Lecture 3.6

Regional Skew

also called generalized skew

Beth Faber, PhD, PE

Flood Frequency Analysis

1

Goals

- Understand the reasons for regionalization in probability estimation, and when it is used
- Consider why it is difficult to estimate the skew coefficient in particular
- Be aware of methods for regionalizing skew, and the USGS effort to update regional skew in the US

2

Topics

- Why Regional Skew?
  - sampling error in skew estimates
  - skew coefficient bias
- Time vs Spatial Sampling Error
- Bulletin 17B/C
  - Guidelines for Conducting Regional Skew Studies
  - Derivation of weighting formula for station skew & regional skew
- Updated Methods for Estimating Regional Skew (17C)
  - Bayesian GLS Regression

3

Why Regionalize?

There are 3 main reasons for using regionalization in probability estimation:

1. We can generate probability estimates for ungaged areas that have no data for frequency analysis
2. We can improve our probability estimates by "*pooling*" data from many gages to gain the effect of a larger sample size
3. Can get more consistent parameters for a region

4

What should we Regionalize?

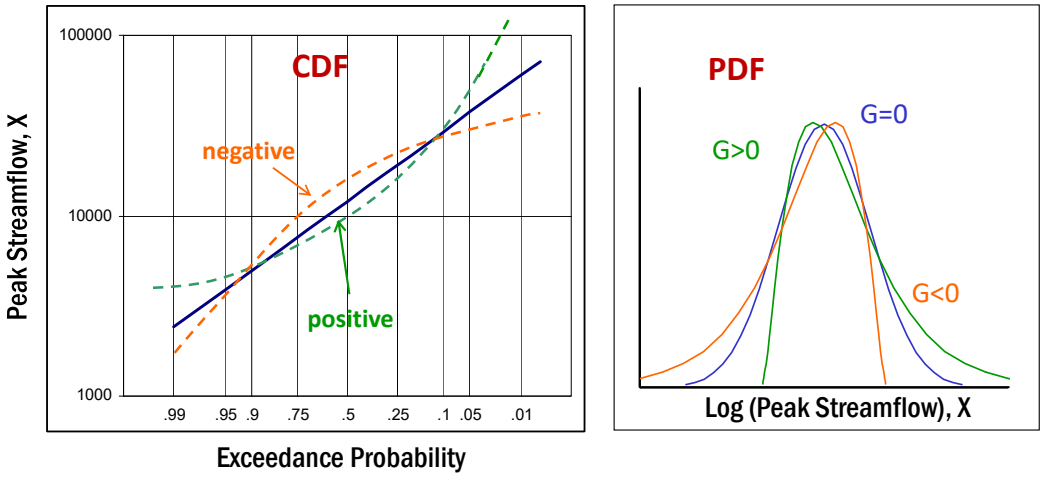
Regionalization is used for various parameters:

- Precipitation frequency for multiple quantiles and durations
  - TP-40, TP-49, NOAA Atlas 14
  - maps for pairs of exceedance probability and duration, i.e. isolines for 1% exceedance 1 hour duration rainfall depth
- Quantiles of peak flow-frequency curves
  - USGS equations: regression based on watershed characteristics
- Skew coefficient of the peak flow frequency curve
  1. dimensionless, so convenient across locations
  2. very difficult to estimate skew coefficient accurately

to make frequency-based precip events

5

Views of Skew



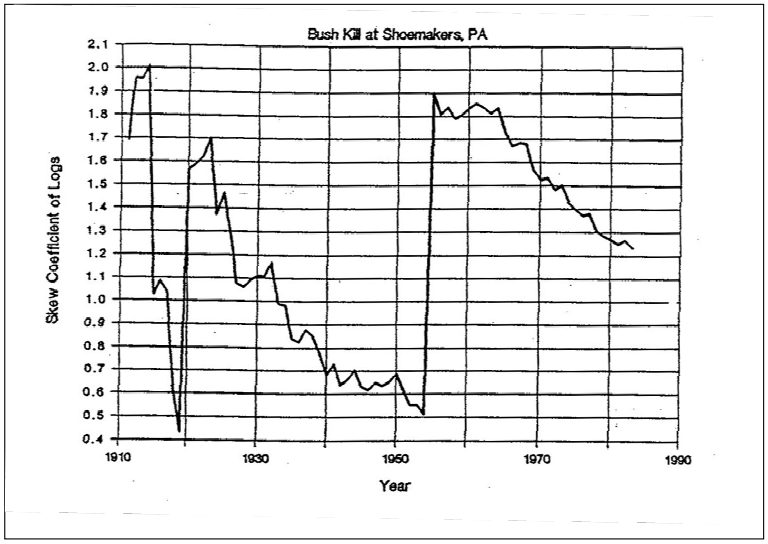
6

A zero skew is a straight line on the cumulative frequency curve, plotted on a Normal probability axis (ie, zero skew defaults the Pearson distribution to the Normal Distribution), and a symmetrical PDF.

A positive skew makes the cumulative frequency curve bend upward, and has a long upper tail in the PDF.

A negative skew makes the cumulative frequency curve bend downward, and has a long lower tail in the PDF.

Why Regional Skew?



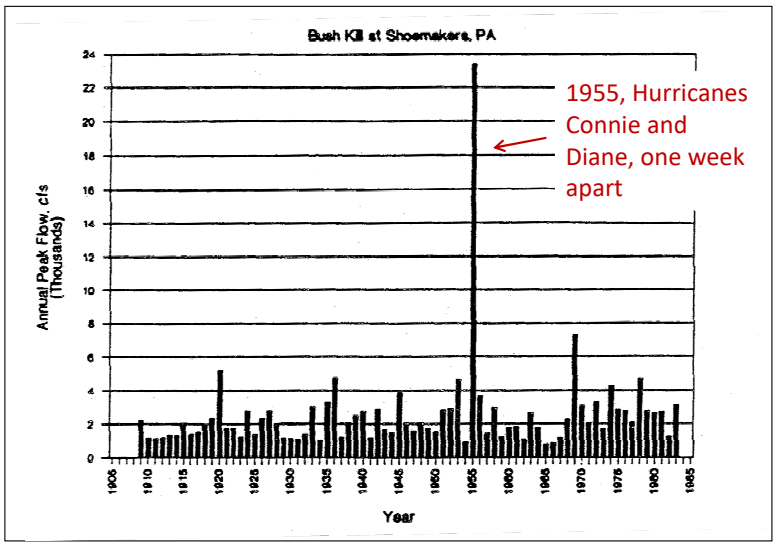
Station skew as a function of record length

7

This figure shows re-estimates of skew for Bush Kill at Shoemakers, PA as each year is added to record. The value of skew is extremely volatile.

1955, Hurricane Connie and Diane, a week apart.

Why Regional Skew?

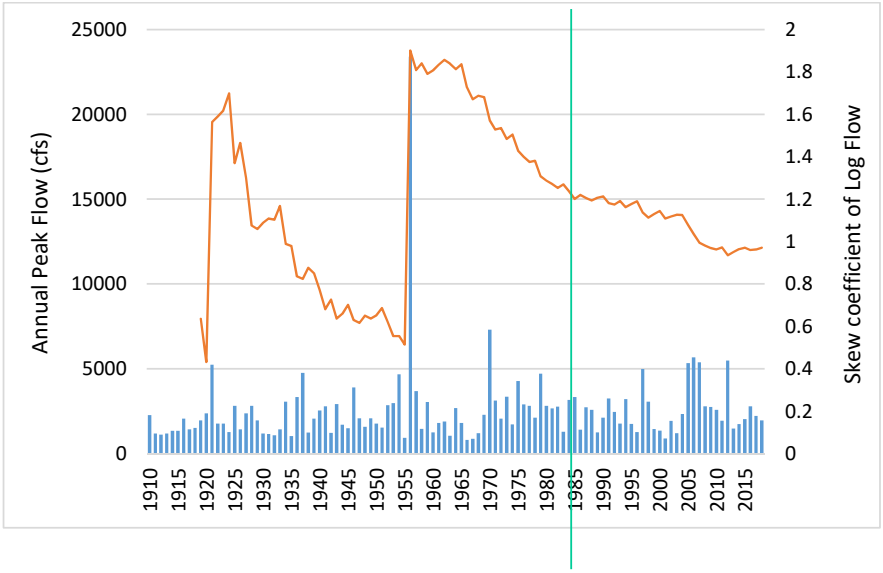


8

This is the time-serial of annual maximum flow that the previous slide is based upon.

1955, Hurricane Connie and Diane, a week apart.

Why Regional Skew? *Skew is volatile!*



9

This is the previous example with more of the recent data set added to the estimate. The green line is where the last slides ended.

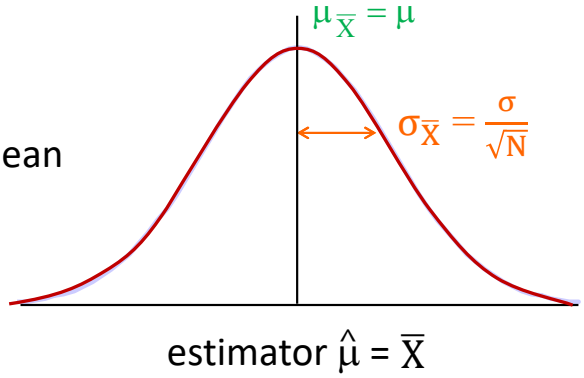
Estimators and Sampling Distributions

Remember that an estimator for the mean was given as:

$$\overline{X} = \sum_{i=1}^{N} \frac{X_i}{N}$$

By the central limit theorem, the sampling distribution of the est.mean is asymptotically Normal

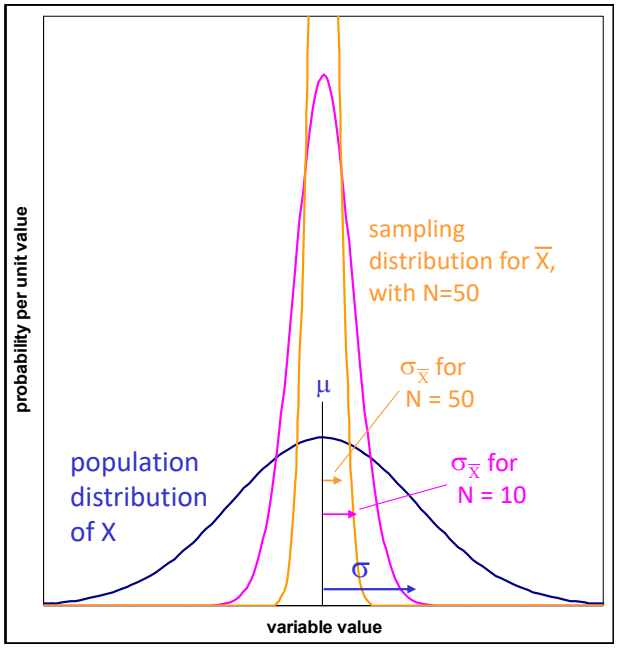
- unbiased, and variance σ^2 proportional to N



estimator $\hat{\mu} = \overline{X}$

10

Recall the sampling distribution material in the lecture on uncertainty. For the estimate of the mean, Xbar, the sampling distribution is unbiased, meaning that it's mean is the true value – the population mean mu. It's standard deviation is proportional to 1/sqrt(N), equal to the population standard deviation divided by the square root of N. The Central Limit Theorem makes this distribution Normal as sample size gets large enough.

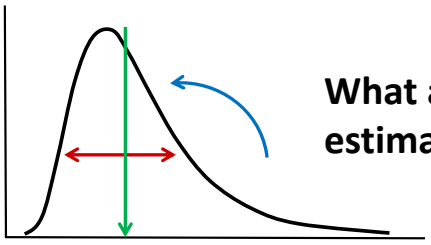


11

The population distribution is shown as the darker PDF. The sampling distribution of the estimate of the mean, X-bar, with sample size N=10 is shown as the lighter PDF (similar to the histogram on the last slide). If sigma is the standard deviation of the population distribution, then sigma divided by the square root of N is the standard deviation of the sampling distribution of X-bar. This distribution is asymptotically Normal due to the Central Limit Theorem, getting closer to Normal as N gets larger.

The lightest PDF is the sampling distribution of the mean with sample size N=50. Note the uncertainty in the estimate is much smaller with a larger sample size. The standard deviation of the population distribution, sigma, is now divided by the square root of 50, rather than the square root of 10. A smaller standard deviation of the sampling distribution means that the error in the estimate is smaller, and therefore the estimate is better.

Estimation Matrix



What am I estimating? How?

How well am I estimating it? *sampling distribution of estimator*

	mean $\bar{X}$	standard deviation S	skew g → estimator
mean (bias?)	μ	σ	biased!
standard deviation (R-MSE)	$\frac{\sigma}{\sqrt{N}}$	$\frac{\sigma}{\sqrt{2N}}$	$\propto \frac{1}{\sqrt{N}}$
skew	0	0	varies

Summary of sampling distribution parameters of distribution moments. Across the top is mean, standard deviation and skew, as parameters of the parent population. Down the side are the parameters of the sampling distributions of the estimators of the parent population parameters.

This will twist your head around!

Mean Squared Error MSE of an Estimate

To quantify how good an estimator is, we use Mean Squared Error **MSE** as a metric

- MSE is the average squared error in the estimate (similar to variance)
- reminder: σ^2 = variance
- the larger the MSE, the smaller the confidence in the estimate
- MSE is proportional to 1/N $\quad \sqrt{\text{MSE}} = \text{RMSE} \propto \frac{1}{\sqrt{N}}$

error = deviation from true value

13

Mean square error is a parallel to variance, but around the true value, rather than around the sample mean. It's a common metric for uncertainty in an estimator.

Mean Square Error of Station Skew

estimated from gage record

- We do ***numerical experiments*** to help us learn how well we can estimate skew with any sample size, N, and any actual skew, γ
- To define the mean square error (MSE) of station skew,
  - Randomly sample from a distribution with KNOWN skew
  - ESTIMATE skew from that random sample
  - compare the ESTIMATE of skew to the KNOWN skew
- How would we do a numerical experiment…?

14

As we've seen earlier, we can study how well we can estimate various parameters with a limited amount of data. We start with a known distribution, randomly sample limited data sets, and re-estimate the parameter of interest. Note how far the estimate is from the known value. Repeating this process lets us develop sampling distributions, or the uncertainty in the estimator and in any estimate. For skew coefficient, the uncertainty is based on both sample size and the actual value of the skew coefficient.

Numerical Experiment for Station Skew

- Randomly generate n=1000 samples of annual peak stream flows, each size N=50, from a distribution with KNOWN skew γ
- Calculate the sample estimate of the skew coefficient for each of the samples (j = 1,2,3….n) as:

$$g_j = \frac{N\sum_{i=1}^{i=N}(X_i - \overline{X}_j)^3}{(N-1)(N-2)S_j^3}$$

X_i = log of the i^{th} of N=50 flows for the j^{th} of n=1000 samples

$\overline{X}_j$ = sample mean for j^{th} of n samples

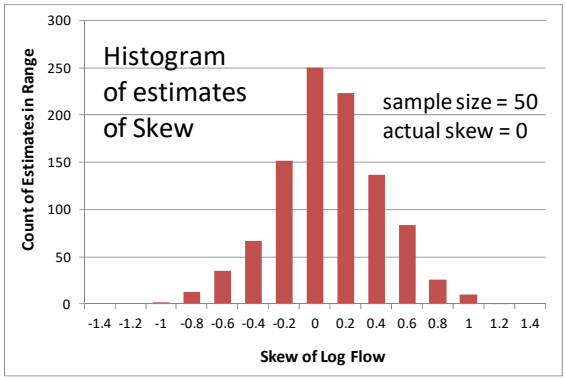
S_j = sample standard deviation for j^{th} of n samples

15

Create random samples of size 50 from a distribution with known skew, and compute the sample skew coefficient for each sample. Repeat with 1000 samples. Note, j is the sample, and Xi is a value from sample j. We'll have an Xbar, S and g for each sample j.

Sampling Distribution of Station Skew

- From the 1000 estimates of skew, one from each sample of size 50, we can create a histogram…



known skew = 0

16

The result of the experiment includes a histogram of 1000 estimates of skew coefficient, each from a sample of size 50 from a know distribution with skew = 0.

Mean Square Error of Station Skew

- Also from the sample of 1000 estimates of skew, the mean square error would then be computed as:

$$MSE_g = \frac{\sum_{j=1}^{j=n} (g_j - \gamma)^2}{n}$$

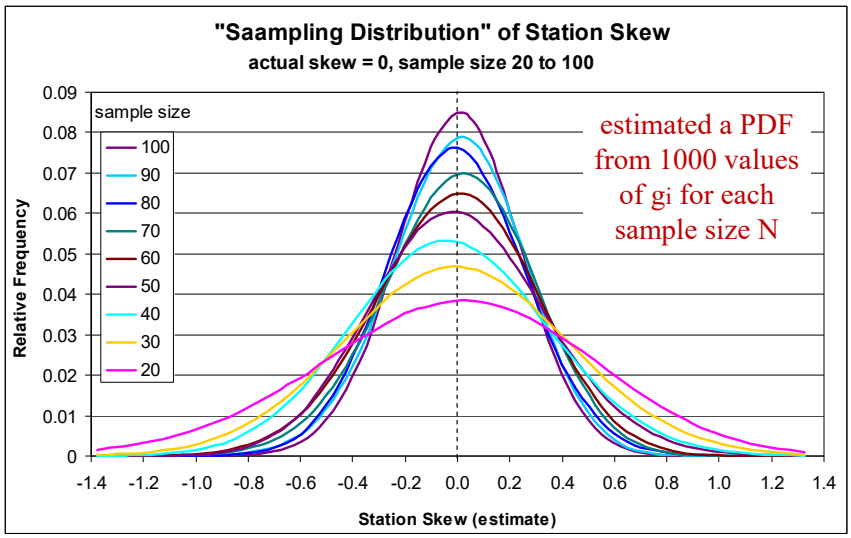
similar to variance, but around "true," not mean

- We repeat this experiment for many sample sizes N and many know skews γ
  - Determine that MSE of station skew is a function of sample size N, skew γ, and distribution f(X)
  - in our case f(X) is LPIII

17

The experiment also results in a mean square error. Each sample's gj is compared to the known skew gamma.

Sampling Distributions of Station Skew



known skew = 0

Station skew error as a function of record length and actual skew, γ

18

Sampling experiment for actual skew = 0 and an LP3 distribution. Sample sizes 20 through 100, with 1000 samples of each. Note, for actual skew = 0, the sampling distributions center at 0, are symmetrical, but are quite wide.

TABLE 1. - SUMMARY OF MEAN SQUARE ERROR OF STATION SKEW AS A FUNCTION OF RECORD LENGTH AND STATION SKEW.

STATION SKEW (G OR G)	RECORD LENGTH, IN YEARS (N OR H) 10	20	30	40	50	60	70	80	90	100
0.0	0.468	0.244	0.167	0.127	0.103	0.087	0.075	0.066	0.059	0.054
0.1	0.476	0.253	0.175	0.134	0.109	0.093	0.080	0.071	0.064	0.058
0.2	0.485	0.262	0.183	0.142	0.116	0.099	0.086	0.077	0.069	0.063
0.3	0.494	0.272	0.192	0.150	0.123	0.105	0.092	0.082	0.074	0.068
0.4	0.504	0.282	0.201	0.158	0.131	0.113	0.099	0.089	0.080	0.073
0.5	0.513	0.293	0.211	0.167	0.139	0.120	0.106	0.095	0.087	0.079
0.6	0.522	0.303	0.221	0.176	0.148	0.128	0.114	0.102	0.093	0.086
0.7	0.532	0.315	0.231	0.186	0.157	0.137	0.122	0.110	0.101	0.093
0.8	0.542	0.326	0.243	0.196	0.167	0.146	0.130	0.118	0.109	0.100
0.9	0.562	0.345	0.259	0.211	0.181	0.159	0.142	0.130	0.119	0.111
1.0	0.603	0.376	0.285	0.235	0.202	0.178	0.160	0.147	0.135	0.126
1.1	0.646	0.410	0.315	0.261	0.225	0.200	0.181	0.166	0.153	0.143
1.2	0.692	0.448	0.347	0.290	0.252	0.225	0.204	0.187	0.174	0.163
1.3	0.741	0.488	0.383	0.322	0.281	0.252	0.230	0.212	0.197	0.185
1.4	0.794	0.533	0.422	0.357	0.314	0.283	0.259	0.240	0.224	0.211
1.5	0.851	0.581	0.465	0.397	0.351	0.318	0.292	0.271	0.254	0.240
1.6	0.912	0.623	0.498	0.425	0.376	0.340	0.313	0.291	0.272	0.257
1.7	0.977	0.667	0.534	0.456	0.403	0.365	0.335	0.311	0.292	0.275
1.8	1.047	0.715	0.572	0.489	0.432	0.391	0.359	0.334	0.313	0.295
1.9	1.122	0.766	0.613	0.523	0.463	0.419	0.385	0.358	0.335	0.316
2.0	1.202	0.821	0.657	0.561	0.496	0.449	0.412	0.383	0.359	0.339
2.1	1.288	0.880	0.704	0.601	0.532	0.481	0.442	0.410	0.385	0.363
2.2	1.380	0.943	0.754	0.644	0.570	0.515	0.473	0.440	0.412	0.389
2.3	1.479	1.010	0.808	0.690	0.610	0.552	0.507	0.471	0.442	0.417
2.4	1.585	1.083	0.866	0.739	0.654	0.592	0.543	0.505	0.473	0.447
2.5	1.698	1.160	0.928	0.792	0.701	0.634	0.582	0.541	0.507	0.479
2.6	1.820	1.243	0.994	0.849	0.751	0.679	0.624	0.580	0.543	0.513
2.7	1.950	1.332	1.066	0.910	0.805	0.728	0.669	0.621	0.582	0.550
2.8	2.089	1.427	1.142	0.975	0.862	0.780	0.716	0.666	0.624	0.589
2.9	2.239	1.529	1.223	1.044	0.924	0.836	0.768	0.713	0.669	0.631
3.0	2.399	1.638	1.311	1.119	0.990	0.895	0.823	0.764	0.716	0.676

19

These is the table of station skew Mean Squared Error (MSE) from Bulletin 17B. The MSE grows with absolute value of skew, and shrinks with record length. Note, the table is the same for negative skew.

Mean Square Error of Station Skew

Mean square error of station skew is a function of record length, N, and station skew, G -- *(note, in B17B, g is stated G)*

$$MSE_G \cong 10^{[A-B[\mathrm{Log}_{10}(N/10)]]}$$

an approximation...

other approximations have been developed

where:

$$A = \begin{cases} -0.33 + 0.08\,|G| & \text{if } |G| < 0.90 \\ -0.52 + 0.30\,|G| & \text{if } |G| > 0.90 \end{cases}$$

$$B = \begin{cases} 0.94 - 0.26\,|G| & \text{if } |G| < 1.50 \\ 0.55 & \text{if } |G| > 1.50 \end{cases}$$

Tabulated for convenience in Table 1 of Bulletin 17B, pg14. (see also Wallis, J.R., Matalas, J.C., and Slack, J.R., "Just a Moment", Water Resources Research, V10(2), 1974, pg 211-219)

Consider the sampling error due to record length a ***Time Sampling Error***

20

An approximation of the MSE of station skew tabulated on the previous slide

Sampling Properties of Skew Estimate

Mean of Sample Coefficient of Skewness

$\hat{\mu}_g$

Distribution		Sample Size 10	20	50	80
Normal	$\gamma = 0$	0.00	0.00	0.00	0.00
Pearson Type III	$\gamma = 0.25$	0.15	0.19	0.23	0.23
	$\gamma = 0.50$	0.31	0.39	0.45	0.47
	$\gamma = 1.00$	0.60	0.76	0.89	0.93
	$\gamma = 2.00$	1.15	1.43	1.68	1.77
	$\gamma = 3.00$	1.59	1.97	2.37	2.54
Upper Bound on skew		3.16	4.47	7.07	8.94

if unbiased, $\mu_g = \gamma$

Source: Adapted from J. R. Wallis, N. C. Matalas, and J. R. Slack, 1974, ***Just a Moment! Appendix,*** National Technical Information Service (PB-231 816), Springfield, VA.

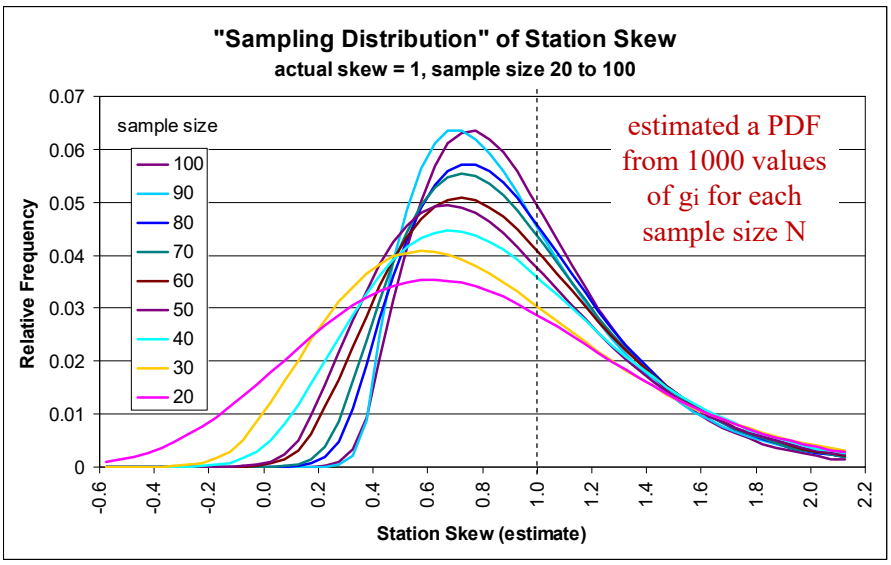
21

From "Just a Moment!" paper by Wallis, Matalas and Slack, 1974. Shows table of the mean of the sampling distribution of the estimate of skew coefficient (by product moment). If unbiased, the values would match the actual skew, but they are lower, biased toward 0. They do improve with sample size.

The bottom row shows the cause of the bias, which is an upper bound on the possible computation of skew coefficient by product moment for a given sample size.

You can demonstrate this upper bound by experimenting with skew samples to get as large a skew coefficient as possible. Try all zeros and a 1.

Underestimate of Skew due to Upper Bound



known skew = 1

22

Sampling experiment for actual skew = 1 and an LP3 distribution. Sample sizes 20 through 100, with 1000 samples of each. Note, for actual skew = 1, the sampling distributions center at are asymmetrical, and the asymmetry gets worse as the sample grows larger, though the mean value does get closer to the true value of 1.

Sampling Properties of Skew Estimate

Standard Deviation of Sample Coefficient of Skewness

similar to RMSE

$\hat{\sigma}_g$

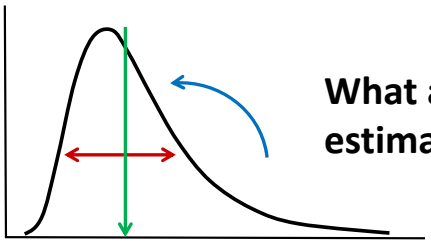
Distribution		Sample Size 10	20	50	80
Normal	$\gamma = 0$	0.69	0.51	0.34	0.26
Pearson Type III	$\gamma = 0.25$	0.69	0.52	0.35	0.28
	$\gamma = 0.50$	0.69	0.53	0.37	0.30
	$\gamma = 1.00$	0.70	0.57	0.44	0.38
	$\gamma = 2.00$	0.72	0.68	0.62	0.58
	$\gamma = 3.00$	0.74	0.76	0.78	0.77

Source: Adapted from J. R. Wallis, N. C. Matalas, and J. R. Slack, 1974, ***Just a Moment! Appendix,*** National Technical Information Service (PB-231 816), Springfield, VA.

23

From "Just a Moment!" paper by Wallis, Matalas and Slack, 1974. Shows table of the standard deviation of the sampling distribution of the estimate of skew coefficient (by product moment). The standard deviation decreases with sample size as expected until absolute value of skew gets large. For large skew coefficient, stops improving with sample size.

Estimation Matrix



What am I estimating? How?

How well am I estimating it? *sampling distribution of estimator*

	mean $\bar{X}$	standard deviation S	skew g → estimator
mean (bias?)	μ	σ	biased!
standard deviation (R-MSE)	$\frac{\sigma}{\sqrt{N}}$	$\frac{\sigma}{\sqrt{2N}}$	$\propto \frac{1}{\sqrt{N}}$
skew	0	0	varies

Summary of sampling distribution parameters of distribution moments. Across the top is mean, standard deviation and skew, as parameters of the parent population. Down the side are the parameters of the sampling distributions of the estimators of the parent population parameters.

Topics

- Why Regional Skew?
  - sampling error in skew estimates
  - skew coefficient bias
- Time vs Spatial Sampling Error
- Bulletin 17B/C
  - Guidelines for Conducting Regional Skew Studies
  - Derivation of weighting formula for station skew & regional skew
- Updated Methods for Estimating Regional Skew (17C)
  - Bayesian GLS Regression

25

Regional Skew

Basic Principle

- Can reduce sampling error by trading space for time
- Average the skew over numerous independent stations with a reasonably long record length *(NOTE: we expect skew to be similar for those stations)*
- Idea: this averaging reduces the time sampling error
  - The assumed record length is the sum of the records at all the stations considered, so large enough to "eliminate" time sampling error
- The sampling error becomes only a function of the number of stations used
  - The sampling error due to a finite number of stations can be considered a spatial sampling error

26

Regional Skew

Assumptions ...thus far

- Spatial sampling error only due to random pattern of storms in a meteorologically homogenous area
  - Consider two equivalent basins in the same region
  - A major storm passes over only one of the basins during the period of record
  - The computed skew coefficient for the two basins will be very different because of the random positioning of the storm – much larger for the station with the storm

27

Regional Skew

Assumptions (continued)

- An average skew or regional value is very effective in accounting for the random nature of storms and its effect on skew
- The greater the number of stations used to obtain average estimates of skew, the smaller the spatial sampling error due to the random nature of storm positioning
- More stations reduces that element of spatial sampling error

28

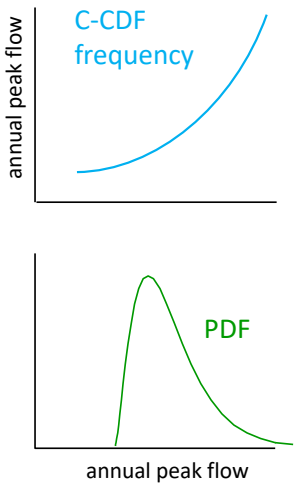
Regional Skew

Problem

- The basins are not truly equivalent
- The spatial sampling error is actually a function of both
  - the number of stations
  - differences in basin characteristics
- Spatial sampling error would exist independent of the number of stations because the skew coefficient is a function of basin shape

29

Factors Influencing Skew



Large positive skew coefficients

long upper tail, can have some very large events

physical:

1. Steep Slopes
2. Generally low basin infiltration rates
3. Fast conveyance through the system
4. A point downstream of two a confluence, with fairly coincident timing on the tributaries

statistical:

5. One or more rare events in a short record
6. Mixed populations (hurricane-nonhurricane; snowmelt-rainflood).

30

Factors Influencing Skew

Large negative skew coefficients

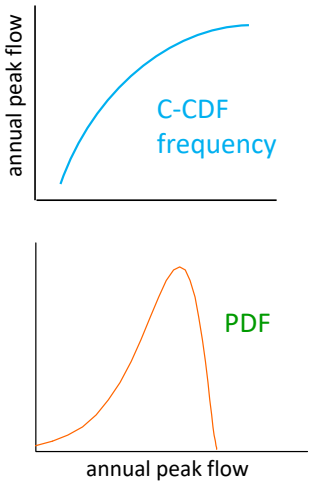
short upper tail, upper bound on large events

physical:

1. Low average basin slopes
2. Large percentage of area controlled by lakes and swamps.
3. High channel losses.
4. Generally high basin infiltration rates.

statistical:

5. One or more very low events.



31

Topics

- Why Regional Skew?
  - sampling error in skew estimates
  - skew coefficient bias
- Time vs Spatial Sampling Error
- Bulletin 17B/C
  - Guidelines for Conducting Regional Skew Studies
  - Derivation of weighting formula for station skew & regional skew
- Updated Methods for Estimating Regional Skew
  - Bayesian GLS Regression

32

Procedure for Developing Generalized Skews (Bulletin 17B)

1. Use at least 40 stations, or all stations in the surrounding region within 100 square miles. Stations should represent a good range of watershed characteristics.
2. Screen out (or adjust) the portions of the records that have been altered significantly by reservoirs, diversions, urbanization, etc.
3. Each station should have 25 years or more of natural flows.
4. Station skew should reflect the adjustments for low outliers, historic information, peaks below the gage base, and zero-flow years.

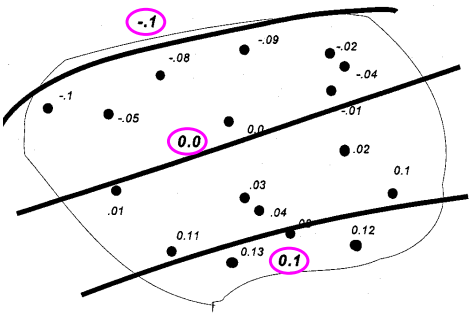
33

Bulletin 17C does not recommend doing regional skew studies, but instead using the studies being performed by USGS. But it's still important to know how studies are done, in case the USGS study is not yet done, or is not representative of the basin of interest.

Procedure for Developing Generalized Skews (Bulletin 17B)

5. Adjust station skew for bias. G*(1+6/N), or G*(1+8.5/N)

6. Method 1: Plot each station skew value at the centroid of the basin and determine if any geographic or topographic trends are present. If any trends are evident, develop an **isoline map**, and compute the MSE



Procedure for Developing Generalized Skews (Bulletin 17B)

7. Method 2: Develop a **prediction equation (regression)** that relates the station skew coefficients to watershed and climatic variables, and compute the MSE.
8. Method 3: Compute the **arithmetic mean** of station skew of at least 20 stations, if possible, in an area of reasonably homogeneous hydrology, and compute MSE.
9. Finally, adopt the results from the method that provides the most accurate estimation of the generalized skew, *i.e., the smallest mean-square error, MSE*.

35

Procedure for Developing Generalized Skews (Bulletin 17B)

Mean square error of regional skew is either:

- The average sum of squared errors from iso-lines (method 1)
- Squared standard error of regression (method 2)
- Squared standard deviation (variance) of sample skews in each area (method 3)

Note assumption behind MSE computation: station skews are CORRECT, and only error is in regionalization

36

Development of First National Skew Map

- Hardison (1974) developed first national skew map based on the following data:
  - Using 1,450 unregulated stations in conterminous United States having at least 25 years of record through 1967.
  - Drainage areas less than 1,000 mi^2 for all stations.
  - Skew coefficients multiplied by 1.26 (which is 1+8.5/N, N=33) to correct for bias.
- Beard (1974) used Hardison's skew map to show that using regional skew provided **improved estimates of T-year events** adding regional skew to Bulletin 17

37

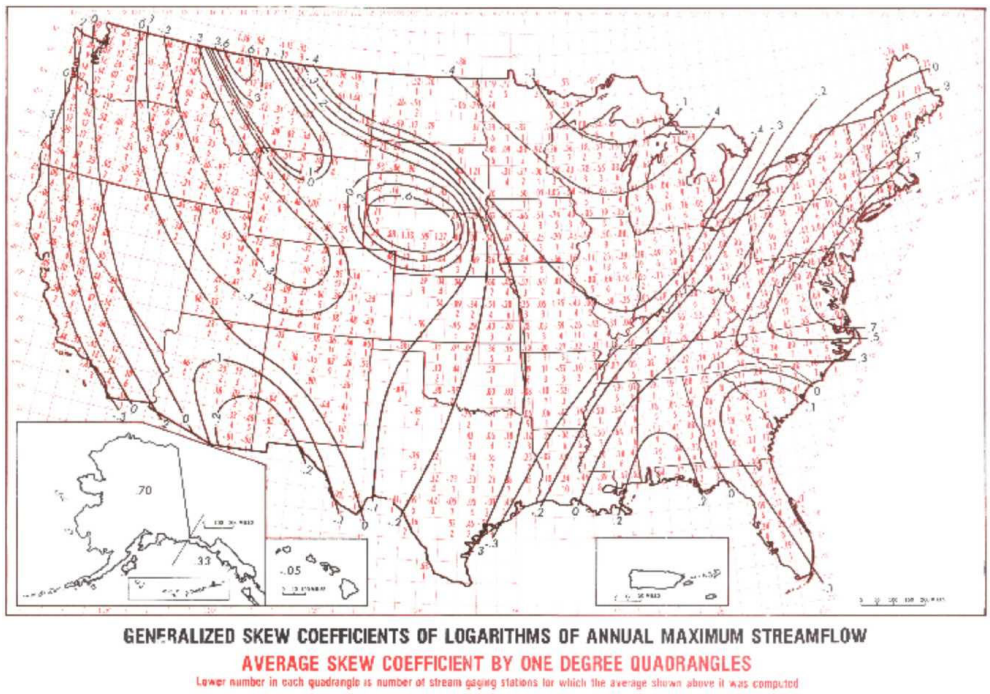
Plate I in Bulletin 17B

- Map based on 2,972 stations that had 25 or more years of record through 1973 and drainage areas less than 3,000 square miles.
- Only 144 low outliers were found (based on the old low-outlier test that was not very sensitive).
- *Historical flood information was not used*
- The national map has a mean-square error **(MSE) of 0.302**, and a root-mean-square error (RMSE) of 0.55

equivalent to 17 years of data

Question: Did Hardison correct skew coefficients for bias in developing Plate I?

38



GENERALIZED SKEW COEFFICIENTS OF LOGARITHMS OF ANNUAL MAXIMUM STREAMFLOW

AVERAGE SKEW COEFFICIENT BY ONE DEGREE QUADRANGLES

Lower number in each quadrangle is number of stream gaging stations for which the average shown above it was computed

39

Plate 1 of Bulletin 17B. See previous slide for description.

Topics

- Why Regional Skew?
  - sampling error in skew estimates
  - skew coefficient bias
- Time vs Spatial Sampling Error
- Bulletin 17B/C
  - Guidelines for Conducting Regional Skew Studies
  - Derivation of weighting formula for station skew & regional skew
- Updated Methods for Estimating Regional Skew
  - Bayesian GLS Regression

40

Weighting Equation for Regional Skew

Bulletin 17B/C suggests we weight Regional skew and Station skew to obtain an average value

$$G_w = W_r G_r + W_s G_s$$

where

W_r and W_s = Weights for averaging skew

G_r = Regional Skew

G_s = Station Skew

Determine weights such that G_w is unbiased and has minimum variance *(minimum MSE)*

41

Weighting Equation for Regional Skew

Unbiasedness

Presuming station and regional estimates are unbiased

$E[G_s] = E[G_r] = G$

where G is the population value, E[ ] is expected value

so, for $G_w = W_s G_s + W_r G_r$,

$E[G_w] = W_s E[G_s] + W_r E[G_r] = (W_s + W_r)G$ which equals G only when $(W_s + W_r) = 1$

42

If estimates are unbiased, the expected value of the estimator is the true value, G. Assumed that estimates of skew have had bias removed. (This is true for regional skew, but is not actually done for station skew.)

For the weighted average to stay unbiased, the weights must add to 1.

Weighting Equation for Regional Skew

So, for **unbiasedness**, the weights must be related as:

$(W_r + W_s) = 1$ OR $W_r = 1 - W_s$

$$\mathbf{G_w = W_s G_s + (1\text{-}W_s) G_r}$$

To **minimize the variance**, note the variance of a sum of independent random variables:

$$S^2_{G_w} = W_s^2 S^2_{G_s} + (1 - W_s)^2 S^2_{G_r}$$

where the S^2 are the sampling variances of G_w, G_s and G_r

43

When using the variance operator, constants can be pulled out but are squared.

Weighting Equation for Regional Skew

Minimizing Variance

Minimize S_G^2 by taking derivatives with respect to W_s, setting to zero, and solving for W_s:

$$S^2_{Gw} = W_s^2 S^2_{Gs} + (1-W_s)^2 S^2_{Gr}$$

$$= W_s^2 S^2_{Gs} + (1 - 2W_s + W_s^2) S^2_{Gr}$$

$$= W_s^2 S^2_{Gs} + S^2_{Gr} - 2W_s S^2_{Gr} + W_s^2 S^2_{Gr}$$

$$\frac{dS^2_{Gw}}{dW_s} = 2W_s S^2_{Gs} + 0 - 2S^2_{Gr} + 2W_s S^2_{Gr} = 0$$

$$2W_s (S^2_{Gs} + S^2_{Gr}) - 2S^2_{Gr} = 0$$

$$W_s = \frac{S^2_{G_r}}{S^2_{G_r} + S^2_{G_s}}$$

44

After multiplying out the terms, take the first derivative and set it to zero to minimize.

Solve for the value of Ws, on the right.

Weighting Equation for Regional Skew

Substituting W_s back into G_w expression:

$G_w = (1-W_s)\, G_r + W_s\, G_s$

$$G_w = \left(1 - \frac{S^2_{G_r}}{S^2_{G_r} + S^2_{G_s}}\right) G_r + \left(\frac{S^2_{G_r}}{S^2_{G_r} + S^2_{G_s}}\right) G_s$$

$$G_w = \left(\frac{S^2_{G_r} + S^2_{G_s} - S^2_{G_r}}{S^2_{G_r} + S^2_{G_s}}\right) G_r + \left(\frac{S^2_{G_r}}{S^2_{G_r} + S^2_{G_s}}\right) G_s$$

the bigger the error, the less the weight

$$G_w = \frac{S^2_{G_s} G_r + S^2_{G_r} G_s}{S^2_{G_s} + S^2_{G_r}}$$

OR

$$G_w = \frac{\frac{1}{S^2_{G_r}} G_r + \frac{1}{S^2_{G_s}} G_s}{\frac{1}{S^2_{G_r}} + \frac{1}{S^2_{G_s}}}$$

45

More algebra to multiply out the terms.

Resolve to the weight being inversely proportional to the variance of the estimator (either station skew or regional skew), such that a larger variance calls for a smaller weight. (The denominator being the sum of the variances keeps the sum of the weights to one.) The equation form on the right makes the inverse relationship more apparent.

Weighting Equation for Regional Skew

The Mean Square Error and variance are approximately the same. Substituting the Mean Square Error for the variance, the weighted skew formula of Bulletin 17B and 17C is obtained

$$G_w = \frac{\frac{1}{MSE_{G_r}} \boldsymbol{G_r} + \frac{1}{MSE_{G_s}} \boldsymbol{G_s}}{\frac{1}{MSE_{G_s}} + \frac{1}{MSE_{G_r}}} \qquad G_w = \frac{N_{G_r} \boldsymbol{G_r} + N_{G_s} \boldsymbol{G_s}}{N_{G_r} + N_{G_s}}$$

17C, using N = equivalent record length in place of MSE

This formula is therefore unbiased and minimizes variance

46

Substitute MSE for variance to get the formula used on the Bulletins.

Weighted Skew, B17C

- In the Expected Moments Algorithm, additional information (historical info, regional skew, etc) is brought in **simultaneously** with the systematic record, not after
- The skew weighting equation is not used at the end, as in 17B, but is part of the skew estimation in each iteration. ***Uses N in place of 1/MSE***

systematic censored regional

$$(7-10)\quad \hat{\gamma}_{j+1} = \frac{\sum_{i\in S}\left(\frac{X_i-\mu}{\sigma}\right)^3 + \sum_{i\in H} E\left[\left(\frac{X_i-\mu}{\sigma}\right)^3 \middle| T,\mu,\sigma,\hat{\gamma}_j\right] + N_R\hat{G}}{N_S + N_H + N_R}$$

47

Topics

- Why Regional Skew?
  - sampling error in skew estimates
  - skew coefficient bias
- Time vs Spatial Sampling Error
- Bulletin 17B/C
  - Guidelines for Conducting Regional Skew Studies
  - Derivation of weighting formula for station skew & regional skew
- Updated Methods for Estimating Regional Skew (17C)
  - Bayesian GLS Regression

48

7. Method 2: Develop a **prediction equation (regression)** that relates the computed skew coefficients to watershed and climatic variables, and compute the MSE.

Challenges to Regression

Remember some assumptions of Ordinary Least Squares (OLS) regression:

1. Variables are uncorrelated
   - In fact, many of the independent variables are cross-correlated, as are the flood records used to estimate as-site skews
2. Data points are of similar "quality," meaning error is similar for each of the data points
   - Each station skew is estimated by a different sample size, making their errors different

49

Considering the regional skew method of performing a regression to relate watershed characteristics and climatic indicators to the skew coefficient.

Other Regression for Regional Skew

OLS = Ordinary Least Squares

- *ignores errors in estimated skews and cross correlation*

WLS = Weighted Least Squares

- *recognizes differing errors, ignores cross correlation*

GLS = Generalized Least Squares

- *recognizes entire error structure*

WLS or GLS regression also give a better estimate of the accuracy of the relationship than OLS -MSE

50

These increasingly sophisticated forms of regression account for some of the ways the assumptions of simple regression are not met.

Regional Skew Study for Illinois

Tasker and Stedinger, 1986

Example Based on 62 Stations in Illinois

- Record lengths varied from 14 years to 68 years.
- Drainage areas varied from 0.08 to 9550 sq miles.
- Unbiased sample skews (multiplied by 1+6/N) were used as response (dependent) variable.
- Explanatory variables included:
  - area
  - channel slope
  - % lakes and ponds
  - forest cover
  - soils index
  - basin indicator variables

51

Regional Skew Study for Illinois

Tasker and Stedinger, 1986

Results of Analysis

Best **OLS** equation:

$G_{OLS} = - 0.545 + 0.085\ ln\ (SL) + 0.154\ ln\ (F+1) – 0.580\ (Z_2)$

MSE = 0.450 (average variance of prediction)

Best **WLS** equation:

$G_{WLS} = - 0.620 + 0.106\ ln\ (SL) + 0.145\ ln\ (F+1) – 0.470\ (Z_2)$

MSE = 0.100 (average variance of prediction)

where SL = main-channel slope, in ft/mi;
F = forest cover, in percentage of drainage area; and
Z_2 = 1, if station in Rock River basin
0, otherwise.

52

The equations with the two forms of regression are similar, but the MSE is quite a bit lower for WLS.

Regional Skew Study for Illinois

Tasker and Stedinger, 1986

Important Results

- Coefficients for OLS and GLS were similar, with same variables selected.
- Significance tests were similar but with enough difference that one might select a different "best" model
- OLS seems to have overestimated variance of prediction.

So, underweights the regional skew

53

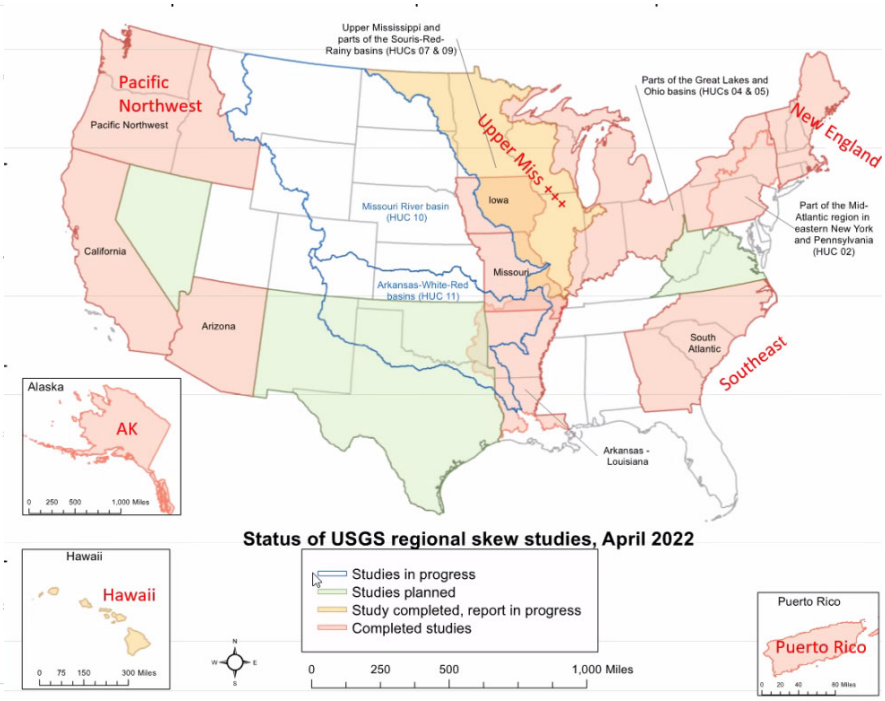
Regional Skew for B17C

- USGS is performing regional skew studies for the entire US, one region at a time
- Using Bayesian-GLS

54

Regional Skew Studies

https://acwi.gov/hydrology/Frequency/b17c/supplementary-materials/reports.html



The USGS is performing skew studies for the entire US using Baysian GLS regression. This map shows the regions for which studies are completed, ongoing, and planned.

Up-to-Date USGS Regional Skew Reports

- Flood Frequency Reports (usgs.gov)

USGS Reports: Flood Frequency and Regional Skew

Below is a list of the most recent flood frequency reports published by the USGS and organized by state. This list includes reports pertaining to regional skew, as well as regional annual exceedance probability equations for both peak flow and flood-duration flows. It will be updated as new reports are published.

As B17C recommends weighting the at-site skew with regional skew, this list provides the most current regional skew study for each state, which include both B17C recommended BGLS regional skew as well as other methodologies. If a B-GLS regional skew is not available, it is recommended that users consult with the USGS to determine the availability of alternate regional skew estimates. If no alternatives are available, then use the B17B map.

Reports which use the flood-frequency methods recommended in Bulletin 17C will be denoted with "B17C."

For additional information, please contact your local Water Science Center or email gs_b17c@usgs.gov.

ALABAMA
Peak Flow
SIR 2007-5204, Magnitude and Frequency of Floods in Alabama, 2003

ALASKA
B17C Peak Flow & B17C B-GLS Regional Skew (combined report)
SIR 2016-5024, Estimating flood magnitude and frequency at gaged and ungaged sites on streams in Alaska and conterminous basins in Canada, based on data through water year 2012

Completed 17 states

Work on 19 states underway

Overview of Methodology

Basic Model

$$\hat{\gamma}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} + \varepsilon_i$$

$$\varepsilon_i = \eta_i + \delta_i$$

site sampling error

underlying model error

Where:

- γ_i = skew at site i
- β_i = regression parameters
- X_i = basin characteristics (i.e. area, slope, precip, etc)
- δ_i = underlying basic model error (true lack of fit)
- η_i = site time sampling error

Complication: We do not know the true value of, γ $\quad \hat{\gamma}_i = \gamma_i + \eta_i$

To employ estimate, need to generate time sampling errors, η_i which are also cross-correlated.

Use a linear regression model.

Delta is the model error that comes from never being able to fit a perfect model

Do not know the true value of skew at each gage location, instead we have an estimate of skew at each gage location based on collected annual peak flow data

-Thus instead of the left side of the equation being equal to the true at-site skew (gamma), we use an estimate (gamma^hat)

-This estimate is equal to the true value of the skew at each site plus an error eta

-This sampling error eta exists because we have finite record lengths at each site from which to estimate our at-site skew

B-GLS Regional Skew Studies

STUDY	SKEW	MSE	ERL	# Sites	Avg Record Yrs
Bulletin 17B	Map	0.3	17	2972	~ 20
B-GLS					
Southeast US	0	0.14	39	342	50
California	Eqn.	0.13	57	158	57
Arizona	-0.1	0.08	85	176	54
Iowa	-0.4	0.16	50	240	52
Missouri	-0.3	0.14	58	108	58
New England	0.4	0.14	56	153	52
Arkansas+Louisiana	-0.2	0.12	59	180	58
Pacific Northwest	-0.1	0.18	41	290	67

These are the studies already done by USGS. The table's first row shows values for the plate 1 skew map in Bulletin 17B. The rest of the rows are the Bayesian GLS studies. Only California ends up using an equation (based on elevation) and the rest ended up with just the constant term of the regression.

ERLs are dependent on both MSE/AVP and regional skew value

Different ERL for different studies because their skew constant is different even though they have similar MSE/AVP

Topics

- Why Regional Skew?
  - sampling error in skew estimates
  - skew coefficient bias
- Time vs Spatial Sampling Error
- Bulletin 17B/C
  - Guidelines for Conducting Regional Skew Studies
  - Derivation of weighting formula for station skew & regional skew
- Updated Methods for Estimating Regional Skew (17C)
  - Bayesian GLS Regression

59

References

- Hydrology Subcommittee of the Interagency Advisory Committee on Water Data, 1982, Bulletin 17B, Guidelines for Determining Flood-flow Frequency: U.S. Geological Survey, Office of Water Data Coordination, Reston, Virginia.
- Hardison, Clayton, 1974, Generalized Skew Coefficients of Annual Floods in the United States, Water Resources Research, v.10, no. 4, p. 745-752.
- HEC, 1983, Generalied Skew Study for the Delaware River Basin: Special Projects Memo 83-1.
- HEC, 1977, Generalized Skew Study for State of New Jersey: Special Projects Memo 480.
- Beard, L. R., **1974**, Flood Flow Frequency Techniques: Technical Report CRWR-1198, Center for Research in Water Resources, University of Texas at Austin.
- Wallis,J.R., N.C.Matalas, and J.R.Slack, **1974**. Just a Moment!, Water Resources Research, v.10, no.2, p.211-219
- Kirby, W.H, **1974**, Algebraic Boundedness of Sample Statistics, Water Resrources Research, v.10, no.2, p.220-222.

References

- Kroll, C.N., and J.R. Stedinger, **1998**, Regional hydrologic analysis: Ordinary and generalized least squares revisited, Water Resources Research, v.34 no.1, p.121-128.
- Reis, D. S. , Jr., J.R. Stedinger, and E.S. Martins, **2005**, Bayesian GLS Regression with application to LP3 Regional Skew Estimation, accepted to *Water Resources Research,* May 2005.
- Stedinger, J.R., and G.D. Tasker, **1985**, Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared, *Water Resources Research, 21*(9), 1421-1432, 1985.
- Stedinger, J.R., and G.D. Tasker, **1986**, Regional Hydrologic Analysis, 2. Model-Error Estimators, Estimation of Sigma, and Log-Pearson Type 3 Distributions, *Water Resources Research, 22*(10), 1487-1499, 1986.

- Reis, D. S. , Jr., J.R. Stedinger, and E.S. Martins, **2005**, Bayesian GLS Regression with application to LP3 Regional Skew Estimation, accepted to *Water Resources Research,* May 2005.

References

- Stedinger, J.R., and G.D. Tasker, **1985**, Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared, *Water Resources Research, 21*(9), 1421-1432, 1985.
- Stedinger, J.R., and G.D. Tasker, **1986**, Regional Hydrologic Analysis, 2. Model-Error Estimators, Estimation of Sigma, and Log-Pearson Type 3 Distributions, *Water Resources Research, 22*(10), 1487-1499, 1986.
- Tasker, G.D., and J.R. Stedinger, **1986**, Estimating Generalized Skew With Weighted Least Squares Regression, *Journal of Water Resources Planning and Management, 112*(2), 225-237, 1986.
- Tasker, G.D., and J.R. Stedinger , **1989**, An Operational GLS Model for Hydrologic Regression, *J. of Hydrology, 111*(1-4), 361-375, 1989.
- Tasker, G.D., S.A. Hodge, and C.S. Barks, **1996**, Region of influence regression for estimating the 50-year flood at ungaged sites, Water resources Bulletin 32(1), 163-170, 1996.