Bayesian Model Averaging for multivariate extremes

Anne Sabourin, Naveau Philippe, Fougères Anne-Laure

BAYESIAN MODEL AVERAGING FOR MULTIVARIATE EXTREMES

A. SABOURIN, P. NAVEAU, AND A.-L. FOUGÈRES

Abstract. The main framework of multivariate extreme value theory is well-known in terms of probability, but inference and model choice remain an active research field. Theoretically, an angular measure on the positive quadrant of the unit sphere can describe the dependence among very high values, but no parametric form can entirely capture it. The practitioner often makes an assertive choice and arbitrarily fits a specific parametric angular measure on the data. Another statistician could come up with another model and a completely different estimate. This leads to the problem of how to merge the two different fitted angular measures. One natural way around this issue is to weigh them according to the marginal model likelihoods. This strategy, the so-called Bayesian Model Averaging (BMA), has been extensively studied in various context, but (to our knowledge) it has never been adapted to angular measures. The main goal of this article is to determine if the BMA approach can offer an added value when analyzing extreme values.

keywords: Bayesian model averaging, multivariate extremes, parametric modelling, spectral measure.

MSC: 62F07 and 62F15 and 62H20 and 62H05 and 62P12

1. Introduction

Assessing the probability of occurrence of joint extreme events has proven to be a major issue for risk management and a complex inferential problem in statistics. To illustrate this point, daily maximum concentrations of three air pollutants, PM10 (Particulate matter), NO (Nitrogen oxide) and NO_2 (Nitrogen dioxide), recorded in Leeds (U.K.) during five winter seasons (1994-1998)[1], are displayed in Figure 1. Visually, asymmetrical relationships seem to be present, the dependence between NO2 and NO may be stronger than that between the two other pairs. For this Leeds data set, at least three different approaches (Cooley et al., 2010; Heffernan and Tawn, 2004; Boldi and Davison, 2007) have already been proposed. Heffernan and Tawn (2004)'s study focuses on conditional distributions, allowing both for asymptotic dependence or independence at extreme levels. On the contrary, Cooley et al. (2010) and Boldi and Davison (2007), under the assumption of asymptotic dependence, characterize the joint distribution of extremes in terms of the so-called *angular measure* (see Section 2 for more details), respectively in a parametric and semi-parametric framework. In this paper, we follow this latter approach, and focus on parametric models. Several such models have already been proposed for the case where the data are dependent at asymptotic levels: see *e.g.* chapter 9 of Beirlant et al. (2004); Tawn (1990) or Coles and Tawn (1991) for the Logistic

Date: Preprint version of the article published in *Extremes*, 2013. Received: 23 May 2012 / Revised: 14 November 2012 / Accepted: 22 November 2012.

[1]Five different air pollutant concentrations (PM10, NO, NO_2, O3, and SO_2) were measured in the city centre of Leeds, see `http://www.airquality.co.uk` for more details. We restrict our analysis to the three most dependent pollutants.

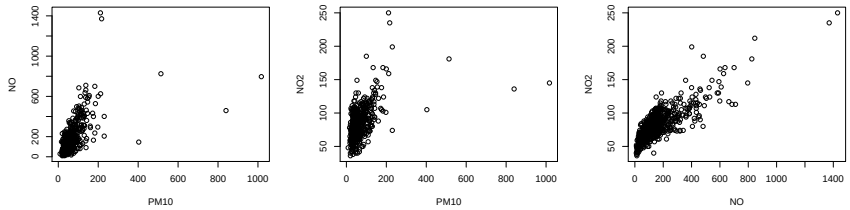

FIGURE 1. Daily maximum concentrations of three air pollutants, PM10, NO and NO_2, recorded in Leeds (U.K.) during five winter seasons (1994-1998)

and Dirichlet families; or Cooley et al. (2010) for the Pairwise Beta model, further generalized by Ballani and Schlather (2011).

In this context, two practitioners working on the same data may well have chosen two distinct models, leading to different estimates of some quantity of interest such as a probability of joint excess of some high multivariate threshold. One could thus reasonably ask if it would be appropriate to merge these results. Our main objective throughout this paper is to investigate how to average the estimates issued from existing parametric families and what are the benefits and the limitations of such an approach.

The Bayesian framework appears to be well tailored for this task because setting priors offers a natural way to integrate results issued from different studies. The so-called *Bayesian Model Averaging* (BMA) method has been extensively studied in other contexts (e.g., Raftery et al., 2005; Madigan and Raftery, 1994; Hoeting et al., 1999). Its adaption to the analysis of multivariate extreme events represents the main aim of this work. To our knowledge, in the field of multivariate extremes, the only publication using BMA is Apputhurai and Stephenson (2011). They combined the cumulative distribution functions of asymptotically dependent and independent models, in the bi-variate case. Our approach differs from theirs in focusing on asymptotically dependent models, and combining the dependence structures themselves (angular measures or exponent functions, see Section 2 for definitions and rationale for such a choice).

In the next section, we recall the necessary background about multivariate extremes. In Section 3, we detail the BMA nuts and bolts within a multivariate extremes context. The BMA scheme is implemented in Section 4 with two different models: the Pairwise Beta model (Cooley et al., 2010) and a nested asymmetric logistic model. A simulation study is performed: data sets are generated from a semi-parametric *Dirichlet mixture model* (DM) introduced by Boldi and Davison (2007) and we compare the predictive performance of the BMA versus a model choice framework. The tri-variate Leeds data set is also revisited. Section 5 offers a few conclusions regarding the advantages and limitations of averaging spectral measures.

2. Background and notations

Spectral measure.
Let $\mathbf{X} = (X_1, \ldots, X_d)^T$ be a positive random vector of dimension d whose margins

follow a unit Fréchet distribution, $\mathbf{P}(X_i \leq x) = \exp(-1/x)$, for all $x > 0$. To describe the extremal behaviour of the vector $\mathbf{X}$, it is mathematically convenient to transform the Cartesian coordinates into pseudo-polar ones by setting

$$R = X_1 + \cdots + X_d \qquad \text{and} \qquad \mathbf{W} = (X_1/R, \ldots, X_d/R)^T$$

where R and $\mathbf{W}$ are often called the radius and the angular vector, respectively. The latter one lies on the unit simplex $\mathbf{S}_d = \{\mathbf{w} : \ w_1 + \cdots + w_d = 1, \quad w_i > 0\}$. With regards to the Leeds data set, we follow the exact same procedure as Cooley et al. (2010) to estimate the marginal distributions of the three pollutants plotted in Figure 1. Each uni-variate series can thus easily be transformed into unit Fréchet distributed ones *via* a probability integral transformation. Observations with the 100 largest radial components [2] (out of 539 non missing triplets) are plotted on the unit simplex $\mathbf{S}_3$ in Figure 2. The points located at the centre of this triangle correspond to events that were equally extreme in the three directions.

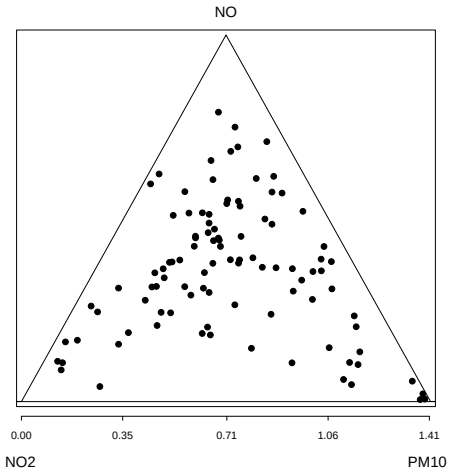

FIGURE 2. Leeds data set: the 100 points with largest radial component $R = x_1 + x_2 + x_3$ (unit Fréchet scale) projected on the unit simplex.

Multivariate extreme value theory tells us that, under mild conditions[3], the dependence structure among excesses above a high radial threshold r can be characterized by the asymptotic distribution H of the angular component:

$$\lim_{r \to \infty} \mathbf{P}\left(\mathbf{W} \in B \mid R > r\right) = H(B)\ , \tag{1}$$

[2]Besides the $\mathbf{L}_1$-norm ($x_1 + x_2 + x_3$), other norms could be used for threshold selection.

[3] The largest values have to belong to the domain of attraction of a max-stable distribution (the distribution G is said to be max-stable if $G^t(t\mathbf{x}) = G(\mathbf{x})$ for any $t > 0$). This type of distribution arises as the natural non-degenerate limit of rescaled i.i.d. component wise maxima of random vectors with unit Fréchet margins (de Haan and Ferreira, 2006; Resnick, 1987, 2007). Within this framework, it is classical to define the exponent function $V(\mathbf{x}) = -\log G(\mathbf{x})$ that satisfies $V(t\mathbf{x}) = t^{-1}V(\mathbf{x})$.

The *spectral measure* $H(.)$ is any probability measure on the simplex $\mathbf{S}_d$ that satisfies the following moment constraint

$$\forall i \in \{1, \ldots, d\}, \int_{\mathbf{S}_d} w_i \, \mathrm{d}H(\mathbf{w}) = \frac{1}{d} \, . \tag{2}$$

Limit measure.
With our normalization choice, the spectral measure is related to a *limit measure* ν, defined on $\mathbf{E} = [\mathbf{0}, \infty]^d \setminus \{\mathbf{0}\}$, in pseudo-polar coordinates, by $\mathrm{d}\nu = \frac{d}{r^2} \, \mathrm{d}r \, \mathrm{d}H$ (see *e.g.* Chapter 6 of Resnick, 2007). The measure ν is homogeneous of order -1, *i.e.* for any measurable subset $A \subset \mathbf{E}$, $\nu(tA) = \frac{1}{t}\nu(A)$. If A is relatively compact in $\mathbf{E}$,

$$\lim_{n \to \infty} n\mathbf{P}\left(\frac{\mathbf{X}}{n} \in A\right) = \nu(A). \tag{3}$$

In particular, (3) holds for any failure set A of the form $A(\mathbf{u}) = \{\mathbf{x} : \ x_1 > u_1, \ldots, x_d > u_d\}$.

Modelling threshold excesses.
Equations (1) and (3) provide the main elements for modelling excesses in practice. Given a data set whose margins have been transformed into unit Fréchet, one may fix a high radial threshold r_0 and retain only observations with radial component exceeding r_0. The corresponding angular data set $\mathscr{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$, as in Figure 2 with $n = 100$ excesses, is assumed to be an i.i.d. sample distributed according to $H(.)$. Then, the statistician has to propose and fit an adequate spectral measure.

In other words, all the inference in this paper is based on the following key assumption, in view of (3): Conditionally on the radial component R exceeding the retained threshold r_0, the random vector X is assumed to be distributed according to some (normalized) limit probability measure $\tilde{\nu}$, with, in polar coordinates, $\mathrm{d}\tilde{\nu} = \frac{d}{r_0} \frac{\mathrm{d}r}{r^2} \, \mathrm{d}H$. The angular and the radial components are thus assumed to be independent on regions $\{r > r_0\}$, and H characterizes $\tilde{\nu}$, so that a a statistical model for excesses above r_0 can be indexed by a set of angular measures. Also, the likelihood is proportional to the density h evaluated at the angular data points $\mathcal{W}$ and inference can be made with the angular components only. This assumption of 'perfect threshold' has a second consequence: The likelihood of an angular measure which mass is concentrated on the axis is zero, because all the angular data points lie in the interior of the positive quadrant. This restricts any likelihood-based inference to asymptotically dependent models, *i.e.* to H-families which put some mass in the interior of the unit simplex only.

Relaxing the 'perfect threshold' assumption is possible if one works with max-stable distributions, but then, the link with questions related to excesses above threshold is not immediate. Another reason for not considering this option in the context of model averaging is the fact that an average of max-stable distributions is not max-stable. More details are given at the end of this section.

Further, one may consider asymptotically independent models with second-order regular variation in the interior of the positive quadrant (Ledford and Tawn, 1996; Ramos and Ledford, 2009). However, flexible parametric models for asymptotically independent data, in problems of dimension greater than three, have only

recently been proposed in an unpublished paper from Qin *et. al* (2008) [4]. For the sake of simplicity, we leave apart this class of models and focus on asymptotically dependent data.

If a vector $\mathbf{u} = (u_1, \ldots, u_d)^T$ defines the boundary of a failure region $A(\mathbf{u}) = \{\mathbf{x} : \ x_1 > u_1, \ldots, x_d > u_d\}$ is such that $\sum_{i=1}^{d} u_i > r_0$ for some large r_0, using (3) and the homogeneity property of the limit measure, the probability of being in the failure region can be approximated with ν:

$$\begin{aligned} \mathbf{P}(X_1 > u_1, \ldots, X_d > u_d) \simeq \nu\big(A(\mathbf{u})\big) &= d \int_A \frac{1}{r^2} \,\mathrm{d}r \,\mathrm{d}H(\mathbf{w}) \\ &= d \int_{\mathbf{S}_d} \int_{r > \max_{i=1:d} \frac{u_i}{w_i}} \frac{1}{r^2} \,\mathrm{d}r \,\mathrm{d}H(\mathbf{w}) \\ &= d \int_{\mathbf{S}_d} \min_{i=1:d} \frac{w_i}{u_i} \,\mathrm{d}H(\mathbf{w}) \end{aligned} \tag{4}$$

A classical way of proposing parametric max-stable models is to define them through their *exponent function* V, (see footnote 3), which is related to ν by

$$\forall \mathbf{x} \in \mathbf{E}, \ V(\mathbf{x}) = \nu\left\{([0, x_1] \times \cdots \times [0, x_d])^c\right\} \ ,$$

were $(\,\cdot\,)^c$ denotes the complementary set in $\mathbf{E}$. In the case where all the mass of the angular measure H is concentrated in the interior of the simplex $\mathbf{S}_d$, and V is regular, Theorem 1 of Coles and Tawn (1991) provides a general relationship to derive the spectral density $h(\mathbf{w})$ from $V(.)$: $h(\mathbf{w}) = -\frac{1}{d}\, \partial_{x_1,\ldots,x_d} V(\mathbf{x})|_{\mathbf{x}=\mathbf{w}}$. One advantage of such models is that (4) has an analytical expression obtained from V by inclusion-exclusion. For three-dimensional sample spaces, it yields

$$\begin{aligned} \nu\big(A(\mathbf{u})\big) = &V(u_1, u_2, u_3) + \cdots \\ &\cdots V(u_1, \infty, \infty) + V(\infty, u_2, \infty) + V(\infty, \infty, u_3) - \ \cdots \\ &\cdots (V(u_1, u_2, \infty) + V(u_1, \infty, u_3) + V(\infty, u_2, u_3)) \ . \end{aligned} \tag{5}$$

One drawback is that the angular likelihood h has to be computed by differentiation of order d.

Multivariate extreme models.
In theory, the only constraint on H is encapsulated by (2), which advocates in favour of fully non-parametric estimation methods (see *e.g.* Einmahl et al., 2001; Einmahl and Segers, 2009; Guillotte et al., 2011; Gudendorf and Segers, 2011). In a Bayesian context, it is computationally difficult to handle moderate dimension problems with semi-parametric spectral measures. For example, Boldi and Davison (2007) introduced a semi parametric Bayesian model based on mixtures of Dirichlet distributions and concluded that "one practical drawback with the approach stems from the use of simulation algorithms, which may converge slowly unless they have been tuned. A second is that the number of parameters increases rapidly with the number of mixture components, so model complexity must be sharply penalized through an information criterion or a prior on the number of mixture components".

[4] Qin X., Smith R.L., Ren R.E., Modelling multivariate extreme dependence, In 2008 Joint Statistical Meetings(JSM) Proceedings, Risk Analysis Section. Alexandria, VA: American Statistical Association: 3089-3096

From a practical point of view, parameters in some well-chosen models may offer interpretable summaries to describe the dependence structure (*e.g.*, a finite number of parameters allows to compare between two time periods), and a feasible strategy to reduce the computational complexity. A seminal example (Gumbel, 1960) of parametric model defined by the exponent function is the logistic one

$$V_{\mathrm{L}}(\mathbf{x}) = \left(\sum_{i=1}^{d} x_i^{-1/\alpha} \right)^{\alpha} \quad (0 < \alpha < 1) \, .$$

The logistic model can be extended to handle asymmetrical behaviours and to capture additional dependencies among subsets of variables (Coles and Tawn, 1991). In particular, the exponent function

$$\begin{aligned} V_{\mathrm{NL}}(\mathbf{x}) = 2^{-\alpha_0} \Big[& \left(x_1^{\frac{-1}{\alpha_0 \alpha_{12}}} + x_2^{\frac{-1}{\alpha_0 \alpha_{12}}} \right)^{\alpha_{12}} + \left(x_1^{\frac{-1}{\alpha_0 \alpha_{13}}} + x_3^{\frac{-1}{\alpha_0 \alpha_{13}}} \right)^{\alpha_{13}} + \cdots \\ & \left(x_2^{\frac{-1}{\alpha_0 \alpha_{23}}} + x_3^{\frac{-1}{\alpha_0 \alpha_{23}}} \right)^{\alpha_{23}} \Big]^{\alpha_0} \quad (0 < \alpha_0 \, , \alpha_{12}, \alpha_{13}, \alpha_{23} < 1) \, , \end{aligned} \tag{6}$$

is a possible generalization which allows for asymmetric pairwise relationships, while concentrating all its mass in the interior of $\mathbf{S}_3$. This model belongs to the class of the *nested asymmetric logistic models*. In the remainder of this paper, we refer to the one defined by (6) as the NL model, we denote $\underline{\alpha} = (\alpha_0, \alpha_{12}, \alpha_{13}, \alpha_{23})$. The expression for the NL density $h_{\mathrm{NL}}(\, \cdot \, | \underline{\alpha})$ on $\mathbf{S}_3$ is given in appendix. One advantage of NL is its low number of parameters and in their interpretability. The scalar α_0 describes the overall dependence among the three coordinates and the α_{ij}'s characterize the additional pairwise dependences. The dependence between a coordinates subset is a decreasing function of the corresponding parameter.

It is also possible to define a multivariate extreme model directly through the spectral density. For example, Cooley et al. (2010) fitted to the Leeds data set the following Pairwise Beta (PB) model

$$h_{\mathrm{PB}}(\mathbf{w} | \beta_0, \{\beta_{ij}\}_{1 \leq i < j \leq d}) = \sum_{1 \leq i < j \leq d} h_{ij}(\mathbf{w} | \beta_0, \beta_{ij}) \quad (\beta_0, \beta_{ij} > 0) \, ,$$

which is a sum of beta functions defined by

$$h_{ij}(\mathbf{w} | \beta_0, \beta_{ij}) \;\; = \;\; K_d(\beta_0) \;\; w_{ij}^{2\beta_0 - 1} \, (1 - w_{ij})^{(d-2)\beta_0 - d + 2} \, \frac{\Gamma(2\beta_{ij})}{\Gamma^2(\beta_{ij})} w_{i/ij}^{\beta_{ij}-1} w_{j/ij}^{\beta_{ij}-1}$$

with $w_{ij} = w_i + w_j$, $w_{i/ij} = \frac{w_i}{w_i + w_j}$ and $K_d(\beta_0) = \frac{2(d-3)!}{d(d-1)} \frac{\Gamma(\beta_0 d + 1)}{\Gamma(2\beta_0 + 1)\Gamma(\beta_0 (d-2))}$. [5] The interpretation for the parameters in the PB model is similar to the NL model's one, except that the strength of the dependence is an increasing function of β_0 and of the β_{ij}'s. Again, we denote $\underline{\beta} = (\beta_0, \beta_{12}, \beta_{13}, \beta_{23})$.

Having at our disposal several spectral models leads us to the question of how to average them with respect to the data set at hand. First, one could wonder what is the meaning of averaging spectral measures in terms of random variables and if directly averaging the corresponding max-stable *distributions* could be a valuable

[5] The difference of a multiplicative factor $\sqrt{d}$ in our normalizing constant compared to the one given by Cooley et al. (2010) is due to the choice of the reference measure: in the aforementioned study, the reference measure is the Lebesgue measure (more precisely the Hausdorff measure) on the simplex itself, whereas we write our densities with respect to its projection on the $d-1$ dimensional euclidean space.

alternative. However, if the random vector M_j follows a max-stable distribution $G_j(\mathbf{x}) = \exp(-V_j(\mathbf{x}))$ with unit Fréchet margins, then the averaged distribution $G(\mathbf{x}) = p_1 G_1(\mathbf{x}) + \cdots + p_J G_J(\mathbf{x})$ (with $\sum_{j=1}^{J} p_j = 1$) is *not* max-stable anymore: it does not satisfy $G^t(t\mathbf{x}) = G(\mathbf{x})$ for any $t > 0$. In contrast, averaging the angular measures H_j still provides another valid angular measure. Indeed, $p_1 H_1(.) + \cdots + p_J H_J(.)$ satisfies (2). In terms of random vectors, averaging angular measures translates into component-wise max-linear combinations. More precisely, if the M_j's are independent, then the max-linear combination $\tilde{M} = p_1 M_1 \vee \cdots \vee p_J M_J$, where $\vee$ denotes the component-wise maximum, has exponent function $\tilde{V}(\mathbf{x}) = p_1 V_1(\mathbf{x}) + \cdots + p_J V_J(\mathbf{x})$. The latter is associated with the spectral measure $\tilde{H} = p_1 H_1 + \cdots + p_J H_J$. This derives immediately from the homogeneity property $(V_j(t\mathbf{x}) = t^{-1} V_j(\mathbf{x}))$ characterizing exponent functions:

$$\mathbf{P}(\tilde{M} \leq \mathbf{x}) = \mathbf{P}\left(\bigvee_{j=1}^{J} p_j M_j \leq \mathbf{x}\right) = \prod_{j=1}^{J} \mathbf{P}(M_j \leq \frac{\mathbf{x}}{p_j})$$

$$= \prod_{j=1}^{J} \exp\left[-V_j\left(\frac{\mathbf{x}}{p_j}\right)\right] = \exp\left[-\sum_{j=1}^{J} p_j V_j(\mathbf{x})\right] .$$

3. Bayesian Model Averaging for spectral measures

In the general context of parametric modeling, the information loss relative to the choice of one particular model may be high. Averaging the estimates stemming from several models, with appropriate weights, can be used to partially overcome this issue. As an example, Raftery et al. (2005) found some evidence in an ensemble weather forecast context that the predictive variance in one model would sometimes not reflect the total predictive uncertainty, whereas the predictive variance in the averaged model accounted better for prevision errors. Madigan and Raftery (1994) found some examples, in a contingency tables context, where averaging models resulted in better predictive performance, as measured by a logarithmic scoring rule. We recall here the basic BMA features within our spectral measure context. For a review of BMA, the reader may refer to Hoeting et al. (1999).

Suppose we have M spectral density models $\mathcal{M}_1, \ldots, \mathcal{M}_M$, such that each model $\mathcal{M}_m = \{h_m(\cdot \mid \theta_m),\ \theta_m \in \Theta_m\}$ has a finite dimensional parameter space Θ_m. In this paper, for illustrative purpose, we set $M = 2$ and h_1 and h_2 correspond respectively to the aforementioned PB spectral measure family h_{PB} and to the NL one h_{NL}.

In a Bayesian framework, beliefs of the statistician about θ_m (*e.g.*, arising from expert knowledge), prior to any observation, are made explicit: each parameter space Θ_m is endowed with a *prior* measure (in our case, a probability measure), denoted π_m. Now, in a Bayesian model averaging setting, the statistical model $\tilde{\mathcal{M}}$ is the *disjoint union* of the individual models: in other words, the parameter space $\tilde{\Theta}$ indexing $\tilde{\mathcal{M}}$ is the disjoint union $\tilde{\Theta} = \bigsqcup_{m=1}^{M} \Theta_m$. We recall that a disjoint union of sets $A_1, \ldots, A_M$ is defined by $\bigsqcup_{m=1}^{M} A_i = \{(m, a_m), 1 \leq m \leq M, a_m \in A_m\}$. In the sequel, the term 'union model' will refer to the BMA model indexed by the disjoint union $\tilde{\Theta}$. A prior on the index set $\{1, \ldots, M\}$ is thus needed: we choose $(p_1, \ldots, p_M)$, with $p_1 + \cdots + p_M = 1$, so that p_m is the *a priori* weight of $\mathcal{M}_m$. In

this work, lacking expert knowledge, we choose a uniform prior: $p_m = 1/M$ for all m. The prior distribution $\tilde{\pi}$ on $\tilde{\mathcal{M}}$ is

$$\tilde{\pi}(\bigsqcup_{m=1}^{M} B_m) = \sum_{m=1}^{M} p_m \pi_m(B_m) \ ,$$

for any collection of measurable sets $(B_1, \ldots, B_M)$ with $B_m \subset \Theta_m$. Suppose that each model has a well defined spectral density $h_m(\cdot \mid \theta_m)$. Given the sample of excesses $\mathscr{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$, a common density estimator is the *Posterior predictive density* [6] which, in the disjoint union model, is defined by

$$\tilde{h}(\mathbf{w} \mid \mathscr{W}) = \sum_{m=1}^{M} p_m(\mathscr{W}) \int_{\Theta_m} h_m(\mathbf{w} \mid \theta_m) \, d(\pi_m|\mathscr{W})(\theta_m) \ , \tag{7}$$

where $\pi_m|\mathscr{W} = \pi_m(\cdot \mid \mathscr{W})$ is the posterior distribution restricted to $\mathcal{M}_m$, and $p_m(\mathscr{W})$ is the posterior weight of $\mathcal{M}_m$. This explains the terminology "model averaging": our density estimate is the average of the density estimates produced in separate Bayesian models. As mentioned above, the family of admissible densities is stable under convex combinations, and it is also stable under integration with respect to any probability measure. Consequently, the posterior predictive density still defines a valid angular measure.

More generally, if the goal is to estimate some quantity of interest Δ, (a measurable function of $\underline{\theta}_m$ in each model $\mathcal{M}_m$, such as *e.g.* the probability of a failure set, then Δ is a random variable which prior and posterior distributions in each model are respectively the image measures $\Delta^*\pi_m$ and $\Delta^*[\pi_m|\mathscr{W}]$. The posterior distribution in the BMA model is thus the average

$$\Delta^*[\pi|\mathscr{W}] = \sum_{m=1}^{M} p_m(\mathscr{W}) \Delta^*[\pi_m|\mathscr{W}]$$

and the mean estimate is $\hat{\delta} = \sum_{m=1}^{M} p_m(\mathscr{W}) \int_{\Theta_m} \Delta(\theta_m) \, d[\pi_m|\mathscr{W}](\theta_m)$.

The Bayes formula provides immediately the expression for the posterior weights: $p_m(\mathscr{W})$ is proportional to the *marginal* likelihood $\mathcal{L}_m(\mathscr{W})$ of the observed angular sample in each model $\mathcal{M}_m$, multiplied by the corresponding prior model weight

$$p_m(\mathscr{W}) = \frac{p_m \mathcal{L}_m(\mathscr{W})}{p_1 \mathcal{L}_1(\mathscr{W}) + \cdots + p_M \mathcal{L}_M(\mathscr{W})} \ ,$$

where

$$\mathcal{L}_m(\mathscr{W}) = \int_{\Theta_m} h_m(\mathscr{W}|\theta_m) \, d\pi_m(\theta_m) \ . \tag{8}$$

In practice, for high dimensional parameter spaces, the main hurdle lies in estimating the integral (8). This can be done either by Monte-Carlo methods or asymptotic approximations, from which the Bayesian Information Criterion (BIC) derives. (see *e.g.* the review from Kass and Raftery, 1995, and the references therein). When the sample size n is not too small compared to the dimension k of

[6] The density estimator defined by (7) is a "Bayes estimator": it minimizes the posterior expected quadratic loss $E_{\tilde{\pi}|\mathscr{W}} \left\{ \left(h(\mathbf{w}|\cdot) - \hat{\mathbf{h}}(\mathbf{w}) \right)^2 \right\}$, at each point $\mathbf{w}$, where the expectancy is taken with respect to the *posterior* density $\tilde{\pi}|\mathscr{W}$ in the union parameter space (see e.g. Robert, 2007, Chap. 2, for details about Bayesian decision theory).

the parameter space, a reasonable trade-off between precision and computational efficiency is the Laplace's approximation method: the logarithm of the integrand $\tilde{l}_m(\theta_m) = \log\left[h_m(\mathbf{w}|\theta_m)\pi_m(\theta_m)\right]$ should be approximately normal around the posterior mode $\hat{\theta}_m$, with covariance matrix $\hat{\Sigma} = (-\mathbf{d}^2\tilde{l})^{-1}$ where $\mathbf{d}^2\tilde{l}$ is the Hessian matrix at $\hat{\theta}$. This yields, by integration, the Laplace approximation

$$\hat{\mathcal{L}}_m(\mathscr{W}) = (2\pi)^{k/2}\left|\hat{\Sigma}\right|^{1/2} h(\mathscr{W}|\hat{\theta}_m)\pi_m(\hat{\theta}_m) \tag{9}$$

Kass and Raftery (1995) suggest that in most cases where $n/k \geq 20$, (9) yields a good precision. More details about the validity of (9) may be found in Kass et al. (1990). For lower sample size, one alternative to obtain the posterior weights would be to implement a MC MC algorithm with reversible jumps between the individual models. The proportion of 'time' spent in each model would provide estimates of posterior weights. The main difficulty with this approach would be to define reasonable 'jumps' proposals, to obtain jumps acceptance rates high enough for the chain's mixing properties to remain acceptable in practice.

Inside each single model, the posterior parameter distribution is classically evaluated by a Metropolis-Hastings algorithm producing an approximate posterior sample $(\theta_{m,1}, \ldots, \theta_{m,N})$. The latter is used to approximate each term $\tilde{h}_m(\mathbf{w}) = \int_{\Theta_m} h_m(\mathbf{w} \mid \theta_m) d(\pi_m|\mathscr{W})(\theta_m)$ in (7) *via*

$$\hat{h}_m(\mathbf{w}) = \frac{1}{N}\sum_{t=1}^{N} h_m(\mathbf{w} \mid \theta_{m,t}) \,. \tag{10}$$

Throughout this paper, the total number of simulations is set to 50×10^3, from which the first 30×10^3, values are discarded. The Heidelberger and Welch test (Heidelberger and Welch, 1981; Cowles and Carlin, 1996) and the Geweke convergence diagnostics (Geweke, 1992) show good convergence properties for this burn-in period.

4. BMA with the PB and NL spectral measures

4.1. **Preliminary: definition of Bayesian PB and NL models.** Before implementing the BMA scheme, we need to separately implement our two models in a Bayesian framework. To our knowledge, this has never been done for the PB model neither for our NL model.

Since none of these models is part of the exponential family, there is no obvious uninformative or invariant prior choices at hand. So, for convenience, we transform the parameter spaces to obtain unconstrained ones. Namely, we set

$$\underline{\theta}_{\mathrm{PB}} = \log(\underline{\beta}) \in \mathbf{R}^4 \,; \quad \underline{\theta}_{\mathrm{NL}} = \mathrm{logit}(\underline{\alpha}) \in \mathbf{R}^4 \,.$$

where, for the NL model, $\mathrm{logit}(x) = \log(x/(1-x))$, which excludes the independent case $\underline{\alpha} = (1,1,1,1)$. Then, the parameters in each model are assumed to be *a priori* mutually independent and normally distributed with common mean equal to 0 and standard deviation equal to 3. Results on simulated data (see Appendix) show that this prior specification does not introduce a strong bias in the estimations.

4.2. **Averaging the PB and NL models: simulation study.**

Comparison with other approaches.
An alternative to the BMA framework would be to select the 'best' model given a data set. The criterion for comparison could be, for example, the posterior weight, or the BIC or AIC. In our case, these three criteria are approximately equivalent: indeed, the differences of scores between two models in terms of AIC or BIC are the same when the two models have same dimension ($k_1 = k_2$), since in such a case $\mathrm{BIC}_{12} - \mathrm{AIC}_{12} = (k_2 - k_1)\log n - 2(k_2 - k_1) = 0$. Selecting the model according to the BIC or the AIC is thus exactly the same. As for the posterior weights, note that the prior model weights were chosen uniform (here $(1/2, 1/2)$). The posterior odds are then equal to the Bayes factor: $B_{12} = p_1(\mathscr{W})/p_2(\mathscr{W}) = \mathcal{L}_1(\mathscr{W})/\mathcal{L}_2(\mathscr{W})$. For large sample sizes the logarithm of the latter can be approximated by the Schwarz criterion $S = \log \mathcal{L}_1(\mathscr{W}) - \log \mathcal{L}_2(\mathscr{W}) - 1/2(k_2 - k_1)\log(n)$, which is $-1/2$ times the the BIC (see *e.g.* Kass and Raftery, 1995; Kass et al., 1990). In view of the asymptotic equivalence of the three criteria, and because posterior weights are anyway computed for the BMA, the 'model selection' alternative considered here consists in selecting the model with highest posterior weight.

The main goal of our simulations is to compare the predictive performance of the BMA against this model selection framework and against single models, in terms of predictive angular densities and estimations of the probability of being in a failure region $A(\mathbf{u})$ as defined in Section 2. The union model is larger than any individual model, and averaging instead of selecting allows to 'integrate' the uncertainty. One could thus expect the predictive performance to be enhanced.

The main theoretical limitation of the averaging approach stems from a concentration phenomenon: if the data arises from the union model (thus, from one of the individual model), the posterior mass should concentrate around the true value and assign more mass to the model containing it. In "misspecified" cases (when the true distribution does not belong to the union model), the posterior is bound to concentrate around "asymptotic carrier" regions of the parameter space, which minimize the Kullback-Leibler divergence to the true distribution (Berk, 1966; Kleijn and van der Vaart, 2006). In our context, this means that, for large sample sizes, the BMA is likely to select a single model, except if the true distribution is at exact "equi-distance" from the two. Consequently, we restrict our study to situations where the sample size is moderate (namely, 80 points).

Predictive scores with simulated data.
In this paragraph, the angular data set $\mathscr{W}$ under consideration is supposed to be simulated according to some angular distribution h_0 (a Dirichlet mixture (DM) distribution, see the next paragraph) on the simplex $\mathbf{S}_3$. The density estimates produced by each inference framework (PB model, NL model, BMA and model selection) are to be compared. We now introduce different scoring rules allowing to do so. The interest of considering several scores is that they rank the predictions according to different criteria. It may happen that one framework be selected by one scoring rule and discarded by another one. The aim here is to check consistency, *i.e.* that our conclusions are relatively independent from the considered score.

As a performance score for a density estimate $\hat{h}$ fitted to $\mathscr{W}$, we consider the logarithmic score

$$LS(\hat{h}, h_0) = -\mathbf{E}_{h_0} \log(\hat{h}(\,\cdot\,)) = -\int_{\mathbf{S}_3} \log(\hat{h}(\mathbf{w}))h_0(\mathbf{w})\,\mathrm{d}\mathbf{w}\,, \tag{11}$$

associated to the Kullback-Leibler divergence between the density estimate and the true distribution. A model with low LS is 'close' to the truth. According to this rule, the best model is the one minimizing the score (note that a zero is not a measure of perfect fit). Since one can simulate from h_0, the latter integral can be evaluated by simple Monte-Carlo

$$\hat{LS}(\hat{h}, h_0) = \frac{-1}{N_{\mathrm{mc}}} \sum_{N=1}^{N_{\mathrm{mc}}} \log(\hat{h}(\mathbf{w}_N))\,; \quad \mathbf{w}_N \overset{i.i.d.}{\sim} h_0\,. \tag{12}$$

In the remainder of this paper, N_{mc} is set to 50×10^3.

The approximation (12) allows us to compare the performance of the predictive densities $\hat{h}_{\mathrm{PB}}, \hat{h}_{\mathrm{NL}}, \hat{h}_{\mathrm{BMA}}$ and $\hat{h}_{\mathrm{Select.}}$ respectively in the PB model, in the NL model, in the BMA and in the model selection framework. The predictive density for the latter is defined as

$$\hat{h}_{\mathrm{Select.}} = \mathbf{1}_{p_{\mathrm{PB}}(\mathscr{W}) \geq 0.5} \hat{h}_{\mathrm{PB}} + \mathbf{1}_{p_{\mathrm{PB}}(\mathscr{W}) < 0.5} \hat{h}_{\mathrm{NL}}\,.$$

One of the main interest in multivariate extreme value theory may reside in the probability of an excess of a high threshold. In this study, we consider the probability of falling in the failure region $A(\mathbf{u})$ with $\mathbf{u} = (100, 100, 100)$. On the Fréchet scale, it corresponds to a marginal excess probability of roughly $\frac{1}{100}$. The quantity of interest Δ is thus a joint probability

$$\Delta(m, \underline{\theta}) = \mathbf{P}(\mathbf{X} > \mathbf{u} | m, \underline{\theta}) \simeq \nu(A(\mathbf{u}) | m, \underline{\theta})$$

where $m \in \{\mathrm{PB}, \mathrm{NL}\}$ (see Section 2). The true probability is

$$\Delta(h_0) = \mathbf{P}_{h_0}(\mathbf{X} > \mathbf{u}) = \nu_0(A(\mathbf{u}))\,,$$

where ν_0 is the true exponent measure, which density in pseudo-polar coordinates is (*e.g.* on the region $\{r > 1\}$) $\mathrm{d}\nu_0 = d\frac{\mathrm{d}r}{r^2} h_0(\mathbf{w})\,\mathrm{d}\mathbf{w}$. Here, the approximation becomes an equality because the radii and angles are simulated independently from each other.

For the PB (*resp.* the true) density, $\nu(A(\mathbf{u})|\underline{\theta}, m)$ (*resp.* $\nu_0(A(\mathbf{u})$) is given by (4) and approximated by Monte-Carlo integration (since we can simulate angular samples from PB distributions and from the true one). In the NL model, it is simply given by (5).

The output of the Bayesian procedure in model m is, for a given data set $\mathscr{W}$, an approximate posterior sample $\{\underline{\theta}_m(t)\}_{1 \leq t \leq T}$ [7]. This posterior is transformed into a posterior Δ-sample $\{\delta_m(t)\}_{1 \leq t \leq T} = \{\Delta(\underline{\theta}_m(t))\}_{1 \leq t \leq T} \in (0, 1)$ of probabilities of failure, so that the posterior predictive distribution $\Delta^*(\pi_m|\mathscr{W})$ (see Section 4) on $(0, 1)$ is approximated by the discrete cumulative distribution function (*cdf*)

$$\hat{F}_m(y) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}_{\delta_m(t) \leq y} \quad (y \in (0, 1))\,.$$

[7] Here, $T = 200$ after discarding the values from the burn-in period and thinning. The thinning interval is set to 100 to reduce the computational time

The posterior predictive *cdf* in the BMA is thus the weighted average

$$\hat{F}_{\mathrm{BMA}}(\,\cdot\,) = p_{\mathrm{PB}}(\mathscr{W})\,\hat{F}_{\mathrm{PB}}(\,\cdot\,) + p_{\mathrm{NL}}(\mathscr{W})\,\hat{F}_{\mathrm{NL}}(\,\cdot\,)\,.$$

We can now compare the different distributions *via* strictly proper scoring rules (Gneiting and Raftery, 2007), adapted to the case where the true distribution is known to be the Dirac mass at $\delta_0 = \Delta(h_0)$. Namely, we consider the continuous ranked probability score ($CRPS$), the predictive model choice criterion[8] ($PMCC$) and the interval score (IS_α) for the central $(1-\alpha)*100\,\%$ interval based on predictive quantiles, with $\alpha = 0.1$. If $\hat{F}$, $\mathbf{E}_{\hat{F}}(\Delta)$, $\mathbf{Var}_{\hat{F}}(\Delta)$, $\hat{q}_{\alpha,l}$ and $\hat{q}_{\alpha,u}$ denote respectively the predictive *cdf* on $(0,1)$, the predictive mean and variance, and the predictive $\alpha/2$ and $1-\alpha/2$ quantiles, and if $\delta_0 \in (0,1)$ is the true value, these (negatively oriented) scores are

$$CRPS(\hat{F},\delta_0) = \int_{(0,1)} (\hat{F}(y) - \mathbf{1}_{\delta_0 \le y})^2 \,\mathrm{d}y\,, \tag{13}$$

$$PMCC(\hat{F},\delta_0) = (\mathbf{E}_{\hat{F}}(\Delta) - \delta_0)^2 + \mathbf{Var}_{\hat{F}}(\Delta)\,, \tag{14}$$

$$IS_\alpha(\hat{F},\delta_0) = \begin{cases} 2\alpha(\hat{q}_{\alpha,u} - \hat{q}_{\alpha,l}) + 4(\hat{q}_{\alpha,l} - \delta_0) & \text{if } \delta_0 \le \hat{q}_{\alpha,l}\,, \\ 2\alpha(\hat{q}_{\alpha,u} - \hat{q}_{\alpha,l}) & \text{if } \hat{q}_{\alpha,l} \le \delta_0 \le \hat{q}_{\alpha,u}\,, \\ 2\alpha(\hat{q}_{\alpha,u} - \hat{q}_{\alpha,l}) + 4(\delta_0 - \hat{q}_{\alpha,u}) & \text{if } \delta_0 > \hat{q}_{\alpha,u}\,. \end{cases} \tag{15}$$

Similarly to the logarithmic score, the 'best' model according to a given scoring rule is the one that minimizes the score.

Experimental setup.
Data sets are generated from Dirichlet mixture (DM) distributions (Boldi and Davison, 2007), which cover a wide variety of distributional shapes. The Dirichlet mixture parameters themselves are drawn according to a simulating rule described in the Appendix. Note that the hyper parameters for this simulating rule were chosen in order to grant significantly positive weights to both models.

We generate 20 DM parameters $\{\underline{\theta}_0^i\}_{1\le i\le 20}$ and for each $\underline{\theta}_0^i$, 5 data sets $\{\mathscr{W}_j^i\}_{1\le j\le 5}$ of size 80 each are generated according to the DM density h_0^i corresponding to $\underline{\theta}_0^i$.

For each of the 100 data sets $\{\mathscr{W}_j^i\}_{1\le i\le 20, 1\le j\le 5}$, separate inference is made in each framework 'fr' (here, fr represents the PB model, the NL model, the BMA and the model selection framework), yielding a density estimate $\hat{h}_j^i|_{fr}$, a *cdf* for the probability of failure $\hat{F}_j^i|_{fr}$ and posterior model weights which are approximated *via* (9). The posterior mode and the hessian matrix are approximated by numerical optimization.

Finally, the scores obtained by each framework are averaged over all the experiments (i,j).

Results.
The first panel of Table 1 shows the average scores (over the 100 data sets) obtained by each model, by the BMA and in the selection framework. The second panel

[8] This scoring rule is not proper in the general case but becomes so when the true distribution is a Dirac mass.

shows the average score differences[9] between the BMA and the three other possible approaches, together with an order of magnitude of the errors involved by the Monte-Carlo approximations used to compute the score differences between the BMA and the other approaches. More details about these errors are given in the Appendix.

For example, line ' BMA/ PB', column 'CRPS' corresponds to

$$CRPS(\text{BMA/PB}) = \sum_{i=1}^{20}\sum_{j=1}^{5} CRPS(\hat{F}_j^i|_{\text{BMA}}, \hat{\delta}^i) - CRPS(\hat{F}_j^i|_{\text{PB}}, \hat{\delta}^i),$$

where $\hat{\delta}^i$ is the Monte-Carlo estimate of the true probability of failure under the Dirichlet distribution with parameter $\underline{\theta}_0^i$, *i.e.*

$$\hat{\delta}^i = \frac{3}{N_{\text{mc}}} \sum_{N=1}^{N_{\text{mc}}} \min_{j\in 1:d} \left\{ \frac{w_{j,N}}{u_j} \right\} \; ; \quad \mathbf{w}_N \overset{i.i.d}{\sim} h_0^i \; ; \quad u_j = \frac{1}{100} \, .$$

Column 'LS' corresponds to

$$\sum_{i=1}^{20}\sum_{j=1}^{5} \hat{LS}(\hat{h}_j^i|_{\text{BMA}}) - \hat{LS}(\hat{h}_j^i|_{\text{PB}}),$$

where $\hat{LS}$ is given by (12).

TABLE 1. Comparison of mean scores with simulated data (error magnitude on score differences between parentheses).

Scores	*LS*	*CRPS*	*PMCC*	*IS*
PB	−107.48	24.04	20.97	45.06
NL	−106.07	21.34	**18.69**	38.08
BMA	**−108.39**	**21.33**	19.56	**37.21**
Select	−107.36	22.95	20.11	42.27
BMA/PB	−0.91 (0.32)	−2.7 (0.04)	−1.41 (0.05)	−7.85 (0.15)
BMA/NL	−2.33 (0.32)	−0.002 (0.09)	0.87 (0.08)	−0.87 (0.21)
BMA/Select	−1.03 (0.32)	−1.62 (0.02)	−0.55 (0.01)	−5.06 (0.12)

In terms of spectral density itself, the BMA approach obtains the best logarithmic score (first column, second panel). The logarithmic score obtained by the selection framework is also better than the ones obtained both by the PB and by the NL models. The gain is less obvious in terms of estimated probabilities of failure, probably because, for this kind of simulated data, the NL model obtains better average scores than the PB model (note that this tendency is reversed in terms of logarithmic scores). In any case, the BMA gives slightly, but consistently, better predictions, with respect to all the considered scores, than the model selection framework (line 7).

The disappointing aspect of these results is the fact that the relative gain or loss is low: roughly, between 1/100 and 1/10 depending on the considered score.

[9]The scores reported in each column have respectively been multiplied by $10^2, 10^5, 10^8$ and 10^5 to improve the readability of the numerical output.

4.3. **Example: Leeds data set.** We separately fit the PB and the NL model on the Leeds data set. Table 2 gathers results in terms of the transformed parameters in each model. $\hat{\theta}_{\text{post}}$ and $\hat{\sigma}_{\text{post}}$ denote the mean and standard deviation of the posterior sample issued from the Metropolis algorithm, $\hat{\theta}_{\text{mode}}$, $\hat{\sigma}_{\text{mode}}$, are the posterior mode and the 'standard deviation' represented by (with the notations of the Laplace approximation (9)) the squared root of the diagonal elements of the inverse hessian $\hat{\Sigma}$. The maximum likelihood estimates $\hat{\theta}_{mle}$ and the estimated standard errors $\hat{\sigma}_{mle}$ are also reported. Our Bayesian analysis corroborates the frequentist

TABLE 2. The PB and NL models fitted to Leeds data: Comparison between frequentist estimates and posterior summary statistics.

	PB model				NL model			
	$\log\beta_0$	$\log\beta_{12}$	$\log\beta_{13}$	$\log\beta_{23}$	$\text{logit}\,\alpha_0$	$\text{logit}\,\alpha_{12}$	$\text{logit}\,\alpha_{13}$	$\text{logit}\,\alpha_{23}$
$\hat{\theta}_{\text{post}}$	0.3	1.27	−0.35	1.22	0.22	0.89	4.57	1.19
$\hat{\sigma}_{\text{post}}$	0.14	0.43	0.23	0.42	0.09	0.49	1.69	0.64
$\hat{\theta}_{\text{mode}}$	0.3	1.3	−0.34	1.14	0.21	0.79	3.67	1.03
$\hat{\sigma}_{\text{mode}}$	0.14	0.43	0.22	0.42	0.09	0.45	1.23	0.42
$\hat{\theta}_{\text{mle}}$	0.3	1.32	−0.34	1.16	0.21	0.81	17.97	1.07
$\hat{\sigma}_{\text{mle}}$	0.14	0.43	0.22	0.43	0.09	0.47	2588.78	0.44

estimates. The unusually high standard deviation of the maximum likelihood estimate for $\text{logit}(\alpha_{13})$ is easily explained: the inverse logit link function is numerically flat (equal to 1) above 17, and $\text{logit}^{-1}(3.67) = 0.98$. As expected, adding a prior re-centres the estimates towards the origin, but the relative discrepancy between the Bayesian and frequentist modes (with respect to the standard deviation of the frequentist ones) is less than 0.12. Also, the posterior mode and mean are close to each other. This suggests that the asymptotic domain of validity of the Bernstein-Von Mises theorem (asymptotic normality of the posterior distribution, see *e.g.* van der Vaart, 2000) is approximately reached.

The posterior predictive spectral densities in the PB and NL models can be computed *via* (10). The squared dots in Figure 3 represent the data displayed in Figure 2. Each panel tells us the same main story, a lot of mass in the middle, more near the middle of the edges joining the pairs (PM10,NO) and (NO,NO2) than between the pair (PM10,NO2). This pattern roughly corresponds to the distribution of the observed angular points over the simplex. Still, the two panels have important differences. For example, the NL model assigns more mass to the regions near the vertices.

For this Leeds data set, the posterior weights are overwhelmingly in favour of the NL model. Table 3 gathers the posterior weights issued from the BIC approximation, the Laplace method and by simple Monte-Carlo integration (parameters are drawn from the prior).

For the Leeds data, BMA teaches us that a well-chosen four parameters NL model belonging to the large Nested Asymmetric Logistic family outperforms the PB model.

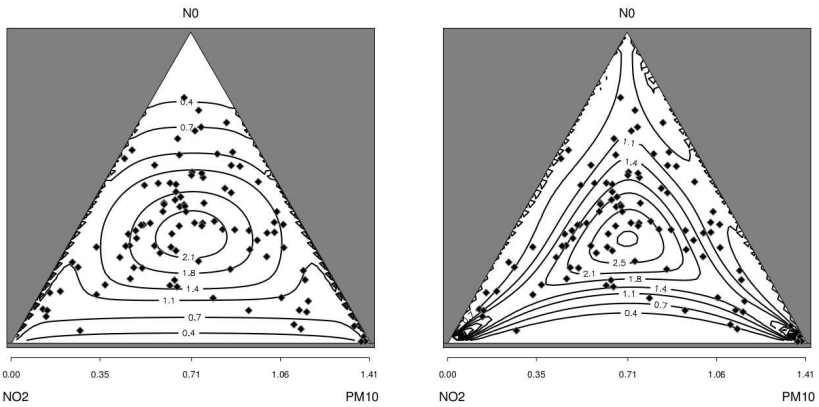

FIGURE 3. Leeds data: posterior predictive densities in the PB model (left panel) and the NL model (right panel).

TABLE 3. Leeds data set: posterior PB model weights and marginal likelihoods.
MC MC steps: 100×10^3.

	Laplace	BIC	Monte-Carlo
PB	$2.2\,10^{32}$	$4.1\,10^{32}$	$2.4\,10^{32}$($4.8\,10^{31}$)
NL	$8.2\,10^{34}$	$1\,10^{35}$	$1.4\,10^{35}$($1.9\,10^{34}$)
$\hat{p}_{\mathrm{PB}}(\mathbf{W})$	0.0027	0.004	0.0017

5. DISCUSSION

This article shows that it is feasible to implement a BMA approach for angular measure models. Simulation studies indicate that this approach can, at best, improve the predictive density estimates over each single model and at worst, be used as a selection tool by identifying a single one. For the four considered scoring rules, the gain represented by the BMA against the model selection framework is significant (in view of the Monte-Carlo errors) but moderate: the order of magnitude of the scores is unchanged.

For the sake of conciseness, we have only considered two models to be averaged. Future BMA roads would be to enlarge the dictionaries of parametric spectral families (e.g. for the PB model, Ballani and Schlather, 2011) and/or to extend the BMA framework to a mixture model setup, i.e. replacing the disjoint union parameter space by a product space. The resulting model would be more flexible, in the sense that the posterior mass would not have to concentrate on one single model for large sample sizes. As a drawback, the dimension of a product space is the sum of the dimensions of individual models, and the curse of dimensionality is likely to impose longer burn-in periods for MC MC algorithms. Also, one could not use posterior samples obtained in distinct models. We recall that, in this paper, we consider situations when separate inference has already been achieved, or can be made in a simple way, and where estimates are to be averaged. The main interest of the BMA approach is to offer a compromise between model flexibility

and parsimony: the estimated distribution (the posterior predictive) is a mixture, while inference is conducted in lower dimensional models.

Also, for our leading example, Heffernan and Tawn (2004)'s study suggests that the pairs ($S0_2$, NO) and (SO_2, PM10) might be asymptotically independent. It should thus be of interest to average general spectral measures associated with asymptotically independent models, as introduced by Ledford and Tawn (1996) and Ramos and Ledford (2009) in the bi-variate case, and extended to general multivariate problems in Qin *et. al.* (see footnote 4). The estimated distributions would not be max-stable anymore, but this would account for a potentially greater source of uncertainty than the one attached to model choice within the asymptotically dependent class.

Supplementary material

An R package is available at
`http://www.lsce.ipsl.fr/Phocea/Pisp/index.php?nom=anne.sabourin`

Acknowledgement

Part of this work has been supported by the EU-FP7 ACQWA Project (www.acqwa.ch), by the PEPER-GIS project, by the ANR-MOPERA project, by the ANR-McSim project and by the MIRACCLE-GICC project. The authors would like to thank Dan Cooley for his help with the PB model, and an anonymous referee for useful remarks.

Appendix 1: Bayesian PB and NL models

Simulation rule for the PB model.
Whereas Cooley *et al.* used an accept-reject method for simulation, the one that we propose here is direct. The PB density can be re-parametrized by setting $\rho_{ij} = w_i + w_j\,,\ w_{i/ij} = w_i/(w_i + w_j)\,,\ s_{ij} = w_{[-(i,j)]}/(1-\rho_{ij})$, where $w_{[-(i,j)]} = (w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{j-1}, w_{j+1}, \ldots, w_d)$.

The transformation $(\rho_{ij}, w_{i/ij}, s_{ij}) \mapsto w$ has Jacobian: $J = J(\rho_{ij}) = \rho_{ij}(1-\rho_{ij})^{d-3}$. Each beta function

$$h_{i,j}(\{w_{ij}, w_{i/ij}, s_{ij}\}\,|\beta_0, \beta_{ij})\,,$$

can be expressed within these new coordinates

$$h_{i,j}(\{\rho_{ij}, w_{i/ij}, s(w_{ij})\}\,|\beta_0, \beta_{ij}) \propto \rho_{ij}^{2\beta_0}\,(1-\rho_{ij})^{(d-2)\beta_0 - 1} \cdots \\ \cdots w_{i/ij}^{\beta_{ij}-1}\,\big(1-w_{i/ij}\big)^{\beta_{i,j}-1}\,\frac{1}{J(\rho_{ij})}\,,$$

which can be written with the standard R package notations as

$$\frac{1}{J(\rho_{ij})}\mathsf{dbeta}(\rho_{ij}, 2\beta_0{+}1, (d{-}2)\beta_0)\,\mathsf{dbeta}(w_{i/ij}, \beta_{i,j}, \beta_{i,j})\,\mathsf{ddirichlet}(s_{ij}, \mathsf{rep}(1, d{-}2))\ .$$

The three factors correspond to two Beta distributions and one uniform distribution on the unit simplex of dimension $d-3$. The following algorithm produces the desired angular variables W according to the density $h_{\mathrm{PB}}(.\mid \beta_0, \beta_{ij})$.

Algorithm 1.

(1) *Choose uniformly a pair* $(i < j)$.

(2) *Generate independently the vectors* R_{ij}, $W_{i/ij}$ *and* S_{ij} *according to the Beta distributions* $\mathcal{B}e(2\beta_0+1,(d-2)\beta_0)$ *and* $\mathcal{B}e(\beta_{ij},\beta_{ij})$, *and the uniform Dirichlet distribution* $\mathcal{D}ir_{d-2}(1,\dots,1)$, *respectively.*

(3) *Define* W *as* $W_i = R_{ij}W_{i/ij}$, $W_j = R_{ij}(1-W_{i/ij})$ *and* $W_{[-(i,j)]} = (1-R_{ij})S_{ij}$.

Angular density in the NL model.
From Coles and Tawn (1991), Theorem 1, with our normalizing convention, the angular density on the simplex is $h_{\mathrm{NL}}(\mathbf{w}|\underline{\alpha}) = -\frac{1}{d}\,\partial_{1,2,3}V_{\mathrm{NL}}(\mathbf{x}|\underline{\alpha})|_{\mathbf{x}=\mathbf{w}}$, where we write $\partial_{i_1,\dots,i_k}(\cdot)$ the partial derivative with respect to $x_{i_1},\dots,x_{i_k}$.

Letting

(16)
$$U_{ij}(\mathbf{x}) = x_i^{\frac{-1}{\alpha_0\alpha_{ij}}} + x_i^{\frac{-1}{\alpha_0\alpha_{ij}}}\ (1\le i<j\le 3)\,, \qquad T(\mathbf{x}) = \left(U_{12}^{\alpha_{12}} + U_{13}^{\alpha_{13}} + U_{23}^{\alpha_{23}}\right)(\mathbf{x})\,,$$

we have $V_{\mathrm{NL}} = 2^{-\alpha_0}T^{\alpha_0}(\mathbf{x})$, so that

$$\partial_{1,2,3}V_{\mathrm{NL}}(\mathbf{x}|\underline{\alpha}) = 2^{-\alpha_0}\alpha_0\left[T^{\alpha_0-1}(\mathbf{x})\partial_{1,2,3}T(\mathbf{x}) + \cdots\right.$$
$$(\alpha_0-1)T^{\alpha_0-2}(\mathbf{x})\left\{\partial_1T(\mathbf{x})\partial_{2,3}T(\mathbf{x}) + \partial_2T(\mathbf{x})\partial_{1,3}T(\mathbf{x}) + \partial_3T(\mathbf{x})\partial_{1,2}T(\mathbf{x})\right\} + \cdots$$
$$\left.(\alpha_0-1)(\alpha_0-2)T^{\alpha_0-3}(\mathbf{x})\partial_1T(\mathbf{x})\partial_2T(\mathbf{x})\partial_3T(\mathbf{x})\right]\,.$$

The simple and double partial derivatives are

$$\partial_iT(\mathbf{x}) = \frac{-1}{\alpha_0}\left(x_i^{\frac{-1}{\alpha_0\alpha_{ij}}-1}U_{ij}^{\alpha_{ij}-1} + x_i^{\frac{-1}{\alpha_0\alpha_{ik}}-1}U_{ik}^{\alpha_{ik}-1}\right) \tag{17}$$

and

$$\partial_{i,j}T(\mathbf{x}) = \frac{\alpha_{ij}-1}{\alpha_0^2\alpha_{ij}}\left(x_i\,x_j\right)^{\frac{-1}{\alpha_{ij}}-2}U_{ij}^{\alpha_{ij}-2}. \tag{18}$$

The third order derivative is thus zero. Finally, we have

$$h_{\mathrm{NL}}(\mathbf{w}|\underline{\alpha}) = \left[\frac{\alpha_0(1-\alpha_0)}{2^{\alpha_0}\,d}T^{\alpha_0-3}\cdots\right.$$
$$\left.\cdots\left\{\sum_{1\le i\ne j\ne k\le 3} T\partial_iT\partial_{j,k}T \;+\; (\alpha_0-2)\partial_1T\partial_2T\partial_3T\right\}\right]_{\mathbf{x}=\mathbf{w}} \tag{19}$$

where all the terms are given in (16), (17) and (18).

Simulation method in the NL model.
We adapt here the method proposed by Stephenson (2003) to our context.

Algorithm 2.

(1) *Generate independently four positive alpha-stable variables* $S, S_{12}, S_{13}, S_{23}$, *with respective index* α_0, α_{12}, α_{13}, $\alpha_{23}\in(0,1)$, i.e. *with Laplace transform* $E(\exp(-tS)) = e^{-t^{\alpha_0}}$ *(*resp. $e^{-t^{\alpha_{ij}}}$*).*

(2) *For* $i\in\{1,2,3\}$:
   (a) *Simulate independently two standard exponentials* $E_{i,ij}, E_{i,ik}$.
   (b) *Set* $X_{i,ij} = \left[\left(\frac{S}{2}\right)^{\frac{1}{\alpha_{ij}}}\frac{S_{ij}}{E_{i,ij}}\right]^{\alpha_{ij}\alpha_0}$ *and* $X_{i,ik} = \left[\left(\frac{S}{2}\right)^{\frac{1}{\alpha_{ik}}}\frac{S_{ik}}{E_{i,ik}}\right]^{\alpha_{ik}\alpha_0}$.
   (c) *Set* $X_i = \max(X_{i,ij}, X_{i,ik})$.

Then, $\mathbf{X} = (X_1, X_2, X_3)$ *has unit Fréchet margins and a multivariate distribution belonging to the* NL *model (6).*

Proof. If $\mathbf{X}$ is generated according to the above algorithm, the conditional variables $\mathbf{X_{i,ij}}|\,(S = s, S_{ij} = s_{ij})$ are independent with distribution

$$\mathbf{P}\left(X_{i,ij} \leq x_i | s, s_{12}\right) = \exp\left(-s_{ij}\left(\frac{s}{2}\right)^{1/\alpha_{ij}}\left(\frac{1}{x_i}\right)^{1/(\alpha_0\alpha_{ij})}\right),$$

So that X has conditional distribution

$$\mathbf{P}\left(\mathbf{X} \leq \mathbf{x} | s, s_{12}\right) = \exp\left(-\sum_{1\leq i<j\leq 3} s_{ij}\left(\frac{s}{2}\right)^{1/\alpha_{ij}} \cdots \right.$$
$$\left. \cdots\left(\left(\frac{1}{x_i}\right)^{1/(\alpha_0\alpha_{ij})} + \left(\frac{1}{x_j}\right)^{1/(\alpha_0\alpha_{ij})}\right)\right).$$

Integrating with respect to the s_{ij}'s and s and using the Laplace transform property of positive α-stable variables yields the desired distribution function. □

The angular components $W_i = X_i/(X_1+X_2+X_3)$ follow immediately. By fixing a high threshold r_0 and retaining only the angular points corresponding to radii $R > r_0$, one obtains a sample on the simplex, approximately following angular distribution with density $h_{\mathrm{NL}}(.|\underline{\alpha})$.

Appendix 2: Results with simulated data from single models

Two data sets of 80 angular points each are simulated, one in the PB model, the other in the NL model. A 50 10^3-iteration Metropolis-Hastings is run, the last 20 10^3 values are kept.

The marginal posterior densities for the four parameters in the PB (*resp.* NL) model, obtained by a kernel smoothing of the *posterior* sample, are shown (solid lines) in Figure 4 (*resp.* Figure 5) , together with the prior densities (thin dotted lines) and the true parameters (vertical thick dotted lines). For all the parameters components, the posterior concentrates around the "true" value.

The posterior predictive density estimates are deduced from the posterior sample according to (10), and plotted in Figure 6. showing remarkable agreement between the estimated (solid lines), and the true distribution contours (dotted lines).

Basic summary statistics for the posterior samples are gathered in Table 4 (θ_0 stands for the "true" transformed parameter, see sub-section 4.3 for other notations), together with maximum likelihood estimates. The three approaches yield comparable results and the true parameter values lie at less than two standard deviations from their respective posterior mean estimates (except for the global dependence parameter α_0 in the NL model, where the discrepancy is about 2.4 for the three estimates).

Appendix 3: Simulation study

We give here a more complete account of the results obtained in Section 4.2.

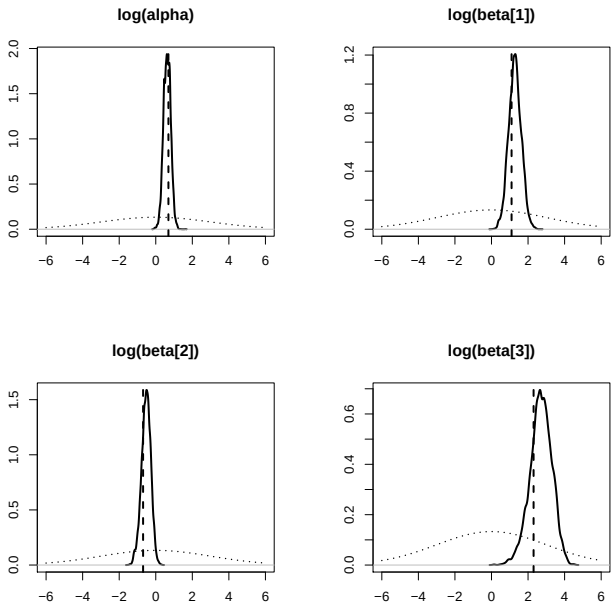

FIGURE 4. PB model: prior and posterior parameter marginal densities with simulated data. Upper left panel: $\text{logit}(\beta_0)$, upper right panel: $\text{logit}(\beta_{12})$, left and right lower panels: $\text{logit}(\beta_{13})$ and $\log(\beta_{23})$.

TABLE 4. The PB and NL models fitted to simulated data: Comparison between frequentist estimates and posterior summary statistics.

	PB model				NL model			
θ	$\log\beta_0$	$\log\beta_{12}$	$\log\beta_{13}$	$\log\beta_{23}$	$\text{logit}\,\alpha_0$	$\text{logit}\,\alpha_{12}$	$\text{logit}\,\alpha_{13}$	$\text{logit}\,\alpha_{23}$
θ_0	0.69	1.1	−0.69	2.3	0.41	−0.85	1.39	−0.41
$\hat{\theta}_{\text{post}}$	0.62	1.28	−0.53	2.7	0.15	−0.5	2.41	0.23
$\hat{\sigma}_{\text{post}}$	0.2	0.35	0.26	0.6	0.11	0.3	0.73	0.4
$\hat{\theta}_{\text{mode}}$	0.62	1.31	−0.52	2.77	0.14	−0.55	2.15	0.13
$\hat{\sigma}_{\text{mode}}$	0.19	0.34	0.25	0.58	0.11	0.28	0.61	0.35
$\hat{\theta}_{\text{mle}}$	0.62	1.32	−0.52	2.88	0.14	−0.56	2.25	0.13
$\hat{\sigma}_{\text{mle}}$	0.19	0.35	0.25	0.61	0.11	0.28	0.67	0.35

Dirichlet mixture model for spectral densities. Recall that the Dirichlet density, which we denote diri, can be parametrized by a mean vector $\boldsymbol{\mu} \in \mathbf{S}_d$ and

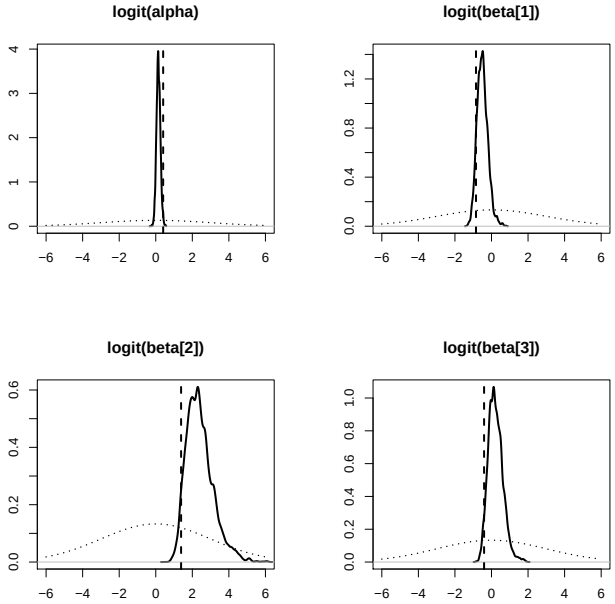

FIGURE 5. NL model: prior and posterior parameter marginal densities with simulated data. Upper left panel: $\log(\alpha_0)$, upper right panel: $\log(\alpha_{12})$, left and right lower panels: $\log(\alpha_{13})$ and $\log(\alpha_{23})$.

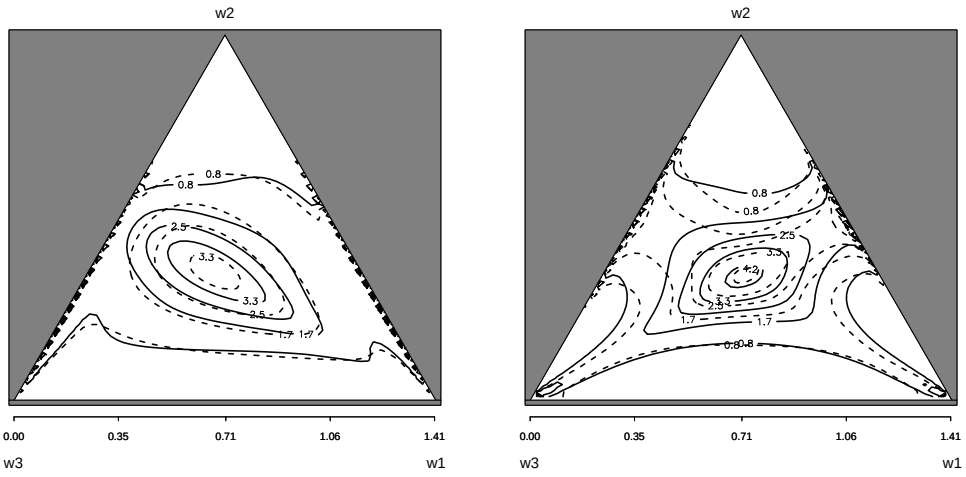

FIGURE 6. Angular measures: Simulation and estimation in the the PB model (left panel) and in the NL model(right panel).

a concentration parameter $\nu > 0$, so that

$$\forall \mathbf{w} \in \mathbf{S}_d,\ \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^{d} \Gamma(\nu\mu_i)} \prod_{i=1}^{d} w_i^{\nu\mu_i - 1}\,.$$

The Dirichlet mixture model is the family of finite mixtures of such densities, with positive weight vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$ $(K \geq 1)$ summing to one, concentration vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$ and mean matrix $\boldsymbol{\mu} = (\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,K})$ where $\boldsymbol{\mu}_{\cdot,k} = (\mu_{1,k}, \ldots, \mu_{d,k})$ is the mean vector for the k^{th} mixture component. In this model, the mean constraint (2) is equivalent to

$$\forall i \in \{1, \ldots, d\}, \ \sum_{k=1}^{K} \omega_k \, \mu_{i,k} = \frac{1}{d} \, . \tag{20}$$

Simulation of random Dirichlet mixture parameters and data sets. For our simulation, we generate 20 Dirichlet mixture parameters

$$\underline{\theta}_0^i = (\boldsymbol{\mu}^i{}_{\cdot,1:K}, \boldsymbol{\omega}^i_{1:K}, \boldsymbol{\nu}^i_{1:K})_{1 \leq i \leq 10}$$

with $K = 10$ components, so that (20) holds for all $\underline{\theta}_0^i$. Each $\underline{\theta}_0^i$ is generated as follows:

- For $k \in \{1, \ldots, K\}$, ν_k is generated under a truncated Gamma distribution with shape equal to 1.4 and scale equal to 10, with an upper bound set to 100.
- For $k \in \{1, \ldots, K-1\}$, $\boldsymbol{\mu}_{\cdot,k}$ is generated (independently) under a Dirichlet distribution with concentration parameter equal to 6 and a mean parameter set to $G_0 = (1/3, 1/3, 1/3)$, truncated to the region $\{\mathbf{w} \in \mathbf{S}_3 : \ \forall i \leq 3, \ w_i > \epsilon\}$ with $\epsilon = 1/100$.
- The first $K-1$ weights are constrained to be equal to each other and the location for the last kernel centre $\boldsymbol{\mu}_{\cdot,K}$ is set in in such a way that (20) holds while keeping the last weight ω_K as close to $1/K$ as possible.

For each $\underline{\theta}_0^i$, five data sets of size 80 each are generated under the corresponding Dirichlet mixture distribution. To avoid numerical errors, angular points with any coordinate less than 10^{-8} are rejected.

Error assessment for the mean differential scores. Since a lot of Monte-Carlo steps are involved in the differential score computations, the second part of Table 1 may only be interpreted as an order of magnitude for the errors. In the remainder of this subsection, an alternative *alt* denotes either the systematic choice of the PB or the NL model, or the model selection framework where the retained estimate is the one produce by the model with greatest posterior weight. If S is a scoring rule, $S(\mathrm{BMA}/alt)$ is the score difference between the BMAand the alternative.

Differential logarithmic score.
Here, we account for the error involved by the Monte-Carlo approximation (12).

For a given alternative *alt*, parameter $\underline{\theta}_0^i$ and data set $\mathbf{W}_j^i$, let $\hat{h}_j^i|_{\mathrm{BMA}}$ (*resp.* $\hat{h}_j^i|_{alt}$) the posterior predictive distributions in the BMA and in the alternative framework. Let $\hat{LS}(\hat{h}_j^i|_{\mathrm{BMA}})$ (*resp.* $\hat{LS}(\hat{h}_j^i|_{alt})$) be the Monte-Carlo estimate of the Logarithmic score as in (12), and let $\hat{\sigma}_j^i(\mathrm{BMA})$ (*resp.* $\hat{\sigma}_j^i(alt)$) be the classical Monte-Carlo error of the estimate. When i is fixed and j varies, the errors $\hat{\sigma}_j^i$ are not independent because they depend on the same Monte-Carlo sample. An estimated upper bound for the standard deviation of the differential score $\hat{LS}_j^i(\mathrm{BMA}/alt)$

is then $\hat{\sigma}_j^i = \hat{\sigma}_j^i(\mathrm{BMA}) + \hat{\sigma}_j^i(alt)$. This is conservative in the sense that this upper bound is only reached in the unrealistic case where $\hat{LS}_j^i(\mathrm{BMA})$ and $\hat{LS}_j^i(alt)$ have correlation equal to -1. In the same way, an upper bound for the standard deviation of the average (letting i fixed) is the average standard deviation: $\hat{\sigma}^i = \frac{1}{5}\sum_{j=1}^5 \hat{\sigma}_i^j$. Further, when i varies, the differential scores are independent from each other (*i.e.* if $i_1 \neq i_2$, $1 \leq j_1, j_2 \leq 5$, then $\hat{LS}_{j_1}^{i_1}(\mathrm{BMA}/alt)$ and $\hat{LS}_{j_2}^{i_2}(\mathrm{BMA}/alt)$ are independent). Consequently, an estimated upper bound for the variance of the average is $\hat{\sigma}(LS(\mathrm{BMA}/alt))^2 = \frac{1}{20^2}\sum_{i=1}^{20}\left[\hat{\sigma}^i\right]^2$. The errors reported in the first column, lines 9-11 of Table 1 are the squared roots of the latter quantity.

Failure region scores: $CRPS, PMCC$ **and** IS**.**
In this paragraph, the error concerns the approximation of the true probability of failure. Let $\hat{\delta}_0^i$, $\hat{\sigma}_0^i$ be respectively the mean Monte-Carlo estimate of the latter (see (4)), and its estimated standard deviation, for a given Dirichlet parameter $\underline{\theta}_0^i$. We define the boundaries of a typical centred error interval: $\delta_{\inf}^i = \hat{\delta}_0^i - \hat{\sigma}_0^i$, $\delta_{\sup}^i = \hat{\delta}_0^i + \hat{\sigma}_0^i$. Now, given a scoring rule S (one of the $CRPS, PMCC$ and IS rules) and an alternative alt, let $S^i(\mathrm{BMA}/alt, \delta_{\inf}^i)$ (*resp* $S^i(\mathrm{BMA}/alt, \delta_{\sup})$), be the mean differential score obtained between the BMA and framework alt, when the true failure probability is set to $\delta_{\inf}^i$ (*resp.* $\delta_{\sup}^i$). For example, for the CRPS differential score between the PB model and the NL model, we set $CRPS^i(\mathrm{BMA/PB}, \delta_{\sup}^i) = \frac{1}{5}\sum_{j=1}^5 CRPS(\hat{F}_j^i|_{\mathrm{BMA}}, \delta_{\sup}^i) - CRPS(\hat{F}_j^i|_{\mathrm{PB}}, \delta_{\sup}^i)$.

An order of magnitude for the fluctuation of the partially averaged score $S^i(\mathrm{BMA}/alt, \hat{\delta}^i)$ is

$$err(S^i, alt) = \left|S^i(\mathrm{BMA}/alt, \delta_{\inf}^i) - S^i(\mathrm{BMA}/alt, \delta_{\sup}^i)\right|/2\,.$$

The final score $\hat{S}(\mathrm{BMA}/alt)$ is the average over $i \in \{1, \ldots, 20\}$ of the $S^i(\mathrm{BMA}/alt, \hat{\delta}_0^i)$'s, and the errors are independent when i varies. The heuristic error magnitude reported in the three last lines and last columns of Table 1 are thus (up to multiplication by the factor appearing in the column titles)

$$err(S, alt) = \left(\frac{1}{20}\sum_{i=1}^{20}\left[err(S^i, alt)\right]^2\right)^{1/2}\,.$$

References

Apputhurai, P. and Stephenson, A. (2011). Accounting for uncertainty in extremal dependence modeling using bayesian model averaging techniques. *Journal of Statistical Planning and Inference*, 141(5):1800–1807.

Ballani, F. and Schlather, M. (2011). A construction principle for multivariate extreme value distributions. *Biometrika*, 98(3).

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of extremes: Theory and applications*. John Wiley & Sons: New York.

Berk, R. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58.

Boldi, M.-O. and Davison, A. C. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229.

Coles, S. and Tawn, J. (1991). Modeling extreme multivariate events. *JR Statist. Soc. B*, 53:377–392.

Cooley, D., Davis, R., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117.

Cowles, M. and Carlin, B. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, pages 883–904.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory, An Introduction*. Springer Series in Operations Research and Financial Engineering.

Einmahl, J., de Haan, L., and Piterbarg, V. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423.

Einmahl, J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *IN BAYESIAN STATISTICS*, pages 169–193. University Press.

Gneiting, T. and Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gudendorf, G. and Segers, J. (2011). Nonparametric estimation of an extreme-value copula in arbitrary dimensions",. *Journal of Multivariate Analysis*, 102:37–47.

Guillotte, S., Perron, F., and Segers, J. (2011). Non-parametric bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 73:377–406.

Gumbel, E. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173.

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

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM*, 24:233–245.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical science*, 14(4):382–401.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the american statistical association*, pages 773–795.

Kass, R., Tierney, L., and Kadane, J. (1990). The validity of posterior expansions based on laplace's method. *Bayesian and Likelihood methods in Statistics and Econometrics*, 7:473–488.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Ledford, A. and Tawn, J. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Raftery, A., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174.

Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):219–241.

Resnick, S. (1987). *Extreme values, regular variation, and point processes, volume 4 of Applied Probability. A Series of the Applied Probability Trust.* Springer-Verlag, New York.

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

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Verlag.

Stephenson, A. (2003). Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59.

Tawn, J. (1990). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245.

van der Vaart, A. (2000). *Asymptotic statistics (Cambridge series in statistical and probabilistic mathematics).* Cambridge University Press.

(Anne Sabourin) Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ, 91191 Gif-sur-Yvette, France or
Universite de Lyon, CNRS UMR 5208,Universite de Lyon 1, Institut Camille Jordan ,43 blvd. du 11 novembre 1918, F-69622 Villeurbanne cedex , France
E-mail address: anne.sabourin@lsce.ipsl.fr

(Philippe Naveau) CNRS-CEA-UVSQ, 91191 Gif-sur-Yvette, France
E-mail address: philippe.naveau@lsce.ipsl.fr

(Anne-Laure FOUGÈRES) Universite de Lyon,CNRS UMR 5208, Universite de Lyon 1, Institut Camille Jordan,43 blvd. du 11 novembre 1918; F-69622 Villeurbanne cedex, France
E-mail address: fougeres@math.univ-lyon1.fr