Regional Frequency Analysis of Floods

Dr. V. V. Srinivas
Associate Professor
Department of Civil Engineering, IISc Bangalore 560 012, India
Overview of Presentation

- Introduction
- Definitions
- Methods for Design Flood Estimation
- Case studies
Introduction

Common problems in hydrologic design

• Ungauged sites

• Inadequate at-site information at gauged sites
This perimeter levee surrounding the Clarence Cannon National Wildlife Refuge, Mo., damaged during the 1993 flooding, allowed excessive water to enter the refuge.

(Photograph courtesy of U.S. Fish and Wildlife Service.)
Introduction

Typical Applications of Predicting Floods

- Planning and design of water-resources systems
- Risk assessment of water control/conveyance structures where failure can cause devastation
  - Culverts, Bridges, Storm Sewers, Dams, Levees, Flood Walls, floodway channels etc.
- Economic evaluation of Engineering projects
- Land use planning and management
- Insurance assessment
Definitions

**Complete duration series:**
All the peak flow data available for a location.

**Annual maximum series**
A series prepared using the largest values occurring in each of the years of the record.
**Definitions**

**Partial duration series [or peak-over-threshold (POT) series]**
A series prepared using data points selected from the observed record so that their magnitude is greater than a predefined threshold value.

**POT(K) series:** PDS in which the number of data points is about K times the number of years of the record.

**Annual exceedence series [or POT (1) series]**
(i) Two consecutive peaks must be separated by at least three times the average time to rise.

At least five clean (i.e., not multi-peaked) events, whose peaks exceed the threshold must be considered.
Figure: Elements of Flood Hydrograph

- **Hydrograph components**
  - $MA$ = base flow recession
  - $AB$ = rising limb
  - $BC$ = crest segment
  - $CD$ = falling limb
  - $DN$ = base flow recession

- Points $B$ and $C$ = inflection points

- **Discharge in m$^3$/s**

- **Time in hours**

- **Time to rise**
(ii) The minimum flow value in the trough between two peaks must be less than two-thirds of the flow value of the first of the two peaks.
Average time to rise = 5 hours

Two consecutive peaks must be separated by at least 15 hours

Peak ‘E’ is the largest and is therefore independent

Peak ‘D’ occurs less than 15 hours before peak ‘E’ and is defined as dependent

Peak ‘F’ is defined as dependent since although it occurs more than 15 hours after the peak ‘E’, the minimum discharge in the trough between the two peaks does not fall by more than two-thirds of the peak discharge for event ‘E’
- Peak ‘C’ is larger than peaks ‘A’ and ‘B’ and is judged independent of peak ‘E’ because
  - it occurs more than 15 hours beforehand
  - the minimum discharge in the trough between the two is less than two-thirds of the discharge for peak C.
- Peak ‘B’ occurs less than 15 hours before peak ‘C’ and is therefore dependent
- Peak ‘A’ is below the threshold and therefore no a POT event

(FEH, 1999)
Definitions

Random variable
A variable is said to be random if its value depends on the outcome of a chance event.

Examples of Event:
- Culvert capacity being exceeded
- Stage of flood flow exceeding deck height of a bridge

Recurrence interval ($\tau$)
Time interval between any two occurrences of an event.

Return Period ($T$)
Expected value of $\tau$ for events equaling or exceeding a specified magnitude $T = E(\tau)$.

Probability of occurrence of an event
Let $p = P(X \geq x_T)$ denote the probability of occurrence of an event $X \geq x_T$.

$p$ can be related to return period $T$ as, $p = 1/T$.

Probability of failure (non-occurrence of the event)

$$F = P(X < x_T) = 1 - P(X \geq x_T) = 1 - \frac{1}{T}.$$
Probability of occurrence of an event

For each observation in peak flow time series, there are two possible outcomes
• Occurrence of event if \( X \geq x_T \)
• Non-occurrence (failure) if \( X < x_T \)

In peak flow time series, observations must be independent of each other.

The probability of a recurrence interval of duration \( \tau = p(1 - p)^{r-1} \)

Return period of the event = \( T = \)

\[
E(\tau) = \sum_{r=1}^{\infty} \tau C_1 p(1 - p)^{r-1} = p \left[ 1 + 2(1 - p) + 3(1 - p)^2 + 4(1 - p)^3 \right]
\]

The expression in parenthesis has the form of the power series expansion with \( x = -(1-p) \) and \( n = -2 \)

Power series expansion, \( (1 + x)^n = 1 + nx + \frac{n(n-1)}{2!} x^2 + \frac{n(n-1)(n-2)}{3!} x^3 + \ldots \)

\[ \therefore T = E(\tau) = p(1 + x)^n = p \left[ 1 - (1 - p) \right]^{-2} = p^{-1} \]

\[ \Rightarrow p = \frac{1}{T} \]
<table>
<thead>
<tr>
<th>Type of structure</th>
<th>Return period (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highway culverts</td>
<td></td>
</tr>
<tr>
<td>Low traffic</td>
<td>5–10</td>
</tr>
<tr>
<td>Intermediate traffic</td>
<td>10–25</td>
</tr>
<tr>
<td>High traffic</td>
<td>50–100</td>
</tr>
<tr>
<td>Highway bridges</td>
<td></td>
</tr>
<tr>
<td>Secondary system</td>
<td>10–50</td>
</tr>
<tr>
<td>Primary system</td>
<td>50–100</td>
</tr>
<tr>
<td>Farm drainage</td>
<td></td>
</tr>
<tr>
<td>Culverts</td>
<td>5–50</td>
</tr>
<tr>
<td>Ditches</td>
<td>5–50</td>
</tr>
<tr>
<td>Urban drainage</td>
<td></td>
</tr>
<tr>
<td>Storm sewers in small cities</td>
<td>2–25</td>
</tr>
<tr>
<td>Storm sewers in large cities</td>
<td>25–50</td>
</tr>
<tr>
<td>Airfields</td>
<td></td>
</tr>
<tr>
<td>Low traffic</td>
<td>5–10</td>
</tr>
<tr>
<td>Intermediate traffic</td>
<td>10–25</td>
</tr>
<tr>
<td>High traffic</td>
<td>50–100</td>
</tr>
</tbody>
</table>

(Ref: Chow et al., 1988)
<table>
<thead>
<tr>
<th>Type of structure</th>
<th>Return period (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levees</td>
<td></td>
</tr>
<tr>
<td>On farms</td>
<td>2–50</td>
</tr>
<tr>
<td>Around cities</td>
<td>50–200</td>
</tr>
<tr>
<td>Dams with no likelihood of loss of life (low hazard)</td>
<td></td>
</tr>
<tr>
<td>Small dams</td>
<td>50–100</td>
</tr>
<tr>
<td>Intermediate dams</td>
<td>100 +</td>
</tr>
<tr>
<td>Large dams</td>
<td>—</td>
</tr>
<tr>
<td>Dams with probable loss of life (significant hazard)</td>
<td></td>
</tr>
<tr>
<td>Small dams</td>
<td>100 +</td>
</tr>
<tr>
<td>Intermediate dams</td>
<td>—</td>
</tr>
<tr>
<td>Large dams</td>
<td>—</td>
</tr>
<tr>
<td>Dams with high likelihood of considerable loss of life (high hazard)</td>
<td></td>
</tr>
<tr>
<td>Small dams</td>
<td>—</td>
</tr>
<tr>
<td>Intermediate dams</td>
<td>—</td>
</tr>
<tr>
<td>Large dams</td>
<td>—</td>
</tr>
</tbody>
</table>
Definitions

Risk

A flood control structure fails if peak flow value exceeds $x_T$

The hydrologic risk of failure can be computed as,

$$R = 1 - \left[ 1 - P( X \geq x_T ) \right]^{n_e}$$

$n_e$ : Expected (or design) life of the structure

$$T = \frac{1}{1 - (1 - R)^{1/n_e}}$$

$$F = P( X < x_T ) = 1 - P( X \geq x_T ) = 1 - \frac{1}{T}$$

<table>
<thead>
<tr>
<th>Risk</th>
<th>$T$ (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>950</td>
</tr>
<tr>
<td>0.15</td>
<td>616</td>
</tr>
<tr>
<td>0.20</td>
<td>449</td>
</tr>
<tr>
<td>0.25</td>
<td>348</td>
</tr>
<tr>
<td>0.30</td>
<td>281</td>
</tr>
<tr>
<td>0.35</td>
<td>233</td>
</tr>
<tr>
<td>0.40</td>
<td>196</td>
</tr>
<tr>
<td>0.45</td>
<td>168</td>
</tr>
<tr>
<td>0.50</td>
<td>145</td>
</tr>
<tr>
<td>0.55</td>
<td>126</td>
</tr>
<tr>
<td>0.60</td>
<td>110</td>
</tr>
<tr>
<td>0.65</td>
<td>96</td>
</tr>
<tr>
<td>0.70</td>
<td>84</td>
</tr>
</tbody>
</table>
## Estimation of Return period (in years)

<table>
<thead>
<tr>
<th>Risk</th>
<th>Design life of structure (years)</th>
<th>100</th>
<th>50</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>950</td>
<td>475</td>
<td></td>
<td>238</td>
</tr>
<tr>
<td>0.15</td>
<td>616</td>
<td>308</td>
<td></td>
<td>154</td>
</tr>
<tr>
<td>0.20</td>
<td>449</td>
<td>225</td>
<td></td>
<td>113</td>
</tr>
<tr>
<td>0.25</td>
<td>348</td>
<td>174</td>
<td></td>
<td>87</td>
</tr>
<tr>
<td>0.30</td>
<td>281</td>
<td>141</td>
<td></td>
<td>71</td>
</tr>
<tr>
<td>0.35</td>
<td>233</td>
<td>117</td>
<td></td>
<td>59</td>
</tr>
<tr>
<td>0.40</td>
<td>196</td>
<td>98</td>
<td></td>
<td>49</td>
</tr>
<tr>
<td>0.45</td>
<td>168</td>
<td>84</td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>0.50</td>
<td>145</td>
<td>73</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>0.55</td>
<td>126</td>
<td>63</td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>0.60</td>
<td>110</td>
<td>55</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>0.65</td>
<td>96</td>
<td>48</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>0.70</td>
<td>84</td>
<td>42</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>0.75</td>
<td>73</td>
<td>37</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>0.80</td>
<td>63</td>
<td>32</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>0.85</td>
<td>53</td>
<td>27</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>0.90</td>
<td>44</td>
<td>22</td>
<td></td>
<td>11</td>
</tr>
</tbody>
</table>
How much at-site data do we need for Hydrologic design?

At least $T$ and preferably $2T$

Institute of Hydrology, UK

Limitations

- Ungauged sites
- Locations having scarcity of flood data

Gauging Stations in Ohio, USA

- Record Length (years)
- Gauging Stations in Indiana, USA

- Record Length (years)

(RL<60 years: 77%) (RL<60 years: 85%)
Gauging stations in Krishna basin

Gauging stations in Godavari basin

(RL< 50 years: 100% of stations)
Methods for Design Flood Estimation

- **Empirical methods**
  
  *Dicken’s formula, Ryve’s formula, Inglis’ formula, Nawab Jung Bahadur formula, Creager’s formula and Rational formula*

- **Frequency analysis methods**

- **Rainfall-runoff methods**

- **Unit hydrograph methods**
Empirical Methods for Design Flood Estimation

(a) Peak discharge formulae involving drainage area only

**Dicken’s formula**

\[ Q_p = CA^{3/4} \]

- \( Q_p \) – Peak discharge in \( m^3/s \)
- \( A \) – Catchment area in \( km^2 \)
- \( C \) - Constant

\( C = 11.5 \) (North India); 14-19.5 (Central India); 22 to 26 (Western Ghats of India)

\( C \) depends on Rainfall and Altitude of catchment

**Ryve’s formula**

\[ Q_p = CA^{2/3} \]

\( C = 6.8 \) for catchment areas within 80 km from the coast

\( = 8.8 \) for catchment areas within 80-2400 km from the coast
Empirical Methods for Design Flood Estimation

(a) Peak discharge formulae involving drainage area only

**Inglis’ formula**

\[ Q_p = \frac{123A}{\sqrt{A} + 10.4} \]

(for fan shaped catchments in old Bombay state)

- \( Q_p \) – Peak discharge in m\(^3\)/s
- \( A \) – Catchment area in km\(^2\)
- \( A' \) – Catchment area in mi\(^2\)
- \( C \) - Constant

**Nawab Jung Bahadur formula**

\[ Q_p = CA'[0.93-(1/14)\log A'] \]

(for Hyderabad Deccan Catchments); \( C=1.77 - 177 \)

**Other formulae:** Creager’s formula; Jarvis formula; Modified Myer’s formula
Empirical Methods for Design Flood Estimation

(b) Peak discharge formulae involving drainage area and its shape

Dredge or Burge formula (for Indian catchments)

\[ Q_p = 19.6 \frac{A}{L^{2/3}} = 19.6 \frac{(BL)}{L^{2/3}} \]

- \( Q_p \) – Peak discharge in m\(^3\)/s
- \( A \) – Catchment area in km\(^2\)
- \( L \) – Average length of catchment area in km
- \( B \) – Average width of the catchment in km

Pettis formula

\[ Q_p = C (P \cdot B)^{5/4} \]

- \( P \) – Probable 100 year maximum 1 day rainfall in cm
- \( C \) - Constant

\( C = 1.5 \) (humid areas); \( 0.2 \) (desert areas)

(for northern USA, Ohio to Connecticut)
Empirical Methods for Design Flood Estimation

(c) Peak discharge formulae involving rainfall intensity and drainage area

**Rational formula**

\[ Q_p = CiA \]

- \( Q_p \) – Peak discharge in cfs
- \( A \) – Catchment area in acres
- \( C \) – Runoff coefficient (0 ≤ C ≤ 1)
- \( i \) – Intensity of rainfall in inches/hr

- **The peak discharge corresponds to time of concentration**

- **Used widely in storm sewer designs, because of its simplicity**

\[ 1 \text{ cfs} = 1 \cdot 008 \text{ acre.in/hr} \]

(Conversion is considered to be included in the value of C)

The assumption of fixed C (i.e., ratio of peak discharge to rainfall rate) for a drainage basin is invalid.
In rational formula, the value of runoff coefficient ‘C’ depends on integrated effects of several factors.

- Character and condition of soil
- Rainfall intensity
- Proximity of the water table
- Degree of soil compaction
- Porosity of subsoil
- Vegetation
- Ground slope
- Ponding character of surface of catchment - Depression storage

Rainfall intensity $i$ is determined based on intensity-duration-frequency (IDF) relationships.

Design duration $D$ is equal to the time of concentration for the catchment.

The frequency $F$ (or corresponding return period, $T$) is chosen by hydrologist.
Fig. Intensity-duration-frequency curves of maximum rainfall in Chicago, USA (Source: Chow et al., 1988)
### TABLE 15.1.1
Runoff coefficients for use in the rational method

<table>
<thead>
<tr>
<th>Character of surface</th>
<th>Return Period (years)</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asphallic</td>
<td></td>
<td>0.73</td>
<td>0.77</td>
<td>0.81</td>
<td>0.86</td>
<td>0.90</td>
<td>0.95</td>
<td>1.00</td>
</tr>
<tr>
<td>Concrete/roof</td>
<td></td>
<td>0.75</td>
<td>0.80</td>
<td>0.83</td>
<td>0.88</td>
<td>0.92</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Grass areas (lawns, parks, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Poor condition</strong></td>
<td>(grass cover less than 50% of the area)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flat, 0–2%</td>
<td></td>
<td>0.32</td>
<td>0.34</td>
<td>0.37</td>
<td>0.40</td>
<td>0.44</td>
<td>0.47</td>
<td>0.58</td>
</tr>
<tr>
<td>Average, 2–7%</td>
<td></td>
<td>0.37</td>
<td>0.40</td>
<td>0.43</td>
<td>0.46</td>
<td>0.49</td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>Steep, over 7%</td>
<td></td>
<td>0.40</td>
<td>0.43</td>
<td>0.45</td>
<td>0.49</td>
<td>0.52</td>
<td>0.55</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>Fair condition</strong></td>
<td>(grass cover on 50% to 75% of the area)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flat, 0–2%</td>
<td></td>
<td>0.25</td>
<td>0.28</td>
<td>0.30</td>
<td>0.34</td>
<td>0.37</td>
<td>0.41</td>
<td>0.53</td>
</tr>
<tr>
<td>Average, 2–7%</td>
<td></td>
<td>0.33</td>
<td>0.36</td>
<td>0.38</td>
<td>0.42</td>
<td>0.45</td>
<td>0.49</td>
<td>0.58</td>
</tr>
<tr>
<td>Steep, over 7%</td>
<td></td>
<td>0.37</td>
<td>0.40</td>
<td>0.42</td>
<td>0.46</td>
<td>0.49</td>
<td>0.53</td>
<td>0.60</td>
</tr>
<tr>
<td><strong>Good condition</strong></td>
<td>(grass cover larger than 75% of the area)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flat, 0–2%</td>
<td></td>
<td>0.21</td>
<td>0.23</td>
<td>0.25</td>
<td>0.29</td>
<td>0.32</td>
<td>0.36</td>
<td>0.49</td>
</tr>
<tr>
<td>Average, 2–7%</td>
<td></td>
<td>0.29</td>
<td>0.32</td>
<td>0.35</td>
<td>0.39</td>
<td>0.42</td>
<td>0.46</td>
<td>0.56</td>
</tr>
<tr>
<td>Steep, over 7%</td>
<td></td>
<td>0.34</td>
<td>0.37</td>
<td>0.40</td>
<td>0.44</td>
<td>0.47</td>
<td>0.51</td>
<td>0.58</td>
</tr>
</tbody>
</table>

(Ref: Chow et al., 1988)
### TABLE 15.1.1
Runoff coefficients for use in the rational method

<table>
<thead>
<tr>
<th>Character of surface</th>
<th>Return Period (years)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>25</td>
<td>50</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>Undeveloped</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultivated Land</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flat, 0–2%</td>
<td>0.31</td>
<td>0.34</td>
<td>0.36</td>
<td>0.40</td>
<td>0.43</td>
<td>0.47</td>
<td>0.57</td>
</tr>
<tr>
<td>Average, 2–7%</td>
<td>0.35</td>
<td>0.38</td>
<td>0.41</td>
<td>0.44</td>
<td>0.48</td>
<td>0.51</td>
<td>0.60</td>
</tr>
<tr>
<td>Steep, over 7%</td>
<td>0.39</td>
<td>0.42</td>
<td>0.44</td>
<td>0.48</td>
<td>0.51</td>
<td>0.54</td>
<td>0.61</td>
</tr>
<tr>
<td>Pasture/Range</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flat, 0–2%</td>
<td>0.25</td>
<td>0.28</td>
<td>0.30</td>
<td>0.34</td>
<td>0.37</td>
<td>0.41</td>
<td>0.53</td>
</tr>
<tr>
<td>Average, 2–7%</td>
<td>0.33</td>
<td>0.36</td>
<td>0.38</td>
<td>0.42</td>
<td>0.45</td>
<td>0.49</td>
<td>0.58</td>
</tr>
<tr>
<td>Steep, over 7%</td>
<td>0.37</td>
<td>0.40</td>
<td>0.42</td>
<td>0.46</td>
<td>0.49</td>
<td>0.53</td>
<td>0.60</td>
</tr>
<tr>
<td>Forest/Woodlands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flat, 0–2%</td>
<td>0.22</td>
<td>0.25</td>
<td>0.28</td>
<td>0.31</td>
<td>0.35</td>
<td>0.39</td>
<td>0.48</td>
</tr>
<tr>
<td>Average, 2–7%</td>
<td>0.31</td>
<td>0.34</td>
<td>0.36</td>
<td>0.40</td>
<td>0.43</td>
<td>0.47</td>
<td>0.56</td>
</tr>
<tr>
<td>Steep, over 7%</td>
<td>0.35</td>
<td>0.39</td>
<td>0.41</td>
<td>0.45</td>
<td>0.48</td>
<td>0.52</td>
<td>0.58</td>
</tr>
</tbody>
</table>

*Note:* The values in the table are the standards used by the City of Austin, Texas. Used with permission.

(Ref: Chow et al., 1988)
Issues with use of Empirical Methods

None of the empirical formulae will give precise results for all locations.

Magnitude of flood in a catchment depends on a large set of predictor variables, and none of the formulae consider adequate number of variables.

Predictor variables - Examples

1) Precipitation

2) Physiographic catchment characteristics
   Drainage Area; Length of longest channel

3) Location attributes
   Latitude; Longitude; Altitude

4) Landuse/landcover
   Built-up Area; Agricultural Area; Forest Area; Water bodies; Waste lands

5) Catchment shape
   Compactness coefficient; Circularity ratio; Form factor; Elongation ratio

6) Drainage Characteristics
   Drainage density; Slope; Soils
**Form factor:** Ratio of area of the basin to square of its axial length \(^{(Large)}\)

**Compactness coefficient:** Ratio of the perimeter of the basin to the circumference of a circle of area equal to the basin area. \(^{(Small)}\)

**Elongation ratio:** Ratio of the diameter of a circle having area the same as the basin area, to the maximum length of the basin. \(^{(Large)}\)

**Circularity ratio:** Ratio of the basin area to the area of a circle of same perimeter as that of the basin. \(^{(Small)}\)
Frequency Analysis Methods for Design Flood Estimation

- Information at target site is augmented with that from gauged sites depicting similar hydrological behavior
  - Data or model parameters
  - Functional relationships between model parameters and watershed attributes

**Region**: A set of gauged sites depicting similar hydrological behavior

**Regionalization**: Process of identifying regions

**Regional Frequency Analysis**: Frequency analysis based on Regional information
Approaches to Regionalization

- **Regression Approach**
  - Thomas and Benson [1970], Wandle [1977], Glatfelter [1984]
  - Choquette [1988]

- **ROI approach**
  - Acreman and Wiltshire [1987], Burn [1990 a,b], Zrinji and Burn [1994]

- **Hierarchical approach & Extension to ROI**
  - Gabriele and Arnell [1991], Zrinji and Burn [1996]

- **Classification and Regression Tree (CART) based Approach**
  - Breiman et al. [1984], Laaha and Bloschl [2006]

- **Canonical correlation analysis**
  - Cavadias [1989, 1990, 1995], Cavadias et al. [2001]
Approaches to Regionalization

- Nearest Neighbor approach
  - Vogel [2005]

- Ensemble Prediction Approach
  - McIntyre et al. [2005]

- Cluster analysis
  - Tasker [1982], Burn [1989], Burn et al. [1997], Burn and Goel [2000], Hall and Minns [1999], Hall et al. [2002], Jingyi and Hall [2004], Rao and Srinivas [2006a,b]
### Frequency Analysis Methods for Design Flood Estimation

When $T \leq 27$ years  (FEH, IH Wallingford, 1999; Vol.3, P.46)

<table>
<thead>
<tr>
<th>Length of record (years)</th>
<th>Site Analysis</th>
<th>Pooled Analysis</th>
<th>Recommended method</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt;T/2$</td>
<td>No</td>
<td>Yes</td>
<td>Pooled analysis</td>
</tr>
<tr>
<td>T/2 to T</td>
<td>For confirmation</td>
<td>Yes</td>
<td>Pooled analysis prevails</td>
</tr>
<tr>
<td>T to 2T</td>
<td>Yes</td>
<td>Yes</td>
<td>Joint (site and Pooled) Analysis</td>
</tr>
<tr>
<td>$&gt;2T$</td>
<td>Yes</td>
<td>For confirmation</td>
<td>Site analysis prevails</td>
</tr>
</tbody>
</table>
## Frequency Analysis Methods for Design Flood Estimation

When $T > 27$ years \hspace{1cm} (FEH, IH Wallingford, 1999; Vol.3, P.46)

<table>
<thead>
<tr>
<th>Length of record (years)</th>
<th>Site Analysis</th>
<th>Pooled Analysis</th>
<th>Recommended method</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;14</td>
<td>No</td>
<td>Yes</td>
<td>Pooled analysis</td>
</tr>
<tr>
<td>14 to $T$</td>
<td>For confirmation</td>
<td>Yes</td>
<td>Pooled analysis prevails</td>
</tr>
<tr>
<td>$T$ to $2T$</td>
<td>Yes</td>
<td>Yes</td>
<td>Joint (site and Pooled) Analysis</td>
</tr>
<tr>
<td>&gt; $2T$</td>
<td>Yes</td>
<td>For confirmation</td>
<td>Site analysis prevails</td>
</tr>
</tbody>
</table>
Assumptions in flood frequency analysis (FFA)

**At-site FFA:** The flood (predictand) data at the target site are
- serially independent
- identically distributed in time

**Regional FFA:** The flood data at any site in a region are
- serially independent
- identically distributed in time

The flood data at different sites in the region
- are independent in space
- have identical frequency distributions, apart from a scale factor

The causal precipitation (predictor) system is stochastic, space-independent and time-independent (Chow et al., 1988).
Rainfall-runoff method for Flood Estimation

Involves estimation of flood frequency from rainfall frequency using a hydrological model which links rainfall to resultant runoff.

The method is appealing as it yields a hydrograph of river flow rather than just an estimate of peak flow.
Unit hydrograph method for Design Flood Estimation

Involves developing regression relationships between the physical characteristics of catchments and parameters of their unit hydrographs, to arrive at a synthetic unit hydrograph for estimation of design flood.
Factors influencing the choice of Flood Estimation Approach

1. Type of Problem

The rainfall-runoff approach should generally be used for estimation of
- Reservoir flood
- Probable maximum flood
- Problems involving storage routing

2. Type of catchment

- Statistical approach should generally be used for large catchments (>1000 km²)
- Rainfall-runoff approach is favored when sub-catchments are disparate/dissimilar

3. Type of data

- Statistical approach is favoured when the gauged record is longer than 2 or 3 years
- Rainfall-runoff approach is favored when there is no continuous flow record, but rainfall and flow data are available for 5 or more flood events

(FEH, 1999)
Points to be noted with reference to case studies in India

- Several of the sub-zones delineated by CWC (1983) are heterogeneous
- Need to delineate alternative homogeneous regions
- Need to consider Partial duration Series
- Regions need not be strictly contiguous
- Regions need not be hard, and can be fuzzy
Selection of attributes

Preparing feature vectors

Forming clusters

Selecting optimum number of regions

Testing the regions for homogeneity

Are the regions Homogeneous?

Yes

Determining values of predictands

No

Adjustment of heterogeneous regions

Steps in Regional Frequency Analysis of watersheds
Selection of attributes

Preparing feature vectors

Forming clusters

Selecting optimum number of regions

Testing the regions for homogeneity

Are the regions Homogeneous?

Yes

Determining values of predictands

No

Adjustment of heterogeneous regions

Steps in Regional Frequency Analysis of watersheds
**Basin Characteristics For Regional Analysis**

- **Physiographic characteristics**
  - Drainage area, Length and slope of main stream

- **Soil characteristics**
  - Infiltration potential; Runoff coefficient, Permeability, Bulk density

- **Land use characteristics**
  - Forests, Agricultural, Suburban or urban land

- **Drainage characteristics**
  - Drainage density

- **Meteorological characteristics**
  - Storm direction, Mean annual precipitation

- **Directional statistics**
  - Seasonality of flood events
Basin Characteristics For Regional Analysis (Contd...)

- **Geographical location attributes**
  Latitude, longitude, and Altitude

- **Geologic features of the basin**
  Fraction of the basin underlain by rock formations

- **Shape indicators of catchments**
  Form factor, Compactness coefficient, Elongation ratio, Circularity ratio

**Feature vector**: vector containing basin characteristics
Steps in Regional Frequency Analysis of watersheds

1. Selection of attributes
2. Preparing feature vectors
3. Forming clusters
4. Selecting optimum number of regions
5. Testing the regions for homogeneity
   - Are the regions Homogeneous?
     - Yes: Determining values of predictands
     - No: Adjustment of heterogeneous regions
The clustering algorithms available in literature can be broadly classified as:

- **Partitional** [e.g., K-means and K-medoids algorithms, fuzzy c-means clustering]
- **Hierarchical** [e.g., single-linkage, complete-linkage and Ward’s algorithms.]
- **Density-based** [e.g., DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points To Identify the Clustering Structure)]
- **Grid-based** [e.g., STING (a STatistical INformation Grid approach) and CLIQUE (Clustering In QUEst).]
- **Model-based** and [e.g., COBWEB]
- **Boundary-detecting**. [e.g., SVC (Support Vector Clustering)]
Regionalization using Soft Clustering Techniques

- **Fuzzy Clustering**
  - Bargaoui et al., [1998], Hall and Minns [1999], Rao and Srinivas[2006]

- **Artificial Neural Network Clustering**
  - **SOFMs**
    - Hall and Minns [1999], Hall et al. [2002], Jingyi and Hall [2004]
  - **Growing Neural Gas Networks**
    - Srinivas and Tripathi [2006]
  - **Fuzzy SOFM**
    - Srinivas et al. [2008]
Fuzzy C-Means Algorithm

Minimize \[ J(U,V : X ) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^{\mu} d^2(X_k, V_i) \]

subject to the following constraints:

\[ 0 < \sum_{k=1}^{N} u_{ik} < N \quad \forall i \in \{1,\ldots,c\} \]
\[ \sum_{i=1}^{c} u_{ik} = 1 \quad \forall k \in \{1,\ldots,N\} \]

\( \mu \in [1,\infty] \) refers to weight exponent or fuzzifier

(1) Initialise fuzzy cluster centroids: \( V_1^{\text{init}}, \ldots, V_c^{\text{init}} \)

(2) Compute fuzzy memberships:

\[ u_{ik} = \frac{\left( \frac{1}{d^2(X_k, V_i^{\text{init}})} \right)^{1/(\mu-1)}}{\sum_{i=1}^{c} \left( \frac{1}{d^2(X_k, V_i^{\text{init}})} \right)^{1/(\mu-1)}} \quad \text{for} \quad 1 \leq i \leq c, \quad 1 \leq k \leq N \]
(3) Update fuzzy cluster centroids

\[ V_i = \frac{\sum_{k=1}^{N} (u_{ik})^\mu X_k}{\sum_{k=1}^{N} (u_{ik})^\mu} \quad \text{for } i = 1, 2, \ldots, c \]

(4) Update the fuzzy membership \( u_{ij} \) as :

\[ u_{ik} = \frac{\left( \frac{1}{d^2(X_k, V_i)} \right)^{1/(\mu-1)}}{\sum_{i=1}^{c} \left( \frac{1}{d^2(X_k, V_i)} \right)^{1/(\mu-1)}} \quad \text{for } 1 \leq i \leq c, \quad 1 \leq k \leq N \]

Repeat steps (3) and (4) until the change in the value of the memberships between two successive iterations becomes sufficiently small.
Conventional methods suggest hardening fuzzy partition matrix

\[
U = \begin{bmatrix}
  u_{11} & \cdots & u_{1k} & \cdots & u_{1N} \\
  \vdots & & \vdots & & \vdots \\
  u_{i1} & \cdots & u_{ik} & \cdots & u_{iN} \\
  \vdots & & \vdots & & \vdots \\
  u_{c1} & \cdots & u_{ck} & \cdots & u_{cN}
\end{bmatrix}_{c \times N}
\]

\[u_{jk} = \max_{1 \leq i \leq c} \{u_{ik}\} = 1 \quad u_{ik} = 0 \quad \forall \ i \neq j\]

\[d_{jk} = \min_{1 \leq i \leq c} \{d_{ik}\} = \min_{1 \leq i \leq c} \|v_i - x_k\| \quad \text{then} \quad u_{jk} = 1 \quad u_{ik} = 0 \quad \forall \ i \neq j\]

Instead, Rao and Srinivas (2006a) suggested using a threshold membership of \(1/c\) to form fuzzy clusters

Srinivas et al. (2008) defined threshold for each feature vector

\[T_k = \max \left\{ \frac{1}{c}, \frac{1}{2} \left[ \max_{1 \leq i \leq c} (u_{ik}) \right] \right\} \]
(1) Initialize weights \{w_{ij}, i=1,\ldots,n, j=1,\ldots,m\} of the connections from the $n$ input nodes to the $m$ output nodes, randomly.
(2) Draw an input vector $X_k$ (randomly without replacement) and compute its distance from j-th output node $W_j$. Find the winning output node as:

$$\omega = \arg \min_j \|X_k(t) - W_j(t)\|$$

for $j = 1, 2, \ldots, m$.
(3) Update the weight vectors of winner and its neighboring nodes:

\[ W_j(t+1) = W_j(t) + \eta(t) h_{j,\omega}(t)[X_k(t) - W_j(t)] \]

Repeat steps (2) and (3) over several iterations, until no changes in the Kohonen layer mapping are observed.
Decrease in topological neighborhood with increase in iterations
Neighborhood function: \( h_{j,\omega}(t) \)

\[
\eta(t) = \eta(0) \exp\left(-\frac{t}{\tau_1}\right) \quad \eta(0) \text{ has value close to 0.1}
\]

\( \tau_1 \), a constant, is set equal to the maximum number of iterations \( t_{\text{max}} \) (1000 in this study).

\[
h_{j,\omega}(t) = \exp\left(-\frac{d_{\omega,j}^2}{2\sigma^2(t)}\right)
\]

\( d_{\omega,j} \) is the topological distance between the winning node \( \omega \) and its neighboring node \( j \)

\[
d_{\omega,j} = \| r_\omega - r_j \| \quad \sigma(t) = \sigma(0) \exp\left(-\frac{t}{\tau_2}\right) \quad \tau_2 = \frac{t_{\text{max}}}{\ln \sigma(0)}
\]

Effective width of the topological neighborhood at time \( t \) : \( \sigma(t) \)

Discrete vector defining the position of the winning node : \( r_\omega \)

Discrete vector defining the position of the neighboring node \( j \) : \( r_j \)

Radius of lattice in the output layer of SOFM : \( \sigma(0) \)
Existing thumb rules to decide number of nodes in output layer of SOFM

- chose nodes to be at least two times $C_{Exp}$ (Hall and Minns, 1999)
- chose nodes to be at least three times $C_{Exp}$ (Hall et al., 2002)

$C_{Exp}$ : Expected number of clusters

**Note:** $C_{Exp}$ in a study region is not known *a priori*.

Proposed method to design output layer of SOFM

For accurate estimation of flood quantiles with non-exceedance probability upto 0.99, regions of size 20 sites are reasonable (p.59, pp.119-123, Hosking and Wallis, 1997).

$$C_{Exp} = \text{Number of sites in study region}/20 = 245/20 \approx 12$$
Case study: Indiana

Count maps for Kohonen layers obtained with classical SOFM
Case study: Ohio

The feature maps generated did not reveal any information that would allow selection of an appropriate number of clusters.

3 options considered:

1-D SOFM with number of nodes in Kohonen layer equal to the expected number of regions Srinivas et al. [2003]

Growing Neural Gas Networks Srinivas and Tripathi [2006]

Fuzzy SOFM Srinivas et al. [2008]
Fuzzy SOFM

Clusters formed at level-2

Prototypes formed at level-1

Srinivas et al. (2008)
Steps in Regional Frequency Analysis of watersheds

1. Selection of attributes
2. Preparing feature vectors
3. Forming clusters
4. Selecting optimum number of regions
5. Testing the regions for homogeneity
   - Are the regions homogeneous?
     - Yes: Determining values of predictands
     - No: Adjustment of heterogeneous regions

Selecting optimum number of regions using cluster validity measures

Partition Coefficient  (Bezdek, 1974)

$$\text{Max } V_{PC}(U) = \frac{1}{N} \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^2$$

Range: $[1/c, 1]$

In Hydrology:
Bargaoui et al., 1998;
Hall and Minns, 1999

Partition Entropy  (Bezdek, 1974)

$$\text{Min } V_{PE}(U) = - \frac{1}{N} \left[ \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik} \log(u_{ik}) \right]$$

Range: $[0, \log c]$

Fuzziness Performance Index  (Roubens, 1982)

$$\text{Min } FPI(U) = 1 - \frac{c \times V_{PC}(U) - 1}{c - 1}$$

Range: $[0, 1]$

In Hydrology:
Guler and Thine, 2004
Normalized Classification Entropy \textbf{(Roubens, 1982)}

\[
\text{Min } NCE(U) = \frac{V_{PE}(U)}{\log(c)} \quad \text{Range: } [0, 1]
\]

Extended Xie-Beni Index \textbf{(Xie and Beni, 1991)}

\[
V_{XB,m}(U,V;X) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ij})^{\mu} \|V_{i} - X_{k}\|^2}{N \min_{i,i \neq k} \|V_{i} - V_{k}\|^2}
\]

Kwon’s Index \textbf{(Kwon, 1998)}

\[
V_{XB,m}(U,V;X) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ij})^{\mu} \|V_{i} - X_{k}\|^2 + \frac{1}{c} \sum_{i=1}^{c} \|V_{i} - \bar{V}\|^2}{N \min_{i,i \neq k} \|V_{i} - V_{k}\|^2}
\]

\text{In Hydrology:}

Guler and Thine, 2004

Rao and Srinivas, 2006
Steps in Regional Frequency Analysis of watersheds

1. Selection of attributes
2. Preparing feature vectors
3. Forming clusters
4. Selecting optimum number of regions
5. Testing the regions for homogeneity
   - Are the regions homogeneous?
     - Yes: Determining values of predictands
     - No: Adjustment of heterogeneous regions

Note: The process may need to be repeated if the regions are not homogeneous.
Definition of Regional Homogeneity

- A region is said to be homogeneous if data at all sites in that region follow the same distribution.

- Degree of homogeneity is decided by comparing the between-site dispersion of the sample $L$-moment ratios for the sites in a region, with dispersion that would be expected for a homogeneous region.
Testing Regions for Homogeneity using Heterogeneity Measures of (Hosking and Wallis, 1997)

- **Heterogeneity measure (HM) based on L-coefficient of variation (L-CV):**
  \[ H_1 = \frac{(V - \mu_V)}{\sigma_V} \]

- **HM based on L-CV and L-Skewness:**
  \[ H_2 = \frac{(V_2 - \mu_{V_2})}{\sigma_{V_2}} \]

- **HM based on L-Skewness and L-Kurtosis:**
  \[ H_3 = \frac{(V_3 - \mu_{V_3})}{\sigma_{V_3}} \]
L-CV at site $i$: $t^{(i)}$

L-skewness at site $i$: $t_3^{(i)}$

L-kurtosis at site $i$: $t_4^{(i)}$

Regional average L-CV: 
\[ t^R = \frac{\sum_{i=1}^{N} n_i t^{(i)}}{\sum_{i=1}^{N} n_i} \]

Regional average L-skewness: 
\[ t_3^R = \frac{\sum_{i=1}^{N} n_i t_3^{(i)}}{\sum_{i=1}^{N} n_i} \]

Regional average L-kurtosis: 
\[ t_4^R = \frac{\sum_{i=1}^{N} n_i t_4^{(i)}}{\sum_{i=1}^{N} n_i} \]
Weighted standard deviation of at-site sample $L$-CVs:

$$V = \left\{ \frac{\sum_{i=1}^{N} n_i (t^{(i)} - t^R)^2}{\sum_{i=1}^{N} n_i} \right\}^{1/2}$$

Weighted average distance from the site $i$ to the group weighted mean in the 2-dimensional space of $L$-CV and $L$-skewness:

$$V_2 = \sum_{i=1}^{N} n_i \left\{ (t^{(i)} - t^R)^2 + (t_3^{(i)} - t_3^R)^2 \right\}^{1/2} \left/ \sum_{i=1}^{N} n_i \right.$$ 

Weighted average distance from the site $i$ to the group weighted mean in the 2-dimensional space of $L$-skewness and $L$-Kurtosis:

$$V_3 = \sum_{i=1}^{N} n_i \left\{ (t_3^{(i)} - t_3^R)^2 + (t_4^{(i)} - t_4^R)^2 \right\}^{1/2} \left/ \sum_{i=1}^{N} n_i \right.$$
\[ \mu_V : \text{Mean of } V \text{ values from Monte-Carlo (MC) simulations} \]

\[ \mu_{V_2} : \text{Mean of } V_2 \text{ values from MC simulations} \]

\[ \mu_{V_3} : \text{Mean of } V_3 \text{ values from MC simulations} \]

\[ \sigma_V : \text{Standard deviation (SD) of } V \text{ values from MC simulations} \]

\[ \sigma_{V_2} : \text{SD of } V_2 \text{ values from MC simulations} \]

\[ \sigma_{V_3} : \text{SD of } V_3 \text{ values from MC simulations} \]

\[ \text{MC simulations} \rightarrow 500 \]
HETEROGENEITY MEASURES
(Hosking and Wallis, 1997)

- Heterogeneity measure (HM) based L-CV:
  \[ H_1 = \frac{(V - \mu_V)}{\sigma_V} \]

- HM based on L-CV and L-Skewness:
  \[ H_2 = \frac{(V_2 - \mu_{V_2})}{\sigma_{V_2}} \]

- HM based on L-Skewness and L-Kurtosis:
  \[ H_3 = \frac{(V_3 - \mu_{V_3})}{\sigma_{V_3}} \]

Acceptably homogeneous: \( H < 1 \)
Possibly heterogeneous: \( 1 \leq H < 2 \)
Definitely heterogeneous: \( H \geq 2 \)
Selection of attributes

Preparing feature vectors

Forming clusters

Selecting optimum number of regions

Testing the regions for homogeneity

Are the regions Homogeneous?

Yes

Determining values of predictands

No

Adjustment of heterogeneous regions

Steps in Regional Frequency Analysis of watersheds
Adjustment of Heterogeneous Clusters

Hosking and Wallis (1997):

- ✓ eliminating one or more sites from the data set
- ✓ dividing a region to form two or more new regions

- × allowing a site to be shared by two or more regions
- × moving one or more sites from a region to other regions
- × dissolving regions by transferring their sites to other regions
- × merging a region with one or more regions
- × merging two or more regions and redefining groups
- × obtaining more data and redefining regions
Adjustment of Heterogeneous Clusters

The primary option that is considered for adjusting a heterogeneous cluster is elimination of one or more sites that are grossly discordant with respect to other sites within the cluster. The grossly discordant sites are identified using the discordancy measure given by Equation

\[ D_i = \frac{1}{3} N_k (u_i - \bar{u})^T S^{-1} (u_i - \bar{u}) \]

where \( D_i \) is discordancy measure for \( i \)th site in a cluster \( k \) having \( N_k \) sites.

\[ u_i = \begin{bmatrix} t^{(i)} & t_3^{(i)} & t_4^{(i)} \end{bmatrix}^T \]

is a vector containing \( L \)-moment ratios (LMR) of predictand at site \( i \)

\( \bar{u} \) is the unweighted group average of the LMR and \( S \) is a covariance matrix

\[ S = \sum_{i=1}^{N_k} (u_i - \bar{u})(u_i - \bar{u})^T \]

\[ \bar{u} = \frac{\sum_{i=1}^{N_k} u_i}{N_k} \]
Critical values for the discordancy statistic $D_t$ (Hosking and Wallis, 1997)

<table>
<thead>
<tr>
<th>$N_k$</th>
<th>critical value of $D_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.333</td>
</tr>
<tr>
<td>6</td>
<td>1.648</td>
</tr>
<tr>
<td>7</td>
<td>1.917</td>
</tr>
<tr>
<td>8</td>
<td>2.140</td>
</tr>
<tr>
<td>9</td>
<td>2.329</td>
</tr>
<tr>
<td>10</td>
<td>2.491</td>
</tr>
<tr>
<td>11</td>
<td>2.632</td>
</tr>
<tr>
<td>12</td>
<td>2.757</td>
</tr>
<tr>
<td>13</td>
<td>2.869</td>
</tr>
<tr>
<td>14</td>
<td>2.971</td>
</tr>
<tr>
<td>$\geq 15$</td>
<td>3.000</td>
</tr>
</tbody>
</table>
Selection of attributes

Preparing feature vectors

Forming clusters

Selecting optimum number of regions

Testing the regions for homogeneity

Are the regions Homogeneous?

Yes

Determining values of predictands

No

Adjustment of heterogeneous regions

Steps in Regional Frequency Analysis of watersheds
Growth curve estimation using index flood method

The flood quantile estimate at site $i$ for $T$-year return period is determined using index flood method (Dalrymple, 1960) by constructing a growth curve as,

$$\hat{Q}_i(T) = \hat{\mu}_i \cdot \hat{q}_k(T)$$

$\hat{\mu}_i$: Index flood value for site $i$ (e.g., mean, or median of flood data)
Probability Distributions to fit peak flows

- **Two parameter distributions**
  - Gumbel distribution (Extreme value type-1)
  - Log-Normal 2 (LN2) parameter distribution
  - Two parameter Gamma distribution

- **Three parameter distributions**
  - Generalized Pareto (GPA) distribution
  - Generalized Extreme value (GEV) distribution
  - Generalized logistic (GLO) distribution
  - Log-Normal 3-parameter (LN3) distribution; Generalized Normal (GNO) distribution
  - Three parameter Gamma distribution; Pearson type-3 (P3) distribution
  - Log- Pearson type-3 (LP3) distribution

- **Four parameter distribution**
  - Kappa distribution
  - Four-parameter Wakeby distribution

- **Five parameter distribution**
  - Wakeby distribution
Which probability distribution fits the flood data?

**Procedures for Identification**

- **Probability papers**
- **Goodness-of-fit tests**
  - Chi-Square test
  - Kolmogorov-Smirnov test
- **L-moment ratio diagram**
Probability Paper

- Flood data are plotted on probability paper to check if a frequency distribution fits the data.
- Probability papers are specifically designed for the frequency distributions.
  - One of the axes represents the values of peak flows.
  - Other axis represents exceedence or non-exceedence probability associated with peak flows, or return period.
- The plotted data appears close to a straight line if the frequency distribution fits the data.
- Flood quantile is estimated graphically by interpolation or extrapolation of the determined linear relationship between abscissa and ordinate.
Lognormal Probability Plot

Unreliability, F(t)

Time, (t)

10.0  100.0  1000.0  10000.0
## Plotting Position Formulae

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Formula</th>
<th>( b )</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>California</td>
<td>-</td>
<td>( P(X \geq x_m) = \frac{m}{n} )</td>
</tr>
<tr>
<td>2</td>
<td>Modified California</td>
<td>-</td>
<td>( P(X \geq x_m) = \frac{m-1}{n} )</td>
</tr>
<tr>
<td>3</td>
<td>Hazen</td>
<td>0.5</td>
<td>( P(X \geq x_m) = \frac{m-0.5}{n} )</td>
</tr>
<tr>
<td>4</td>
<td>Gringorten</td>
<td>0.4</td>
<td>( P(X \geq x_m) = \frac{m-0.44}{n+0.12} )</td>
</tr>
<tr>
<td>5</td>
<td>Blom</td>
<td>3/8</td>
<td>( P(X \geq x_m) = \frac{m-3/8}{n+1/4} )</td>
</tr>
<tr>
<td>6</td>
<td>Tukey</td>
<td>1/3</td>
<td>( P(X \geq x_m) = \frac{m-1/3}{n+1/3} )</td>
</tr>
<tr>
<td>7</td>
<td>Chegodayev</td>
<td>0.3</td>
<td>( P(X \geq x_m) = \frac{m-0.3}{n+0.4} )</td>
</tr>
<tr>
<td>8</td>
<td>Weibull</td>
<td>0</td>
<td>( P(X \geq x_m) = \frac{m}{n+1} )</td>
</tr>
</tbody>
</table>

\( P(X \geq x_m) = \frac{m-b}{n+1-2b} \)
### Results from study of Cunnane (1978)

<table>
<thead>
<tr>
<th>Probability distribution</th>
<th>Best plotting position formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-normal / Normal</td>
<td>Blom $P(X \geq x_m) = \frac{m - 3/8}{n + 1/4}$</td>
</tr>
<tr>
<td>EV-1 (Extreme value type-1)</td>
<td>Gringorten $P(X \geq x_m) = \frac{m - 0.44}{n + 0.12}$</td>
</tr>
<tr>
<td>Exponential</td>
<td>Gringorten $P(X \geq x_m) = \frac{m - 0.44}{n + 0.12}$</td>
</tr>
<tr>
<td>Uniform</td>
<td>Weibull $P(X \geq x_m) = \frac{m}{n + 1}$</td>
</tr>
<tr>
<td>LP-3 (Log Pearson type 3)</td>
<td>$P(X \geq x_m) = \frac{m - b}{n + 1 - 2b}$ for $C_s &gt; 0$ and $b &gt; 3/8$, $b &lt; 3/8$ for $C_s &lt; 0$</td>
</tr>
</tbody>
</table>

$$F(x) = 1 - P(X \geq x_m)$$
Goodness-of-fit tests

**Chi-square test**

The flood data at target site are divided into \( m \) class intervals.

\[
\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}
\]

- \( O_i \): The observed number of occurrences in class interval \( i \)
- \( E_i \): The expected number of occurrences in class interval \( i \)

If the class intervals are chosen such that each class interval corresponds to an equal probability, then \( E_i = n / m \), where \( n \) is the sample size of data.

\[
\chi^2 = \frac{m}{n} \sum_{i=1}^{m} O_i^2 - n
\]
The distribution of $\chi^2$ is chi-square with $\nu = m - p - 1$ degrees of freedom.

$p$ : The number of parameters of distribution that is being tested to fit the data.

\[ \chi^2 = \sum_{i=1}^{\nu} z_i^2 \]

$z_i$ : random variable among independent standard normal random variables

The hypothesis that the data are from a specified distribution is rejected at $\alpha$ significance level if following Equation is satisfied.

\[ \chi^2 > \chi^2_{\nu, 1-\alpha} \]
Kolmogorov-Smirnov Goodness-of-fit test

The Kolmogorov-Smirnov test compares the sample cumulative frequency function and theoretical cumulative distribution function (determined from the theoretical frequency distribution under test) to judge whether a chosen distribution fits the data acceptably closely.

The observed flood data at target site are divided into \( m \) class intervals, and the maximum deviation over all class intervals is determined as

\[
D = \max_{1 \leq i \leq m} \left| F \left( x_i \right) - F_s \left( x_i \right) \right|
\]

\( x_i \): The upper limit of \( i \)th class interval

The hypothesis that the data are from a specified distribution is rejected at \( \alpha \) significance level if following Equation is satisfied.

\[
D \geq D_{n,\alpha}
\]
L-Moment Ratio Diagram

L-Kurtosis vs. L-skewness graph with several curves representing different distributions: GLO, GPA, GEV, LNB, PE3, G, and E. The x-axis represents L-skewness ranging from -1.00 to 1.00, and the y-axis represents L-kurtosis ranging from -0.1 to 1.00.
Wang (1996) derived the following direct sample estimators of the L-moments

\[ \hat{\lambda}_1 = \frac{1}{nC_1} \sum_{j=1}^{n} x_{j:n} \]

\[ \hat{\lambda}_2 = \frac{1}{2} \frac{1}{nC_2} \sum_{j=1}^{n} \left( j^{-1}C_1 - n^{-j}C_1 \right) x_{j:n} \]

\[ \hat{\lambda}_3 = \frac{1}{3} \frac{1}{nC_3} \sum_{j=1}^{n} \left( j^{-1}C_2 - 2j^{-1}C_1 n^{-j}C_1 + n^{-j}C_2 \right) x_{j:n} \]

\[ \hat{\lambda}_4 = \frac{1}{4} \frac{1}{nC_4} \sum_{j=1}^{n} \left( j^{-1}C_3 - 3j^{-1}C_2 n^{-j}C_1 + 3j^{-1}C_1 n^{-j}C_2 - n^{-j}C_3 \right) x_{j:n} \]
Coefficient of L–variation \((L-CV)\), \(t = \frac{l_2}{l_1}\)

L–skewness, \(t_3 = \frac{l_3}{l_2}\)

L–kurtosis, \(t_4 = \frac{l_4}{l_2}\)

In general, \(t_r = \frac{l_r}{l_2}\)
Regional Goodness-of-fit Test

$N_k$ : Number of sites in region $k$

$n_i$ : Number of peak flow values at site $i$

$t^{(i)} t_3^{(i)} t_4^{(i)}$ : L-moment ratios computed using peak flow values at site $i$

Compute the regional average L-moment ratios

$$t^r = \frac{\sum_{i=1}^{N_k} n_i t^{(i)}}{\sum_{i=1}^{N_k} n_i}$$

$$t_3^r = \frac{\sum_{i=1}^{N_k} n_i t_3^{(i)}}{\sum_{i=1}^{N_k} n_i}$$

$$t_4^r = \frac{\sum_{i=1}^{N_k} n_i t_4^{(i)}}{\sum_{i=1}^{N_k} n_i}$$
set \( l_1^r = 1 \) by dividing flood values at each site by their mean

Fit the distribution being tested using the regional average \( L \)-moments

\[ \tau_{4}^{DIST} \] L-kurtosis of the fitted distribution

\( t_{4}^r \) L-kurtosis of the observed data

The goodness of fit is judged by the difference between \( \tau_{4}^{DIST} \) and \( t_{4}^r \)

Compare the difference with standard deviation (i.e., sampling variability) of \( t_{4}^r \) obtained by repeated simulation of the homogeneous region using kappa distribution

Regional goodness-of-fit measure

\[ Z_{DIST}^{DIST} = \frac{\left( \tau_{4}^{DIST} - t_{4}^r + B_{4} \right)}{\sigma_{4}} \]
Regional goodness-of-fit measure

\[ Z^{DIST} = \frac{(\tau^{DIST}_4 - t^r_4 + B_4)}{\sigma_4} \]

\[ B_4 = \frac{\sum_{m=1}^{N_{sim}} (t^r_4)^m - t^r_4}{N_{sim}} \theta \]

\[ \sigma_4 = \sqrt{\frac{\sum_{m=1}^{N_{sim}} (t^r_4)^m - N_{sim}B_4^2}{(N_{sim} - 1)}} \]

\[ t^r_4 = \text{Regional average } L\text{-kurtosis computed using the synthetic data simulated for } m\text{th realization} \]

Simulate a large number of realizations \((N_{sim} = 500)\) of the region using kappa distribution. Each realization constitutes a homogeneous region, with synthetic data of flood simulated at each of the sites equal to its real-world counterpart. Further, in each realization, the data simulated at any site in the region is serially independent and the data simulated at different sites in the region are not cross-correlated.
Fig. 2.5. $L$-moment ratio diagram. Two- and three-parameter distributions are shown as points and lines, respectively. Key to distributions: E – exponential, G – Gumbel, L – logistic, N – Normal, U – uniform, GPA – generalized Pareto, GEV – generalized extreme-value, GLO – generalized logistic, LN3 – lognormal, PE3 – Pearson type III. The shaded area contains the possible values of $\tau_3$ and $\tau_4$, given by (2.45).

Hosking & Wallis (1997)
Example: Generalized extreme value distribution

Probability density function

\[ f(x) = \alpha^{-1} e^{-(1-k)y - e^{-y}}, \quad \begin{cases} 
-\infty < x \leq \xi + \alpha/k, & \text{if } k > 0 \\
-\infty < x < \infty, & \text{if } k = 0 \\
\xi + \alpha/k \leq x < \infty, & \text{if } k < 0
\end{cases} \]

Cumulative distribution function

\[ F(x) = e^{-e^{-y}} \]
Shape parameter

\[ k = 7.8590c + 2.9554c^2, \quad c = \frac{2}{3 + t^r_3} - \frac{\ln 2}{\ln 3} \]

Scale parameter

\[ \alpha = \frac{l^r_2 \cdot k}{(1 - 2^{-k})\Gamma(1+k)} \]

Location parameter

\[ \xi = l^r_1 + \frac{\alpha}{k} \left[ \Gamma(1+k) - 1 \right] \]

How to compute \( \Gamma(1+k) \)?

\[ \Gamma(1+y) = \int_0^\infty t^y e^{-t} \, dt, \quad 1+y > 0 \]

Properties of Gamma function

\[ \Gamma(y+1) = y\Gamma(y), \quad \text{if} \ y > 0 \]

\[ \Gamma(y) = \Gamma(y+1)/y, \quad \text{if} \ y < 1 \]

\[ \Gamma(n) = (n-1)! \quad \text{if} \ n \text{ is a positive integer} \]

\[ \Gamma(-0.75) = \frac{\Gamma(-0.75+1)}{-0.75} = \frac{\Gamma(0.25)}{-0.75} \]

\[ = \frac{\Gamma(1+0.25)}{-0.75 \times 0.25} \]

\[ \Gamma(3.15) = \Gamma(1+2.15) = 2.15\Gamma(2.15) = 2.15\Gamma(1+1.15) = 2.15 \times 1.15\Gamma(1.15) = 2.15 \times 1.15 \Gamma(1+0.15) \]
\( \Gamma(1 + y) = 1 + b_1 y + b_2 y^2 + \ldots + b_8 y^8 + \epsilon(y) \) if \( 0 \leq y \leq 1 \)

\[
\begin{align*}
  b_1 &= -0.577191652 & b_5 &= -0.756704078 \\
  b_2 &= 0.988205891 & b_6 &= 0.482199394 \\
  b_3 &= -0.897056937 & b_7 &= -0.193527818 \\
  b_4 &= 0.918206857 & b_8 &= 0.035868343 \\
\end{align*}
\]

\[|\epsilon(y)| \leq 3 \times 10^{-7}\]
Application Examples

REGIONALIZATION OF INDIANA WATERSHEDS
# Attributes considered for regionalization

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drainage Area</td>
<td>0.28 – 28,813 km²</td>
</tr>
<tr>
<td>Mean Annual Precipitation</td>
<td>86.36 – 116.84 cm</td>
</tr>
<tr>
<td>Main channel Slope</td>
<td>0.17 – 50.57 m/km</td>
</tr>
<tr>
<td>Main channel Length</td>
<td>0.48 – 506.94 km</td>
</tr>
<tr>
<td>Basin Elevation</td>
<td>125.6 – 362.7 m</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>0% – 11%</td>
</tr>
<tr>
<td>Soil Runoff coefficient</td>
<td>0.30 – 1.00</td>
</tr>
<tr>
<td>Forest cover in Drainage Area</td>
<td>0.0 – 88.4%</td>
</tr>
<tr>
<td><strong>I(24,2)</strong></td>
<td>6.60 – 8.51 cm</td>
</tr>
</tbody>
</table>

*Storage – percentage of the contributing drainage area covered by lakes, ponds or wetlands

**I(24,2) – 24-hour rainfall having a recurrence interval of 2 years.

Source: Indiana Department of Natural Resources (IDNR), Division of Water; USGS
### Characteristics of the regions formed using Regression Approach (Glatfelter, 1984)

<table>
<thead>
<tr>
<th>Region number</th>
<th>NS</th>
<th>RS</th>
<th>Heterogeneity Measure</th>
<th>Region type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>H₁</td>
<td>H₂</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>598</td>
<td>4.85</td>
<td>1.39</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>1191</td>
<td>4.99</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>3449</td>
<td>2.09</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>1294</td>
<td>1.08</td>
<td>1.65</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>852</td>
<td>2.96</td>
<td>-0.24</td>
</tr>
<tr>
<td>6</td>
<td>32</td>
<td>913</td>
<td>4.57</td>
<td>2.96</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>901</td>
<td>10.81</td>
<td>2.31</td>
</tr>
</tbody>
</table>

**NS**: Number of stations  
**RS**: Region size in station years
Characteristics of the regions formed using Fuzzy c-means clustering (Rao and Srinivas, 2006)

<table>
<thead>
<tr>
<th>Region number</th>
<th>NS</th>
<th>RS</th>
<th>Heterogeneity Measure</th>
<th>Region type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>H₁</td>
<td>H₂</td>
</tr>
<tr>
<td>1</td>
<td>52</td>
<td>911</td>
<td>0.63</td>
<td>0.83</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>2010</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>837</td>
<td>0.23</td>
<td>0.95</td>
</tr>
<tr>
<td>4</td>
<td>75</td>
<td>3274</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>975</td>
<td>0.97</td>
<td>-0.13</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>431</td>
<td>13.51</td>
<td>5.71</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>1012</td>
<td>0.48</td>
<td>0.03</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>160</td>
<td>0.99</td>
<td>-0.24</td>
</tr>
</tbody>
</table>

NS : Number of stations
RS : Region size in station years
### Characteristics of the regions formed using Fuzzy SOFM (Srinivas et al., 2008):

<table>
<thead>
<tr>
<th>Region number</th>
<th>NS</th>
<th>RS</th>
<th>Heterogeneity Measure</th>
<th>Region type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>674</td>
<td>0.65 -0.30 -0.93</td>
<td>Homogeneous</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>1869</td>
<td>0.76 0.41 -0.37</td>
<td>Homogeneous</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>608</td>
<td>-0.32 0.93 0.31</td>
<td>Homogeneous</td>
</tr>
<tr>
<td>4</td>
<td>62</td>
<td>2820</td>
<td>0.86 0.69 -0.46</td>
<td>Homogeneous</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>1121</td>
<td>0.93 -0.11 -0.82</td>
<td>Homogeneous</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>467</td>
<td>13.77 6.05 2.41</td>
<td>Heterogeneous</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>990</td>
<td>0.82 0.42 1.27</td>
<td>Homogeneous</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>188</td>
<td>0.54 -0.69 -1.99</td>
<td>Homogeneous</td>
</tr>
</tbody>
</table>

**NS**: Number of stations  
**RS**: Region size in station years
Glatfelter Regions

Regions by Fuzzy SOFM clustering
Assessment of Predictive Performance using Leave one-out Cross Validation
Flood data and basin attributes for sites $s = 1, \ldots, N$

Take $s = 1$

Consider site $s$ as ungauged

For SOFM method compute centroids of region(s) $V_i = 1, \ldots, N_R$

For SOFM method assign site $s$ to the regions using the procedure given in section 2.6

For Glatfelter regions assign site $s$ to a region based on its geographical location

For CCA method find neighborhoods for site $s$ using procedure described in Appendix B of Ouarda et al. (2001) for chosen confidence level

Determine growth curves of region(s)/neighborhoods of $s$ using index flood method

Determine regression relationship between basin attributes and mean annual flood for region(s)/neighborhood

For SOFM method compute flood quantiles of site $s$ using Eq. (36), whereas for CCA and Glatfelter method use Eq. (35)

Is $s > N$?

No

$s = s + 1$

Yes

Compute performance measures (e.g., R-bias, R-RMSE, NMSE)
\[ R - bias = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{\hat{Q}_i^a - \hat{Q}_i^R}{Q_i^a} \right) \]

\[ R - RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \frac{\hat{Q}_i^a - \hat{Q}_i^R}{Q_i^a} \right)^2} \]

\[ NMSE = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{Q}_i^a - \hat{Q}_i^R \right)^2 \]

At-site flood quantile

Regional quantile
Leave one-out cross validation

<table>
<thead>
<tr>
<th>T</th>
<th>Two-level SOFM</th>
<th>Regional Regression</th>
<th>CCA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-bias</td>
<td>R-RMSE</td>
<td>NMSE</td>
</tr>
<tr>
<td>2</td>
<td>-0.139</td>
<td>0.618</td>
<td>0.065</td>
</tr>
<tr>
<td>10</td>
<td>-0.141</td>
<td>0.592</td>
<td>0.073</td>
</tr>
<tr>
<td>20</td>
<td>-0.145</td>
<td>0.605</td>
<td>0.080</td>
</tr>
<tr>
<td>100</td>
<td>-0.163</td>
<td>0.688</td>
<td>0.107</td>
</tr>
</tbody>
</table>

Ordinary least square regional regression equation is used to estimate index flood value for the target ungauged site
REGIONALIZATION OF Ohio WATERSHEDS
Koltun (2003) GNG Regions
Application Example

REGIONALIZATION OF WATERSHEDS
IN GODAVARI BASIN
Forming Watershed Clusters using Cluster Analysis

- **Hard clustering**
  - Mosley [1981], Tasker [1982], Burn et al. [1997] (Hierarchical)
  - Burn [1989], Burn and Goel [2000] (Partitional)
  - Rao and Srinivas [2006] (Hybrid)

- **Fuzzy Clustering**
  - Bargaoui et al., [1998], Hall and Minns [1999], Rao and Srinivas [2006]

- **Artificial Neural Network Clustering**
  - **SOFMs**
    - Hall and Minns [1999], Hall et al. [2002], Jingyi and Hall [2004]
  - **Growing Neural Gas Networks**
    - Srinivas and Tripathi [2006]
  - **Fuzzy SOFM**
    - Srinivas et al. [2008]
Study Area

Fig. Stream network and gauges in Godavari basin (Area=3,12,813 km²)
Performance Measures Considered to Compare the Regions

\[ R - bias(T) = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{\hat{q}_i^a(T) - \hat{q}_i(T)}{\hat{q}_i(T)} \right) \times 100 \]

\[ R - RMSE(T) = \left[ \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \frac{\hat{q}_i^a(T) - \hat{q}_i(T)}{\hat{q}_i(T)} \right)^2} \right] \times 100 \]

\[ NMSE(T) = \frac{\frac{1}{N} \sum_{i=1}^{N} [\hat{q}_i^a(T) - \hat{q}_i(T)]^2}{\frac{1}{(N-1)} \sum_{i=1}^{N} [\hat{q}_i^a(T) - \bar{\hat{q}}_i^a(T)]^2} \]

\( \hat{q}_i(T) \) and \( \hat{q}_i^a(T) \) are regional and at-site estimates of \( T \)-year quantile of predictand at site \( i \)

\( \bar{\hat{q}}_i^a(T) \) represents mean of \( \hat{q}_i^a(T) \) for all the sites in the study region.
Performance comparison – New Regions and CWC sub-zones in Godavari river basin.

<table>
<thead>
<tr>
<th>$T$</th>
<th>Newly delineated regions</th>
<th>Sub-zones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-bias</td>
<td>R-RMSE</td>
</tr>
<tr>
<td>5</td>
<td>-3.635</td>
<td>6.788</td>
</tr>
<tr>
<td>10</td>
<td>-5.564</td>
<td>10.653</td>
</tr>
<tr>
<td>25</td>
<td>-6.254</td>
<td>17.363</td>
</tr>
<tr>
<td>50</td>
<td>-6.205</td>
<td>22.802</td>
</tr>
<tr>
<td>100</td>
<td>-6.066</td>
<td>28.474</td>
</tr>
</tbody>
</table>
Conclusions

- Cluster analysis based on soft computing techniques is effective in determining homogeneous regions for flood frequency analysis.

- In regionalization using FCM algorithm the fuzzy memberships of watersheds in clusters are useful in adjusting the regions.

- It is not always possible to interpret clusters from the output of an SOFM, and Fuzzy SOFM and GNG network are feasible alternatives.

- The effort needed to determine regions using conventional regression analysis is considerable, and the approach is inefficient.
Publications

Journal:


Text Book:
THANK YOU