define a neural copula. Soft-monotonicity should not be used to define the PIT of an EGP as there is no guarantee that it will behave as a distribution everywhere on its support. In particular, there could be non-positive density in the upper tail of the PIT, which would in turn alter the upper tail of the GP. Hard-monotonicity corresponds to neural networks where the architecture is modified to enforce monotonicity. The most common way to do so is to constrain the weights of the network to be non-negative, which is what we use for the neural EGP. See Chilinski et Silva (2018) for more details on modelling distributions via neural networks with non-negative weights constraint, and Sartor et al. (2025) for other ways to enforce hard-monotonicity.

An advantage of SGs in risk assessment is their ability to jointly model rainy and dry periods. Different types of extreme events can compound, such as in a drought-to-flood event, so analysing them in a multivariate framework is relevant (Brunner 2023). Our focus is on the upper tail of rainfall, but results for the dry periods are presented. The intermittency of rainfall is modelled by a specific component in our SG, but it could be removed from the model to apply it to strictly positive variables (e.g., temperature or wind speed).

This paper introduces a neural network-based SG that extends the EGP framework to capture the whole range of rainfall behaviour, while explicitly modelling temporal dependence and intermittency. Such a model can serve a range of practical applications, from flood risk analysis to reservoir operation and urban drainage design, where accurate modelling of both extremes and dry spells

is essential. Section 4.2 presents the method of the neural EGP and the other components of our SG. Section 4.3 presents an application of the model to a daily time series from mainland France. However, the architecture of the model is not specific to this location. The model is not specific either to daily time step, although it was not tested for finer time resolutions. Section 4.4 discusses the current state and the potential future developments of the model. Section 4.5 concludes.

4.2 Methods

4.2.1 Neural EGP

Extreme values, defined as exceedances over a sufficiently high threshold, are modelled using the GP distribution, whose cdf is given by

$$G(y) = 1 - \left(1 + \frac{\xi\, y}{\sigma}\right)^{-1/\xi}. \tag{4.1}$$

The threshold of the GP can be lowered to sub-extreme levels by extending it with a PIT function F with support on $[0, 1]$ (Papastathopoulos et Tawn 2013). This results in an EGP distribution, denoted W, with the cdf defined as

$$W(y) = F(G(y)), \tag{4.2}$$

and the corresponding pdf given by

$$w(y) = f(G(y))\; g(y), \tag{4.3}$$

where f is the derivative of F with respect to $G(y)$, and $g(y)$ is the pdf of the GP. Lowering the threshold to zero allows an EGP to model the whole range of rainfall, from low to extreme values (Naveau et al. 2016).

A neural network can model a cdf if its output is monotonic with respect to its input, which include the observations. The pdf is obtained by differentiating the cdf with respect to the observations, thanks to the automatic differentiation capabilities of deep learning. This makes the model trainable by its likelihood (Chilinski et Silva 2018). The objective is to build a SG for time series, so we model the conditional distribution of y_t with respect to past values. We denote t as the current time step from the perspective of the model, during both training and simulation. $t - 1$ is the previous value, and so on.

A conditional pdf is given by $f(x_1|x_2) = f(x_1, x_2)/f(x_2)$, where $f(\cdot)$ is the pdf of the corresponding arguments. Let $F(\cdot)$ be a monotonic convex neural network, with inputs x_1 and x_2, and a single value corresponding to an unnormalized probability as output. The first-order derivative gives the unnormalized pdf $f(x_2) = \partial F(x_1, x_2)/\partial x_2$. The second-order derivative with respect to x_1 gives the joint distribution pdf, with $f(x_1, x_2) = \partial^2 F(x_1, x_2)/(\partial x_1\, \partial x_2)$. Thus the conditional pdf $f(x_1|x_2)$ is obtained from the neural network. To build a neural EGP, the network acts as both a PIT and a joint distribution, with

$$f(G(y_t)|y_{t-1}) \propto \frac{\partial^2 F(G(y_t), y_{t-1})/(\partial G(y_t)\, \partial y_{t-1})}{\partial F(G(y_t), y_{t-1})/\partial y_{t-1}}. \tag{4.4}$$

This could be extended to a distribution conditional on J past values with $f(G(y_t)|y_{t-1}, \ldots, y_{t-J})$, but doing so would require more gradient evaluation for each dimension. Instead, we treat past values $y_{t-1}, \ldots, y_{t-J}$ as independent and compute separate conditional distributions for each. The resulting EGP conditional on each past value is proportional to those independent distributions, with

$$w(y_t|y_{t-1}, \ldots, y_{t-J}) \propto \sum_{j=1}^{J} \frac{\partial^2 F(G(y_t), y_{t-j})/(\partial G(y_t)\, \partial y_{t-j})}{\partial F(G(y_t), y_{t-j})/\partial y_{t-j}}\, g(y_t), \tag{4.5}$$

which is here written as the full EGP density w, including the GP density (see equation 4.3). The advantage of doing so is that it only requires two gradient evaluations, regardless of J. The derivatives of the denominator in equation 4.5 are computed simultaneously in one gradient evaluation, then the derivatives in the numerator are also computed simultaneously in a second gradient evaluation.

The loss function of the neural EGP is the negative log-likelihood obtained from the density in equation 4.5. This loss function targets both the asymptotic pattern of the upper tail with the GP and the temporal dependence. Without this conditional distribution, the loss would only target a time-varying marginal distribution at time t. The model would partially reproduce the temporal dependence if the GP and PIT inputs have information about the past, but there would be no guarantee that the model converges to reproduce the temporal dependence as well as its architecture permits.

Instead of conditioning the EGP directly on the J past values, we use the history vector $\boldsymbol{h}$, which concatenated the J past values, a representation of the time of year, and convolution in time. Section 4.2.4 will present in detail how $\boldsymbol{h}$ is defined. The neural EGP is conditioned on the vector $\boldsymbol{u} = U(\boldsymbol{h})$, where $U(\cdot)$ is a neural network.

We follow Chilinski et Silva (2018) to normalize the output of $F(\cdot)$ to $[0, 1]$, which requires the input to have a know maximal value. Therefore, the last activation of $U(\cdot)$ must have a maximal value of 1, so the sigmoid function is used. The maximal possible output of the monotonic neural network is given by $F(\mathbf{1})$, where $\mathbf{1}$ is a vector of ones of the dimension of the input. The output is normalized with $F^* = F(\boldsymbol{y})/F(\mathbf{1})$, which is now a valid cdf on $[0, 1]$. Note that the monotonic neural network could be used as a distribution without this normalization, as the unnormalized pdf could be sampled from. Note that the model does not need this normalization in principle, as an unnormalized density is sufficient to train and simulate from, but it helps training the model as the unnormalized density can reach very high values, causing numerical errors. The normalized version of equation 4.5 becomes

$$w(y_t|\boldsymbol{u}) = \sum_{n=1}^{N} \frac{\partial^2 \left[F(G(y_t), u_n)/F(\mathbf{1})\right] / (\partial G(y_t)\, \partial u_n)}{\partial \left[F(G(y_t), u_n)/F(\mathbf{1})\right] / \partial u_n}\, g(y_t), \tag{4.6}$$

where $\boldsymbol{u}$ is of size N. The transformation of $\boldsymbol{h}$ with $U(\cdot)$ is not specific to this normalization and is used to increase the representational power of the model, but we also take advantage of it to normalize the PIT.

Figure 4.1 illustrates the forward pass of the neural EGP. The scale parameter σ is first computed to obtain probabilities of the GP with $G(y)$, which in turn are in input of the monotonic layers. The scale parameter σ of the GP is also obtained by a neural network, but without monotonic layers. The component in blue modelling the intermittency of non-zero values will be presented in section 4.2.3. The next section 4.2.2 will present the neural network components modelling the PIT, with monotonic convex blocs.

The shape parameter ξ of the GP is not estimated jointly in the neural EGP. It is estimated by fitting the GEV distribution on yearly block maxima, then in input of the neural EGP. This parameter is notably hard to estimate in extreme value distributions. In the case of the present model, ξ has a negligible influence on the loss function (via the likelihood), so it is not straightforward to estimate it in the neural network. In the forward pass, ξ is in input of the GP cdf (figure 4.1). Similarly, it is used as input to the GP quantile function during simulation from the trained model. It is the only parameter held constant over time.

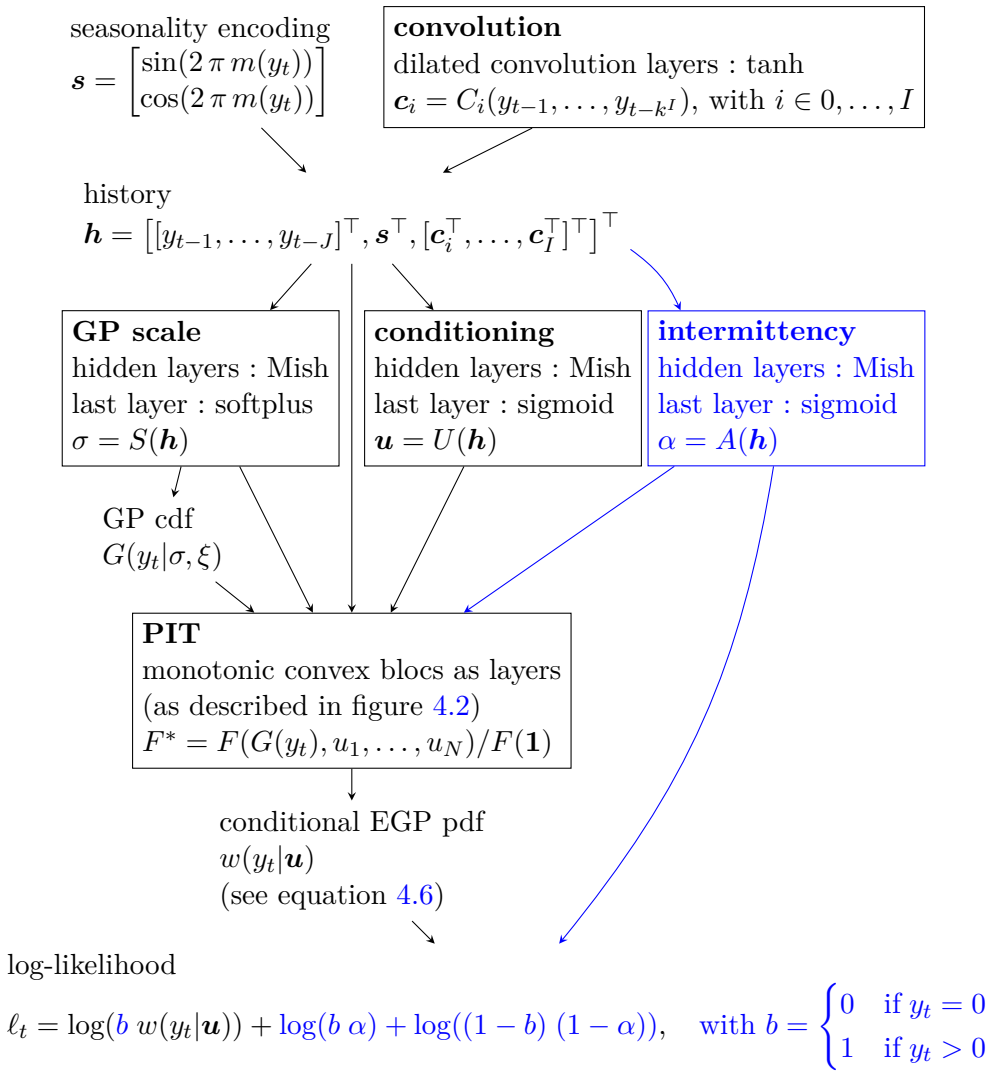


FIGURE 4.1 – Forward pass of the neural EGP for a single timestep t. The four neural network components of the model are denoted by boxes, which indicate the layer types with their activations (fully-connected layers if not specified otherwise) and outputs. Blue elements model the rainfall intermittency via the probability of dry timestep (here days), and would be removed for a non-intermittent variable. $u_{t-j} = \tanh(y_{t-j})$ for the normalized version of the PIT, see equation 4.6. Note that the subscript t is indicated for $\boldsymbol{y}$ to distinguish between past values $y_1, \ldots, y_{t-1}$ and y_t, but $\boldsymbol{s}$, $\boldsymbol{c}$, α, σ and b are also varying in time.

4.2.2 Convex monotonic neural network

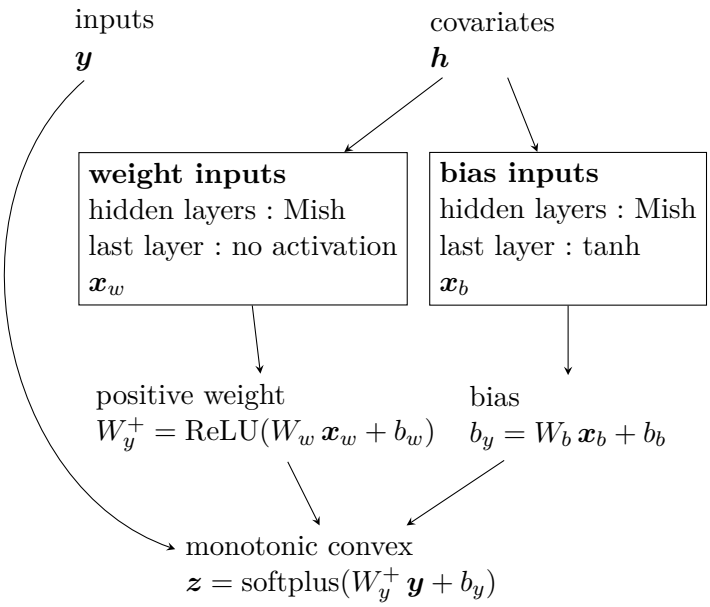


FIGURE 4.2 – Monotonic convex bloc, which can be stacked as layers to model a multivariate cdf. The two boxes represent standard fully-connected networks with outputs $\boldsymbol{x}_w$ and $\boldsymbol{x}_b$, from which the positive weight matrix W_y^+ and the bias vector b_y are computed, respectively. The higher-order derivatives of $\partial^K \boldsymbol{z}/\partial y_1, \ldots, \partial y_K$ are positive, so that this bloc, or several layered blocs, can represent multivariate cdfs and the corresponding pdfs.

To model a joint or conditional distribution, a neural network needs to be monotonic and convex, so that the higher-order derivatives are positive (CHILINSKI et SILVA 2018). Monotonicity can be enforced by having the weights of the network be positive. Convex activations, such as the softplus function, guarantee that the higher-order derivatives remain positive, so that they can represent densities. We define a new neural network block with these constraints, which can be stacked as layers to model a joint or conditional distribution. Figure 4.2 presents this bloc, which takes $\boldsymbol{y}$ as the variables of the multivariate cdf, and $\boldsymbol{h}$ as covariates (which is the history vector, described in section 4.2.4). The output of the bloc is only convex and monotonic with respect to $\boldsymbol{y}$, not $\boldsymbol{h}$. The covariates $\boldsymbol{h}$ are in input of standard neural network layer stacks (e.g. fully-connected layers), from which the weight W_y^+ and bias b_y of the monotonic convex operation are computed. This way, the constraints of monotonicity and convexity are not on the covariates.

4.2.3 Intermittency of rainfall

The probability α of non-zero value (i.e. rainy timestep) accounts for the intermittency of rainfall. This is similar to combining the EGP with a mixed-uniform distribution having a probability mass of $1 - \alpha$ at 0 (Papalexiou et al. 2023). The resulting log-likelihood combining the EGP and the probability of non-zero value is

$$\ell = \log(b\, w(y)) + \log(b\, \alpha) + \log((1-b)\,(1-\alpha)), \qquad \text{with } b = \begin{cases} 0 & \text{if } y = 0 \\ 1 & \text{if } y > 0 \end{cases}. \tag{4.7}$$

This likelihood is equal or proportional to elements on the the right hand side, depending on using the normalized of unnormalized PIT, respectively. The loss of the neural network is the corresponding negative log-likelihood.

The probability α is in input of the neural networks computing the GP scale parameter σ and the PIT (figure 4.1). The EGP is also directly conditioned on α, similarly to the J past observation values. This is not indicated in equation 4.5 as it is specific to the intermittency, but it is indicated on figure 4.1.

Note that zeros are not generated by the EGP, and the density of the EGP for dry timesteps is not included in the log-likelihood of equation 4.7.

4.2.4 Temporal convolution

The EGP distribution is varying over time to capture the temporal dependence of rainfall. This includes the GP scale parameter σ, the vector $\boldsymbol{u}$ and the PIT, which are all dependent on the past values encoded in $\boldsymbol{h}$. One exception to this is the GP shape parameter ξ, which is constant over time. Temporal neural convolutions are used to represent the patterns of past values in $\boldsymbol{h}$. These convolutions are causal, meaning that for a given time step, they are computed only on past values (see figure 4.3). The different convolution levels are dilated to capture long-range temporal dependencies, but keeping a constant kernel size k. Dilations exponentially increase the range of convolution by skipping values, as depicted in figure 4.3 (Espeholt et al. 2022). For dilations rates k^i with $i \in 0, 1, \ldots, I$, the convolutions are computed on the past k^{I+1} values. Each ith dilation rate corresponds to a convolution layer C_i, outputting the feature vector $\boldsymbol{c}_i$. Instead of only taking the feature vector $\boldsymbol{c}_I$ of the last layer as final output of the convolutions, the feature vectors of each dilation rate are concatenated in the "history" vector $\boldsymbol{h}$ and used in subsequent parts of the model

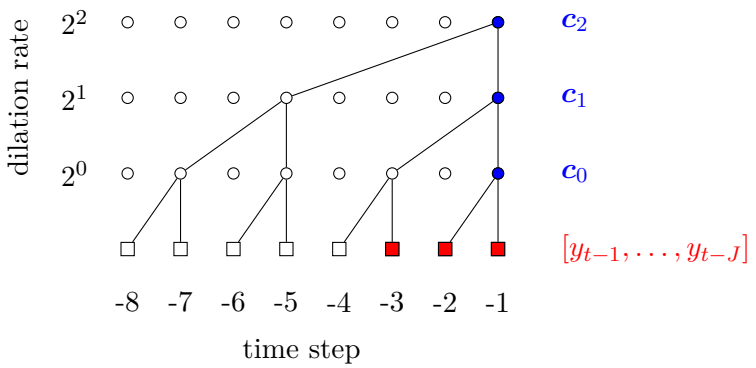


FIGURE 4.3 – Dilated causal convolution in time. Example of a kernel size $k = 2$ and $I = 2$, with dilation rates up to k^I. The square of the bottom row represent the observations, while the circles and edges represent the convolution levels. The convolutions a time $t = 0$ are computed on past values up to $t - 8$. The filled nodes indicate observations and convolution outputs that are concatenated in $\boldsymbol{h}$ (along the season vector $\boldsymbol{s}$). Each node in blue represents a feature vector of convolution. Nodes in red represent a single value of the observations, which are concatenated in $\boldsymbol{h}$ up to $t - J$ ($J = 3$ in this example).

(see figure 4.1). If $\boldsymbol{h}$ included only the last convolution level with dilation factor k^I, the patterns of every previous convolution level would be indirectly included in it, with a loss of definition in the finer patterns. In other applications of convolutions, for example in recognising hand-written digits in an image, taking only the output of the last layer is appropriate because the pattern of interest is the entire picture without distinction, rather than a focus on a particular subset. In the case of the neural EGP, past values and convolution outputs that are closer to the present time step t are more relevant than those further back in time, in most cases. Likewise, the short-term patterns of the lower convolution levels are also potentially more relevant than the longer-term patterns of the higher convolution levels. For the same reason, past values $y_{t-1}, \ldots, y_{t-J}$ are also concatenated in $\boldsymbol{h}$. The filled nodes in figure 4.3 represent the values concatenated in $\boldsymbol{h}$.

The annual cycle (i.e. rainfall seasonality) is encoded in $\boldsymbol{s}$ as a two-value vector, defined by

$$\boldsymbol{s} = \begin{bmatrix} \sin(2\,\pi\, m(y_t)) \\ \cos(2\,\pi\, m(y_t)) \end{bmatrix}, \tag{4.8}$$

where $m(y_t)$ is the day of the year of y_t mapped to $[0, 1]$. It is also concatenated in $\boldsymbol{h}$.

The history vector $\boldsymbol{h}$ of dimension H is given by

$$\boldsymbol{h} = \left[[y_{t-1}, \ldots, y_{t-J}]^\top, \boldsymbol{s}^\top, [\boldsymbol{c}_0^\top, \ldots, \boldsymbol{c}_I^\top]^\top \right]^\top. \tag{4.9}$$

In addition to k, I and J, H depends on the dimension of each convolution feature vector $\boldsymbol{c}_i$.

4.2.5 Time series generation

Simulation from the model is done by sampling from the PIT, which gives the probability of the GP, from which the corresponding quantile is computed. The main difference between the forward pass presented in section 4.2.1 and the simulation procedure is that the former computes the density of the whole EGP with equation 4.6, while the latter computes only the density of the PIT, with

$$f(p|\boldsymbol{u}) = \sum_{n=1}^{N} \frac{\partial^2 \left[F(p, u_n)/F(\mathbf{1}) \right] / (\partial p \, \partial u_n)}{\partial \left[F(p, u_n)/F(\mathbf{1}) \right] / \partial u_n}, \tag{4.10}$$

where p is a proposal for $G(y_t)$. At each simulated timestep, a vector of uniformly distributed proposals $\boldsymbol{p} \sim \mathcal{U}(0, 1)$ is generated. For each element of $\boldsymbol{p}$, the density of the PIT $f(p)$ is then computed with equation 4.10 A zero is then simulated for this timestep with a probability of $1 - \alpha$, or a non-zero value is drawn from $\boldsymbol{p}$ with a probability of $alpha \; f(p) / \sum f(\boldsymbol{p})$, where $f(\cdot)$ is applied element-wise to $\boldsymbol{p}$. $G^{-1}(p)$ gives the non-zero value corresponding to the drawn marginal probability p.

Simulation is initialized by a warm-up period of k^{I+1} values, which is as long as the range of the convolution layers. These initial values are discarded.

4.3 Application to daily rainfall

The performance of the model is evaluated by its ability to accurately reproduce patterns of the observations in simulations. These patterns include the marginal distribution, annual variability, dry periods, temporal dependence, and IDF pattern. Rainfall amounts, denoted as y, were normalized and are therefore presented in a dimensionless form.

4.3.1 Data

The SG is demonstrated with a daily rainfall time series of 62 years, from 1960 to 2023. These are observations from the station of Mount Aigoual, located in the south of mainland France (44.1°N 3.58°E). This station is in the southern part of the mountain range of the Massif Central, which experiences frequent extreme rainfall events with long durations (Blanchet et Creutin 2017).

4.3.2 Model parametrization

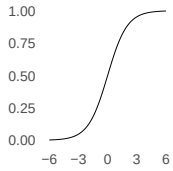
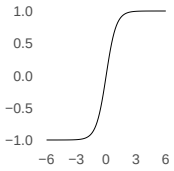
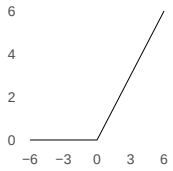
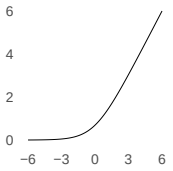
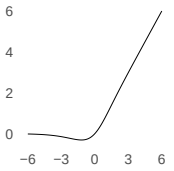
Table 4.1 details the parametrization of the neural networks components of the model (which correspond to the boxes in figures 4.1 and 4.2). The equations and graph of the different activation functions are presented in table 4.2 (details on the Mish activation can be found in Misra (2020)). The sigmoid activation is used as the last layer of the intermittency component because α is a probability. Likewise, the last layer of the scale component has the softplus activation because the GP scale σ must be strictly positive. The EGP is conditional on the history vector $\boldsymbol{h}$, which includes the $J = 10$ last values, the seasonality and the convolutions, all encoded in $\boldsymbol{u}$.

Table 4.1 – Layers parametrization of the neural network components. See figures 4.2 and 4.1 for how these components are linked.

name	layer type	layers dimensions	activations	output
convolution	dilated convolution with $k = 2$	$[8, 8, 8, 8, 8, 4, 4, 4, 4, 4]$	tanh	$\boldsymbol{c}_i$ with $i = 0, \dots, 9$
GP scale	fully-connected	$[32, 16, 8]$	Mish	
	fully-connected	1	softplus	σ
conditioning	fully-connected	$[64, 32]$	Mish	
	fully-connected	$N = 16$	sigmoid	$\boldsymbol{u}$
intermittency	fully-connected	$[32, 16, 8]$	Mish	
	fully-connected	1	sigmoid	α
PIT*	monotonic convex	$[32, 32, 32, 32, 1]$	softplus	$F(G(y_t), \boldsymbol{u})$, $F(\mathbf{1})$
weight inputs	fully-connected	32	Mish	
	fully-connected	16	none	$\boldsymbol{x}_w$
bias inputs	fully-connected	32	Mish	
	fully-connected	16	tanh	$\boldsymbol{x}_b$

* Each of the five monotonic convex blocs of the PIT as its own "weight inputs" and "bias inputs" components, as described in figure 4.2.

TABLE 4.2 – Activations used in the model.

sigmoid	**tanh**	**ReLU**	**softplus**	**Mish**
$1/(1+e^{-x})$	$\tanh(x)$	$\max(0,x)$	$\log(1+e^{x})$	$x\tanh(\text{softplus}(x))$
				
$(0,1)$	$(-1,1)$	$\mathbb{R}_{\geq 0}$	$\mathbb{R}_{>0}$	$\mathbb{R}_{>\approx -0.31}$

4.3.3 Marginal distribution

Figure 4.4 shows the 62 years of daily rainfall values, and a simulated time series of the same duration.

Figure 4.5 presents the density of the PIT obtained from the monotonic convex network, for one time step. The PIT density values greater than 1 towards $y = 0$ correspond to an increase of the EGP density for the lower values, compared to the GP. This PIT shape with more density for the lower values is as expected for rainfall (NAVEAU et al. 2016 ; GAMET et JALBERT 2022).

Figure 4.6 shows the densities of the PIT during simulations. Time steps for which the value sampled from the PIT has a high density tend to increase the density for high values for the subsequent time steps, thus generating a sequence of positive rainfall values corresponding to a rainfall event spanning several time steps. Time steps not showing the PIT density and with a probability of 0 correspond to dry time steps.

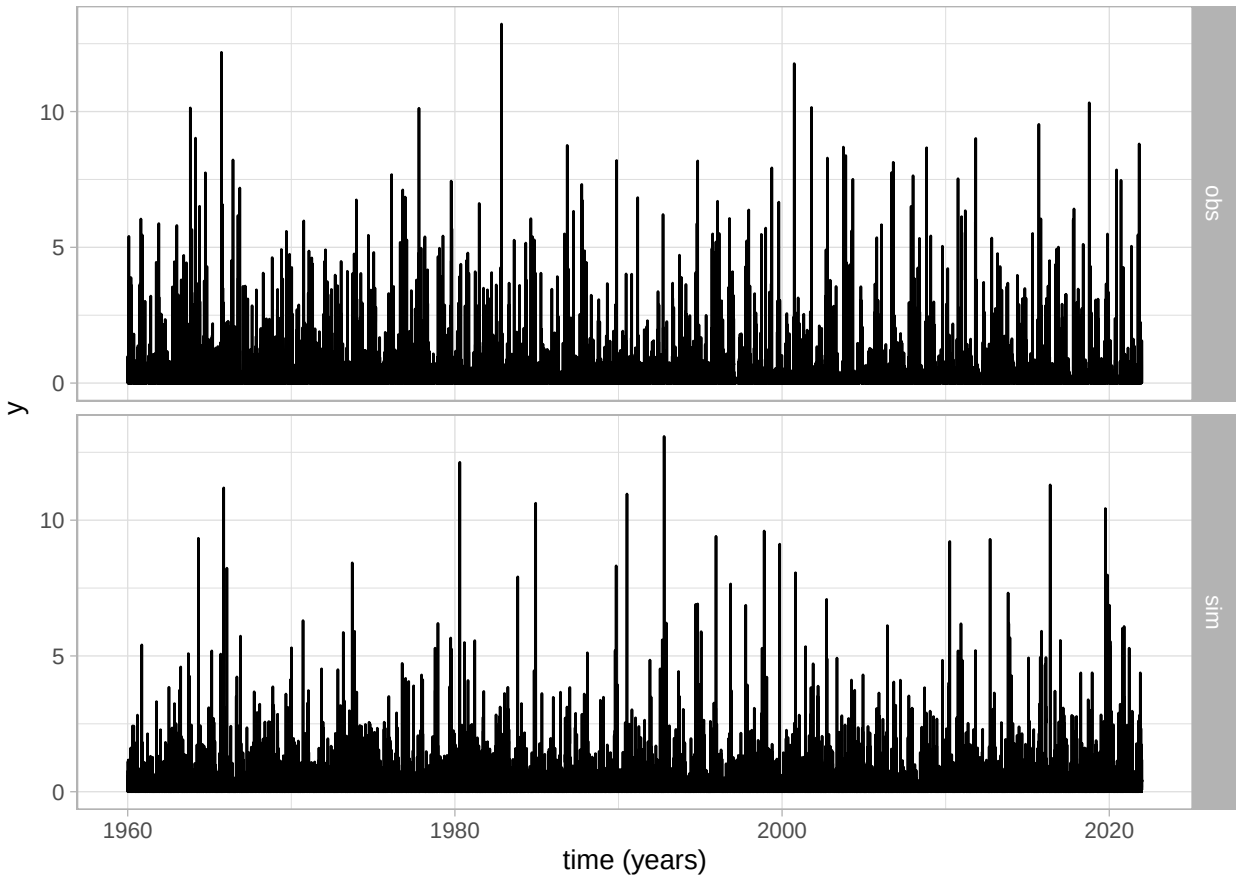


FIGURE 4.4 – Observations (top) and simulations (bottom) of 62 years of daily rainfall values.

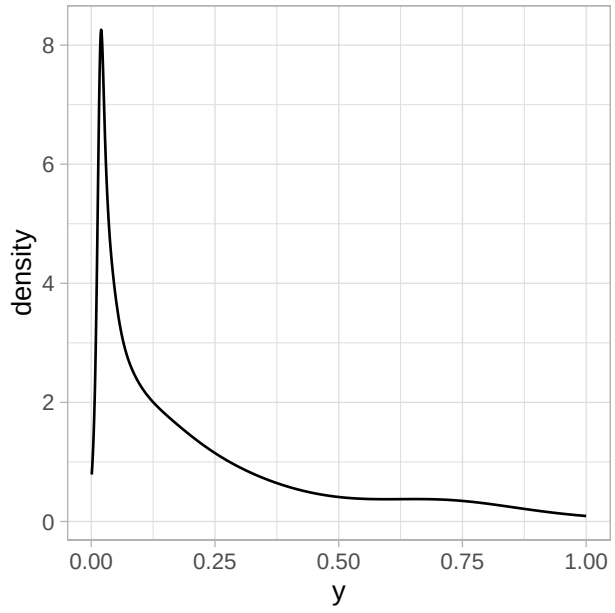


FIGURE 4.5 – Density of the neural PIT for one time step.

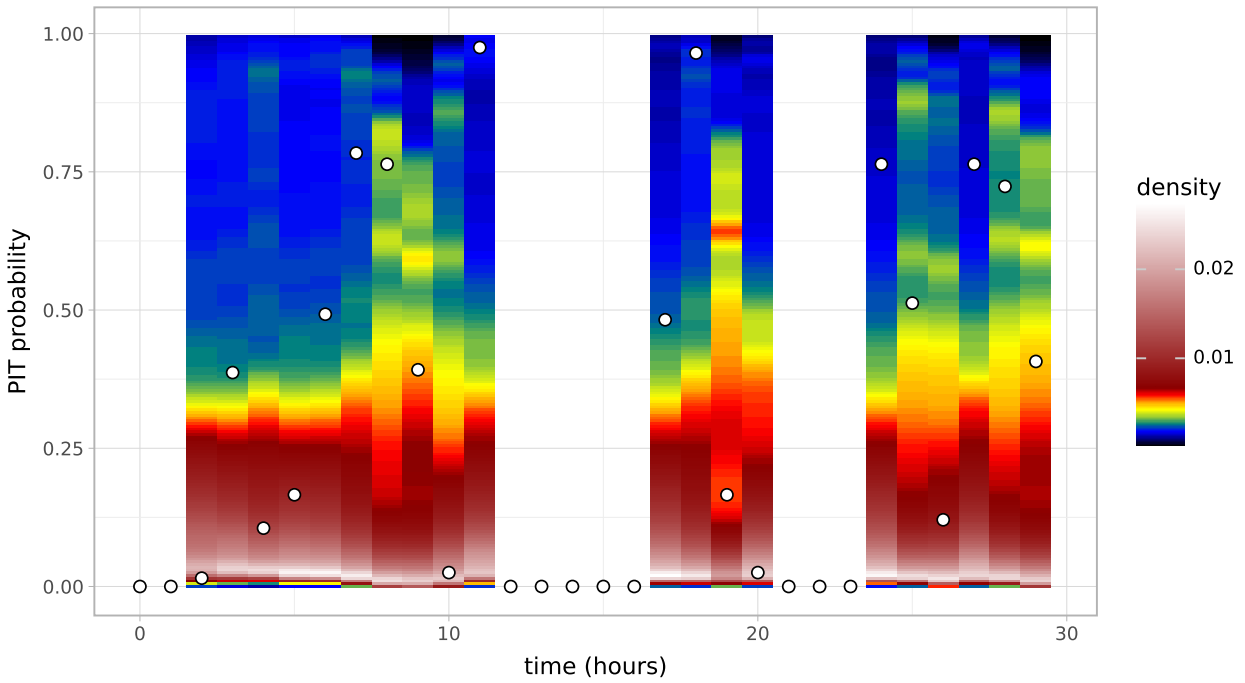


FIGURE 4.6 – Density of the PIT for 30 simulated values. For each simulated time step, the Hawkes process is first sampled, then the PIT density is sampled in case of a rainy time step. The color scale of the PIT density is non-linear for readability.

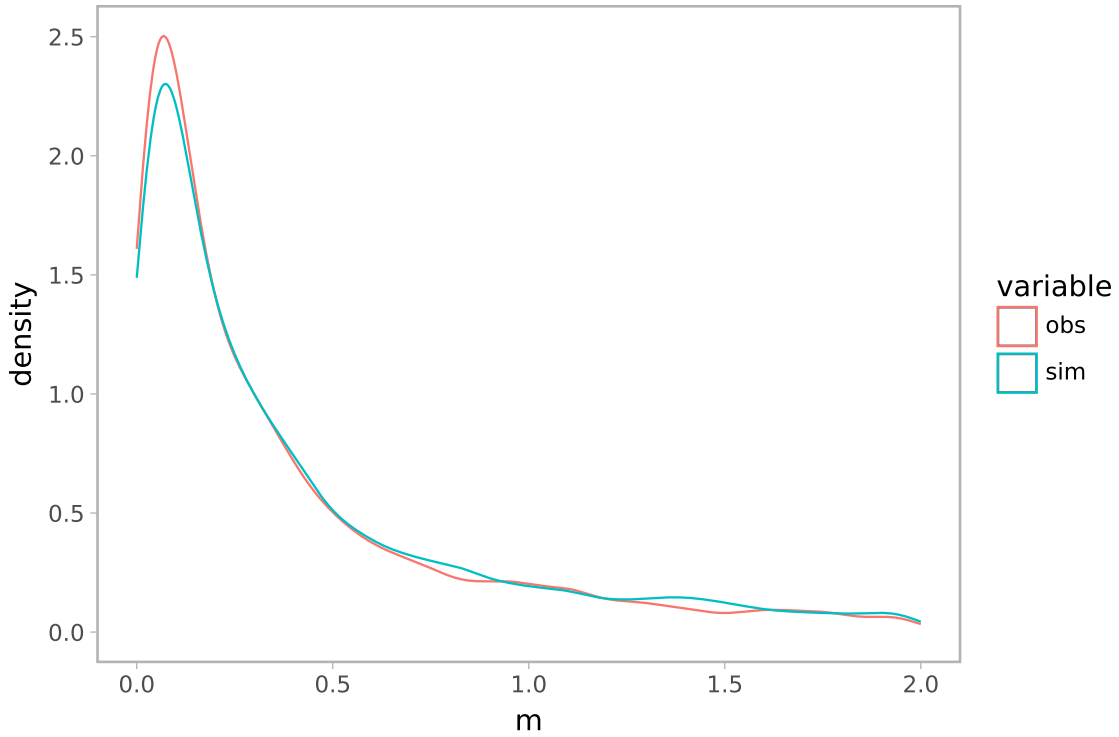


FIGURE 4.7 – Empirical distribution of the observed (red) and simulated (blue) non-zero values. High values are not displayed for readability.

4.3.4 Annual variability

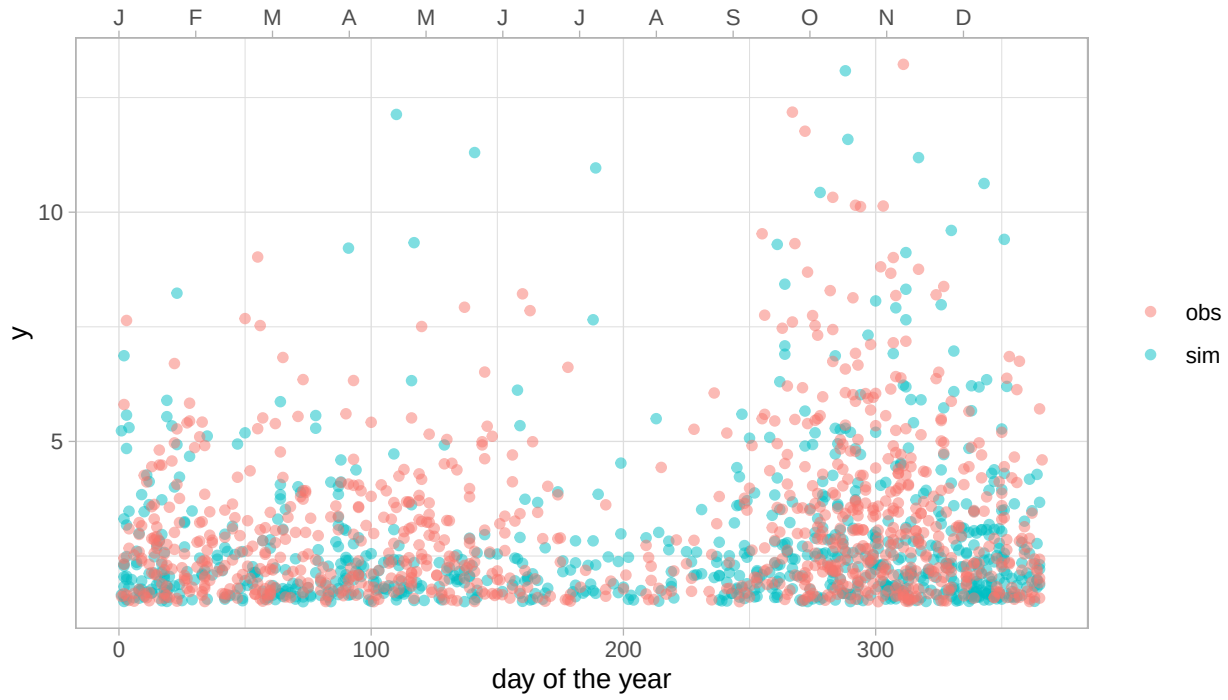


FIGURE 4.8 – Variability of non-zero observations (red) and simulations (blue) according to the annual cycle. Low values are not displayed for readability.

Figure 4.8 presents the 62 years of observation and simulations of the same length along the annual cycle. The annual variability is reproduced in the simulations, with lower values during the summer (i.e. middle of the year), then the most extreme values during autumn.

4.3.5 Distribution of dry periods

Figure 4.9 presents the empirical distribution of dry periods (i.e. successive zero-values) for simulations and observations. The figure is shown twice, with a modified vertical axis on the right for readability. The histograms of observed and simulated dry periods reveal a bias of the model, where too many short dry periods of a few days are simulated. Likewise, the simulations lack the longer dry periods seen in observations, with duration greater than 20 days. Note that these histograms as diagnostic are not perfect, as a dry period in simulations could be "cut" in two by a single very low positive value. With a threshold of 1.5 (in dimensionless rainfall), both histograms overlap for

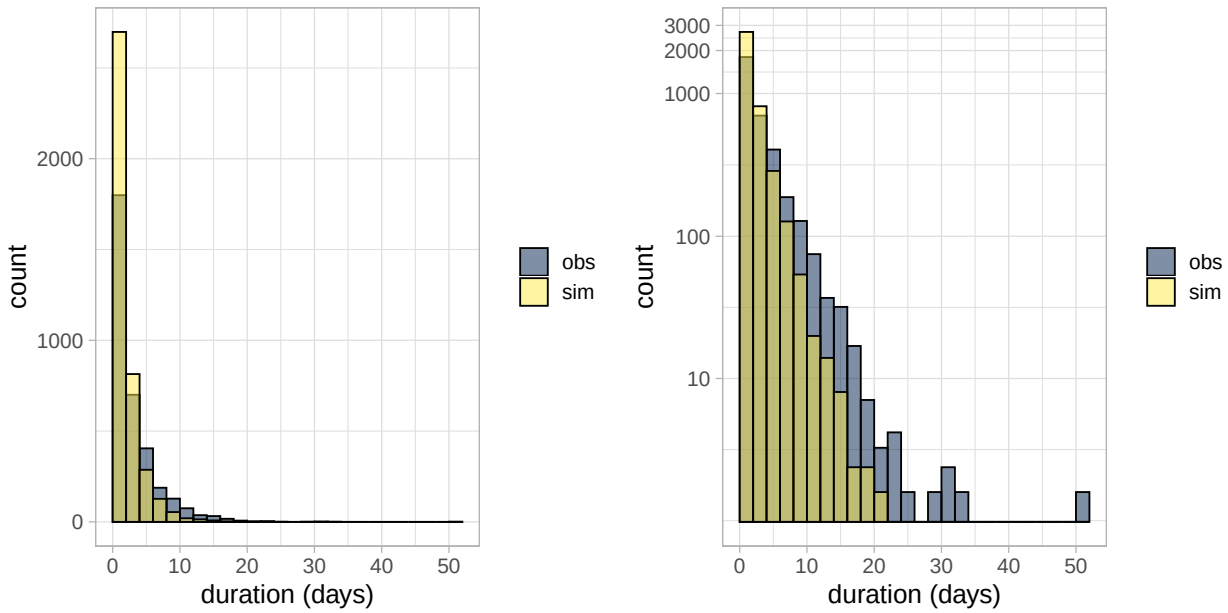


FIGURE 4.9 – Distribution of the duration of dry periods in observations (dark blue) and simulations (yellow). The third greenish color not indicated in the legend is the overlapping of both histograms. The same plot is represented on the right with a modified vertical axis in "pseudo" log, for readability of the higher durations.

short and long durations (not shown). That being said, the simulations do not fully reproduce the duration of dry spells.

4.3.6 Autocorrelation and extremal dependence

The reproduction of the temporal dependence is assessed by the autocorrelation and a measure of the extremal dependence for several time lags. Pearson's ρ is used on the whole time series. This diagnostic might show/could give a good result while hiding a bias in the serial dependence in the upper tail, so a measure of the extremal dependence between successive high values is also used. The tail dependence coefficient λ of SCHMID et SCHMIDT (2007) is computed on the successive values above a high threshold. This coefficient is based on Spearman's ρ.

Figure 4.10 shows that the temporal dependence is reproduced in the simulations for both the whole time series (ρ, right) and the upper tail (λ, left). The lag 1 ρ value is slightly higher in observations compared to simulations, which appears as a small difference, but these are computed on 62 years

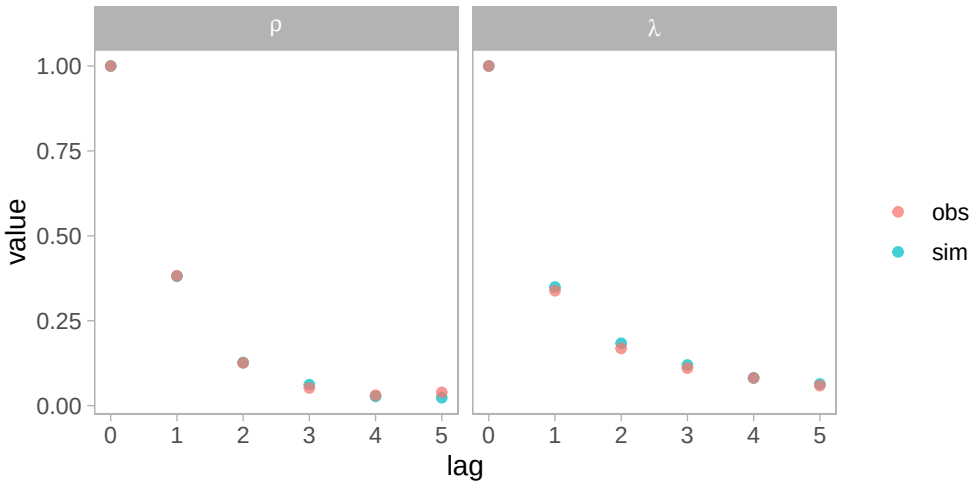


FIGURE 4.10 – Autocorrelation (Pearson's ρ, left) and extremal dependence (Schmid and Schmidt's λ, right) for observations (red) and simulations (blue) up to a 5 days lag.

of daily values, so it shows some bias. The reason for this bias could be a lack of flexibility of the neural PIT, or in the intermittency process. This figure shows that the conditional neural EGP successfully targets the temporal dependence.

4.3.7 Intensity-duration-frequency of extreme events

Figure 4.11 shows the IDF curves for the observations and 30 simulations. The parametrization of the IDF model follows ULRICH et al. 2020. The curves show that the model overall reproduces the IDF pattern in simulations. Some bias remains and can be seen for the 0.9 probability curves, which are slightly underestimated in simulations. The variability between the 30 simulations increases along the probability, which is to be expected for extreme values.

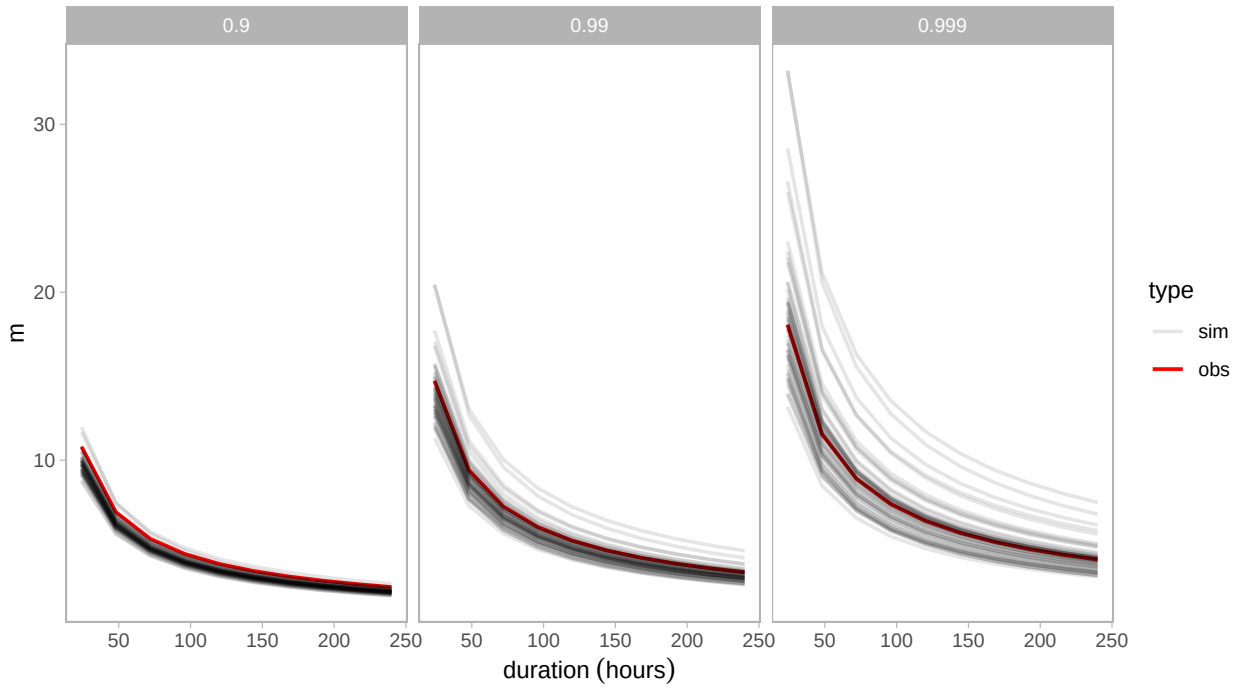


FIGURE 4.11 – IDF curves for the 62 years of observations (red) and 30 simulations of 62 years (grey), for durations of 1 to 10 days.

4.4 Discussion and perspectives

Several parametric SGs for rainfall estimate parameters per month to account for the annual variability (EVIN et al. 2018 ; BENEYTO et al. 2023 ; PAPALEXIOU et al. 2023). In principle, it is better to fit the model only once on the whole dataset. Fitting models separately for individual months or seasons is an artificial way to reproduce the annual variability pattern, but it reduces the ability of the model to reproduce patterns that could only be "seen" by looking at the whole time series. This season fitting approach increases uncertainties due to the reduced size of the data subsets used for parameters estimation. It also introduces artificial discontinuities in simulations as the parameters would jump from once value to the other at the changes of month or season. Our model shows that accounting for the annual cycle is more straightforward with deep learning, as encoding the season in the input and having convolutions in time is enough to reproduce this pattern. This allows fitting the model only once, on the whole time series.

The IDF model is an extension of the GEV block maxima model where the probability and intensity

of rainfall (or another variable) is linked to the duration of events (Koutsoyiannis et al. 1998). This model is common for extreme rainfall modelling and has been developed in several directions over the years, such as spatial (Ulrich et al. 2020) or non-stationary (Ouarda et al. 2018) variants. Simulations from the SG and observations should be indistinguishable from the perspective of the IDF model. In other words, reproducing the IDF pattern is one diagnostic of the reliability of the SG for risk assessment. The IDF curves of simulations correspond overall to those of simulations, but a slight bias remains, which hints at some refinements needed for the model. As a comparison, the parametric rainfall GS of Chapon et al. (2025) had IDF curves of simulations that did not present the same curvature as those of the observation. This shows that the present deep learning model improves the reproduction of the IDF pattern, compared to a parametric GS modelling the same type of data. The SG is applied to daily data in the present work, which also limits the pertinence of the IDF model as diagnostic. Future development of the neural EGP could use hourly data instead.

Being unable to estimate the tail index ξ of the GP simultaneously in the deep learning model, and having to estimate it separately in a previous step, is another limitation of our model. Estimating ξ jointly with the SG would allow it to vary according to the annual cycle. Coles et Pericchi (2003) showed that letting ξ vary according to the annual cycle improves the inference for rainfall extremes.

Other directions for future work could be to account for trend non-stationarity in the context of a changing climate, and account for the impact of climatic patterns via stochastic covariates such as climatic indices.

Lastly, the model could be extended to multiple stations. The challenge in this would be to simulate jointly for the different stations at a given time step. One way to achieve this would be to model the PIT of the EGP such that it capture both temporal and spatial dependencies. In the current model, the intermittency process is separate from the PIT. In the case of multiple stations, this process would also have to be a joint distribution between the stations. A possible improvement of the model would be to integrate the intermittency process in the PIT to truly have a single stochastic component, instead of separate components for the intermittency and the non-zero values.

4.5 Conclusion

An new type of EGP distribution is proposed by using a neural network as the probability integral transform extending the GP. This neural EGP takes advantage of deep learning strength in

modelling complex non-linear patterns, while retaining an extreme value distribution in the upper tail. The conditional form of the neural EGP explicitly targets the temporal dependence in the loss function. A stochastic generator is defined from the neural EGP, to simulate daily rainfall time series. Simulations reproduce several patterns of the observations relevant for risk assessment. These patterns are the marginal distribution, the annual variability and the temporal dependence, including the extremal dependence of successive high values. The intermittency of rainfall and IDF patterns of simulations present some bias, compared to the observations. However the basis of the model works as intended and demonstrates the potential of monotonic convex neural networks for stochastic generators.

Code availability

The model is implemented TensorFlow and R, with the code available at *(GitHub link added in published version).*

Acknowledgements

We thank the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs Program, and the French Nuclear Safety and Radiation Protection Authority (ASNR) for funding this research. We also thank Enzo Pinheiro and Freddy Houndekindo for their useful advices about neural networks.

Chapitre 5

Discussion et conclusion générale

Cette dernière section concluant la thèse synthétise les résultats des trois articles en section 5.1, discute les limites des modèles proposés en section 5.2, détaille l'applicabilité des modèles en section 5.3, puis identifie des pistes de développements futurs en section 5.4.

5.1 Synthèse

Les trois articles de la thèse proposent des nouveaux modèles pour imputer des données manquantes ou simuler des séries temporelles de pluie. L'objectif est de développer des modèles adaptés aux valeurs extrêmes. Ces modèles reposent sur des distributions multivariées, spatiale dans le cas du modèle d'imputation et temporelles dans le cas des générateurs stochastiques.

Dans le cas des modèles paramétriques, les copules en vigne permettent de définir des distributions multivariées adaptées aux valeurs extrêmes, car elles peuvent prendre en compte différents scénarios de dépendance extrémale pour chaque couple de variable. Les vine copulas sont des modèles emboités, ce qui leur permet d'avoir une dimension variable, en fonction de leur structure. Cette propriété est exploitée pour deux application différentes, dans le modèle d'imputation et dans le générateur stochastique paramétrique.

Les deux générateurs stochastiques de série temporelles de pluie, paramétrique et par apprentissage profond, ont le même objectif, mais utilisent des méthodes différentes. Leur point commun est d'être construits à partir d'une distribution EGP, qui est dans les deux cas liée à une distribution multivariée temporelle, afin d'obtenir une distribution conditionnelle aux valeurs passées.

Ces trois modèles remplissent globalement leurs objectifs, et le principe à la base de chaque méthode fonctionne. Cependant il s'agit d'approches nouvelles qui auraient besoin de développements supplémentaires avant d'être considérées pour des application réelles. Les deux GSs présentés répondent à la problématique de développer des modèles permettant de simuler des séries temporelles de pluie, tout en modélisant le patron asymptotique des valeurs hautes. L'objectif de la thèse est donc globalement rempli, malgré les limites de ces modèles.

5.2 Limites des modèles

Le modèle d'imputation des séries temporelles de surcote marine présenté dans le premier article (chapitre 2) a rempli son objectif en améliorant de modèle existant pour cette tâche, cependant ce modèle présente plusieurs limites. La première limite fondamentale porte sur les données. Le choix a été fait d'imputer la surcote car c'est l'approche utilisée par Hamdi et al. (2019), cependant il ne s'agit pas d'une variable transformée mais calculée à partir du niveau marin et de la marée. Indépendamment du modèle d'imputation, cela aurait davantage de sens d'imputer le niveau marin pour ensuite calculer la surcote. La marée n'a pas de valeurs manquantes car elle est modélisée par des harmoniques. Le principe du modèle présenté dans l'article est de modéliser la distribution multivariée entre différentes séries temporelles de surcote correspondant à différentes stations, mais cette distribution est uniquement spatiale. Comme la fréquence des données est relativement élevée (environ 12 heures), un modèle d'imputation pourrait utiliser l'information des observations avant et après une valeur manquante, à la fois dans pour la série temporelles à imputer et dans celles des stations voisines. Enfin, la distribution marginal paramétrique utilisée n'était pas une distribution des valeurs extrêmes. Les deux générateurs stochastiques développés par la suite ne modélisent plus la dimension spatiale, cependant les développements sur la prise en compte de la dépendance temporelle et l'utilisation d'une distribution marginale des valeurs extrêmes auraient bénéficié à ce premier modèle d'imputation.

Le premier GS paramétrique présenté dans le chapitre 3 a donné un résultat satisfaisant, mis à part une reproduction uniquement partielle du patron IDF dans les simulations. La cause de cette différence entre les patrons IDF observés et simulés n'est pas identifiée. La principale limite de ce modèle est le fait qu'il soit purement paramétrique, ce qui le rend compliqué à ajuster et n'est pas facilement modifiable. Le modèle est codé en Stan pour permettre d'obtenir les distributions a priori des paramètres, mais les difficultés d'ajustement font que les paramètres ont finalement été ajustés par maximum de vraisemblance. De plus, le temps de simulations était relativement long,

notamment à cause de l'échantillonnage de la copule en vigne temporelle. Ces limites expliquent pourquoi il a été décidé d'utiliser l'apprentissage profond pour le second GS.

Le deuxième GS par apprentissage profond et troisième article présenté dans le chapitre 4 a rempli son objectif, mais a trois limites identifiées. La première est que le paramètre de forme ξ de la distribution GP n'est pas estimé lors de l'entrainement du modèle, et est donc estimé dans une étape préalable. Ce problème semble lié à l'entrainement par propagation-inverse du gradient et à la fonction objectif par vraisemblance, le paramètre ξ n'ayant pas suffisamment d'impact sur la vraisemblance comparé au reste du modèle. Lors de tests d'estimation de ce paramètre par back-propagation, il tend à augmenter constamment de manière quasi-linéaire, indépendamment du reste du modèle, et fini par se stabiliser à une niveau anormalement haut pour des données de pluie. Une second limite est le fait que le modèle à deux processus stochastiques séparés, un processus binaire pour la génération des valeurs sans pluie et un second via la distribution EGP pour les valeurs non-nulles. Avoir ces deux processus stochastiques séparés est commun pour les GS, mais avoir à la place un seul processus stochastique capable de générer à la fois les zéros et les valeurs non-nulles serait préférable, en particulier pour étendre le modèle au cadre spatial. Si le modèle était étendu à plusieurs stations, l'EGP neurale pourrait prendre en compte la dépendance entre les stations sans développement additionnel pour les valeurs non-nulles. Par contre il faudrait modifier la modélisation de l'intermittence pour prendre en compte la dépendance spatiale du processus binaire. Enfin, une troisième limite est l'ajustement de la distribution par couches neurales monotones convexes, qui n'est pas aussi aisé que pour une architecture de couches classique. Il reste encore à identifier la meilleure façon de régulariser l'entrainement du modèle pour qu'il converge de façon plus sûre à la distribution ciblée.

5.3 Applicabilité des modèles

Les trois modèles présentés dans la thèse ne sont pas applicables en l'état, car ils nécessiteraient à la fois des améliorations sur le plan des méthodes, et une meilleur implémentation en terme de code pour les modèles paramétriques (i.e. le modèle d'imputation et le GS paramétrique). Néanmoins, ces trois modèles étaient développés en essayant d'aller au delà de la preuve de concept, et d'atteindre un niveau de développement suffisant pour une application réelle.

Les deux modèles paramétriques nécessiteraient des développements supplémentaires pour être applicable, avec la modélisation de la dépendance temporelle dans le cas du modèle d'imputation, et une modification pour reproduire totalement le patron IDF dans le cas du GS paramétrique.

Le GS par apprentissage profond pourrait être considéré pour application réelle si le problème d'ajustement de la PIT était identifié. On peut alors s'attendre à ce que le patron IDF soit totalement reproduit dans les simulation. Le GS pourrait être alors considéré comme une alternative au modèle IDF. Le GS aurait l'avantage de reproduire des patrons statistiques que le modèle IDF ne considère pas. Le GS par apprentissage profond est donc le meilleur candidat parmi les trois modèle pour une application. C'est aussi le modèle dont l'implémentation en code est la plus réutilisable et modifiable (par exemple pour développer une version non-stationnaire du modèle), tout en ayant les meilleurs performances en temps d'ajustement et de simulation.

5.4 Perspectives de développements futurs

Les deux GS, paramétriques et par apprentissage profond, ont en commun d'avoir deux processus stochastiques séparés pour la modélisation binaire des valeurs sans pluie et l'EGP modélisant les valeurs non-nulles. Cela complique la modélisation et limite la potentielle généralisation au cadre spatial, car il faudrait alors modéliser la dépendance spatiale dans les deux processus. Une façon de contourner cette limitation serait de profiter du fait que la pluie est positive, en remplaçant les valeurs nulles par du bruit. En générant des valeurs à partir d'une distribution au support $(-\infty, 0]$, par exemple une distribution tronquée, et en remplaçant les valeurs nulles des observations par ces valeurs aléatoire, la série temporelle de pluie modifiée aurait comme support $\mathbb{R}$. Le modèle serait entrainé sur ces données modifiées. Lors de la simulation, une valeur simulée inférieur à 0 serait remplacée par une valeur générée à partir de la distribution tronquée, afin de ne pas introduire de biais. Au lieu de remplacer les zéros par du bruit, il serait aussi possible de les remplacer par des valeurs générées à partir d'une fonction. En s'inspirant des fonctions d'intensité des processus de Hawkes, les valeurs générées pourraient déprendre des observations passées. Cette option serait de l'ingénierie des caractéristiques (feature engineering), qui aurait pour but d'aider l'entrainement du modèle.

Les aspects de non-stationnarité temporelle via les forçages climatiques ont été mis de coté dans le cadre de cette thèse, car l'objectif était le développement de nouveaux modèles. En effet, il serait compliqué de proposer un nouveau modèle qui intégrerait déjà des covariables climatiques, car une mauvaise performance du modèle pourrait être masquée par l'information des covariables qui contribuerait majoritairement au résultat. Il semble préférable de proposer dans un premier temps le meilleur modèle possible utilisant uniquement la variable d'intérêt, ici la pluie, puis dans un second temps d'ajouter des covariables, une fois que la base du modèle est satisfaisante. Il y aurait deux

façons d'utiliser des covariables climatiques pour un générateur stochastique temporel. La manière la plus simple serait d'avoir des projections de ces covariables indépendamment du modèle, et il suffit alors de simuler la variable d'intérêt conditionnellement à ces projections. Ces covariables pourraient par exemple être les données du Coupled Model Intercomparison Project (O'Neill et al. 2016), qui commencent en 1900 et vont jusqu'en 2100. La seconde option serait d'utiliser des covariables couvrant uniquement la période des observations de la variable d'intérêt, et de générer aussi des valeurs pour les covariables lors de la simulation. Dans un modèle d'apprentissage profond, une possibilité serait de transformer les covariables en un espace latent et de générer des valeurs à partir de cette espace latent plutôt que de directement simuler les covariables. Un modèle de diffusion conditionnelle pourrait être utilisé pour créer cet espace latent, à partir duquel des valeurs seraient générées. Cela simplifierait la modélisation des covariables et leur simulation.

Un direction d'amélioration pour l'EGP neural serait d'utiliser un modèle Bayésien d'apprentissage profond. En plus d'être entrainé par la méthode classique de back-propagation du gradient, un réseau de neurone peut être entrainé par différentes méthodes Bayésienne (Jospin et al. 2022). Le modèle présenté dans l'article 3 n'est stochastique que dans la simulation, qui passe par la génération de valeurs aléatoires à partir d'une distribution. Le modèle en lui-même ne tient pas compte de l'incertitude de ses paramètres et de sa forme, alors que cela influence le résultat. L'incertitude sur les paramètres pourrait être prise en compte en obtenant leurs distributions a posteriori, tandis que l'incertitude sur la forme du modèle pourrait être prise en compte avec un algorithme modifiant le modèle pendant l'entrainement, comme c'est le cas pour l'algorithme reversible jump Markov chain Monte Carlo du premier article d'imputation. La modélisation Bayésienne et l'apprentissage profond on été utilisés séparément dans cette thèse, mais une bonne option serait d'utiliser ces deux frameworks ensemble. Une application intéressante d'un algorithme modifiant la forme du modèle serait pour l'inclusion de covariables, afin d'améliorer l'interprétabilité du modèle. Un modèle avec sauts réversibles ou algorithme birth-death peut sélectionner les meilleures covariables, ainsi que leur lien fonctionnel avec la variable d'intérêt (El Adlouni et Ouarda 2009). Cela revient à définir un métamodèle dont la distribution a posteriori est échantillonnée, donnant ainsi une probabilité à chaque version du modèle. Une autre option plus facile à implémenter que des sauts réversibles serait l'utilisation de distributions a priori de rétrécissement (shrinkage priors), dont la densité est majoritairement à 0 et 1, permettant ainsi d'ajouter ou de retirer des paramètres au modèle (Piironen et Vehtari 2017a ; Piironen et Vehtari 2017b).

La distribution par réseau de neurones monotone convexe de l'EGP neurale pourrait être étendue au cadre stationnaire, en modélisant alors une distribution spatio-temporelle. La différence avec la

distribution de l'article serait que les dimensions spatiales à temps t ne pourraient pas être considérés indépendantes pendant l'entrainement et la simulation, comme les sont les valeurs passées à $t-1,\ldots,t-J$. Pour K dimensions spatiales, il faudrait alors calculer la dérivée d'ordre K pour obtenir la densité. Cela deviendrait rapidement rédhibitoire en temps de calcul. Une solution serait alors d'utiliser une vraisemblance composite pour les dimension spatiale, considérant la vraisemblance des dimensions par paires, et la vraisemblance totale du modèle étant la somme des ces vraisemblances par paires (Chilinski et Silva 2018). Si le nombre de stations ou dimensions spatiales est élevée, il serait aussi possible de considérer la dépendance dans la distribution uniquement pour les stations proches. Le reste de la dépendance serait alors modélisée par les autres éléments de l'architecture du réseau de neurones, par exemple des couches de convolution spatiale dans le cas de données sur grille.

Bibliographie

Aas, K., C. Czado, A. Frigessi et H. Bakken (2009). "Pair-copula constructions of multiple dependence". In : *Insurance : Mathematics and Economics* 44.2, p. 182-198. doi : 10.1016/j.insmatheco.2007.02.001.

Ahn, K.-H. (2020). "Coupled annual and daily multivariate and multisite stochastic weather generator to preserve low- and high-frequency variability to assess climate vulnerability". In : *Journal of Hydrology* 581, p. 124443. doi : 10.1016/j.jhydrol.2019.124443.

— (2021). "Streamflow estimation at partially gaged sites using multiple-dependence conditions via vine copulas". In : *Hydrology and Earth System Sciences* 25.8, p. 4319-4333. doi : 10.5194/hess-25-4319-2021.

Andreevsky, M., Y. Hamdi, S. Griolet, P. Bernardara et R. Frau (2020). "Regional frequency analysis of extreme storm surges using the extremogram approach". In : *Natural Hazards and Earth System Sciences* 20.6, p. 1705-1717. doi : 10.5194/nhess-20-1705-2020.

Bedford, T. et R. M. Cooke (2002). "Vines–a new graphical model for dependent random variables". In : *The Annals of Statistics* 30.4. doi : 10.1214/aos/1031689016.

Ben Aissia, M.-A., F. Chebana et T. B. M. J. Ouarda (2017). "Multivariate missing data in hydrology – review and applications". In : *Advances in Water Resources* 110, p. 299-309. doi : 10.1016/j.advwatres.2017.10.002.

Beneyto, C., J. A. Aranda et F. Francés (2023). "Exploring the uncertainty of weather generators' extreme estimates in different practical available information scenarios". In : *Hydrological Sciences Journal* 68.9, p. 1203-1212. doi : 10.1080/02626667.2023.2208754.

Blanchet, J. et J.-D. Creutin (2017). "Co-occurrence of extreme daily rainfall in the French Mediterranean region". In : *Water Resources Research* 53.11, p. 9330-9349. doi : 10.1002/2017WR020717.

Brechmann, E. C., C. Czado et K. Aas (2012). "Truncated regular vines in high dimensions with application to financial data". In : *Canadian Journal of Statistics* 40.1, p. 68-85. doi : 10.1002/cjs.10141.

Brunner, M. I. (2023). "Floods and droughts : A multivariate perspective". In : *Hydrology and Earth System Sciences* 27.13, p. 2479-2497. doi : 10.5194/hess-27-2479-2023.

Brunner, M. I., J. Seibert et A. Favre (2016). "Bivariate return periods and their importance for flood peak and volume estimation". In : *WIREs Water* 3.6, p. 819-833. doi : 10.1002/wat2.1173.

Bárdossy, A. et G. G. S. Pegram (2009). "Copula based multisite model for daily precipitation simulation". In : *Hydrology and Earth System Sciences* 13.12, p. 2299-2314. doi : 10.5194/hess-13-2299-2009.

Cardona, O.-D., M. K. Van Aalst, J. Birkmann, M. Fordham, G. McGregor, R. Perez, R. S. Pulwarty, E. L. F. Schipper, B. T. Sinh, H. Décamps, M. Keim, I. Davis, K. L. Ebi, A. Lavell, R. Mechler, V. Murray, M. Pelling, J. Pohl, A.-O. Smith et F. Thomalla (2012). "Determinants of Risk : Exposure and Vulnerability". In : *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Sous la dir. de C. B. Field, V. Barros, T. F. Stocker et Q. Dahe. 1re éd. Cambridge University Press, p. 65-108. doi : 10.1017/CBO9781139177245.005.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li et A. Riddell (2017). "Stan : A probabilistic programming language". In : *Journal of Statistical Software* 76.1. doi : 10.18637/jss.v076.i01.

Cavadias, G. S., T. B. M. J. Ouarda, B. Bobée et C. Girard (2001). "A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins". In : *Hydrological Sciences Journal* 46.4, p. 499-512. doi : 10.1080/02626660109492846.

Chapon, A., T. B. Ouarda et N. Bertrand (2025). "Stochastic generator for rainfall with a Hawkes process marked by an extended generalized Pareto and a vine copula". In : *Environmental Modelling & Software* 191, p. 106490. doi : 10.1016/j.envsoft.2025.106490.

Chapon, A., T. B. Ouarda et Y. Hamdi (2023). "Imputation of missing values in environmental time series by D-vine copulas". In : *Weather and Climate Extremes*, p. 100591. doi : 10.1016/j.wace.2023.100591.

Chebana, F. et T. B. M. J. Ouarda (2021). "Multivariate non-stationary hydrological frequency analysis". In : *Journal of Hydrology* 593. doi : 10.1016/j.jhydrol.2020.125907.

Chilinski, P. et R. Silva (2018). *Neural likelihoods via cumulative distribution functions*. Version Number : 2. doi : 10.48550/ARXIV.1811.00974.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London : Springer.

Coles, S. et L. Pericchi (2003). “Anticipating catastrophes through extreme value modelling”. In : *Journal of the Royal Statistical Society : Series C (Applied Statistics)* 52.4, p. 405-416. doi : 10.1111/1467-9876.00413.

Craiu, R. V. et J. S. Rosenthal (2014). “Bayesian computation via Markov chain Monte Carlo”. In : *Annual Review of Statistics and Its Application* 1.1, p. 179-201. doi : 10.1146/annurev-statistics-022513-115540.

Davis, C. (2015). *sgt : Skewed generalized t distribution tree*.

Davis, R. A., C. Klüppelberg et C. Steinkohl (2013). “Statistical inference for max-stable processes in space and time”. In : *Journal of the Royal Statistical Society Series B : Statistical Methodology* 75.5, p. 791-819. doi : 10.1111/rssb.12012.

Di Lascio, F. M. L., S. Giannerini et A. Reale (2015). “Exploring copulas for the imputation of complex dependent data”. In : *Statistical Methods and Applications* 24.1, p. 159-175. doi : 10.1007/s10260-014-0287-2.

Dissmann, J., E. C. Brechmann, C. Czado et D. Kurowicka (2013). “Selecting and estimating regular vine copulae and application to financial returns”. In : *Computational Statistics and Data Analysis* 59, p. 52-69. doi : 10.1016/j.csda.2012.08.010.

El Adlouni, S. et T. B. M. J. Ouarda (2009). “Joint Bayesian model selection and parameter estimation of the generalized extreme value model with covariates using birth-death Markov chain Monte Carlo”. In : *Water Resources Research* 45.6. doi : 10.1029/2007WR006427.

Embrechts, P., F. Lindskog et A. McNeil (2003). “Modelling Dependence with Copulas and Applications to Risk Management”. In : *Handbook of heavy tailed distibutions in finance*. Amsterdam : Elsevier, p. 329-384.

Espeholt, L., S. Agrawal, C. Sønderby, M. Kumar, J. Heek, C. Bromberg, C. Gazen, R. Carver, M. Andrychowicz, J. Hickey, A. Bell et N. Kalchbrenner (2022). “Deep learning for twelve hour precipitation forecasts”. In : *Nature Communications* 13.1, p. 5145. doi : 10.1038/s41467-022-32483-x.

Evin, G., A.-C. Favre et B. Hingray (2018). “Stochastic generation of multi-site daily precipitation focusing on extreme events”. In : *Hydrology and Earth System Sciences* 22.1, p. 655-672. doi : 10.5194/hess-22-655-2018.

Gamet, P. et J. Jalbert (2022). “A flexible extended generalized Pareto distribution for tail estimation”. In : *Environmetrics* 33.6. doi : 10.1002/env.2744.

Gao, Y., C. Merz, G. Lischeid et M. Schneider (2018). “A review on missing hydrological data processing”. In : *Environmental Earth Sciences* 77. doi : 10.1007/s12665-018-7228-6.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari et D. B. Rubin (2015). *Bayesian Data Analysis*. 3e éd. New York : Chapman et Hall/CRC.

Genest, C. et A.-C. Favre (2007). "Everything you always wanted to know about copula modeling but were afraid to ask". In : *Journal of Hydrologic Engineering* 12.4, p. 347-368. doi : 10.1061/(ASCE)1084-0699(2007)12:4(347).

Green, P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In : *Biometrika* 84.4, p. 711-732. doi : 10.1093/BIOMET/82.4.711.

Gruber, L. F. et C. Czado (2018). "Bayesian model selection of regular vine copulas". In : *Bayesian Analysis* 13.4. doi : 10.1214/17-BA1089.

Grösser, J. et O. Okhrin (2022). "Copulae : An overview and recent developments". In : *WIREs Computational Statistics* 14.3. doi : 10.1002/wics.1557.

Hamdi, Y., C.-M. Duluc, L. Bardet et V. Rebour (2019). "Development of a target-site-based regional frequency model using historical information". In : *Natural Hazards* 98.3, p. 895-913. doi : 10.1007/s11069-018-3237-8.

Hamzah, F. B., F. M. Hamzah, S. F. M. Razali, O. Jaafar et N. A. Jamil (2020). "Imputation methods for recovering streamflow observation : A methodological review". In : *Cogent Environmental Science* 6.1. Sous la dir. de F. Li, p. 1745133. doi : 10.1080/23311843.2020.1745133.

Han, X., T. B. M. J. Ouarda, A. Rahman, K. Haddad, R. Mehrotra et A. Sharma (2020). "A network approach for delineating homogeneous regions in regional flood frequency analysis". In : *Water Resources Research* 56.3. doi : 10.1029/2019WR025910.

Hasler, C., R. V. Craiu et L.-P. Rivest (2018). "Vine copulas for imputation of monotone non-response". In : *International Statistical Review* 86.3, p. 488-511. doi : 10.1111/insr.12263.

Hawkes, A. G. (2018). "Hawkes processes and their applications to finance : A review". In : *Quantitative Finance* 18.2, p. 193-198. doi : 10.1080/14697688.2017.1403131.

Hersbach, H, B Bell, P Berrisford, G Biavati, A Horányi, J Muñoz Sabater, J Nicolas, C Peubey, R Radu, I Rozum, D Schepers, A Simmons, C Soci, D Dee et J.-N. Thépaut (2018). "ERA5 hourly data on single levels from 1959 to present". In : *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. doi : 10.24381/cds.adbb2d47.

Hollenbach, F. M., I. Bojinov, S. Minhas, N. W. Metternich, S. Minhas, M. D. Ward et A. Volfovsky (2014). "Multiple imputation using Gaussian copulas". In : Publisher : arXiv Version Number : 3. doi : 10.48550/ARXIV.1411.0647.

Hurrell, J. W., Y. Kushnir, G. Ottersen et M. Visbeck (2003). "An overview of the North Atlantic Oscillation". In : *The North Atlantic Oscillation : Climatic Significance and Environmental Impact*. American Geophysical Union (AGU), p. 1-35.

Huser, R., T. Opitz et J. Wadsworth (2024). "Modeling of spatial extremes in environmental data science : Time to move away from max-stable processes". In : doi : 10.48550/ARXIV.2401.17430.

Hyndman, R. J. (1996). "Computing and graphing highest density regions". In : *The American Statistician* 50.2, p. 120-126. doi : 10.2307/2684423.

Hyndman, R. J., J. Einbeck et M. P. Wand (2021). *hdrcde : Highest density regions and conditional density estimation*.

James, G., D. Witten, T. Hastie et R. Tibshirani (2017). *An Introduction to Statistical Learning*. New York : Springer.

Jane, R., L. Dalla Valle, D. Simmonds et A. Raby (2016). "A copula-based approach for the estimation of wave height records through spatial correlation". In : *Coastal Engineering* 117, p. 1-18. doi : 10.1016/j.coastaleng.2016.06.008.

Ji, H. K., M. Mirzaei, S. H. Lai, A. Dehghani et A. Dehghani (2024). "Implementing generative adversarial network (GAN) as a data-driven multi-site stochastic weather generator for flood frequency estimation". In : *Environmental Modelling & Software* 172, p. 105896. doi : 10.1016/j.envsoft.2023.105896.

Joe, H. (2015). *Dependence modeling with copulas*. Monographs on statistics and applied probability 134. Boca Raton : CRC Press, Taylor & Francis Group.

Johannesson, A. V., S. Siegert, R. Huser, H. Bakka et B. Hrafnkelsson (2022). "Approximate Bayesian inference for analysis of spatiotemporal flood frequency data". In : *The Annals of Applied Statistics* 16.2. doi : 10.1214/21-AOAS1525.

Jospin, L. V., H. Laga, F. Boussaid, W. Buntine et M. Bennamoun (2022). "Hands-on Bayesian neural networks—A tutorial for deep learning users". In : *IEEE Computational Intelligence Magazine* 17.2, p. 29-48. doi : 10.1109/MCI.2022.3155327.

Jurado, O. E., J. Ulrich, M. Scheibel et H. W. Rust (2020). "Evaluating the performance of a max-stable process for estimating intensity-duration-frequency curves". In : *Water* 12.12, p. 3314. doi : 10.3390/w12123314.

Kalteh, A. M. et P. Hjorth (2009). "Imputation of missing values in a precipitation–runoff process database". In : *Hydrology Research* 40.4, p. 420-432. doi : 10.2166/nh.2009.001.

Karras, T., M. Aittala, T. Aila et S. Laine (2022). *Elucidating the design space of diffusion-based generative models*. Version Number : 2. doi : 10.48550/ARXIV.2206.00364.

Kerman, S. C. et J. B. McDonald (2013). "Skewness–kurtosis bounds for the skewed generalized t and related distributions". In : *Statistics and Probability Letters* 83.9, p. 2129-2134. doi : 10.1016/j.spl.2013.05.028.

Knoben, W. J. M., J. E. Freer et R. A. Woods (2019). "Technical note : Inherent benchmark or not ? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores". In : *Hydrology and Earth System Sciences* 23.10, p. 4323-4331. doi : 10.5194/hess-23-4323-2019.

Koutsoyiannis, D., D. Kozonis et A. Manetas (1998). "A mathematical framework for studying rainfall intensity-duration-frequency relationships". In : *Journal of Hydrology* 206.1-2, p. 118-135. doi : 10.1016/S0022-1694(98)00097-3.

Krouma, M., D. Specq, L. Magnusson, C. Ardilouze, L. Batté et P. Yiou (2024). "Improving subseasonal forecast of precipitation in Europe by combining a stochastic weather generator with dynamical models". In : *Quarterly Journal of the Royal Meteorological Society* 150.762, p. 2744-2764. doi : 10.1002/qj.4733.

Lang, M., T. Ouarda et B. Bobée (1999). "Towards operational guidelines for over-threshold modeling". In : *Journal of Hydrology* 225.3-4, p. 103-117. doi : 10.1016/S0022-1694(99)00167-5.

Langousis, A. et D. Veneziano (2007). "Intensity-duration-frequency curves from scaling representations of rainfall". In : *Water Resources Research* 43.2, 2006WR005245. doi : 10.1029/2006WR005245.

Laub, P. J., T. Taimre et P. K. Pollett (2015). "Hawkes processes". In : Publisher : arXiv Version Number : 1. doi : 10.48550/ARXIV.1507.02822.

Lee, T., T. B. M. J. Ouarda et O. Seidou (2023). "Characterizing and forecasting climate indices using time series models". In : *Theoretical and Applied Climatology* 152.1-2, p. 455-471. doi : 10.1007/s00704-023-04434-z.

Lee, T., J.-Y. Shin, J.-S. Kim et V. P. Singh (2020). "Stochastic simulation on reproducing long-term memory of hydroclimatological variables using deep learning model". In : *Journal of Hydrology* 582, p. 124540. doi : 10.1016/j.jhydrol.2019.124540.

Li, X., C. Genest et J. Jalbert (2021). "A self-exciting marked point process model for drought analysis". In : *Environmetrics* 32.8, e2697. doi : 10.1002/env.2697.

Little, T. D., T. D. Jorgensen, K. M. Lang et E. W. G. Moore (2014). "On the joys of missing data". In : *Journal of Pediatric Psychology* 39.2, p. 151-162. doi : 10.1093/jpepsy/jst048.

Massei, N., B. Dieppois, D. M. Hannah, D. A. Lavers, M. Fossa, B. Laignel et M. Debret (2017). "Multi-time-scale hydroclimate dynamics of a regional watershed and links to large-scale atmospheric circulation : Application to the Seine river catchment, France". In : *Journal of Hydrology* 546, p. 262-275. doi : 10.1016/j.jhydrol.2017.01.008.

McDonald, J. B. (1984). "Some generalized functions for the size distribution of income". In : *Econometrica* 52.3, p. 647. doi : 10.2307/1913469.

Min, A. et C. Czado (2010). "Bayesian inference for multivariate copulas using pair-copula constructions". In : *Journal of Financial Econometrics* 8.4, p. 511-546. doi : 10.1093/jfinec/nbp031.

— (2011). "Bayesian model selection for D-vine pair-copula constructions". In : *Canadian Journal of Statistics* 39.2, p. 239-258. doi : 10.1002/cjs.10098.

Misra, D. (2020). *Mish : A self regularized non-monotonic activation function*. arXiv :1908.08681 [cs]. doi : 10.48550/arXiv.1908.08681.

Morales Napoles, O. (2009). "Bayesian belief nets and vines in aviation safety and other applications". Thèse de doct. Technische Universiteit Delft.

Nadarajah, S. et S. Kotz (2006). "R programs for computing truncated distributions". In : *Journal of Statistical Software* 16. doi : 10.18637/jss.v016.c02.

Nagler, T., U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler et T. Erhardt (2022). *VineCopula : Statistical inference of vine copulas*.

Nagler, T. et T. Vatter (2023). *rvinecopulib : High performance algorithms for vine copula modeling*.

Naveau, P., R. Huser, P. Ribereau et A. Hannart (2016). "Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection". In : *Water Resources Research* 52.4, p. 2753-2769. doi : 10.1002/2015WR018552.

O'Neill, B. C., C. Tebaldi, D. P. Van Vuuren, V. Eyring, P. Friedlingstein, G. Hurtt, R. Knutti, E. Kriegler, J.-F. Lamarque, J. Lowe, G. A. Meehl, R. Moss, K. Riahi et B. M. Sanderson (2016). "The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6". In : *Geoscientific Model Development* 9.9, p. 3461-3482. doi : 10.5194/gmd-9-3461-2016.

Ouarda, T. B. M. J., C. Charron, S. Mahdi et L. A. Yousef (2021). "Climate teleconnections, interannual variability, and evolution of the rainfall regime in a tropical Caribbean island : case study of Barbados". In : *Theoretical and Applied Climatology* 145.1-2, p. 619-638. doi : 10.1007/s00704-021-03653-6.

Ouarda, T. B. M. J., C. Charron et A. St-Hilaire (2019). "Uncertainty of stationary and nonstationary models for rainfall frequency analysis". In : *International Journal of Climatology* 40.4, p. 2373-2392. doi : 10.1002/joc.6339.

Ouarda, T. B. M. J., L. A. Yousef et C. Charron (2018). "Non-stationary intensity-duration-frequency curves integrating information concerning teleconnections and climate change". In : *International Journal of Climatology* 39.4, p. 2306-2323. doi : 10.1002/joc.5953.

Pan, X., G. Yildirim, A. Rahman, K. Haddad et T. B. M. J. Ouarda (2023). "Peaks-over-threshold-based regional flood frequency analysis using regularised linear models". In : *Water* 15.21, p. 3808. doi : 10.3390/w15213808.

Papalexiou, S. M. (2022). "Rainfall generation revisited : Introducing CoSMoS-2s and advancing copula-based intermittent time series modeling". In : *Water Resources Research* 58.6. DOI : 10.1029/2021WR031641.

Papalexiou, S. M., F. Serinaldi et M. P. Clark (2023). "Large-domain multisite precipitation generation : Operational blueprint and demonstration for 1,000 sites". In : *Water Resources Research* 59.3, e2022WR034094. DOI : 10.1029/2022WR034094.

Papastathopoulos, I. et J. A. Tawn (2013). "Extended generalised Pareto models for tail estimation". In : *Journal of Statistical Planning and Inference* 143.1, p. 131-143. DOI : 10.1016/j.jspi.2012.07.001.

Parey, S., T. T. H. Hoang et D. Dacunha-Castelle (2010). "Different ways to compute temperature return levels in the climate change context". In : *Environmetrics* 21.7-8, p. 698-718. DOI : 10.1002/env.1060.

Piironen, J. et A. Vehtari (2017a). "On the hyperprior choice for the global shrinkage parameter in the horseshoe prior". In : Publisher : arXiv Version Number : 2. DOI : 10.48550/ARXIV.1610.05559.

— (2017b). "Sparsity information and regularization in the horseshoe and other shrinkage priors". In : *Electronic Journal of Statistics* 11.2. DOI : 10.1214/17-EJS1337SI.

Price, I., A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam et M. Willson (2023). *GenCast : Diffusion-based ensemble forecasting for medium-range weather*. Version Number : 2. DOI : 10.48550/ARXIV.2312.15796.

Rasul, K., C. Seward, I. Schuster et R. Vollgraf (2021). "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting". In : Publisher : arXiv Version Number : 2. DOI : 10.48550/ARXIV.2101.12072.

Reiss, R. et M. Thomas (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. 3e éd. Birkhäuser.

Roberts, G. O. et J. S. Rosenthal (2009). "Examples of adaptive MCMC". In : *Journal of Computational and Graphical Statistics* 18.2, p. 349-367. DOI : 10.1198/jcgs.2009.06134.

Saint Criq, L., E. Gaume, Y. Hamdi et T. B. M. J. Ouarda (2022). "Extreme sea level estimation combining systematic observed skew surges and historical record sea levels". In : *Water Resources Research* 58.3. DOI : 10.1029/2021WR030873.

Sarhadi, A. et E. D. Soulis (2017). "Time-varying extreme rainfall intensity-duration-frequency curves in a changing climate". In : *Geophysical Research Letters* 44.5. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/2016GL072201, p. 2454-2463. DOI : 10.1002/2016GL072201.

Sartor, D., A. Sinigaglia et G. A. Susto (2025). *Advancing constrained monotonic neural networks : Achieving universal approximation beyond bounded activations*. Version Number : 2. doi : 10.48550/ARXIV.2505.02537.

Scher, S. et S. Pessenteiner (2021). "Technical Note : Temporal disaggregation of spatial rainfall fields with generative adversarial networks". In : *Hydrology and Earth System Sciences* 25.6, p. 3207-3225. doi : 10.5194/hess-25-3207-2021.

Schmid, F. et R. Schmidt (2007). "Multivariate conditional versions of Spearman's rho and related measures of tail dependence". In : *Journal of Multivariate Analysis* 98.6, p. 1123-1140. doi : 10.1016/j.jmva.2006.05.005.

Serinaldi, F. (2015). "Dismissing return periods !" In : *Stochastic Environmental Research and Risk Assessment* 29.4, p. 1179-1189. doi : 10.1007/s00477-014-0916-1.

Serinaldi, F. et C. G. Kilsby (2014). "Rainfall extremes : Toward reconciliation after the battle of distributions". In : *Water Resources Research* 50.1, p. 336-352. doi : 10.1002/2013WR014211.

— (2015). "Stationarity is undead : Uncertainty dominates the distribution of extremes". In : *Advances in Water Resources* 77, p. 17-36. doi : 10.1016/j.advwatres.2014.12.013.

Sklar, A. (1959). "Fonctions de répartition à N dimensions et leurs marges". In : *Publications de l'Institut de Statistique de L'Université de Paris 8*, p. 229-231.

Smith, R (2003). "Statistics of extremes, with applications in environment, insurance, and finance". In : *Extreme Values in Finance, Telecommunications, and the Environment*. 1st, p. 78.

Song, Y., J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon et B. Poole (2020). *Score-based generative modeling through stochastic differential equations*. Version Number : 2. doi : 10.48550/ARXIV.2011.13456.

Stephenson, A. G., E. A. Lehmann et A. Phatak (2016). "A max-stable process model for rainfall extremes at different accumulation durations". In : *Weather and Climate Extremes* 13, p. 44-53. doi : 10.1016/j.wace.2016.07.002.

Sun, X., B. Renard, M. Thyer, S. Westra et M. Lang (2015). "A global analysis of the asymmetric effect of ENSO on extreme precipitation". In : *Journal of Hydrology* 530, p. 51-65. doi : 10.1016/j.jhydrol.2015.09.016.

Sønderby, C. K., L. Espeholt, J. Heek, M. Dehghani, A. Oliver, T. Salimans, S. Agrawal, J. Hickey et N. Kalchbrenner (2020). *MetNet : A neural weather model for precipitation forecasting*. Version Number : 2. doi : 10.48550/ARXIV.2003.12140.

Tencaliec, P., A. Favre, P. Naveau, C. Prieur et G. Nicolet (2020). "Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount". In : *Environmetrics* 31.2. doi : 10.1002/env.2582.

Tootoonchi, F., M. Sadegh, J. O. Haerter, O. Räty, T. Grabs et C. Teutschbein (2022). "Copulas for hydroclimatic analysis : A practice-oriented overview". In : *WIREs Water* 9.2. doi : 10.1002/wat2.1579.

Toulemonde, G., J. Carreau et V. Guinot (2020). "Space–Time Simulations of Extreme Rainfall : Why and How ?" In : *Mathematical Modeling of Random and Deterministic Phenomena*. Sous la dir. de S. M. Manou-Abi, S. Dabo-Niang et J. Salone. 1re éd. Wiley, p. 53-71. doi : 10.1002/9781119706922.ch3.

Ulrich, J., O. E. Jurado, M. Peter, M. Scheibel et H. W. Rust (2020). "Estimating IDF curves consistently over durations with spatial covariates". In : *Water* 12.11. doi : 10.3390/w12113119.

Valle, D. et D. Kaplan (2019). "Quantifying the impacts of dams on riverine hydrology under non-stationary conditions using incomplete data and Gaussian copula models". In : *Science of The Total Environment* 677, p. 599-611. doi : 10.1016/j.scitotenv.2019.04.377.

Vernieuwe, H., S. Vandenberghe, B. De Baets et N. E. C. Verhoest (2015). "A continuous rainfall model based on vine copulas". In : *Hydrology and Earth System Sciences* 19.6, p. 2685-2699. doi : 10.5194/hess-19-2685-2015.

Vorobevskii, I., J. Park, D. Kim, K. Barfus et R. Kronenberg (2024). "Simulating sub-hourly rainfall data for current and future periods using two statistical disaggregation models : case studies from Germany and South Korea". In : *Hydrology and Earth System Sciences* 28.2, p. 391-416. doi : 10.5194/hess-28-391-2024.

Ward, P. J., M. C. De Ruiter, J. Mård, K. Schröter, A. Van Loon, T. Veldkamp, N. Von Uexkull, N. Wanders, A. AghaKouchak, K. Arnbjerg-Nielsen, L. Capewell, M. Carmen Llasat, R. Day, B. Dewals, G. Di Baldassarre, L. S. Huning, H. Kreibich, M. Mazzoleni, E. Savelli, C. Teutschbein, H. Van Den Berg, A. Van Der Heijden, J. M. Vincken, M. J. Waterloo et M. Wens (2020). "The need to integrate flood and drought disaster risk reduction strategies". In : *Water Security* 11, p. 100070. doi : 10.1016/j.wasec.2020.100070.

Wilks, D. (1998). "Multisite generalization of a daily stochastic precipitation generation model". In : *Journal of Hydrology* 210.1-4, p. 178-191. doi : 10.1016/S0022-1694(98)00186-3.

Wójcik, R. et T. Buishand (2003). "Simulation of 6-hourly rainfall and temperature by two resampling schemes". In : *Journal of Hydrology* 273.1-4, p. 69-80. doi : 10.1016/S0022-1694(02)00355-4.

Yan, J. (2007). "Enjoy the joy of copulas : With a package copula". In : *Journal of Statistical Software* 21.4. doi : 10.18637/jss.v021.i04.

Yiou, P. (2014). "AnaWEGE : a weather generator based on analogues of atmospheric circulation". In : *Geoscientific Model Development* 7.2, p. 531-543. DOI : 10.5194/gmd-7-531-2014.

Zeng, Z. et T. Wang (2022). *Neural copula : A unified framework for estimating generic high-dimensional copula functions*. Version Number : 3. DOI : 10.48550/ARXIV.2205.15031.