Spatially smooth regional estimation of the flood frequency curve (with uncertainty)

D. Ganora, P. Claps, F. Laio

Politecnico di Torino, Italy
Flood frequency curve estimation in ungauged basins

- Regional approach
- Ease of use
- Handle short records
Regionalization approach

Statistical regionalization through morpho-climatic basin-scale descriptors

Model (multiple regression)

Measured variable

Basin’s descriptors
Regionalization approach

Statistical regionalization through morpho-climatic basin-scale descriptors

Model (multiple regression)

Measured variable
Basin’s descriptors

Ungauged site’s descriptors

Estimated variable
Estimated uncertainty
Approach characteristics

- Challenging some standard approach assumptions
  - choice of a priori probability distribution
  - regions definition

<table>
<thead>
<tr>
<th>New gauging stations</th>
<th>available years</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>20</td>
<td>70</td>
</tr>
</tbody>
</table>
Approach characteristics

- Challenging some standard approach assumptions
  - choice of a priori probability distribution
  - regions definition
- Uncertainty (variance) of final estimates

<table>
<thead>
<tr>
<th>New gauging stations</th>
<th>available years</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>70</td>
</tr>
</tbody>
</table>
Approach characteristics

- Challenging some standard approach assumptions
  - choice of a priori probability distribution
  - regions definition

- Uncertainty (variance) of final estimates

- Maximization of available at-site information:
  - non-systematic data
  - “poorly” gauged sites (short records)
Approach characteristics

- Challenging some standard approach assumptions
  - choice of a priori probability distribution
  - regions definition
- Uncertainty (variance) of final estimates
- Maximization of available at-site information:
  - non-systematic data
  - “poorly” gauged sites (short records)

Suitable method for Northwestern Italy
(about 100 new stations)
Data handling

Use of $L$-moments statistics

- Possibility to incorporate non-systematic data:
  - historical floods
  - occasional extreme events during gap in time series

$L$-moments variance via simplified formulae
Use of \( L \)-moments statistics

- Possibility to incorporate non-systematic data:
  - historical floods
  - occasional extreme events during gap in time series
- \( L \)-moments variance via simplified formulae

Flood frequency curve

\[
Q_T = Q_{\text{ind}} \cdot P(T, L_{CV}, L_{CA})
\]

\[
\begin{align*}
Q_{\text{ind}} &\rightarrow \text{scale (index-flood)} \\
L_{CV} &\rightarrow \text{dispersion} \\
L_{CA} &\rightarrow \text{skewness}
\end{align*}
\]
Which variable to regionalize?

Usually: a priori choice of distribution

- Quantile regionalization
- Regionalization of the parameters of the distribution
Which variable to regionalize?

Usually: a priori choice of distribution
- Quantile regionalization
- Regionalization of the parameters of the distribution

![Graph showing multiple frequency distributions for Chisone at S. Martino.](image)

Figure 2.8: Example of sample flood data for the river Chisone at S. Martino and superposition of different theoretical frequency distributions: Gamma (GAM), generalized extreme value (GEV), lognormal (LN3), Gumbel (G), generalized logistic (GL) and generalized Pareto (GP). Black dots represent empirical data, circled ones correspond to non-systematic events; plotting positions come from equation (2.30).

where

\[ m \] is the equivalent sample length (as in section 2.2.1),
\[ g \] is the total number of flood events available (both systematic and non-systematic),
\[ k \] is the total number of events that exceed the threshold \( x_0 \),
\[ s \] is the length of the systematic sample and \( e \) is the number of measures of the systematic sample that exceed the threshold.

An example is shown in figure 2.8 for the river Chisone at S. Martino, where the sample points are plotted using the plotting position of equation (2.30) and the circled dots highlight the non-systematic measurements. The example shows that all the distributions have a similar behavior up to a 100-years return period, except for the Gumbel distribution that is a less flexible distribution with only two parameters.
Which variable to regionalize?

**Usually: a priori choice of distribution**
- Quantile regionalization
- Regionalization of the parameters of the distribution

**Distribution-free statistics**
- $L$-moments regionalization
- A posteriori reconstruction of frequency distribution
How to regionalize?

Usually: homogeneous regions

Difficult uncertainty estimation due to:
- regions creation
- regions border effects
How to regionalize?

Usually: homogeneous regions

Difficult uncertainty estimation due to:
- regions creation
- regions border effects

Smooth variability of $L$-moments in the descriptors space
- No homogeneity required
- Easier uncertainty evaluation
Regional model definition

- Multiple linear regression
- Error structure as in *Stedinger & Tasker (1985)*
  - Model error $Y_T = X \beta + \delta$
  - Sampling error $Y = Y_T + \eta$
Regional model definition

- Multiple linear regression
- Error structure as in *Stedinger & Tasker (1985)*
  - Model error $Y_T = X \beta + \delta$
  - Sampling error $Y = Y_T + \eta$

**GLS concurrent estimation of regression coefficients and model variance**

\[
Y = X\beta + \varepsilon
\]
\[
\Lambda = \sigma_\delta^2 I + \Sigma
\]
Regional model definition

- Multiple linear regression
- Error structure as in *Stedinger & Tasker (1985)*
  - Model error $Y_T = X \beta + \delta$
  - Sampling error $Y = Y_T + \eta$

**GLS concurrent estimation of regression coefficients and model variance**

$$Y = X \beta + \varepsilon$$

$$\Lambda = \sigma^2_\delta I + \Sigma$$

**Variance of prediction**

$$VP = \sigma^2_Y = \sigma^2_\delta + x (X^T \Lambda^{-1} X)^{-1} x^T$$
Case study: 70 basins - 35 descriptors

Application to a large descriptors set: all the combination with 1 to 4 descriptors (+ intercept) are calculated
Case study: 70 basins - 35 descriptors

**Application to a large descriptors set:** all the combination with 1 to 4 descriptors (+ intercept) are calculated

- Model validation
  - Multicollinearity test
  - Student’s test
  - Residuals check
Case study: 70 basins - 35 descriptors

Application to a large descriptors set: all the combination with 1 to 4 descriptors (+ intercept) are calculated

- Model validation
  - Multicollinearity test
  - Student’s test
  - Residuals check
- Model selection (among all combinations)
  - Model variance $\sigma^2_\delta$
  - Average variance of predictions ($AVP$)
  - Simplest model
Example: index-flood estimation

\[
\log (Q_{\text{ind}}) = -8.76 + 7.99 \times 10^{-1} \cdot A + 1.09 \cdot IDFa \\
+ 9.53 \times 10^{-1} \cdot MAP + 7.85 \times 10^{-1} \cdot perm
\]
Example: $L_{CV}$

\[ L_{CV} = 6.44 \times 10^{-1} - 4.28 \times 10^{-7} \cdot X_c - 5.00 \times 10^{-4} \cdot P - 1.44 \times 10^{-4} \cdot H_{min} \]
Example: $L_{CA}$

\[
L_{CA} = 9.38 \times 10^{-1} - 1.40 \times 10^{-2} \cdot IDFa - 1.39 \times 10^{-1} \cdot P - 2.65 \times 10^{-4} \cdot H_{min}
\]
Quantile estimation

- $Q_{ind}$, $L_{CV}$ and $L_{CA}$ → different suitable distributions
- Flood frequency curve as the average
- Confidence bands through Monte Carlo simulations (using also $\sigma_{Q_{ind}}^2$, $\sigma_{L_{CV}}^2$ and $\sigma_{L_{CA}}^2$)
Quantile estimation

- \( Q_{ind}, L_{CV} \) and \( L_{CA} \) \( \rightarrow \) different suitable distributions
- Flood frequency curve as the average
- Confidence bands through Monte Carlo simulations (using also \( \sigma^2_{Q_{ind}}, \sigma^2_{L_{CV}} \) and \( \sigma^2_{L_{CA}} \))

10 – Chisone a S.Martino

Sample estimation

Regional estimation

![Sample estimation graph](image1)

![Regional estimation graph](image2)
Final remarks: a practical tool

Shaded area = sample estimate with variance lower than regional one