Statistical Methods for Understanding Hydrologic Change

Bocci, Chiara
Department of Statistics "Giuseppe Parenti", University of Firenze
Viale Morgagni 59
Firenze (50134), Italy
E-mail: bocci@ds.unifi.it

Caporali, Enrica
Department of Civil and Environmental Engineering, University of Firenze
Via S. Marta 3, I
Firenze (50139), Italy
E-mail: enrica.caporali@unifi.it

Petrucci, Alessandra
Department of Statistics "Giuseppe Parenti", University of Firenze
Viale Morgagni 59
Firenze (50134), Italy
E-mail: alessandra.petrucci@unifi.it

Introduction

Extreme value models and techniques are widely applied in environmental studies to define protection systems against the effects of extreme levels of environmental processes. Environmental extreme events such as floods, earthquakes, hurricanes, may have a massive impact on everyday life for the consequences and damage that they cause. For this reason there is considerable attention in studying, understanding and predicting the nature of such phenomena and the problems caused by them, not least because of the possible link between extreme climate events and climate change. Regarding the matter related to the climate change science, a certain importance is cover by the implication of changes in the hydrological cycle.

Among all hydrologic processes, rainfall is a very important variable as it is a fundamental component of flood risk mitigation and drought assessment, as well as water resources availability and management. Specific technical studies, summarized in the most recent Intergovernmental Panel on Climate Change reports (Pachauri and Reisinger, 2007; Bates et al., 2008) and by the Environmental European Agency (EEA, 2007) highlight several emergencies related to water that the community has to face in the next future to cope with a changing climate. Furthermore special attention has been paid to the potential changes in the extreme events that could accompany global climate change; as it constitutes a primary concern in estimating the impacts of climate change. Extreme events, such as heat waves, heavy rain, hailstorms, snowfall, and droughts, are in fact responsible for a large part of climate-related damages (Meehl et al., 2000), and their impact is of great concern for the community and stakeholders (Easterling et al., 2000). A number of theoretical modeling and empirical analyses have also suggested that notable changes in the frequency and intensity of extreme events, including intense rainfall and floods, may occur even when there are only small changes in climate (Katz and Brown, 1992; Wagner, 1996).

In this framework, in the past two decades there has been an increasing interest for statistical methods that model rare events (Coles, 2001; Smith, 2003). Statistical modeling of extreme values has flourished since about the mid-1980s. Such analysis, for instance, can help by estimating both the rate and magnitude of rare events, so that precautionary measures can be taken to prevent catastrophic phenomena, plan for their impact and mitigate their effects.

The Generalized Extreme Value distribution (GEV) is widely adopted model for extreme events in the univariate context. It's motivation derives from asymptotic arguments that are based on reasonably wide classes of stationary processes.

For modeling extremes of non-stationary sequences it is commonplace to still use the GEV as a basic model, but to handle the issue of non-stationarity by regression modeling of the GEV parameters. Traditionally this has been done using parametric models (Coles 2001, chapter 6), but there has been considerable recent interest in the possibility of nonparametric or semiparametric modeling of extreme value model parameters. For example, Davison and Ramesh (2002) and Chavez-Demoulin and Davison (2005) have demonstrated the usefulness of nonparametric regression, or smoothing, for certain types of extreme value models. The former used a local likelihood approach, while the latter used smoothing splines. Nevertheless, the literature on smoothing in extremes models remains scarce and in its infancy. Whilst the theory and statistical practice of univariate extremes is well developed, there is much less guidance for the modeling of spatial extremes. This creates problems because many environmental processes - such as rainfall have a natural spatial domain. The spatial analogue of univariate or multivariate extreme value models is the class of max-stable processes. (e.g de Haan and Pickands, 1986; Resnick, 1987). Max-stable processes were first developed by de Haan (1984) and have a similar asymptotic motivation, but expanded to a spatial domain, as the GEV distribution in the univariate case. They provide a general and useful approach to model extreme processes incorporating temporal or, more commonly, spatial dependence. On the statistical side, a parametric class of max-stable processes, together with a simple approach for inference, is provided by Smith (1990). Statistical methods for max-stable processes and data analysis of practical problems are discussed further by Coles (1993) and Coles and Tawn (1996). However, likelihood methods for such models are complicated by the intractability of density functions in all but the most trivial cases, although some alternative nonparametric estimators have been proposed by de Haan and Pereira (2006).

Recently Padoan and Wand (2008) propose the use of mixed model-based splines for extremal models developing nonparametric estimation for a smoothly varying location parameter within the GEV model. A compelling feature of this approach is that the smoothing parameters correspond to variance components, so maximum likelihood or Bayesian techniques can be applied for model fitting, assessment and inference (e.g. Ruppert,Wand and Carroll, 2003).

Here we implement a geoadditive mixed model for extremes with a temporal random effect. We assume that the observations follow generalized extreme value distributions whose locations are spatially dependent where the dependence is captured using the geoadditive model. The analyzed territory is the catchment area of Arno River in Tuscany in Central Italy. The dataset is composed by the time series of annual maxima of daily rainfall recorded in about 400 rain gauges, spatially distributed over an area of about 8.830 km^2. The record period covers mainly the second half of 20th century.

The characteristics of the dataset to which we apply our model are presented in the next section. Then we outline the model and display the estimation of the parameters and the preliminary results are discussed. We conclude with some brief remarks.

Dataset Description

The investigation is developed on the catchment area of Arno River almost entirely situated within Tuscany, Central Italy. The area is expected to suffer significantly from global climate change (Burlando and Rosso, 2002). The area is characterized by a climate that ranges from temperate to Mediterranean maritime, and by a complex physical topography. It presents plain areas near the sea and around the main metropolitan areas, hilly internal zones, and the mountainous area of the

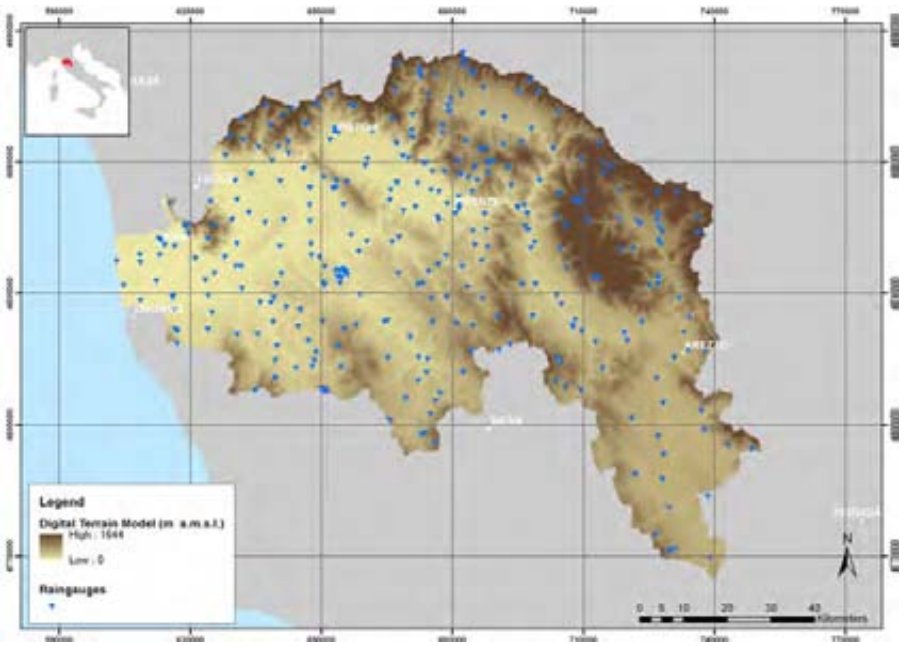

Figure 1: Geographical location and orography (i.e. Digital Terrain Model) of the catchment area of Arno River in Central Italy.

Apennines (Figure 1). The river is 241 km long and the catchment area is of about 8830 km^2 and has a mean elevation of 353 m a.m.s.l..

The precipitation regime is greatly influenced by the topography. Total annual precipitation ranges from 720 mm to 1690 mm (Figure 2). Heavy storms mainly occur in autumn following dry summers. Most of the territory of Arno River basin have suffered in the past from many severe hydro-geological events (Fatichi and Caporali, 2009), with high levels of risk due to the vulnerability of a unique artistic and cultural heritage (Caporali et al., 2004).

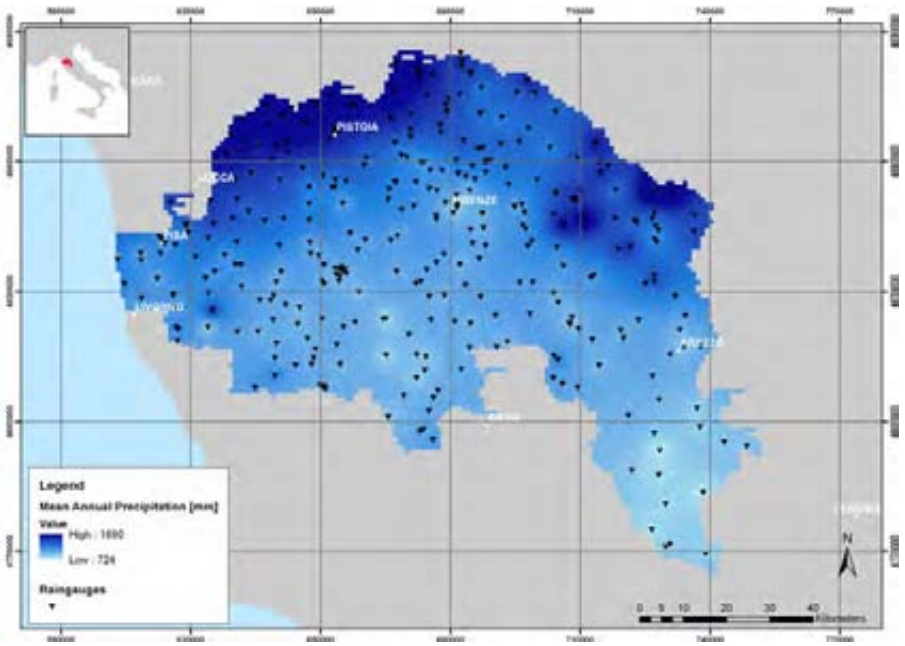

Figure 2: Average Total Annual Precipitation and rain gauges distribution of Arno River basin (daily precipitation dataset; recorded period 1916-2008).

The time series of annual maxima of daily rainfall recorded in 415 rain gauges are analysed. The registrations cover the period 1916-2008 and the available rain gauges series length ranges from 1 to 81 years. Using the time series minimum length suggested by WMO (1983), only stations with at least 30 hydrologic years of data, even not consecutive, were considered. In addition, in order to have enough rain gauges observations to estimate each year specific effect, we reduce the time series length to the post Second World War period: 1951-2000. The final dataset is composed by the data recorded from 1951 to 2000 at 118 rain gauges for a total of 4903 observations.

Geoadditive Mixed Models for Sample Extremes

Extreme value theory begins with a sequence $Y_1, Y_2, \ldots$ of independent and identically distributed random variables and, for a given n asks about parametric models for $M_n = maxY_1, \ldots, Y_n$. If the distribution of the Y_i is specified, the exact distribution of M_n is known. In the absence of such specification, extreme value theory considers the existence of $\lim_{n\to\infty} P\left[\frac{M_n - b_n}{a_n} \leq y\right] \equiv F(y)$ for two sequences of real numbers $a_n > 0$, b_n. If $F(y)$ is a non-degenerate distribution function, it belongs to either the Gumbel, the Fréchet or the Weibull class of distributions, which can all be usefully expressed under the umbrella of the GEV(μ, ψ, ξ) .

$$(1) \quad F(y; \mu, \psi, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{y-\mu}{\psi}\right)\right]^{\frac{1}{\xi}}\right\}, \qquad -\infty < \mu, \xi < \infty, \psi > 0$$

for $y : 1 + \xi\frac{(y-\mu)}{\psi} > 0$ and μ, ψ and ξ are respectively location, scale and shape parameters. The GEV distribution is heavy-tailed and its probability density function decreases at a slow rate when the shape parameter ξ is positive. On the other hand, the GEV distribution has a bounded upper tail for a negative shape parameter. Note that n is not specified; the GEV is viewed as an approximate distribution to model the maximum of a sufficiently long sequence of random variables.

Now suppose we observe n sample maxima $y_1, \ldots, y_n$ as well as corresponding covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The y_i are obtained from approximately equi-sized samples of a variable of interest. A common situation is y_i corresponding to the annual maximum of a daily measurement, such as rainfall in a particular town, for year i $(1 \leq i \leq n)$. General GEV regression models (e.g. Coles, 2001) take the form

$$y_i|\mathbf{x}_i \sim \text{GEV}(\mu(\mathbf{x}_i), \psi(\mathbf{x}_i), \xi(\mathbf{x}_i))$$

where, for example, $\mu(\mathbf{x}_i) = g([\mathbf{X}\boldsymbol{\beta}]_i)$, g is a link function, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\mathbf{X}$ is a design matrix associated with the $\mathbf{x}_i$s. Similar structures may be imposed upon $\psi(\mathbf{x}_i)$ and $\xi(\mathbf{x}_i)$. The regression coefficients can be estimated via maximum likelihood. The classic literature illustrate GEV regression with parametric models, however recent works present more flexible non-parametric approaches (Chavez-Demoulin and Davison (2005)).

Padoan and Wand (2008) discuss how generalized additive models (GAM) with penalized splines can be carried out in a mixed model framework for the GEV family. Assuming that the location parameter in the GEV distribution is smooth on an interval $[a, b]$ in the x_i domain then the simplest time-nonhomogeneous nonparametric regression model is given by

$$y_i|x_i \sim \text{GEV}(\mu(x_i), \psi, \xi)$$

with a mixed model-based penalised spline model for μ

$$\eta(x) = g(\mu(x)) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k z_k(x), \qquad u_1, \ldots, u_K \text{ i.i.d. } N(0, \sigma_u^2)$$

where g is a link function and $z_1, \ldots, z_K$ is an appropriate set of spline basis functions.

Let $\mathbf{y} = (y_1, \ldots, y_n)$ and define the design matrices $\mathbf{X} = [1\ x_i]_{1 \le i \le n}$, $\mathbf{Z} = [z_k(x_i)]_{1 \le i \le n, 1 \le k \le K}$ associated with fixed effects $\boldsymbol{\beta} = [\beta_0, \beta_1]$ and random effects $\mathbf{u} = [u_1, \ldots, u_K]$. Given $\mathbf{u}$, the y_i are conditionally independent with distribution,

$$\mathbf{y}|\mathbf{u} \sim \mathrm{GEV}(g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}), \psi, \xi). \tag{2}$$

Note that $\boldsymbol{\mu} \equiv g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})$ is related to the conditional mean of $\mathbf{y}$ given $\mathbf{u}$ via

$$\mathrm{E}(\mathbf{y}|\mathbf{u}) = \begin{cases} \boldsymbol{\mu} + \mathbf{1}\psi\left[\Gamma(1-\xi) - 1\right]/\xi & \text{for } \xi \neq 0 \\ \boldsymbol{\mu} + \mathbf{1}\psi\gamma & \text{for } \xi = 0 \end{cases}$$

where $\mathbf{1}$ is a vector of n one values, Γ is the Gamma function and $\gamma = 0.57721566...$ is Euler's constant.

The addition of other explicative variables in regression model (2) is straightforward: smoothing components and random effect components are added in the random effects term $\mathbf{Z}$, while linear components can be incorporated as fixed effects in the $\mathbf{X}$ term. Moreover, the mixed model structure provides a unified and modular framework that allows to easily extend the model to include various kind of generalization and evolution.

Geoadditive models, introduced by Kammand and Wand (2003), are a particular specification of GAM that models the spatial distribution of y with a bivariate penalized spline on the spatial coordinates. Suppose to observe n sample maxima y_{ij} at spatial location $\mathbf{s}_{ij}$, $\mathbf{s} \in \mathbb{R}^2$, $j = 1, \ldots, p$ and at time $i = 1, \ldots, t$. In order to model both the spatial and the temporal influence on the annual rainfall maxima, we consider a geoadditive mixed model for extremes with a temporal random effect:

$$\begin{cases} y_{ij}|s_{ij} \sim \mathrm{GEV}(\mu(s_{ij}), \psi, \xi) \\ \mu(\mathbf{s}_{ij}) = \beta_0 + \mathbf{s}_{ij}^T\boldsymbol{\beta}_s + \sum_{k=1}^{K} u_k b_{tps}(\mathbf{s}_{ij}, \boldsymbol{\kappa}_k) + \gamma_i, \end{cases} \tag{3}$$

where b_{tps} are the low-rank thin plate spline basis functions with K knots and γ_i is the time specific random effect. The model (3) can be written as a mixed model

$$\mathbf{y}|(\mathbf{u}, \boldsymbol{\gamma}) \sim \mathrm{GEV}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{D}\boldsymbol{\gamma}, \psi, \xi). \tag{4}$$

with

$$\mathrm{E}\begin{bmatrix}\mathbf{u} \\ \boldsymbol{\gamma}\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix}, \qquad \mathrm{Cov}\begin{bmatrix}\mathbf{u} \\ \boldsymbol{\gamma}\end{bmatrix} = \begin{bmatrix}\sigma_u^2\mathbf{I}_K & 0 \\ 0 & \sigma_\gamma^2\mathbf{I}_t\end{bmatrix}.$$

where

$$\begin{aligned} \boldsymbol{\beta} &= \left[\beta_0, \boldsymbol{\beta}_s^T\right], \\ \mathbf{u} &= [u_1, ..., u_K], \\ \boldsymbol{\gamma} &= [\gamma_1, ..., \gamma_t], \\ \mathbf{X} &= \left[1, \mathbf{s}_{ij}^T\right]_{1 \le ij \le n}, \\ \mathbf{D} &= [d_{ij}]_{1 \le ij \le n}, \end{aligned}$$

with d_{ij} an indicator taking value 1 if we observe a rainfall maxima at rain gauge j in year i and 0 otherwise, and $\mathbf{Z}$ is the matrix containing the spline basis functions, that is

$$\mathbf{Z} = [b_{tps}(\mathbf{s}_{ij}, \boldsymbol{\kappa}_k)]_{1 \le ij \le n, 1 \le k \le K} = [C(\mathbf{s}_{ij} - \boldsymbol{\kappa}_k)]_{1 \le ij \le n, 1 \le k \le K} \cdot [C(\boldsymbol{\kappa}_h - \boldsymbol{\kappa}_k)]_{1 \le h,k \le K}^{-1/2},$$

where $C(\mathbf{v}) = \|\mathbf{v}\|^2 \log\|\mathbf{v}\|$ and $\boldsymbol{\kappa}_1, ..., \boldsymbol{\kappa}_K$ are the spline knots locations.

Model Implementation

The geoadditive mixed model for extremes (4) can be naturally formulated as a hierarchical Bayesian model and estimated under the Bayesian paradigm. Following the specifications of Padoan (2008) and Crainiceanu at al. (2003), our complete hierarchical Bayesian formulation is

$$
\begin{aligned}
&\text{1st level} \quad y_i|\,(\mathbf{u},\boldsymbol{\gamma}) \overset{\text{ind}}{\sim} \text{GEV}(\,[\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}+\mathbf{D}\boldsymbol{\gamma}]_i\,,\psi,\xi),\\
&\qquad\qquad \mathbf{u}|\sigma_u^2 \sim N(0,\sigma_\gamma^2\mathbf{I}_K),\\
&\qquad\qquad \boldsymbol{\gamma}|\sigma_\gamma^2 \sim N(0,\sigma_\gamma^2\mathbf{I}_t),\\
&\text{2st level} \quad \boldsymbol{\beta} \sim N(0,10^4\mathbf{I})\\
&\qquad\qquad \xi \sim \text{Unif}(-5,5)\\
&\qquad\qquad \psi \sim \text{InvGamma}(10^{-4},10^{-4})\\
&\qquad\qquad \sigma_u^2 \sim \text{InvGamma}(10^{-4},10^{-4})\\
&\text{3st level} \quad \sigma_\gamma^2 \sim \text{InvGamma}(10^{-4},10^{-4}).
\end{aligned}
$$

where the parameters setting of the priors distributions for ξ, ψ, $\boldsymbol{\beta}$, σ_u^2, σ_γ^2, corresponds to non-informative priors.

Given the complexity of the proposed hierarchical models, we employ `OpenBUGS` Bayesian MCMC inference package to do the model fitting. We access `OpenBUGS` using the package `BRugs` (Thomas et al., 2006) in the `R` computing environment (R Development Core Team, 2011). We implement the MCMC analysis with a burn-in period of 40000 iterations and then we retain 10000 iterations, that are thinned by a factor of 5, resulting in a sample of size 2000 collected for inference. Finally, the last setting concern the thin plate spline knots that are selected setting $K = 30$ and using the *clara* space filling algorithm of Kaufman and Rousseeuw (1990), available in the `R` package `cluster` (the resulting knots location is presented in Figure 3).

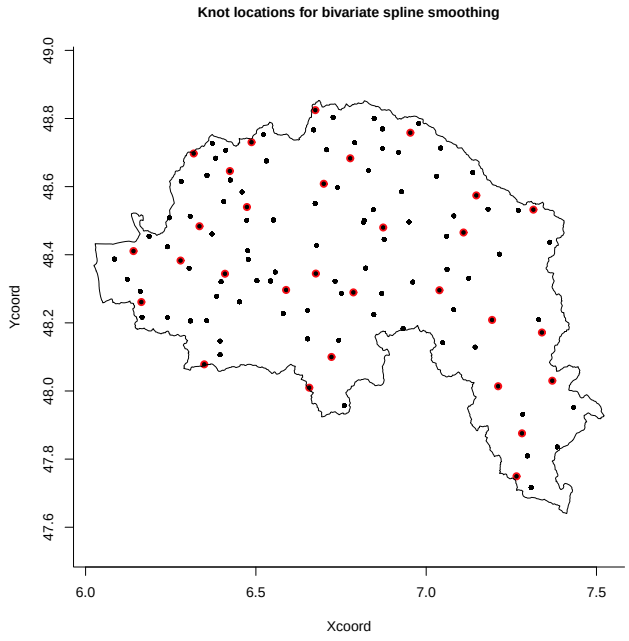

Figure 3: Knots location (in red) for the spline component. Black dots indicate the rain gauges sites.

Table 1: Estimated parameters of the GEV geoadditive mixed model for the annual maxima of daily rainfall.

Parameter*	Posterior Mean	95% Credible Interval
β_0	11.31	(8.75;13.82)
β_{s1}	-1.74	(-4.39; 1.21)
β_{s2}	1.02	(0.62;1.38)
ξ	0.11	(0.09;0.12)
ψ	15.13	(14.79;15.45)
σ_γ	7.75	(6.35;9.51)
σ_u	27.24	(20.66;35.86)

*Intercept and coordinates coefficients are required by model structure.

The estimated parameters are presented in Table 1, that provides their posterior means along with the corresponding 95% credible intervals. The posterior mean of the ξ takes value of 0.122 with 95% credible interval (0.112, 0.129), indicating the GEV distributions of annual maximum rainfalls in the Arno catchment belong to the Gumbel family and have heavy upper tails.

The resulting spatial smoothing component and time specific component of $\mu(s_{ij})$ are presented in Figures 4 and 5. Observing the map, it is evident the presence of a spatial trend in the rainfall extreme dynamic, even after controlling for the year effect. The spline seems to capture well the spatial dependence as it produce the same same patter that is shown in Figure 2. The time influence is pointed out by the estimated year specific random effects, that present a strong variability through years.

Finally, in order to asses the usefulness of our model we plot the predicted values of $\mathrm{E}(\mathbf{y}|\mathbf{u},\boldsymbol{\gamma})$ against the observed values. The results, presented in Figure 6, show a good prediction performance.

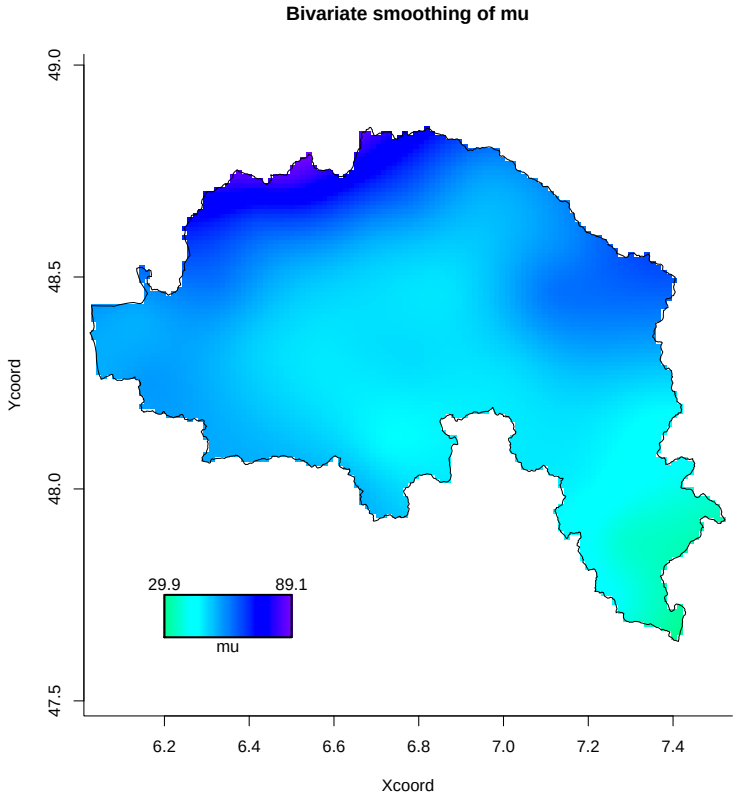

Figure 4: Estimated spatial component of $\mu(s_{ij})$.

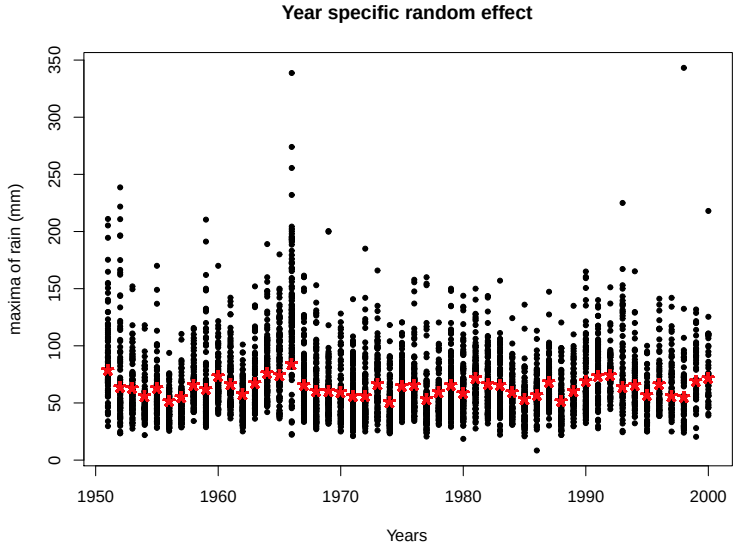

Figure 5: Estimated year specific random effects of $\mu(s_{ij})$ (in red). Black dots indicate the observed values.

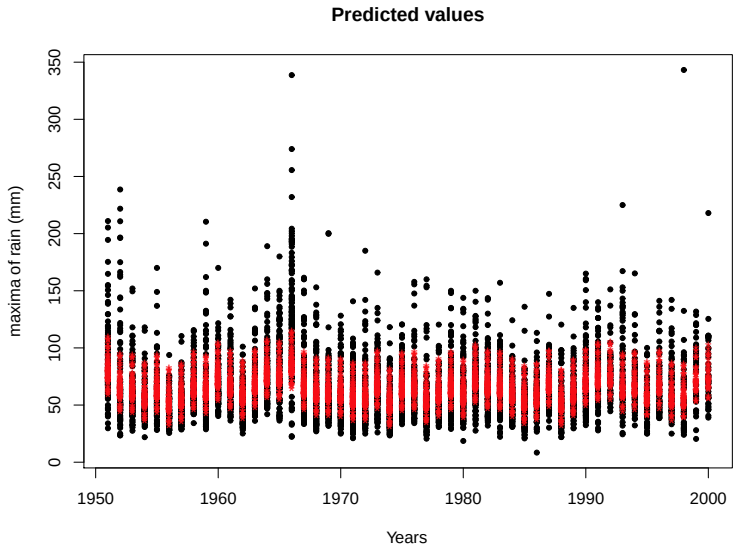

Figure 6: Predicted values of $\mathrm{E}(\mathbf{y}|\mathbf{u},\boldsymbol{\gamma})$ (in red). Black dots indicate the observed values.

Conclusions

We have implemented a geoadditive modeling approach for explaining a collection of spatially referenced time series of extreme values. We assume that the observations follow generalized extreme value distributions whose locations are spatially dependent.

The results show that this model allows us to capture both the spatial and the temporal dynamics of the rainfall extreme dynamic.

Under this approach we expect to reach a better understand of the occurrence of extreme events which are of practical interest in climate change studies particularly when related to intense rainfalls and floods, and hydraulic risk management.

REFERENCES

Bates B.C., Kundzewicz Z.W., Wu S. and Palutikof J.P. (Eds.) (2008). Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change. IPCC Secretariat. Geneva, 210 pp.

Burlando, P. and Rosso, R. (2002). Effects of transient climate change on basin hydrology. Precipitation scenarios for the Arno River, central Italy, Hydrological Process., 16, 1151-1175.

Caporali E., Rinaldi, M. and Casagli, N. (2005). The Arno River Floods. Giornale di Geologia Applicata, Vol. 1, 177:192. DOI: 10.1474/GGA.2005-01.0-18.0018. ISSN: 1825-6635.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. Applied Statistics, 54, 207222.

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

Easterling D.R., Meehl G.A., Permesan C., Changnon S.A., Karl T.R. and Mearns L.O. (2000). Climate extremes: observations, modelling and impacts. Science 289: 20682074.

EEA, European Environment Agency (2007). Climate change and water adaptation issues. EEA Technical report N 2/2007, February.

Fatichi S. and Caporali E. (2009). A comprehensive analysis of changes in precipitation regime in Tuscany. International Journal of Climatology, 29(13), 1883-1893.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. Applied Statistics, 52, 118.

Katz R.W. and Brown B.G. (1992). Extreme events in a changing climate: variability is more important than averages. Climate Change 21: 289302.

Katz R.W., Brush B.G. and Parlange M. (2005). Statistics of extremes: Modeling ecological disturbances. Ecology 86: 11241134.

Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Meehl G.A., Karl T., Easterling D., Changnon S., Pielke R. Jr., Changnon D., Evans J., Groisman P., Knutson T.R., Kunkel K.E., Mearns L.O., Parmesan C., Pulwarty R., Root T., Sylves R.T., Whetton P. and Zwiers F. (2000). An introduction to trends in extreme weather and climate events: observations, socioeconomic impacts, terrestrial ecological impacts, and model projections. Bulletin of the American Meteorological Society 81(3): 413416.

Pachauri, R.K. and Reisinger, A. (Eds.) (2007). Climate change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC Secretariat: Geneva, 104 pp

Padoan, S.A. (2008). Computational methods for complex problems in extreme value theory. Ph.D. thesis, Ph.D. in Statistical Science, Department of Statistical Science, University of Padova.

Padoan, S.A. and Wand, M.P. (2008). Mixed model-based additive models for sample extremes. Statistics and Probability Letters, 78, 2850- 2858.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Sang, H. and Gelfand, A.E. (2009). Hierarchical modeling for extreme values observed over space and time. Environmental and Ecological Statistics, 16, 407-426.

Thomas, A., O'Hara, B., Ligges, U. and Sturtz, S. (2006). Making BUGS Open. R News 6 (1), 12-17.

Wagner D. (1996). Scenarios of extreme temperature events. Climatic Change 33: 385407.

WMO (1983). Document no. 100, Guide to climatological practices. Secretariat of the World Meteorological Organization, Geneva, Switzerland.