A Regional Bayesian Hierarchical Model for Flood Frequency Analysis over the Willamette River Basin

Hongxiang Yan (yan2@pdx.edu) and Hamid Moradkhani (hamidm@cecs.pdx.edu)
Department of Civil and Environmental Engineering, Portland State University

Abstract

In this study, we propose a regional Bayesian hierarchical model for flood frequency analysis. Besides the three classical layers (data, process, and prior) of the Bayesian hierarchical model, we add a new layer "L-moments layer" in the flood frequency analysis. The L-moments layer uses the L-moments theory to select the best-fit probability distribution based on the available data. This new layer can overcome the subjective selection of the distribution from the extreme value theory and let the data decide their own distribution. By adding this new layer, we can combine the merits of the regional flood frequency method and the Bayesian method together. The performance of the Bayesian model is assessed by a case study over the Willamette River Basin in the PNW,US. The uncertainty of different quantiles can be quantified from the posterior distributions using Markov Chain Monte Carlo algorithm. Temporal changes for 100-year flood quantiles are also examined using a 20- and 30-year moving window method. The calculated shifts in flood risk can facilitate future water resource management.

Bayesian Hierarchical Model Structure

Any inference of Bayesian hierarchical model comes from the posterior distribution, which is given by the Bayes Law:

$$p(\boldsymbol{\theta}\,|\,\mathbf{z}) \propto p(\mathbf{z}\,|\,\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is the vector of the distribution parameters, $\mathbf{z}$ is the vector of observations, $p(\mathbf{z}\,|\,\boldsymbol{\theta})$ is the likelihood function, and $p(\boldsymbol{\theta})$ is the prior distribution. For hierarchical model, usually more than one layer is considered, so the equation becomes:

$$p(\boldsymbol{\theta}\,|\,\mathbf{z}) \propto p_1(\mathbf{z}\,|\,\boldsymbol{\theta}_1) \times p_2(\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2) \times p_3(\boldsymbol{\theta}_2)$$

where $p_1(\mathbf{z}\,|\,\boldsymbol{\theta}_1)$ represents the data layer, $p_2(\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2)$ is the process layer, and $p_3(\boldsymbol{\theta}_2)$ represents the prior layer. Instead of the subjective selection of probability distribution in the data layer, a fourth layer "L-moments layer" is added in our hierarchical model to indentify the best-fit distribution for the observation data.

Bayesian Hierarchical Model Framework

The four layers of our regional Bayesian hierarchical model can be seen in Figure 1. The details of L-moments layer can be seen in the next section.

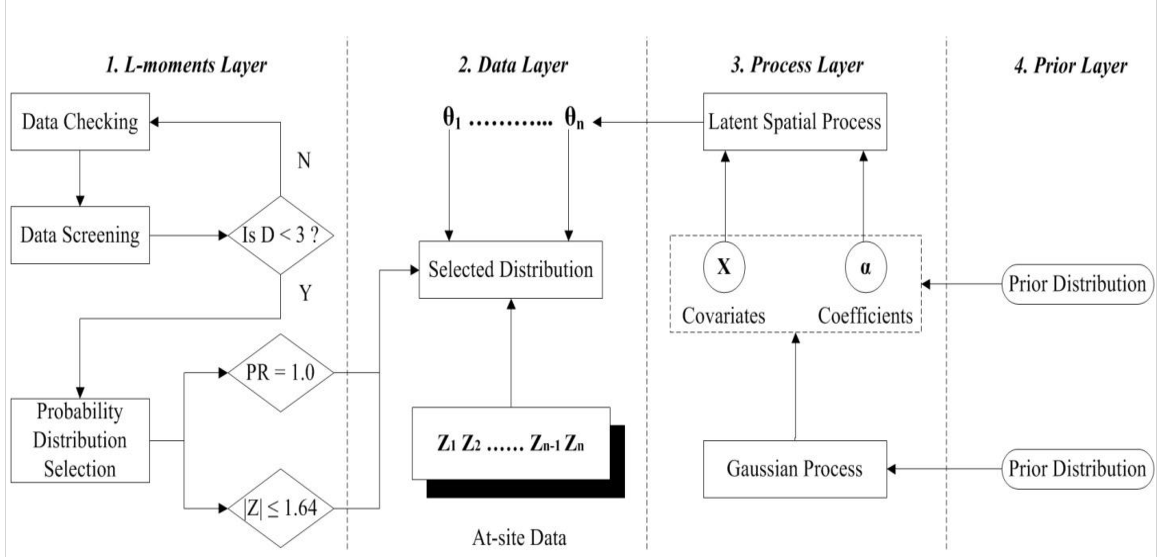

Figure 1. Flowchart for the Bayesian hierarchical model structure.

L-moments Layer

Hosking and Wallis (1997) gives a comprehensive study of L-moments, so the mathematics concerning L-moments is left out of this poster.

D Value:

$$D_i = \frac{1}{3} N (\mathbf{u}_i - \bar{\mathbf{u}})^T \mathbf{A}^{-1} (\mathbf{u}_i - \bar{\mathbf{u}})$$

The D value is for data discordancy test, and if $D_i > 3$ at site i, then this site is considered to be discordant from the region and the data should be rechecked carefully.

Two methods for the selection of best-fit distribution:

1. PR (Performance Ratio)

PR is based on the average weighted orthogonal distance (AWOD) between sample and theoretical L-moments. The best-fit distribution has a PR value 1 (Kroll and Vogel, 2002).

$$PR = AWOD_{Best} / AWOD$$

2. Z^{DIST} Value

$$Z^{DIST} = (\tau_4^{DIST} - t_4^R + B_4)/\sigma_4$$

The fit is considered to be adequate if $|Z^{DIST}|$ is close to zero, with a reasonable criterion being $|Z^{DIST}| \le 1.64$.

Case Study: Area and Data

The drainage area for the Willamette River Basin (WRB) is 29,728 km^2, and covers about 12% of Oregon. A total of 51 USGS streamflow gages were carefully selected for this study. The annual peak flow data for this study were up through the water year 2012.

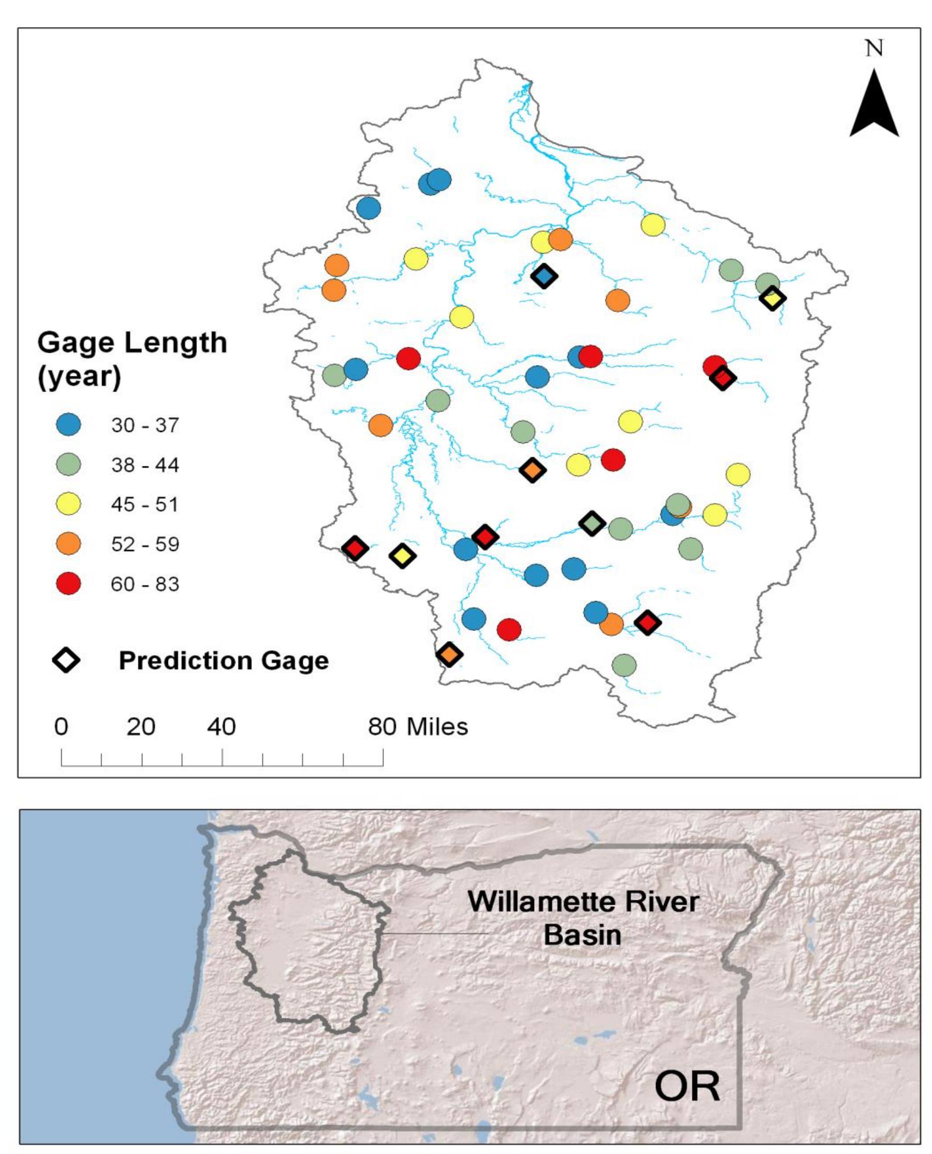

Figure 2. The Willamette River Basin and the stream gages.

Because the goodness-of-fit tests are not reliable for a small number of observations, only those data sets which have more than 30 observations are considered.

Best-fit Distribution by L-moments Layer

Figure 3 presents the L-moments diagram for the 51 gages. Both the PR and Z^{DIST} values indicate that the generalized extreme value (GEV) distribution performs the best and is chosen as the likelihood function in the data layer.

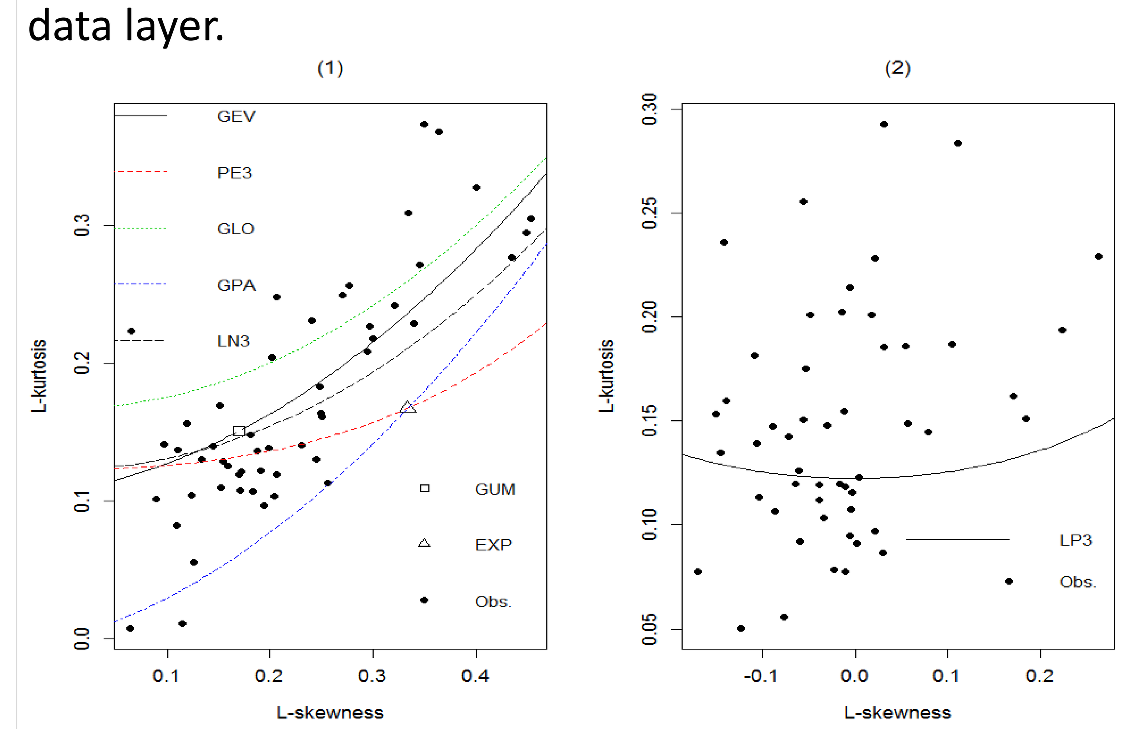

Figure 3. L-moments diagrams of L-skewness versus L-kurtosis for 51 gages. Gumbel (GUM), exponential (EXP), generalized logistic (GLO), Pearson Type III (PE3), lognormal (LN3), generalized Pareto (GPA), log-Pearson Type III (LP3).

Covariates Selection and Model Comparison

In the process layers, the geostatistic variogram model is used for this case study. How to select the appropriate covariates, however, remains no specific criterion in literature. In this study, we propose a covariates selection process by combing maximum likelihood estimate (MLE), exploratory data analysis (EDA), and deviance information criterion (DIC) (Spiegelhalter et al., 2002) together (Figure 4).

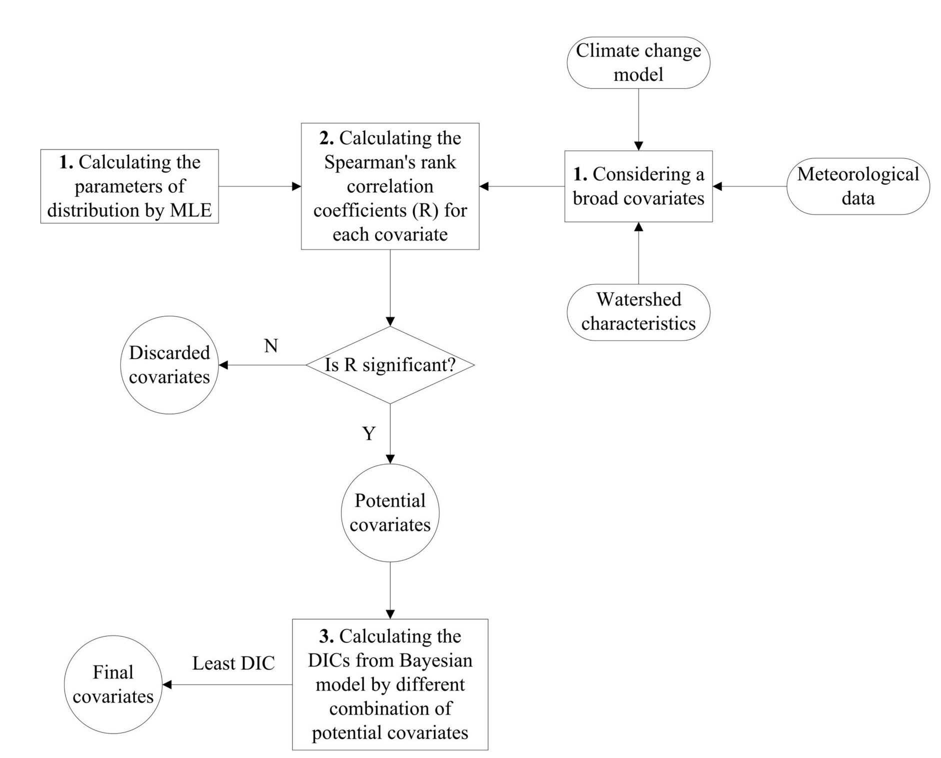

Figure 4. Flowchart for covariates and model selection.

MCMC Structure

The Metropolis-Hasting algorithm was used to draw samples. Three parallel chains with different initials values were run for each model. Each simulation was performed for 200,000 iterations. The burn-in size for each simulation was 50,000, and the thinning factor was 50. The scale reduction factor $\hat{R}$ for all the models were less than 1.05, which suggests a good convergence of the model.

Posterior Density

A total of 16 covariates were considered at first, and our DICs indicated that combing only the drainage area gave the best model. Figure 5 shows the posterior densities of parameters using two USGS gages (14144800 and 14146500) as instances.

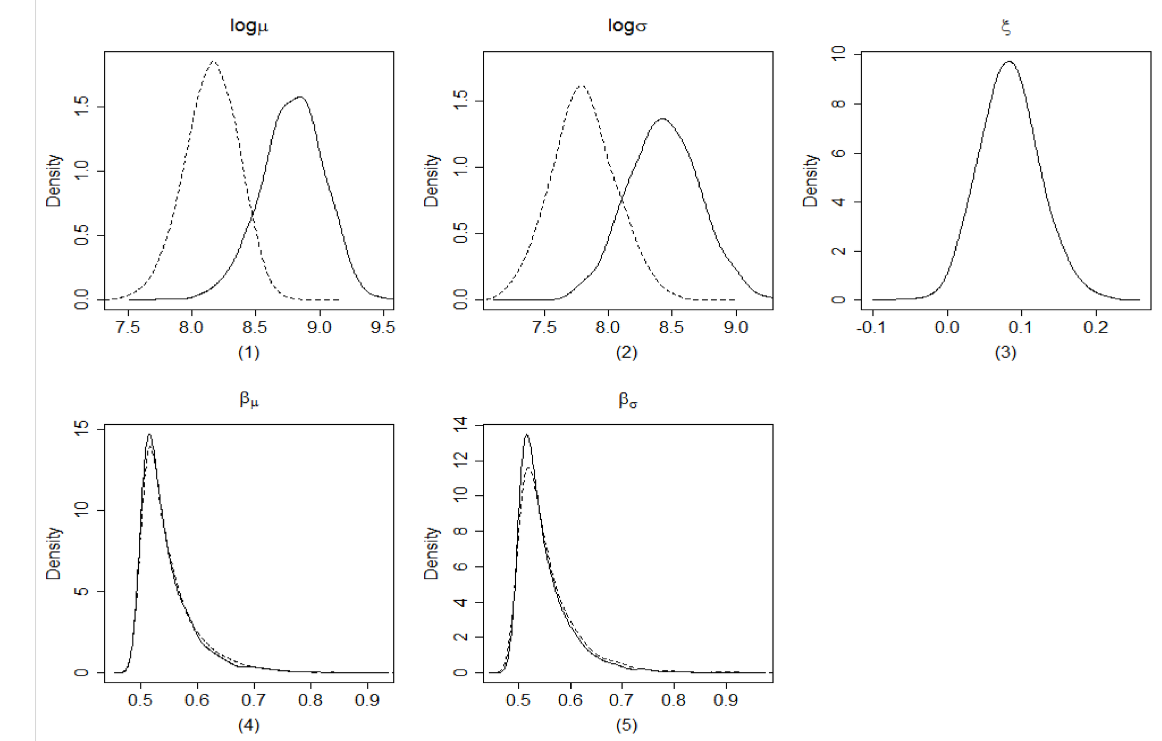

Figure 5. Posterior density plots from the best model.

The 100-year Flood Quantile

The 100-year flood quantiles and the 95% confidence interval were calculated by MLE and hierarchical model (Figure 6). Out of the 51 gages, a total of 36 gages (71%) show a reduction of uncertainty.

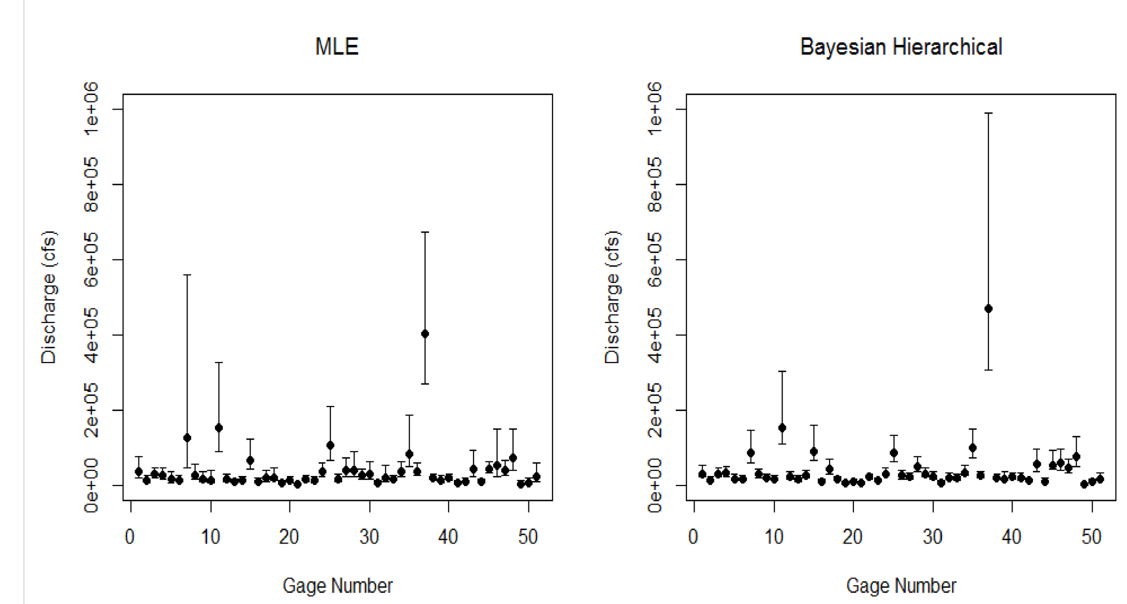

Figure 6. Point estimation (MLE and posterior mode) and the corresponding 95% intervals for 100-year flood to 51 gages.

Conclusion

1. A regional Bayesian hierarchical model is developed for flood frequency analysis and a covariate selection process is proposed for the Bayesian hierarchical model.
2. The spatial Bayesian model can reduce the uncertainties for flood quantiles and provide a more reliable and accuracy uncertainty estimation.
3. The results of temporal change analysis indicate that the 100-year flood quantile has been increasing from 1941 to 2000.

Literature Cited

Hosking, J.R.M., and Wallis J.R. (1997). "Regional frequency analysis: An approach based on L – moments", Cambridge University Press.

Kroll, C.N. and Vogel, R.M. (2002). "Probability distribution of low streamflow series in the United States." Journal of Hydrologic Engineering, 7(2), 137-146.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002). "Bayesian measures of model complexity and fit." Journal of the Royal Statistical Society: Series B, 64(4), 583-639.