



# High-Order Composite Likelihood Inference for Max-Stable Distributions and Processes

Stefano CASTRUCCIO, Raphaël HUSER, and Marc G. GENTON

In multivariate or spatial extremes, inference for max-stable processes observed at a large collection of points is a very challenging problem and current approaches typically rely on less expensive composite likelihoods constructed from small subsets of data. In this work, we explore the limits of modern state-of-the-art computational facilities to perform full likelihood inference and to efficiently evaluate high-order composite likelihoods. With extensive simulations, we assess the loss of information of composite likelihood estimators with respect to a full likelihood approach for some widely used multivariate or spatial extreme models, we discuss how to choose composite likelihood truncation to improve the efficiency, and we also provide recommendations for practitioners. This article has supplementary material online.

**Key Words:** Composite likelihood; Efficiency; Max-stable process; Parallel computing; Spatial extremes.

## 1. INTRODUCTION

In the environmental statistics community, the development of models and inference methods for multivariate or spatial extremes is a very active area of research (Davison, Padoan, and Ribatet 2012). The max-stability property that underpins these extremal models justifies, at least asymptotically, extrapolation beyond the temporal scale of the observed data (Davison, Huser, and Thibaud 2013). In this framework, widely used max-stable processes include the Smith (1990), Schlather (2002), Brown–Resnick (Brown and Resnick 1977; Kabluchko, Schlather, and de Haan 2009), and extremal  $t$  (Nikoloulopoulos, Joe, and Li 2009; Opitz 2013) models. However, their practical use on big datasets has been hampered by the computationally challenging nature of likelihood functions, whose evaluation for frequentist or Bayesian inference requires computing and storing a prohibitively large amount of information. This challenge, arising even with moderate sample sizes, contrasts with that of the classical Gaussian-based geostatistics literature, where, thanks to recent advances, hundreds of thousands of locations, or even millions, can be handled using a wide

---

Stefano Castruccio, Lecturer, School of Mathematics and Statistics, Newcastle University, Newcastle Upon Tyne NE1 7RU, United Kingdom (E-mail: [stefano.castruccio@ncl.ac.uk](mailto:stefano.castruccio@ncl.ac.uk)). Raphaël Huser, Assistant Professor, (E-mail: [raphael.huser@kaust.edu.sa](mailto:raphael.huser@kaust.edu.sa)) and Marc G. Genton, Professor, (E-mail: [marc.genton@kaust.edu.sa](mailto:marc.genton@kaust.edu.sa)), CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.

© 2016 *American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*

*Journal of Computational and Graphical Statistics*, Volume 25, Number 4, Pages 1212–1229

DOI: [10.1080/10618600.2015.1086656](https://doi.org/10.1080/10618600.2015.1086656)

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/r/jcgs](http://www.tandfonline.com/r/jcgs).

variety of techniques; see the review by Sun, Li, and Genton (2012), and references therein, as well as recent work such as Nychka et al. (2015) and Sun and Stein (2016). Indeed, a single likelihood evaluation for a max-stable process observed at  $Q$  locations requires  $B_Q$  operations and storage with  $B_Q/Q = O(\log(Q))$ , which, compared to the “Big Data” problem of fitting a large multivariate normal distribution that requires  $O(Q^3)$  operations and  $O(Q^2)$  storage, is on a different level of computational complexity.

The cumulative distribution function of a max-stable process  $Z(\mathbf{x})$  observed at  $Q$  locations  $\mathbf{x}_1, \dots, \mathbf{x}_Q \in \mathcal{X} \subset \mathbb{R}^2$  may be written as

$$\mathbb{P}\{Z(\mathbf{x}_1) \leq z_1, \dots, Z(\mathbf{x}_Q) \leq z_Q\} = \exp[-V\{T(\mathbf{z})\}], \quad (1)$$

where  $T(\mathbf{z})$  is a suitable marginal transformation,  $\mathbf{z} = (z_1, \dots, z_Q)^\top$  and  $V(\mathbf{z})$  is a function called the exponent measure; see, for example, Huser, Davison, and Genton (2016). The corresponding density can be obtained by computing the derivative of (1) with respect to all elements of  $\mathbf{z}$ . This results in a combinatorial explosion of terms as a function of  $Q$  (e.g., the number of monitoring stations in spatial applications), which requires substantial computational power and a prohibitively large amount of information storage if  $Q$  is large (more details in Section 3).

To circumvent these computational challenges, the common approach has been to rely on composite likelihoods (Lindsay 1988; Varin, Reid, and Firth 2011), based on pairs (Padoan, Ribatet, and Sisson 2010) or triples (Genton, Ma, and Sang 2011; Huser and Davison 2013; Sang and Genton 2014). These misspecified likelihoods enable consistent but less efficient estimation of parameters under mild regularity conditions, and are computationally convenient: the number of terms involved decreases dramatically, and each of these terms is less expensive to compute. Although the loss of information entailed by pairwise or triplewise approaches has been largely investigated in various contexts (see, e.g., Cox and Reid 2004; Hjort and Varin 2008; Davis and Yau 2011; Bevilacqua et al. 2012; Eidsvik et al. 2014), little has been done to assess their performance, and that of higher-order counterparts, in the context of extremes. Huser, Davison, and Genton (2016) compared several likelihood inference methods for multivariate extremes, but they focused on the multivariate logistic model only, and did not consider composite likelihoods of orders higher than two (except for the full likelihood). Bienvenüe and Robert (2016) proposed partition-based composite likelihoods, but they assessed their performance for blocks of maximum dimension five, focusing on a clustered max-stable process and on the Schlather (2002) process. Furthermore, they did not compare composite likelihoods of different orders, and did not provide any assessment of the associated computational burden. Our article not only complements the existing literature by conducting an extensive simulation study assessing the practical performance of low- and high-order composite likelihoods for several max-stable models, but it also offers a computational perspective on this problem by exploring the limits of likelihood-based inference, suggesting how state-of-the-art computational resources may be efficiently used in this framework, and giving recommendations to practitioners. We also demonstrate that suitably truncated low-order composite likelihoods can improve the inference, while decreasing the computational time and thus allowing analyses in higher dimensions. Our simulation study focuses on max-stable models of increasing computational complexity: we start with the widely studied multivariate logistic model (Gumbel 1960), then, we consider the model for spatial extremes advocated by Reich and

Shaby (2012), which can be viewed as a spatial generalization of the logistic model; finally, we investigate a related, likely more realistic, stationary process, the Brown–Resnick model.

The rest of the article is organized as follows: Section 2 gives some background on multivariate or spatial extremes and introduces the models used in this work. Section 3 discusses inference for extremes based on full or composite likelihoods and describes the related computational challenges. In Section 4, the performance of composite likelihoods is assessed by simulation, and truncation strategies are discussed. Section 5 provides estimates of memory and computational requirements in high dimensions, as well as practical recommendations, and Section 6 concludes with perspectives on inference for extremes with current and future computer architectures.

## 2. EXTREME-VALUE THEORY AND MODELS

Suppose that  $Y_n$  is a sequence of independent and identically distributed (iid) random variables. Furthermore, let there exist sequences of normalizing constants,  $a_n > 0$  and  $b_n$ , such that  $Z_n = a_n^{-1}(\max_{i=1,\dots,n} Y_i - b_n)$  converges in distribution to a nondegenerate random variable  $Z$ , as  $n \rightarrow \infty$ . Then, the distribution of  $Z$  has to be of the form

$$\mathbb{P}(Z \leq z) = \exp \left[ - \left\{ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right\}_+^{-1/\xi} \right], \quad (2)$$

where  $a_+ = \max(0, a)$  and  $\mu, \sigma > 0$  and  $\xi$  are denominated the location, scale, and shape parameters, respectively. The case  $\xi = 0$  is formally undefined but is understood as the limit as  $\xi \rightarrow 0$ . The right-hand side of (2) is referred to as the generalized extreme-value family, or in short  $\text{GEV}(\mu, \sigma, \xi)$ , and has been extensively used as a model for block maxima; see, for example, Coles (2001) or the review article by Davison and Huser (2015).

Likewise, in the multivariate setting, suppose that  $\mathbf{Y}_1 = (Y_{1;1}, \dots, Y_{1;Q})^\top$ ,  $\mathbf{Y}_2 = (Y_{2;1}, \dots, Y_{2;Q})^\top, \dots$ , is a sequence of iid random vectors with unit Fréchet margins, that is, each component is distributed according to (2) with  $\mu = \sigma = \xi = 1$ ; the general case is easily treated by using the probability integral transform to put data on the unit Fréchet scale. Then, as  $n \rightarrow \infty$ , the vector of renormalized componentwise maxima  $\mathbf{Z}_n = n^{-1}(\max_{i=1,\dots,n} Y_{i;1}, \dots, \max_{i=1,\dots,n} Y_{i;Q})^\top$  converges in distribution (see, e.g., Coles 2001) to a random vector  $\mathbf{Z}$  distributed according to a multivariate extreme-value (or equivalently, max-stable) distribution (1), with the marginal transformation  $T(\mathbf{z})$  being the identity, and where the exponent measure  $V(\mathbf{z})$  is a positive function such that  $V(a^{-1}\mathbf{z}) = aV(\mathbf{z})$  for any  $a > 0$  and  $\mathbf{z} > \mathbf{0}$  (i.e.,  $V$  is homogenous of order  $-1$ ). This function satisfies the marginal constraints  $V(z\mathbf{e}_q^{-1}) = z^{-1}$  for any  $z > 0$  and  $q = 1, \dots, Q$ , where  $\mathbf{e}_q$  denotes the  $q$ th canonical basis vector in  $\mathbb{R}^Q$ , and  $1/0 = \infty$  by convention. Numerous multivariate models have been proposed in the literature (see, e.g., Joe 1997; Tawn 1988, 1990; Cooley, Davis, and Naveau 2010; Ballani and Schlather 2011; Segers 2012).

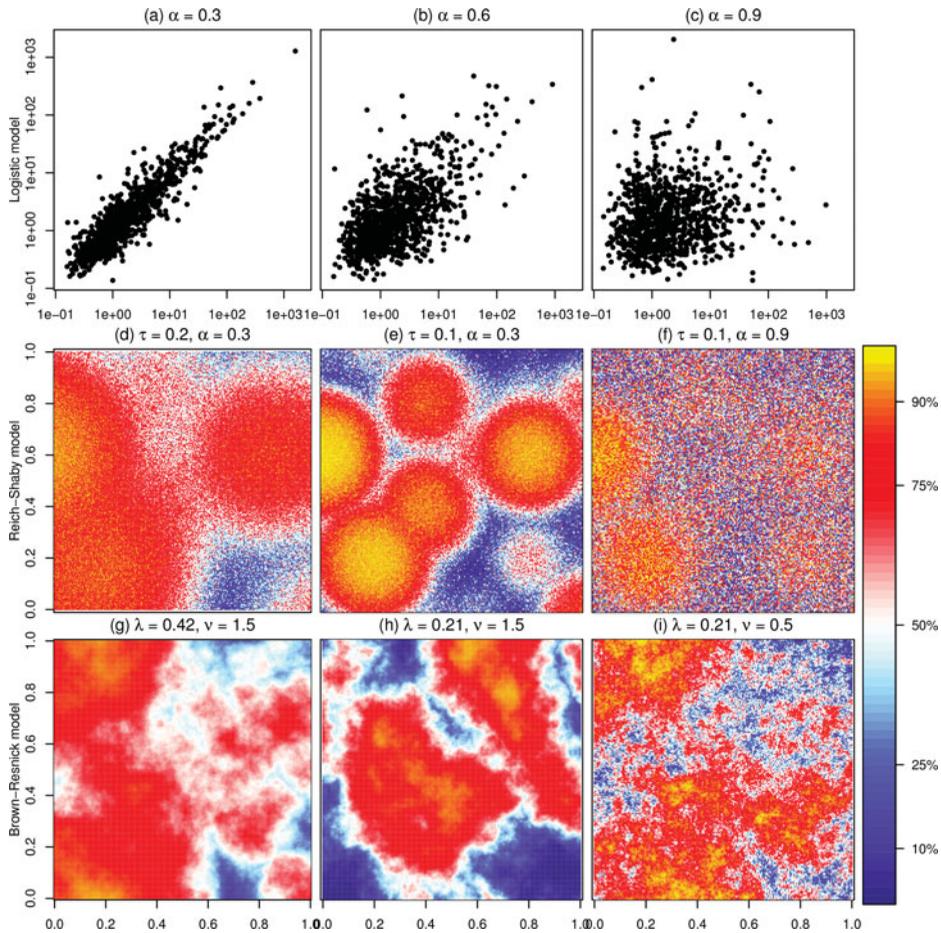


Figure 1. Examples of realizations for the two-dimensional logistic model ((a)–(c)), Reich–Shaby process with 36 knots on a regular grid  $[0, 0.2, \dots, 1]^2$  ((d)–(f)), and Brown–Resnick process with semivariogram  $\gamma(\mathbf{h}) = (\|\mathbf{h}\|/\lambda)^\nu$  ((g)–(i)). The latter are simulated using the `RandomFields` package (Schlather et al. 2015) of the statistical software R. Marginal quantiles are indicated by the color scale. Parameters are indicated above the corresponding panels. The grid spacing is 0.005.

The first proposed and most widely used model is the symmetric logistic model (Gumbel 1960), which assumes that

$$V(\mathbf{z}) = \left( \sum_{q=1}^Q z_q^{-1/\alpha} \right)^\alpha, \tag{3}$$

for some dependence parameter  $0 < \alpha \leq 1$ . When  $\alpha = 1$ , the components of  $\mathbf{Z}$  are mutually independent, whereas perfect dependence is attained as  $\alpha \rightarrow 0$ ; see typical realizations in Figure 1(a)–(c). By symmetry, components of  $\mathbf{Z}$  are equidependent. A possible asymmetric extension of (3) is to consider a max-mixture of logistic variables. More precisely, for  $l = 1, \dots, L$ , consider independent random vectors  $\mathbf{Z}_l = (Z_{l;1}, \dots, Z_{l;Q})^\top$  distributed according to (1) and (3) with dependence parameters  $0 < \alpha_l \leq 1$ . Then, the random

vector constructed as  $(\max_{l=1,\dots,L} w_l Z_{l;1}, \dots, \max_{l=1,\dots,L} w_l Z_{l;Q})^\top$ , with nonnegative weights  $w_1, \dots, w_L$  such that  $\sum_{l=1}^L w_l = 1$ , is distributed as (1) with

$$V(\mathbf{z}) = \sum_{l=1}^L \left\{ \sum_{q=1}^Q \left( \frac{z_q}{w_l} \right)^{-1/\alpha_l} \right\}^{\alpha_l}. \tag{4}$$

This model is different from, but closely related to, the asymmetric logistic model proposed by Tawn (1990). Although (4) is more flexible than (3), the number of parameters equals  $2L - 1$  and thus increases dramatically for large  $L$ , and, for practical use, the data structure should be exploited to construct simpler models.

In the spatial framework, Reich and Shaby (2012) proposed a max-stable process with finite-dimensional distributions that may be expressed through (4). Specifically, let  $Z(\mathbf{x})$  ( $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^2$ ) be a spatial process defined on the plane as  $Z(\mathbf{x}) = U(\mathbf{x})\theta(\mathbf{x})$ , where  $U(\mathbf{x}) \stackrel{\text{iid}}{\sim} \text{GEV}(1, \alpha, \alpha)$  is a random noise process, and  $\theta(\mathbf{x}) = \{\sum_{l=1}^L A_l w_l(\mathbf{x})^{1/\alpha}\}^\alpha$ , with  $\sum_{l=1}^L w_l(\mathbf{x}) = 1$  and  $w_l(\mathbf{x}) \geq 0$  for any  $\mathbf{x} \in \mathcal{X}$ . The variables  $A_l$  are underlying independent random effects distributed according to the  $\alpha$ -stable distribution ( $0 < \alpha \leq 1$ ) (see Stephenson 2009 for some background on  $\alpha$ -stable variables), and the deterministic weights  $w_l(\mathbf{x})$  capture the latent spatial structure. Then, for any finite subset of locations  $\{\mathbf{x}_1, \dots, \mathbf{x}_Q\} \subset \mathcal{X}$ , it can be shown that the random vector  $\{Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_Q)\}^\top$  is distributed according to (1) with

$$V(\mathbf{z}) = \sum_{l=1}^L \left[ \sum_{q=1}^Q \left\{ \frac{z_q}{w_l(\mathbf{x}_q)} \right\}^{-1/\alpha} \right]^\alpha, \tag{5}$$

which is a special case of (4) with fewer parameters. The weights may be further written as  $w_l(\mathbf{x}) = k_l(\mathbf{x})\{\sum_{l=1}^L k_l(\mathbf{x})\}^{-1}$ , where  $k_l(\mathbf{x}) = \frac{1}{2\pi\tau^2} \exp\{-\frac{1}{2\tau^2}(\mathbf{x} - \mathbf{v}_l)^\top(\mathbf{x} - \mathbf{v}_l)\}$ , for  $\tau > 0$ , and some fixed knots  $\mathbf{v}_1, \dots, \mathbf{v}_L \in \mathcal{X}$ . For known knots, this model has only two unknown parameters:  $\alpha$  controls the amount of noise in the resulting random field, and  $\tau$  is a range parameter. Smaller values of  $\alpha$  result in less noisy processes (with  $\alpha \rightarrow 0$  corresponding to deterministic profiles), while increasing  $\tau$  implies larger-scale processes; see typical realizations in Figure 1(d)–(f). In particular, it can be shown that, as  $\alpha \rightarrow 0$  and as regularly spaced knots get denser (with  $L \rightarrow \infty$ ), model (5) converges to the stationary max-stable model proposed by Smith (1990). Although appealing, the Reich–Shaby model has several peculiarities. First, the process is stationary and isotropic only as  $L \rightarrow \infty$ . Second, as illustrated in Figure 1(e), realizations from (5) may be strongly affected by the locations of the knots, if the range parameter  $\tau$  is small compared to the knot spacings. And finally, realized random fields are either discontinuous in every point (with  $\alpha \in (0, 1]$ ) or analytic almost everywhere (with  $\alpha \rightarrow 0$ ). Despite these features, which may be seen as limitations in real applications, this model is simple, intuitive, computationally convenient, and inference can be performed at practically the same cost as for the logistic distribution. These computational advantages have proven to be especially useful in Bayesian hierarchical modeling of spatial extremes (Reich and Shaby 2012).

A popular and likely more realistic stationary max-stable model is the Brown–Resnick process (Brown and Resnick 1977; Kabluchko, Schlather, and de Haan 2009), which may be represented as  $Z(\mathbf{x}) = \sup_{i \in \mathbb{N}} W_i(\mathbf{x})/T_i$ ; the  $W_i(\mathbf{x})$ ’s are independent copies of a random

process  $W(\mathbf{x}) = \exp\{\varepsilon(\mathbf{x}) - \gamma(\mathbf{x})\}$ , where  $\varepsilon(\mathbf{x})$  is an intrinsically stationary Gaussian random field with semivariogram  $\gamma(\mathbf{h})$  and  $\varepsilon(\mathbf{0}) = 0$  almost surely, and the  $T_i$ 's are occurrences of a unit-rate Poisson process on the positive real line. Such a process has finite-dimensional distributions that may be written as (1) with exponent measure

$$V(\mathbf{z}) = \sum_{q=1}^Q \frac{\Phi_{Q-1}(\boldsymbol{\eta}_q; \mathbf{0}, \mathbf{R}_q)}{z_q}, \quad (6)$$

where  $\Phi_D(\mathbf{a}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  is the cumulative distribution function of a  $D$ -variate normal variable with mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$ , evaluated at  $\mathbf{a}$ . Quantities  $\boldsymbol{\eta}_q$  and  $\mathbf{R}_q$  appearing in (6) are defined in the supplementary material. The Brown–Resnick process with fractional Brownian semivariogram, that is,  $\gamma(\mathbf{h}) = (\|\mathbf{h}\|/\lambda)^\nu$ ,  $\lambda > 0$ ,  $\nu \in (0, 2]$ , is illustrated in Figure 1(g)–(i). In this case,  $\lambda$  is a range parameter, whereas  $\nu$  is a smoothness parameter with higher values indicating smoother processes. In particular, the isotropic Smith (1990) process (with analytical storm profiles) is recovered when  $\nu = 2$ . Hence, both the Reich–Shaby and Brown–Resnick models can be viewed as nonsmooth extensions of the Smith model, but they are not nested in each other.

### 3. INFERENCE AND COMPUTATIONAL CHALLENGES

Likelihood inference for multivariate or spatial extremes is computationally challenging. Although the cumulative distribution function can be easily written as (1), the density for a single vector  $\mathbf{z} = (z_1, \dots, z_Q)^\top$  is much more complicated since it corresponds to the derivative of (1) with respect to all components of  $\mathbf{z}$ . Assume that  $V(\mathbf{z}) = V(\mathbf{z} | \boldsymbol{\theta})$ , that is, the exponent measure is parameterized by some  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ , and that all derivatives of  $V(\mathbf{z} | \boldsymbol{\theta})$  exist. The likelihood in the case of unit Fréchet margins, where  $T(\mathbf{z})$  in (1) is the identity, is

$$L_Q(\boldsymbol{\theta} | \mathbf{z}) = \exp\{-V(\mathbf{z} | \boldsymbol{\theta})\} \sum_{\mathcal{P} \in \mathcal{P}_{\mathbf{z}}} \prod_{S \in \mathcal{P}} \{-V_S(\mathbf{z} | \boldsymbol{\theta})\}, \quad (7)$$

where  $\mathcal{P}_{\mathbf{z}}$  is the collection of all partitions of  $\mathbf{z}$ ,  $S \neq \emptyset$  is a particular set in partition  $\mathcal{P} \in \mathcal{P}_{\mathbf{z}}$ , and  $V_S(\mathbf{z} | \boldsymbol{\theta})$  is the partial derivative of  $V(\mathbf{z} | \boldsymbol{\theta})$  with respect to the elements in the set  $S$ . The main issues regarding the computation of (7) are the following:

1. Closed-form expressions for  $V(\mathbf{z} | \boldsymbol{\theta})$  and  $V_S(\mathbf{z} | \boldsymbol{\theta})$  are not always available. Although Monte Carlo methods could in principle be used to approximate these functions for some models, this would certainly result in a prohibitively large amount of computations, especially in large dimensions. In this article, we focus on the three models, (3), (5), and (6), where closed forms are known.
2. The  $2^Q - 1$  partial derivatives of  $V(\mathbf{z} | \boldsymbol{\theta})$ , corresponding to all nonempty subsets  $S$  of  $\{1, \dots, Q\}$  must be computed. Depending on the class of models, this step can be more or less demanding. In this work, the partial derivative evaluation is increasingly complex as we consider models (3), (5), and (6).

3. The vector of partial derivatives must be assembled to compute the sum across all partitions. As the cardinality of  $\mathcal{P}_{\mathbf{z}}$  equals  $B_Q$ , the Bell number of order  $Q$  (Graham, Knuth, and Patashnik 1988), which grows more than exponentially with  $Q$ , the storage of very large data structures is required.

Depending on the model considered, issues 2 and 3 could prevent likelihood evaluation for high-dimensional data, and several solutions can be devised to improve computations. If issue 2 is the most computationally demanding operation, the evaluation of the  $2^Q - 1$  derivatives can be performed in parallel across processors. Since in our simulation study (see Section 4), multiple independent experiments are performed, we have found that parallelization across experiments is the most efficient solution, but if the goal is to analyze one dataset, this option can greatly reduce the computational time if  $Q$  is large. Combining the derivatives into a sum across all partitions in issue 3 is the main limitation in high dimensions and cannot be easily improved with parallelization. For each likelihood evaluation, the vector  $\mathbf{D}_Q$  of all the  $2^Q - 1$  partial derivatives is computed, and to efficiently assemble the sum across all partitions, an efficient strategy is to precompute  $\mathbf{P}_Q$ , a  $B_Q \times 2$  cell array (with elements stored as 16-bit integers instead of double precision), with each row consisting of a different partition. The first column consists of the sets in a partition, and the second consists of the indexes of the sets in  $\mathbf{D}_Q$ . This approach allows for efficient storage, but large  $Q$ 's make it impossible to store  $\mathbf{P}_Q$ . To avoid this, it is possible to divide  $\mathcal{P}_{\mathbf{z}}$  into groups and to compute the sum dynamically for each group. In our experience, the computation of a larger number of smaller subsets of partitions is computationally very demanding (see Section 5 for a discussion about groups of Stirling partitions) and, since this needs to be performed for each likelihood evaluation, it has proven to be very slow. In Section 5, we show that a full likelihood evaluation, implemented efficiently on a powerful computer, can be performed in maximum dimension  $Q = 11$  in reasonable time. Parallelization across experiments implies the creation of a copy of  $\mathbf{P}_Q$  for each processor but if the goal is the analysis of one dataset, a single data structure is created and the computation could possibly be increased to  $Q = 12$  or  $Q = 13$  if a large memory is available; computation for higher dimensions is completely out of reach with current technologies.

To avoid these shortcomings, the standard approach (Padoan, Ribatet, and Sisson 2010; Genton, Ma, and Sang 2011; Huser and Davison 2013; Bienvenüe and Robert 2016; Sang and Genton 2014) is to consider composite likelihoods. If we denote by  $\mathbf{z}_q$  a vector of  $q$  elements from a  $Q$ -dimensional vector  $\mathbf{z}$  and  $C_{\mathbf{z}_q}(Q)$  the collection of all possible subvectors  $\mathbf{z}_q$ , then a composite likelihood of order  $q$  may be expressed as

$$\text{CL}_q(\boldsymbol{\theta} \mid \mathbf{z}) = \prod_{\mathbf{z}_q \in C_{\mathbf{z}_q}(Q)} L_q(\boldsymbol{\theta} \mid \mathbf{z}_q)^{\omega_{\mathbf{z}_q}}, \quad (8)$$

where each contribution  $L_q(\boldsymbol{\theta} \mid \mathbf{z}_q)$  is a likelihood term of order  $q$ , as defined in (7), and the real numbers  $\omega_{\mathbf{z}_q}$  are positive weights that do not necessarily sum up to one. In general, evaluation of  $L_q(\boldsymbol{\theta} \mid \mathbf{z}_q)$  is more convenient than  $L_Q(\boldsymbol{\theta} \mid \mathbf{z})$  for small  $q$ 's, but small  $q$ 's imply large sets  $C_{\mathbf{z}_q}(Q)$ , so the computational time is not necessarily monotonically increasing with  $q$ , as is shown in Section 5. Since composite likelihoods are built from valid likelihood terms, they inherit some of the large-sample properties from the full likelihood. More

precisely, suppose that  $m$  independent replicates  $\mathbf{z}_1, \dots, \mathbf{z}_m$  of an extreme-value distributed vector are observed, and consider the full and composite likelihoods constructed from (7) and (8) as

$$L_Q(\boldsymbol{\theta} \mid \mathbf{z}_1, \dots, \mathbf{z}_m) = \prod_{i=1}^m L_Q(\boldsymbol{\theta} \mid \mathbf{z}_i), \quad \text{CL}_Q(\boldsymbol{\theta} \mid \mathbf{z}_1, \dots, \mathbf{z}_m) = \prod_{i=1}^m \text{CL}_Q(\boldsymbol{\theta} \mid \mathbf{z}_i). \quad (9)$$

Let  $\hat{\boldsymbol{\theta}}$  and  $\hat{\boldsymbol{\theta}}_C$  denote the estimators of  $\boldsymbol{\theta}$  maximizing the full likelihood and composite likelihood in (9), respectively. Then, under mild regularity conditions,  $\hat{\boldsymbol{\theta}}$  and  $\hat{\boldsymbol{\theta}}_C$  are both strongly consistent as  $m \rightarrow \infty$ , asymptotically Gaussian, and converge at the same rate, namely,  $\sqrt{m}$  (see Padoan, Ribatet, and Sisson 2010). However, the variability of  $\hat{\boldsymbol{\theta}}_C$  is typically larger than that of  $\hat{\boldsymbol{\theta}}$ , and it depends on the choice of weights  $\omega_{\mathbf{z}_q}$  in (8). When weights are ignored, that is,  $\omega_{\mathbf{z}_q} = 1$ , we refer to the corresponding composite likelihood as a *complete composite likelihood*; when the  $\omega_{\mathbf{z}_q}$ 's are not all equal, we use the expression *weighted composite likelihood*; when binary weights are used, that is,  $\omega_{\mathbf{z}_q} = 0$  or 1, we refer to it as a *truncated composite likelihood*. The latter is also called tapered composite likelihood by Sang and Genton (2014). The performance of complete composite likelihoods is studied in Section 4.2 for the models introduced in Section 2, and truncated alternatives are explored in Section 4.3. As we will show, truncation allows us to drastically reduce the computational time, and in some cases also to improve the efficiency. Section 5 contains more detailed discussion on the computational cost of composite likelihoods.

## 4. PERFORMANCE OF COMPOSITE LIKELIHOODS

### 4.1 GENERAL SETTING

In this section, we detail simulation studies for the three models mentioned in Section 2: we consider them in order of increasing computational complexity, considering first the logistic model (3), then the Reich–Shaby model (5) and finally the Brown–Resnick model (6). The goal of this section is to assess the improvement of high-order composite likelihoods compared to the traditional pairwise and triplewise approaches, in terms of root mean squared error. We deliberately choose a relatively small number  $Q$  of locations ( $Q = 11$  for the logistic and the Reich–Shaby models, and  $Q = 9$  for the Brown–Resnick model), so that full likelihoods can be computed in a reasonable time, and we provide estimated projections of the computational cost in higher dimensions in Section 5.

All simulations (obtained with the same random seed) were performed on a fully dedicated 39-node cluster with 20 cores (and 512 Gb of RAM) per node, for a total of 780 cores. The algorithms were implemented in MATLAB, and each likelihood maximization was performed using a Nelder–Mead algorithm, allowing at most 1000 iterations and a tolerance of convergence of 0.01 between successive iterates. Given the small dimensionality of the parameter space ( $p = 1$  or 2), these specifications were sufficient to achieve accurate results for all experiments.

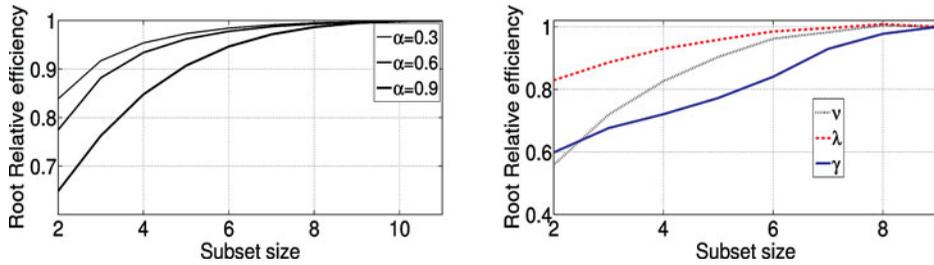


Figure 2. Left: Root relative efficiency for the simulation study for the logistic model computed over 1000 simulations of  $m = 50$  replicates each, with  $\alpha = 0.3, 0.6, 0.9$  (with increasingly thicker lines). The range of the absolute bias to standard deviation ratio is (3.4%, 3.8%) (median 3.8%), (1.3%, 4.3%) (median 4.0%), and (0.0%, 8.1%) (median 4.1%) for  $\alpha = 0.3, 0.6, 0.9$ , respectively. The plot of the standard deviation ratio (not shown) is graphically undistinguishable. Right: Root relative efficiency of complete composite likelihood estimators for  $\lambda$  (in dashed red) and for  $\nu$  (in dotted black), for 100 experiments of  $m = 50$  replicates of the Brown–Resnick process with semivariogram  $\gamma(\mathbf{h}) = (\|\mathbf{h}\|/\lambda)^\nu$  for  $\lambda = 0.42$  and  $\nu = 1.5$ , observed at nine locations. Results are plotted against the subset size  $q$  of composite likelihoods. The range of the absolute bias to standard deviation ratio is (1.9%, 13.5%) (median 12.2%) for  $\alpha$  and (9.7%, 15.7%) (median 14.2%) for  $\lambda$ . The plot of the standard deviation ratio (not shown) is graphically undistinguishable. The ratio of  $L^2$  distances (taken over the range from  $\|\mathbf{h}\| = 0$  to  $\|\mathbf{h}\| = \sqrt{2}$ ) between the estimated and the true semivariogram  $\gamma(\mathbf{h})$  at  $Q = 9$  as a function of  $q$  is also shown in solid blue.

## 4.2 COMPLETE COMPOSITE LIKELIHOODS

*The Logistic Model.* For  $\alpha = 0.3, 0.6, 0.9$  (strong, mild, and weak dependence, respectively), we perform 1000 independent experiments. For each experiment  $j = 1, \dots, 1000$ , we simulate  $m = 50$  independent replicates (a realistic number for most environmental applications) from a  $Q = 11$ -dimensional logistic distribution and obtain the maximum complete composite likelihood estimator of order  $q$ ,  $\hat{\alpha}_{j,q}$ , from (9) with subsets of cardinality  $q = 2, \dots, Q$ . Following Huser and Davison (2013), the root mean squared error  $\text{rmse}_q = \sqrt{\text{b}_q^2 + \text{sd}_q^2}$  is computed, where the bias is  $\text{b}_q = \bar{\alpha}_q - \alpha$ , ( $\bar{\alpha}_q = \sum_{j=1}^{1000} \hat{\alpha}_{j,q}/1000$ ), and the standard deviation is  $\text{sd}_q = \sqrt{\sum_{j=1}^{1000} (\hat{\alpha}_{j,q} - \bar{\alpha}_q)^2/999}$ . The root relative efficiencies defined as  $\text{re}_q = \text{rmse}_Q/\text{rmse}_q$  are shown in Figure 2 (left), and the range (and the sample median) of absolute bias to standard deviation ratio is reported in the caption. Overall, complete composite likelihood estimators improve as  $q$  increases. Moreover, as the dependence decreases (i.e.,  $\alpha$  increases), the gain of high-order composite likelihood estimators with respect to low-order counterparts is more evident. Overall, the root mean squared error of the full likelihood estimator is between 16% and 35% (respectively, between 8% and 23%) smaller than that of the pairwise (respectively, triplewise) complete composite likelihood estimator.

*The Reich–Shaby Model.* The logistic model (3) considered previously applies to multivariate extremes, but has no notion of spatial dependence. As discussed in Section 2, the Reich–Shaby model is a spatial extension of the logistic model, whose exponent measure and derivatives are relatively convenient to compute (see supplementary material).

To be consistent with the logistic model, 1000 experiments are performed, each with  $m = 50$  independent replicates. For each experiment,  $Q = 11$  stations are uniformly generated on  $[0, 1]^2$ , and data are simulated at those (fixed) locations with knots  $\mathbf{v}_1, \dots, \mathbf{v}_{36}$  on a regular grid  $[0, 0.2, \dots, 1]^2$ . Different levels of noise, with  $\alpha = 0.3, 0.6, 0.9$  (little to very

Table 1. Root relative efficiency (%) for  $\hat{\alpha}$  and  $\hat{\tau}$  computed over 1000 simulations of  $m = 50$  replicates each, for different values of  $\alpha$  and  $\tau$  in the Reich–Shaby model. The range of the absolute bias to standard deviation ratio is (0.2%, 27.4%) (median 6.0%) for  $\alpha$  and (0.9%, 17.8%) (median 7.9%) for  $\tau$

$\tau$	$\alpha$	$\hat{\alpha}/\hat{\tau}$									
		$q = 2$	$q = 3$	$q = 4$	$q = 5$	$q = 6$	$q = 7$	$q = 8$	$q = 9$	$q = 10$	$q = 11$
0.1	0.3	76/65	82/77	87/82	90/87	93/91	95/95	97/97	99/99	100/100	100/100
	0.6	77/55	83/71	87/78	90/84	92/89	94/95	96/96	98/97	99/99	100/100
	0.9	79/74	83/86	84/82	85/91	81/89	90/93	90/89	96/100	97/97	100/100
0.2	0.3	55/29	69/44	79/62	85/75	90/84	94/91	96/95	98/98	100/99	100/100
	0.6	75/54	84/64	89/74	92/82	95/88	97/93	98/96	99/97	100/98	100/100
	0.9	80/72	84/80	87/83	89/79	92/87	94/94	95/98	96/103	98/105	100/100
0.4	0.3	72/56	81/63	87/70	92/76	95/82	97/87	98/92	99/95	100/97	100/100
	0.6	77/74	86/86	91/91	95/94	97/97	99/98	100/100	100/100	100/101	100/100
	0.9	59/94	83/93	72/97	74/100	76/100	71/100	99/106	100/103	100/105	100/100

noisy), and different dependence ranges, with  $\tau = 0.1, 0.2, 0.4$  (short- to long-range), are considered. Holding knots fixed, we estimate the two parameters  $\alpha$  and  $\tau$  using maximum complete composite likelihoods estimators of order  $q = 2, \dots, Q$ , defined in (9). The root relative efficiency  $\text{rre}_q$  is reported in Table 1. Estimates of  $\alpha$  show a strictly monotonic improvement with higher-order composite likelihoods, as noticed for the logistic case, while for  $\tau$  the efficiency is not strictly monotonic for high  $q$ 's. Overall, the root mean squared error of the full likelihood estimator is 20% to 45% (respectively, 14% to 31%) smaller than that of the pairwise (respectively, triplewise) complete composite likelihood estimator for  $\hat{\alpha}$  and 6% to 71% (respectively, 7% to 66%) for  $\hat{\tau}$ .

Besides the increase in root relative efficiency, we found evidence that the use of high-order likelihoods also results in a decrease in correlation between  $\hat{\alpha}$  and  $\hat{\tau}$ , especially when this correlation is nonnegligible (see supplementary material).

*The Brown–Resnick Model.* After an experiment on a relatively simple model for spatial extremes, we conduct a simulation study in a more realistic and more computationally intensive context: 100 experiments are performed, each consisting of  $m = 50$  replicates of a Brown–Resnick process with semivariogram  $\gamma(\mathbf{h}) = (\|\mathbf{h}\|/\lambda)^\nu$  and  $\lambda = 0.42$ ,  $\nu = 1.5$  (a case considered by Huser and Davison (2013), and illustrated in Figure 1(g)), on  $Q = 9$  stations uniformly generated in the unit square. The dimensionality and number of experiments are reduced from the two previous settings since evaluating partial derivatives of the exponent measure for the Brown–Resnick model requires expensive computations of high-dimensional normal cumulative distribution functions (see (6) and the supplementary material), which rely on quasi-Monte Carlo approaches (Genz 1992; Genz and Bretz 2002, 2009). The range  $\lambda > 0$  and smoothness  $\nu \in (0, 2]$  parameters are estimated using maximum complete composite likelihood estimators of order  $q = 2, \dots, Q$ . Since the complete simulation study in this case requires approximately 2 weeks, it was not possible to repeat it for other parameter values. As we can see from the results in Figure 2 (right), both parameters show an increase in efficiency with subset size  $q$ . This is especially striking for  $\nu$ : the pairwise (respectively, triplewise) likelihood estimator is approximately 44% (respectively, 28%) less efficient than the full likelihood estimator. By comparison, the

loss in efficiency for the range  $\lambda$  is about 17% (respectively, 13%) for pairwise (respectively, triplewise) likelihood estimators. High values of  $q$  did not result in a noticeable change in the estimated (negative) correlation between  $\hat{\lambda}$  and  $\hat{\nu}$  (results not shown). Besides the interest in parameter estimation, we also quantify how close the estimated spatial structure is to the true one. In Figure 2 (right), we report  $\|\gamma - \hat{\gamma}_Q\|_2 / \|\gamma - \hat{\gamma}_q\|_2$ , where  $\|\cdot\|_2$  denotes the  $L^2$  norm taken over the range from  $\|\mathbf{h}\| = 0$  to  $\|\mathbf{h}\| = \sqrt{2}$ , and  $\hat{\gamma}_q$  is the variogram estimated with a composite likelihood of order  $q$ . It is apparent that the use of high-order likelihoods results in a better agreement with the true spatial dependence, and that the increase in efficiency is more striking compared to parameter estimates. Overall, the patterns for  $\nu$  and  $\lambda$  seem coherent with what was obtained for the logistic and the Reich–Shaby models, so we believe these results can be extrapolated by analogy to Brown–Resnick processes with other parameter combinations.

### 4.3 TRUNCATED COMPOSITE LIKELIHOODS

In Section 4.2, we showed that the use of high-order complete composite likelihoods, or full likelihoods when possible, results in a better estimation performance. However, to further improve the estimation and to reduce the computational burden, an option is to consider the larger class of weighted composite likelihoods and to investigate how to make the best choice of weights in (8). One solution could be to choose weights minimizing the trace or the determinant of the asymptotic variance of  $\hat{\theta}_C$  (Padoan, Ribatet, and Sisson 2010; Sang and Genton 2014), but this is challenging to implement and computationally demanding. Furthermore, if weights are all nonzero, this approach might improve the estimation efficiency, but it does not reduce the computational time, since it still requires evaluating all the elements in  $C_{z_q}(Q)$ . An alternative solution is to select only some elements of the collection  $C_{z_q}(Q)$  in (8) (i.e., choose binary weights) depending on their set distances, on the ground that close locations are generally more informative about dependence parameters than are distant, less correlated, ones. We propose to rank all elements in  $C_{z_q}(Q)$  according to their maximum set distance (i.e., the maximum distance among all pairs in the subset), and to consider only a percentage of the ranked vector.

To investigate the efficiency of this approach, we perform 1000 experiments with  $m = 50$  replicates of the Reich–Shaby model. For  $t = 0.1, 0.2, \dots, 1$ , we consider the first  $[t \times |C_{z_q}(Q)|]$  elements of the ranked collection  $C_{z_q}(Q)$ , and we compute the root relative efficiency  $\text{rre}_q$  (with respect to the full likelihood) as in Section 4.2. In other words, we increase the number of elements of  $C_{z_q}(Q)$  from 10% to 100% and investigate how the root mean squared error is affected for different values of  $q$ . The 100% case considers the whole set  $C_{z_q}(Q)$  and therefore corresponds to the complete composite likelihood case of Section 4.2. This simulation study requires maximizing a likelihood for every value of  $t$ , of  $q = 2, \dots, Q - 1$  with  $Q = 11$  (the full likelihood case for  $q = Q$  has  $|C_{z_Q}(Q)| = 1$ ) and experiment  $j = 1, \dots, 1000$ , for a total of about  $10^5$  maximizations. This simulation study requires a week of computation on the fully dedicated cluster for each parameter choice. Table 2 reports the results for  $\alpha = 0.6$  and  $\tau = 0.2$ , but similar patterns were observed for all other parameter configurations. Regarding the estimation of  $\alpha$ , the most efficient estimators are for  $t = 100\%$  when  $q \geq 6$ . For  $q = 2$ , the best estimate is at  $t = 70\%$  (which confirms

Table 2. Root relative efficiency with respect to the full likelihood for  $\hat{\alpha}$  and  $\hat{\tau}$ , computed over 1000 simulations of  $m = 50$  replicates each of the Reich–Shaby model with  $\tau = 0.2$  and  $\alpha = 0.6$ , considering  $[t \times |C_{z_q}(Q)|]$  elements  $t = 0.1, 0.2, \dots, 1$ . In bold, the maximum efficiency across  $t$  for all composite likelihood orders. In the last two rows, the smallest  $t$  (%) for the  $q$ -set to beat the best result for  $(q - 1)$ -set. In parentheses, the ratio ( $\times 100$ ) of the elapsed times (averaged across experiments) between these two combinations: values less than 100 mean that it is less time demanding to use an optimal  $t$  for  $(q - 1)$ -sets rather than considering  $q$ -sets, and vice versa. The range of the absolute bias to standard deviation ratio is (0.0%, 20.2%) (median 3.3%) for  $\alpha$  and (6.2%, 21.0%) (median 12.7%) for  $\tau$

		$\hat{\alpha}/\hat{\tau}$								
$q$	$t$ (%)	2	3	4	5	6	7	8	9	10
10		52/15	79/48	84/66	87/72	89/76	92/82	94/87	95/92	96/89
20		63/22	84/64	87/72	89/75	92/81	94/87	96/90	96/92	98/94
30		70/36	85/69	87/74	91/78	93/86	95/88	96/92	98/93	98/95
40		74/46	<b>86/70</b>	89/74	91/80	94/86	95/89	97/93	98/95	98/97
50		78/54	<b>86/69</b>	89/74	92/81	94/86	96/90	97/93	98/96	100/98
60		<b>79/61</b>	86/68	<b>89/75</b>	92/81	94/87	96/90	98/94	99/97	100/98
70		<b>79/61</b>	85/67	<b>89/74</b>	92/81	95/87	97/91	98/95	99/97	100/97
80		78/58	85/66	89/74	92/81	95/87	97/92	98/95	99/98	100/97
90		77/56	84/65	89/74	<b>93/82</b>	95/88	97/92	<b>98/96</b>	<b>99/98</b>	<b>100/98</b>
100		75/53	84/64	89/74	93/82	<b>95/88</b>	<b>97/92</b>	<b>98/95</b>	<b>99/97</b>	<b>100/98</b>
$\hat{\alpha}$			10(99)	20(61)	20(116)	20(202)	30(199)	40(187)	60(148)	80(136)
$\hat{\tau}$			20(50)	20(49)	20(100)	20(202)	20(298)	40(187)	50(157)	40(266)

that the complete pairwise likelihood is not always the best solution, as observed by Sang and Genton (2014), while for  $q = 3, 4, 5$  the best is at  $t = 50\%, 70\%, 90\%$ . Therefore, truncating  $C_{z_q}(Q)$  slightly improves the estimation when  $q$  is small, but for higher orders, complete composite likelihoods give better results. In the last two rows, we also report the smallest truncation factor of  $t$  for the  $q$ -set that corresponds to the best efficiency result with the best  $(q - 1)$ -set and in parentheses the ratio ( $\times 100$ ) of the elapsed times (averaged across experiments) between these two combinations. For  $q = 3, 4$ , an optimal choice of the truncation for  $q - 1$  allows for faster likelihood evaluation than considering  $q$ -sets. This is not the case when  $q \geq 5$ , but the efficiency gain at these cardinalities is considerably smaller than for small  $q$ 's. For the estimation of  $\tau$ , the truncation improves estimation when subsets are of size  $q = 2-5, 8, 9$ , although the optimal estimators for a given  $q$  are never better than the ones for  $q + 1$  at  $t = 100\%$ . From the last row, we see that it is more efficient to choose an optimal  $t$  for a given  $q - 1$  than evaluating the composite likelihood for  $q$  if  $q = 3, 4$ , while the opposite is true for  $q \geq 5$ . We therefore conclude that, for low-order composite likelihoods, truncation can improve the estimation precision while also decreasing the computation, especially for  $\tau$ .

A similar study was carried out for the Brown–Resnick process, with results shown in tables similar to Table 2 and reported in the supplementary material. Similarly to the Reich–Shaby case, the best relative efficiency is achieved once a truncation is performed, and there is a diminished improvement in performance with a truncation of high-order composite likelihoods for both  $\lambda$  and  $\nu$ . Similarly, for low-order composite likelihoods a truncation at  $q - 1$  is computationally more convenient than considering likelihoods of order  $q$ .

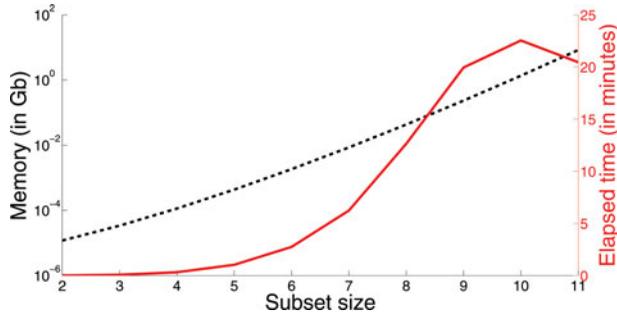


Figure 3. Average memory required (on log scale and in dashed black) and elapsed time (in solid red) across simulations per individual likelihood evaluation. Simulation studies for 1000 experiments, each one with  $m = 50$  replicates, for the logistic model with  $\alpha = 0.6$ .

## 5. COMPUTATIONAL COST OF COMPOSITE LIKELIHOODS

### 5.1 MEMORY AND CPU REQUIREMENTS

As demonstrated in Section 4, high-order composite likelihood estimators generally perform better than pairwise or triplewise approaches, in terms of root mean squared error. This improved performance, however, has a computational cost, which increases rapidly with the composite likelihood order  $q$ .

To illustrate this fact, Figure 3 displays the computational requirements of complete composite likelihood estimators as a function of the order  $q$ , for the simulation study described in Section 4.2 for the logistic model with  $\alpha = 0.6$ . The figure shows the memory required to store the cell array containing all partitions,  $\mathbf{P}_q$ , and the elapsed time (averaged across experiments) for different values of  $q$ . This shows that the required memory, which does not depend on the model, is the main limitation for the use of high-order composite likelihoods; in comparison, for the logistic model, the computational complexity is not as problematic (but it could be for other models). Composite likelihoods of order  $q$  require storing  $\mathbf{P}_q$  containing  $B_q$  elements and cannot be larger than  $B_{11} = 678,570$  with our facilities (since a copy of  $\mathbf{P}_q$  had to be created for each of the 780 CPUs used for parallel computing), although if one uses shared memory or if the goal is to analyze a single dataset,  $q$  could be 12 or 13.

A possible solution could be to use the divide-and-conquer strategy by splitting the computation for all partitions according to some criterion, the most natural being the number of sets within the partition (so-called Stirling partitions, see Graham, Knuth, and Patashnik 1988). This would mitigate the storage problem, but the dynamical computation of the subsets has proven to be, in our experience, very inefficient. This suggests that even with an efficient coding on a powerful computer, the full likelihood on a single dataset may be computed in dimensions  $Q = 12$  or  $Q = 13$ , and similarly, that composite likelihoods are limited to order  $q = 13$ . Given the very steep increase in memory usage with  $q$ , these estimated upper bounds should not change noticeably in coming years.

Figure 3 also reveals an interesting fact concerning complete composite likelihoods: the computational time does not always increase monotonically with subset size. The full likelihood approach is less demanding than 9-sets or 10-sets because an evaluation with 9-

Table 3. Estimated elapsed time per likelihood evaluation on a single CPU, for different values of  $q$  and  $Q$  for  $m = 50$  in the logistic case (s = seconds, m = minutes, h = hours, d = days). When the required time is more than 1 day, the required truncation (in %) to decrease the computational time to this threshold are in parentheses. Estimated times over 1 month are indicated as  $> 30d$ , and below 1 sec as  $< 1s$

$Q \backslash q$	2	3	4	5	6	7	8
11	<1s	<1s	1s	2s	4s	10s	19s
15	<1s	<1s	3s	13s	44s	3m	12m
20	<1s	1s	8s	1m	6m	40m	4h
50	<1s	14s	6m	2h	2d(62)	>30d(3)	>30d(0.14)
100	2s	2m	2h	4d(27)	>30d(0.83)	>30d(0.02)	>30d(0)
500	54s	4h	>30d(2)	>30d(0.01)	>30d(0)	>30d(0)	>30d(0)
1000	4m	1d(71)	>30d(0.13)	>30d(0)	>30d(0)	>30d(0)	>30d(0)
5000	2h	>30d(0.56)	>30d(0)	>30d(0)	>30d(0)	>30d(0)	>30d(0)
10,000	6h	>30d(0.07)	>30d(0)	>30d(0)	>30d(0)	>30d(0)	>30d(0)
100,000	25d(4)	>30d(0)	>30d(0)	>30d(0)	>30d(0)	>30d(0)	>30d(0)

or 10-sets requires the computation of  $\binom{11}{9} \times B_9$  or  $\binom{11}{10} \times B_{10}$  terms in (8), respectively, which is more computationally demanding than summing over  $B_{11}$  terms once.

## 5.2 PROJECTIONS FOR HIGHER DIMENSIONS

While the results discussed in Section 4 and the computational limitations in the previous subsection are important to understand the benefits of high-order likelihood inference, they do not provide insights on how to perform an analysis with a large dataset. Although an analysis with  $q$  larger than 13 is not feasible with current computers, an important question concerns computational requirements for composite likelihood evaluation with relatively small  $q$  but large  $Q$ . Table 3 shows the estimated elapsed time per likelihood evaluation on a single CPU in the setting of Section 4.2 in the logistic case (a similar table for the Reich-Shaby model can be found in the supplementary material), with  $m = 50$  and  $\alpha = 0.6$ , for different values of  $q$  and  $Q$ . The elapsed time  $e_{q,Q}$  is computed directly if it is less than 25 sec. For larger computational times, we estimate it as  $e_{q,\tilde{Q}} \times \frac{|C_{z_q}(Q)|}{|C_{z_q}(\tilde{Q})|}$ , where  $\tilde{Q}$  is the largest  $Q$  for which the computation requires less than 25 sec. Since increasing  $Q$  will increase the size of  $C_{z_q}(Q)$ , but not the computational cost per element, we believe this simple extrapolation provides a realistic indication of the projected time for higher dimensions. Whenever the projected time is more than 1 day, it is clearly not feasible to perform a likelihood maximization, and we therefore also report the truncation proportion  $t$  (recall Section 4.3) needed to reduce the projected time to 1 day.

It is clear from the results that when  $Q$  is large, the complete composite likelihood evaluation becomes problematic, even for small orders  $q$ . By contrast, when  $Q = 11, 15, 20$ , it is possible to evaluate composite likelihoods with  $q = 8$  in a relatively short time without truncation. As the dimension  $Q$  increases, a truncation may be necessary to reduce the computational time (especially for high  $qs$ ), and for very large dimensions ( $Q = 10,000, 100,000$ ), pairwise likelihood seems to be the only viable solution. This table was obtained with a single 2.4GHz processor, and the results should not improve significantly with multiple processors. This is because  $|C_{z_q}(Q)|$  becomes very large for even small values of  $Q$ , therefore

requiring a very large number of fast independent operations. Under this setting, parallel computing will likely perform worse than a single CPU.

### 5.3 RECOMMENDATIONS TO PRACTITIONERS

We now summarize some findings, which may be helpful to practitioners, who have to find the best compromise between statistical and computational efficiency:

- Truncation decreases the computational time and also improves the statistical efficiency for low-order composite likelihoods.
- Computational time may also be decreased by providing good starting values to maximize more efficient higher-order composite likelihoods, for example, by using low-order composite likelihood estimates as starting values.
- For small dimensional datasets ( $Q = 11, 15, 20$ ), a high-order composite likelihood is possible to compute and improves the inference, although there is a diminished return as  $q$  increase and the memory usage should always be monitored. For high-dimensional datasets ( $Q > 20$ ), composite likelihoods can be computed only for relatively small  $q$ 's (unless a hard truncation is applied).
- When the dimensionality is too high for a complete composite likelihood evaluation, as it is likely the case in typical environmental applications, a truncation is absolutely needed. The choice of the truncation proportion will likely be dictated from computational convenience rather than parameter estimate optimality, but truncated high-order composite likelihoods also result in better performance than a complete low-order composite likelihood.

## 6. DISCUSSION

In this work, we have tackled the challenging problem of inference for max-stable processes from a computational perspective and have explored the limits of likelihood-based inference. We have shown that high-performance computing can be a powerful tool to perform efficient inference for multivariate or spatial extremes, but even with a large memory and an efficient use of computational resources, full likelihood inference seems to be limited to dimension  $Q = 12$  or  $Q = 13$  with current technologies; for higher dimensions, composite likelihoods with suitably selected components are an efficient alternative, although, for the same reasons, they are limited to order  $q = 12$  or  $q = 13$ . With extensive simulation studies based on three increasingly complex classes of multivariate or spatial extremes, we have quantified the loss of efficiency of widely used pairwise and triplewise likelihood approaches and found that high-order composite likelihoods can lead to substantial improvements in root mean squared error. Furthermore, we have shown that truncation of composite likelihoods not only reduces the computational burden, but in some cases also leads to better estimation of the parameters, especially for low-order composite likelihoods, confirming results obtained in previous studies. Finally, we have also given guidance on the choice of subsets to include into truncated composite likelihoods, advocating a simpler approach based on the maximum subset distance. This choice is, however, not optimal and

other possibilities could be to choose the subsets based on the average, or minimal, subset distance.

We do not expect that computers with contemporary architectures will render the “full likelihood problem” easy to solve for large dimensions in the foreseeable future: an increase in available processors would only linearly decrease the computational time, and memory storage is not foreseen to increase enough to allow the storage of, say, data structures with  $B_{30} = O(10^{23})$  elements. We therefore conclude that a direct full likelihood approach is not feasible unless important methodological advances are made to deal with this problem, and truncated composite likelihoods with subsets of size  $q = 4\text{--}6$  provide an efficient and relatively cheap solution in the settings we have examined. For very large datasets, however, the only solution might be to use pairwise likelihoods.

Given the computationally demanding nature of these approaches, it is natural to wonder whether other likelihoods can be derived from alternative representations of extreme events. We have assumed that max-stable data (typically pointwise maxima of random processes) are the only information available for inference. However, if additional information, such as the times of occurrence of maxima or the original processes, is incorporated, considerably more efficient strategies based on point processes can be devised; see Stephenson and Tawn (2005), Wadsworth and Tawn (2014), and Thibaud et al. (2015). Alternatively, recent advances in the simulation of max-stable processes (Dieker and Mikosch 2015; Dombry, Engelke, and Oesting 2016) suggest that simulation-based likelihood inference might be an alternative (Koch 2014).

In this work, we focused on three max-stable parametric families, but the considerations on computational feasibility are the same for every process whose cumulative distribution function has the form (1). Furthermore, we believe that similar efficiency results should be expected for composite likelihoods applied to other max-stable models, such as the Schlather (2002) or the extremal  $t$  model (Nikoloulopoulos, Joe, and Li 2009; Opitz 2013), under similar scenarios. It would also be worth investigating the case of multivariate max-stable processes (Oesting, Schlather, and Friederichs 2016; Genton, Padoan, and Sang 2015), or asymptotically independent models (Wadsworth and Tawn 2012), where the rate of decay toward independence has to be estimated.

## SUPPLEMENTARY MATERIALS

**Supplements:** Correlation between  $\hat{\alpha}$  and  $\hat{\tau}$  in the Reich–Shaby model; Performance of truncated composite likelihood estimators for the Brown–Resnick model; Computational time projections for the Reich–Shaby model; Partition-based likelihood approximation; Closed form expression of the partial derivatives for the three models. (SupplementsCHG.pdf)

**Code:** Simulation code. (CodeCHG.zip)

*[Received June 2015. Revised August 2015.]*

## REFERENCES

- Ballani, F., and Schlather, M. (2011), “A Construction Principle for Multivariate Extreme Value Distributions,” *Biometrika*, 98, 633–645. [1214]
- Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012), “Estimating Space and Space-Time Covariance Functions for Large Data Sets: A Weighted Composite Likelihood Approach,” *Journal of the American Statistical Association*, 107, 268–280. [1213]
- Bienvenüe, A., and Robert, C. Y. (2016), “Likelihood Based Inference for High-dimensional Extreme Value Distributions,” arXiv:1403.0065v3. [1213,1218]
- Brown, B. M., and Resnick, S. I. (1977), “Extreme Values of Independent Stochastic Processes,” *Journal of Applied Probability*, 14, 732–739. [1212,1216]
- Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, New York: Springer. [1214]
- Cooley, D., Davis, R. A., and Naveau, P. (2010), “The Pairwise Beta Distribution: A Flexible Parametric Multivariate Model for Extremes,” *Journal of Multivariate Analysis*, 101, 2103–2117. [1214]
- Cox, D. R., and Reid, N. (2004), “A Note on Pseudolikelihood Constructed From Marginal Densities,” *Biometrika*, 91, 729–737. [1213]
- Davis, R., and Yau, C. (2011), “Comments on Pairwise Likelihood in Time Series Models,” *Statistica Sinica*, 21, 255–278. [1213]
- Davison, A. C., and Huser, R. (2015), “Statistics of Extremes,” *Annual Review of Statistics and its Application*, 2, 203–235. [1214]
- Davison, A. C., Huser, R., and Thibaud, E. (2013), “Geostatistics of Dependent and Asymptotically Independent Extremes,” *Mathematical Geosciences*, 45, 511–529. [1212]
- Davison, A. C., Padoan, S. A., and Ribatet, M. (2012), “Statistical Modeling of Spatial Extremes,” *Statistical Science*, 27, 161–186. [1212]
- Dieker, A., and Mikosch, T. (2015), “Exact Simulation of Brown–Resnick Random Fields at a Finite Number of Locations,” *Extremes*, 18, 301–314. [1227]
- Dombry, C., Engelke, S., and Oesting, M. (2016), “Exact Simulation of Max-Stable Processes,” *Biometrika*, 103, in press. [1227]
- Eidsvik, J., Shaby, B. A., Reich, B. J., Matthew, W., and Niemi, J. (2014), “Estimation and Prediction in Spatial Models With Block Composite Likelihoods,” *Journal of Computational and Graphical Statistics*, 23, 295–315. [1213]
- Genton, M. G., Ma, Y., and Sang, H. (2011), “On the Likelihood Function of Gaussian Max-Stable Processes,” *Biometrika*, 98, 481–488. [1213,1218]
- Genton, M. G., Padoan, S., and Sang, H. (2015), “Multivariate Max-Stable Spatial Processes,” *Biometrika*, 102, 215–230. [1227]
- Genz, A. (1992), “Numerical Computation of Multivariate Normal Probabilities,” *Journal of Computational and Graphical Statistics*, 1, 141–149. [1221]
- Genz, A., and Bretz, F. (2002), “Methods for the Computation of Multivariate t-Probabilities,” *Journal of Computational and Graphical Statistics*, 11, 950–971. [1221]
- Genz, A., and Bretz, F. (2009), *Computation of Multivariate Normal and t Probabilities*, Berlin: Springer. [1221]
- Graham, R., Knuth, D. E., and Patashnik, O. (1988), *Concrete Mathematics*, Reading, MA: Addison-Wesley. [1218,1224]
- Gumbel, E. J. (1960), “Bivariate Exponential Distributions,” *Journal of the American Statistical Association*, 55, 698–707. [1213,1215]
- Hjort, N., and Varin, C. (2008), “ML, PL, QL in Markov Chain Models,” *Scandinavian Journal of Statistics*, 35, 64–82. [1213]
- Huser, R., and Davison, A. C. (2013), “Composite Likelihood Estimation for the Brown-Resnick Process,” *Biometrika*, 100, 511–518. [1213,1218,1220,1221]

- Huser, R., Davison, A. C., and Genton, M. G. (2016), “Likelihood Estimators for Multivariate Extremes,” *Extremes*, 19, 79–103. [1213]
- Joe, H. (1997), *Multivariate Models and Dependence Concepts*, London: Chapman & Hall. [1214]
- Kabluchko, Z., Schlather, M., and de Haan, L. (2009), “Stationary Max-Stable Fields Associated to Negative Definite Functions,” *The Annals of Probability*, 37, 2042–2065. [1212,1216]
- Koch, E. (2014), “Estimation of Max-Stable Processes by Simulated Maximum Likelihood,” unpublished manuscript. [1227]
- Lindsay, B. (1988), “Composite Likelihood Methods,” *Contemporary Mathematics*, 80, 220–239. [1213]
- Nikoloulopoulos, A. K., Joe, H., and Li, H. (2009), “Extreme Value Properties of Multivariate  $t$  Copulas,” *Extremes*, 12, 129–148. [1212,1227]
- Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015), “A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets,” *Journal of Computational and Graphical Statistics*, 24, 579–599. [1213]
- Oesting, M., Schlather, M., and Friederichs, P. (2016), “Statistical Post-Processing of Forecasts for Extremes Using Bivariate Brown-Resnick Processes with an Application to Wind Gusts,” arXiv:1312.4584v2. [1227]
- Opitz, T. (2013), “Extremal  $t$  Processes: Elliptical Domain of Attraction and a spectral representation,” *Journal of Multivariate Analysis*, 122, 409–413. [1212,1227]
- Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010), “Likelihood-Based Inference for Max-Stable Processes,” *Journal of the American Statistical Association*, 105, 263–277. [1213,1218,1219,1222]
- Reich, B. J., and Shaby, B. A. (2012), “A Hierarchical Max-Stable Spatial Model for Extreme Precipitation,” *The Annals of Applied Statistics*, 6, 1430–1451. [1214,1216]
- Sang, H., and Genton, M. G. (2014), “Tapered Composite Likelihood for Spatial Max-Stable Models,” *Spatial Statistics*, 8, 86–103. [1213,1218,1219,1222]
- Schlather, M. (2002), “Models for Stationary Max-Stable Random Fields,” *Extremes*, 5, 33–44. [1212,1213,1227]
- Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., and Strokorb, K. (2015), “Analysis, Simulation and Prediction of Multivariate Random Fields with Package RandomFields,” *Journal of Statistical Software*, 62, 1–25. [xxxx]
- Segers, J. (2012), “Max-Stable Models for Multivariate Extremes,” *REVSTAT*, 10, 61–82. [1214]
- Smith, R. L. (1990), “Max-Stable Processes and Spatial Extremes,” unpublished manuscript. [1212,1216,1217]
- Stephenson, A., and Tawn, J. A. (2005), “Exploiting Occurrence Times in Likelihood Inference for Component-wise Maxima,” *Biometrika*, 1, 213–227. [1227]
- Stephenson, A. G. (2009), “High-Dimensional Parametric Modelling of Multivariate Extreme Events,” *Australian & New Zealand Journal of Statistics*, 51, 77–88. [1216]
- Sun, Y., Li, B., and Genton, M. G. (2012), “Geostatistics for Large Datasets,” in *Space-Time Processes and Challenges Related to Environmental Problems* (Vol. 207), eds. E. Porcu, J. M. Montero, M. Schlather, Berlin: Springer, 55–77. [1213]
- Sun, Y., and Stein, M. L. (2016), “Statistically and Computationally Efficient Estimating Equations for Large Spatial Datasets,” *Journal of Computational and Graphical Statistics*, 25, 187–208. [1213]
- Tawn, J. A. (1988), “Bivariate Extreme Value Theory: Models and Estimation,” *Biometrika*, 75, 397–415. [1214]
- (1990), “Modelling Multivariate Extreme Value Distributions,” *Biometrika*, 77, 245–253. [1214,1216]
- Thibaud, E., Aalto, J., Cooley, D. S., Davison, A. C., and Heikkinen, J. (2015), “Bayesian Inference for the Brown-Resnick Process, With an Application to Extreme Low Temperatures,” arXiv:1506.07836. [1227]
- Varin, C., Reid, N., and Firth, D. (2011), “An Overview of Composite Likelihood Methods,” *Statistica Sinica*, 21, 5–42. [1213]
- Wadsworth, J. L., and Tawn, J. A. (2012), “Dependence Modelling for Spatial Extremes,” *Biometrika*, 99, 253–272. [1227]
- (2014), “Efficient Inference for Spatial Extreme Value Processes Associated to Log-Gaussian Random Functions,” *Biometrika*, 101, 1–15. [1227]