Multivariate nonstationary hydrologic frequency analysis: A Bayesian hierarchical approach

Cameron Bracken, Balaji Rajagopalan
Department of Civil, Environmental and Architectural Engineering
University of Colorado at Boulder

Kathleen Holman
Bureau of Reclamation

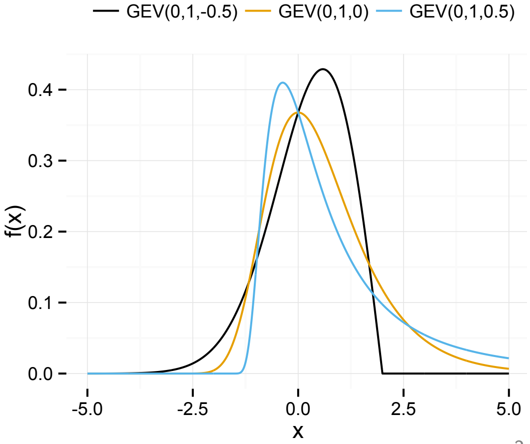
February 2016

What is hydrologic frequency analysis?

Process of estimating recurrence probabilities of rare hydrologic events (floods, heavy rainfall, etc.).

General procedure:

1. Generate extreme data. For example take the maximum daily flow value from each year from a daily flow dataset.
2. Fit a probability distribution. For example generalized extreme value.
3. Compute return levels (quantiles). A 100-year return level will be the (1-1/100)th quantile.



Lay of the land

- Bayesian hierarchical modeling of precipitation and streamflow extremes
  - Active area of research in the last 10-15 years
  - Alternative to regional frequency analysis
  - End goal is to estimate distributions of return levels
- Hydrologic frequency models come in many flavors
  - Single site and spatial
  - Stationary and nonstationary
- Typically these analyses are conducted independently

Lay of the land

- Bayesian hierarchical modeling of precipitation and streamflow extremes
  - Active area of research in the last 10-15 years
  - Alternative to regional frequency analysis
  - End goal is to estimate distributions of return levels
- Hydrologic frequency models come in many flavors
  - Single site and spatial
  - Stationary and nonstationary
- Typically these analyses are conducted independently
- **How should a multivariate frequency analysis be conducted?**

Lay of the land

- Bayesian hierarchical modeling of precipitation and streamflow extremes
  - Active area of research in the last 10-15 years
  - Alternative to regional frequency analysis
  - End goal is to estimate distributions of return levels
- Hydrologic frequency models come in many flavors
  - Single site and spatial
  - Stationary and nonstationary
- Typically these analyses are conducted independently
- **How should a multivariate frequency analysis be conducted?**
- **What multivariate frequency models are appropriate?**

Lay of the land

- Bayesian hierarchical modeling of precipitation and streamflow extremes
  - Active area of research in the last 10-15 years
  - Alternative to regional frequency analysis
  - End goal is to estimate distributions of return levels
- Hydrologic frequency models come in many flavors
  - Single site and spatial
  - Stationary and nonstationary
- Typically these analyses are conducted independently
- **How should a multivariate frequency analysis be conducted?**
- **What multivariate frequency models are appropriate?**
- **What is gained by a multivariate analysis?**

Statistics of Extremes

Given daily data, if we select the maximum value in each year, those data follow a generalized extreme value (GEV) distribution:

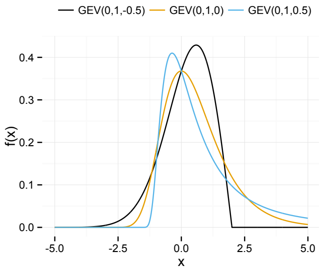
$$\mathrm{GEV}(x;\mu,\sigma,\xi) = \frac{1}{\sigma} b^{(-1/\xi)-1} \exp\left\{-b^{-1/\xi}\right\}$$

$b = 1 + \xi\left(\frac{x-\mu}{\sigma}\right)$, μ: Location, σ: Scale, ξ: Shape.

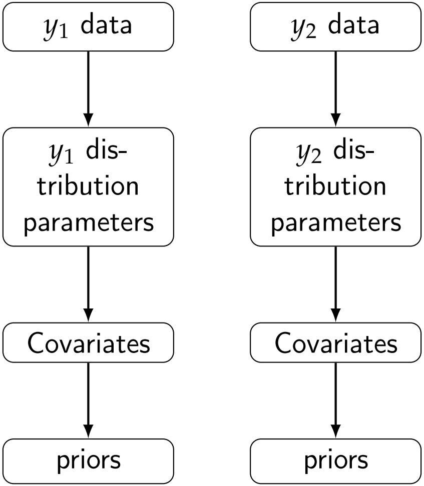
Return Level (quantile function):

$$z_r = \mu + \frac{\sigma}{\xi}\left[(-\log(1-1/r))^{-\xi} - 1\right]$$

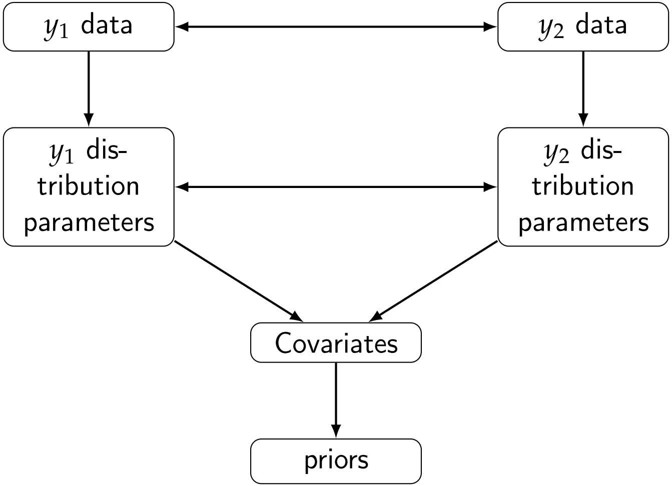
Where r is the return period in years (100 years for example).



Typical model framework



Model framework



General Multivariate Model Structure

Let $y_1, \ldots, y_n$ be n block maxima variables we wish to conduct frequency analysis on.

$$
\begin{aligned}
(y_1(t), \ldots, y_n(t)) &\sim C_g(\Sigma; \{\boldsymbol{\mu}(t), \boldsymbol{\sigma}(t), \boldsymbol{\xi}(t)\}) && && (1)\\
y_i(t) &\sim GEV(\mu_i(t), \sigma_i(t), \xi_i(t)), && i = 1 \ldots n && (2)\\
\mu_i(t) &= g(\mathbf{x}_i(t)^T, \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t), \boldsymbol{\xi}(t)), && i = 1 \ldots n && (3)\\
\sigma_i(t) &= g(\mathbf{x}_i(t)^T, \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t), \boldsymbol{\xi}(t)), && i = 1 \ldots n && (4)\\
\xi_i(t) &= g(\mathbf{x}_i(t)^T, \boldsymbol{\mu}(t), \boldsymbol{\sigma}(t), \boldsymbol{\xi}(t)), && i = 1 \ldots n && (5)
\end{aligned}
$$

where C_g is a gaussian elliptical copula joint distribution and $g(\cdot)$ is a (possibly nonlinear) function of covariates and parameters of other variables.

$$\boldsymbol{\mu}(t) = [\mu_i(t)]_{i=1}^n, \;\; \boldsymbol{\sigma}(t) = [\sigma_i(t)]_{i=1}^n, \;\; \boldsymbol{\xi}(t) = [\xi_i(t)]_{i=1}^n$$

Copula Dependence

The copula dependence matrix, Σ is a symmetric positive definite matrix capturing the strength of dependence between each pairwise variable. The i,jth element of Σ measures the dependence between variables i and j and can take values between -1 and 1. By definition the dependence between a variable and itself is unity so the diagonal elements of Σ are 1's

$$\Sigma = \begin{bmatrix} 1 & \nu_{12} & \cdots & \nu_{1,n-1} & \nu_{1n} \\ \nu_{12} & 1 & & & \nu_{2n} \\ \nu_{13} & & \ddots & & \vdots \\ \vdots & & & 1 & \nu_{n-1,n} \\ \nu_{1n} & \nu_{2n} & \cdots & \nu_{n-1,n} & 1 \end{bmatrix} \tag{6}$$

Note that since Σ is symmetric, there are $n(n-1)/2$ dependence parameters to fit (values in the lower or upper triangle of Σ).

Application 1 - Streamflow and precipitation

- 32 years of winter (DJF) 3-day flow maxima (Neuman et. al 2015):

$$z(t) \sim GEV(\mu_z(t), \sigma_z(t), \xi_z(t))$$

Application 1 - Streamflow and precipitation

- 32 years of winter (DJF) 3-day flow maxima (Neuman et. al 2015):

$$z(t) \sim GEV(\mu_z(t), \sigma_z(t), \xi_z(t))$$

- 32 years of winter (DJF) 3-day precipitation maxima (GHCNd):

$$y(s_i, t) \sim GEV(\mu_y(t), \sigma_y(t), \xi_y(t)),\ i = 1, ..., n$$

Application 1 - Streamflow and precipitation

- 32 years of winter (DJF) 3-day flow maxima (Neuman et. al 2015):

$$z(t) \sim GEV(\mu_z(t), \sigma_z(t), \xi_z(t))$$

- 32 years of winter (DJF) 3-day precipitation maxima (GHCNd):

$$y(s_i, t) \sim GEV(\mu_y(t), \sigma_y(t), \xi_y(t)),\ i = 1, ..., n$$

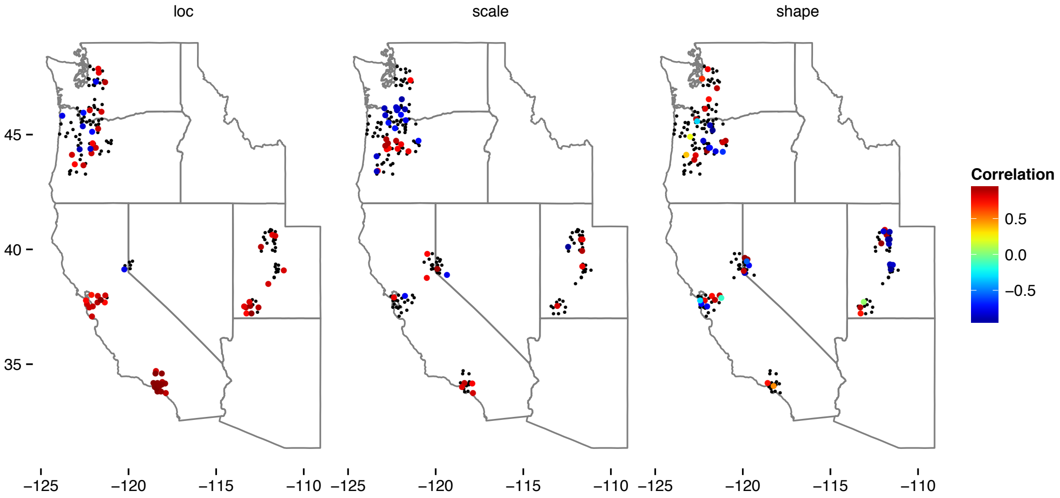
- Covariates:

$$x(t) = (\text{seasonal total precip, enso, pdo, amo})$$

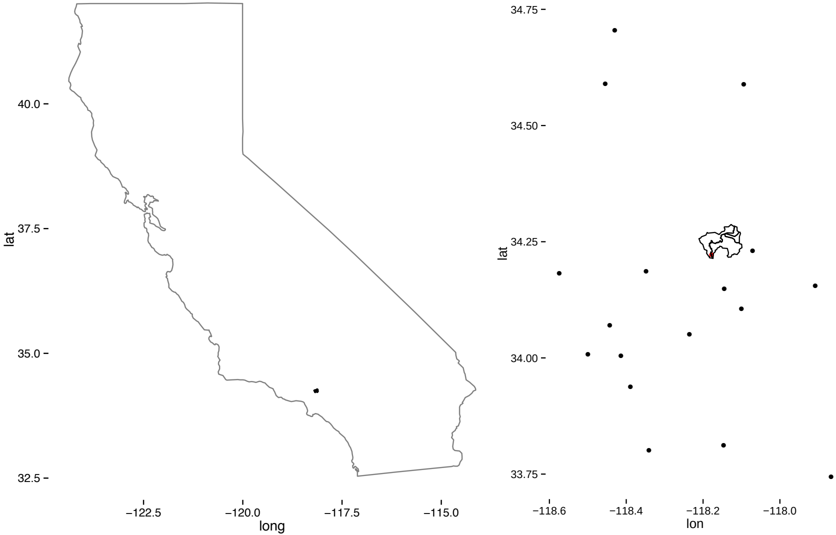
Exploratory Analysis - Are extremes linearly linked at the parameter level?

1. Fit nonstationary GEV models to flow gage and surrounding precip gages using maximum likelihood
2. Correlate the nonstationary parameter estimates
3. High correlation implies GEV parameters are related linearly

Exploratory Analysis - Are extremes linearly linked at the parameter level?



Study area



40.0
37.5
35.0
32.5
lat
−122.5
−120.0
−117.5
−115.0
long
34.75
34.50
34.25
34.00
33.75
lat
−118.6
−118.4
−118.2
−118.0
lon

Model Structure

$$(y(s_1,t),\ldots,y(s_n,t),z(t)) \sim C_g(\Sigma,\{\boldsymbol{\mu}(t),\boldsymbol{\sigma}(t),\boldsymbol{\xi}(t)\})$$

Regional nonstationary precip model:

$$y(s_i,t) \sim GEV(\mu_y(t),\sigma_y(t),\xi_y)$$

Model Structure

$$(y(s_1,t),\ldots,y(s_n,t),z(t)) \sim C_g(\Sigma,\{\boldsymbol{\mu}(t),\boldsymbol{\sigma}(t),\boldsymbol{\xi}(t)\})$$

Regional nonstationary precip model:

$$y(s_i,t) \sim GEV(\mu_y(t),\sigma_y(t),\xi_y)$$

$$\boxed{\mu_y(t) = \mathrm{x}^T(t)\boldsymbol{\beta}_\mu}$$

$$\boxed{\sigma_y(t) = \mathrm{x}^T(t)\boldsymbol{\beta}_\sigma}$$

$$\xi_y(t) = \xi_y$$

Model Structure

$$(y(s_1,t),\ldots,y(s_n,t),z(t)) \sim C_g(\Sigma,\{\boldsymbol{\mu}(t),\boldsymbol{\sigma}(t),\boldsymbol{\xi}(t)\})$$

Regional nonstationary precip model:

$$y(s_i,t) \sim GEV(\mu_y(t),\sigma_y(t),\xi_y)$$

$$\boxed{\mu_y(t) = \mathrm{x}^T(t)\boldsymbol{\beta}_\mu}$$

$$\boxed{\sigma_y(t) = \mathrm{x}^T(t)\boldsymbol{\beta}_\sigma}$$

$$\xi_y(t) = \xi_y$$

Nonstationary flow model:

$$z(t) \sim GEV(\mu_z(t),\sigma_z(t),\xi_z)$$

Model Structure

$$(y(s_1,t),\ldots,y(s_n,t),z(t)) \sim C_g(\Sigma,\{\boldsymbol{\mu}(t),\boldsymbol{\sigma}(t),\boldsymbol{\xi}(t)\})$$

Regional nonstationary precip model:

$$y(s_i,t) \sim GEV(\mu_y(t),\sigma_y(t),\xi_y)$$

$$\boxed{\mu_y(t) = \mathrm{x}^T(t)\boldsymbol{\beta}_\mu}$$

$$\boxed{\sigma_y(t) = \mathrm{x}^T(t)\boldsymbol{\beta}_\sigma}$$

$$\xi_y(t) = \xi_y$$

Nonstationary flow model:

$$z(t) \sim GEV(\mu_z(t),\sigma_z(t),\xi_z)$$

$$\boxed{\mu_z(t) = a + \mu_y(t)b}$$

$$\boxed{\sigma_z(t) = c + \sigma_y(t)d}$$

$$\xi_z(t) = \xi_z$$

Model Structure

$$(y(s_1,t),\ldots,y(s_n,t),z(t)) \sim C_g(\Sigma,\{\boldsymbol{\mu}(t),\boldsymbol{\sigma}(t),\boldsymbol{\xi}(t)\})$$

Regional nonstationary precip model:

$$y(s_i,t) \sim GEV(\mu_y(t),\sigma_y(t),\xi_y)$$

$$\boxed{\mu_y(t) = \mathrm{x}^T(t)\boldsymbol{\beta}_\mu}$$

$$\boxed{\sigma_y(t) = \mathrm{x}^T(t)\boldsymbol{\beta}_\sigma}$$

$$\xi_y(t) = \xi_y$$

Nonstationary flow model:

$$z(t) \sim GEV(\mu_z(t),\sigma_z(t),\xi_z)$$

$$\boxed{\mu_z(t) = a + \mu_y(t)b}$$

$$\boxed{\sigma_z(t) = c + \sigma_y(t)d}$$

$$\xi_z(t) = \xi_z$$

a,b,c,d are latent regression coefficients.

From nonstationary GEV parameter estimates we can compute nonstationary return levels.

Copula dependence

The copula dependence matrix Σ is a positive definite symetric matrix with diagonal elements equal to 1 and all other elements are between -1 and 1.

$$\Sigma = \begin{bmatrix} D & \mathbf{v} \\ \mathbf{v} & 1 \end{bmatrix}$$

$$\mathbf{v} = [v_{z1}]_{i=1}^{n}$$

v_{z1} is the correlation between the flow gage and precip station i.

$$D = \exp(||\mathbf{x}_i - \mathbf{x}_j||/a)$$

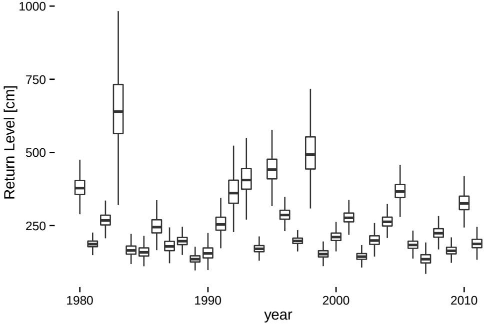
a is the precipitation range parameter.

Model fit and priors

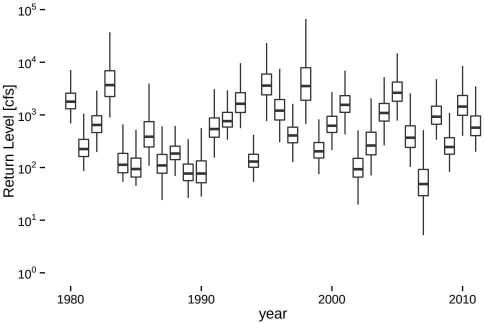
- Fit using a univariate slice sampler within Gibbs
- Uninformative uniform priors, except for $\xi \sim N(0, 0.3)$.
- 100,000 samples, 20,000 warmup iterations, 3 chains, thinned by 20, resulting in 12,000 posterior samples.
- $\hat{R} < 1.1$ for all parameters

Results: 100 year return levels

Precip:

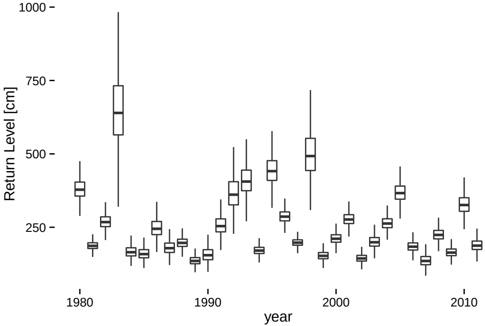


Flow:

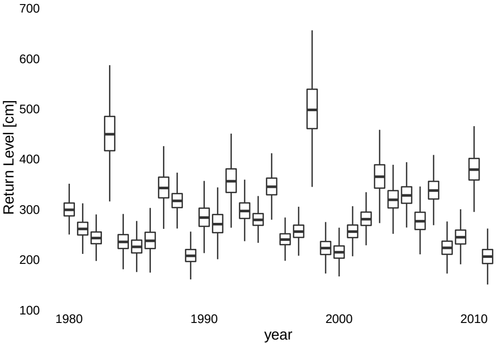


Results: Precipitation return levels (100 year)

Coupled Precip:

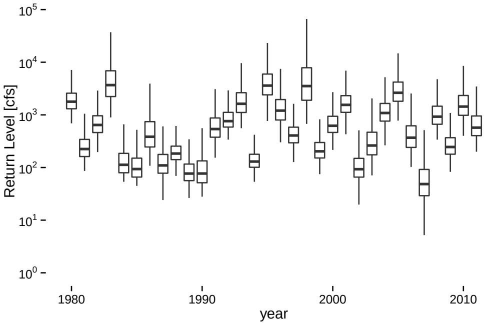


Uncoupled precip:

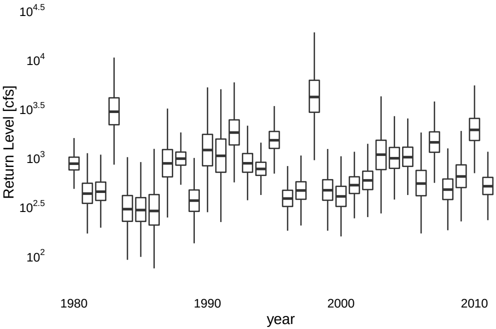


Results: Flow return levels (100 year)

Coupled flow:



Uncoupled flow:



Application 2 - Reservoir frequency analysis

Taylor Park Reservoir, Colorado, USA.

- 35 years of annual 1-day flow maxima:

$$z(t) \sim GEV(\mu_z(t), \sigma_z, \xi_z)$$

Application 2 - Reservoir frequency analysis

Taylor Park Reservoir, Colorado, USA.

- 35 years of annual 1-day flow maxima:

$$z(t) \sim GEV(\mu_z(t), \sigma_z, \xi_z)$$

- 35 years of annual 1-day peak SWE (GHCNd):

$$y(t) \sim GEV(\mu_y(t), \sigma_y, \xi_y),\ i = 1, ..., n$$

Application 2 - Reservoir frequency analysis

Taylor Park Reservoir, Colorado, USA.

- 35 years of annual 1-day flow maxima:

$$z(t) \sim GEV(\mu_z(t), \sigma_z, \xi_z)$$

- 35 years of annual 1-day peak SWE (GHCNd):

$$y(t) \sim GEV(\mu_y(t), \sigma_y, \xi_y),\ i = 1, ..., n$$

Application 2 - Reservoir frequency analysis

Taylor Park Reservoir, Colorado, USA.

- 35 years of annual 1-day flow maxima:

$$z(t) \sim GEV(\mu_z(t), \sigma_z, \xi_z)$$

- 35 years of annual 1-day peak SWE (GHCNd):

$$y(t) \sim GEV(\mu_y(t), \sigma_y, \xi_y),\ i = 1, ..., n$$

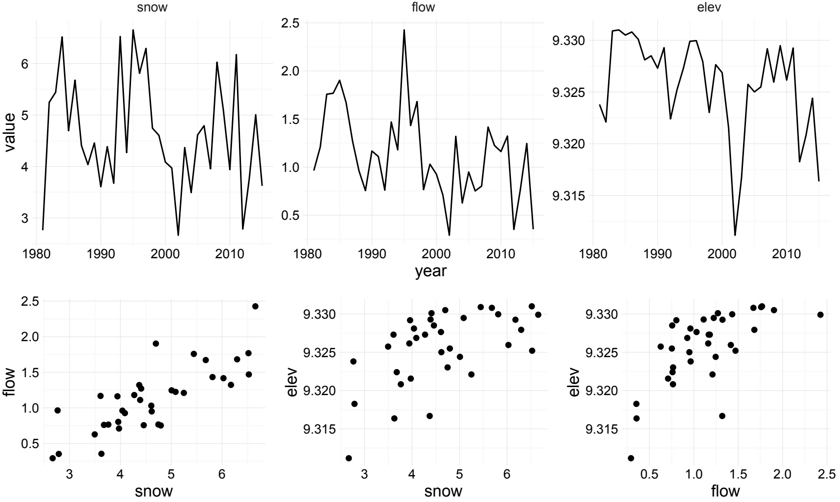
- 35 years of annual 1-day peak reservoir elevation:

$$h(t) \sim GEV(\mu_h(t), \sigma_h, \xi_h),\ i = 1, ..., n$$

- Covariates:

$$x(t) = (\text{linear trend, enso, pdo, amo})$$

Application 2 - Reservoir frequency analysis



Application 2 - Model structure

$$
\begin{aligned}
(y(t), z(t), h(t)) &\sim C_g(\Sigma; \{\mu_y(t), \sigma_y, \xi_y, \mu_z(t), \sigma_z, \xi_z, \mu_h(t), \sigma_h, \xi_h\}) && (7)\\
y(t) &\sim GEV(\mu_y(t), \sigma_y, \xi_y) && (8)\\
z(t) &\sim GEV(\mu_z(t), \sigma_z, \xi_z) && (9)\\
h(t) &\sim GEV(\mu_h(t), \sigma_h, \xi_h) && (10)\\
\mu_y(t) &= \mu_{y0} + x(t)^T \beta_y && (11)\\
\mu_z(t) &= \mu_{z0} + x(t)^T \beta_z && (12)\\
\mu_h(t) &= a - \exp(-b\mu_z(t)) && (13)
\end{aligned}
$$

where $x(t)^T$ is a vector of climate covariates.

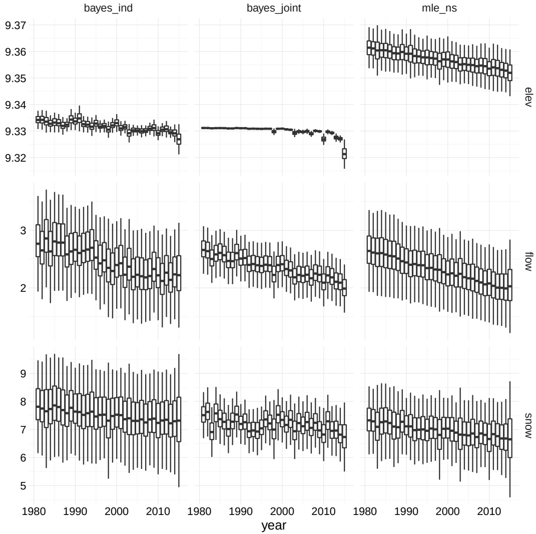
Application 2 - Copula dependence

The copula dependence matrix is

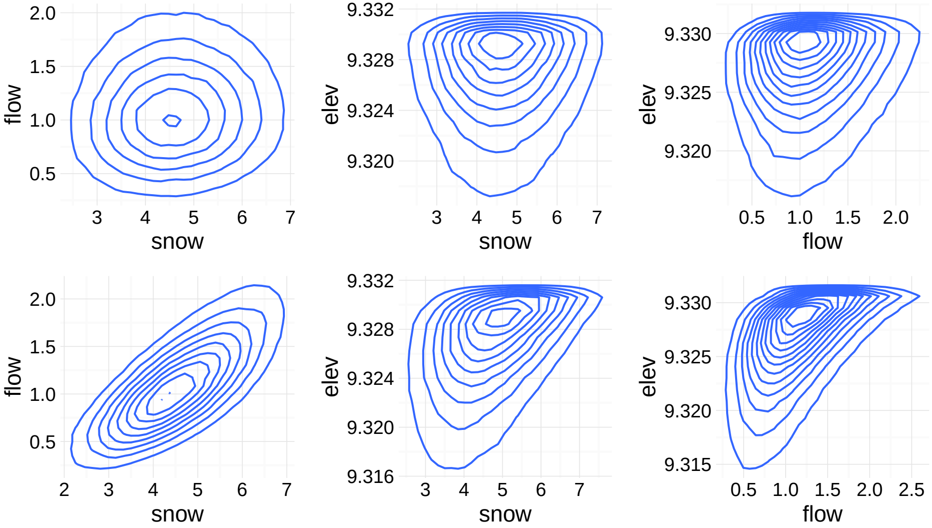
$$\Sigma = \begin{bmatrix} 1 & \nu_{yz} & \nu_{yh} \\ \nu_{yz} & 1 & \nu_{zh} \\ \nu_{yh} & \nu_{zh} & 1 \end{bmatrix} \tag{14}$$

where ν_{ij} represents the dependence (correlation) between variable i and j.

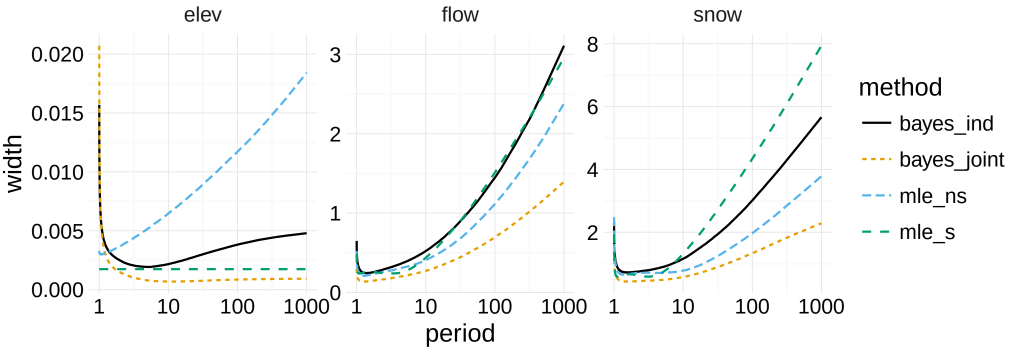
Results - Nonstationary return levels



Results - Joint distributions



Results - Uncertainty



Conclusions

Pros:

- Multivariate frequency analysis allows multiple variable to lend strength across space and time
- May decrease uncertainty
- Multivariate simulation
- Nonstationary risk estimation
- Potential for seasonal forecasting and future projections of risk

Cons:

- May increase uncertainty
- Data availability
- Computation time
- Need to tailor the model structure to each analysis

Thanks!