Bayesian inference from composite likelihoods, with an application to spatial extremes

D. Cooley[†], A. Davison[‡], M. Ribatet[*]

[†] Department of Statistics, Colorado State University, USA

[‡] Institute of Mathematics, EPFL, Switzerland

[*] Department of Mathematics, University of Montpellier 2, France

Motivation/Key questions

MOTIVATIONS:

- The likelihood is central to both frequentist and Bayesian inference
- But in many settings it may not be feasible to calculate it
- In a frequentist framework, one can benefit using composite likelihood
- However there is much less guidance in a Bayesian framework

KEY QUESTIONS

- Is it possible to use composite likelihood within a "Bayesian framework" ?
- If so what do we actually get? A kind of "posterior distribution"?
- Has this "posterior distribution" a statistical meaning?

Outline

Institut de Mathématiques et de Modélisation de Montpellier

1. Composite posterior distribution

- For a set of events $\{\mathcal{A}_i : i \in I\}$ and positive weights $\{w_i : i \in I\}$, $I \subset \mathbb{N}$, the corresponding composite likelihood is

$$L_c(\theta; y) = \prod_{j=1}^{n} \prod_{i \in I} f(y_j \in \mathcal{A}_i; \theta)^{w_i}.$$

- By analogy with the usual posterior distribution, a composite posterior distribution is

$$\pi_c(\theta \mid y) = \frac{L_c(\theta; y)\pi(\theta)}{\int L_c(\theta; y)\pi(\theta)\mathrm{d}\theta}, \qquad \pi(\cdot) \text{ prior distribution.}$$

☞ This is just a definition—not Bayes' theorem!

- Under mild conditions $\int L_c(\theta; y)\pi(\theta)\mathrm{d}\theta < \infty$, e.g., $\sup_\theta f(y \in \mathcal{A}_i; \theta) < \infty$, $i \in I$.

2. A motivating example

The seeming assumption *of independence between the $\mathcal{A}_i$ give misleading inference*

Example 1 (Gaussian process/Multivariate normal)**.** Consider the following model

$$Y \mid \mu, \tau \sim \text{Gauss. Proc.}(\mu, \gamma),$$
$$\mu \sim N(a, b),$$
$$\tau \sim \text{IG}(c, d),$$

where the covariance function $\gamma(h) = \tau \exp(-h/\omega)$, ω known. Suppose we observe this Gaussian process at k locations $x_1, \ldots, x_k$.

Because the priors distributions are conjugate, it is easily shown that the full conditional distribution are

$$\pi(\mu \mid \cdots) \sim N(\tilde{\mu}, \tilde{\sigma}^2),$$

$$\pi(\tau \mid \cdots) \sim IG\left\{c + \frac{k}{2}, d + \frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right\},$$

where

$$\tilde{\sigma}^2 = (b^{-1} + \tau^{-1} 1^T \Sigma^{-1} 1)^{-1}, \quad \tilde{\mu} = \tilde{\sigma}^2 (ab^{-1} + \tau^{-1} 1^T \Sigma^{-1} y),$$

$$\Sigma = \{\gamma(x_i - x_j)\}_{i,j}.$$

The full conditional pairwise distributions are

$$\pi_p(\mu \mid \cdots) \sim N(\tilde{\mu}_p, \tilde{\sigma}_p^2),$$
$$\pi_p(\tau \mid \cdots) \sim IG\left\{c + \frac{k(k-1)}{2}, d + \frac{1}{2}(y_p - \mu)^T \Sigma_p^{-1} (y_p - \mu)\right\},$$

with $\tilde{\sigma}_p^2 = (b^{-1} + \tau^{-1} 1^T \Sigma_p^{-1} 1)^{-1}$ and $\tilde{\mu} = \tilde{\sigma}_p^2 (ab^{-1} + \tau^{-1} 1^T \Sigma_p^{-1} y_p)$ where Σ_p is a block diagonal matrix with blocks

$$\begin{bmatrix} 1 & \tau^{-1}\gamma(x_i - x_j) \\ \tau^{-1}\gamma(x_i - x_j) & 1 \end{bmatrix}, \qquad 1 \leq i < j \leq k,$$

and

$$y_p = (y_1, y_2, y_1, y_3, \ldots, y_1, y_k, y_2, y_3, \ldots, y_2, y_k, \ldots, y_{k-1}, y_k).$$

In particular when τ is fixed,

$$\frac{\tilde{\sigma}_p^2}{\tilde{\sigma}^2} \leq \frac{(1+\tau)(\tau+bk)}{(1+\tau)\tau + b\tau k(k-1)} \downarrow 0, \qquad k \to \infty.$$

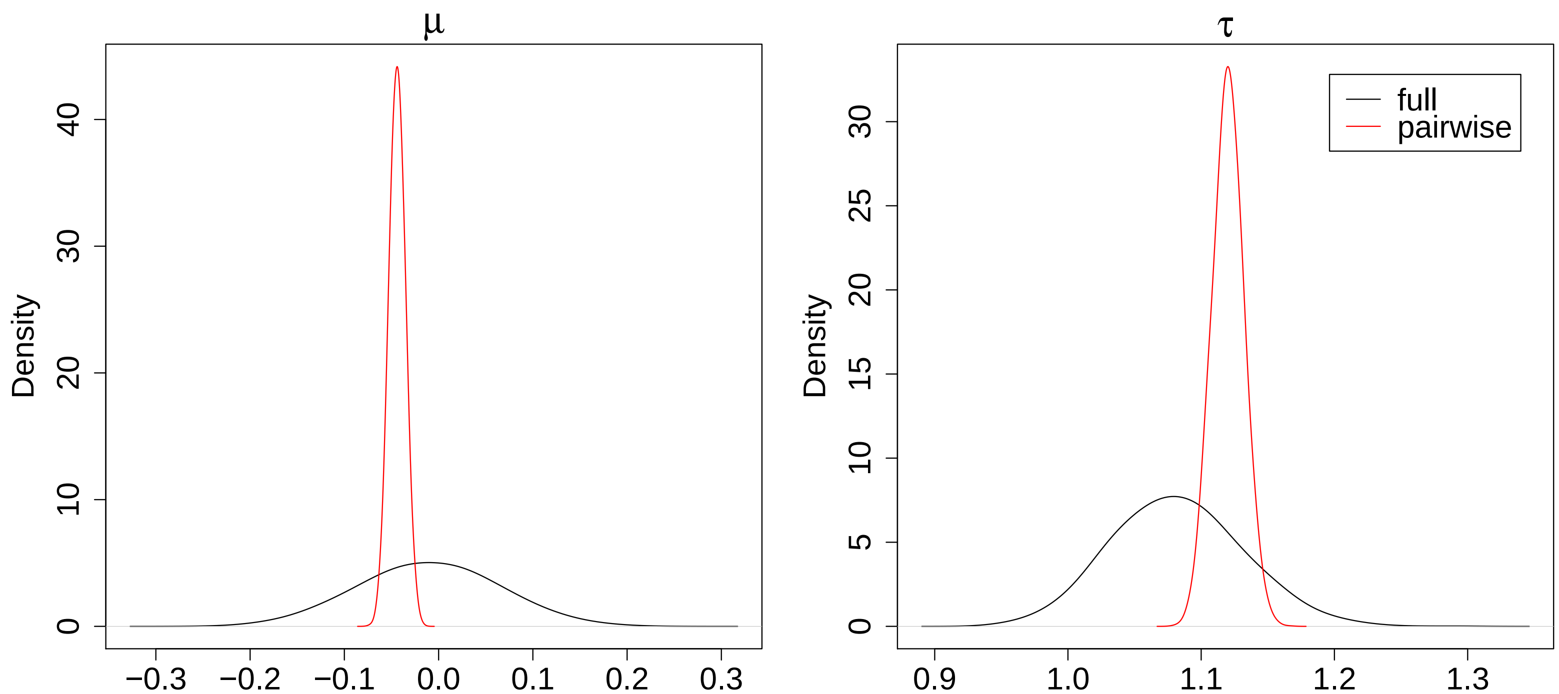

Figure 1: *Marginal full and pairwise posterior densities for example 1.*

- The previous example clearly shows that the composite posterior distributions will be too narrow
- Our goal is to adjust the composite posterior distribution such that it has a statistical meaning, i.e., credible intervals with reasonable coverage
- The adjustments are based on the developments of Chandler and Bate (2007) and Rotnitzky and Jewell (1990) on hypothesis testings under missspecification

☞ Note that we won't try to approximate the full posterior distribution! (although some information is captured by L_c)

3. Adjustment of the composite posterior distribution

Guideline for an adjustment

- The adjustments rely on the same asymptotic argument
- Since we are working under misspecification we have

$$2\left\{\ell_c(\hat{\theta}_c) - \ell_c(\hat{\theta}_{\psi_0})\right\} \xrightarrow{\text{d}} \sum_{i=1}^{p} \lambda_i X_i, \qquad n \to \infty, \qquad (1)$$

while under the usual setting we have

$$2\left\{\ell(\hat{\theta}) - \ell(\hat{\theta}_{\psi_0})\right\} \xrightarrow{\text{d}} \chi_p^2, \qquad n \to \infty. \qquad (2)$$

- The idea consists in "tweaking" L_c such that (2) (approximately) holds.

Magnitude/Vertical Adjustment (R & J)

- The magnitude adjustment for the composite posterior distribution is

$$\pi_{\text{magn}}(\theta \mid y) \propto L_c(\theta; y)^k \pi(\theta),$$

  with $k = p / \sum_{i=1}^p \lambda_i$ where λ_i are the eigenvalues of $-\mathbb{E}[\nabla^2 \ell_c(\theta_0)]^{-1}\text{Var}[\nabla \ell_c(\theta_0)]$.

- By construction we have

$$\mathbb{E}\left[2\left\{\ell_{\text{magn}}(\hat{\theta}) - \ell_{\text{magn}}(\hat{\theta}_0)\right\}\right] \longrightarrow k \sum_{i=1}^p \lambda_i = p = \mathbb{E}[\chi_p^2],$$

  but higher moments will differ from that of a χ_p^2.

☞ This adjustment can be thought as a *tempering* of L_c to flatten the composite likelihood function.

Curvature adjustment (C & B)

- Another approach is to modify the curvature of L_c around $\hat{\theta}_c$, i.e.,

$$\pi_{\text{curv}}(\theta \mid y) \propto L_{\text{curv}}(\theta; y)\pi(\theta),$$

where

$$L_{\text{curv}}(\theta; y) = L_c\{\hat{\theta}_c + C(\theta - \hat{\theta}_c); y\}, \qquad C^T H C = H J^{-1} H,$$

with $H = -\mathbb{E}[\nabla^2 \ell_c(\theta_0; Y)]$ and $J = \text{Var}[\nabla \ell_c(\theta_0; Y)]$.

- By construction

$$n^{-1}\nabla^2 \ell_{\text{curv}}(\hat{\theta}_c; y) \xrightarrow{\text{a.s.}} H J^{-1} H, \qquad n \to \infty,$$

and the usual asymptotic distribution of the likelihood ratio statistic holds—not only the first moment.

Asymptotic posterior distributions

It is not difficult to show that

- the posterior distributions are approximately (n large enough, usual regularity conditions)

$$\pi_c(\theta \mid y) \dot\sim N(\theta_0, n^{-1}H^{-1})$$

$$\pi_{\text{magn}}(\theta \mid y) \dot\sim N\{\theta_0, (np)^{-1}\text{tr}(H^{-1}J)H^{-1}\},$$

$$\pi_{\text{curv}}(\theta \mid y) \dot\sim N(\theta_0, n^{-1}H^{-1}JH^{-1}).$$

- In particular we do not approximate the true posterior $\pi(\theta \mid y)$.

MCMC samplers

- Standard MCMC samplers can be used once the adjusted composite likelihood is known, e.g.,

Algorithm 1: Adjusted Metropolis–Hastings algorithm.

Input : $\hat\theta_c$, $\hat H(\hat\theta_c)$, $\hat J(\hat\theta_c)$, $\theta_1 \in \Theta$, a proposal distribution $q(\cdot \mid \theta)$ and an adjusted composite likelihood $L_{\text{adj}}(\cdot; y)$
Output: A realisation of length $N+1$ from a Markov chain

for $t \leftarrow 1$ **to** N **do**
  $\theta^{(p)} \sim q(\cdot \mid \theta^{(t)})$;
  $\alpha_{\text{adj}}(\theta^{(t)}, \theta^{(p)}) \leftarrow \min\left\{1, \frac{L_{\text{adj}}(\theta^{(p)};y)\pi(\theta^{(p)})q(\theta^{(t)}|\theta^{(p)})}{L_{\text{adj}}(\theta^{(t)};y)\pi(\theta^{(t)})q(\theta^{(p)}|\theta^{(t)})}\right\}$;
  $U \sim U(0,1)$;
  if $\alpha_{\text{adj}}(\theta^{(t)}, \theta^{(p)}) \leq U$ **then**
    $\theta^{(t+1)} \leftarrow \theta^{(p)}$;
  else
    $\theta^{(t+1)} \leftarrow \theta^{(t)}$;
  end
end
return $\{\theta^{(t)}\}_{t=1,\ldots,N+1}$;

- The adjusted Gibbs sampler can be implemented similarly but an adaptive version can be implemented.

i3m Institut de Mathématiques et de Modélisation de Montpellier

Algorithm 2: Adaptive adjusted Gibbs sampler.

Input : $\theta^{(1)} \in \Theta$
Output: A realisation of length $N+1$ from a Markov chain

for $t \leftarrow 1$ **to** N **do**
  for $j \leftarrow 1$ **to** G **do**
    Get the restricted maximum composite likelihood estimate $\hat{\theta}_{j,c}$ with θ_{-j} held fixed at $\theta_{-j}^{(t)}$;
    Get $\hat{H}_{j,j}(\hat{\theta}_j) = \nabla^2 \ell_c(\hat{\theta}_{j,c} \mid \theta_{-j}^{(t)}, y)$ and $\hat{J}_{j,j}(\hat{\theta}_j)$, the sample covariance matrix of $\nabla \ell_c(\hat{\theta}_{j,c} \mid \theta_{-j}^{(t)}, y_i), i = 1, \ldots n$, and define the adjusted composite log-likelihood $\ell_{\text{adj}}(\theta_j \mid \theta_{-j}^{(t)}, y)$ from either (8) or (10);
    Draw $\theta_j^{(t+1)}$ from $L_{\text{adj}}(\theta_j; y, \theta_{-j}^{(t)})\pi(\theta_j \mid \theta_{-j})$ (using Metropolis–Hastings updates if necessary);
  end
end
return $\{\theta^{(t)}\}_{t=1,\ldots,N+1}$;

- This algorithm samples from a well-defined posterior distribution since it considers the completion

$$\pi(\theta, \hat{\theta} \mid y) \propto \prod_{j=1}^{p} \underbrace{\pi(\hat{\theta}_j \mid \theta, y)}_{j\text{-th completion}} \underbrace{\pi(\theta \mid y)}_{\text{target density}},$$

where $\pi(\hat{\theta}_j \mid \theta, y) = \delta_{\arg\max L_c(\theta_j \mid \theta_{-j}, y)}(\hat{\theta}_j)$.

4. Simulation Study

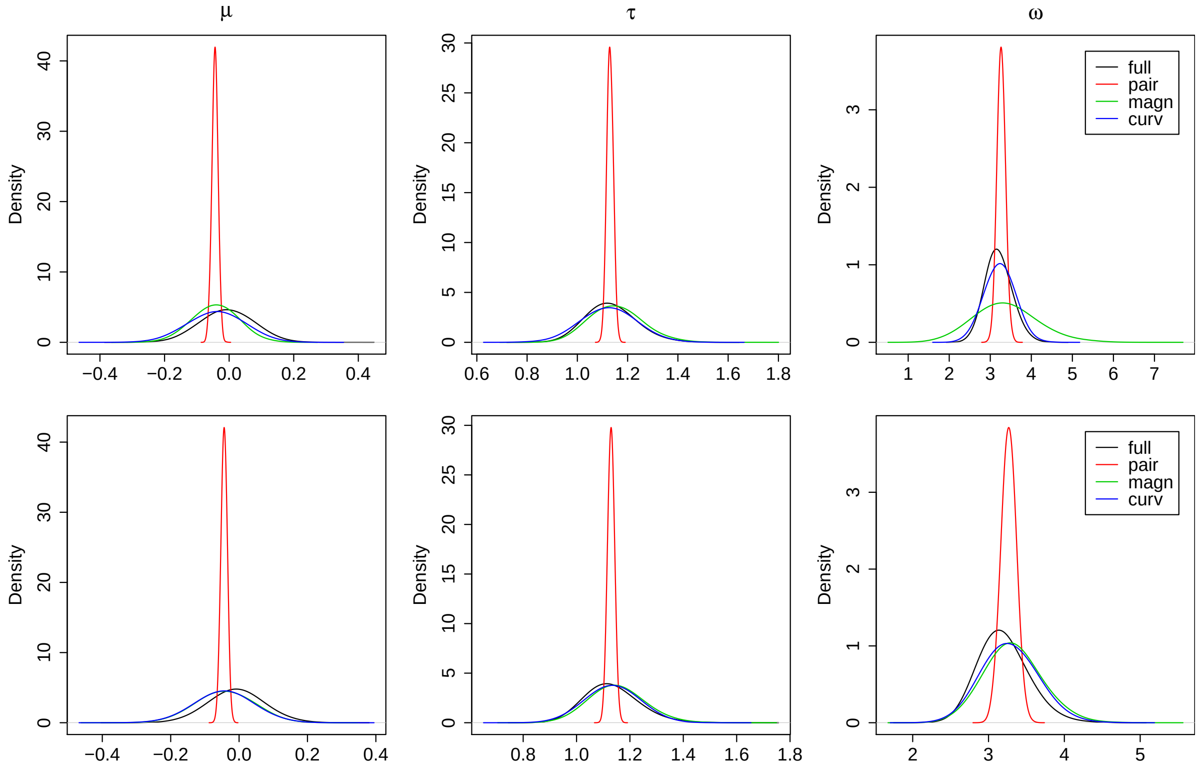

Figure 2: *Comparison between the marginal full posterior (black), the marginal pairwise posterior (red) and the marginal adjusted pairwise posterior densities based on the magnitude (green) and curvature (blue) adjustments. The posterior distributions are derived from $n = 50$ realisations of a Gaussian process having an exponential covariance function with $\mu = 0$, $\tau = 1$ and $\omega = 3$ and observed at $K = 20$ locations. Top row: Metropolis–Hastings algorithm. Bottom row: Adaptive adjusted Gibbs sampler.*

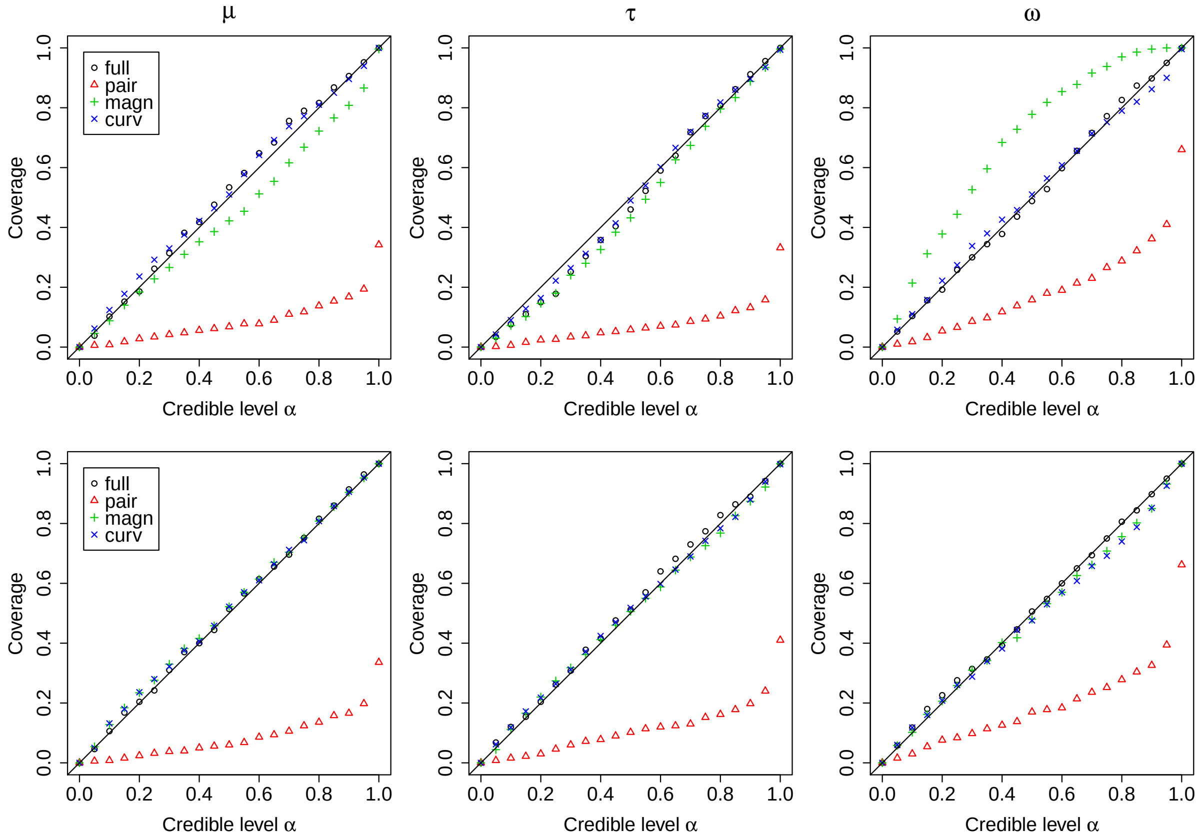

Figure 3: *Variation of the empirical coverages with the credible level* α*, based on* 500 *replicates of the Gaussian process simulation with* $\mu = 0$*,* $\tau = 1$ *and* $\omega = 3$*, for the full, the non adjusted pairwise and the magnitude/curvature adjusted posteriors. Top row: Overall Gibbs sampler. Bottom row: Adaptive Gibbs sampler.*

5. Extreme precipitations around Zurich

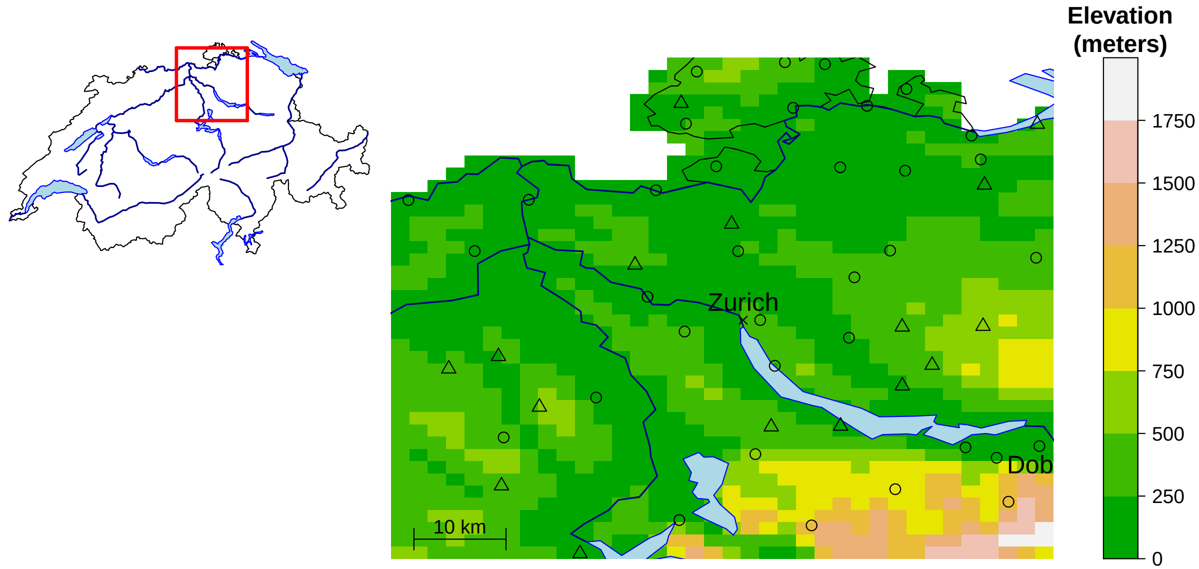

Figure 4: *Map of the study region. The stations used for inference/validation are depicted by circles/triangles.*

- Maximum daily rainfall amounts (mm)
- 51 stations (16 kept for validation)
- Record period: 1962–2008
- Elevation: 322 to 910 meters a.m.s.l.

Model

$$\begin{aligned}
Y \mid \mu, \sigma, \xi, \Sigma &\sim \text{Smith's max-stable model},\\
\mu \mid \beta_\mu, \tau_\mu, \omega_\mu &\sim \text{Gauss. proc.}(X_\mu\beta_\mu, \gamma_\mu),\\
\log\sigma \mid \beta_\sigma, \tau_\sigma, \omega_\sigma &\sim \text{Gauss. proc.}(X_\sigma\beta_\sigma, \gamma_\sigma),\\
\xi \mid \beta_\xi, \tau_\xi, \omega_\xi &\sim \text{Gauss. proc.}(X_\xi\beta_\xi, \gamma_\xi),\\
\text{Priors} &\sim \text{Ind. uninf. (but proper)},
\end{aligned}$$

where μ, σ, ξ are random functions specifying the GEV parameters for all $x \in \mathcal{X}$, $X.$ are design matrices, $\beta.$ regression coefficients and $\gamma.$ covariance functions.

- This model will be compared to
  - a BHM with a condional independence assumption (independence likelihood);
  - a max-stable model with deterministic trend surfaces.

Realization from each model

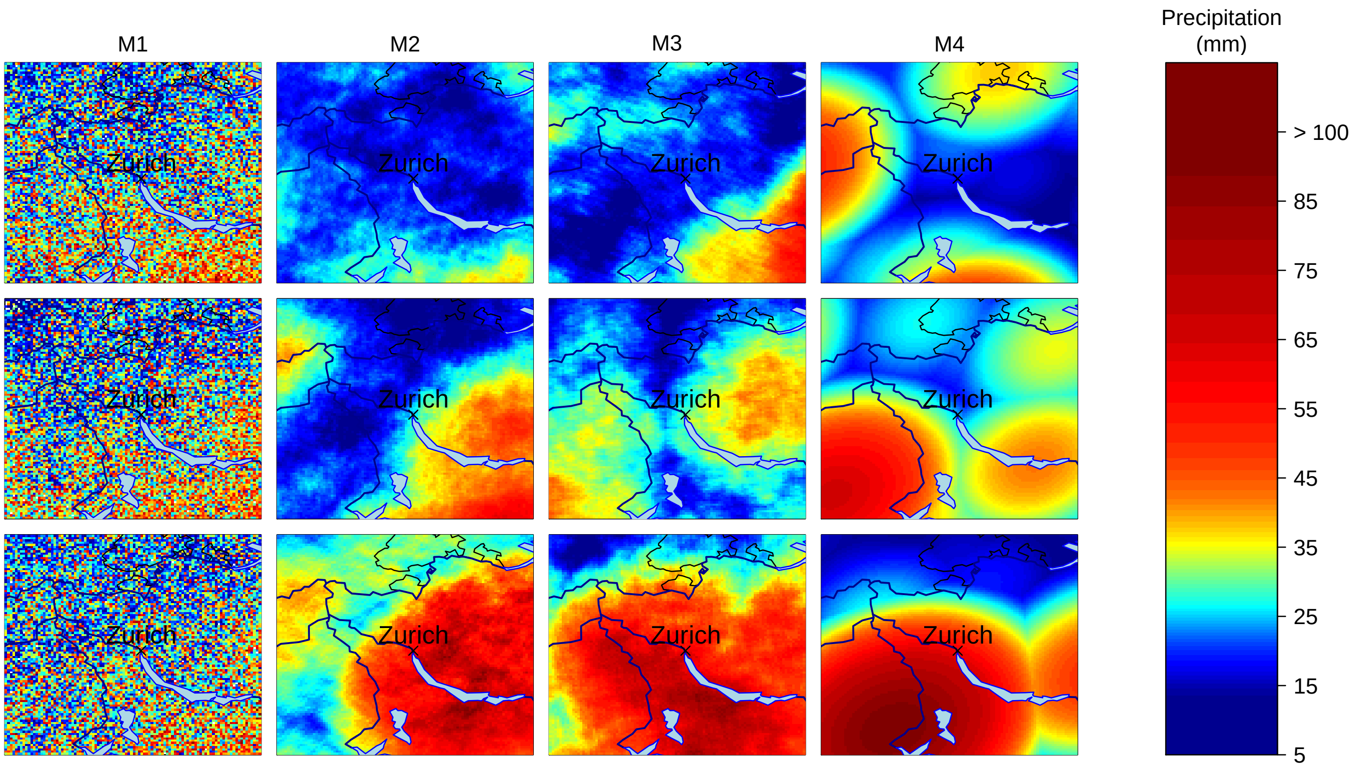

Figure 5: *Three realizations of random fields over the study region for the conditional independent model (M1), hierarchical models without any adjustment (M2) and with the curvature adjustment (M3) and a simple max-stable model with deterministic trend surfaces (M4). The three rows show realizations corresponding to different risk scenarios according to the values of* S_{Zurich} *expected to be exceeded once every* 1.05, 2 *and* 20 *years (from top to bottom).*

Distribution of $\sup_{x \in b(x_0, r)} Y(x)$ for each model

Institut de Mathématiques et de Modélisation de Montpellier

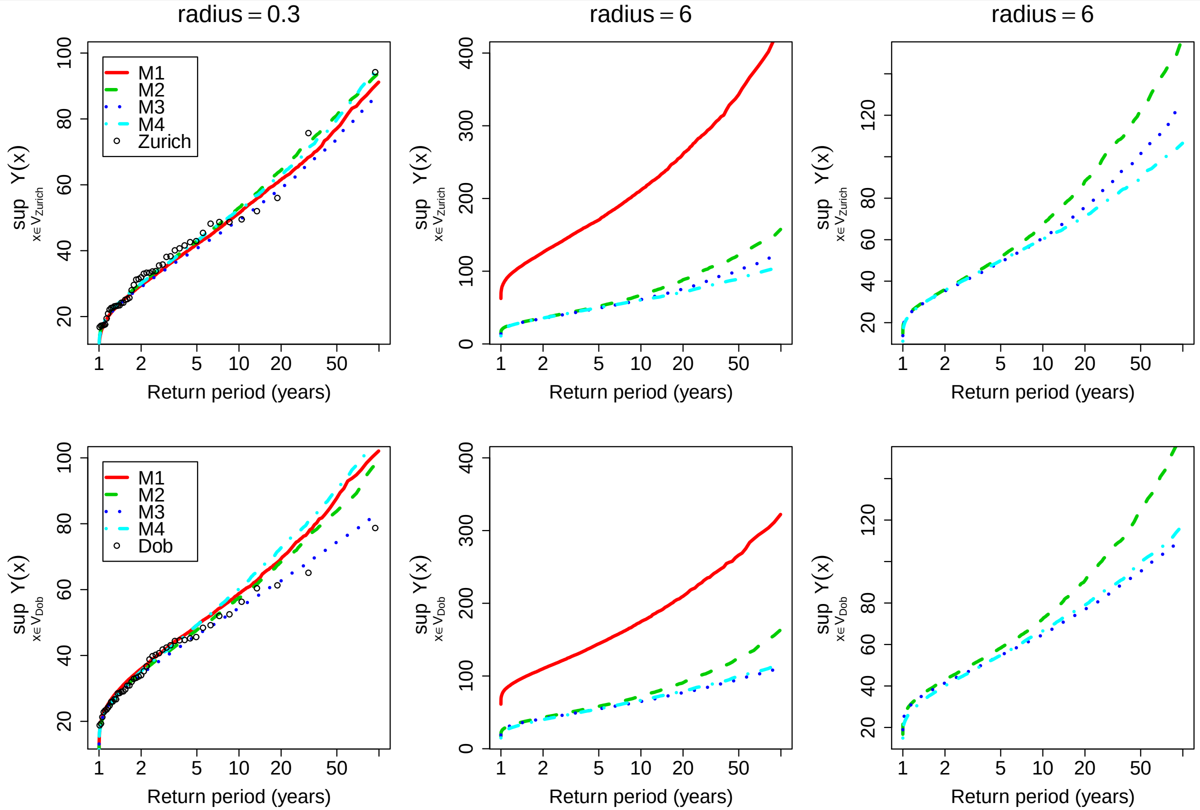

Figure 6: *Comparison between the return level curves (mm) computed on neighborhoods centered at the Zurich (top) and DOB gauging stations (bottom) and having radius* 0.3 *and* 6 *km (left and middle panels) for the conditional independent model (M1), the hierarchical models without any adjustment (M2) and with the curvature adjustment (M3) and a simple max-stable model with deterministic trend surfaces (M4). The left panels compares the return level curves to the observations available at the gauging stations. The right panel is the same as the middle one but shows only the max-stable based models.*

6. References

Chandler, R. E. and Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, 94(1):167–183.

Monahan, J. and Boos, D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika*, 79(2):271–278.

Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, 21:149–164.

Ribatet, M., Cooley, D., and Davison, A. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22:813–845.

Rotnitzky, A. and Jewell, N. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77:495–497.

Survey on the French presidential election. . .

"On election day, are you more likely to vote for the candidate who gives you a headache or the candidate who gives you a stomachache?"

THANK YOU !