Non-stationary Process Mixtures for Extreme Streamflow Forecasting in the Central US

Reetam Majumder

Joint work with Brian J. Reich

JSM 2023

NC STATE UNIVERSITY

Motivation

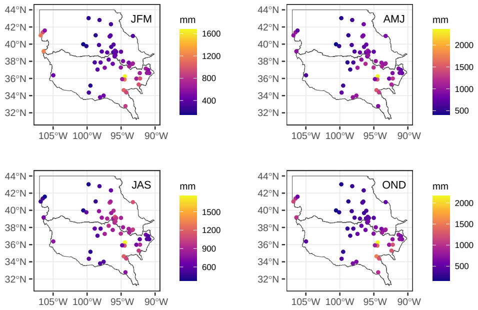


Figure 1: 0.99 quantiles of seasonal precipitation for each HCDN site. **Source:** NClimGrid.

- The Central US (CUS) is characterized by severe convective storms, and precipitation trends that could potentially influence flooding
- Extreme streamflow is a key indicator of flood risk
- The USGS Hydro Climatic Data Network (HCDN) provides streamflow data for watersheds which are minimally impacted by anthropogenic activity.

Goals

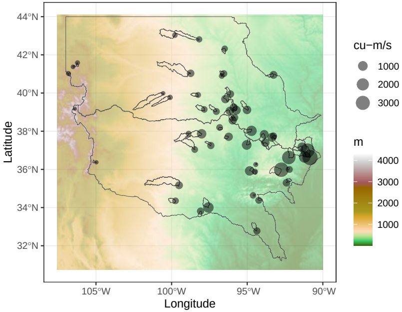


Figure 2: Sample 0.99 quantile of annual streamflow maxima from 1972–2021. **Source**: HCDN.

- HCDN data for the CUS: 55 watersheds across HUC-02 Regions 10L and 11
- **Challenge**: Expressive spatial extremes processes often have intractable likelihoods, making computation challenging.
- **Goal**: Develop a flexible and tractable spatial extremes model for climate-informed estimation of annual streamflow maxima.

Marginal distribution of streamflow maxima

Let the annual streamflow maxima for year t and site $\mathbf{s}$ following a generalized extreme value distribution:

$$Y_t(\mathbf{s}) \sim \text{GEV}\{\mu_t(\mathbf{s}), \sigma_t(\mathbf{s}), \xi_t(\mathbf{s})\},$$

whose cumulative distribution function (CDF) $F_{t,\mathbf{s}}(y) := \mathbb{P}[Y_t(\mathbf{s}) < y]$ is

$$\mathbb{P}\left[Y_t(\mathbf{s}) < y\right] = \exp\left\{-\left[1 + \xi_t(\mathbf{s})\left(\frac{y - \mu_t(\mathbf{s})}{\sigma_t(\mathbf{s})}\right)\right]^{-1/\xi_t(\mathbf{s})}\right\}. \quad (1)$$

The CDF is defined over the set $\{y : 1 + \xi_t(\mathbf{s})(y - \mu_t(\mathbf{s}))/\sigma_t(\mathbf{s}) > 0\}$

Let Z_{1t} and Z_{2t} be the annual precipitation for the two HUC-02 regions (10L and 11); define $X_{1t}(\mathbf{s})$ as:

$$X_{1t}(\mathbf{s}) = \mathbb{I}\{\mathbf{s} \in \text{Region 10L}\}Z_{1t} + \mathbb{I}\{\mathbf{s} \in \text{Region 11}\}Z_{2t}$$

Denote $X_{it}(\mathbf{s}), i = 2, \ldots, 5$ as the seasonal precipitation for site $\mathbf{s}$ for year t.

GEV parameters vary spatially and depend on precipitation:

$$\mu_t(\mathbf{s}) = \mu_0(\mathbf{s}) + \sum_{i=1}^{5} \mu_i(\mathbf{s})X_{it}(\mathbf{s}), \qquad \sigma_t(\mathbf{s}) = \sigma(\mathbf{s}), \qquad \xi_t(\mathbf{s}) = \xi(\mathbf{s}). \quad (2)$$

The joint distribution of extreme streamflow

Given streamflow data ($y_{1:n}$), marginal parameters ($\boldsymbol{\theta}_1$), and spatial process parameters ($\boldsymbol{\theta}_2$), our Bayesian hierarchical model is:

$$\text{Prior model: } \boldsymbol{\theta}_1 \sim p(\boldsymbol{\theta}_1) \perp \boldsymbol{\theta}_2 \sim p(\boldsymbol{\theta}_2),$$

$$\text{Data model: } f_y(y_1, ..., y_n | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \underbrace{f_u(u_1, ..., u_n | \boldsymbol{\theta}_2)}_{\text{spatial dependence}} \underbrace{\prod_{i=1}^{n} \left| \frac{dF(y_i | \boldsymbol{\theta}_1)}{dy_i} \right|}_{\text{marginal GEV likelihoods}} .$$

The CDF transformed variables $U_t(\mathbf{s}) := F_{t,\mathbf{s}}\left(Y_t(\mathbf{s})\right)$ share common uniform marginal distributions but are spatially correlated

This change-of-variables in the likelihood separates residual spatial dependence in $U_t(\mathbf{s})$ from spatial dependence induced by spatial variation in the GEV parameters.

The latter is modeled using Gaussian process priors on the components of $\boldsymbol{\theta}_1$

The non-stationary process mixture model (NPMM)

A spatial dependence model on $U_t(\mathbf{s})$ is obtained via the transformation $U_t(\mathbf{s}) = G_{t,\mathbf{s}}(V_t(\mathbf{s}))$:

$$V_t(\mathbf{s}) = \delta_t(\mathbf{s})R_t(\mathbf{s}) + (1 - \delta_t(\mathbf{s}))W_t(\mathbf{s}), \tag{3}$$

where $R_t(\mathbf{s})$ is a max-stable process, $W_t(\mathbf{s})$ is a Gaussian process. We call this a process mixture model

$\delta_t(\mathbf{s}) \in [0, 1]$ are weight parameters depending on regional annual precipitation,

$$\delta_t(\mathbf{s}) = \mathbb{I}\{\mathbf{s} \in \text{Region 10L}\}\delta_{1t} + \mathbb{I}\{\mathbf{s} \in \text{Region 11}\}\delta_{2t} \tag{4}$$

$$g^{-1}(\delta_{it}) = \beta_{i0} + \beta_{i1}Z_{it}, i = 1, 2. \tag{5}$$

Dependence of $\delta_t(\mathbf{s})$ on precipitation introduces non-stationarity[1]

If $\delta_t(\mathbf{s}) = \delta$, the NPMM simplifies to a stationary PMM[2]

[1] Majumder and Reich (2023), *Spat. Stat.*

[2] Majumder, Reich, and Shaby (2022), *arXiv:2208.03344*. Huser and Wadsworth (2019), *J. Am. Stat. Assoc.*

Asymptotic properties of the NPMM

Extremal spatial dependence often measured in terms of the upper-tail coefficient:

$$\chi_u(\mathbf{s}_1, \mathbf{s}_2) := \text{Prob}\{U(\mathbf{s}_1) > u | U(\mathbf{s}_2) > u\}, \tag{6}$$

where $u \in (0, 1)$ is a threshold. $U(\mathbf{s}_1)$ and $U(\mathbf{s}_2)$ are defined to be asymptotically dependent if

$$\chi(\mathbf{s}_1, \mathbf{s}_2) = \lim_{u \to 1} \chi_u(\mathbf{s}_1, \mathbf{s}_2) \tag{7}$$

is positive, and independent if $\chi(\mathbf{s}_1, \mathbf{s}_2) = 0$. For the PMM/NPMM,

$\delta < 0.5 \implies$ asymptotic independence, and

$\delta > 0.5 \implies$ asymptotic dependence

Inference involves a Vecchia approximated density regression (VADeR) approach for the intractable joint likelihood

Inference for the NPMM

1. Use a Vecchia approximation[3] to approximate the joint likelihood as:

$$f_u(u_1, \ldots, u_n | \boldsymbol{\theta}_2) = \prod_{i=1}^{n} f_i(u_i | \boldsymbol{\theta}_2, u_1, \ldots, u_{i-1}) \approx \prod_{i=1}^{n} f_i(u_i | \boldsymbol{\theta}_2, u_{(i)}), \tag{8}$$

   $u_{(i)} \subseteq \{u_1, \ldots, u_{i-1}\}$. The subset of locations $\mathbf{s}_{(i)}$ are often the nearest neighbors
2. Obtain density estimates of each term $f_i(u_i | \boldsymbol{\theta}_2, u_{(i)})$ using a semi-parametric quantile regression (SPQR) model[4]
3. Use the surrogate likelihood in a Bayesian framework to obtain posterior estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$

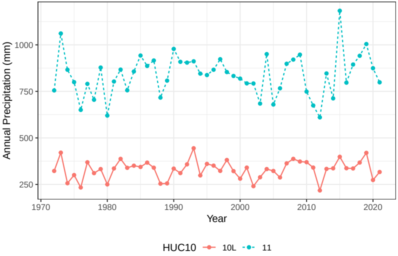
[3] Vecchia (1988), *J. R. Stat. Soc. B*. Stein, Chi, and Welty (2004), *J. R. Stat. Soc. B*.
[4] Xu and Reich (2021), *Biometrics*

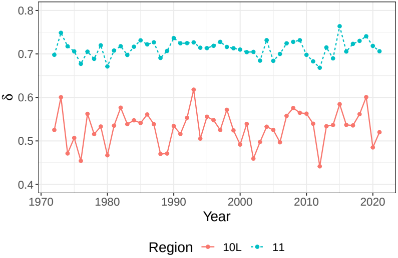
Posterior of spatial process parameters for extreme streamflow

Precipitation and streamflow for HUC-02 regions 10L and 11 from 1972–2021:

Left: Time series of annual NClimGrid precipitation (in *mm*)

Right: Posterior means of δ_{1t} and δ_{2t} corresponding to regions 10L and 11





δ_{1t} and δ_{2t} do not change with changes in basin-wide annual precipitation

δ_{1t}, δ_{2t} have posterior means of 0.53 and 0.71 for the 50 year period (asymptotically dependent)

- Precipitation is a significant predictor of streamflow maxima
- Spring (AMJ) precipitation is the most significant predictor at most locations.

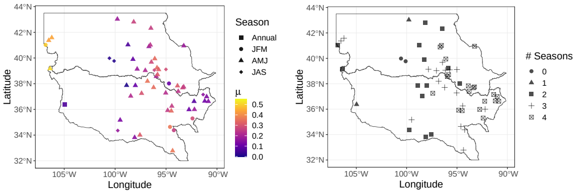
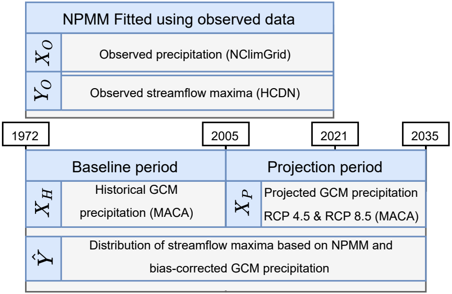


Figure 3: Estimates of $\mu(s) = \max(\mu_j(s))$ for $j = 2:5$ corresponding to the 4 seasons with shapes denoting the season with the highest slope value (left), and number of seasons (excluding annual) where $\mathbb{P}[\mu(s) > 0] > 0.90$ (right).

Scale and shape parameters estimates also show spatial variation

Posterior mean of shape parameter is positive at 54 out of 55 locations

Streamflow projections for the CUS



We use bias-corrected climate model precipitation output from CMIP5[5] as covariates in the posterior predictive distribution of streamflow maxima to get projections for 2006–2035

6 CMIP5 models (3 wet + 3 dry) considered for each representative climate pathway (RCP) scenario, viz. RCP 4.5 and RCP 8.5

[5] Taylor, Stouffer, and Meehl (2012), *B. Am. Meteorol. Soc.*

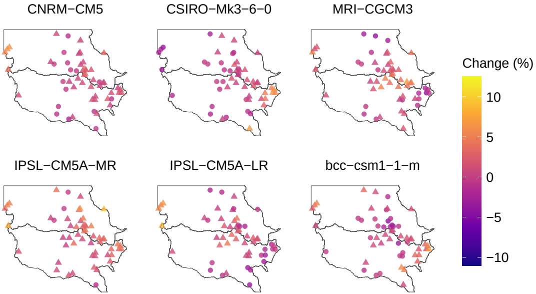


Figure 4: Percentage change in observed 0.90 quantile under RCP 4.5. Triangles denote an increase while circles denote a decrease.

We compare annual streamflow maxima for 2006–2035 against 1972–2005

Changes from -10.3% to 12.3% for the 0.90 quantile of annual streamflow maxima

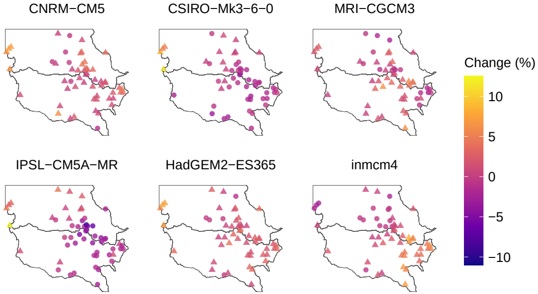


Figure 5: Percentage change in observed 0.90 quantile under RCP 8.5. Triangles denote an increase while circles denote a decrease.

All 6 models under RCP 4.5 and 4 models under RCP 8.5 estimate that more than 50% locations have increased streamflow in the projection period.

Less pronounced but similar results for the 0.99 quantile

Summary

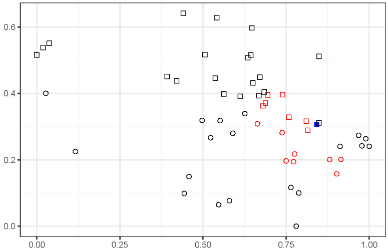
- Significance: Precipitation is estimated to be a significant predictor of extremal streamflow in the CUS and shows a strong seasonal component
- Non-stationarity: The asymptotic dependence properties of the two HUC-02 regions are estimated to be different from each other, and each show inter-annual variability
- Projections: Annual streamflow maxima is projected to increase in the near future
- Methodology: The NPMM is flexible (desirable asymptotic properties), and tractable (computational cost increases linearly in number of locations). The density estimation approach can be used for any intractable spatial process
- Brian's talk (Tuesday afternoon) will go into more details of the methodology

Acknowledgments: Southeast National Synthesis Wildfire and National Climate Adaptation Science Center (G21AC10045), NSF (DMS2152887, CBET2151651, DMS1929348).

References

- R. Majumder and B. J. Reich. A deep learning synthetic likelihood approximation of a non-stationary spatial model for extreme streamflow forecasting. *Spatial Statistics*, page 100755, 2023.
- R. Majumder, B. J. Reich, B. A. Shaby. Modeling extremal streamflow using deep learning approximations and a flexible spatial process. *ArXiv*, 2022.
- L. Zhang, M. D. Risser, E. M. Molter, M. F. Wehner, T. A. O'Brien. Accounting for the spatial structure of weather systems in detected changes in precipitation extremes. *Weather and Climate Extremes*, 38:100499, 2022.
- C. Awasthi, S. A. Archfield, K. R. Ryberg, J. E. Kiang, A. Sankarasubramanian. Projecting flood frequency curves under near-term climate change. *Water Resources Research*, 58(8):e2021WR031246, 2022.
- R. Huser, J. L. Wadsworth. Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444, 2019.
- S. G. Xu, B. J. Reich. Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics* 00:1–14, 2021.
- A. V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312, 1988.
- M. L. Stein, Z. Chi, L. J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(2):275–296, 2004.

Vecchia approximation of intractable likelihood



Problem: Evaluate the NPMM likelihood for any values of $\boldsymbol{\theta}_2$ and $(y_1, \ldots, y_n)$

Approach: Density estimation of surrogate univariate conditional likelihoods based on a Vecchia decomposition of the joint distribution $f_u(u_1, \ldots, u_n | \boldsymbol{\theta}_2)$:

$$f_u(u_1, \ldots, u_n | \boldsymbol{\theta}_2) = \prod_{i=1}^{n} f_i(u_i | \boldsymbol{\theta}_2, u_1, \ldots, u_{i-1}) \approx \prod_{i=1}^{n} f_i(u_i | \boldsymbol{\theta}_2, u_{(i)}), \tag{9}$$

$u_{(i)} \subseteq \{u_1, \ldots, u_{i-1}\}$. The subset of locations $\mathbf{s}_{(i)}$ are often the nearest neighbors.

Inference for the NPMM

Density regression is carried out for each of the $n-1$ terms separately using neural networks in a semi-parametric quantile regression (SPQR) model[6]:

$$f_i(u_i|\mathbf{x}_i, \mathcal{W}) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i, \mathcal{W}_i) B_k(u_i), \tag{10}$$

$$\pi_{ik}(\mathbf{x}_i, \mathcal{W}_i) = f_i^{NN}(\mathbf{x}_i, \mathcal{W}_i), \text{ for } i = 2, \ldots, n. \tag{11}$$

- $\mathbf{x}_i = (u_{(i)}, \boldsymbol{\theta}_2)$ are treated as covariates, with u_i as the response variable
- Each NN maximizes the log-likelihood of a univariate conditional (RHS of (10))
- NNs are trained using synthetic data (surrogate likelihood)
- Given a value of $\boldsymbol{\theta}_2$ and $u_{(i)} = F(y_{(i)})$, we can then evaluate $f_u(u_1, ..., u_n|\boldsymbol{\theta}_2)$ as a product of surrogate conditional distributions
- Can be used in an MCMC to estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

[6] Xu and Reich (2021), *Biometrics*.