A SUB-ASYMPTOTIC MODEL FOR BIVARIATE THRESHOLD EXCEEDANCES

Mirco Lescart, Anna Kiriliouk, Philippe Naveau

LIDAM Discussion Paper ISBA
2026 / 11

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

arXiv:2604.12405v1 [stat.ME] 14 Apr 2026

A sub-asymptotic model for bivariate threshold exceedances

Mirco Lescart* Anna Kiriliouk* Philippe Naveau†

April 15, 2026

Abstract

Extreme value theory offers a statistical framework for quantifying the risk of rare events, with the generalized Pareto (GP) distribution providing the canonical limit model for univariate threshold exceedances. In many applications, however, extremes are intrinsically multivariate, requiring models that capture both marginal tail behaviours and joint extremal dependencies. Under asymptotic dependence, the multivariate GP distribution represents a suitable modelling family, but when asymptotic independence arises, sub-asymptotic models are needed. In this work, we propose and study a flexible sub-asymptotic parametric class to model bivariate threshold exceedances. Our new model accommodates a broad range of tail dependence behaviours and contains the standardised multivariate GP distribution as a limiting case while retaining margins that converge to univariate GP tails. Our formulation allows extremal dependence to evolve naturally with the marginal parameters on the original data scale, facilitating direct computation and interpretation of failure probabilities. Model inference is done via a likelihood-free neural Bayes estimation approach, with tailored prior specifications. An extensive simulation study and an application to Belgian rainfall extremes illustrate the estimation framework and the flexibility of the model.

Key words: bivariate extremes, peaks-over-thresholds, multivariate generalized Pareto distribution, asymptotic independence, neural Bayes, rainfall

1 Introduction

Extreme value theory (EVT) provides a mathematical framework to statistically model the probability distributions of extreme (rare) events (Beirlant et al., 2004). A key model within EVT is the univariate *generalized Pareto (GP)* distribution, which characterizes exceedances over a high threshold through a known and explicit parametric form. Under mild conditions, any non-degenerate limit distribution of (suitably standardised) threshold exceedances must be GP, just as the Gaussian distribution arises as the universal limit of (suitably standardised) sample means. The GP distribution underpins the peaks-over-threshold approach (Davison and Smith, 1990), widely applied in hydrology, finance, and engineering to quantify risks associated with rare events such as floods, financial crashes, or structural failures (Embrechts et al., 2013; Bousquet and Bernardara, 2021).

In practice, many extreme phenomena are inherently multivariate: flooding may result from concurrent rainfall across several catchments, financial crises arise from assets crashing

*LIDAM/ISBA, UCLouvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium. E-mails: mirco.lescart@uclouvain.be, anna.kiriliouk@uclouvain.be

†Laboratoire des Sciences du Climat et de l'Environnement (IPSL, CNRS, CEA, UVSQ). E-mail: naveau@lsce.ipsl.fr

jointly, and infrastructure failures may stem from multiple stresses. Such examples highlight the need for models that capture both the marginal tail behaviour of individual components and their joint extremal dependence.

To extend the univariate GP model, the family of *multivariate generalized Pareto (MGP)* distributions was proposed as the limiting law of (suitably standardised) multivariate threshold exceedances (Rootzén and Tajvidi, 2006; Rootzén et al., 2018; Kiriliouk et al., 2019). In this setting, an exceedance is a data point for which *at least one* of the components lies above a high threshold; it is common to say that the MGP distribution is supported on an *L-shaped region* (see Figure 3). Like its univariate counterpart, the MGP distribution is *threshold-stable*, in the sense that exceedances above higher thresholds remain MGP, preserving the marginal tail behaviour and extremal dependence structure. For *asymptotically dependent (AD)* variables, for which extremes tend to occur jointly, threshold-stability is an additional restriction. Under *asymptotic independence (AI)*, threshold-stability cannot arise. In practice, such cases are frequent (see, e.g. Huser et al., 2025) and a more refined knowledge of dependence is needed. To illustrate this, Figure 3 displays three scatterplots of simulated bivariate AI samples. Visually, the dependencies among extremes seem to change from weak (left panel), to moderate (centre panel), and finally strong (right panel). However, by construction, all three samples exhibit asymptotic independence with the same level of *residual* tail dependence. To move beyond describing extremal dependence solely through asymptotic limit behaviour, a recent strategy is to also model intermediate to high quantiles. This approach allows for flexible tail decay and avoids the need to commit a priori to either AI or AD.

Recent contributions to the development of sub-asymptotic models include unified frameworks for AI and AD modelling of bivariate threshold exceedances (Wadsworth et al., 2017; Engelke et al., 2019), as well as spatial constructions that accommodate distance-varying dependence regimes (Huser et al., 2017; Krupskii et al., 2018; Huser and Wadsworth, 2019; Morris et al., 2017; Hazra et al., 2025; Shi et al., 2026). Conditional extreme-value models (Heffernan and Tawn, 2004; Wadsworth and Tawn, 2022) offer a complementary approach, conditioning on a single variable exceeding a high threshold to describe joint tail behaviour. Bacro et al. (2025) also proposed a framework for constructing multivariate distributions with GP margins, exploiting the representation of a univariate GP variable as a ratio of exponential and gamma random variables. While their approach provides a tractable link between marginal and joint tail behaviour and can reach both AI and AD in the limit, its ability to model intermediate to strong AD is limited (see Section 2.2.3). Also, sub-asymptotic models should ideally contain commonly used asymptotic models as a boundary or limiting case (see also Zhang et al., 2025), ensuring coherence with asymptotic theory.

In this work, we introduce a flexible sub-asymptotic bivariate model that extends the multivariate GP representation through sums of exponential and gamma components. Our approach specifically targets the L-shaped region, enabling a coherent description of both moderate and extreme co-occurrences, and converges to the (standardised) MGP distribution as a limiting case. Although our model shares some aspects with Bacro et al. (2025), it provides a much wider range of extremal dependence.

Unlike many existing approaches, our framework does not require transforming margins to a common scale prior to modelling dependence. While probability integral transforms may be convenient in iid settings, they typically require explicit modelling of covariate-dependent marginal distributions in non-stationary applications, which can complicate interpretation and lead to uncertainty propagation (Kakampakou et al., 2024). Working on the original scale allows different variables to retain distinct Pareto tails, as illustrated in our analysis of

extreme precipitation over Belgium (Section 6). In addition, it facilitates the interpretation and computation of failure probabilities (i.e., probabilities of the data falling in a specified extreme set) in terms of the observed variables (Kiriliouk and Zhou, 2022).

Concerning the estimation of our model parameters, we avoid using classical methods such as censored likelihood (Smith et al., 1997; Wadsworth and Tawn, 2014; Bacro et al., 2025), or weighted least squares (Einmahl et al., 2012, 2018). Instead, we leverage the parameter parsimony and the simplicity of simulating random draws from our model. This allows us to adapt neural Bayes estimation techniques (Sainsbury-Dale et al., 2024; Richards et al., 2024; André et al., 2025). This approach has become popular for models with intractable likelihood; see also Rödder et al. (2025) for a characterization of its consistency and efficiency.

The remainder of the paper is organized as follows. Section 2 reviews univariate and bivariate GP models, asymptotic tail dependence coefficients and the main ingredients of the model of Bacro et al. (2025). The sub-asymptotic bivariate GP distribution is introduced in Section 3, where we derive marginal distributions, tail dependence properties and sub-asymptotic features of the model. Section 4 describes the estimation procedure and Section 5 contains a simulation study. An application to rainfall data in Belgium is provided in Section 6, and Section 7 concludes. Finally, the main proofs are detailed in Section 8. The Supplementary Material contains proofs of the lemmas stated in Section 8 and additional numerical results for the Belgian rainfall application.

Throughout, bold symbols will refer to multivariate quantities, $\mathbf{1}$ denotes a vector of ones (of a dimension clear from the context), and mathematical operations on vectors such as addition, multiplication and comparison are considered component-wise.

2 Background

2.1 Univariate peaks-over-thresholds modelling

The univariate *peaks-over-thresholds* approach (Davison and Smith, 1990) provides the theoretical and practical foundation for modelling threshold exceedances via a *generalized Pareto (GP)* distribution. First, recall that the survival function of the GP distribution is given by

$$\overline{H}_{\xi,\sigma}(z) = \begin{cases} \left(1 + \dfrac{\xi z}{\sigma}\right)^{-1/\xi}, & \text{if } \xi \neq 0, \\ e^{-z/\sigma}, & \text{if } \xi = 0, \end{cases} \qquad z \geq 0,\ 1 + \frac{\xi z}{\sigma} > 0,$$

where $\xi \in \mathbb{R}$ is the shape parameter or *tail index* and $\sigma > 0$ is the scale parameter. If a random variable Z has cdf $H_{\xi,\sigma}$, we also write $Z \sim \mathrm{GP}(\xi, \sigma)$. For $\xi > 0$, the survival function $\overline{H}_{\xi,\sigma}$ is regularly varying tail of order $1/\xi$ (see Section 8.1 for a definition).

Let X be a random variable with cdf F with upper endpoint $x^* = \infty$. Suppose that there exists an auxiliary function c such that, for any $x > 0$,

$$\lim_{u \to \infty} \mathbb{P}\left[\frac{X - u}{c(u)} > x \mid X > u\right] = (1 + \xi x)^{-1/\xi}. \tag{2.1}$$

Assumption (2.1) is equivalent to the fact that F is in the max-domain of attraction of a generalized extreme-value distribution (Pickands III, 1975; Balkema and de Haan, 1974). Interpreting $c(u)$ as a scale parameter $\sigma > 0$, we get, for $x > u$ and u sufficiently large, $\mathbb{P}[X > x] \approx \mathbb{P}[X > u]\overline{H}_{\xi,\sigma}(x - u)$, which can be used to model the tail of X.

The GP distribution is *threshold-stable*: for any positive scalar v, the conditional random variable $Z - v \mid Z > v$ is again GP distributed with the same shape parameter ξ and with scale $\sigma + \xi v$. Finally, it admits a convenient representation as the ratio of a standard exponential random variable over an independent gamma-distributed random variable.

Lemma 2.1 (Exponential-gamma representation of the GP distribution)**.** *Any GP distributed random variable with a positive shape parameter ξ can be written as the ratio of two independent random variables, a standard exponential random variable E and a gamma random variable G, with shape parameter α and rate parameter β, i.e., with density $f_G(g) = \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g}$ for $g > 0$,*

$$\frac{E}{G} \sim \mathrm{GP}(\xi = 1/\alpha, \sigma = \beta/\alpha).$$

For a proof of Lemma 2.1, see, for example, Reiss et al. (1997, page 157). The shape parameter of the gamma distribution being positive, we cannot obtain GP distributions with negative tail indices from the representation in Lemma 2.1. In the remainder of this work, we focus on $\xi \geq 0$. Note that $\xi = 0$ is obtained by taking $\beta = \alpha c$ for some $c > 0$ and $\alpha \to \infty$, in which case G converges to $1/c$ in probability (see the proof of Lemma 3.3), and E/G is exponential with scale c.

2.2 Bivariate peaks-over-thresholds modelling

We recall some convenient (asymptotic) summary measures of extremal dependence before introducing the model in Bacro et al. (2025) and our bivariate sub-asymptotic GP distribution.

2.2.1 Summary measures of extremes

Let $\boldsymbol{X} = (X_1, X_2)^T$ be a bivariate random vector with joint cdf F and continuous marginal distributions F_1, F_2. A measure of the upper tail dependence between X_1 and X_2 is

$$\chi(q) = \frac{\mathbb{P}\left(F_1(X_1) > q,\ F_2(X_2) > q\right)}{1 - q}, \qquad q \in [0, 1). \tag{2.2}$$

The *(upper) tail dependence coefficient* $\chi = \lim_{q \uparrow 1} \chi(q)$ then quantifies to what extent the components of $\boldsymbol{X}$ can be extreme simultaneously; we say that X_1 and X_2 exhibit *asymptotic dependence (AD)* when $\chi > 0$, and *asymptotic independence (AI)* when $\chi = 0$, see Coles et al. (1999). The function $\chi(q)$ can be seen as a sub-asymptotic measure of tail dependence.

When χ equals zero, the decay of joint extremes can be described by the *residual tail dependence coefficient* $\eta = \lim_{q \uparrow 1} \eta(q)$, where

$$\eta(q) = \frac{\log(1 - q)}{\log \mathbb{P}\left(F_1(X_1) > q,\ F_2(X_2) > q\right)}, \qquad q \in [0, 1).$$

The residual tail dependence coefficient is derived from the Ledford and Tawn (1996) model assumption that

$$\mathbb{P}\left(F_1(X_1) > q,\ F_2(X_2) > q\right) = \ell(1 - q)(1 - q)^{1/\eta},$$

where ℓ denotes a slowly varying function at zero, i.e., $\lim_{t \downarrow 0} \ell(ct)/\ell(t) = 1$ for all $c > 0$. A value $\eta < 1/2$ indicates negative residual dependence, meaning that the joint tail decays faster than under independence; $\eta = 1/2$ means there is no residual tail dependence. Values in the

interval $(1/2, 1)$ indicate positive residual dependence, with joint tails decaying more slowly than in the independent case. Finally, the boundary case $\eta = 1$ is the only one compatible with asymptotic dependence, where the probability of simultaneous extremes remains non-zero in the limit. The coefficient $\eta(q)$ quantifies the amount of sub-asymptotic residual tail dependence.

Evaluated for a full range of $q \in [0, 1)$, the coefficients $\chi(q)$ and $\eta(q)$ are particularly informative when convergence to the asymptotic regime is slow. They are related via

$$\eta(q) = \frac{\log(1-q)}{\log(1-q) + \log \chi(q)}. \tag{2.3}$$

Sometimes, we will add a subscript, writing $\chi_{\boldsymbol{X}}$ or $\eta_{\boldsymbol{X}}$, when confusion may arise.

2.2.2 Bivariate generalized Pareto distributions

The bivariate generalized Pareto distribution is the asymptotically justified model for the (conditional) threshold exceedances of a bivariate random vector as the threshold grows to infinity. We introduce it for a vector $\boldsymbol{X}_E$ whose margins follow unit exponential distributions; a full treatment can be found in Rootzén et al. (2018) or Rootzén et al. (2018).

Define $\mathcal{L} = \{\boldsymbol{x} \in \mathbb{R}^2 : \max(\boldsymbol{x}) > 0\}$. The vector $\boldsymbol{X}_E$ is in the domain of attraction of a bivariate generalized Pareto vector $\boldsymbol{Z} = (Z_1, Z_2)^T$ (Rootzén and Tajvidi, 2006) if, for any $\boldsymbol{z} \in \mathcal{L}$,

$$\mathbb{P}[\boldsymbol{Z} \leq \boldsymbol{z}] = \lim_{u \to \infty} \mathbb{P}[\boldsymbol{X}_E - u\mathbf{1} \leq \boldsymbol{z} \mid \max(\boldsymbol{X}_E) > u].$$

The bivariate GP distribution is *threshold-stable*: for any scalar $v \geq 0$, the conditional vector $\boldsymbol{Z} - v\mathbf{1} \mid \max(\boldsymbol{Z}) > v$ is again bivariate GP. In addition, any bivariate GP vector $\boldsymbol{Z}$ can be written as the sum of a unit exponential random variable E and a bivariate (*"spectral"*) random vector $\boldsymbol{S}$ whose component-wise minimum equals zero with probability one,

$$\boldsymbol{Z} = E - \boldsymbol{S}, \qquad E \sim \text{Exp}(1),\ \min(\boldsymbol{S}) = 0 \text{ a.s.},\ E \perp\!\!\!\perp \boldsymbol{S},$$

see Equation (3.2) in (Kiriliouk et al., 2019). Since the margins of $\boldsymbol{X}_E$ are unit exponential, we find that $Z_j \mid Z_j > 0$ is also unit exponential for $j = 1, 2$; we say that $\boldsymbol{Z}$ is *standardised*. We can obtain a general bivariate GP vector $\boldsymbol{Z}'$ whose conditional margins $Z'_j \mid Z'_j > 0$ are $\text{GP}(\xi_j, \sigma_j)$ via the transformation

$$Z'_j = \sigma_j \frac{\exp(\xi_j Z_j) - 1}{\xi_j}, \quad j = 1, 2.$$

The bivariate GP distribution can be used to model data whose components exhibit asymptotic dependence. The value of the tail dependence coefficient χ depends on the choice of $\boldsymbol{S}$, see Kiriliouk et al. (2019, Supplementary material, Section B).

2.2.3 Other multivariate extensions of the univariate GP model

The exponential-gamma representation of a univariate GP distribution (Lemma 2.1) provides a natural building block for multivariate extensions. For example, Bopp and Shaby (2017) introduce temporal dependence through a Markov chain structure for the gamma random variable, while Yadav et al. (2021) replace the exponential numerator by a gamma or a

generalized inverse Gaussian distribution, coupled with a latent process to account for spatial dependence.

Recently, Bacro et al. (2025) proposed a multivariate generalization of the representation in Lemma 2.1 by combining independent exponential numerators with shared or individual Gamma denominators, leading to a model that allows for both asymptotic dependence and independence. It does not require marginal standardization and has directly interpretable GP margins. We present the simplest form of their so-called gamma-mixture GP model, defined by the bivariate vector

$$\left(\frac{E_1}{G+G_1}, \frac{E_2}{G+G_2}\right),$$

where E_1 and E_2 are standard exponential random variables, and G, G_1, and G_2 follow gamma distributions with unit scale and shape parameters α, α_1 and α_2, respectively. All variables are assumed to be mutually independent. By construction, each margin is GP distributed with tail index $\xi_j = (\alpha + \alpha_j)^{-1}$. The residual tail dependence coefficient is

$$\eta = \frac{\alpha + \max(\alpha_1, \alpha_2)}{\alpha + 2\max(\alpha_1, \alpha_2)},$$

and in the limiting setting $\alpha_1, \alpha_2 \to 0$ one obtains $\chi = 2^{-\alpha}$. Despite its flexibility, the model has some practical drawbacks. In particular, in the asymptotic dependence regime (when the tail indices of both components coincide at $\xi = 1/\alpha$), the tail dependence coefficient χ is fully determined by ξ, and quickly decreases to zero as ξ increases. Bacro et al. (2025) proposed to use *beta scaling*, i.e., multiplying $G+G_j$ by beta random variables with appropriate parameters, to obtain stronger asymptotic dependence. However, this requires an a priori commitment to asymptotic dependence. Moreover, sub-asymptotic dependence remains weak as the exponential numerators are independent.

The construction proposed in the next section retains the idea of extending the representation of Lemma 2.1 but adds an additive shift and possibly dependent numerators, allowing a richer range of extremal dependence and support on $\mathcal{L}$ (as opposed to $(0, \infty)^2$ for the model of Bacro et al. (2025)), in line with the bivariate GP distribution.

3 A sub-asymptotic model for threshold exceedances

We propose a new model for bivariate extremes which generalizes both the bivariate GP and the Bacro et al. (2025) models. Our construction captures asymptotic dependence and asymptotic independence in a unified framework, with explicit expressions for the tail coefficients χ and η, and operates directly on the original scale of the data (no pre-processing or marginal standardization is required). We start by describing the model and its asymptotic properties in Section 3.1 before moving to its sub-asymptotic behaviour in Section 3.2.

3.1 Model construction and asymptotic properties

Definition 3.1 introduces our bivariate model construction. We will study its marginal distributions, including their speed of convergence towards the GP distribution, and its asymptotic tail dependence properties.

Definition 3.1. The bivariate random vector $\boldsymbol{Y} = (Y_1, Y_2)^T$ is said to follow a *sub-asymptotic bivariate GP (sBGP) distribution* if it can be written as

$$\boldsymbol{Y} = \left(\beta_1 \left(\frac{wE + (1-w)E_1}{G + G_1} - S_1\right), \beta_2 \left(\frac{wE + (1-w)E_2}{G + G_2} - S_2\right)\right)^T,$$

where β_1, β_2 are positive scale parameters, $w \in [0, 1]$ is a weight, E_1, E_2 and E represent standard exponential random variables and G, G_1 and G_2 follow gamma distributions with unit scales and shape parameters α, α_1 and α_2, respectively. The bivariate vector $(S_1, S_2)^T$ is a non-negative random vector such that $\min(S_1, S_2) = 0$. All random variables are mutually independent and the non-negative constants $\alpha_1, \alpha_2, \alpha$ satisfy $\alpha + \min(\alpha_1, \alpha_2) > 0$.

By Breiman's lemma (Breiman, 1965), we know how multiplication by an independent, lighter-tailed variable (here, a weighted sum of exponentials) preserves the tail order. Hence, the vector $\boldsymbol{Y}$ has heavy-tailed margins because $1/(G + G_j)$ is regularly varying of order $(\alpha + \alpha_j)$ for $j = 1, 2$. The parameters $\alpha_1, \alpha_2, \alpha$ control the relative importance of individual and shared gamma components, drive residual tail dependence, and determine the marginal tail indices; see Proposition 3.2(iii). The weight w controls the balance between shared and individual exponential components: as w increases from 0 to 1, the strength of (sub-)asymptotic dependence increases (see also Section 3.2). Table 1 lists the values of the coefficients χ and η for key parameter configurations, which are obtained in Proposition 3.4.

Case	$w = 0$	$w \in (0, 1)$	$w = 1$
$\max(\alpha_1, \alpha_2) > 0$ (AI, $\chi = 0$)		$\eta = \dfrac{\alpha + \max(\alpha_1, \alpha_2)}{\alpha + 2\max(\alpha_1, \alpha_2)}$	
$\alpha_1 = \alpha_2 = 0$ (AD, $\eta = 1$)	$\chi = 2^{-\alpha}$	$\chi \in \left(2^{-\alpha}, 1\right)$	$\chi = 1$

Table 1: Tail dependence coefficients χ and η as functions of the weight w and the parameters α_1, α_2, α.

The *shift vector* $(S_1, S_2)^T$ allows for distinct sub-asymptotic dependence shapes. For example, we can use

$$S_j = \max(T_1, T_2) - T_j, \qquad T_j \overset{\text{iid}}{\sim} N(0, \sigma_T^2), \qquad j = 1, 2. \tag{3.1}$$

This construction implies that $S_j \geq 0$ and automatically enforces $\min(S_1, S_2) = 0$, ensuring that $\boldsymbol{Y}$ is supported on the canonical L-shaped region $\mathcal{L}$. The parameter σ_T controls the dispersion of $\boldsymbol{S}$, allowing one to model finite-sample dependence without altering the asymptotic tail dependence coefficients.

Marginal distributions.

Proposition 3.2 (Marginal density, moments, and upper GP tail behaviour of Y_j.)**.** *Let* $\boldsymbol{Y}$ *be as in Definition 3.1,* $\boldsymbol{S}$ *as in* (3.1) *and fix* $j \in \{1, 2\}$.

(i) ***Density.*** *The marginal density of* Y_j *is*

$$f_{Y_j}(y) = \frac{1}{2\beta_j}\Big\{f_{V_j}(y/\beta_j) + \int_0^\infty f_{V_j}(y/\beta_j + s)\, f_{S_j|S_j>0}(s)\, \mathrm{d}s\Big\}. \tag{3.2}$$

where

$$V_j = \frac{wE + (1-w)E_j}{G + G_j},$$

f_{V_j} *is given in Lemma 8.1, and the distribution of* $S_j \mid S_j > 0$ *is given in* (8.3).

(ii) ***Moments.*** *Let* $\xi_j = (\alpha + \alpha_j)^{-1}$. *The mean and variance are*

$$\mathbb{E}[Y_j] = \beta_j \left(\frac{\xi_j}{1-\xi_j} - \mathbb{E}[S_j] \right), \qquad \text{if } \xi_j \in (0,1), \tag{3.3}$$

$$\mathrm{Var}[Y_j] = \beta_j^2 \left(\frac{w^2 + (1-w)^2 + 2\xi_j w(1-w)}{(\xi_j^{-1}-1)^2(1-2\xi_j)} + \mathrm{Var}[S_j] \right), \qquad \text{if } \xi_j \in (0, \tfrac{1}{2}). \tag{3.4}$$

(iii) ***Upper GP tail convergence.*** *If* $\xi_j = (\alpha + \alpha_j)^{-1}$ *and* $\sigma_j = \beta_j C^{\xi_j} \xi_j$ *with* $C = \frac{w^{1/\xi_j+1} - (1-w)^{1/\xi_j+1}}{2w-1}$, *then*

$$\frac{\mathbb{P}(Y_j > x)}{\overline{H}_{\xi_j,\sigma_j}(x)} = 1 - \left(\frac{\Delta + o(1)}{x} \right), \qquad \text{as } x \to \infty,$$

where $\Delta = \dfrac{\mathbb{E}[S_j] + D - C^{\xi_j}}{\xi_j/\beta_j}$ *and* $D = \dfrac{w^{1/\xi_j+2} - (1-w)^{1/\xi_j+2}}{w^{1/\xi_j+1} - (1-w)^{1/\xi_j+1}}$.

If $w = 0$ or $w = 1$, the random variable V_j follows a $\mathrm{GP}(\xi_j, \xi_j)$ distribution (Lemma 2.1). Note that the parameter w does not appear in the mean of Y_j (Proposition 3.2(ii)). Figure 1 shows how w and σ_T modify the *body* of the marginal density without changing its asymptotic tail. In the left panel, increasing w makes the positive part closer to a Pareto shape: when $w = 1$, the leading term reduces to $E/(G_j + G)$, yielding a Pareto-like density, whereas values of w near $1/2$ replace the numerator by a hypo-exponential mixture, slightly smoothing the peak. In the right panel, larger σ_T spreads mass toward negative values through the additive shift S_j, slowing the convergence toward the GP tail while preserving its slope.

Connections with the multivariate GP distribution. The multivariate GP distribution arises as a boundary case of our proposed construction. Recall that the gamma variables $G_j + G$ have shape parameter $\alpha + \alpha_j$. When $\alpha + \alpha_j \to \infty$, the relative fluctuations of $G_j + G$ vanish, and these variables behave as deterministic constants after rescaling. In addition, setting $w = 1$ removes the mixture component in the numerator, leaving a purely additive exponential representation. In this regime, the vector $\boldsymbol{Y}$ converges to the classical additive form of a multivariate GP distribution with standardised margins.

Lemma 3.3 (Multivariate upper GP limit)**.** *Let* $\mathbf{Y} = (Y_1, Y_2)^T$ *be a sBGP random vector as in Definition 3.1 with* $w = 1$. *Let* $\boldsymbol{S}^{(\beta)} = (S_1^{(\beta)}, S_2^{(\beta)})^T$ *be the associated shift vector, possibly depending on* $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$, *and set* $\beta_j = \sigma_j(\alpha + \alpha_j)$ *for some* $\sigma_j > 0$. *Assume that* $\alpha + \alpha_j \to \infty$ *for* $j = 1, 2$ *and that*

$$\beta_j S_j^{(\beta)} \xrightarrow{p} S_j^0, \qquad j = 1, 2,$$

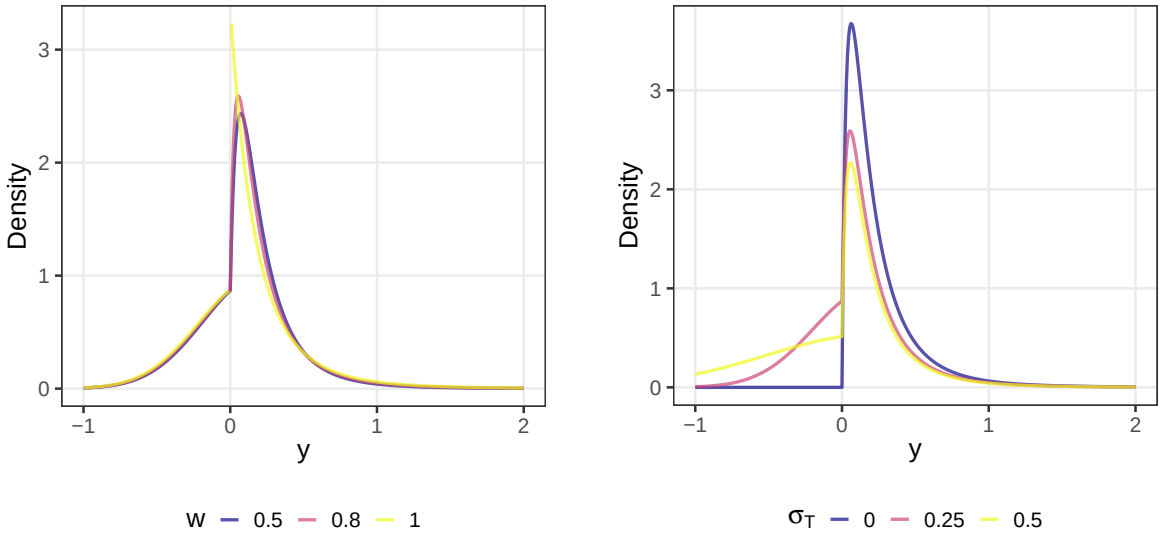


Figure 1: Probability density functions of Y_j in Definition 3.1 with $\xi_j = 0.2$ and $\beta_j = 1$. In the left panel, the parameter $\sigma_T = 0.25$ is fixed and the weight $w \in \{0.5, 0.8, 1\}$ (blue, red, yellow lines), while $w = 0.8$ and $\sigma_T \in \{0, 0.25, 0.5\}$ (blue, red, yellow lines) in the right panel.

for some non-negative random vector $\boldsymbol{S}^0 = (S_1^0, S_2^0)^T$ *with* $\min(S_1^0, S_2^0) = 0$ *a.s. Then*

$$(Y_1, Y_2) \xrightarrow{p} (\sigma_1 E - S_1^0,\ \sigma_2 E - S_2^0),$$

where $E \sim \text{Exp}(1)$. *The limit vector is bivariate GP with shape vector* $\boldsymbol{\xi} = (0, 0)^T$ *and scale vector* $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)^T$.

In the multivariate GP limit, the shift vector $(S_1, S_2)^T$ allows for distinct asymptotic dependence shapes (Kiriliouk et al., 2019, Supplementary material, Section D). For the general model as in Definition 3.1, $(S_1, S_2)^T$ has no impact on asymptotic model properties, driving sub-asymptotic dependence only. When $w = \beta_1 = \beta_2 = 1$, the sBGP model of Definition 3.1 can be rewritten as

$$(Y_1, Y_2) = \left(\frac{E - \tilde{S}_1}{G + G_1},\ \frac{E - \tilde{S}_2}{G + G_2} \right),$$

where $\tilde{S}_j = (G + G_j) S_j$ for $j = 1, 2$. Since $\tilde{S}_1, \tilde{S}_2 \geq 0$ and $\min(\tilde{S}_1, \tilde{S}_2) = 0$, we find a bivariate standardised GP vector $(E - \tilde{S}_1, E - \tilde{S}_2)$ in which each component is divided by a gamma random variable.

Tail dependence properties. The tail dependence of $\boldsymbol{Y}$ is governed by the shared exponential- and gamma-distributed variables. We now derive the values of the coefficients χ and η that are listed in Table 1. We see that the distinction between AD and AI depends solely on the configuration of the gamma-distributed variables. This is consistent with the fact that the gamma-distributed denominator (i.e. the multiplicative inverse gamma-distributed factor) has heavier tails than the weighted sum of exponentials in the numerator.

In particular, we find that the residual tail dependence coefficient η is entirely determined by the gamma shape parameters $\alpha, \alpha_1, \alpha_2$ and is independent of the weight w. By contrast, in

case of asymptotic dependence, the tail dependence coefficient χ can reach a full spectrum of tail dependence (i.e. $\chi \in [0, 1)$) thanks to the weight w.

Proposition 3.4 (Tail dependence structure)**.** *Let $\boldsymbol{Y}$ be as in Definition 3.1, with marginal tail indices $\xi_j = (\alpha + \alpha_j)^{-1}$ for $j = 1, 2$.*

(i) ***Residual tail dependence*** *If $\alpha_1 > 0$ or $\alpha_2 > 0$, then Y_1 and Y_2 are asymptotically independent, so that $\chi = 0$ and*

$$\eta = \frac{\alpha + \max(\alpha_1, \alpha_2)}{\alpha + 2\max(\alpha_1, \alpha_2)},$$

In particular, $\eta = \frac{1}{2}$ if $\alpha = 0$.

(ii) ***Tail dependence.*** *If $\alpha_1 = \alpha_2 = 0$, then by assumption, $\alpha > 0$, and Y_1, Y_2 are asymptotically dependent, so that $\eta = 1$ and*

$$\chi = \frac{2w - 1}{3w - 1} \frac{2w^{\alpha+1} - 2^{-\alpha}(1 - w)^{\alpha+1}}{w^{\alpha+1} - (1 - w)^{\alpha+1}},$$

with $\chi = 2^{-\alpha}$ when $w = 0$ and $\chi = 1$ when $w = 1$.

(iii) ***Constraints induced by marginal asymmetry.*** *The residual tail dependence coefficient satisfies*

$$\eta \in \left[\frac{1}{2}, \frac{1}{2 - \frac{\min(\xi_1, \xi_2)}{\max(\xi_1, \xi_2)}}\right],$$

while $\chi > 0$ only if $\xi_1 = \xi_2$.

Proposition 3.4(iii) illustrates that increasing asymmetry between the marginal tail indices ξ_1 and ξ_2 reduces the maximal attainable value of η. When $\xi_1 = \xi_2$, the model attains full flexibility, with $\eta \in [1/2, 1]$ spanning the range from independence to asymptotic dependence; within the latter regime ($\eta = 1$), the tail dependence coefficient χ varies continuously in $[2^{-\alpha}, 1]$. Note that asymptotic dependence cannot arise when $\xi_1 \neq \xi_2$.

The weight w does not impact the value of η when the model exhibits AI but has a dominant role in shaping the tail dependence coefficient χ in case of AD, as illustrated in Figure 2. The left panel shows that varying w leads to large changes in χ, whereas changes in the marginal tail index $\xi = 1/\alpha$ have a comparatively minor effect. The right panel illustrates how quickly η increases to 1 as a function of the shared coefficient α.

3.2 Sub-asymptotic model features

While Section 3.1 focused on asymptotic properties, convergence to the limiting coefficients (χ, η) can be slow, so that dependence at practically relevant thresholds may deviate substantially from the asymptotic regime. To characterize such behaviour, we examine the model's *sub-asymptotic* features, which describe extremal dependence at intermediate quantile levels that are typically accessible in finite samples.

Asymptotic coefficients alone are therefore insufficient to summarize extremal dependence in practice: distinct processes can share the same limiting coefficients (χ, η) while exhibiting

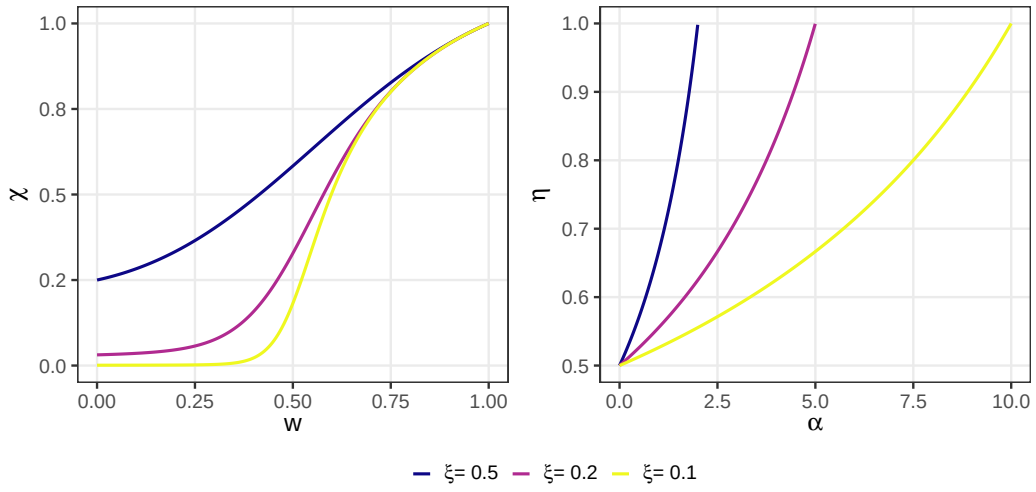


Figure 2: Limiting tail dependence coefficients χ (left panel) and η (right panel) as functions of the weight w in Definition 3.1, for the special case of identical tail indices $\xi_1 = \xi_2 = \xi$.

markedly different dependence strengths at moderately high quantile levels. This phenomenon is illustrated in Figure 3, which shows three simulated datasets with identical $(\chi, \eta) = (0, 0.9)$ but different mixture weights w, resulting in clearly different sub-asymptotic dependence patterns despite identical asymptotic behaviour. Hence, model diagnostics will focus on the full curves $\chi(q)$ and $\eta(q)$ rather than just their limits χ and η (see also Huser et al., 2025; Dell'Oro and Gaetan, 2025). The curves $\eta(q)$ and $\chi(q)$ contain equivalent information, see (2.3). We only focus on $\chi(q)$ in what follows.

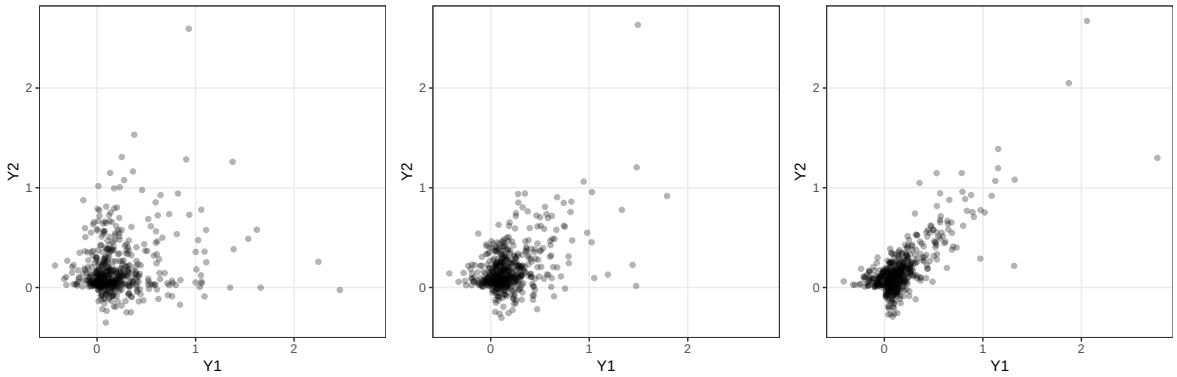


Figure 3: Scatterplots of samples of size $n = 1000$ simulated from the sBGP model (Definition 3.1) with identical coefficients $(\chi, \eta) = (0, 0.9)$ and parameters $(\alpha, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_T) = (4.44, 0.56, 0.56, 1, 1, 0.1)$. From left to right: $w = 0.1$, $w = 0.5$, $w = 0.9$.

No closed-form expression is available for $\chi(q)$ when $q < 1$. We approximate it empirically using a large simulated sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ from the model,

$$\hat{\chi}(q) = \frac{1}{N(1-q)} \sum_{i=1}^{N} \mathbf{1}\{R_{i1} > (N+1)q,\ R_{i2} > (N+1)q\}, \tag{3.5}$$

where R_{ij} denotes the rank of Y_{ij} among $(Y_{1j}, \ldots, Y_{Nj})$ for $j = 1, 2$. This estimator corresponds to the proportion of joint exceedances above the empirical q-quantiles, normalized by the expected number of marginal exceedances $N(1-q)$. Figure 4 displays the resulting $\chi(q)$ curves. For small w (left panel), curves corresponding to different η values remain nearly indistinguishable except extremely close to $q = 1$, reflecting weak sub-asymptotic separation. As w increases (middle and right panels), the weighted exponential sum induces stronger joint extremes and the curves separate earlier, making η practically identifiable from finite samples. The orange curves correspond to the three datasets of Figure 3: although they share identical asymptotic coefficients, their $\chi(q)$ trajectories differ substantially at intermediate quantiles, highlighting the importance of full $\chi(q)$ diagnostics."

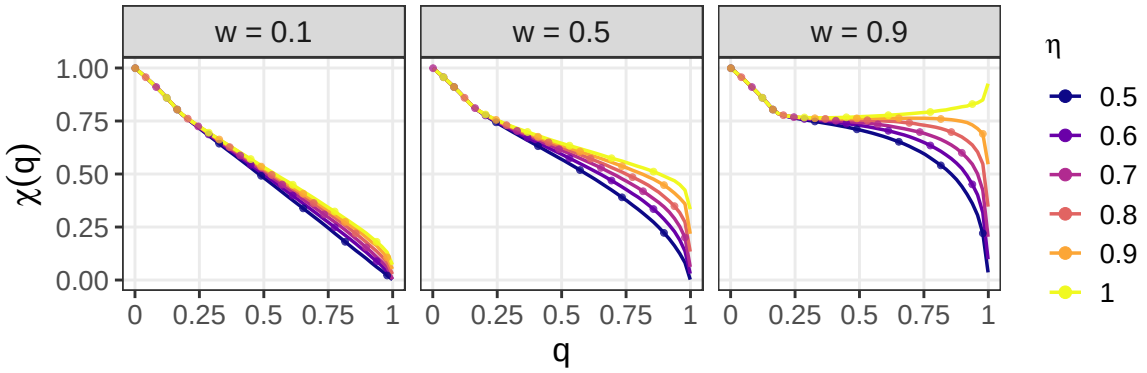


Figure 4: Empirical $\chi(q)$ curves for the sBGP model (3.1). The parameters $(\alpha, \alpha_1, \alpha_2)$ are chosen such that $\xi_1 = \xi_2 = 0.2$ and $\eta \in \{0.5, 0.6, \ldots, 1\}$, with $\sigma_T = 0.1$. Left: $w = 0.1$; Middle: $w = 0.5$; Right: $w = 0.9$. For each pair (w, η), $\chi(q)$ is estimated using (3.5), based on a single synthetic sample of size $N = 10^6$, with q ranging up to 0.999.

In case of asymptotic dependence, Figure 5 further illustrates the slow convergence of $\eta(q)$ to its asymptotic limit $\eta = 1$, for $w = 0$, $\sigma_T = 0$, and $\xi_1 = \xi_2 = 0.2$ (equivalently, $\alpha_1 = \alpha_2 = 0$ and $\alpha = 5$), for which a closed-form expression of $\eta(q)$ is available (Bacro et al., 2025). Hence, for practically observable quantiles (e.g., $q \leq 0.9999$), $\eta(q)$ may misleadingly suggest asymptotic independence. A positive shift σ_T lowers the entire curve but does not accelerate convergence, i.e.,

$$\eta(q; \sigma_T > 0) < \eta(q; \sigma_T = 0), \qquad q < 1,$$

while the limit is unchanged.

Inflection point in the $\chi(q)$ curve. The $\chi(q)$ curves in Figure 4 display an inflection around low to moderate quantiles. This feature is a consequence of the L-shaped support $\mathcal{L}$. As q increases, the conditioning set in the $\chi(q)$ curve definition (2.2) progressively aligns with the geometry of the support. The transition around the marginal threshold creates the observed change in curvature, which is therefore an intrinsic property of the model rather than an artifact. Beyond this point, the curve continues smoothly toward its limiting value.

The shift vector $\boldsymbol{S}$ (as defined in (3.1)) drives this behaviour by ensuring the model is supported on $\mathcal{L}$, capturing observations where one component is extreme while the other remains moderate. Consequently, $\boldsymbol{S}$ governs the finite-sample dependence and the shape of the $\chi(q)$ curve, while leaving the asymptotic coefficients (χ, η) unchanged.

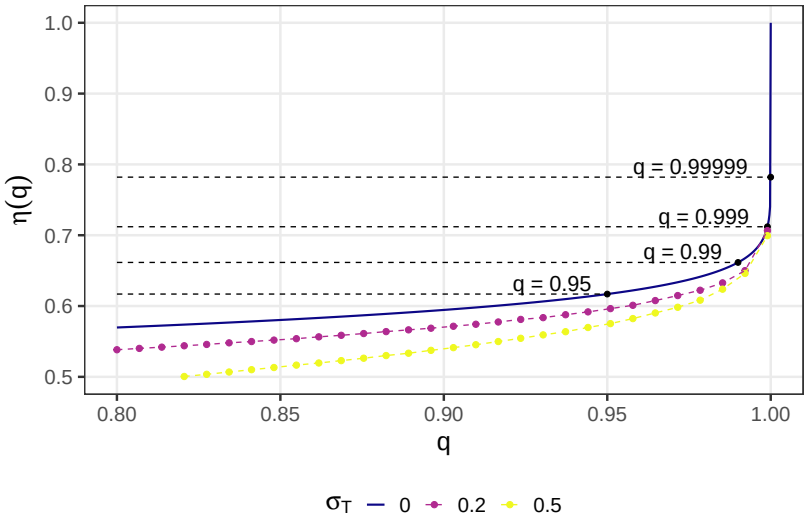


Figure 5: Residual tail-dependence curves $\eta(q)$ for $w = 0$ in the asymptotically dependent case with $\alpha_1 = \alpha_2 = 0$ and $\alpha = 5$ (so that $\xi_1 = \xi_2 = 0.2$ and $\chi = 1/32$). Blue: theoretical curve with $\sigma_T = 0$; Purple and yellow: empirical curves with respectively $\sigma_T = 0.2$ and $\sigma_T = 0.5$.

4 Neural Bayes Estimation

Inference for complex extreme value models is generally challenging, as their likelihood functions often lack closed-form expressions or require computationally expensive numerical approximations. In such scenarios, likelihood-free methods relying on simulation from the model become particularly valuable; examples include approximate Bayesian computation (Sisson et al., 2018), synthetic likelihoods (Wood, 2010), and pseudo-marginal methods (Andrieu and Roberts, 2009). More recently, neural network–based likelihood-free approaches, such as the neural Bayes estimator (NBE) introduced in Sainsbury-Dale et al. (2024), have been increasingly adopted because of their computational efficiency (after initial training) and their ability to learn informative summary statistics automatically, rather than relying on manually chosen ones. Recent work by André et al. (2025) further demonstrated the effectiveness of the NBE framework for inference in complex bivariate extreme-value models.

The NBE approximates the Bayes estimator using a neural network trained on simulated data. Unlike classical inference methods that rely on explicit likelihoods or carefully selected summary statistics, the NBE directly learns a mapping from the data to the parameter estimates, automatically identifying informative features.

Let $\Theta \subseteq \mathbb{R}^p$ denote the parameter space and $f(\cdot \mid \boldsymbol{\theta})$ the density of the sBGP distribution, depending on the unknown parameter vector $\boldsymbol{\theta} \in \Theta$. Following Sainsbury-Dale et al. (2024), the Bayes estimator $\hat{\boldsymbol{\theta}}$ minimizes the Bayes risk $\int_\Theta R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})\, \pi(\mathrm{d}\boldsymbol{\theta})$, where $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{Y} \sim f(\cdot \mid \boldsymbol{\theta})}[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{Y}))]$ is the expected loss under $f(\cdot \mid \boldsymbol{\theta})$, and π is a prior distribution on $\boldsymbol{\theta}$. Since this quantity is generally intractable, the NBE approximates it by empirical risk minimization over simulated observations

$$\boldsymbol{Y}_1^{(k)}, \ldots, \boldsymbol{Y}_{M_k}^{(k)} \overset{\text{iid}}{\sim} f(\cdot \mid \boldsymbol{\theta}^{(k)}), \quad \boldsymbol{\theta}^{(k)} \sim \pi(\cdot), \quad k = 1, \ldots, K$$

where $\boldsymbol{Y}_i^{(k)} = (Y_{i1}^{(k)}, Y_{i2}^{(k)})$ for $i = 1, \ldots, M_k$ and the sample size M_k is randomly drawn from

a discrete distribution to improve generalization across datasets of varying size. The neural network represents the estimator $\hat{\boldsymbol{\theta}}(\cdot)$ and is trained by minimizing the empirical squared loss

$$\hat{\boldsymbol{\theta}}_{\mathrm{NBE}} = \arg\min_{\hat{\boldsymbol{\theta}}\in\mathcal{F}} \left\| D^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\boldsymbol{Y})) \right\|_2^2 = \arg\min_{\hat{\boldsymbol{\theta}}\in\mathcal{F}} \sum_{k=1}^{K} \left\| D^{-1}\left(\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}}(Y^{(k)})\right) \right\|^2. \tag{4.1}$$

where $Y^{(k)}$ is the $M_k \times 2$ matrix with rows $\boldsymbol{Y}_1^{(k)}, \ldots, \boldsymbol{Y}_{M_k}^{(k)}$ and $\mathcal{F}$ denotes the class of functions represented by the neural network. To account for the different scales of the parameters, we normalize the loss using a diagonal scaling matrix D, whose entries correspond to typical magnitudes (or empirical standard deviations) of each component of $\boldsymbol{\theta}$.

In addition to the classical loss defined in (4.1), we also consider a penalized estimator, denoted $\hat{\boldsymbol{\theta}}^{(p)}_{\mathrm{NBE}}$, which minimizes the loss function

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}, \boldsymbol{Y}) = \left\| D^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\boldsymbol{Y})) \right\|_2^2 + \lambda D_{11}^{-2} \left(\hat{\eta}(\boldsymbol{Y}) - \hat{\eta}_{emp}(\boldsymbol{Y})\right)^2. \tag{4.2}$$

The additional term penalizes discrepancies between the estimated residual tail dependence coefficient $\hat{\eta}(\boldsymbol{Y})$ and its empirical counterpart $\hat{\eta}_{emp}(\boldsymbol{Y})$, where $\hat{\eta}_{emp}(\boldsymbol{Y})$ denotes a Hill-type estimator computed from pseudo-observations $U_{ij} = R_{ij}/(n+1)$, with R_{ij} the rank of X_{ij} among $(X_{1j}, \ldots, X_{nj})$. Setting $Z_i = \min\{(1-U_{i1})^{-1}, (1-U_{i2})^{-1}\}$,

$$\hat{\eta}_{emp}(\boldsymbol{Y}) = \frac{1}{k} \sum_{j=1}^{k} \log \frac{Z_{(n-j+1)}}{Z_{(n-k)}}, \tag{4.3}$$

where $k = \lfloor n/10 \rfloor$. This acts as a regularization toward a standard non-parametric estimator of η, improving stability when η is weakly identifiable. In practice, we set $\lambda = 0.5$.

Permutation invariance across replicates is enforced through the DeepSets architecture (Zaheer et al., 2017). Each observation is first transformed by a feature map $\psi(\cdot)$, the resulting representations are aggregated via the mean, and the summary vector is mapped to parameter estimates by $\phi(\cdot)$:

$$\hat{\boldsymbol{\theta}}(Y^{(k)}) = \phi\left(\frac{1}{M_k} \sum_{i=1}^{M_k} \psi\left(\boldsymbol{Y}_i^{(k)}\right) \right).$$

Once trained, the NBE delivers fast, amortized parameter inference without further simulation.

Prior choice. The performance of amortised, simulation-based inference depends critically on the choice of the parameter prior, which shapes the synthetic training distribution and determines which regions of the parameter space are learned accurately. More generally, accuracy depends not only on the number of simulations but also on their quality and relevance to the inferential task; well-designed synthetic data improves generalisation in deep models (He et al., 2019; Ding et al., 2019).

For the parameters $(\beta_1, \beta_2, \sigma_T, w)$ the choice is straightforward, as they are only weakly interdependent, and the distribution of $\boldsymbol{S}$ is chosen as in (3.1). In contrast, $(\alpha, \alpha_1, \alpha_2)$ determine the tail indices $\xi_j = 1/(\alpha + \alpha_j)$ and the residual tail dependence coefficient η. Unrealistically large tail indices generate synthetic samples with excessively large values, causing the training procedure to fail. Hence, we place the prior directly on (ξ_1, ξ_2, η), the real quantities of interest, and recover $(\alpha, \alpha_1, \alpha_2)$ via the transformations in Proposition 3.4.

To enforce the constraint linking ξ_1, ξ_2 and η (Proposition 3.4 (iii)) while maintaining a symmetric treatment of the margins (i.e. the same univariate prior for ξ_1 and ξ_2), we sample η and one of the tail indices freely, and then draw the remaining tail index conditionally. This strategy avoids drawing ξ_1, ξ_2 independently and then η conditionally, which would constrain the admissible range of η too strongly when ξ_1 and ξ_2 are substantially different. However, asymptotic dependence corresponds to the limiting case $\alpha_1 = \alpha_2 = 0$, such a prior never produces samples with $\eta = 1$, causing the network to underperform when tail dependence is strong. To address this, we adopt a mixture prior that draws from the scheme described above (the *AI scheme*) with probability 0.9 and fixes $\eta = 1$ in the remaining 10% of cases (the *AD scheme*). This ensures that the training set includes a non-negligible proportion of asymptotically dependent configurations, allowing the estimator to learn effectively across the full dependence spectrum.

Formally, the sampling scheme (using a conditional prior for ξ_2) is

$$\pi_\eta \sim 0.9\,\mathrm{Unif}(\tfrac{1}{2}, 1) + 0.1\,\delta_{\{\eta=1\}}, \quad \pi_{\xi_1} \sim \mathrm{Unif}(0, \tfrac{1}{2}), \quad \pi_{\xi_2|\eta,\xi_1} \sim \mathrm{Unif}\left(\frac{\xi_1(2\eta-1)}{\eta}, \frac{\xi_1\eta}{2\eta-1}\right),$$

where $\delta_{\{\eta=1\}}$ denotes the Dirac measure at $\eta = 1$. In addition, we take

$$\pi_{\sigma_T} \sim \mathrm{Unif}(0,1), \quad \pi_w \sim \mathrm{Unif}(0,1), \quad \pi_{\beta_j} \sim \mathrm{Unif}(0, 1000) \quad (j = 1, 2),$$

After drawing (ξ_1, ξ_2, η), we invert the mapping to $(\alpha, \alpha_1, \alpha_2)$ to generate training samples. The uniform prior $\mathrm{Unif}(0, \frac{1}{2})$ on ξ_1 and the conditional construction for ξ_2 restrict tail indices to realistic values, preventing synthetic samples with excessively heavy tails. Similarly, the prior $\mathrm{Unif}(0, 1)$ for σ_T was chosen because larger values generate overly dispersed datasets.

Finally, to improve robustness to varying sample sizes, each synthetic dataset used for training is generated with a sample size drawn from $\mathrm{Unif}(100, 1000)$.

Network architecture. As mentioned earlier, we use a DeepSets-based network. It consists of two components: a feature extractor ψ applied independently to each replicate, and a final mapping ϕ from the aggregated representation to parameter estimates. The architecture uses ReLU activations throughout, except for softplus activations on positive parameters and sigmoid activations for η and w, constraining them to $[0.5, 1]$ and $[0, 1]$, respectively.

The outer (feature-extraction) network ψ consists of four fully connected layers with mapping dimensions $2 \mapsto 64 \mapsto 64 \mapsto 128 \mapsto 128$. The inner mapping ϕ takes as input the concatenation of the aggregated representation with a vector of summary statistics $S = \{\hat{\chi}(q)\}$ for $q \in \{0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.95, 0.98\}$, where $\hat{\chi}(q)$ denotes the empirical estimator defined in (3.5). This results in an input dimension of 136. It then passes through four layers with mapping dimensions $136 \mapsto 128 \mapsto 64 \mapsto 64 \mapsto 7$, with ReLU activations at all hidden layers and parameter-specific activations at the output layer. This concatenation of learned summary features and expert-chosen features follows a similar strategy to that employed in Dell'Oro and Gaetan (2025), where empirical $\chi(q)$ values were likewise incorporated as additional expert summary statistics to complement learned representations. The network is trained with the Adam optimizer (learning rate 10^{-4}, batch size 32).

Remark 4.1. Due to the prior specification and the sigmoid activations, the estimator cannot attain boundary values of the supports of η and w. Estimates are therefore biased near the boundaries; for instance, η tends to be underestimated at 1 and overestimated at $1/2$, with similar effects for w at 0 and 1.

5 Simulation Studies

We evaluate the performance of the Neural Bayes Estimator for the sBGP distribution introduced in Section 3.1, with a focus on estimation accuracy and the recovery of extremal dependence features. The estimation procedure was implemented using the `NeuralEstimators` package (Sainsbury-Dale et al., 2024). Training required approximately one hour on a single NVIDIA Volta V100 GPU (64 GB RAM), while inference takes only a few milliseconds per dataset on a standard laptop.

We examine the estimator's performance in a fixed-parameter regime, where the true parameter vector remains constant across simulations. This setting allows us to assess both estimation accuracy and uncertainty quantification under different (known) dependence structures. We consider three configurations

$$\boldsymbol{\theta}^{(1)} = (3, 0, 0, 20, 30, 0.1, 0.8),\ \boldsymbol{\theta}^{(2)} = (2, 1, 1, 20, 30, 0.1, 0.6),\ \boldsymbol{\theta}^{(3)} = (1, 2, 2, 20, 30, 0.1, 0.2), \tag{5.1}$$

where $\boldsymbol{\theta} = (\alpha, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_T, w)$. All three cases share the same values for $(\beta_1, \beta_2, \sigma_T)$, so that differences across configurations are entirely driven by the first four parameters, which govern the marginal tail indices, the residual dependence coefficient, and the strength of sub-asymptotic association.

These configurations span distinct regimes of extremal dependence. For $\boldsymbol{\theta}^{(1)}$, we obtain $\eta = 1$, corresponding to asymptotic dependence; combined with $w = 0.8$, this yields strong extremal dependence across all quantile levels. For $\boldsymbol{\theta}^{(2)}$, we have $\eta = 0.75$, corresponding to asymptotic independence with positive residual dependence, while $w = 0.6$ induces moderate sub-asymptotic association. Finally, $\boldsymbol{\theta}^{(3)}$ yields $\eta = 0.6$, corresponding to asymptotic independence with weaker residual dependence, further attenuated at intermediate levels by the small value $w = 0.2$. These cases therefore represent, respectively, an asymptotic dependence regime, an intermediate regime with residual tail dependence, and a regime of weak dependence both asymptotically and at sub-asymptotic levels. Figure 7 (top) shows samples from the sBGP distribution under the three configurations.

For each configuration $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, 3$, we generate $K = 1000$ independent datasets of size $N = 1000$. We then apply the NBE procedure to obtain a point estimate $\hat{\boldsymbol{\theta}}_k^{(i)}$ and the associated dependence curve $\hat{\chi}_k^{(i)}(q)$, $k = 1, \ldots, K$. Uncertainty is assessed conditionally on each fitted model via a non-parametric bootstrap with $B = 200$ resamples, yielding empirical confidence intervals for both model parameters and associated dependence measures. The results are summarized below in terms of estimation accuracy of point estimates and calibration of the resulting confidence intervals.

Estimation accuracy. Figure 6 displays boxplots of the point estimates $\hat{\boldsymbol{\theta}}_k^{(i)}$ across the K replicated datasets for each configuration. Overall, the NBE demonstrates good estimation accuracy, with parameter estimates closely centred around their true values. Estimation variability for η increases in the second configuration and becomes more pronounced in the third, which is consistent with the progressively weaker sub-asymptotic dependence induced by smaller values of w (see Section 3.2). Finally, η is never estimated exactly equal to 1 due to the upper-bound constraint discussed in Remark 4.1, which prevents the estimator from attaining boundary configurations. This behaviour is expected and does not affect the overall recovery of the dependence structure.

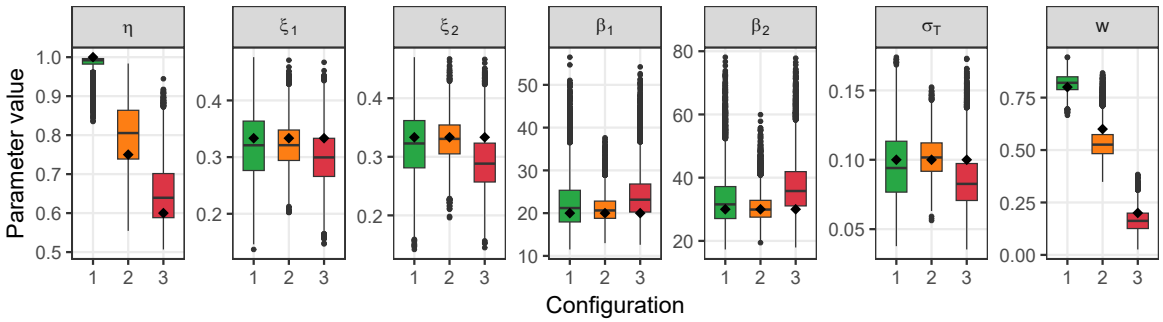


Figure 6: Boxplots for each parameter $(\eta, \xi_1, \xi_2, \beta_1, \beta_2, \sigma_T, w)$ under three dependence settings: $\boldsymbol{\theta}^{(1)}$ (green), $\boldsymbol{\theta}^{(2)}$ (orange), and $\boldsymbol{\theta}^{(3)}$ (red), as defined in (5.1), based on $K = 1000$ samples of size $N = 1000$. Diamond markers indicate true values.

To assess whether estimation errors in the asymptotic dependence parameters propagate to sub-asymptotic dependence measures, we also examine recovery of the function $\chi(q)$ over the full range $q \in (0, 1]$. For each fitted parameter vector $\hat{\boldsymbol{\theta}}_k^{(i)}$, we generate a synthetic sample of size 10^4 and compute the corresponding empirical estimator $\hat{\chi}(q)$ defined in Equation (3.5). Since no closed-form expression of $\chi(q)$ is available for the model, this simulation-based evaluation is used throughout the paper to approximate the dependence curves numerically. Reference curves are obtained from large samples of size 10^5 generated under the true parameters $\boldsymbol{\theta}^{(i)}$. Figure 7 displays a random subset of 100 estimated curves for visual clarity. Across all three configurations, the estimated curves closely match the reference dependence structure, indicating that $\chi(q)$ is accurately recovered even in regimes where individual dependence parameters exhibit increased variability.

Confidence intervals. Uncertainty is quantified using a non-parametric bootstrap applied conditionally on each fitted model. From each of the K datasets and for each parameter configuration, we generate $B = 200$ bootstrap resamples, refit the NBE, and recompute both parameter estimates and the associated $\chi(q)$ curves. This yields empirical 95% confidence intervals for model parameters and associated dependence measures. Table 2 reports empirical coverage probabilities and mean confidence interval widths for a representative subset of parameters across the three configurations.

Overall, the confidence intervals are well calibrated, with empirical coverage close to the nominal level for most parameters. In the first configuration, where $\eta = 1$, coverage for η is not meaningful since the true value lies on the boundary of the parameter space and cannot be attained by the estimator (Remark 4.1); in this case, uncertainty is more appropriately assessed through χ. In the remaining configurations, coverage for η improves substantially. Interval widths increase as w decreases, reflecting the higher uncertainty associated with weaker sub-asymptotic dependence (see Section 3.2). The marginal and scale parameters $(\xi_1, \xi_2, \beta_1, \beta_2, \sigma_T)$ exhibit stable coverage across configurations, with interval widths adapting sensibly to the scale of each parameter. Finally, Table 2 shows that empirical coverage of the $\chi(q)$ curves remains high across quantiles, indicating reliable uncertainty quantification for tail dependence measures even in settings where parameter-level intervals widen.

All results presented in this section are based on the non-penalized estimator defined in (4.1). The impact of the penalized version, as well as additional experiments based on

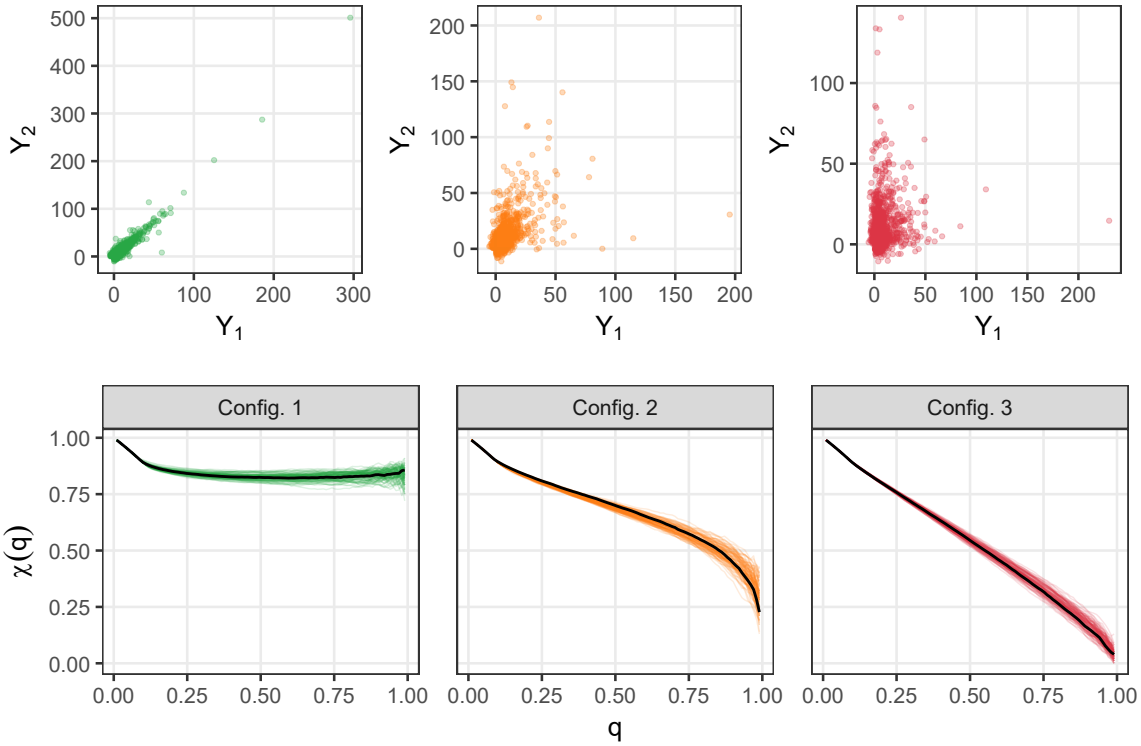


Figure 7: Top: samples from the sBGP distribution (Definition 3.1) with parameters $\boldsymbol{\theta}^{(1)}$ (green), $\boldsymbol{\theta}^{(2)}$ (orange), and $\boldsymbol{\theta}^{(3)}$ (red) (see (5.1)). Bottom: estimated $\chi(q)$ curves. Solid lines correspond to reference curves empirically estimated from samples of size 10^5 generated under the true parameters. Coloured lines show 100 estimated curves based on samples of size 10^4 generated under fitted parameter values.

	Config. 1			Config. 2			Config. 3		
$\hat{\theta}_j$	True	Cov.	Width	True	Cov.	Width	True	Cov.	Width
$\hat{\eta}$	1.00	NA	0.06	0.75	0.94	0.24	0.60	0.92	0.22
$\hat{\xi}_1$	0.33	0.90	0.15	1.50	0.90	0.10	0.33	0.79	0.13
$\hat{\xi}_2$	0.33	0.90	0.15	1.50	0.92	0.09	0.33	0.72	0.13
$\hat{\beta}_1$	20.00	0.90	15.60	20.00	0.93	8.88	20.00	0.81	14.80
$\hat{\beta}_2$	30.00	0.91	22.40	30.00	0.97	12.00	30.00	0.73	22.60
$\hat{\sigma}_T$	0.10	0.90	0.07	0.10	0.99	0.04	0.10	0.80	0.06
$\hat{w}$	0.80	0.92	0.12	0.60	0.76	0.21	0.20	0.81	0.15
$\chi(0.50)$	0.82	0.95	0.05	0.70	0.87	0.05	0.55	0.98	0.05
$\chi(0.70)$	0.82	0.96	0.06	0.60	0.96	0.07	0.37	0.98	0.08
$\chi(0.80)$	0.83	0.99	0.07	0.54	0.99	0.09	0.27	0.96	0.09
$\chi(0.90)$	0.84	0.98	0.09	0.45	0.98	0.12	0.16	0.93	0.10
$\chi(0.95)$	0.84	0.99	0.11	0.37	0.96	0.16	0.10	1.00	0.10
$\chi(0.99)$	0.85	1.00	0.18	0.23	1.00	0.25	0.03	1.00	0.10

Table 2: Empirical coverage probability (Cov.) and mean confidence interval width (Width) for model parameters and tail dependence curves $\chi(q)$ under three fixed configurations (see (5.1)), based on $N = 1000$, $K = 1000$, $B = 200$ bootstrap resamples per estimate, together with the true values.

randomly generated parameter configurations, are reported in the Supplementary Material.

6 Application to extreme precipitation in Belgium

We illustrate the capabilities of the sBGP model for extreme precipitation over Belgium, using gridded data from the ERA5 reanalysis dataset (Hersbach et al., 2020). This dataset provides hourly precipitation estimates on a 0.25° grid from 1950 to the present, derived from a global reanalysis of meteorological observations. Our aim is to assess the model's ability to capture a range of extremal dependence behaviours, from strong asymptotic dependence to near-independence, while also accurately representing marginal distributions. We use the penalized version of the neural Bayes estimator defined in (4.2). This penalization is optional and does not alter the underlying model, but stabilizes inference in practice; results based on the non-penalized estimator are given in the Supplementary Material.

From the full temporal coverage (1950—2024), we compute weekly maxima of daily precipitation (in millimetres) for all grid points in the rectangle $[0^\circ, 7^\circ]\text{E} \times [47^\circ, 52^\circ]\text{N}$; this region is shown in Figure A.17 of the Supplementary material. This block size is chosen to mitigate short-range temporal dependence. We restrict attention to the autumn season (21 September–21 December), which is associated with more homogeneous large-scale precipitation processes. At each grid point, we extract threshold exceedances relative to the empirical 70^{th} percentile. Other thresholds were also examined and led to qualitatively similar fits and interpretations; see Section A.3 in the Supplementary Material.

We start with a reference grid point centred at (4.25° E, 50.75° N) that contains Brussels. We first consider all pairs formed between this reference location and grid points within the

surrounding region, and fit the proposed bivariate model defined in Definition 3.1 to each pair. This allows us to examine how the estimated model parameters vary across space. Next, we select three representative pairs within the region, and apply our model to evaluate the extremal dependence curve $\chi(q)$ and the marginal tail distributions. Finally, we compare the fit of our model with that of a multivariate GP distribution.

6.1 Spatial patterns of bivariate precipitation extremes

For each of the 506 pairs consisting of Brussels and another grid point in the region, we fit the sBGP model to the threshold exceedances. Figure 8 illustrates several key parameters as spatial fields, i.e., each map reports, at a given location, the parameter estimate obtained for the pair between Brussels and that grid point. Maps for the remaining parameters are provided in Section A.3 of the Supplementary Material.

In the following, the extremal dependence coefficient $\hat{\eta}$ is displayed only for locations where $\hat{w} > 0.5$; as illustrated in Figure 4, estimates of η become unreliable when w is small.

The parameters $\hat{w}$ and $\hat{\sigma}_T$ vary coherently with distance from the reference grid point: fitted pairs tend to display weaker dependence as distance increases, with $\hat{w}$ decreasing and $\hat{\sigma}_T$ increasing. For locations with sufficiently large $\hat{w}$, the estimated extremal dependence coefficient $\hat{\eta}$ tends to decrease with distance from the reference grid point, reflecting a gradual weakening of extremal dependence.

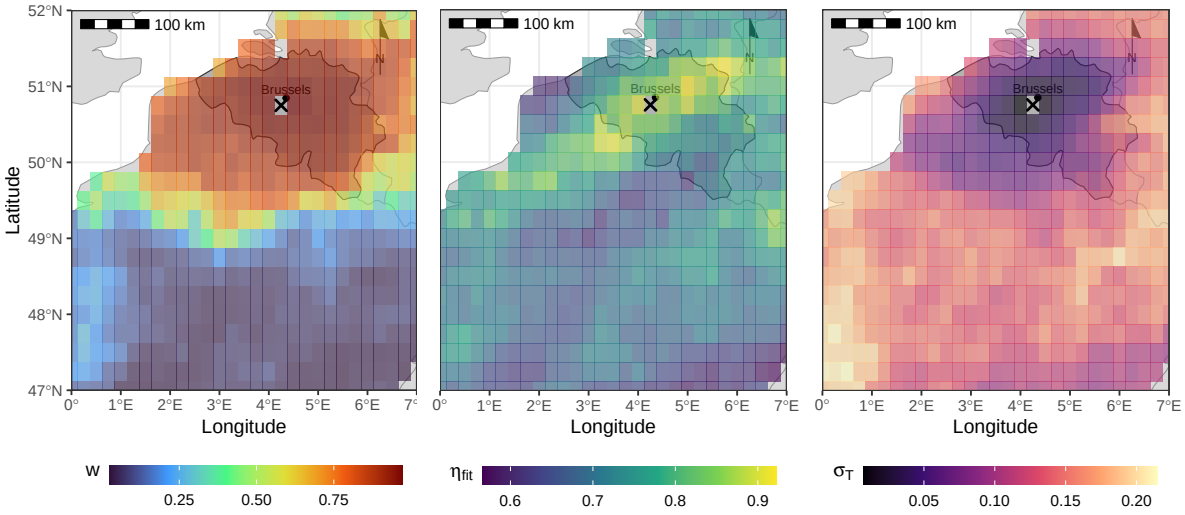


Figure 8: Spatial fields of estimated dependence parameters for bivariate precipitation extremes between Brussels (reference location, black cross) and each grid point. Each panel displays the parameter estimate obtained from fitting the sBGP model to threshold exceedances of the corresponding bivariate pair. From left to right: $\hat{w}$, $\hat{\eta}$, and $\hat{\sigma}_T$.

6.2 Diagnostics for selected pairs

We select two pairs to illustrate different dependence regimes: (Brussels, Nivelles) and (Ostend, Spa). These pairs differ in distance between sites, inland versus coastal position, and regional

topographic exposure. The selected grid points are shown in Section A.3 of the Supplementary Material, with corresponding exceedances displayed in Figure 9.

For all $t = 1, \ldots, T$, let

$$\boldsymbol{Z}_t^{(i)} = \begin{cases} (Z_t^{\text{Bru}}, Z_t^{\text{Niv}}), & i = 1, \\ (Z_t^{\text{Ost}}, Z_t^{\text{Spa}}), & i = 2, \end{cases}$$

denote the weekly rainfall observations.

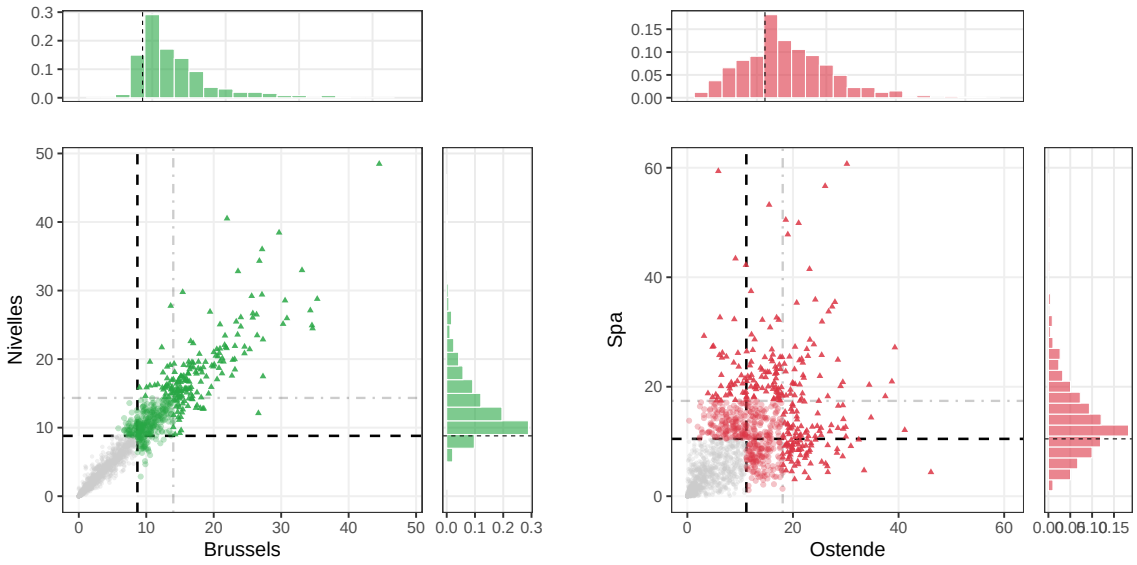


Figure 9: Bivariate scatterplots of threshold exceedances of weekly precipitation maxima for Brussels–Nivelles and Ostend–Spa. Green points denote exceedances above the 70th percentile (see (6.1)), used for fitting the sBGP model, and green triangles denote exceedances above the 90th percentile, used for the comparison with the bivariate GP model in Section 2.2.2. Black and gray dashed lines indicate the marginal 70th- and 90th-percentile thresholds, respectively. The marginal histograms display the distributions of marginal exceedances above the 70th percentile.

Let

$$\begin{aligned} \tilde{\mathcal{Z}}_{0.7}^{(1)} &= \{(Z_t^{\text{Bru}}, Z_t^{\text{Niv}}) : \hat{F}_{\text{Bru}}(Z_t^{\text{Bru}}) > 0.7 \text{ or } \hat{F}_{\text{Niv}}(Z_t^{\text{Niv}}) > 0.7\}, \\ \tilde{\mathcal{Z}}_{0.7}^{(2)} &= \{(Z_t^{\text{Ost}}, Z_t^{\text{Spa}}) : \hat{F}_{\text{Ost}}(Z_t^{\text{Ost}}) > 0.7 \text{ or } \hat{F}_{\text{Spa}}(Z_t^{\text{Spa}}) > 0.7\}. \end{aligned} \tag{6.1}$$

denote the subsets of exceedances for the pairs Brussels–Nivelles and Ostend–Spa, respectively. Here, $\hat{F}_j$ denotes the empirical cumulative distribution function of station j. At this threshold, the numbers of exceedances are 522 for Brussels–Nivelles and 671 for Ostend–Spa.

We fit the sBGP model to the exceedances of the two selected pairs, $\tilde{\mathcal{Z}}_{0.7}^{(1)}$ and $\tilde{\mathcal{Z}}_{0.7}^{(2)}$. Parameter estimates with 95% confidence intervals are reported in Table 3. Marginal tail parameters (ξ_j, β_j) vary moderately across locations. The dependence parameters, however, display clear contrasts between the two regimes: for Brussels–Nivelles, $\hat{\eta} = 0.85$ with $\hat{w} = 0.95$; for Ostend–Spa, $\hat{\eta} = 0.72$ with $\hat{w} = 0.16$. The shift parameter $\hat{\sigma}_T$ also differs substantially between the two pairs, from 0.03 for the nearby inland pair to 0.14 for the distant coastal–inland pair. Results for alternative thresholds are reported in the Supplementary Material.

	Brussels–Nivelles	Ostend–Spa
η	0.85 (0.72, 0.93)	0.72 (0.64, 0.83)
ξ_1	0.11 (0.09, 0.15)	0.14 (0.11, 0.17)
ξ_2	0.10 (0.08, 0.14)	0.14 (0.11, 0.18)
β_1	49.46 (35.32, 62.48)	41.68 (28.32, 56.27)
β_2	61.54 (39.39, 73.06)	46.79 (35.46, 60.85)
σ_T	0.03 (0.02, 0.04)	0.14 (0.10, 0.20)
w	0.95 (0.91, 0.96)	0.16 (0.08, 0.31)

Table 3: Parameter estimates (with 95% confidence intervals) for the sBGP model (Definition 3.1) fitted to the exceedance sets $\tilde{\mathcal{Z}}_{0.7}^{(1)}$ and $\tilde{\mathcal{Z}}_{0.7}^{(2)}$ defined above.

Overall, the Brussels–Nivelles pair displays pronounced sub-asymptotic dependence ($w \approx 0.95$), while the Ostend–Spa pair shows much weaker sub-asymptotic dependence ($w \approx 0.16$). The latter pair also has a smaller estimated residual tail dependence coefficient. However, as explained in Section 3.2, when w is small, the focus should be on sub-asymptotic measures such as $\chi(q)$ or $\eta(q)$.

All dependence diagnostics in this subsection, including $\chi(q)$ curves, are computed from exceedances in $\tilde{\mathcal{Z}}_{0.7}^{(1)}$ and $\tilde{\mathcal{Z}}_{0.7}^{(2)}$, so that the quantiles q refer to empirical quantiles within each exceedance subset rather than the full sample. Figure 10 presents the estimated $\hat{\chi}_{\tilde{\mathcal{Z}}_{0.7}}(q)$ curves, defined in Equation (3.5), obtained from $B = 100$ nonparametric bootstrap replications of the exceedance data. For each bootstrap resample, model parameters were re-estimated, yielding a bootstrap sample of parameter values, and the associated $\chi(q)$ curves were evaluated numerically using the same simulation-based procedure as in the simulation study.

Brussels–Nivelles exhibits strong dependence up to around the 0.9 quantile of the exceedance subset $\tilde{\mathcal{Z}}_{0.7}^{(1)}$, after which the $\chi(q)$ curve drops abruptly, revealing a weakening of dependence in the upper tail. Ostend–Spa shows a marked decline at the most extreme quantiles, indicating near-independence in the tail. These distinct patterns are well reproduced by the fitted model.

For both pairs, the $\chi_{\tilde{\mathcal{Z}}_{0.7}}(q)$ curves display visible discontinuities at low quantiles. This feature reflects the censoring inherent in the L-shaped support of the model, where at least one variable remains below its threshold. The transition point corresponds to the proportion of observations that are not simultaneously extreme in both margins. A detailed discussion of this mechanism is provided in Section 3.2.

Marginal adequacy is assessed via QQ-plots (Figure 11) comparing the empirical exceedance distributions in $\tilde{\mathcal{Z}}_{0.7}^{(1)}$ and $\tilde{\mathcal{Z}}_{0.7}^{(2)}$ to their corresponding fitted marginal models. Overall, deviations from the identity line are minor, indicating satisfactory marginal fits. In particular, the model reproduces well the upper-tail quantiles, which are of primary importance in applications.

In addition to the QQ-plots, Figure 12 compares the empirical marginal densities with their fitted counterparts. The estimated model densities, computed from Equation (3.2), closely follow the empirical histograms, confirming that the marginal fits adequately reproduce both the bulk and the upper tail of the data.

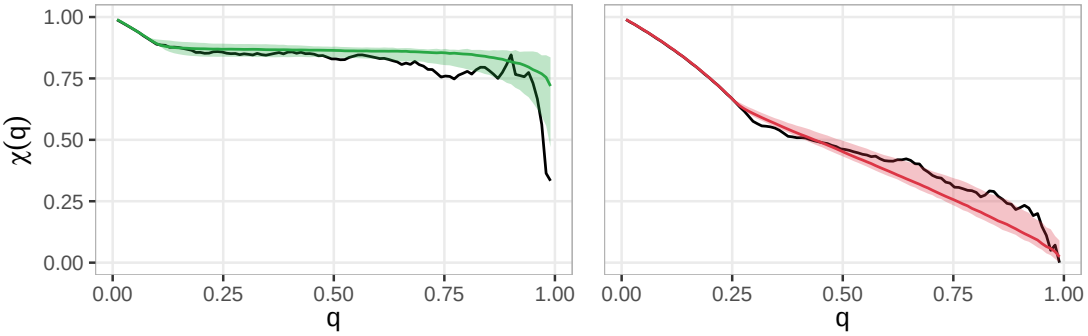


Figure 10: Estimated $\chi_{\tilde{\mathcal{Z}}_{0.7}}(q)$ curves for Brussels–Nivelles and Ostend–Spa, where $\chi(q)$ is estimated using (3.5). Quantiles q are defined relative to the exceedance subset $\tilde{\mathcal{Z}}_{0.7}$ (i.e., within the exceedance region, $u_j = F_j^{-1}(0.7)$). The black line shows the empirical $\chi_{\tilde{\mathcal{Z}}_{0.7}}(q)$ from observed exceedances; the green and red lines are the model estimates; the shaded area denotes the 95% bootstrap confidence region.

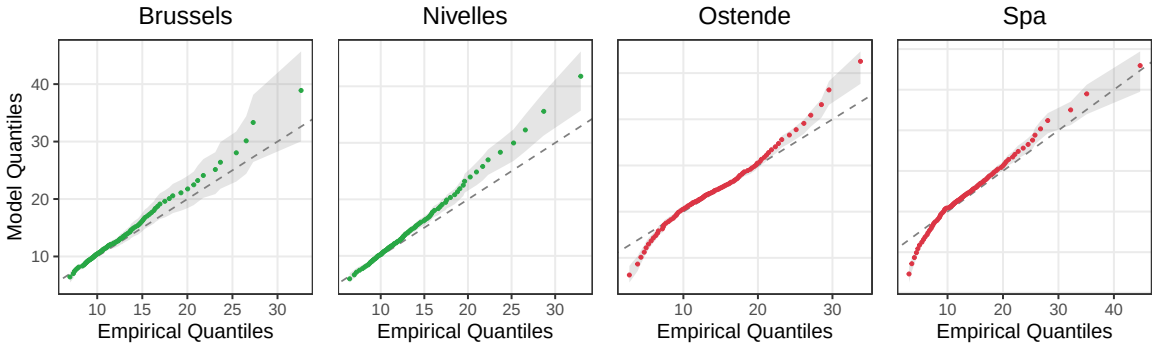
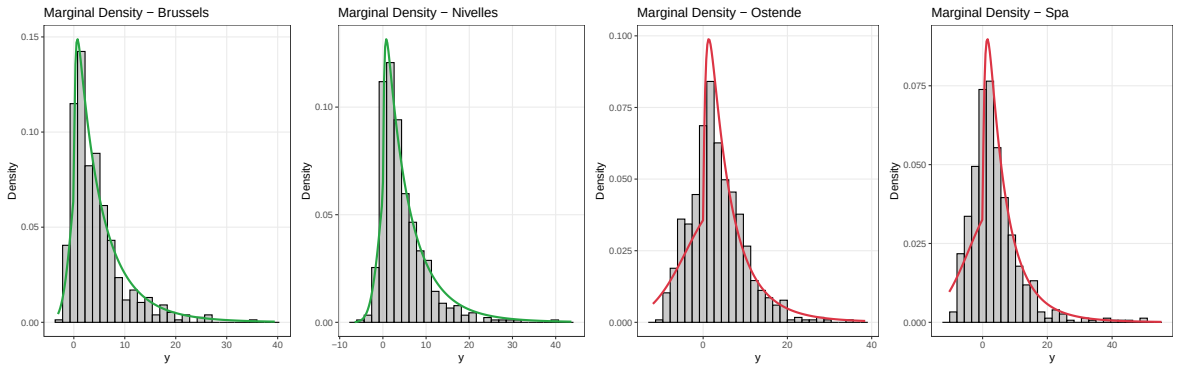


Figure 11: QQ-plots comparing empirical marginal exceedances to fitted marginal distributions for Brussels–Nivelles and Ostend–Spa under the sBGP model (Definition 3.1). Each panel corresponds to one margin of the respective pair. All plots are based on exceedances above the marginal threshold ($u_j = F_j^{-1}(0.7)$), with dashed vertical lines marking the threshold values. Theoretical quantiles are shown on the original data scale, obtained by adding the threshold to the modeled exceedances.



Bruxelles Nivelles Ostende Spa

Figure 12: Empirical histograms (gray bars) and fitted marginal densities (solid lines) for Brussels–Nivelles and Ostend–Spa. Each panel corresponds to one margin of the respective pair. Model-based densities are computed from the fitted sBGP model (Definition 3.1) according to Equation (3.2), using exceedances above the marginal threshold ($u_j = F_j^{-1}(0.7)$).

6.3 Comparison with the bivariate GP distribution

To benchmark the proposed approach, we compare it with the bivariate GP distribution (see Section 2.2.2), the standard asymptotic model for threshold exceedances under asymptotic dependence. In both models, the shift (spectral) vector $\mathbf{S}$ is defined as $\boldsymbol{S} = \max(\boldsymbol{T}) - \boldsymbol{T}$. In our model, $\boldsymbol{T}$ is specified as in (3.1), whereas in the bivariate GP model it is specified through an independent Gumbel generator with common shape and scale parameters, i.e.

$$T_1, T_2 \overset{\text{iid}}{\sim} \text{Gumbel}(a_T, b_T);$$

other choices of $\boldsymbol{T}$ led to similar results.

The bivariate GP distribution is threshold-stable and theoretically valid only at sufficiently high quantiles for asymptotically dependent vectors; we therefore restrict the comparison to the pair Brussels–Nivelles and fit both models using exceedances above the empirical 90th percentile (Kiriliouk et al., 2019). The corresponding exceedance subset is defined as

$$\tilde{\mathcal{Z}}_{0.9}^{(1)} = \{(Z_t^{\text{Bru}}, Z_t^{\text{Niv}}) : \hat{F}_{\text{Bru}}(Z_t^{\text{Bru}}) > 0.9 \text{ or } \hat{F}_{\text{Niv}}(Z_t^{\text{Niv}}) > 0.9\},$$

that is, the set of observations where at least one margin exceeds its 90th-percentile threshold. At this higher threshold, the number of exceedances for the Brussels–Nivelles pair is 187.

Results. Table 4 reports the parameter estimates for both models fitted to the Brussels–Nivelles pair. For the proposed model, the estimated dependence and tail parameters are $\hat{\eta} = 0.66$ and $\hat{w} = 0.94$, together with moderate marginal tail indices ($\hat{\xi}_1 = 0.13$, $\hat{\xi}_2 = 0.13$). The corresponding scale parameters are $\hat{\beta}_1 = 37.8$ and $\hat{\beta}_2 = 36.6$, while the shift variability parameter $\hat{\sigma}_T = 0.06$ indicates a moderate level of sub-asymptotic variability. For the GP distribution, parameter estimates are reported for comparison and exhibit slightly larger tail indices together with a strong dependence parameter $\hat{\alpha} = 2.4133$, consistent with its asymptotic dependence structure.

Confidence intervals were obtained via bootstrap procedures: a nonparametric bootstrap was used for the proposed model, whereas a parametric bootstrap was employed for the GP distribution. In the latter case, bootstrap samples were generated from the fitted distribution and refitted to obtain the empirical distribution of the estimators. The $\chi(q)$ curves for the GP distribution were then evaluated numerically using the same simulation-based procedure as described in the simulation study, from which 95% confidence bands were derived.

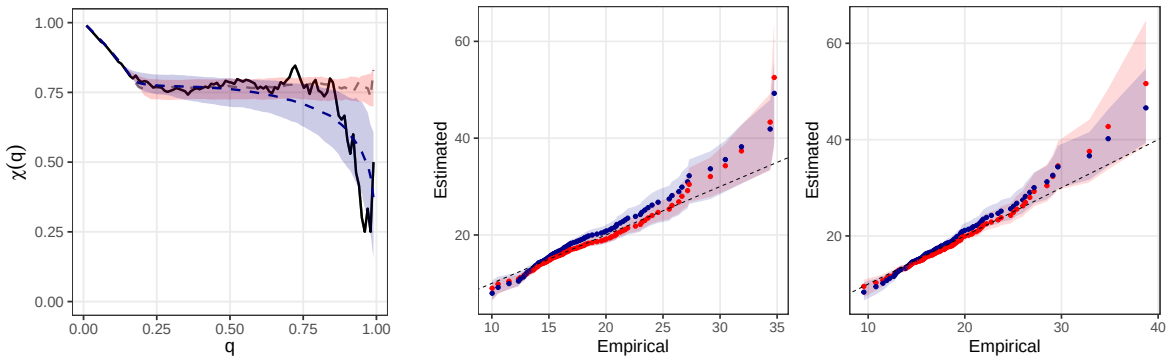
Figure 13 compares the estimated $\hat{\chi}_{\tilde{\boldsymbol{Z}}_{0.9}^{(1)}}(q)$ curves and the marginal QQ plots for both models. While the proposed model reproduces the empirical marginal distributions satisfactorily, the GP distribution exhibits discrepancies in the marginal QQ plots, indicating a poorer fit. This behaviour is more likely attributable to structural model misspecification rather than estimation issues: the GP model relies on threshold stability, an assumption that is not supported by the data in this case. When threshold stability is violated, fitting the GP distribution at a fixed level may induce biases in the parameter estimates and, in this case, an overestimation of the marginal tail indices ξ_j, resulting in overly heavy estimated tails and a degraded marginal fit. Regarding extremal dependence, the proposed model closely follows the empirical $\chi(q)$ curve across most quantiles, capturing the gradual decrease observed beyond $q \approx 0.9$. In contrast, the GP model yields an almost constant $\chi(q)$ function, as expected under asymptotic dependence, but is less able to reflect the observed weakening of dependence.

	Our model	**MGPD**
η	0.66 (0.57, 0.82)	—
ξ_1	0.15 (0.09, 0.22)	0.26 (0.09, 0.40)
ξ_2	0.14 (0.08, 0.22)	0.27 (0.11, 0.41)
β_1^*	6.70 (4.98, 7.79)	4.37 (3.69, 5.43)
β_2^*	6.28 (5.21, 7.51)	4.29 (3.36, 5.16)
σ_T	0.07 (0.04, 0.13)	—
w	0.95 (0.87, 0.98)	—
a_T	—	2.41 (1.97, 2.91)
b_T	—	-0.02 (-0.18, 0.08)

Table 4: Parameter estimates with 95% confidence intervals for the sBGP and the bivariate GP model, fitted to the pair Brussels–Nivelles using exceedances above the 90th percentile. Parameters that are not defined for a given model are indicated by "—". For the proposed model, β_j^* denotes the rescaled version of β_j, obtained by dividing by the expectation of the first latent component (see Eq. (3.3)), in order to ensure comparability with the MGPD parametrisation. For the MGPD, β_j^* corresponds to the standard scale parameter.

7 Discussion

We introduced a flexible sub-asymptotic model for bivariate threshold exceedances that extends classical generalized Pareto constructions. The proposed model accommodates both asymptotic dependence and asymptotic independence within a unified framework while preserving margins that converge to generalized Pareto tails. By explicitly modelling the sub-asymptotic regime, the model captures a wide range of dependence behaviours at practically relevant thresholds. Inference is performed through a neural Bayes estimator trained on simulated data, providing an



(a) $\chi(q)$ curves. (b) Marginal QQ plots.

Figure 13: Comparison between the sBGP and the bivariate GP model for Brussels–Nivelles. (a) Estimated $\chi_{\tilde{\mathcal{Z}}_{0.9}^{(1)}}(q)$ curves and (b) marginal QQ plots. Blue lines and shaded regions correspond to the proposed model, and red lines to the MGPD. Dashed lines represent model-based estimates, solid lines the empirical values, and shaded areas indicate 95% bootstrap confidence regions. Both models were fitted to exceedances $\tilde{\mathcal{Z}}_{0.9}^{(1)}$ (threshold $u_j = F_j^{-1}(0.9)$).

efficient likelihood-free estimation strategy. Simulation experiments and the rainfall application illustrate that the model captures a broad range of extremal dependence behaviours while remaining interpretable and computationally tractable.

In theory, the construction admits a natural extension to dimensions $d > 2$. Let $E_J \sim \text{Exp}(1)$ and $G_J \sim \Gamma(\alpha_J, 1)$ for each subset $J \subseteq D = \{1, \ldots, d\}$ with $|J| \geq 1$, all mutually independent. For $j \in D$, let

$$S_j = \max(T_1, \ldots, T_d) - T_j, \quad (T_1, \ldots, T_d) \sim \mathcal{N}_d(0, \sigma I_d),$$

we may define the components of a d-variate sub-asymptotic GP vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ as

$$Y_j = \beta_j \left(\frac{\sum_{J \subseteq D: j \in J} w_J E_J}{\sum_{J \subseteq D: j \in J} G_J} - S_j \right),$$

where $w_J \geq 0$ and $\sum_{J \subseteq D: j \in J} w_J = 1$. In practice, however, the full model involves many parameters and may suffer from identifiability issues, suggesting that parsimonious specifications are preferable. Hence, future work will focus on extending the sBGP distributions to $d > 2$ while retaining a limited number of parameters.

8 Proofs

This section contains proofs of all main results stated in the main paper. Some of these rely on additional lemmas whose proofs have been deferred to the Supplementary Material.

Recall that a function $f : (0, \infty) \to (0, \infty)$ is *regularly varying* (at infinity) of order $\rho \in \mathbb{R}$, written $f \in \text{RV}(\rho)$, if for all $x > 0$,

$$\lim_{t \to \infty} \frac{f(tx)}{f(t)} = x^{\rho}.$$

A cumulative distribution function (cdf) F with upper endpoint $x^* = \infty$ has regularly varying (upper) tail of order $\rho \geq 0$ if the survival function $\overline{F}(x) = 1 - F(x)$ is in $\mathrm{RV}(-\rho)$. A regularly varying F is necessarily *heavy-tailed*; for a thorough inventory of upper tail behaviour classes, see Engelke et al. (2019).

8.1 Marginal behaviour

Recall the model in Definition 3.1. Write

$$Y_j = \beta_j(V_j - S_j), \qquad \text{where } V_j = \frac{wE + (1-w)E_j}{G + G_j} \qquad j = 1, 2,$$

and $\xi_j = (\alpha + \alpha_j)^{-1}$. Below, Lemma 8.1 derives the survival function and density of V_j and its first two moments, and Lemma 8.2 gives an asymptotic expansion of the survival function of V_j. These are used in the proof of Proposition 3.2.

Lemma 8.1 (Distribution of V_j)**.**

1. *For $x \geq 0$ and $w \in [0,1]$,*

$$\overline{F}_{V_j}(x) = \begin{cases} \dfrac{w(1+x/w)^{-1/\xi_j} - (1-w)(1+x/(1-w))^{-1/\xi_j}}{2w-1}, & w \in (0,1) \setminus \{\frac{1}{2}\}, \\ (1+2x)^{-1/\xi_j - 1}\left(1 + 2x(1+\xi^{-1})\right), & w = \frac{1}{2}. \end{cases}$$

The corresponding density is

$$f_{V_j}(x) = \begin{cases} \dfrac{\mathbb{1}\{x \geq 0\}}{\xi_j(2w-1)}\left[\left(1+\dfrac{x}{w}\right)^{-1/\xi_j - 1} - \left(1+\dfrac{x}{1-w}\right)^{-1/\xi_j - 1}\right], & w \in (0,1) \setminus \{\frac{1}{2}\}, \\ \mathbb{1}\{x \geq 0\}\dfrac{4x}{\xi_j}\left(1+\dfrac{1}{\xi_j}\right)(1+2x)^{-1/\xi_j - 2}, & w = \frac{1}{2}. \end{cases}$$

2. *The expected value and the variance of V_j are*

$$\begin{cases} \mathbb{E}[V_j] = \dfrac{\xi_j}{1-\xi_j}, & 0 < \xi_j < 1, \\ \mathrm{Var}[V_j] = \dfrac{w^2 + (1-w)^2 + 2\xi_j w(1-w)}{(\xi_j^{-1} - 1)^2(1-2\xi_j)}, & 0 < \xi_j < \frac{1}{2}. \end{cases}$$

Lemma 8.2 (Asymptotic expansion of $\overline{F}_{V_j}$)**.** *For $x \to \infty$ and $w \in [0,1]$,*

$$\overline{F}_{V_j}(x) = C x^{-1/\xi_j}\left(1 - \frac{D}{\xi_j x} + o(x^{-1})\right), \tag{8.1}$$

where, for $w \in [0,1] \setminus \frac{1}{2}$,

$$C = \frac{w^{1/\xi_j+1} - (1-w)^{1/\xi_j+1}}{2w-1} \quad \text{and} \quad D = \frac{w^{1/\xi_j+2} - (1-w)^{1/\xi_j+2}}{w^{1/\xi_j+1} - (1-w)^{1/\xi_j+1}}. \tag{8.2}$$

If $w \in \{0,1\}$, $C = D = 1$. Expressions for $w = 1/2$ are given in the proof.

Proof of Proposition 3.2.

(i) We state the results for $j = 1$ without loss of generality. Recall that $S_1 = \max(T_1, T_2) - T_1 = \max(0, T_2 - T_1)$ so that $\mathbb{P}[S_1 = 0] = 1/2$ for $T_1, T_2 \overset{\text{iid}}{\sim} N(0, \sigma_T^2)$ as in (3.1). We have

$$\begin{aligned} F_{Y_1}(y) &= \frac{1}{2}\left\{\mathbb{P}[Y_1 \le y \mid S_1 = 0] + \mathbb{P}[Y_1 \le y \mid S_1 > 0]\right\} \\ &= \frac{1}{2}\left\{F_{V_1}(y/\beta_1) + \mathbb{E}[F_{V_1}(y/\beta_1 + S_1) \mid S_1 > 0]\right\} \\ &= \frac{1}{2}\left\{F_{V_1}(y/\beta_1) + \int_0^\infty F_{V_1}(y/\beta_1 + s) f_{S_1|S_1>0}(s)\,\mathrm{d}s\right\} \end{aligned}$$

Differentiating with respect to y given the density expression in (3.2). In addition,

$$\mathbb{P}[S_1 \le s \mid S_1 > 0] = \frac{\mathbb{P}[0 \le T_2 - T_1 \le s]}{\mathbb{P}[T_2 > T_1]} = 2\Phi\left(\frac{s}{\sqrt{2}\sigma_T}\right) - 1, \tag{8.3}$$

which is a half-normal distribution (i.e. folded normal with mean zero) with scale $\sqrt{2}\sigma_T$.

(ii) The mean and variance of V_j are given in Lemma 8.1. If $T_1, T_2 \overset{\text{iid}}{\sim} N(0, \sigma_T^2)$, then

$$\mathbb{E}[S_j] = \frac{\sigma_T}{\sqrt{\pi}}, \qquad \mathrm{Var}[S_j] = (1 - 1/\pi)\sigma_T^2$$

(iii) First, note that

$$\lim_{x\to\infty} \frac{\overline{F}_{V_j - S_j}(x)}{\overline{F}_{V_j}(x)} = \lim_{x\to\infty} \mathbb{E}\left[\frac{\overline{F}_{V_j}(x + S_j)}{\overline{F}_{V_j}(x)}\right] = \mathbb{E}\left[\lim_{x\to\infty} \frac{\overline{F}_{V_j}(x + S_j)}{\overline{F}_{V_j}(x)}\right],$$

where the last step follows from dominated convergence. Using expansion (8.1), we get, for $s > 0$ and $x \to \infty$,

$$\begin{aligned} \frac{\overline{F}_{V_j}(x+s)}{\overline{F}_{V_j}(x)} &= \left(1 + \frac{s}{x}\right)^{-1/\xi_j} \left(\frac{1 - \frac{D}{\xi_j(x+s)} + o(x^{-1})}{1 - \frac{D}{\xi_j x} + o(x^{-1})}\right) \\ &= \left(1 + \frac{s}{x}\right)^{-1/\xi_j} (1 + o(x^{-1})) = 1 - \frac{s}{\xi_j x} + o(x^{-1}), \end{aligned}$$

where the last equality is based on expansion (A.2) in the Supplementary Material, and

$$\frac{\overline{F}_{Y_j}(x)}{\overline{F}_{\beta_j V_j}(x)} = \frac{\overline{F}_{V_j - S_j}(x/\beta_j)}{\overline{F}_{V_j}(x/\beta_j)} = 1 - \frac{\mathbb{E}[S_j]}{(\xi_j/\beta_j)x} + o(x^{-1}), \qquad x \to \infty. \tag{8.4}$$

Second, if $Z \sim \mathrm{GP}(\xi_j, \sigma_j)$ with $\sigma_j = \beta_j C^{\xi_j}\xi_j$, then $\overline{F}_Z(x) = \overline{H}_{\xi_j,\sigma_j}(x) = \left(1 + x/(\beta_j C^{\xi_j})\right)^{-1/\xi_j}$. Hence, using (8.1) for the numerator and (A.2) again for the denominator, we find, as $x \to \infty$,

$$\frac{\overline{F}_{\beta_j V_j}(x)}{\overline{F}_Z(x)} = \frac{C(x/\beta_j)^{-1/\xi_j}\left(1 - \frac{\beta_j D}{\xi_j x} + o(x^{-1})\right)}{\left(C^{-\xi_j}(x/\beta_j)\right)^{-1/\xi_j}(1 - \frac{\beta_j C^{\xi_j}}{\xi_j x} + o(x^{-1}))} = 1 - \left(\frac{D - C^{\xi_j}}{(\xi_j/\beta_j)x}\right) + o\left(x^{-1}\right). \tag{8.5}$$

Combining (8.4) and (8.5), we find, as $x \to \infty$,

$$\frac{\overline{F}_{Y_j}(x)}{\overline{F}_Z(x)} = \frac{\overline{F}_{Y_j}(x)}{\overline{F}_{\beta_j V_j}(x)} \frac{\overline{F}_{\beta_j V_j}(x)}{\overline{F}_Z(x)} = \left(1 - \frac{\mathbb{E}[S_j]}{(\xi_j/\beta_j)x} + o(x^{-1})\right)\left(1 - \left(\frac{D - C^{\xi_j}}{(\xi_j/\beta_j)x}\right) + o(x^{-1})\right).$$

□

8.2 Connection with the multivariate GP distribution

Proof of Lemma 3.3. For $\beta_j = \sigma_j(\alpha + \alpha_j)$, we have

$$\frac{G + G_j}{\sigma_j(\alpha + \alpha_j)} = \frac{G + G_j}{\beta_j} \sim \Gamma(\alpha + \alpha_j, \beta_j),$$

with moment-generating function

$$M(t) = \left(1 - \frac{t}{\sigma_j(\alpha + \alpha_j)}\right)^{-(\alpha+\alpha_j)}, \quad t < \sigma_j(\alpha + \alpha_j).$$

As $(\alpha + \alpha_j) \to \infty$, $M(t) \to \exp(t/\sigma_j)$, which corresponds to a degenerate random variable equal to $1/\sigma_j$. Hence, $\frac{G+G_j}{\beta_j} \xrightarrow{p} 1/\sigma_j$ for $j = 1, 2$. By the continuous mapping theorem,

$$\beta_j \frac{E}{G + G_j} \xrightarrow{p} \sigma_j E, \qquad j = 1, 2. \tag{8.6}$$

Combining (8.6) with the assumption that $\beta_j S_j^{(\beta)} \xrightarrow{p} S_j^0$ and using again the continuous mapping theorem, we obtain

$$\beta_j\Big(\frac{E}{G + G_j} - S_j^{(\beta)}\Big) \xrightarrow{p} \sigma_j E - S_j^0, \qquad j = 1, 2.$$

Joint convergence in probability follows immediately, yielding the stated limit for (Y_1, Y_2). □

8.3 Tail dependence coefficients

We now derive the residual tail dependence coefficient η and the tail dependence coefficient χ. Without loss of generality, we take $\beta_1 = \beta_2 = 1$ for notational simplicity. As before, we write $\boldsymbol{Y} = \boldsymbol{V} - \boldsymbol{S}$ for $\boldsymbol{S} = (S_1, S_2)^T$ and $\boldsymbol{V} = (V_1, V_2)^T$ with

$$V_j = \frac{wE + (1 - w)E_j}{G + G_j}, \qquad j = 1, 2. \tag{8.7}$$

In the lemmas below, we start by deriving the joint distribution function of $\boldsymbol{V}$ and its asymptotic expansion. We then show that $\eta_{\boldsymbol{Y}} = \eta_{\boldsymbol{V}}$ and $\chi_{\boldsymbol{Y}} = \chi_{\boldsymbol{V}}$ and finally state the proof of Proposition 3.4.

Lemma 8.3 (Joint distribution of $\boldsymbol{V}$)**.** *Let $x_1, x_2 \geq 0$ and $w \in (0, 1) \setminus \{\frac{1}{2}\}$. Consider the vector $\boldsymbol{V} = (V_1, V_2)$ defined in* (8.7)*. Then the joint survival function admits the decomposition*

$$\mathbb{P}(V_1 > x_1, V_2 > x_2) = A_1(x_1, x_2) + A_2(x_1, x_2) + A_{12}(x_1, x_2),$$

where

$$A_{12}(x_1, x_2) = \left(1 + \frac{x_1 + x_2}{1 - w}\right)^{-\alpha} \prod_{j=1}^{2} \left(1 + \frac{x_j}{1 - w}\right)^{-\alpha_j},$$

and, for $j = 1, 2$,

$$A_j(x_1, x_2) = \frac{w(1-w)^{\alpha_{3-j}}}{1 - 2w} \left(1 + \frac{x_j}{w}\right)^{-(\alpha + \alpha_j)} x_{3-j}^{-\alpha_{3-j}} \, \mathbb{E}[\mathcal{I}_j(x_1, x_2)],$$

with

$$\mathcal{I}_1(x_1, x_2) = \int_0^{H_1} \left(\exp\left(\frac{1 - 2w}{w} \min\{\tilde{D}_1,\, H_1 - y\}\right) - 1\right) \frac{y^{\alpha_2 - 1}}{\Gamma(\alpha_2 + 1)} \exp\left(-\frac{1 - w}{x_2} y\right) dy,$$

where

$$\tilde{D}_1 = \frac{x_1 w}{(1 - w)(x_1 + w)}(G + G_1), \qquad H_1 = \tilde{D}_1 + E_2 - \frac{x_2 w}{(1 - w)(x_1 + w)} G,$$

and $\mathcal{I}_2$ *defined similarly as* $\mathcal{I}_1$ *by swapping indices* 1 *and* 2.

Lemma 8.4 (Joint tail distribution of $\boldsymbol{V}$)**.** *We have, for a constant* $C > 0$,

$$\mathbb{P}\left(F_{V_1}(V_1) > q, F_{V_2}(V_2) > q\right) \sim C\,(1 - q)^{1 + \frac{\max(\alpha_1, \alpha_2)}{\alpha + \max(\alpha_1, \alpha_2)}}, \qquad q \uparrow 1.$$

Lemma 8.5 (Shift–invariance)**.** *We have*

$$\chi_{\boldsymbol{Y}} = \chi_{\boldsymbol{V}}, \qquad \eta_{\boldsymbol{Y}} = \eta_{\boldsymbol{V}}.$$

Proof of Proposition 3.4. We compute $\eta_{\boldsymbol{V}}$ and $\chi_{\boldsymbol{V}}$ which by Lemma 8.5 give $\eta_{\boldsymbol{Y}}$ and $\chi_{\boldsymbol{Y}}$. Also, suppose that $w \in [0, 1] \setminus \{\frac{1}{2}\}$; the case $w = \frac{1}{2}$ can be obtained by applying l'Hôpital's rule.

(i) Lemma 8.4 shows that the joint tail of $\boldsymbol{V}$ is of Ledford–Tawn form $L(1 - q)(1 - q)^{1/\eta}$ with L slowly varying at 0 and

$$\frac{1}{\eta} = 1 + \frac{\max(\alpha_1, \alpha_2)}{\alpha + \max(\alpha_1, \alpha_2)} = \frac{\alpha + 2\max(\alpha_1, \alpha_2)}{\alpha + \max(\alpha_1, \alpha_2)}, \qquad \eta = \frac{\alpha + \max(\alpha_1, \alpha_2)}{\alpha + 2\max(\alpha_1, \alpha_2)}.$$

(ii) Suppose that $\alpha_1 = \alpha_2 = 0$, in which case V_1 and V_2 have the same distribution. We can isolate E in the joint survival function of $\boldsymbol{V}$ (Lemma 8.3) and apply elementary algebra to obtain

$$\mathbb{P}(V_j > x) = \mathbb{E}\left[\exp\left(-\frac{x}{w} G + \frac{1 - w}{w} E_j\right) \mathbb{1}\{xG > (1 - w)E_j\}\right] + \mathbb{P}\left(xG < (1 - w)E_j\right),$$

and

$$\mathbb{P}(V_1 > x, V_2 > x) = \mathbb{E}\left[\exp\left(-\frac{x}{w} G + \frac{1 - w}{w} \min(E_1, E_2)\right) \mathbb{1}\{xG > (1 - w)\min(E_1, E_2)\}\right]$$

$$+ \mathbb{P}\left(xG < (1-w)\min(E_1, E_2)\right). \tag{8.8}$$

Since $E_j \sim \mathrm{Exp}(1)$ and $\min(E_1, E_2) \sim \mathrm{Exp}(2)$, both expressions share the same structure. One can then compute:

$$\mathbb{P}(V_1 > x) = p_1(x), \qquad \mathbb{P}(V_1 > x, V_2 > x) = p_2(x),$$

with

$$p_s(x) = \frac{sw}{(s+1)w-1}\left(1+\frac{x}{w}\right)^{-\alpha} - \frac{1-w}{(s+1)w-1}\left(1+\frac{s\,x}{1-w}\right)^{-\alpha}$$

for $s = 1, 2$. This expression holds for $w \neq (s+1)^{-1}$; if not, we take the continuous extension. Finally, by the definition of χ, taking the ratio $p_2(x)/p_1(x)$ and letting $x \to \infty$ yields the stated closed-form expression.

(iii) Assume without loss of generality that $\alpha_2 \geq \alpha_1$, i.e., $\xi_1 \geq \xi_2$. Using $\xi_j = 1/(\alpha + \alpha_j)$ and $\eta = (\alpha + \alpha_2)/(\alpha + 2\alpha_2)$, we express

$$\alpha = \frac{2\eta - 1}{\eta\,\xi_2}, \qquad \alpha_1 = \frac{1}{\xi_1} - \frac{2\eta - 1}{\eta\,\xi_2}.$$

The constraint on the shape parameter $\alpha_1 \geq 0$ leads to the inequality

$$\eta \leq \frac{\xi_1}{2\xi_1 - \xi_2}.$$

On the other hand, since $\eta = (\alpha + \alpha_2)/(\alpha + 2\alpha_2) \geq 1/2$ for all $\alpha, \alpha_2 \geq 0$, we obtain

$$\eta \in \left[\frac{1}{2},\ \frac{\xi_1}{2\xi_1 - \xi_2}\right].$$

Repeating the same reasoning with $\alpha_1 \geq \alpha_2$ (i.e., $\xi_2 \geq \xi_1$) yields the symmetric bound with indices swapped, completing the proof.

□

Acknowledgements

This work was supported by the Fonds de la Recherche Scientifique–FNRS under Grant No F.4036.24.

Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

Part of Naveau's research work was supported by the Agence Nationale de la Recherche via the SICIM and SHARE PEPR Maths-Vives project (France 2030 ANR-24-EXMA-0008), EXSTA grant (ANR-23-CE40-0009-01), PORC-EPIC, the PEPR TRACCS program (PC4 EXTENDING, ANR-22-EXTR-0005), and the PEPR IRIMONT (France 2030 ANR-22-EXIR-0003). He has also benefited from the Geolearning research chair.

Supplementary material

A.1 Additional proofs

Proof of Lemma 8.1.

1. For $w \in (0,1)$, the numerator $N = wE + (1-w)E_j$ of V_j is a sum of two independent exponential variables with different scale parameters. Its distribution is hypo-exponential with cdf
$$F_N(u) = 1 - \frac{(1-w)e^{-u/w} - we^{-u/(1-w)}}{2w-1}.$$
Let f_{G+G_j} denote the density of $G + G_j \sim \Gamma(1/\xi_j, 1)$. Straightforward calculations give the cdf of $V_j = N/(G+G_j)$, for $x \geq 0$,
$$\begin{aligned} F_{V_j}(x) = \mathbb{P}(V_j \leq x) &= \int_0^\infty F_N(xt) \cdot f_{G+G_j}(t)\ \mathrm{d}t \\ &= 1 - \frac{w(1+\frac{x}{w})^{-1/\xi_j} - (1-w)(1+\frac{x}{1-w})^{-1/\xi_j}}{2w-1}. \end{aligned} \tag{A.1}$$
Differentiating (A.1) with respect to x yields the stated density. The cases $w \in \{0,1\}$ can be obtained directly from Lemma 2.1. The expression for $w = \frac{1}{2}$ can be obtained by applying l'Hôpital's rule to the above formula.

2. The mean and variance follow from the facts that, writing $V_j = NM$ for $M = (G+G_j)^{-1}$,
$$\begin{aligned} \mathbb{E}[N] &= 1, & \mathbb{E}[N^2] &= 1 + w^2 + (1-w)^2, \\ \mathbb{E}[M] &= \frac{1}{\xi_j^{-1} - 1}, & \mathbb{E}[M^2] &= \frac{1}{(\xi_j^{-1}-1)(\xi_j^{-1}-2)}. \end{aligned}$$

□

Proof of Lemma 8.2. For $\lambda > 0$ and $\alpha > 0$, we have
$$\frac{(1+\lambda x)^{-\alpha}}{(\lambda x)^{-\alpha}} = 1 - \alpha(\lambda x)^{-1} + o(x^{-1}), \qquad \text{as } x \to \infty. \tag{A.2}$$
Expanding $(1+\lambda x)^{-1/\xi_j}$ in (A.1) according to (A.2) for $\lambda = 1/w$ and $\lambda = 1/(1-w)$ and collecting constants, we get
$$\overline{F}_{V_j}(x) = Cx^{-1/\xi_j}\left(1 - \frac{D}{\xi_j x} + o(x^{-1})\right), \qquad \text{as } x \to \infty,$$
for $w \in [0,1] \setminus \{\frac{1}{2}\}$, where C and D are given in (8.2). When $w = \frac{1}{2}$, l'Hôpital's rule gives
$$C = 2^{-1/\xi_j}(\xi_j^{-1} + 1) \quad \text{and} \quad D = \frac{1+2\xi_j}{2+2\xi_j}.$$

□

Proof of Lemma 8.3. We write

$$\mathbb{P}\left[V_1 > x_1, V_2 > x_2\right] = \mathbb{P}\left[E > \max_{j=1,2}\left\{\frac{x_j(G+G_j)-(1-w)E_j}{w}\right\}\right].$$

Conditioning on all variables except E and defining $M_j := x_j(G+G_j)-(1-w)E_j$, we obtain

$$\mathbb{P}(V_1 > x_1, V_2 > x_2) = \mathbb{E}\left[e^{-(M_1 \vee M_2)/w}\mathbb{1}\{M_1 \vee M_2 > 0\}\right] + \mathbb{P}\{M_1 \vee M_2 < 0\}.$$

Decomposing on $\{M_1 > M_2\}$ and $\{M_2 \geq M_1\}$ yields

$$\mathbb{P}(V_1 > x_1, V_2 > x_2) = A_1 + A_2 + A_{12},$$

with $A_{12} = \mathbb{P}\{M_1 \vee M_2 < 0\}$,

$$A_1 = \mathbb{E}\left[e^{-M_1/w}\mathbb{1}\{M_1 > M_2 \vee 0\}\right], \text{ and } A_2 = \mathbb{E}\left[e^{-M_2/w}\mathbb{1}\{M_2 > M_1 \vee 0\}\right].$$

Term A_{12}: Since $M_1 \vee M_2 < 0$ iff $(1-w)E_j > x_j(G+G_j)$ for $j = 1, 2$, conditioning on (G, G_1, G_2) and using independence of exponentials gives

$$A_{12} = \mathbb{E}\left[e^{-\frac{x_1+x_2}{1-w}G}e^{-\frac{x_1}{1-w}G_1}e^{-\frac{x_2}{1-w}G_2}\right].$$

Applying the Laplace transform of Gamma variables yields

$$A_{12} = \left(1+\frac{x_1+x_2}{1-w}\right)^{-\alpha}\left(1+\frac{x_1}{1-w}\right)^{-\alpha_1}\left(1+\frac{x_2}{1-w}\right)^{-\alpha_2}.$$

Term A_1: Using $M_j = x_j(G+G_j)-(1-w)E_j$, we write

$$A_1 = \mathbb{E}\left[e^{-\frac{x_1}{w}(G+G_1)}e^{\frac{1-w}{w}E_1}\mathbb{1}\{M_1 > M_2,\ M_1 > 0\}\right].$$

The constraints are equivalent to $E_1 \leq B$, where

$$B = \min\left\{\frac{x_1}{1-w}(G+G_1),\ \frac{x_1}{1-w}(G+G_1)+E_2-\frac{x_2}{1-w}(G+G_2)\right\}.$$

Conditioning on all variables except E_1 and evaluating the exponential integral (valid for $w \neq \frac{1}{2}$) yields

$$A_1 = \frac{w}{1-2w}\,\mathbb{E}\left[e^{-\frac{x_1}{w}(G+G_1)}\left(e^{\frac{1-2w}{w}B_+}-1\right)\right], \qquad B_+ = \max(B, 0).$$

Introduce the rescaled variables $\tilde{G} = \frac{w}{x_1+w}G$ and $\tilde{G}_1 = \frac{w}{x_1+w}G_1$. Using scaling invariance of Gamma distributions,

$$A_1 = \frac{w}{1-2w}\left(1+\frac{x_1}{w}\right)^{-(\alpha+\alpha_1)}\mathbb{E}\left[e^{\frac{1-2w}{w}\tilde{B}_+}-1\right],$$

where $\tilde{B}$ denotes the corresponding rescaled bound. Since the integrand vanishes on $\{\tilde{B} \leq 0\}$,

$$\mathbb{E}\left[e^{\frac{1-2w}{w}\tilde{B}_+}-1\right] = \mathbb{E}\left[(e^{\frac{1-2w}{w}\tilde{B}_+}-1)\mathbb{1}\{\tilde{B} > 0\}\right].$$

Conditioning on (G, G_1, E_2) and integrating out G_2 gives

$$A_1 = \frac{w(1-w)^{\alpha_2}}{1-2w}\left(1+\frac{x_1}{w}\right)^{-(\alpha+\alpha_1)} x_2^{-\alpha_2}\,\mathbb{E}[\mathcal{I}_1], \tag{A.3}$$

where $\mathcal{I}_1$ is the integral term

$$\mathcal{I}_1 = \int_0^{H_1}\left(\exp\left(\frac{1-2w}{w}\min\{\tilde{D}_1,\, H_1-y\}\right)-1\right)\frac{y^{\alpha_2-1}}{\Gamma(\alpha_2+1)}\exp\left(-\frac{1-w}{x_2}y\right)\,dy,$$

with

$$\tilde{D}_1 = \frac{x_1 w}{(1-w)(x_1+w)}(G+G_1), \qquad H_1 = \tilde{D}_1 + E_2 - \frac{x_2 w}{(1-w)(x_1+w)}G.$$

Term A_2: The expression for A_2 follows by symmetry after exchanging indices 1 and 2.

Combining A_1, A_2 and A_{12} yields the stated survival function. □

Proof of Lemma 8.4. Let $(x_1, x_2) = (u_1(q), u_2(q))$ with $q \uparrow 1$. Since V_j has a GP tail (see the proof of Proposition 3.2 (iii)), it is regularly varying with index $\alpha+\alpha_j$, hence for some constants $C_j > 0$,

$$u_j(q) := F_{V_j}^{-1}(q) \sim C_j\,(1-q)^{-1/(\alpha+\alpha_j)}, \qquad q \uparrow 1,$$

and therefore $u_j(q)^{-\beta} \sim C_j^{-\beta}\,(1-q)^{\beta/(\alpha+\alpha_j)}$ for any $\beta > 0$. In particular, assume without loss of generality that $\alpha_1 < \alpha_2$, in which case $u_2(q)/u_1(q) \to 0$. We analyse successively the three contributions A_1, A_2, A_{12} from

$$\mathbb{P}(V_1 > x_1, V_2 > x_2) = A_1(x_1,x_2) + A_2(x_1,x_2) + A_{12}(x_1,x_2),$$

obtained in Lemma 8.3.

Contribution of A_{12}: Using $(1+x/a)^{-k} \sim (x/a)^{-k}$ and $x_1 + x_2 \sim x_1$, we obtain

$$A_{12}(x_1,x_2) \sim (1-w)^{\alpha+\alpha_1+\alpha_2}\,x_1^{-(\alpha+\alpha_1)}x_2^{-\alpha_2} \sim c_{12}\,(1-q)^{\,1+\frac{\alpha_2}{\alpha+\alpha_2}},$$

for some constant $c_{12} \in (0,\infty)$.

Contribution of A_1: From (A.3) and $(1+x_1/w)^{-(\alpha+\alpha_1)} \sim w^{\alpha+\alpha_1}x_1^{-(\alpha+\alpha_1)}$, we get

$$A_1(x_1,x_2) \sim K_1\,x_1^{-(\alpha+\alpha_1)}x_2^{-\alpha_2}\,\mathbb{E}\left[\mathcal{I}_1(x_1,x_2)\right], \qquad K_1 := \frac{w(1-w)^{\alpha_2}}{1-2w}\,w^{\alpha+\alpha_1}.$$

In the regime $x_1, x_2 \to \infty$ with $x_2/x_1 \to 0$, $\mathcal{I}_1(x_1,x_2) \to \mathcal{I}_{1,\infty}$ a.s. and is dominated by an integrable envelope (Gamma moments), so dominated convergence yields $\mathbb{E}[\mathcal{I}_1(x_1,x_2)] \to \kappa_1 \in (0,\infty)$. Consequently,

$$A_1(x_1,x_2) \sim c_1\,x_1^{-(\alpha+\alpha_1)}x_2^{-\alpha_2} \sim c_1\,(1-q)^{\,1+\frac{\alpha_2}{\alpha+\alpha_2}},$$

for some $c_1 \in (0,\infty)$.

Contribution of A_2: By symmetry (swap indices 1 and 2 in (A.3)),

$$A_2(x_1, x_2) = \frac{w(1-w)^{\alpha_1}}{1-2w} \left(1 + \frac{x_2}{w}\right)^{-(\alpha+\alpha_2)} x_1^{-\alpha_1} \, \mathbb{E}[\mathcal{I}_2(x_1, x_2)].$$

In the regime $x_1, x_2 \to \infty$ with $x_2/x_1 \to 0$, the constraint defining $\mathcal{I}_2$ forces $G = O(x_2/x_1)$; with the scaling $G = (x_2/x_1)U$ and the small-argument expansion of the Gamma density, one obtains

$$\mathbb{E}[\mathcal{I}_2(x_1, x_2)] \sim \kappa_2 \left(\frac{x_2}{x_1}\right)^{\alpha}, \qquad \kappa_2 \in (0, \infty),$$

and therefore

$$A_2(x_1, x_2) \sim c_2 \, x_1^{-(\alpha+\alpha_2)} x_2^{-\alpha_1} \sim c_2 \, (1-q)^{1+\frac{\alpha_2}{\alpha+\alpha_2}},$$

for some $c_2 \in (0, \infty)$.

If $\alpha_1 > \alpha_2$, the same expansions hold with the roles of α_1 and α_2 reversed. Hence, we find

$$\bar{F}_{V_1,V_2}\left(u_1(q), u_2(q)\right) \sim (c_1 + c_2 + c_{12}) \, (1-q)^{1+\frac{\max(\alpha_1,\alpha_2)}{\alpha+\max(\alpha_1,\alpha_2)}}. \tag{A.4}$$

□

Proof of Lemma 8.5. Since V_j is regularly varying and $S_j \geq 0$, by dominated convergence, we have $\bar{F}_{Y_j}(x) = \bar{F}_{V_j}(x)(1 + o(1))$ as $x \to \infty$. By Resnick (2007, Proposition 2.6(vi)), this yields $w_j(q) := F_{Y_j}^{-1}(q) = F_{V_j}^{-1}(q)(1 + o(1))$ as $q \uparrow 1$. From the proof of Lemma 8.4, recall that $u_j(q) := F_{V_j}^{-1}(q) \sim C_j(1-q)^{-1/(\alpha+\alpha_j)}$. In particular, the proof of Lemma 8.4 gives

$$\bar{F}_{V_1,V_2}\left(u_1(q)\{1 + o(1)\}, u_2(q)\{1 + o(1)\}\right) \sim \bar{F}_{V_1,V_2}\left(u_1(q), u_2(q)\right) \tag{A.5}$$

Next, note that

$$\bar{F}_{Y_1,Y_2}(w_1(q), w_2(q)) = \mathbb{E}\left[\bar{F}_{V_1,V_2}(w_1(q) + S_1, \; w_2(q) + S_2)\right].$$

Define

$$R(q) := \frac{\bar{F}_{V_1,V_2}(w_1(q) + S_1, \; w_2(q) + S_2)}{\bar{F}_{V_1,V_2}(w_1(q), w_2(q))}.$$

Conditionally on (S_1, S_2), the shifts become deterministic constants. Since $w_j(q) \to \infty$, we have $S_j = o(w_j(q))$ a.s., so that equation A.5 yields $R(q) \to 1$ almost surely. Moreover, since $0 \leq R(q) \leq 1$, we have $\mathbb{E}[R(q)] \to 1$ by bounded convergence and

$$\bar{F}_{Y_1,Y_2}(w_1(q), w_2(q)) \sim \bar{F}_{V_1,V_2}(w_1(q), w_2(q)), \qquad q \uparrow 1.$$

Finally, using again equation A.5,

$$\bar{F}_{Y_1,Y_2}(F_{Y_1}^{-1}(q), F_{Y_2}^{-1}(q)) \sim \bar{F}_{V_1,V_2}\left(F_{V_1}^{-1}(q), F_{V_2}^{-1}(q)\right), \qquad q \uparrow 1.$$

so that dividing by $(1-q)$ and letting $q \uparrow 1$ gives $\chi_{\boldsymbol{Y}} = \chi_{\boldsymbol{V}}$. If $\chi_{\boldsymbol{V}} = 0$, then taking logarithms gives also $\eta_{\boldsymbol{Y}} = \eta_{\boldsymbol{V}}$. □

A.2 Simulation Study: additional results

This appendix provides additional simulation results that complement the analysis in Section 5. We report results under randomized parameter configurations, as well as results for the penalized estimator in the fixed-parameter setting.

A.2.1 Randomized parameters

We assess the global performance of the NBE under randomly generated parameter configurations. For each of the $K = 10^3$ replications, we draw the parameters

$$(\eta, \xi_1, \xi_2, \beta_1, \beta_2, \sigma_T, w)$$

independently from their prior distributions specified in Section 4. The values of η, ξ_1 and ξ_2 are then transformed into the model parameters $(\alpha, \alpha_1, \alpha_2)$ according to the construction described in Section 3.1, and a dataset of size $N = 1000$ is simulated from the resulting model. Each dataset is then provided to the NBE for estimation.

Figure A.14 compares true and estimated values for all parameters. The points cluster tightly around the diagonal, indicating accurate recovery of both the model parameters and the extremal dependence structure. Estimates of η exhibit higher variability for lower values, and stabilize as η increases. This pattern is expected since weaker dependence (smaller η) corresponds to lighter joint tails, providing less information for reliable estimation. In contrast, when η approaches 1, the stronger dependence structure yields more stable and precise estimates across replications.

A.2.2 Simulation study: Penalised NBE (fixed-parameter setting)

In Figures A.15 and A.16 and Table A.5, we report additional results for the fixed-parameter simulation setting using the penalized NBE. The experimental design is identical to that described in Section 5, with the only difference being the use of the penalized loss function.

Overall, the penalised NBE introduces small biases, leading to a moderate reduction in coverage for several parameters, in particular (η, ξ_1, ξ_2) and σ_T. This effect is more pronounced in the weakest dependence setting (Config. 3). In contrast, the estimation of $\chi(q)$ remains stable across most quantiles, with coverage close to the nominal level, except at the most extreme levels (e.g., $q = 0.99$ in Config. 1). These results highlight a trade-off induced by penalisation: improved structural coherence of the estimates at the cost of a modest loss in estimation quality due to the introduction of small biases.

A.3 Application to Belgian Weekly Rainfall: additional results

A.3.1 Parameter stability and QQ-plots

We present some supplementary figures for the Belgian rainfall application.

Figure A.17 shows the spatial domain used in the analysis, corresponding to all ERA5 grid points in the rectangle $[0^\circ, 7^\circ]\text{E} \times [47^\circ, 52^\circ]\text{N}$. The central reference grid point covering Brussels is marked, along with neighbouring points. Coordinates correspond to the centres of the ERA5 grid cells.

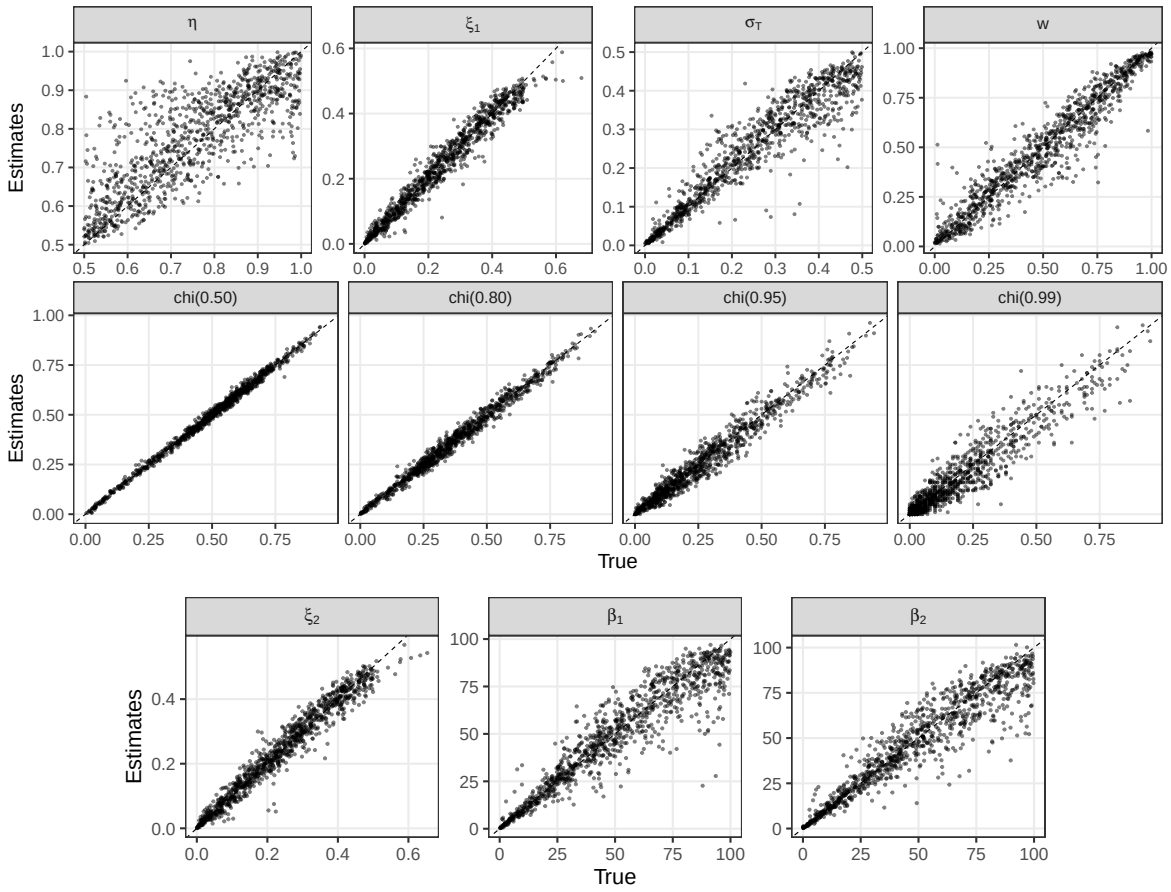


Figure A.14: True vs estimated values for η, ξ_1, σ_T, w (top), $\chi(q)$ with $q \in \{0.5, 0.8, 0.95, 0.99\}$ (middle), and ξ_2, β_1, β_2 (bottom) in the randomized parameter setting with $n = 1000$. Each point corresponds to one of the $K = 1000$ Monte Carlo replications.

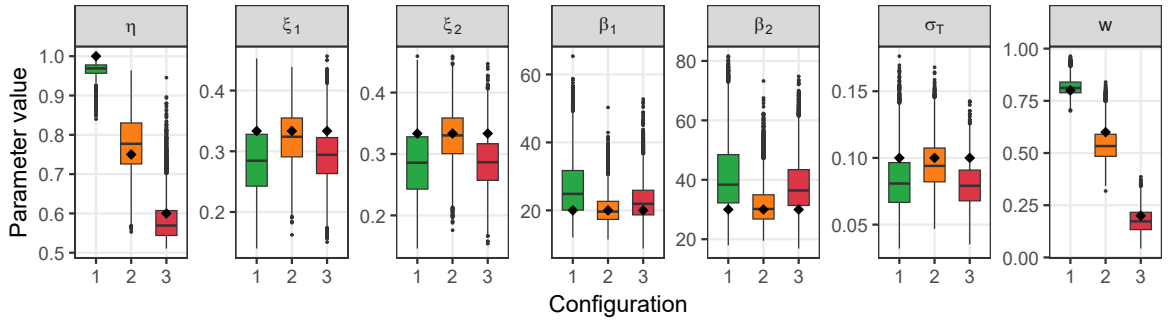


Figure A.15: Boxplots of parameter estimates $(\eta, \xi_1, \xi_2, \beta_1, \beta_2, \sigma_T, w)$ obtained with the penalised NBE under three dependence configurations: $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, and $\boldsymbol{\theta}^{(3)}$ (see (5.1)), based on $K = 1000$ samples of size $N = 1000$. Diamond markers indicate true values.

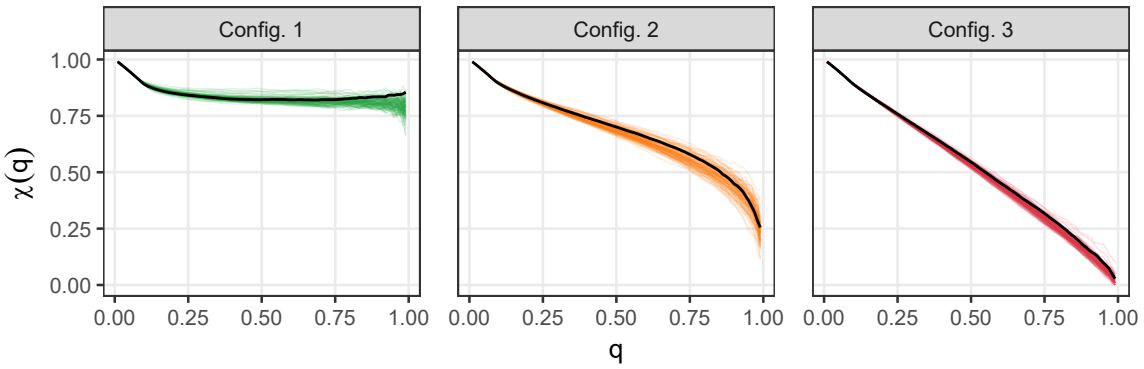


Figure A.16: Estimated $\chi(q)$ curves obtained with the penalised NBE, where $\chi(q)$ is estimated using (3.5). Solid lines correspond to reference curves empirically estimated from samples of size 10^5 generated under the true parameters. Coloured lines show 100 estimated curves based on samples of size 10^4 generated under fitted parameter values.

	Config. 1			Config. 2			Config. 3		
$\hat{\theta}_j$	True	Cov.	Width	True	Cov.	Width	True	Cov.	Width
$\hat{\eta}$	1.00	0.00	0.05	0.75	0.90	0.20	0.60	0.86	0.14
$\hat{\xi}_1$	0.33	0.73	0.15	1.50	0.89	0.11	0.33	0.68	0.12
$\hat{\xi}_2$	0.33	0.72	0.15	1.50	0.90	0.11	0.33	0.69	0.12
$\hat{\beta}_1$	20.00	0.79	20.9	20.00	0.96	10.8	20.00	0.87	15.2
$\hat{\beta}_2$	30.00	0.63	28.8	30.00	0.94	17.1	30.00	0.76	23.5
$\hat{\sigma}_T$	0.10	0.78	0.06	0.10	0.89	0.04	0.10	0.59	0.04
$\hat{w}$	0.80	0.86	0.10	0.60	0.78	0.21	0.20	0.92	0.15
$\chi(0.50)$	0.82	1.00	0.05	0.70	1.00	0.05	0.55	1.00	0.04
$\chi(0.70)$	0.82	1.00	0.06	0.60	1.00	0.08	0.37	1.00	0.06
$\chi(0.80)$	0.83	1.00	0.06	0.54	1.00	0.11	0.27	1.00	0.06
$\chi(0.90)$	0.84	1.00	0.07	0.45	1.00	0.13	0.16	1.00	0.07
$\chi(0.95)$	0.84	0.77	0.09	0.37	1.00	0.16	0.10	1.00	0.07
$\chi(0.99)$	0.85	0.51	0.16	0.23	1.00	0.22	0.03	1.00	0.06

Table A.5: Empirical coverage probability (Cov.) and mean confidence interval width (Width) for model parameters and tail dependence curves $\chi(q)$ obtained with the penalised NBE under three fixed configurations defined in (5.1) ($N = 1000$, $K = 1000$, $B = 200$ bootstrap resamples per estimate), together with the true values.

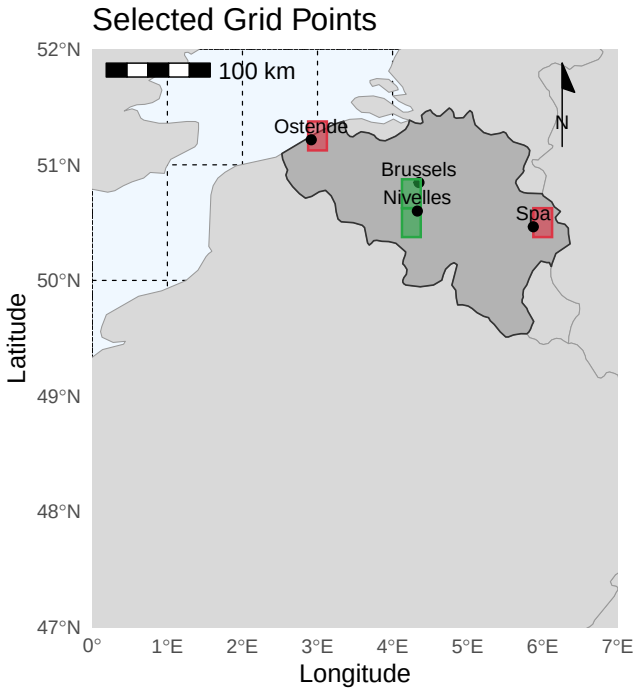


Figure A.17: Reference grid point (black point, Brussels) and surrounding ERA5 grid points within the analysis domain $[0^\circ, 7^\circ]$E $\times$ $[47^\circ, 52^\circ]$N. Coordinates indicate grid cell centres.

Next, Figure A.18 reports parameter estimates as functions of the marginal threshold, for all pairs involving the Brussels reference grid point. These plots provide a visual check of threshold stability for both marginal and dependence parameters.

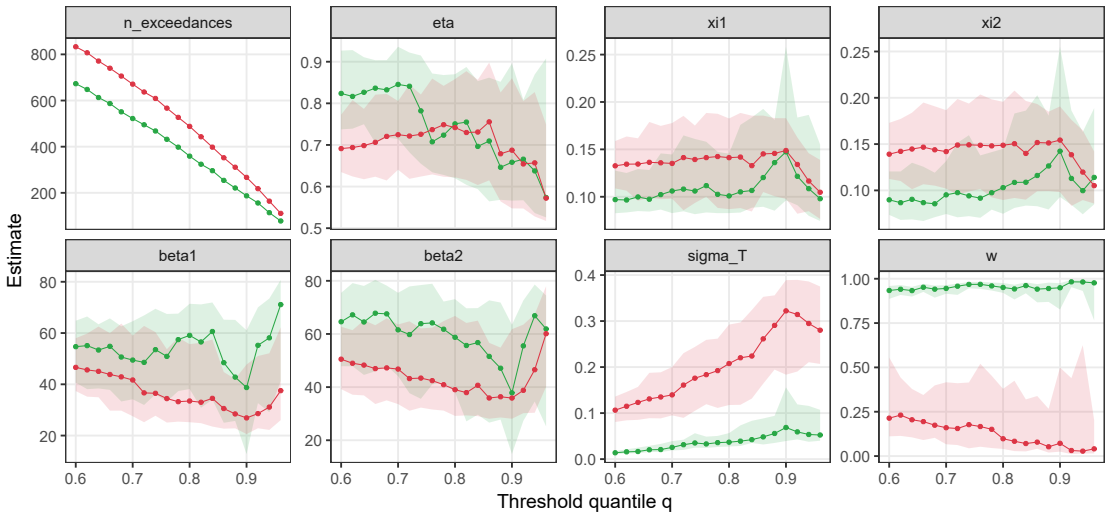


Figure A.18: Estimated parameter values of the sBGP model (Definition 3.1) as a function of the threshold quantile q for Brussels–Nivelles and Ostend–Spa. Green corresponds to the Brussels–Nivelles pair and red to the Ostend–Spa pair. Shaded areas denote the 95% bootstrap confidence intervals.

Finally, Figure A.19 complements Figure 8 in the main paper, illustrating several additional parameters as spatial fields.

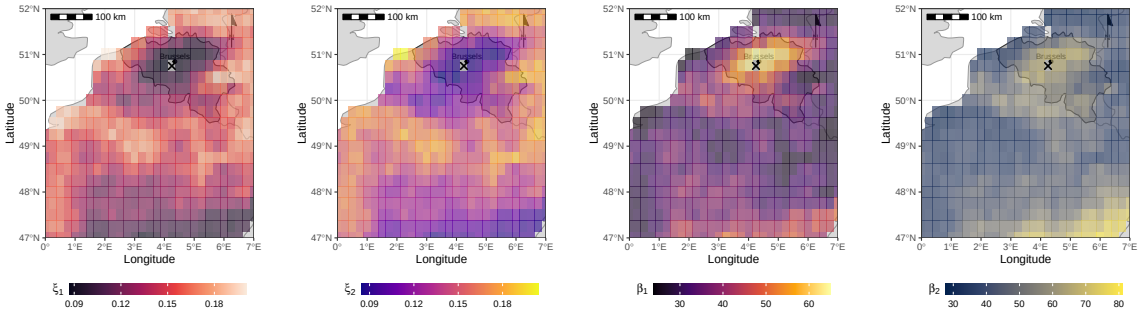


Figure A.19: Spatial fields of estimated parameters of the sBGP model (Definition 3.1) for bivariate precipitation extremes between Brussels (reference location, black cross) and each grid point.

A.3.2 Non-Penalised NBE

Figure A.20 shows the spatial fields of estimated parameters for the Belgian rainfall application obtained using the NBE trained without penalisation. These results are provided for comparison with the penalised version discussed in the main text, and illustrate the impact of the loss specification on both parameter estimates and goodness-of-fit diagnostics.

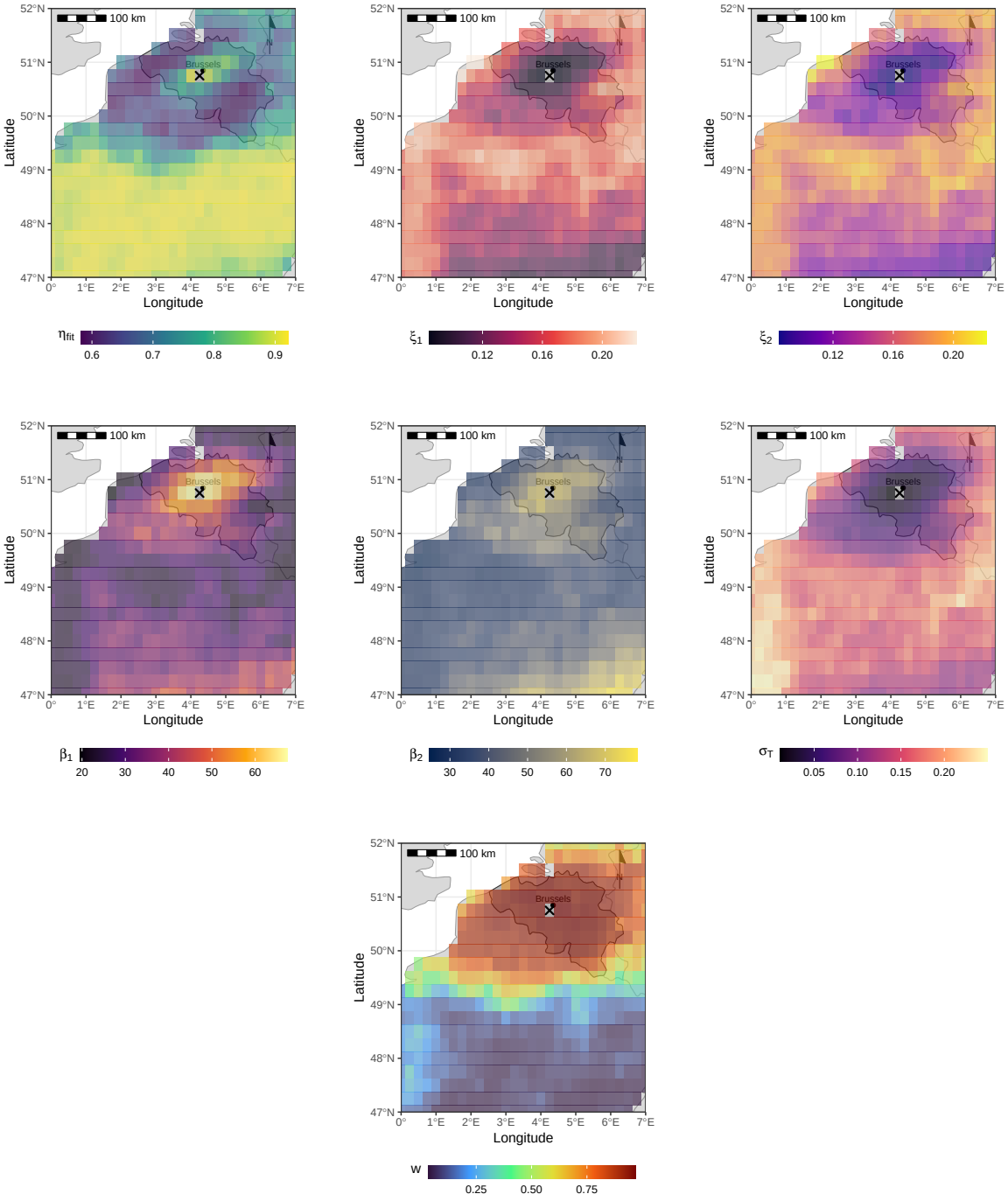


Figure A.20: Spatial fields of estimated dependence parameters of the sBGP model (Definition 3.1) for bivariate precipitation extremes between Brussels (reference location, black cross) and each grid point, obtained using the non-penalised NBE.

	Brussels–Nivelles	Ostend–Spa
η	0.77 (0.67, 0.85)	0.92 (0.84, 0.94)
ξ_1	0.09 (0.06, 0.14)	0.08 (0.07, 0.11)
ξ_2	0.08 (0.06, 0.13)	0.08 (0.07, 0.11)
β_1	47.11 (28.19, 65.55)	66.45 (49.17, 74.91)
β_2	60.46 (35.26, 74.40)	71.79 (52.62, 76.53)
σ_T	0.02 (0.02, 0.04)	0.07 (0.06, 0.10)
w	0.94 (0.89, 0.97)	0.10 (0.04, 0.29)

Table A.6: Parameter estimates (with 95% confidence intervals) for the sBGP model (Definition 3.1) fitted to the two representative pairs at the 70th-percentile threshold using the non-penalised NBE.

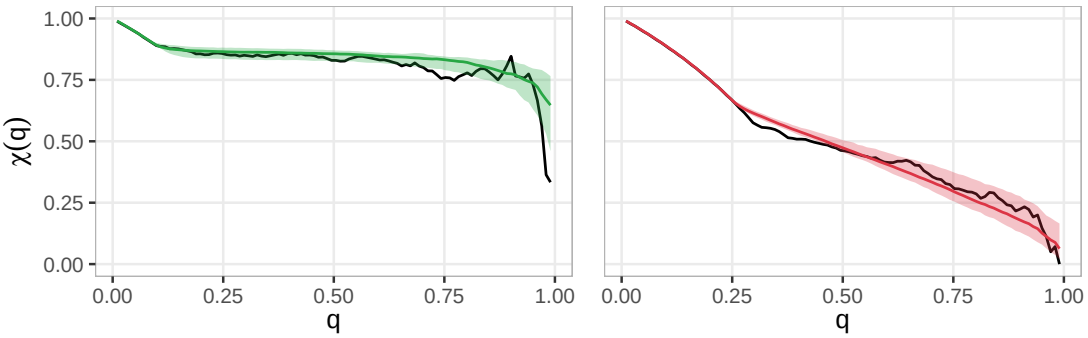


Figure A.21: Estimated $\chi_{\tilde{\mathcal{Z}}_{0.7}}(q)$ curves for Brussels–Nivelles and Ostend–Spa using the non-penalised NBE, where $\chi(q)$ is estimated using (3.5). Quantiles q are defined relative to the exceedance subset $\tilde{\mathcal{Z}}_{0.7}$ (i.e., within the exceedance region, $u_j = F_j^{-1}(0.7)$). The black line shows the empirical $\chi_{\tilde{\mathcal{Z}}_{0.7}}(q)$ from observed exceedances; the green and red lines are the model estimates; the shaded area denotes the 95% bootstrap confidence region.

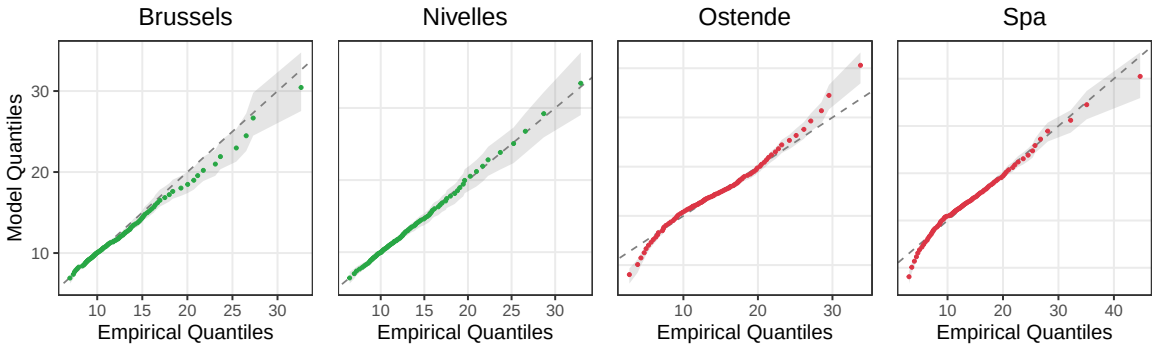


Figure A.22: QQ-plots comparing empirical marginal exceedances to fitted marginal distributions for Brussels–Nivelles and Ostend–Spa under the sBGP model (Definition 3.1) using the non-penalised NBE. Each panel corresponds to one margin of the respective pair. All plots are based on exceedances above the marginal threshold ($u_j = F_j^{-1}(0.7)$), with dashed vertical lines marking the threshold values. Theoretical quantiles are shown on the original data scale, obtained by adding the threshold to the modeled exceedances.

REFERENCES

André, L. M., J. L. Wadsworth, and R. Huser (2025). Neural Bayes estimation and selection of complex bivariate extremal dependence models. *Extremes*, 1–40.

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*(2), 697–725.

Bacro, J.-N., C. Gaetan, T. Opitz, and G. Toulemonde (2025). Multivariate peaks-over-threshold with latent variable representations of generalized Pareto vectors. *Extremes 28*(2), 273–300.

Balkema, A. A. and L. de Haan (1974). Residual life time at great age. *The Annals of Probability 2*(5), 792–804.

Beirlant, J., Y. Goegebeur, J. Segers, and J. Teugels (2004). *Statistics of extremes: theory and applications*. Wiley.

Bopp, G. P. and B. A. Shaby (2017). An exponential–gamma mixture model for extreme Santa Ana winds. *Environmetrics 28*(8), e2476.

Bousquet, N. and P. Bernardara (2021). *Extreme value theory with applications to natural hazards*. Springer.

Breiman, L. (1965). On some limit theorems similar to the arc-sin law. *Theory of Probability & Its Applications 10*(2), 323–331.

Coles, S., J. Heffernan, and J. Tawn (1999). Dependence Measures for Extreme Value Analyses. *Extremes 2*(4), 339–365.

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B: Statistical Methodology 52*(3), 393–425.

Dell'Oro, L. and C. Gaetan (2025). Flexible space–time models for extreme data. *Spatial Statistics 68*, 100916.

Ding, J., X. Li, X. Kang, and V. N. Gudivada (2019). A Case Study of the Augmentation and Evaluation of Training Data for Deep Learning. *J. Data and Information Quality 11*(4), 20:1–20:22.

Einmahl, J. H., A. Kiriliouk, and J. Segers (2018). A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes 21*, 205–233.

Einmahl, J. H. J., A. Krajina, and J. Segers (2012). An M-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics 40*(3), 1764–1793.

Embrechts, P., C. Klüppelberg, and T. Mikosch (2013). *Modelling extremal events: for insurance and finance*, Volume 33. Springer Science & Business Media.

Engelke, S., T. Opitz, and J. Wadsworth (2019). Extremal dependence of random scale constructions. *Extremes 22*(4), 623–666.

Hazra, A., R. Huser, and D. Bolin (2025). Efficient modeling of spatial extremes over large geographical domains. *Journal of Computational and Graphical Statistics 34*(3), 795–811.

He, T., S. Yu, Z. Wang, J. Li, and Z. Chen (2019). From data quality to model quality: An exploratory study on deep learning. In *Proceedings of the 11th Asia-Pacific Symposium on Internetware*, pp. 1–6.

Heffernan, J. E. and J. A. Tawn (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology 66*(3), 497–546.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al. (2020). The ERA5 global reanalysis. *Quarterly journal of the royal meteorological society 146*(730), 1999–2049.

Huser, R., T. Opitz, and E. Thibaud (2017). Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spatial Statistics 21*, 166–186.

Huser, R., T. Opitz, and J. L. Wadsworth (2025). Modeling of spatial extremes in environmental data science: Time to move away from max-stable processes. *Environmental Data Science 4*, e3.

Huser, R. and J. L. Wadsworth (2019). Modeling Spatial Processes with Unknown Extremal Dependence Class. *Journal of the American Statistical Association 114*(525), 434–444.

Kakampakou, L., E. S. Simpson, and J. L. Wadsworth (2024). Spatial extremal modelling: A case study on the interplay between margins and dependence. *Stat 13*(4), e70021.

Kiriliouk, A., H. Rootzén, J. Segers, and J. L. Wadsworth (2019). Peaks Over Thresholds Modeling With Multivariate Generalized Pareto Distributions. *Technometrics 61*(1), 123–135.

Kiriliouk, A. and C. Zhou (2022). Estimating probabilities of multivariate failure sets based on pairwise tail dependence coefficients. Available at `https://arxiv.org/abs/2210.12618`.

Krupskii, P., R. Huser, and M. G. Genton (2018). Factor copula models for replicated spatial data. *Journal of the American Statistical Association 113*(521), 467–479.

Ledford, A. W. and J. A. Tawn (1996). Statistics for Near Independence in Multivariate Extreme Values. *Biometrika 83*(1), 169–187.

Morris, S. A., B. J. Reich, E. Thibaud, and D. Cooley (2017). A space-time skew-t model for threshold exceedances. *Biometrics 73*(3), 749–758.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics 3*(1), 119–131.

Reiss, R.-D., M. Thomas, and R. Reiss (1997). *Statistical analysis of extreme values*, Volume 2. Springer.

Resnick, S. (2007). *Heavy-Tail Phenomena*. Springer.

Richards, J., M. Sainsbury-Dale, A. Zammit-Mangion, and R. Huser (2024). Neural Bayes estimators for censored inference with peaks-over-threshold models. *Journal of Machine Learning Research 25*(390), 1–49.

Rödder, A., M. Hentschel, and S. Engelke (2025). Theoretical guarantees for neural estimators in parametric statistics. Available at `https://arxiv.org/abs/2506.18508`.

Rootzén, H., J. Segers, and J. L. Wadsworth (2018). Multivariate generalized Pareto distributions: parametrizations, representations and properties. *Journal of Multivariate Analysis 165*(1), 117–131.

Rootzén, H. and N. Tajvidi (2006). Multivariate generalized Pareto distributions. *Bernoulli 12*(5), 917–930.

Rootzén, H., J. Segers, and J. L. Wadsworth (2018). Multivariate peaks over thresholds models. *Extremes 21*(1), 115–145.

Rootzén, H., J. Segers, and J. L. Wadsworth (2018). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis 165*, 117–131.

Sainsbury-Dale, M., A. Zammit-Mangion, and R. Huser (2024). Likelihood-Free Parameter Estimation with Neural Bayes Estimators. *The American Statistician 78*(1), 1–14.

Shi, M., L. Zhang, M. D. Risser, and B. A. Shaby (2026). Spatial scale-aware tail dependence modeling for high-dimensional spatial extremes. *Journal of the American Statistical Association* (just-accepted), 1–23.

Sisson, S. A., Y. Fan, and M. Beaumont (2018). *Handbook of approximate Bayesian computation*. CRC press.

Smith, R. L., J. A. Tawn, and S. G. Coles (1997). Markov chain models for threshold exceedances. *Biometrika 84*(2), 249–268.

Wadsworth, J., J. Tawn, A. Davison, and D. Elton (2017). Modelling across extremal dependence classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology 79*(1), 149–175.

Wadsworth, J. L. and J. Tawn (2022). Higher-dimensional spatial extremes via single-site conditioning. *Spatial Statistics 51*, 100677.

Wadsworth, J. L. and J. A. Tawn (2014). Efficient inference for spatial extreme value processes associated to log-gaussian random functions. *Biometrika 101*(1), 1–15.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature 466*(7310), 1102–1104.

Yadav, R., R. Huser, and T. Opitz (2021). Spatial hierarchical modeling of threshold exceedances using rate mixtures. *Environmetrics 32*(3), e2662.

Zaheer, M., S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola (2017). Deep sets. *Advances in neural information processing systems 30*.

Zhang, L., C. Rohrbeck, and T. Opitz (2025). Subasymptotic models for spatial extremes. In P. N. M. de Carvalho, R. Huser and B. Reich (Eds.), *Handbook on Statistics of Extremes*. CRC Press.