Extreme Rainfall Events in Connecticut with Weighted Likelihood

Xiaojing Wang[1], Mekonnen Gebremicahel[2], and Jun Yan[1]

[1]Department of Statistics, University of Connecticut
[2]Department of Civil and Environmental Engineering, University of Connecticut

ABSTRACT Annual extreme rainfall events are constructed from raw precipitation data of every 15 minutes at 12 stations with various length in the state of Connecticut. Three characteristics, the volume, duration, and peak intensity, are modeled by a multivariate distribution specified by three marginal distributions and a dependence structure via copula. A major issue in this application is that the sample size at most stations are small, ranging from 10 to 33, because the 15-minute precipitation data are only available fairly recently. For each station, we propose to estimate the model parameters by maximizing a weighted likelihood, which assigns weight to data at stations nearby, borrowing strengths from them. The weights are assigned by some kernel function whose bandwidth is chosen by cross-validation in terms of predictive loglikelihood. The analysis of the extreme rainfall events in Connecticut shows substantial improvement in predictive loglikelihood by kernel weighting.

1 Introduction

- Hydrologic designs require accurate estimates of design rainfall
- The conventional approach of estimating design rainfall is subject to erro rs, since it does not take into account the dependence among rainfall intensity, depth, and duration.
- Recently, *Kao and Govindaraju* (2008) demonstrated that the trivariate copula perfo rms well for these trivariate random variables at the hourly scale.
- In this study, we test the applicability of trivariate copula models to 15 -min rainfall datasets.
- Analysis of the data at such short time scale is challenging, due to the s hort record length of the data.
- We use weighted likelihood approach to overcome this challenge

2 Data

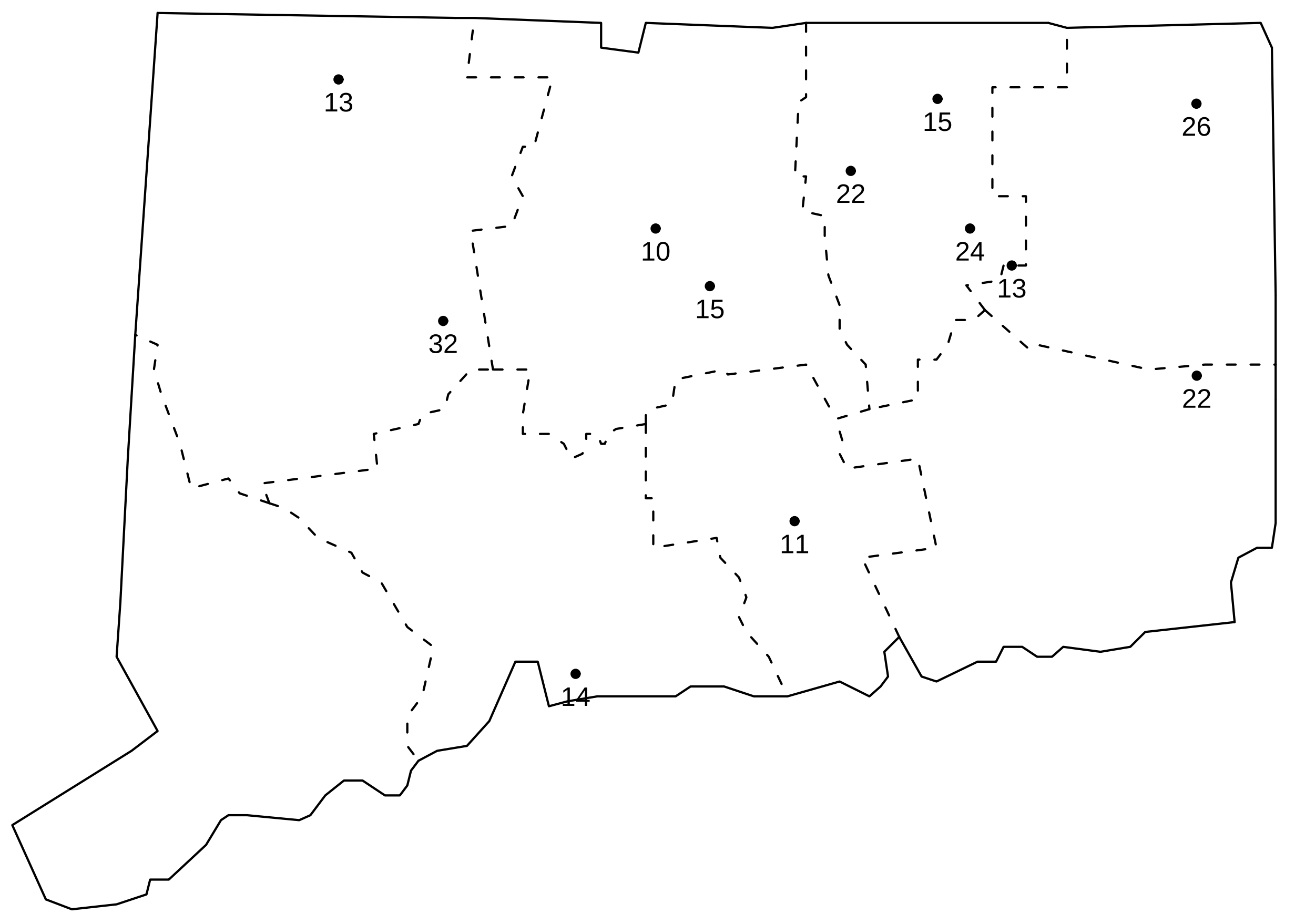

Figure 1: Location of stations in Connecticut and their sample sizes.

- Raw precipitation data of every 15 minutes for 16 stations in Connecticut from the NCDC.
- At each station, rainfall event records are obtained with the rainfall volume, duration and peak 15-minute intensity.
- The annual extreme rainfall event at a given station in a given year is chosen to be the rainfall event which possesses the largest joint cumulative probability of volume and peak intensity.

3 Copula Model

Let θ be the vector of parameters of multivariate CDF H, containing both marginal parameters in F_i, $i = 1, \ldots, p$, and copula parameters in C. The PDF of H is

$$h(\mathbf{x}; \theta) = c\{F_1(x_1), \ldots, F_p(x_p)\} \prod_{i=1}^{p} f_i(x_i), \tag{1}$$

where c is the PDF of C, and f_i is the PDF of F_i, $i = 1, \ldots, p$. Given a random sample $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ from H, the loglikelihood is then

$$L(\theta; \mathcal{X}) = \sum_{i=1}^{n(\mathcal{X})} \log h(\mathbf{X}_i; \theta), \tag{2}$$

where $n(\mathcal{X})$ is the sample size of $\mathcal{X}$.

A metaellipitical copula is the copula determined by an elliptical distribution, characterizes by a dispersion matrix Σ, whose diagonal elements are all 1. Let G_Σ be the CDF of a elliptical distribution with dispersion matrix Σ, and F be the CDF of all the margins. The implicitly determined metaelliptical copula C is

$$C(\mathbf{u}; \Sigma) = G\{F^{-1}(u_1), \ldots, F^{-1}(u_p)\}, \tag{3}$$

where $\mathbf{u} = (u_1, \ldots, u_p) \in (0, 1)^p$. Metaelliptical copulas provide greater flexibility in pairwise dependence structure through the dispersion matrix Σ than another class of copulas, Archimedean copulas, which restricts that all pairs of variables share the same dependence structure (*Genest et al.*, 2007).

4 Weighted Likelihood

Since the sample sizes at some stations are so small, estimation based on observations at these stations alone can be unreliable and even numerically infeasible. We approach the problem by pooling observations from stations nearby to construct the weighted likelihood (*Hu and Zidek*, 2002). Let $\mathcal{X}_s = \{\mathbf{X}_{s,1}, \ldots, \mathbf{X}_{s,n_s}\}$, $s = 1, \ldots, S$, be a random sample of size n_s for station s. Let $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_S\}$ be the collection of all data from all stations. We use a kernel K as weight function. The weighted loglikelihood for station s with bandwidth h is

$$L_s(\theta; \mathcal{X}; h) = \sum_{t=1}^{S} K\left(\frac{d_{s,t}}{h}\right) L(\theta; \mathcal{X}_t), \tag{4}$$

where $d_{s,t}$ is the distance between station s and t. The parameter θ for station s is estimated by the maximizer $\hat{\theta}_s(h)$ of $L_s(\theta; \mathcal{X}, h)$. As $h \to 0$, only data at station s is used to fit H. In this case, $\hat{\theta}_s(0)$ maximizes the loglikelihood based on the local data alone, but the variation of the estimator can be large, especially for stations with a small number of observations. As $h \to \infty$, all data will be used to fit the distribution H and the fitted H will be the same for all stations.

For station s, let $\mathcal{X}_s^{(k)}$ be the data in fold k and $\mathcal{X}_s^{(-k)}$ all the data except those in fold k. Let $\hat{\theta}_s^{(-k)}(h)$ be the estimate of θ with bandwidth h at station s, based on $\mathcal{X}_s^{(-k)}$. Define cross validation score at station station s with bandwidth h as

$$\mathrm{CV}_s(h) = \sum_{k=1}^{K} L(\hat{\theta}_s^{(-k)}(h); \mathcal{X}_s^{(k)}). \tag{5}$$

The overall cross-validation score of bandwidth h is then

$$\mathrm{CV}(h) = \sum_{s=1}^{S} \mathrm{CV}_s(h). \tag{6}$$

We choose h that maximizes $\mathrm{CV}(h)$ because it leads to a model with the highest predictive capability.

5 Connecticut Rainfall Event Analysis

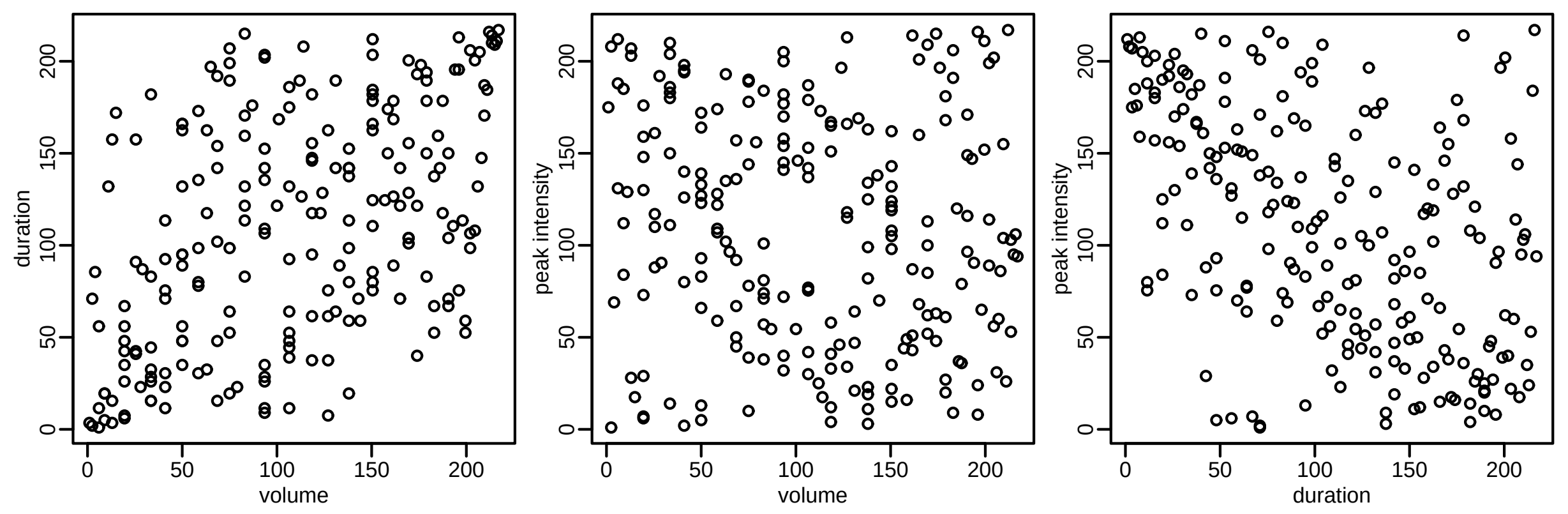

Figure 2: Scatter plots pairs of ranks among variable volume, duration and peak intensity based on the pooled data.

Dependence among the three variables turns out to indeed exist. As an exploratory analysis, we pool the data from all stations and plot pairwise ranks of the three variables in Figure 2. These plots suggest positive dependence between volume and duration, and negative dependence between duration and peak intensity. Formal multivariate independence test proposed by *Genest and Rémillard* (2004) is carried out with the pooled data with the R package `copula` (*Yan and Kojadinovic*, 2008). The test results confirmed visual impression observed in Figure 2.

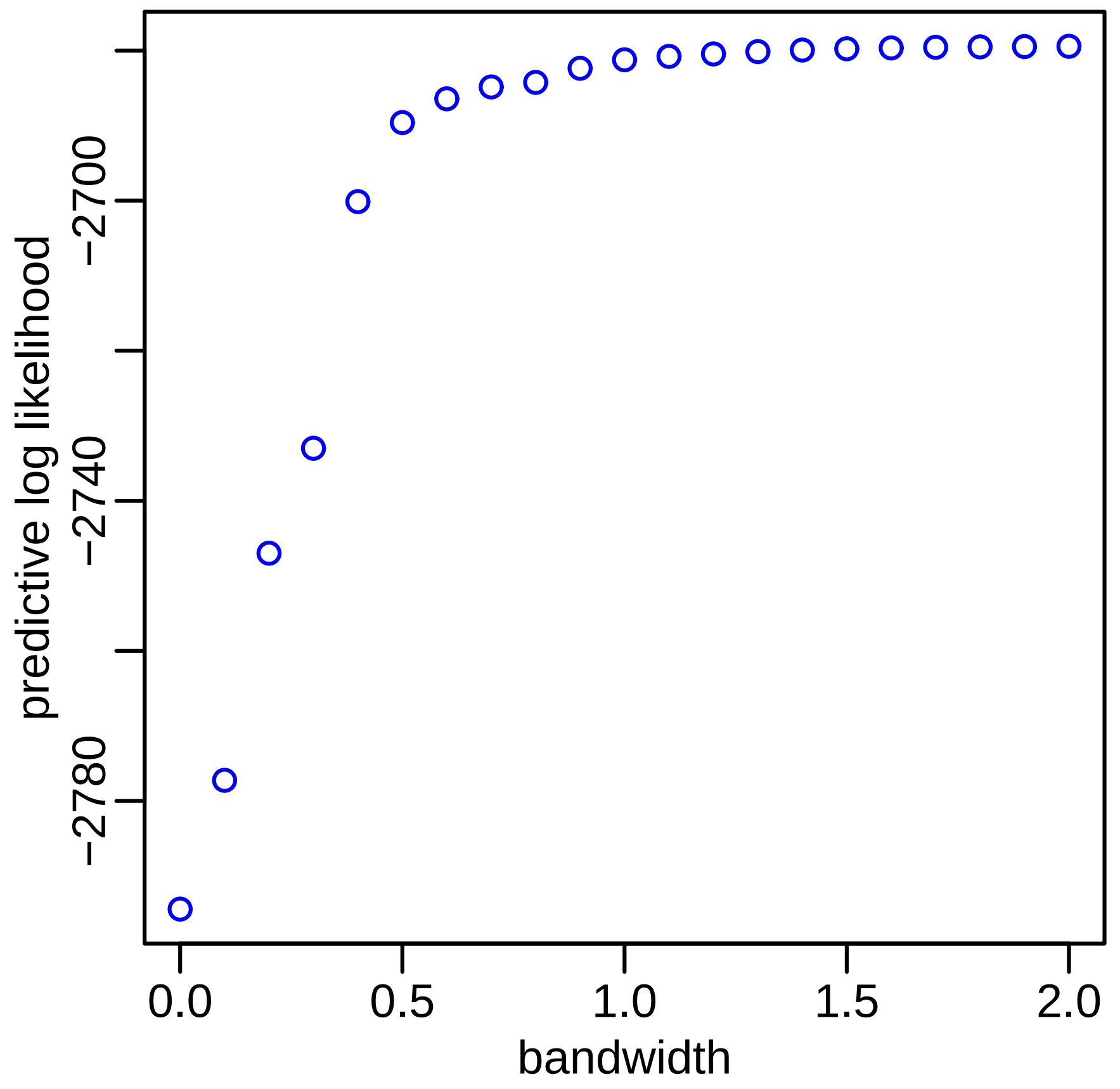

Figure 3: Predictive loglikelihood via 5-fold cross validation for bandwidth over a grid.

The 5-fold overall cross-validation score for bandwidth over a grid from 0 to 2 is shown in Figure 3. Recall that bandwidth 0 means no borrowing strength from sites nearby. As the bandwidth increases, the predictive loglikelihood increases at the beginning, and then levels off after the bandwidth goes beyond 1.0. The difference in predictive loglikelihood at bandwidth 1.0 and bandwidth 0 is approximately 100, which is quite substantial. Since the cross-validation scores is approaching an asymptote in this application, we cannot choose bandwidth to maximize it. Instead, we choose the smallest bandwidth such that the cross-validation score is close enough to the asymptote within certain tolerance. The chosen bandwidth is 1.0.

Table 1: Parameter estimates and standard errors from maximizing weighted likelihood with bandwidth 1. In each cell, the upper entry is the estimate and the lower entry is the standard error.

Station	n	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{3,1}$	$\beta_{3,3}$	ρ_{12}	ρ_{13}	ρ_{23}
061488	11	5.579	0.413	1.897	7.098	2.278	0.525	0.546	-0.102	-0.494
		0.095	0.068	0.566	2.457	0.119	0.085	0.187	0.261	0.202
062169	14	5.584	0.437	1.987	6.845	2.278	0.524	0.549	-0.055	-0.453
		0.102	0.073	0.609	2.411	0.122	0.087	0.188	0.264	0.215
063449	10	5.575	0.415	1.922	6.976	2.287	0.522	0.543	-0.107	-0.486
		0.095	0.068	0.576	2.412	0.119	0.085	0.187	0.260	0.204
063451	15	5.576	0.413	1.906	7.038	2.285	0.523	0.543	-0.109	-0.491
		0.094	0.067	0.569	2.429	0.119	0.085	0.187	0.260	0.203
063857	22	5.575	0.384	1.812	7.421	2.284	0.523	0.539	-0.166	-0.537
		0.085	0.060	0.520	2.495	0.114	0.082	0.187	0.256	0.187
064488	13	5.575	0.396	1.821	7.342	2.281	0.526	0.547	-0.144	-0.523
		0.089	0.063	0.531	2.505	0.117	0.084	0.185	0.258	0.193
065445	13	5.559	0.429	2.031	6.620	2.312	0.500	0.532	-0.106	-0.471
		0.099	0.070	0.612	2.275	0.114	0.081	0.188	0.254	0.206
066942	22	5.575	0.404	1.857	7.196	2.286	0.524	0.544	-0.131	-0.508
		0.091	0.065	0.546	2.462	0.118	0.084	0.186	0.258	0.197
067959	15	5.574	0.397	1.827	7.295	2.286	0.525	0.547	-0.149	-0.520
		0.089	0.064	0.533	2.482	0.117	0.084	0.185	0.256	0.193
068138	24	5.576	0.398	1.827	7.313	2.282	0.526	0.548	-0.142	-0.520
		0.089	0.064	0.533	2.493	0.117	0.084	0.185	0.257	0.193
068330	32	5.573	0.430	2.000	6.747	2.292	0.515	0.541	-0.083	-0.463
		0.100	0.071	0.608	2.348	0.119	0.084	0.188	0.260	0.210
069388	26	5.570	0.381	1.792	7.461	2.285	0.528	0.542	-0.166	-0.536
		0.083	0.059	0.509	2.482	0.114	0.082	0.183	0.253	0.185

Table 1 presents the parameter estimates and their standard errors for each station with a common bandwidth 1.0. The parameter estimates are similar from station to station, but there are still considerable differences. Such estimates borrow strength from data at stations nearby and at the same time, each station keeps its own identity.

The positive association of volume and duration is significant at all stations as indicated by the standard errors. The point estimates are the same in the first two digits. The negative association of duration and peak intensity is also significant at all stations, but the point estimates are not very similar, ranging from 0.453 to 0.537. The association of volume and peak intensity is estimated as negative, but not significant at all stations.

References

Genest, C., and B. Rémillard (2004), Test of independence and randomness based on the empirical copula process, *TEST*, *13*(2), 335–369.

Genest, C., A. C. Favre, J. Béliveau, and C. Jacques (2007), Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data, *Water Resources Research*, *43*(9).

Hu, F., and J. V. Zidek (2002), The weighted likelihood, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *30*(3), 347–371.

Kao, S. C., and R. S. Govindaraju (2008), Trivariate statistical analysis of extreme rainfall events via the plackett family of copulas, *Water Resources Research*, *44*(2).

Yan, J., and I. Kojadinovic (2008), `copula:` *Multivariate dependence with copulas*, r package version 0.8-0.