A spatio-temporal model for extreme precipitation simulated by a climate model

Jonathan Jalbert
Postdoctoral fellow at McGill University, Montréal
Anne-Catherine Favre, Claude Bélisle and Jean-François Angers

Workshop:
Uncertainty and Causality Assessment in Modeling Extreme and Rare Events

April 27th, 2016
NCAR, Boulder, CO.
Introduction

The 2011 Richelieu River flood.
- Damages estimated at USD 100 million.

The 2013 Alberta Floods.
- Damages estimated at USD 5 billion.
Questions naturally come up with these and other major floods:
- Are climate changes responsible for these particular events?
- Should flooding defences be modified?

With climate, *in situ* experimentations are impossible. Climate models are therefore the only tools for providing partial answers to the precedent questions.

The goal of the talk is to present a spatio-temporal statistical model especially suited for extreme precipitation simulated by a climate model:
- non-stationarity in transient time series;
- large spatial simulation domain;
- spatial dependence among grid points.
Questions naturally come up with these and other major floods:

- Are climate changes responsible for these particular events?
- Should flooding defences be modified?

With climate, *in situ* experimentations are impossible. Climate models are therefore the only tools for providing partial answers to the precedent questions.

The goal of the talk is to present a spatio-temporal statistical model especially suited for extreme precipitation simulated by a climate model:

- non-stationarity in transient time series;
- large spatial simulation domain;
- spatial dependence among grid points.
The dataset consists in the daily precipitation outputs from a run of the Canadian Regional Climate Model (CRCM).

- 12,570 land grid points;
- Daily precipitation series for the period \([1961, 2100]\) at every grid point.

Let \(Y_{ik\ell}\) be the precipitation depth (mm) of day \(\ell\) of year \(k\) at grid point \(i\), where

- \(1 \leq i \leq 12,570\);
- \(1 \leq k \leq 140\);
- \(1 \leq \ell \leq 365\).
Let $M_{ik}$ be the annual maximum of year $k$ at grid point $i$:

$$M_{ik} = \max_{1 \leq \ell \leq 365} Y_{ik\ell}.$$ 

For 66% of the grid points, the series of maxima

$$(M_{ik} : 1 \leq k \leq 140)$$

exhibits temporal non-stationarity (the grid points in red in the following figure).
Non-Stationarity

The *standard* approach consists in modeling the maxima series as follows:

\[ M_{ik} \overset{\mathcal{L}}{\approx} \text{GEV} \left( \mu_{ik}, \sigma_{ik}, \xi_i \right). \]  

(1)

The non-stationarity is estimated only with the maxima series.

Suppose that there exist sequences of constants \( \nu_{ik} \) and \( \tau_{ik} \) such that

\[
\left( M'_{ik} = \frac{M_{ik} - \nu_{ik}}{\tau_{ik}} : 1 \leq k \leq 140 \right);
\]  

(2)

can be assumed identically distributed along index \( k \) for each grid point \( i \).

Therefore,

\[ M'_{ik} \overset{\mathcal{L}}{\approx} \text{GEV} \left( \mu_i, \sigma_i, \xi_i \right). \]  

(3)
Non-Stationarity

The *standard* approach consists in modeling the maxima series as follows:

\[ M_{ik} \overset{\mathcal{L}}{=} \mathcal{GEV} (\mu_{ik}, \sigma_{ik}, \xi_i). \] (1)

The non-stationarity is estimated only with the maxima series.

Suppose that there exist sequences of constants \( \nu_{ik} \) and \( \tau_{ik} \) such that

\[
\left( M'_{ik} = \frac{M_{ik} - \nu_{ik}}{\tau_{ik}} : 1 \leq k \leq 140 \right); \tag{2}
\]

can be assumed identically distributed along index \( k \) for each grid point \( i \). Therefore,

\[ M'_{ik} \overset{\mathcal{L}}{=} \mathcal{GEV} (\mu_i, \sigma_i, \xi_i). \] (3)
Non-Stationarity

Let $Z_{ik}$ be the vector of precipitation exceedences over the threshold $u_{ik}$ of year $k$ at grid point $i$:

$$Z_{ik} = (Y_{ikl} : Y_{ikl} > u_{ik}, 1 \leq l \leq 140);$$

and let

$$\nu_{ik} = \mathbb{E}(Z_{ik}) \text{ and } \tau_{ik}^2 = \mathbb{Var}(Z_{ik}).$$

The threshold has to be chosen in order that the transformation:

$$M'_{ik} = \frac{M_{ik} - \nu_{ik}}{\tau_{ik}};$$

removes the trend in the maxima series $M'_i$.

The threshold is needed to isolate the trend in the tail from the one in the bulk of the distribution. Its definition does not rely on asymptotic convergence requirements as in the Peaks-Over-Threshold model.
Non-Stationarity

Let $Z_{ik}$ be the vector of precipitation exceedences over the threshold $u_{ik}$ of year $k$ at grid point $i$:

$$Z_{ik} = (Y_{ikl} : Y_{ikl} > u_{ik}, 1 \leq l \leq 140);$$  \hspace{1cm} (4)

and let

$$
\nu_{ik} = \mathbb{E}(Z_{ik}) \text{ and } \tau_{ik}^2 = \text{Var}(Z_{ik}). \hspace{1cm} (5)
$$

The threshold has to be chosen in order that the transformation:

$$
M'_{ik} = \frac{M_{ik} - \nu_{ik}}{\tau_{ik}}; \hspace{1cm} (6)
$$

removes the trend in the maxima series $M'_i$.

The threshold is needed to isolate the trend in the tail from the one in the bulk of the distribution. Its definition does not rely on asymptotic convergence requirements as in the Peaks-Over-Threshold model.
We chose the 80th annual empirical quantiles at each grid point as the thresholds.

Then, the non-stationarity hypothesis of the preprocessed maxima series was rejected for only 1.5% of the grid points (the grid points in red in the following figure).

Pointwise Mann-Kendall stationarity test ($\alpha = 5\%$) for the preprocessed maxima series
Benefits of the proposed preprocessing approach are:

- if there exist constants $a_{ik} > 0$ and $b_{ik}$ such that

$$\frac{M_{ik} - b_{ik}}{a_{ik}} \overset{\mathcal{L}}{\to} \text{GEV}(0, 1, \xi_i);$$

then there exists constants $a'_{ik} > 0$ and $b'_{ik}$ such that

$$\frac{M'_{ik} - b'_{ik}}{a'_{ik}} \overset{\mathcal{L}}{\to} \text{GEV}(0, 1, \xi_i).$$

- The model for the untransformed maxima is tractable:

$$M_{ik} \overset{\mathcal{L}}{\approx} \text{GEV} \left( \tau_{ik} \mu_i + \nu_{ik}, \tau_{ik} \sigma_i, \xi_i \right). \quad (7)$$
Following the idea of Cooley and Sain (2010), the spatial dependence is taken into account by modeling spatial variation in the GEV parameters mainly for two reasons:

- such a latent variable approach is very flexible;
- the local properties of extremal distributions (such as return levels) are well reproduced.

However such an approach can neither model nor predict an event occurring simultaneously at several grid points.
Spatial latent model

Local estimates of the GEV location parameter could definitely benefit from neighboring site values.

Local estimates of the GEV location parameter.
Spatial latent model

Local estimates of the GEV scale parameter.

Local parameter estimations could definitely benefit from neighboring site values.
Spatial latent model

Local estimates of the GEV shape parameter.

Local parameter estimations could definitely benefit from neighboring site values.
The most widely used spatial latent model is Gaussian field. For example,

$$\mu \sim \mathcal{N}_n(\eta, \Sigma);$$

where $\mu = (\mu_1, \ldots, \mu_n)$, $\eta$ is a mean vector and $\Sigma$ is a covariance matrix.

Since the random variables $\mu$ lie on a regular lattice, a special case of Gaussian field is more appropriate: the Gaussian Markov random fields. Such a field inherits the Markov property:

$$f_{[\mu_i | \mu_{-i} = \mu_{-i}]}(\mu_i) = f_{[\mu_i | \mu_{\delta_i} = \mu_{\delta_i}]}(\mu_i);$$

where $\delta_i$ is the set of neighbors of grid point $i$. 
In Gaussian Markov random fields, the parametrization uses the precision matrix $Q$ instead of the covariance matrix $\Sigma$ because of the important simplification that comes from the conditional independence assumption:

$$\mu_i \perp \mu_j \mid \mu_{-i,-j} \iff q_{ij} = 0; \quad (10)$$

where $q_{ij}$ is the element $(i,j)$ of the precision matrix $Q$.

The set of conditional distributions $(f_{[\mu_i | \mu_{\delta_i} = \mu_{\delta_i}]} : 1 \leq i \leq n)$ must respect some constraints in order to yield a valid joint distribution:

$$f_{[\mu_i | \mu_{\delta_i} = \mu_{\delta_i}]}(\mu_i) = \mathcal{N} \left\{ \mu_i \mid \eta_i - \frac{1}{q_{ii}} \sum_{j \in \delta_i} q_{ij} (\mu_j - \eta_j), \frac{1}{q_{ii}} \right\}. \quad (11)$$
Gaussian Markov random fields

For a stationary field in space, the set of conditional distributions can be rewritten as follows:

\[
f_{\mu_i|\mu_{\delta_i}=\mu_{\delta_i}}(\mu_i) = N \left\{ \mu_i \left| \eta_i + \rho \sum_{j \in \delta_i} (\mu_j - \eta_j), \zeta^2 \right. \right\}; \tag{12}
\]

where \(0 \leq \rho \leq 1\) and \(\zeta^2 > 0\).

\[
\text{Cor} (\mu_i, \mu_j|\mu_{-i,-j} = \mu_{-i,-j}) = \rho, \text{ for } j \in \delta_i. \tag{13}
\]

It can be shown that marginal bivariate correlation coefficients between neighbors are necessarily less than 0.8\(^1\):

\[
\text{Cor} (\mu_i, \mu_j) \leq 0.8, \text{ for } j \in \delta_i. \tag{14}
\]

The spatial correlation that can be modeled is therefore limited.

\(^1\)Besag and Kooperberg (1995)
Intrinsic Gaussian Markov random fields

Following Besag and Kooperberg (1995) and Rue and Held (2005), let $n_i = \text{Card}(\delta_i)$, be the number of neighbors of grid point $i$.

$$f[\mu_i | \mu_{\delta_i} = \mu_{\delta_i}](\mu_i) = \mathcal{N} \left( \mu_i \left| \frac{1}{n_i} \sum_{j \in \delta_i} \mu_j, \frac{1}{\kappa_{\mu_i} n_i} \right. \right);$$

(15)

where $\kappa_{\mu_i} > 0$ is a precision parameter.

The set of conditional distribution gives an improper joint distribution:

$$f_{\mu}(\mu) \propto \kappa_{\mu}^{(n-1)/2} \exp \left( -\frac{\kappa_{\mu}}{2} \mu^\top W \mu \right);$$

(16)

where

$$w_{ij} = \begin{cases} n_i & \text{if } j = i; \\ -1 & \text{if } j \in \delta_i; \\ 0 & \text{otherwise}. \end{cases}$$

(17)
Intrinsic Gaussian Markov random fields

Following Besag and Kooperberg (1995) and Rue and Held (2005), let $n_i = \text{Card}(\delta_i)$, be the number of neighbors of grid point $i$.

$$f[\mu_i | \mu_{\delta_i} = \mu_{\delta_i}](\mu_i) = \mathcal{N}\left(\mu_i \left| \frac{1}{n_i} \sum_{j \in \delta_i} \mu_j, \frac{1}{\kappa_{\mu} n_i}\right)\right);$$

(15)

where $\kappa_{\mu} > 0$ is a precision parameter.

The set of conditional distribution gives an improper joint distribution:

$$f_{\mu}(\mu) \propto \kappa_{\mu}^{(n-1)/2} \exp \left(-\frac{\kappa_{\mu}}{2} \mu^\top W \mu \right);$$

(16)

where

$$w_{ij} = \begin{cases} 
  n_i & \text{if } j = i; \\
  -1 & \text{if } j \in \delta_i; \\
  0 & \text{otherwise.}
\end{cases}$$

(17)
Intrinsic Gaussian Markov random fields

Under the Bayesian paradigm, the intrinsic Gaussian Markov random field used as a prior leads to a proper posterior if $f_{\kappa \mu}(\kappa \mu)$ is proper.

\[ E(\mu_i) \text{ is undefined; } \forall \text{Var}(\mu_i) = \infty. \quad (18) \]

\[ f_{[\mu_i - \mu_j]}(\delta) = \mathcal{N}\left(\delta \mid 0, \frac{1}{\kappa \mu}\right); \text{ for } j \in \delta_i. \quad (19) \]

The model is therefore:

\[ f[M'_{ik} | (\mu_i, \sigma_i, \xi_i)](m_{ik}) \overset{\approx}{\sim} \mathcal{GEV}(m'_{ik} | \mu_i, \sigma_i, \xi_i); \]

\[ f(\mu, \sigma, \xi)(\mu, \sigma, \xi) \propto (\kappa_\mu \kappa_\sigma \kappa_\xi)^{n-1} \exp \left\{ -\frac{\kappa_\mu}{2} \mu^\top W \mu - \frac{\kappa_\sigma}{2} \sigma^\top W \sigma - \frac{\kappa_\xi}{2} \xi^\top W \xi \right\}; \]
Under the Bayesian paradigm, the intrinsic Gaussian Markov random field used as a prior leads to a proper posterior if $f_{\kappa \mu}(\kappa \mu)$ is proper.

$$ \mathbb{E} (\mu_i) \text{ is undefined; } \quad \text{Var}(\mu_i) = \infty. \quad (18) $$

$$ f_{[\mu_i - \mu_j]}(\delta) = \mathcal{N} \left( \delta \left| 0, \frac{1}{\kappa \mu} \right. \right) ; \text{ for } j \in \delta_i. \quad (19) $$

The model is therefore:

$$ f_{[M'_{ik} | (\mu_i, \sigma_i, \xi_i)]}(m_{ik}) \sim \mathcal{G} \mathcal{E} \mathcal{V} \left( m'_{ik} | \mu_i, \sigma_i, \xi_i \right); $$

$$ f(\mu, \sigma, \xi)(\mu, \sigma, \xi) \propto \left( \kappa_\mu \kappa_\sigma \kappa_\xi \right)^{n-1} \frac{1}{2} \exp \left\{ - \frac{\kappa_\mu}{2} \mu \, ^\top \, W \, \mu - \frac{\kappa_\sigma}{2} \sigma \, ^\top \, W \, \sigma - \frac{\kappa_\xi}{2} \xi \, ^\top \, W \, \xi \right\}; $$
Model fit

A spatio-temporal model for extreme precipitation simulated by a climate model

Jonathan Jalbert

2016/04/27
A chain of length 6000 was generated where the first 1000 iterations were discarded as the burn-in period. It took less than 40 minutes of computation time on a 2.53 GHz processor.

- algorithms for sparse matrix;
- parallel MCMC.
According to the *Deviance Information Criterion*, the spatial modeling of $\xi$ is relevant.

Regional estimates of the GEV shape parameter.
The grid point $i$ containing the Richelieu River.
Effective return levels

The grid point $i$ containing the Richelieu River.

- **20-year**
- **100-year**
The statistical model developed was well suited for climate model outputs, specifically for:

- transient time series;
- data that lie on a regular grid.

The model’s simplicity and intuitive interpretation along with its fast adjustment make it more appealing than Regional Frequency Analysis where uncertainty description is difficult.

Nevertheless, the model could be enhanced

- by considering space-varying textures for the latent fields;
- by integrating several climate simulations for a better description of future climate uncertainty.
Conclusion

The statistical model developed was well suited for climate model outputs, specifically for:
- transient time series;
- data that lie on a regular grid.

The model’s simplicity and intuitive interpretation along with its fast adjustment make it more appealing than Regional Frequency Analysis where uncertainty description is difficult.

Nevertheless, the model could be enhanced
- by considering space-varying textures for the latent fields;
- by integrating several climate simulations for a better description of future climate uncertainty.