FLOOD FREQUENCY ANALYSIS AT DEMING, FERNDALE, AND EVERSON

Delbert D. Franz, Linsley, Kraeger Associates, Limited
12 April 2005

EXECUTIVE SUMMARY

Flood frequencies have been estimated at Ferndale, Everson Main Street, and Deming all based on the annual flood-peak series available at Ferndale, the only reliable flood-peak series in the lower Nooksack watershed. The annual flood-peak series at Ferndale could not be used to estimate the flood frequency at Ferndale because the population is mixed. Some peaks were affected by the overflow to Canada near Everson but most were not. The annual flood-peak series at Deming could not be used for flood frequencies at Deming because the flow rating there is unreliable and there are large potential errors in peak flow. Therefore an approach unique to the requirements of the lower Nooksack River was devised. The available unsteady-flow model of the lower Nooksack River was used to predict the peak flow at Deming from the peak flow at Ferndale using calibrated data from 1990 through 2003. Given this function, the annual peak-flood series at Ferndale was transformed into an annual peak-flood series at Deming with the full effect of the overflow at Everson taken into account.

A conventional flood-frequency analysis on the computed annual-flood peak series at Deming found that the Log-Pearson Type III (LP3) distribution fitted the data adequately. The estimated 100-year flood peak at Deming was $74,500 ft^3/s$ and the 500-year flood peak was $94,900 ft^3/s$. The flood peak of record, in November of 1990 was given a return period of about 35 years.

The next stage of the analysis extended the peak flow data from the historical events by inflating the 1990 and the 1995 flood to add additional points for estimating two additional functions. The first function predicted the peak flow at Everson Main Street given the peak flow at Deming and the second predicted the peak flow at Ferndale given the peak flow at Deming. In both cases well defined empirical functions were fitted to the data.

In the final step these two functions were used to derive flood frequency curves at Ferndale and at Everson Main Street using well known methods from probability theory. Two approaches were used: one treated the prediction functions as exact and the second included the effect of the variance of an estimated residual term. The difference between the two approaches proved to be generally less than a fraction of one per cent at the flows of interest. Thus the scatter about the fitted functions had a nil effect on the conclusions.

The results at Ferndale clearly showed the effect of the overflow when predicting the higher return period flows. Not including the effect of the overflow near Everson distorts predictions of the 100-year and higher return period events. The 100-year return- period flow at Ferndale was estimated as $60,900 ft^3/s$ and the 500-year return-period flow as $70,300 ft^3/s$. The flood peak of record of November 1990 was given a return period of 50 years. No previous defensible estimates of the flood frequencies at Everson Main Street have been made. The current method gave $11,200 ft^3/s$ for the 100-year average return period flow and $25,100 ft^3/s$ for the 500-year average return period flow. The peak overflow as estimated for the November 1990 flood was given a return period of about 35 years.

The method used here produces estimates that are internally consistent, that include the effect of the overflow at Everson, and also make good use of the available data. As such, these are currently the most reliable estimates available.

FLOOD FREQUENCY ANALYSIS AT DEMING, FERNDALE, AND EVERSON

Delbert D. Franz, Linsley, Kraeger Associates, Limited
12 April 2005

Introduction

Estimates of the peak flow with return periods ranging from 50 to 500 years will be required at Deming, Ferndale, and Everson Main Street for several different applications. A hydrograph to accompany the peak flow values must also be developed but the methods used to develop a hydrograph are not covered here. We focus on making use of the available information to estimate internally consistent values for the flood peaks of a given return period at these three locations.

Flood frequency analysis on the lower Nooksack River using traditional statistical methods encounters several difficulties. These include:

1. The flow rating at Deming is known to be unstable and the conversion of a stage record to a flow record at this station has been a continuing challenge. The results of the calibration of the 1990 flood event show that the errors can be large, as much as 70 percent of flow values that had already been adjusted from the published record. The station is rated as poor with regard to data reliability, the lowest rating in the scale used by the USGS for streamflow stations. There is little hope that traditional statistical analysis of such a record can be extrapolated with any expectation of a valid result.
2. When the flows at Deming are near 45,000 ft^3/s, overflow at Everson begins and increases rapidly with increases in flow at Deming. The relative increase is on the order of four to one in the range of flows experienced at Everson Main Street in the past. That is, an increase of one per cent in the peak flow at Deming creates an increase of about four per cent in the peak flow at Everson Main Street. There are no reliable measurements of overflow at Everson Main Street. Consequently there is no record of peak flows to analyze.
3. The flow record at Ferndale is rated as good. This is the penultimate rating of the USGS. Thus its record of peak flows is markedly more reliable than the record of peaks at Deming. However, extrapolation to large events from a fitted frequency relationship may be invalid if the effect of the overflows at Everson is not included at some point. The larger peak flows are affected by overflow to Canada near Everson. Thus traditional flood frequency analysis at Ferndale fails because there are really two peak-flow populations: those with no overflow at Everson and those with an overflow at Everson.

We are in the interesting dilemma that neither the peak-flow series at Deming nor the peak-flow series at Ferndale is suitable for traditional flood frequency analysis. The record at Deming fails because it has large errors in its values, and the record at Ferndale fails because it is a mixed population. Even if the record at Deming were excellent, we would have to use non-traditional methods to obtain flood frequency estimates at Everson Main Street and at Ferndale. The source of this problem, even with excellent data at Deming, is the uncommon location of the watershed divide, at the bank of the Nooksack River near Everson, between the Fraser River drainage and the Nooksack River drainage. Consequently, large flows to Canada occur during the floods of interest. The nature and extent of the overflow surface has a strong effect on the peak flow at Ferndale. Thus extrapolation to floods with return periods of 100 years or more only makes sense if the effect of these overflows is included in the extrapolation. Both the depth of overflow and the lateral extent of the overflow will increase with increasing return period.

Even though traditional flood-frequency analysis fails, there is another characteristic of the watershed that makes flood-frequency analysis possible. That is, the principal source of the larger floods at Ferndale is from the watershed upstream of Deming. Even though the areas above and below Deming are of comparable size, the unit-area contribution from upstream of Deming is much larger. The reasons are:

1. The drainage area above Deming is at a higher elevation reaching to Mt. Baker itself. Thus the slopes are steep, the surface may be rocky or even ice and snow. The major floods are often caused by warm and moist air from tropical frontal systems being cooled as the moving front reaches the mountains near and upstream of the Deming gage.
2. The tributaries below Everson generally drain flat farm land that was once wetland. Tile drains are in frequent use to lower the local ground-water table. Consequently, not only are the precipitation amounts

smaller but the land surface moves water much more slowly than is the case upstream of the Deming gage.

The final key to developing flood frequencies at these three locations is the available calibrated unsteady-flow model of the lower Nooksack River. This model is used to develope relationships between sets of peak-flow data as outlined in the proposed methodology. Without this model no meaningful frequency analysis of the existing data would be possible.

Proposed Methodology

The following methodology makes it possible to develop estimates of return periods of peak flows at these three locations that are consistent and make good use of the existing data:

1. Develop a relationship between flood peak at Ferndale and flood peak at Deming for events selected from the 1990, 1995, 2002 and 2003 floods. These are the events for which we have calibrated and adjusted the flows at Deming to get reasonable mimicry of the hydrograph at Ferndale, stage values at Huntingdon at the International Border, and high-water marks. Selecting events above 15,000 ft^3/s at Deming yields 15 events. Six of these events have overflow simulated at Everson. All of these events have been calibrated to high-water marks and to the observed hydrograph at Ferndale. Use these 15 events to estimate a relationship between the peak flow at Deming and the peak flow at Ferndale. That is, we will be predicting the peak flow at Deming given the peak flow at Ferndale.

2. Use the relationship from step 1, to predict the peak flow at Deming for each peak flow in the annual flood-peak series at Ferndale. This yields an annual peak-flow series at Deming that is consistent with the simulation of the system using the lower Nooksack River unsteady-flow model. The peak flows estimated for Deming will then include the effect of any overflow at Everson. This annual peak-flow series is suitable for frequency analysis and is used in place of the observed annual peak-flows which are subject to large uncertainties.

3. Do a conventional frequency analysis on the annual peak-flow series at Deming developed in step 2. This series is homogeneous because the effect of the overflows at Everson is included in the relationship from step 1. The results of this analysis can be extrapolated to the 100-year and greater return periods with an expectation of consistent results.

4. We will need peak flows at Deming of 100,000 ft^3/s to reach values close to the expected range of a 500-year event. The maximum peak flow available is only about 63,000 ft^3/s for the larger of the two events in November of 1990. Thus we will inflate the 1990 flood to create peak flows in the range we need. The 1990 flood was unusual in that it had a large volume and this large volume does not appear to be typical. The flood peak in late November of 1995 is the second highest in the series we have and had a smaller volume. Thus we used inflated values for these two floods to create additional peak flows at Deming, Ferndale, and Everson Main Street so that we can extend the relationship in Step 1 to a maximum peak of near 100,000 ft^3/s at Deming. For each of these two floods we also used two different assumptions on the inflation applied to the tributaries. The first assumption was that the tributaries were left un-inflated, that is, they had the value used for each of the historical peaks. The second assumption was that the tributaries would experience the same inflation as the flows at Deming. These were thought to represent a reasonable range of possibilities for the larger floods.

5. The model representation upstream of Everson was checked at the highest flow to verify if all potential overflow surfaces were included. A survey of the bank elevations not currently included as overflow points revealed that there was a free board of about one foot minimum at all points. Thus the model as used in the calibration of the historical events was suitable for the simulation of the inflated events.

6. Develop a function that predicts the peak flow at Ferndale given the peak flow at Deming from the 15 historical peaks augmented by the results of the simulation runs in Step 4.
7. Also develop a function that predicts the peak flow at Everson Main Street given the peak flow at Deming using the historical data as well as the inflated data results that had overflows at Everson.
8. Use the functions developed in steps 6 and 7 to derive a flood-peak distribution at Ferndale and Everson Main Street from the flood-peak distribution computed at Deming in Step 3.

The available unsteady-flow simulation model, and the nature of floods in the lower Nooksack River make this "bootstrap" methodology possible. The goodness of the approach will depend on the nature of the various functions relating peak flows and on the length of record at Ferndale. Any regionalization of a flood-peak series must be done at Deming, using the predicted annual flood-peak series there.

This approach makes maximum use of the reliable peak-flow data at Ferndale. It also produces results that are consistent with our current knowledge of the overflows that take place at Everson. No other approach has been found that provides flood frequency estimates at all three locations that are internally consistent.

Creating the Basic Peak-Flow Data

The data to use in estimating the three different functions relating peak flows comes from three different sources: simulation of historical data with a given standard hydraulic geometry, simulation of the inflated 1990 floods, and simulation of inflated 1995 floods. The standard hydraulic geometry was set at 2002 which includes the new bridge at Everson and the Lagerway Dike. No flood fights and no levee failures were modeled. These conditions were held fixed for all flood events so that all would be consistent. Table 1 gives the flood peaks selected from the historical events modeled. Table 2 gives the peaks for the various inflated values of the 1990 flood, and Table 3 gives the peaks for the various inflated values of the 1995 flood.

TABLE 1: Simulated Historical Flood Peaks

Flood Peak Year	Flow at Ferndale (ft^3/s)	Flow at Deming (ft^3/s)	Flow at Everson MS (ft^3/s)
1990	56,510	62,920	5,925
1990	28,417	40,177	
1990	47,815	55,476	2,736
1995	29,817	31,472	
1995	16,468	19,288	
1995	13,398	15,203	
1995	15,309	17,211	
1995	15,348	17,331	
1995	46,286	52,830	1,739
1995	—	46,610	135
2002	29,901	39,221	
2002	28,698	30,491	
2003	13,080	16,636	
2003	39,296	49,344	569
2003	38,864	48,519	401
2003	28,809	35,023	

In Table 1 the last entry for 1995 has no flow at Ferndale because there was no distinct peak even though there was a clear secondary peak at Deming. This secondary peak was included to help define the flow level at Deming that initiates flow to Everson Main Street. Also because these are simulated results and are for a standard geometry they may not be the same as calibration results for the same event.

In Tables 2 and 3 the peak flow at Deming for the same inflation factor varies slightly because the simulation software changes the time step dynamically as it computes the flows. These changes are not always the same when flows are changed. Since the tributary flows changed the time-step pattern near the peak at Deming, the peak flow there will differ slightly between the two sets of runs for each event. Consequently there is a small variation in the peak flow at Deming when one would expect them to be the same. The simulation software does not insure that the computations always match the time of peak flow but only that there will be a close approach to the peak value. Clearly the effects are negligible and pale in light of other uncertainties involved in this analysis.

TABLE 2: Flood Peaks from Simulation of Inflated 1990 Flood

Inflation Factor	Tribs Inflated	Flow at Ferndale (ft^3/s)	Flow at Deming (ft^3/s)	Flow at Everson MS (ft^3/s)
0.90	no	53,067	55,561	3,174
1.05	no	58,216	66,083	7,332
1.10	no	59,867	69,227	8,712
1.15	no	61,548	72,303	10,092
1.20	no	62,939	75,555	11,910
1.25	no	64,196	78,712	14,019
1.30	no	65,036	81,817	15,985
1.40	no	66,790	88,035	20,139
1.60	no	70,573	100,068	28,784
1.05	yes	58,307	66,118	7,357
1.10	yes	60,100	69,224	8,753
1.15	yes	61,902	72,344	10,144
1.20	yes	63,400	75,535	11,982
1.30	yes	65,606	81,819	16,036
1.40	yes	67,641	88,085	20,222
1.60	yes	71,846	100,063	28,929

TABLE 3: Flood Peaks from Simulation of Inflated 1995 Flood

Inflation Factor	Tribs Inflated	Flow at Ferndale (ft^3/s)	Flow at Deming (ft^3/s)	Flow at Everson MS (ft^3/s)
1.05	no	48,219	55,462	2,850
1.10	no	49,930	58,107	3,965
1.15	no	51,448	60,716	5,142
1.20	no	52,854	63,376	6,335
1.25	no	54,205	66,032	7,510
1.30	no	55,496	68,587	8,671
1.40	no	57,998	73,954	11,207
1.60	no	62,673	84,470	18,100
1.05	yes	48,388	55,469	2,896
1.10	yes	50,219	58,114	4,031
1.20	yes	53,360	63,386	6,464
1.30	yes	56,219	68,675	8,849
1.40	yes	58,961	73,952	11,506
1.60	yes	64,047	84,491	18,543

Functional Form Selection and Fitting Technique

We will be estimating a function from a sample of data. We do not know the true functional relationship between the various pairs of flood peaks. However, we assume the existence of the true population of flood peaks from which we have a sample. This is the typical technique used in making estimates from data using results from mathematical statistics. A population is assumed to exist that follows some probability model, that is, although there is variability, there is a regularity to this variability. In the language of the mathematics of probability, the annual flood peaks are random variables and they derive from some probability distribution. Since we are also looking at flood peaks in pairs, we assume there is some joint distribution between the members of a pair as well. We do not know the underlying probability model and we must define it as a part of the analysis of the data. Often we assume a convenient probability model and

then estimate various parameters of that model and decide if the estimated model and the data are consistent with each other.

To make this more specific, let us look at the joint flood peaks at Deming and Everson Main Street. We assume that for the true population of these flood peaks, which might be approximated closely by a large sample, say of 10,000 pairs of data, we have a relationship like

$$Q_{E_i} = h(Q_{D_i}) + u_i \tag{1.0}$$

where Q_E = peak at Everson Main Street; Q_D = peak at Deming; h = population function relating the two peaks; and u = a random variable giving a disturbance term for the relationship. The disturbance term comes about because even for an unlimited sample, that is, the assumed population, there may be no function that relates the peaks exactly. The subscript i varies over the entire sample which we assume is so large that we have the complete population in hand.

In order to estimate the function h from a finite sample of data, we need to make several assumptions:

1. We need to select the characteristics of the disturbance term. The simplest assumptions commonly used are that $E(u_i) = 0$, that is, the expected value of the disturbance term is the same for all i, and is zero; and $E(u_i u_j) = \sigma^2$ for $i = j$ and 0 otherwise. That is, the variance of the disturbance term is constant and is given by σ^2 and all covariances are 0. We further assume that the disturbance term is normally distributed.
2. We need to select a functional form for h and typically that is taken as some polynomial in the variable used as the predictor, in this example, the peak flow at Deming. However, we do not do that here. Polynomials are simple and easy to use but a small change in one part of the range of approximation will affect the polynomial at all points. That is, using a single polynomial for the entire range yields an approximation that changes globally even when only a local change is made. Instead of using a single polynomial for the entire range, we will break the range into sub-ranges, which we call panels, and we will use polynomials of first degree in each panel. The boundaries between the panels are called breakpoints and the resulting piecewise polynomial is required to be continuous at each interior breakpoint. However, the first derivative may be discontinuous at each interior breakpoint. This piecewise polynomial is called a linear spline and has a broken-line graph when plotted. The linear spline is defined in terms of basis functions that are known once the breakpoints for the linear spline are set. That is,

$$h(x) = a_1 H_{1,m}(x) + a_2 H_{2,m}(x) + \cdots + a_m H_{m,m}(x) \tag{2.0}$$

   where $a_j, j = 1, \ldots, m$ are coefficients to be estimated; and $H_{j,m}, j = 1, \ldots, m$ are the basis functions defined on the breakpoints. See Appendix A for a more extensive discussion of the linear-spline.
3. We need to select a means for fitting the linear spline to the data. For this we use techniques based on minimizing the sum of squares of the residuals between what the linear spline predicts and what is observed. Least-squares fitting or regression analysis, as it is often called, is outlined in Appendix A as well.

The set of assumptions in item 1 defines ordinary least squares (OLS). The disturbance term is homoscedastic, that is, has constant variance. This assumption, when applied to the fitting of the linear spline predicting the peak flow at Everson Main Street from the peak flow at Deming encounters a physical absurdity. A constant variance applied to the disturbance term implies that the peak flow at Everson Main Street could become negative as the flow there approaches zero. Of course the normal distribution is unbounded so that a negative flow is possible at any point in the range. However, the probability that such an outcome will appear is vanishingly small so long as the standard deviation is about 1/4 the predicted flow or less. Thus what we need for fitting the peak flow at Everson is to have the variance of the disturbance term become small as the peak flow at Everson becomes small. Ordinary least squares is then extended to generalized least squares (GLS) to include the effect of a heteroscedastic disturbance term. Appendix A and the reference given therein should be consulted for more details.

Analysis of Simulated Historical Peaks

We have 15 simulated historical peaks to provide a function that will predict the peak at Deming given a peak at Ferndale. We used a two-panel linear spline, that is one interior break point and two

exterior breakpoints, to fit these data. Figure 1 shows the data and the two-panel linear spline fitted to it using ordinary least squares (OLS). Testing the residuals did not reject the assumption of a homoscedastic disturbance term in this case. The dotted lines in the figure show the limits set by a band two standard deviations wide about the regression line.

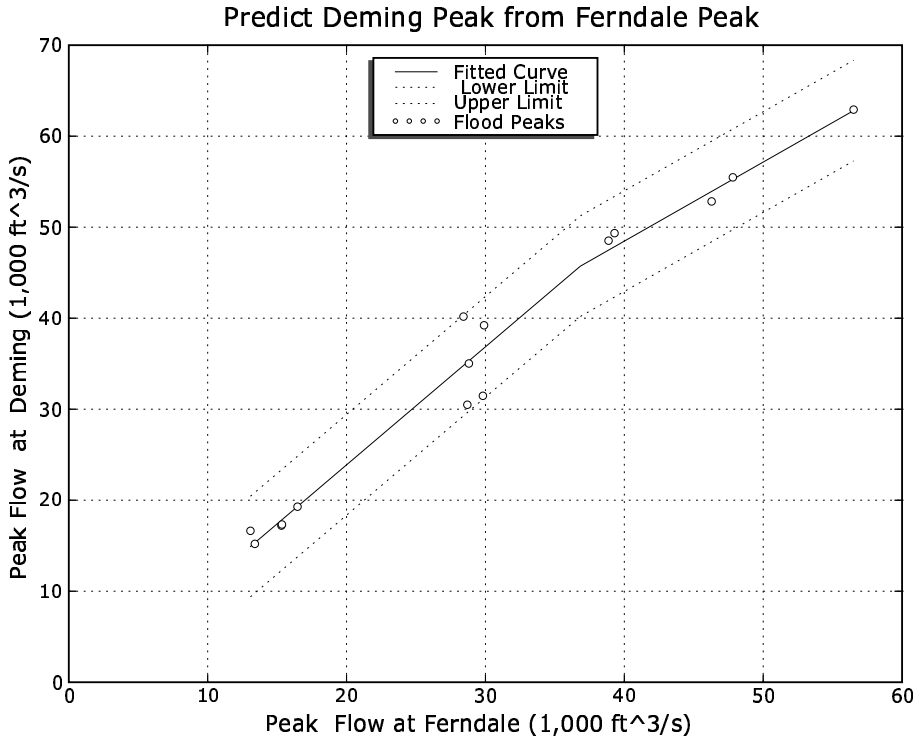

Figure 1: Predict Deming Peak given Ferndale Peak

The two-panel linear spline explains about 97 per cent of the variance in the flows at Deming. Table 4 gives the parameters of the fit. The middle breakpoint was adjusted so that the flow level predicted at Deming would match the flow level at Deming when overflow to Everson Main Street begins. This value is established below. Allowing this breakpoint to vary to improve the fit gains little. The two linear splines were barely distinguishable when plotted. Thus the middle breakpoint was set so that the flow at Deming would be at the level that initiates overflow at Everson Main Street. The thinking here was that a change in pattern would be expected near that level.

However, the change in slope differs from our initial expectation. The slope of the second segment of the linear spline is less than 1.0 and is less than the slope of the first segment of the linear spline. When overflows at Everson begin, we expect that the difference in peak flows between Deming and Ferndale would increase but the data in the range of the simulated historical peaks show otherwise. As the peak at Ferndale increases, we get a lesser rate of increase in the peak at Deming even with overflows removing water from the system near Everson. Although counter intuitive the data support this variation. There are five data points that clearly fix the slope of the second segment of the linear spline. All of these points have overflow at Everson Main Street. Thus it seems reasonable to conclude that the unexpected variation is correct.

Flow over banks and levees begins when the flow at Deming reaches about 40,000 ft^3/s. Thus some flood-plain filling has already begun before overflows start at Everson. This would seem to increase the difference between the peaks at the two stations. However, one of the scenarios tested in earlier work involved preventing all overflows at Everson. This forced more water into the flood plains. But instead of reducing peak flows down the Nooksack, this change tended to keep them more nearly the same. The levee system is not complete and there are several locations for entry and exit of water at the higher flows.

Consequently at certain levels, the flood plains seem to not attenuate the flow as much as takes place when the flow is confined to the main channel itself. It appears that the flood plains sometimes provide additional flow paths that do not increase the attenuation of the flood peak but may decrease it. When the flows get larger than any yet recorded, the effect of the overflow at Everson becomes evident and the variation between the two peaks returns to what is expected as we will see below when the extended data are analyzed.

TABLE 4: Deming given Ferndale Linear-Spline Fit

Flow at Ferndale (ft^3/s)	Flow at Deming (ft^3/s)	Student t Statistic
13080	14911	10.8
36833	45716	30.2
56510	62809	26.2

Table 4 also shows the Student t statistic for each coefficient. The theory of OLS shows that the estimates of the coefficients of the linear spline are distributed with a Student t distribution with a degree of freedom of $n - m$, where n is the sample size. Typically m is the number of parameters estimated in the functional form being fitted to the data. However, in this case, we also varied the middle breakpoint in order to match the flow at Deming when overflow begins at Everson Main Street. Thus there were really four parameters being estimated: three in the functional from and one breakpoint location. Thus the degrees of freedom in this case was $15 - 4 = 11$ and the two-sided t significance level was 2.2. In all cases the computed statistic for each coefficient of the linear spline was signifiant at this level. The results in Table 4 support the conclusion that the relationship found is valid and is a good representation of the variation inherent in the data we have available.

Estimating Annual Peak-Flows at Deming

Table 5 shows the annual peak flows at the Ferndale gaging station and those estimated at Deming from Ferndale using the relationship shown in Figure 1. Table 5 also contains estimates of the overflow at Everson Main Street based on a relationship between the peak flow at Deming and the peak flow at Everson Main Street to be developed below. The relationship between the peaks shown in Figure 1 is monotone increasing so that the rank of the floods at Deming is the same as the corresponding flood at Ferndale. This follows from the methodology and based on calibration to the larger floods, is supported by the data. Table 1 does not show any peak flows higher at Ferndale than at Deming.

A review of the available information on observed overflows at Everson Main Street verifies that overflow occurred for the years where the peak flow at Everson Main Street has an asterisk as a superscript. Eleven flows were predicted as annual peaks for Everson Main Street and nine of the eleven are supported by the scanty historical records available on overflows. The two smallest overflows shown in Table 5 may not have occurred. On the other hand there are other options:

1. The relationships used to predict peak flows at Deming and at Everson Main Street are based on flow records from 1990 through 2003. Conditions now may differ from those in 1971 but should include the event in 1997.
2. The overflows shown in Table 5 for water years 1971 and 1997 are quite small. The flow shown would not have flooded Everson Main Street because the culvert on the main tributary of Johnson Creek would convey this flow. This means that only Emmerson Road may have experienced a small flow over its surface. This would not have impeded traffic nor would there have been any damage to surrounding property. The possibility also exists that the overflow occurred late at night and being so small was not noted.
3. The conditions used for the simulations creating the flows used as a basis for the predictions were set in 2002. This includes the Lagerway Dike as well as the new bridge at Everson. Previous work has shown that the November 1990 overflow at Everson Main Street would have had a flow increase of about 2.5 per cent from these two sources. The effect of the new bridge was to reduce the overflow at Everson Main Street but the effect of the Lagerway Dike was to increase the overflow by more than the bridge

TABLE 5: Annual Flood Peaks at Ferndale, Deming, and Everson Main Street

Water Year	Peak at Ferndale (ft^3/s)	Peak at Deming (ft^3/s)	Peak at Everson MS (ft^3/s)	Water Year	Peak at Ferndale (ft^3/s)	Peak at Deming (ft^3/s)
1991	57,000	63,235	6,246*	1964	23,300	28,165
1951	55,000	61,497	5,482*	1957	23,000	27,776
1990	47,800	55,243	2,732*	2000	22,200	26,739
1996	47,200	54,722	2,503*	1960	22,000	26,479
1976	46,700	54,287	2,312*	1974	21,800	26,220
2005	42,250	50,422	707*	1995	21,700	26,090
1946	41,600	49,857	622*	1967	21,400	25,701
1984	41,500	49,770	609*	1989	21,000	25,182
2004	39,900	48,380	400*	1975	20,800	24,923
1971	38,100	46,817	165	1955	20,700	24,793
1997	38,100	46,817	165	1977	20,600	24,664
1980	36,400	45,154		2003	20,100	24,015
1987	36,000	44,636		1965	20,000	23,886
1956	35,000	43,339		1953	19,300	22,978
1983	34,200	42,301		1993	19,000	22,589
1961	30,800	37,892		1962	18,800	22,329
2002	30,300	37,244		1979	18,800	22,329
1959	30,200	37,114		1954	18,500	21,940
1986	29,900	36,724		1994	18,500	21,940
1981	29,700	36,465		1952	18,300	21,681
1969	28,100	34,390		1958	18,300	21,681
1950	27,500	33,612		1992	18,100	21,421
1982	27,200	33,223		1988	17,700	20,903
1963	26,000	31,667		1998	17,600	20,773
1972	24,800	30,111		1966	17,500	20,643
1973	24,800	30,111		1970	17,300	20,384
1999	24,600	29,851		1985	16,300	19,087
1968	23,900	28,943		2001	14,300	16,493
1978	23,900	28,943				

reduced the flow. It is therefore possible that there was no overflow for the 1971 event and these changes caused an overflow. However, other reasons would apply to the 1997 overflow shown in Table 5.

Given the data available, the results for the overflow at Everson Main Street, give strong support to the use of this methodology. Nine out of the largest eleven overflow peaks in the last 57 years of record are corroborated. This also adds credibility to the predicted peaks at Deming based on the record at Ferndale.

Frequency Analysis at Deming

The analysis of the annual peak-flow series at Deming given here follows the methods used by various Federal agencies since the adoption of the Log-Pearson Type III (LP3) distribution for flood peaks. There are 57 flood peaks in the series in Table 5. Thus let $n = 57$, then the mean value, μ_{log}, the standard deviation, σ_{log}, and the coefficient of skewness, γ_{log} are:

$$\mu_{log} = \frac{\sum_{i=1}^{n} \log(Q_i)}{n} = 4.4907 \tag{3.0}$$

$$\sigma_{log} = \sqrt{\frac{\sum_{i=1}^{n} (log(Q_i) - \mu_{log})^2}{n-1}} = 0.1498 \tag{4.0}$$

$$\gamma_{log}^{station} = \frac{n\sum_{i=0}^{n}(\log(Q_i) - \mu_{log})^3}{(n-1)(n-2)\sigma_{log}^3} = 0.4209 \tag{5.0}$$

Here Q_i is the Deming flood peak in the $i-$th year in the series. Because the coefficient of skew is strongly influenced by the station data, Bulletin 17B provides a map on Plate 1 of generalized skew that may be used to compute a weighted skew coefficient. The Plate 1 skew coefficient for both Ferndale and Deming is 0.0. This implies that the annual flood peaks in the Nooksack follow the log-normal distribution. The mean-squared error of the generalized skew from Plate 1 is given as 0.302 and the mean-square error of the station skew is computed as:

$$MSE_{stationskew} = 10^{-0.33+0.08\gamma_{log}^{station}\log([n+1]/10)} = 0.1191 \tag{6.0}$$

$$MSE_{generalizedskew} = 0.302 \tag{7.0}$$

$$\gamma_{log}^{generalized} = 0.0 \tag{8.0}$$

Bulletin 17D recommends that the station and generalized skew be weighted in inverse proportion to their estimated mean-square errors. This gives the following equation for the skew coefficient to use for the Log-Pearson Type III distribution:

$$\gamma = \frac{MSE_{generalizedskew}\gamma_{log}^{station} + MSE_{stationskew}\gamma_{log}^{generalized}}{MSE_{generalizedskew} + MSE_{stationskew}} = 0.3019 \tag{9.0}$$

We have dropped the subscripts and superscripts on the final skew coefficient as we will with the mean and standard deviation after we convert the later two from base 10 logarithms to natural logarithms. We will be integrating various functions and the natural logarithms are "natural" in that context. The skew coefficient is dimensionless and is not changed by changes in scale. Consequently the skew coefficient is the same whether we use base 10 logarithms or natural logarithms. However, the mean and standard deviation are affected by changes in scale. We then get

$$\mu = \ln(10)\mu_{log} = 10.3403 \tag{10.0}$$

$$\sigma = \ln(10)\sigma_{log} = 0.3450 \tag{11.0}$$

The parameters for the Log-Pearson Type III distribution are estimated using the method of moments:

$$\alpha = \frac{4}{\gamma^2} = 43.9008 \tag{12.0}$$

$$\beta = \frac{2}{\gamma\sigma} = 19.2044 \tag{13.0}$$

$$\xi = \mu - 2\frac{\sigma}{\gamma} = 8.0543 \tag{14.0}$$

The skew is such that this distribution has a finite lower bound and that is

$$Q_{D_{min}} = e^{\xi} = 3147.3 \tag{15.0}$$

In these terms the probability density function for the Log-Pearson Type III distribution becomes

$$f_{lnQ_D}(x) = \beta\left[[\beta(x-\xi)]^{\alpha-1}\,\frac{e^{-\beta(x-\xi)}}{\Gamma(\alpha)}\right] \tag{16.0}$$

This gives the probability density function for the natural logarithm of the flow. We will also need the probability density function in terms of the flow itself. This is

$$f_D(x) = \frac{f_{lnQ_D}(\ln(x))}{x} \tag{17.0}$$

The probability density function given by Eq. 17.0 is shown in Figure 2. The range shown there for the flow almost reaches the 3000-year return period level.

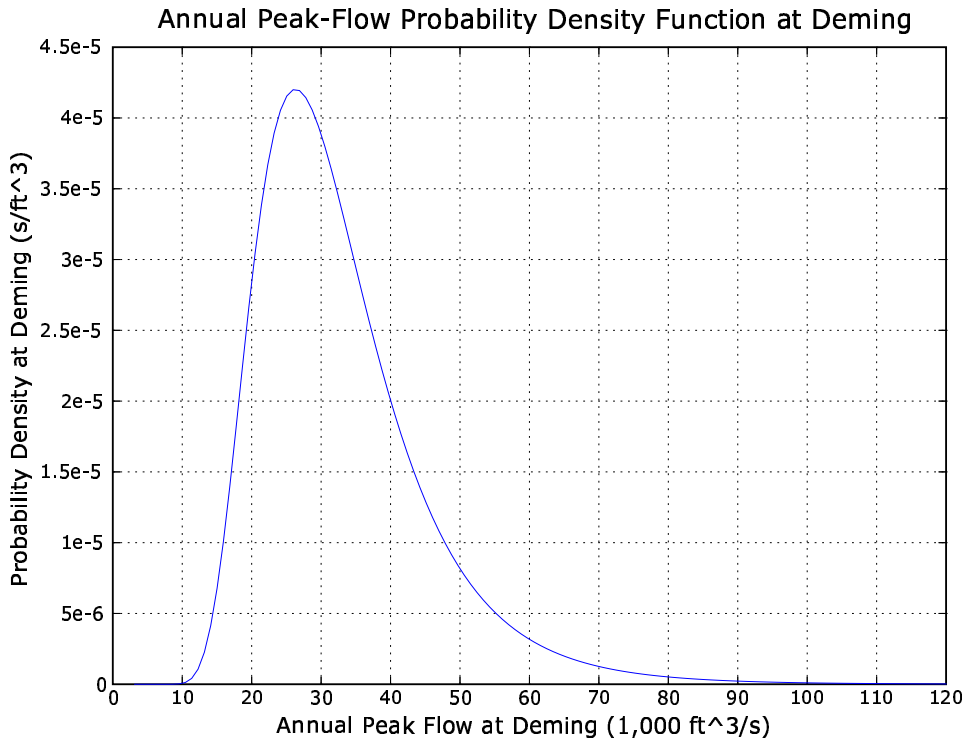

Figure 2: Annual Peak-Flow Probability Density Function at Deming

The return period for an annual peak at Deming is then given by

$$T_D(Q) = \frac{1}{1 - \int_{Q_{D_{min}}}^{Q} f_{Q_D}(x)\,dx} \tag{18.0}$$

and a simple iterative solution will yield the flow for a given return period. Table 6 gives the results for a range of return periods. The largest flood of record, in November of 1990, is shown to have a return period of about 35 years. The flood peak for an average return period of 100 years is about 74,500 ft^3/s.

The results in Table 6 show that the largest flood of record, representing at least a 57-year long time period, is given a return period of about 35 years. Part of this assignment is influenced by there being two annual flood peaks that are nearly the same. The flood peak in 1951 is only about three per cent less than the flood peak in November 1990. The flood peak in water year 1951 is given a return period of 32 years by the fitted curve, which is not far from what is expected for the second peak in a series of 57. This raises the question of the validity of the LP3 in fitting the annual-flood peak series at Deming. The chi-square goodness of fit test was, therefore, applied to the fitted LP3 distribution. Eight categories were used in the test, with each range of flows having a probability close to 0.125. This ensures that the expected value of the number of floods in each category was greater than 5. The computed value of the test statistic was 10.1 and the critical value with seven degrees of freedom and a significance level of 0.05 was 14.1. Thus this test gives no evidence to reject the LP3 as the distribution. We have estimated three parameters so the degrees of freedom could be reduced to five even though we have not met the rigorous requirements of the chi-square

TABLE 6: Frequency Analysis results at Deming

Return Period (years)	Peak Flow at Deming (ft^3/s)
1.01	14,971
1.25	23,064
2.00	30,423
4.00	38,660
5.00	41,128
10.0	48,634
15.0	53,016
20.0	56,151
25.0	58,605
35.0	62,351
50.0	66,394
75.0	71,091
100.0	74,497
150.0	79,409
200.0	82,979
300.0	88,138
400.0	91,893
500.0	94,861

test in this case. This gives a critical value of 11.1 and again gives no support to rejecting the hypothesis that the LP3 fits the data.

The Kolmogorov-Smirnov test for goodness of fit has a critical value of 0.18 for a sample of size 57. The maximum difference between the sample CDF and the CDF from the LP3 fit was 0.11. Again this test gives no support to rejecting the LP3 fit to the data as estimated at Deming. Thus this distribution is taken as the appropriate probability model to use in describing the distribution of annual peak floods at Deming.

Even though the goodness of fit tests do not support rejecting the LP3, we need some estimates of the variability in results that we could expect with a record length of 57. We used a method that falls under the broad category of bootstrap methods as outlined by Efron and Tibshirani(1993). These are computer intensive methods and are only possible because the cost of computation has plummeted in the last thirty years. The computational effort of bootstrap methods is large but then that is what computers are for! In this approach the estimated probability model, in our case, the particular LP3 distribution, is used as a surrogate for the unknown probability model. We then draw samples of size 57 from the LP3 distribution and treat these samples as we would the original sample of the peak flows at Deming. However, we can now draw thousands of samples, each being different than any other, and analyse them to estimate the variation we can expect for the annual-flood peak series at Deming.

The following steps outline the procedure:

1. Use the estimated probability model as a surrogate for the unknown true probability model. The true probability model we denote by P. The surrogate probability model is then denoted by $\hat{P}$. We then resample from $\hat{P}$. That is, we select samples of size 57 from the fitted LP3, and use these samples to develop the limits that we want. The original sample is denoted by S and the samples we take from the fitted LP3 are denoted by $\hat{S}$. A particular sample is then given as $\hat{S}_i$ for the i-th sample.
2. For each $\hat{S}_i$:
   - 2.1 Compute sample statistics and fit a LP3 distribution to it but do not apply the generalized skew correction. That was already done for S and needs only be done once since its effect is implicit in $\hat{P}$. That correction was designed to improve the station estimate of skew to better approximate the skew for the population. Applying it twice would bias the results. Store each sample's parameters for later analysis.
   - 2.2 Estimate the flows for a range of return periods and store them for later analysis.

3. When all samples are complete, sort the flows for each return period from largest to smallest. For example, if we draw 10,000 samples we will have 10,000 estimates for each return period. Any one of these samples could occur given the probability model, $\hat{P}$. Do the same for the parameters stored in step 2.1
4. Compute 0.95 limits for return periods stored in step 2.2. If we have a sample size of 10,000, then the upper limit is the $0.025 \times 10,000 = 250$-th from the top in each return period's sorted list. The lower limit is than given by the 250-th from the end of the list for each return period. Do the same for the parameters from each sample.

Table 7 gives the results for the various return periods when 100,000 samples were drawn from $\hat{P}$. We also give the mean, standard deviation, and coefficient of variation at each return period. The table shows the large variability that is inherent in flood frequency estimates from 57 years of record. The upper confidence limit for the 100-year flood is about 34 per cent above the mean value. Thus the flow for the 100-year flood in Table 6 has only about one significant digit at most!

TABLE 7: Variability of Return Periods at Deming

Return Period (Years)	Mean Value (ft^3/s)	Standard Deviation (ft^3/s)	Coefficient of Variation	Lower Limit (ft^3/s)	Upper Limit (ft^3/s)
1.01	15,027	1,576	0.1049	12,186	18,268
1.10	19,916	1,115	0.0560	17,850	22,216
1.25	23,164	1,143	0.0493	21,018	25,497
1.50	26,420	1,284	0.0486	23,983	29,057
1.75	28,703	1,409	0.0491	26,037	31,594
2.00	30,502	1,516	0.0497	27,622	33,613
3.00	35,429	1,856	0.0524	31,944	39,234
4.00	38,672	2,140	0.0553	34,669	43,057
5.00	41,116	2,400	0.0584	36,653	46,090
7.00	44,742	2,874	0.0642	39,481	50,699
10.00	48,560	3,497	0.0720	42,286	55,833
15.00	52,920	4,370	0.0826	45,121	62,201
20.00	56,048	5,099	0.0910	47,079	66,940
25.00	58,503	5,731	0.0980	48,442	70,805
30.00	60,531	6,291	0.1039	49,584	74,142
35.00	62,263	6,795	0.1091	50,568	77,063
40.00	63,778	7,256	0.1138	51,399	79,575
45.00	65,125	7,681	0.1179	52,104	81,865
50.00	66,340	8,076	0.1217	52,737	84,185
60.00	68,464	8,794	0.1285	53,824	88,178
70.00	70,283	9,436	0.1343	54,726	91,554
80.00	71,876	10,019	0.1394	55,526	94,686
90.00	73,297	10,554	0.1440	56,196	97,414
100.00	74,579	11,049	0.1482	56,864	99,839
150.00	79,622	13,110	0.1647	59,080	109,919
200.00	83,313	14,730	0.1768	60,581	117,419
250.00	86,245	16,082	0.1865	61,834	123,552
300.00	88,687	17,252	0.1945	62,811	128,945
350.00	90,786	18,289	0.2014	63,564	133,640
400.00	92,629	19,223	0.2075	64,323	138,027
500.00	95,766	20,864	0.2179	65,337	145,514

Table 7a gives the various results for parameters of the samples drawn from $\hat{P}$. The first column in this table gives the result from the samples. The minimum statistic, the lower 0.95 confidence limit of the

statistic, the average value of the statistic, the upper 0.95 confidence limit, and the maximum are shown in the table. Six columns follow with the first three giving the results for the mean, standard deviation, and skew coefficient for the flows, and the last three giving these same summaries for the natural logarithm of the flows. The later are used in the fitting process for the LP3 distribution. This table again shows the high variability in sampling especially for the standard deviation and skew coefficient. The skew coefficient strongly influences the shape of the Pearson Type 3 PDF and the range shown includes shapes ranging from the hyper-exponential at a value of about 1.8, to exponential for a value close to 1.0, a unimodal shape skewed to the right for positive skews less than 1.0, a near normal PDF for skews close to zero, then to PDF's with an upper limit and skewed to the left for the larger negative skews. Limits on the magnitude of numbers in the computer required that the samples be defined as log-normal whenever the skew coefficient of the natural logarithm of flow was less than 0.07 in absolute value. About 13 per cent of the samples came from the log-normal distribution. About 19 per cent of the samples had a negative skew coefficient for the natural logarithm of flow.

TABLE 7a: Variability of Sample Statistics at Deming

Result from Samples	Mean of Q (ft^3/s)	Std. Dev. of Q (ft^3/s)	Skew Coef. of Q	Mean of $ln(Q)$ (ft^3/s)	Std. Dev. of $ln(Q)$ (ft^3/s)	Skew Coef. of $ln(Q)$
Minimum	27,680	6,666	-0.055	10.18	0.2271	-0.9115
0.95 Lower	29,864	8,887	0.383	10.25	0.2795	-0.3245
Average	32,932	12,239	1.304	10.34	0.3435	0.2776
0.95 Upper	36,371	16,960	2.939	10.43	0.4114	0.9709
Maximum	39,520	28,736	5.774	10.52	0.4638	1.7956

Figure 3 shows the fitted LP3 distribution as a straight line with the horizontal scale transformed accordingly. The upper and lower limits are shown together with the observed annual flood peaks plotted using the following plotting position

$$\hat{p}_j = \frac{j-a}{n+1-2a} \tag{20.0}$$

with $a = 0.4$, the value recommended by Cunanne(1978), yielding approximately quantile unbiased positions. A quantile is the name given to a flow exceeded with a given probability. A quantile-unbiased plotting position places a ranked observation at the probability of exceedance of the expected value at that rank. To make this a bit clearer, imagine that we draw repeated samples from a defined population distribution like we did above to estimate limits on return period, but now we keep 57 different lists of results. The first list contains the largest flow from each sample, the second list contains the second-largest flows, and on to the last list which contains the smallest flow out of each sample. Each list contains then a sample of an order statistic. On order statistic is just the value in a sample at a given rank from largest to smallest. It is a statistic because it depends on the sample, for example, each sample will have a different maximum flow value.

Let $X_{(j)}$ denote the $j-$th order statistic out of a sample of n and then $X_{(1)}$ denotes the random variable that represents the largest value from a sample. This differs from some statistics texts but is in agreement with the usage of rank in flood frequency analysis where rank 1 is the largest flood in the sample. Let $f_{X_{(j)}}$ be the probability density function of the $j-$th order statistic in the sample; f_D be the probability density function (PDF) of the LP3 distribution fitted at Deming; and F_D be the cumulative distribution function (CDF) of f_D. Then, results in Lundgren(1968, p.404) show that

$$f_{X_{(j)}}(y) = n f_D(y) \frac{(n-1)!}{(n-j)!\,(j-1)!} F_D(y)^{n-j}\,[1-F_D(y)]^{j-1} \tag{21.0}$$

gives the PDF of the $j-th$ order statistic in a sample of size n. In this case we do not need to sample, we can define the distribution of each rank in the sample using Eq. 21.0. The expected value for each rank in a sample is then computed from

$$E(X_{(j)}) = \int y f_{X_{(j)}}(y)\, dy \tag{21.1}$$

The quantile-unbiased plotting position is defined by

$$p_j = \int_{E(X_{(j)})}^{\infty} f_D(x)\, dx \tag{21.2}$$

and the results for the LP3 fitted at Deming appear in Table 7b. The last two columns show that the approximation in Eq. 20.0 is close for the flood-peak distribution defined at Deming. It is interesting to note also that the largest flood of the series falls below the expected value for the largest flood out of a sample of 57. The second largest flood observed, the 1951 flood, is closer to the expected value for that rank out of a sample of 57. The smallest five floods are all greater than their expected value as defined by the LP3 fitted at Deming.

TABLE 7b: Selected Results for Order Statistics at Deming

Rank of the Statistic	Expected Value (ft^3/s)	Standard Deviation (ft^3/s)	Exact Plot Position	Cunanne Plot Position	Rank of the Statistic	Expected Value (ft^3/s)	Standard Deviation (ft^3/s)	Exact Plot Position	Cunanne Plot Position
1	75,480	16,360	0.0092	0.0105	30	30,030	1,700	0.5151	0.5175
2	63,380	9,160	0.0261	0.0280	31	29,580	1,670	0.5325	0.5350
3	57,700	6,860	0.0434	0.0455	32	29,140	1,640	0.5500	0.5524
4	54,010	5,670	0.0608	0.0629	33	28,710	1,610	0.5675	0.5699
5	51,280	4,920	0.0783	0.0804	34	28,280	1,590	0.5850	0.5874
6	49,110	4,390	0.0957	0.0979	35	27,860	1,560	0.6025	0.6049
7	47,300	4,000	0.1132	0.1154	36	27,440	1,540	0.6200	0.6224
8	45,750	3,690	0.1306	0.1329	37	27,020	1,520	0.6375	0.6399
9	44,400	3,440	0.1481	0.1503	38	26,600	1,500	0.6550	0.6573
10	43,190	3,230	0.1656	0.1678	39	26,180	1,480	0.6725	0.6748
11	42,100	3,050	0.1830	0.1853	40	25,770	1,460	0.6900	0.6923
12	41,100	2,900	0.2005	0.2028	41	25,350	1,440	0.7075	0.7098
13	40,180	2,760	0.2180	0.2203	42	24,930	1,420	0.7250	0.7273
14	39,330	2,650	0.2354	0.2378	43	24,500	1,410	0.7425	0.7448
15	38,530	2,540	0.2529	0.2552	44	24,070	1,390	0.7600	0.7622
16	37,780	2,450	0.2704	0.2727	45	23,640	1,380	0.7775	0.7797
17	37,070	2,360	0.2878	0.2902	46	23,190	1,370	0.7951	0.7972
18	36,400	2,290	0.3053	0.3077	47	22,730	1,360	0.8126	0.8147
19	35,760	2,210	0.3228	0.3252	48	22,260	1,360	0.8301	0.8322
20	35,150	2,150	0.3403	0.3427	49	21,770	1,350	0.8477	0.8497
21	34,560	2,090	0.3577	0.3601	50	21,260	1,350	0.8652	0.8671
22	33,990	2,030	0.3752	0.3776	51	20,720	1,350	0.8828	0.8846
23	33,450	1,980	0.3927	0.3951	52	20,130	1,360	0.9003	0.9021
24	32,920	1,930	0.4102	0.4126	53	19,500	1,380	0.9179	0.9196
25	32,410	1,890	0.4277	0.4301	54	18,780	1,410	0.9355	0.9371
26	31,910	1,850	0.4451	0.4476	55	17,940	1,460	0.9531	0.9545
27	31,420	1,810	0.4626	0.4650	56	16,880	1,550	0.9707	0.9720
28	30,950	1,770	0.4801	0.4825	57	15,250	1,800	0.9881	0.9895
29	30,490	1,740	0.4976	0.5000					

The often used Weibull plotting position, obtained with $a = 0$ in Eq. 20.0, is based on a different criterion and it gives a return period of 58 years for the largest observation. The differences shown by alternative plotting positions are largest for the top two observations in a sample, where the differences may be large.

At the smaller flows the differences are smaller. The criterion for the Weibull plotting formula is an unbiased estimate of the probability intervals between ranked observations. A result in mathematical statistics shows that

$$E\left[F_D(X_{(j)}) - F_D(X_{(j-1)})\right] = \frac{1}{n+1} \tag{22.0}$$

and the Weibull plotting position uses this criterion. The ranked flows tend to divide the probability mass of 1 into $(n+1)$ equal parts. Thus the Weibull plotting position gives, on average, unbiased exceedance probabilities.

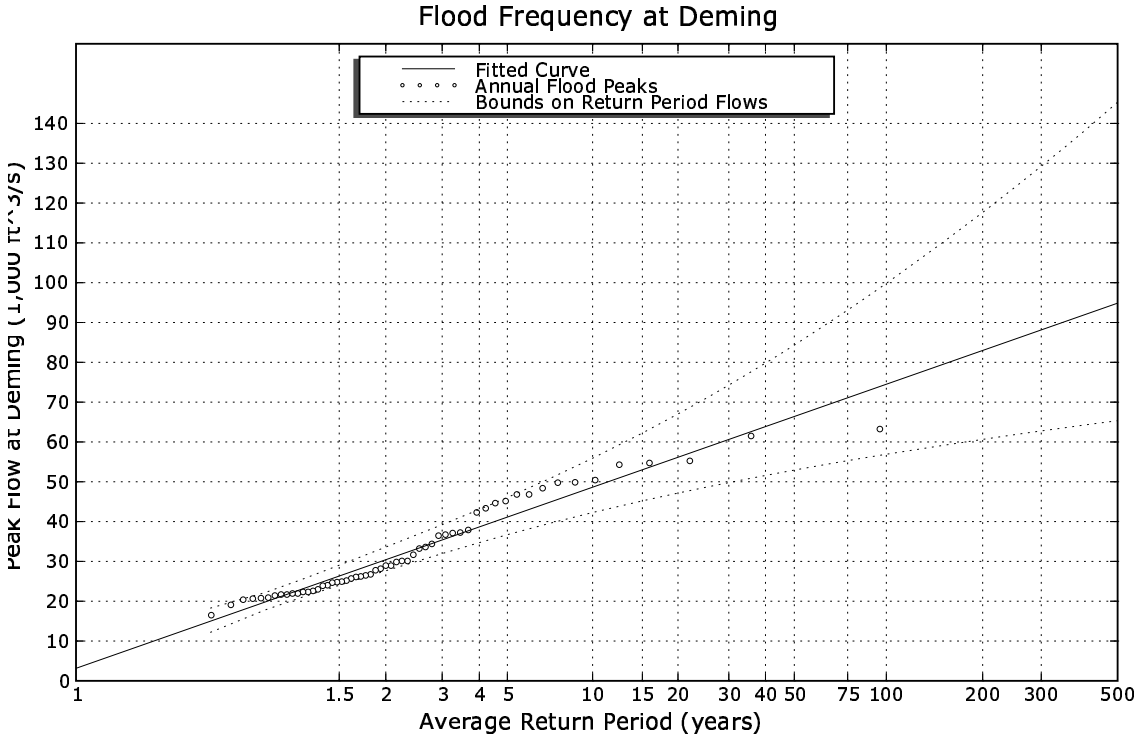

Figure 3 : Flood Frequency at Deming

In order to estimate the flood frequency distribution at Ferndale and at Everson Main Street, we now develop functions to predict the peak flow at Everson Main Street given the peak flow at Deming and also to predict the peak flow at Ferndale given the peak flow at Deming. For this purpose we use a combination of the simulated historical peaks and the peaks created by selective inflation of the 1990 and 1995 floods.

Analysis of Overflows at Everson Main Street

There are 36 joint values of peak flow at Deming and peak flow at Everson Main Street. These data are shown plotted in Figure 4 using a shifted value of the peak at Deming. The flow at Deming at which overflow begins at Everson Main Street was computed as a part of the analysis. In this case a three-panel linear spline appeared to fit the data well. The point of zero flow was identified by trial and error such that a value of zero shifted flow would yield a value of zero flow at Everson Main Street. The peak flow at Deming at which flow begins at Everson Main Street was $45786.6 ft^3/s$. Thus the first segment of the linear spline when extrapolated would pass through the origin for both flows. The middle two break points were optimized to make the residual sum of squares about the linear spline as small as possible. This process was automated in Mathcad, so that only a few trials were needed to define the optimum fit with a well defined point of zero overflow.

As Figure 4 shows, the three-segment linear spline fit the data well. The standard deviation of the residuals was estimated as $202 ft^3/s$. Table 8 gives the results for breakpoints (shifted flow at Deming), the coefficients of the linear spline (flow at Everson Main Street), and the Student t statistic for each coefficient. In this case we had 36 data points and we estimated four basis function coefficients, two breakpoint locations, and the peak flow at Deming when flow at Everson Main Street begins. That is seven parameters. Thus the number of degrees of freedom for the Student t significance level was 29, and this gave a significance value of 2.04

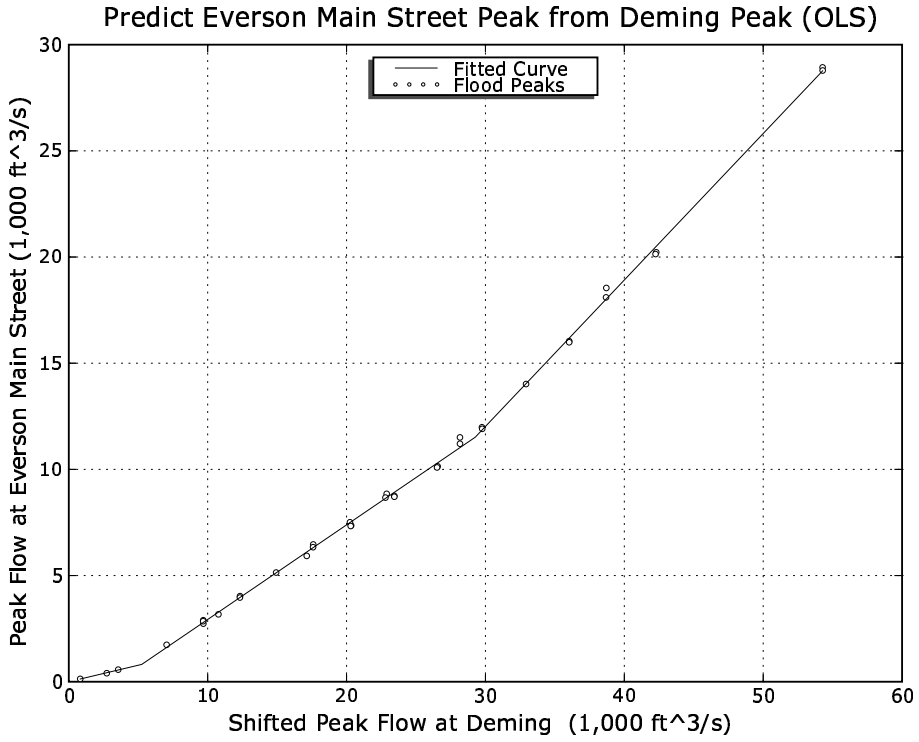

Figure 4: Predict Everson Main Street Peak from Deming Peak (OLS)

TABLE 8: Everson Main Street given Deming (OLS)

Shifted Flow at Deming (ft^3/s)	Flow at Everson MS (ft^3/s)	Student t Statistic
823.40	127.92	0.76
5252.5	815.93	9.43
29264.	11513.	181.
54281.	28768.	238.

Note that the first coefficient in Table 8, 127.92 appears to be insignificantly different than zero. Its computed Student t statistic is much less than the significance value of 2.05. However, this fit is a clear case where OLS is not appropriate. If the variance of the residual term is taken to be the same at all flow levels, then as the overflow at Everson Main Street approaches zero, as it must, because zero flow there is very much a possible outcome even when flooding occurs elsewhere, then a coefficient will be shown to be insignificantly different from zero when in fact it is not! This is a side effect of using OLS when it is not appropriate, that is, that the estimated sampling variances for the coefficients in the regression equation will

be larger than they should be. This outcome alone suggests that we attempt to find a simple functional form that shows how the variance of the residual term should vary with the shifted flow at Deming.

As a first step, consider the residuals computed from the fitted linear spline. If the standard deviation were really a constant at all shifted flow levels at Deming, then we would expect that the absolute value of the residual would not show much dependence on the shifted flow. If there is a clear trend and if a fitted line to the scatter shows a slope that differs from zero, we can conclude that the assumption of constant variance for the residual term is invalid.

Figure 5 shows the trend of the absolute value of the residuals. The residuals show a clear trend, getting larger as the shifted flow at Deming increases. Fitting a linear function to these residuals shows that the slope is statistically different than zero. This implies that the assumption of constant variance for the residual term is not supported. To estimate the variation of the variance of the residual term, the sample of 36 was broken into sub-samples of six points, all taken in ascending order, and the standard deviation was computed for each sub-sample. Figure 6 shows these values plotted at the average value of the shifted peak flow at Deming. There is a clear trend and fitting a straight line to the six points with OLS reveals a slope that differs from zero but an intercept that is not significantly different than zero at the 5 per cent level. Thus the six points were fitted by least squares, forcing the intercept to be exactly zero. This gives a slope of 0.16 and this fitted line is shown in Figure 6. These results show that the variance of the residual term should vary as $\sigma^2[0.16(Q_D - Q_0)^2]$ instead of being constant. This means that the variance about the fitted line will increase as the shifted flow at Deming increases. This seems reasonable and makes physical sense. Furthermore as the shifted flow at Deming approaches zero, the variance of the residual terms will also approach zero. This is also physically realistic in that the probability of negative flows at Main Street should be vanishingly small.

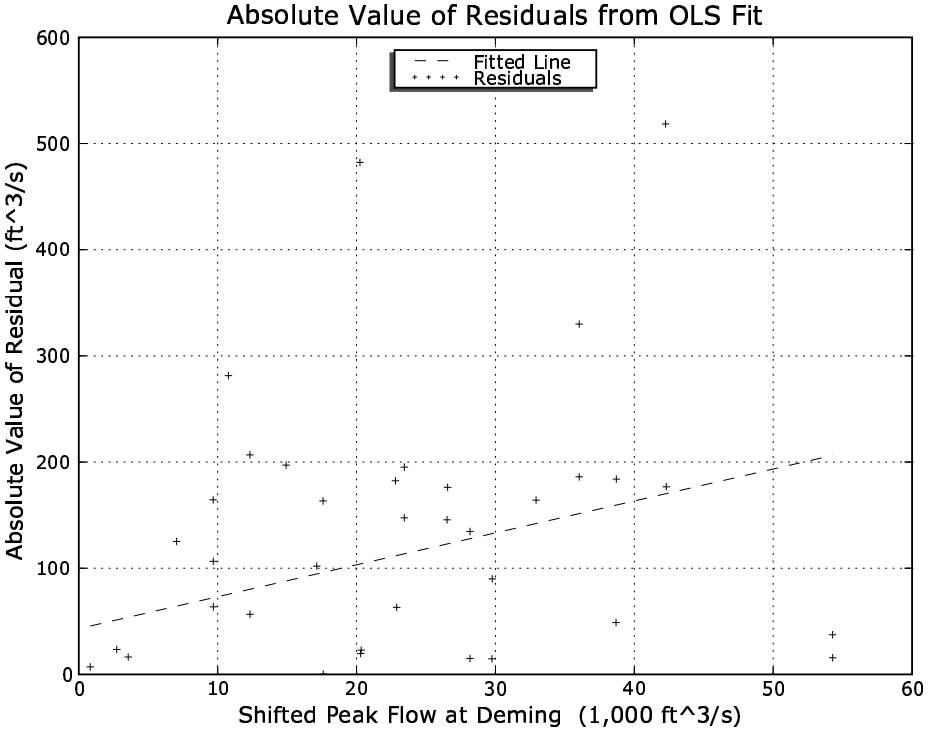

Figure 5: Absolute Value of Residuals from OLS Fit

Using Generalized Least Squares as outlined in Appendix A and the reference contained therein, yields the results shown in Figure 7 and Table 9. We see that the variance of the residual terms increases as it should and now we see that all coefficients are statistically different than zero. The peak flow at Deming

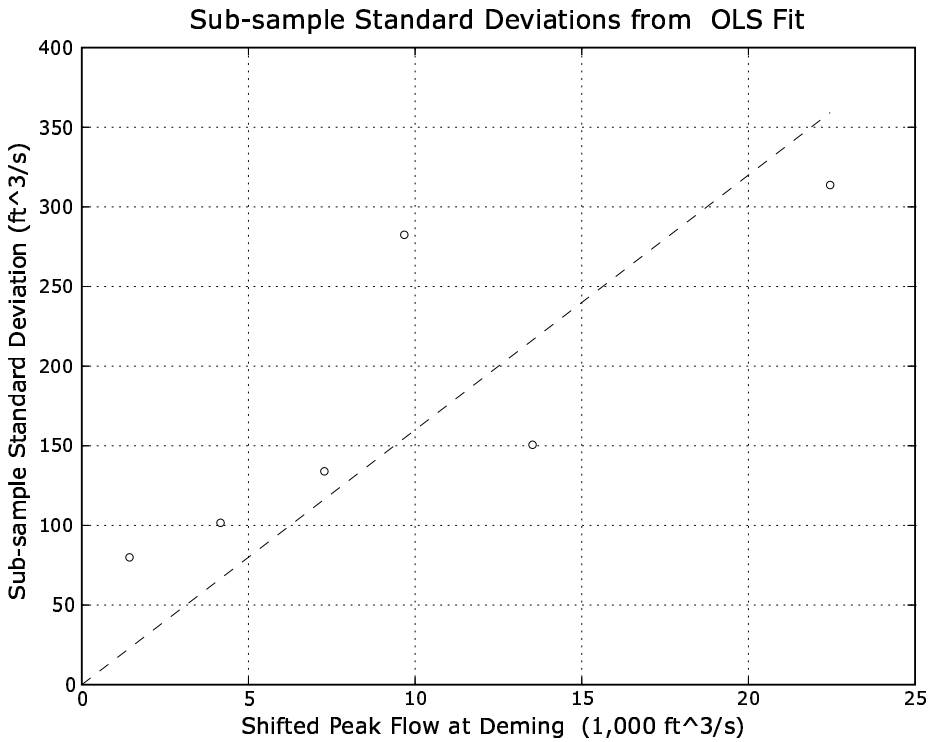

Figure 6: Subsample Standard Deviations from OLS Fit

at which flow begins at Everson Main Street was $45715.89 ft^3/s$. This would be expected to change as the fitting method changes.

TABLE 9 : Everson Main Street given Deming (GLS)

Shifted Flow at Deming (ft^3/s)	Flow at Everson MS (ft^3/s)	Standard Deviation (ft^3/s)	Student t Statistic
894.11	134.38	7.17	18.9
5034.0	756.56	40.4	29.4
28920.	11260.	232.	194.
54352.	28718.	436.	128.

Figure 7 shows that a linear spline is an excellent candidate for the functional form relating the peak flow at Deming and the peak flow at Everson Main Street. The variation between these two peaks is linear for extended ranges of flow. These ranges are well defined by the data. The changes in the slope of the linear spline must be related to the nature of the flow over the banks of the Nooksack. This is flow over a side weir with an irregular crest. There are more than 40 different locations included in the model for flow over the bank of the Nooksack. These differ in minimum crest elevation and in the shape of the crest. A check of the output shows that not all of them were active with the largest flow used in the extended data series. An important characteristic of flow over side weirs is that the flow increases rapidly with the head on the weir. The crest shape at the minimum crest elevation is that of a flat triangular weir. Flow over such weirs increases in proportion to the 2.5 power of the head on the weir. As the peak flows at Deming increase there is the dual effect of an increase of flow over surfaces already active and the initiation of flow over surfaces not previously reached by the water in the Nooksack River. As the rate of increase in the already active surfaces decreases, the newly activated surfaces are at their maximum rate of increase in flow. It may be that these

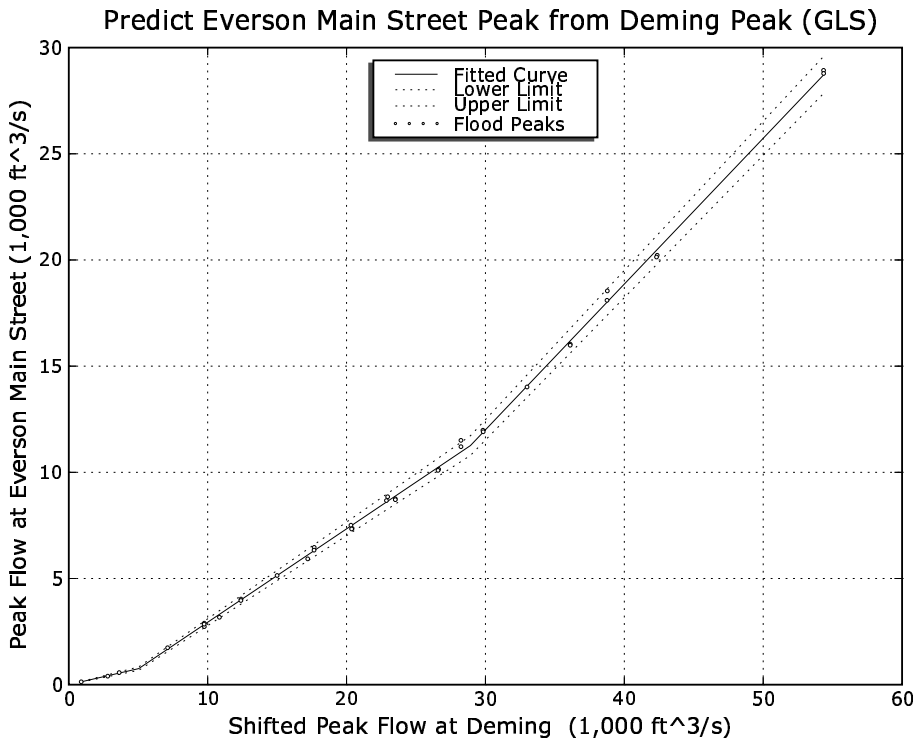

Figure 7: Predict Everson Main Street Peak from Deming Peak (GLS)

two effects result in the multiple linear relationships shown in Figure 7. An explanation is interesting but not needed because the data clearly show this behavior.

Predicting Ferndale Peak given the Peak at Deming

The relationship developed in the previous section, predicting the peak flow at Everson Main Street given the peak flow at Deming, will be used to estimate the average return periods for various flow levels at Everson Main Street. However, before we do that we want to develop a predictive relationship for the peak flow at Ferndale given the peak flow at Deming. Then we will develop the frequency relationships at these two locations. The Deming-Ferndale relationship will be developed using the extend set of data that includes inflated flows for the 1990 and the 1995 floods. We will need peak flows that are near $100,000 ft^3/s$ at Deming and the historical peak flow only reaches about $63,000 ft^3/s$ there. Extrapolating the historical fit is invalid because as the peak flows at Deming reach levels not experienced we can expect that new overflow surfaces will become active. Thus the relationship between the peaks at Ferndale and at Deming would be expected to change. The linear-spline fitted to predict the peak flow at Everson Main Street from the peak flow at Deming does show two distinct and well defined breaks in slope. The first break is near a peak flow of $50,750 ft^3/s$ at Deming and the second is near a peak flow of $74,640 ft^3/s$ at Deming. The flow at the first interior breakpoint has been seen several times in the historical flood series. However, the second interior breakpoint is near the 100-year flood level and has not yet been experienced on the lower Nooksack River.

There are 45 data points for the relationship between the peak flow at Deming and the peak flow at Ferndale. A three-panel linear spline fitted the data well. The first interior breakpoint was fixed at the flow level at Deming that initiates flow at Everson Main Street and the second interior breakpoint was allowed to vary to improve the fit. Thus there were five estimated parameters: four coefficients and one breakpoint location. Tests of the absolute residuals and of the variation of the standard deviation for sub-samples did not show significant reason to reject the assumption of a homoscedastic disturbance term. The standard deviation of the residuals was $2003 ft^3/s$. The results are shown in Figure 8 and Table 10. The three-panel linear spline explains 98.6 per cent of the variance in the peak flows at Ferndale. Also all of the coefficients

are statistically non-zero. The Student t for 40 degrees of freedom is 2.02 and all coefficients have a t statistic far larger than that.

TABLE 10 : Ferndale given Deming (OLS)

Flow at Deming (ft^3/s)	Flow at Ferndale (ft^3/s)	Standard Deviation (ft^3/s)	Student t Statistic
15203.0	13983	2003	15.0
45715.9	35906	2003	33.5
59692.1	53598	2003	90.9
100068.	72426	2003	75.4

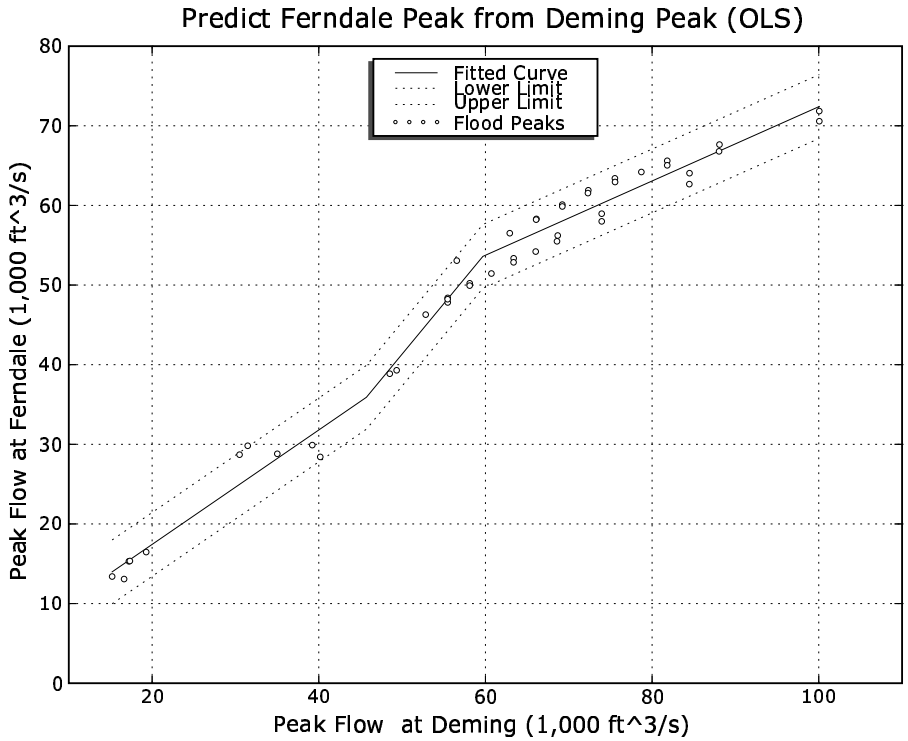

Figure 8: Predict Ferndale Peak from Deming Peak (GLS)

Estimating Flood Frequency at Ferndale and Everson Main Street From Deming

With the prediction equations established that relate the peak flows at Ferndale and Everson Main Street to the peak flow at Deming, we can transform the flood frequency results at Deming to the flood frequency results at the other locations. The first and simplest way to accomplish this transform is to note that each of the prediction equations is monotone increasing and single valued. Let $h_{E|D}(x) =$ the prediction equation found above for the peak flow at Everson Main Street given the peak flow at Deming; $h_{F|D}(x) =$ the prediction equation found above for the peak flow at Ferndale given the peak flow at Deming; $Q_D =$the peak flow at Deming; $Q_E =$ the peak flow at Everson Main Street; $Q_F =$the peak flow at Ferndale; $P_D(\cdots)$ be the function that gives the probability of an event at Deming; $P_E(\cdots)$ be the function that gives the probability of an event at Everson Main Street; and $P_F(\cdots)$ be the function that gives the probability of an

event at Ferndale. The three dots in the later three functions will be replaced by an event where an event is, for example, $Q_D > q$ where q is some number such as $75,000$. Then because the prediction functions are monotone increasing and single valued we get for the peak flow at Everson Main Street

$$P_E[Q_E > h_{E|D}(q)] = P_D(Q_D > q) \tag{23.0}$$

and

$$P_F[Q_F > h_{F|D}(q)] = P_D(Q_D > q) \tag{24.0}$$

for the peak flows at Ferndale. The probability functions used here are just the complement of the cumulative distribution function (CDF) at each location. The results for these two locations are given in Table 11.

TABLE 11: Flood Frequency at Everson Main Street and Ferndale (CDF)

Return Period (years)	Peak Flow at Everson MS (ft^3/s)	Peak Flow at Ferndale (ft^3/s)
1.01	0	13,817
1.25	0	19,631
2.00	0	24,918
4.00	0	30,837
5.00	0	32,609
10.0	439	39,599
15.0	1,753	45,147
20.0	3,131	49,115
25.0	4,211	52,222
35.0	5,858	54,838
50.0	7,635	56,723
75.0	9,701	58,914
100.0	11,199	60,502
150.0	14,536	62,793
200.0	16,987	64,457
300.0	20,528	66,863
400.0	23,106	68,614
500.0	25,144	69,998

The peak flow at Deming that initiates overflow at Everson Main Street has a return period of 7.6 years. Thus some overflow can be expected on average about once every eight years. Based on past history of overflows it appears that an overflow occurs about every five years but this may include more than one overflow in a year as occurred in November of 1990. The results in Table 5 yield an overflow about once every 5 years as well if the two uncertain events are included. Excluding them gives an interval of slightly more than 6 years. The overflow peak in November of 1990 has a return period of about 35 years which is the same as the return period for the peak flow at Deming for that event as is expected given the close relationship between the two peak flows.

The results in Table 11 treat the peak-flow relationships as being exact, that is, free of any error or uncertainty. This is not true as the relationships were developed from a least-squares fitting process to data that clearly is only approximate. What is the effect of the variation about the regression line found in this process? We estimate that effect by developing the flood frequencies at Everson Main Street and Ferndale including the estimated effect of the variance of the residual term about the regression line. It is probable that the variance as estimated above is an under estimate because all of the data came from a simulation model. However, we can access the effect of the variance and then determine if it has a major effect on the outcomes.

The regression line plus the scatter about the line implied by the assumed normal distribution of the residual, defines a conditional probability density function for the predicted flow given the flow at Deming. The probability density function, $f_X(x)$ gives the probability of a value occurring in a small interval centered on the value x. Consequently if we think of the total probability, 1.0, being some tangible substance, such as butter, the probability density function defines how a fixed amount of the substance is distributed across the possible values. As such it gives the density, the amount of probability per unit length, for a given random variable X. A normal distribution is completely defined when its mean value and its standard deviation are known. The mean value is just the flow at Everson Main Street or at Ferndale given by the prediction equation. The standard deviation is the standard deviation of the residual term. Let $f_N(x, \mu, \sigma)$ be the probability density function of the normal distribution with mean μ and standard deviation σ.

The conditional probability density function for Everson Main Street is then:

$$f_{E|D}(q_E, q_D) = f_N[q_E, h_{E|D}(q_D - q_0), 0.08(q_D - q_0)] \tag{25.0}$$

where $q_0 = 45715.89 ft^3/s$=peak flow at Deming that initiates flow at Everson Main Street. The coefficient 0.08 gives the variation of the standard deviation with shifted peak flow at Deming developed when fitting the predictive relationship. The conditional probability density function at Ferndale is a bit simpler, since the residual was found to be homoscedastic. It is then

$$f_{F|D}(q_F, q_D) = f_N[q_F, h_{F|D}(q_D - q_0), 2003.0] \tag{26.0}$$

where 2003.0 is the standard deviation of the residuals determined earlier. With these conditional probability density functions and with the density function for the flow at Deming, we define the joint probability density function for the respective peak-flow pairs as:

$$f_{E,D}(q_E, q_D) = f_{E|D}(q_E, q_D) f_D(q_D) \tag{27.0}$$

for the peak flow at Everson and the peak flow at Deming, and

$$f_{F,D}(q_F, q_D) = f_{F|D}(q_F, q_D) f_D(q_D) \tag{28.0}$$

for the peak flow at Ferndale and the peak flow at Deming.

We can now compute the probability density functions for the peak flow at Everson Main Street and at Ferndale by finding the marginal density function for each of these flows from the respective joint density functions. That is, for Everson Main Street,

$$f_E(q_E) = \int_{Q_{D_{min}}}^{\infty} f_{E,D}(q_E, q_D)\, dq_D \tag{29.0}$$

where $Q_{D_{min}}$ is the minimum peak flow at Deming defined by the flood frequency analysis. Equation 29.0 only gives the continuous part of the mixed distribution at Everson Main Street. The probability of zero peak flow there is given by

$$P_{E=0} = \int_{q_0}^{\infty} f_D(q_D)\, dq_D \tag{30.0}$$

To derive the result at Ferndale we assume that the first segment of the linear spline for $h_{F|D}$ can be extrapolated to the minimum flow for Deming to yield the minimum peak flow at Ferndale. Define this value as $Q_{F_{min}}$. This was found to be $5322 ft^3/s$. We then get

$$f_F(q_F) = \int_{Q_{F_{min}}}^{\infty} f_{F,D}(q_F, q_D)\, dq_D \tag{31.0}$$

The upper limit in Eqs. 29.0 and 31.0 were replaced by large flows and then the equations were integrated numerically to compute values of the marginal density functions. These values were then fitted with a cubic spline and then integrated from their lower limit to their upper limit to verify that we had obtained all but a minute part of the probability under the density function. In each case the discrepancy was less than 1

unit out of more than 2 million. Thus the return periods that would be affected by this discrepancy would have to be on the order of 500,000 years or more! It is then valid to compute the flows for each of the return periods shown in Table 11 but now including the effect of the scatter inherent in the predictive relationships used to estimate for the flood frequency at Everson Main Street and Ferndale from the flood frequency at Deming.

Table 12 shows the results and shows the relative difference between the two estimates. As can be seen the differences are small except at the smaller return period events. Even there the differences are small relative to the uncertainties in the basic flood-peak data itself. The differences treat the results from the joint PDF as the base, that is, a positive percentage difference means that the flow from using the CDF is larger than the flow using the JPDF. Clearly the differences between flows at Everson Main Street are within the noise level of fitting and numerical integration. However, the small return-period results for Ferndale show differences larger than the computational noise. The effect of including the variance about the regression line is to decrease the small-return period flows and to increase the large-return period flows. This comes about in part because including the variance of the residual term in the development, results in a minimum flow at Ferndale that is smaller than the minimum flow computed from the regression line. The CDF results take the computed minimum flow as the true minimum. However, once we allow variation about the regression line there is some probability that a flow will be less than the minimum shown from the regression line. This puts more of the probability mass below the computed minimum and thus the lower return period flows can be somewhat smaller in order to have a given probability mass below them on the axis. This shift however must be compensated by a shift at other locations along the flow axis because the probability mass must be 1.0. Consequently the large-return period events must increase somewhat to again have a given probability mass below them as well.

TABLE 12: Flood Frequency at Everson Main Street and Ferndale (JPDF)

Return Period (years)	Peak Flow at Everson MS (ft^3/s)	Percent Difference (CDF - JPDF)	Peak Flow at Ferndale (ft^3/s)	Percent Difference (CDF-JPDF)
1.01	0	0.000	12,589	9.756
1.25	0	0.000	19,459	0.885
2.00	0	0.000	25,030	-0.445
4.00	0	0.000	31,059	-0.715
5.00	0	0.000	32,860	-0.762
10.0	437	0.246	39,770	-0.431
15.0	1,752	0.056	45,296	-0.331
20.0	3,130	0.037	49,254	-0.282
25.0	4,209	0.029	52,067	0.297
35.0	5,857	0.021	54,918	-0.145
50.0	7,634	0.014	57,072	-0.612
75.0	9,700	0.010	59,301	-0.654
100.0	11,225	-0.23	60,886	-0.631
150.0	14,533	0.023	63,168	-0.594
200.0	16,983	0.027	64,826	-0.568
300.0	20,521	0.035	67,221	-0.533
400.0	23,096	0.044	68,965	-0.509
500.0	25,131	0.052	70,344	-0.491

A standard frequency analysis was done at Ferndale to compare its results with the results shown in Table 12. The details are not given but the key parameters were: $\mu = 10.1589$, $\sigma = 0.3445$,$\gamma = 0.4136$, $\alpha = 23.3853$, $\beta = 14.0378$, and $\xi = 8.4930$. The minimum annual peak flow from this analysis was: 4881. It is perhaps accidential but never-the-less interesting that this minimum flow is somewhat less than the minimum flow predicted by $h_{F|D}$ which was 5322! This suggests that the flood frequency using the joint probability density function is to be preferred.

Figure 9 shows the results at Ferndale with a frequency curve from the at-station analysis at Ferndale, the frequency curve derived from the joint probability analysis, and with the annual flood peaks shown at the return period given by the Cunanne formula given above. The horizontal axis giving the return period has been set so that the flood frequency derived from Deming will plot as a straight line. We can then see the dramatic effect of *not* including the overflow at Everson when extrapolating to the larger return periods. Clearly flood frequency analysis, being purely statistical, knows nothing of hydrology or hydraulics. The analyst must include the effect of special factors to the degree possible.

Figure 10 then gives the flood frequency curve at Everson Main Street with the annual flood peaks again shown at their Cunanne return period. It proved impossible to render this curve as a straight line and still plot the annual flood peaks. Therefore, the return-period scale in this figure is the same as used at Deming, that is, so that the Deming flood frequency curve would be a straight line.

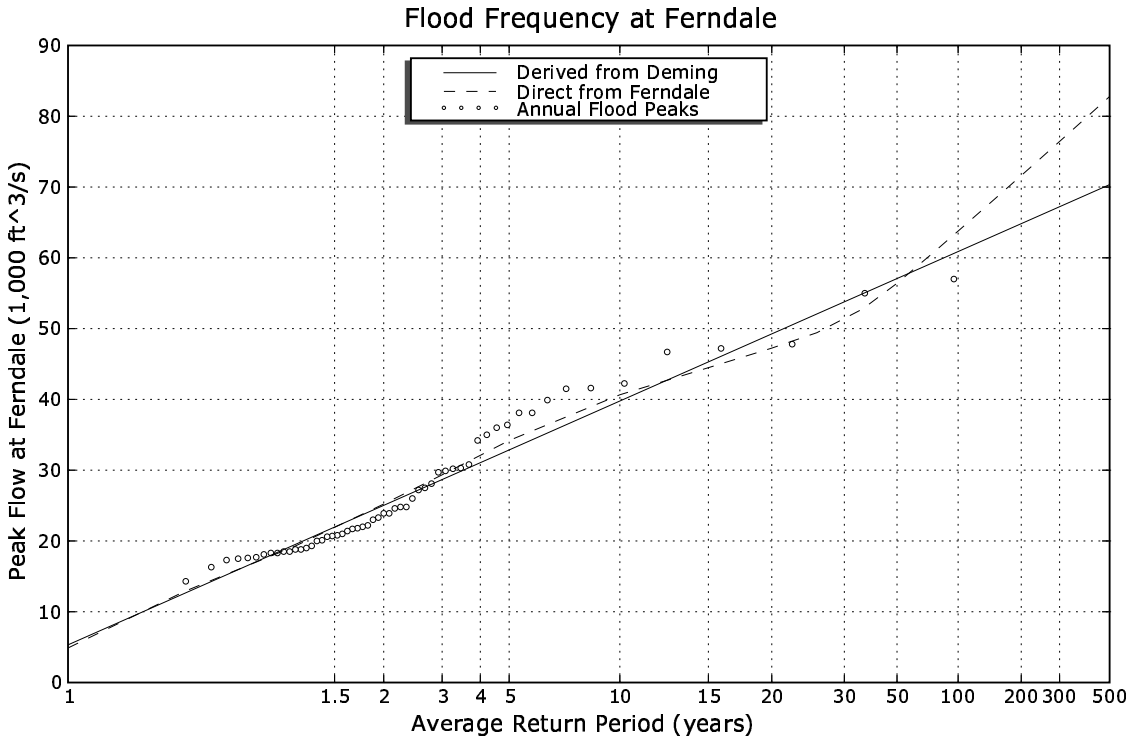

Figure 9: Flood Frequency at Ferndale

The frequency results at Everson Main Street are the first ever computed. There are no records of flow at this location and all flows are based on runs of the model of the lower Nooksack River. The sample of only 11 peak flows is quite small. It appears that the fit based on Deming assigns a much smaller return period to the peak flow of record than the record length of 57 years indicates. This is the same result as at Deming, as would be expected given that the peak of record at Deming also produced the peak of record at Everson Main Street. That will continue to be the case assuming that the current unsteady flow model captures the major features of the overflows at Everson. The results to date support that assumption.

Discussion of Results

The development of flood frequency estimates at Ferndale and at Everson Main Street is an involved process. However, the Nooksack River with major overflow to Canada occurring near Everson, is not a standard River either! The methods used were designed to make good use of the available data and the available unsteady-flow model of the lower Nooksack River. Here are some major observations based on the process of producing the results outlined above:

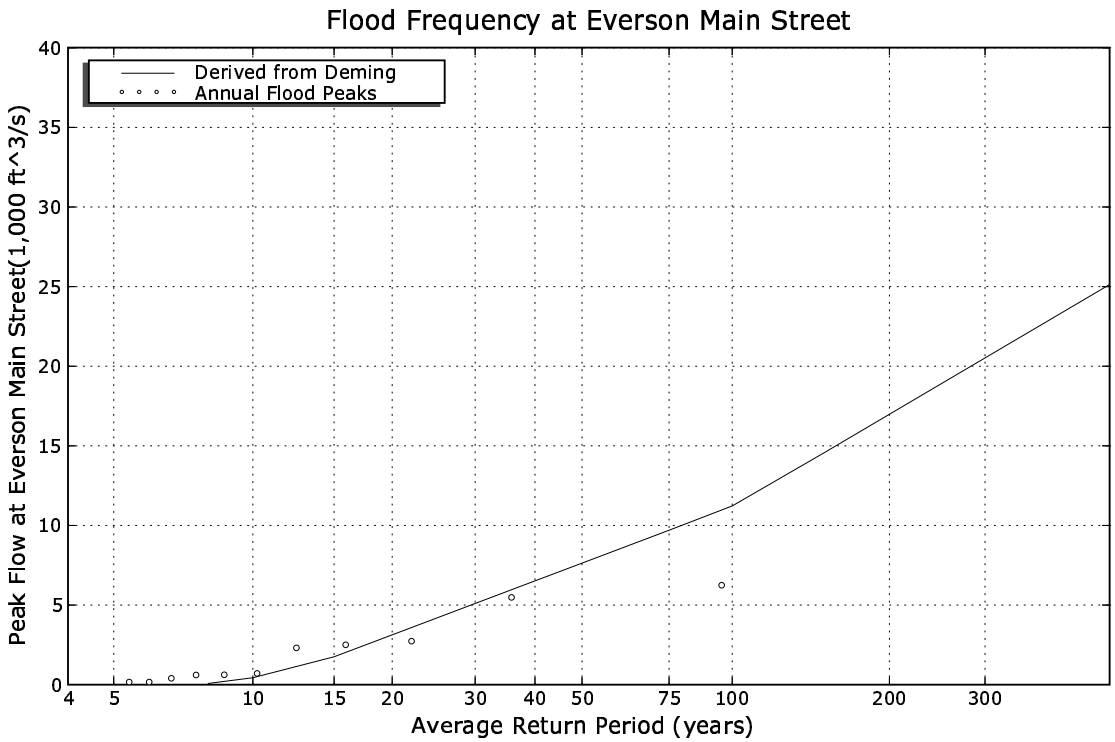

Figure 10: Flood Frequency at Everson Main Street

1. Variations in fitting, assumptions on the flood distribution to use, and other details could contribute changes of a few percent to the results. The precision used in the presentation of results is far greater than the data itself. The high precision was used for consistency purposes only. Any final values should be rounded to perhaps two significant figures at most.
2. The estimated variances of the residuals as found in the regression lines for the flows at Ferndale and Everson Main Street as functions of the flows at Deming, may be under estimates. The peak flows involved were all derived from the unsteady-flow model of the lower Nooksack River. A similar set of observed flows would probably show more variation. However, no such set is available nor will one become available because the gaging station at Deming has an inherently unstable rating. Only time will establish if the new gaging station at Cedarville will prove to be more reliable. It is also unlikely that a record of flows at Everson Main Street will become available. In any case, the effect of scatter about the regression line is small so any underestimate implicit in the methods used here has nil effect on the conclusions.
3. Changes in the peak overflow at Everson Main Street are sensitive to changes at Deming. When the overflow at Everson Main Street is near $1000 ft^3/s$ a change of one per cent in the peak flow at Deming can change the peak flow at Everson Main Street by as much as 20 per cent. This relative change in peak flow at Everson Main Street attenuates to about 8 percent when the flow is about $3000 ft^3/s$. The relative change then continues to decline to four percent when the peak flow at Everson Main Street reaches about $6400 ft^3/s$. Even at the highest flows computed the relative changes is still about three precent for a one percent change at Deming. Given the usual uncertainties in a peak flow even under ideal conditions, implies an uncertainty in the peak flow at Everson Main Street of 20 percent or more if based on values at Deming.
4. Reasonable checks on the consistency of the sample data with the fitted relationship shows no reason to discard either the LP3 distribution at Deming or the derived distributions at Ferndale and Everson Main Street.

Conclusions and Recommendations

The flood frequencies derived for Ferndale and Everson Main Street are the best estimates currently available. They are internally consistent with each other and with the peak flows at Deming. In so far as the unsteady-flow model of the lower Nooksack River captures the major processes in the system, the flood frequencies are both valid and useful. All checks on the unsteady-flow model to date support the belief that it does in fact represent the response of the stream to floods seen in the recent past and also larger floods than any yet seen. Consequently the results found here are recommended for use in further work in developing a flood hydrograph at Deming to use in future studies of the stream system.

APPENDIX A: Technical Background

Selection of an Approximating Function

We have three functions to define to carry out the steps outlined above. These functions will be based on the limited data we have available. As a consequence the functions need to be simple but yet able to represent the major part of the variation of the value being predicted. Polynomials are simple functions and we can increase the degree of the polynomial to get a closer fit. However, polynomials have the disadvantage of being sensitive to changes anywhere in their range of approximation. That is, a change in one data point can cause changes in the polynomial far from that data point. Another way to state it, is that a polynomial is a global approximation. To keep the simplicity of the polynomial but to also make the approximation more local, we will be using piecewise polynomials in the analysis below. That is, we break the range of approximation into several sub-ranges, with the division points at the sub-range limits being called breakpoints. They so to speak, break the range into parts. It is convenient to call the interval between adjacent breakpoints a panel. Thus is there is only one panel the piecewise polynomial is really only a single polynomial. However, if there is more than one panel, that is, we have one or more breakpoints interior to the range, then we have a piecewise polynomial. At each interior breakpoint we need to establish some conditions on the two polynomials that have the breakpoint in common. These conditions are defined in terms of the continuity of the function value and its derivatives at the breakpoint. For example, if we use a first degree polynomial in each panel, then a natural condition is to require that the approximation be continuous at each interior break point. However, the slope of the lines on each side of the breakpoint may differ. If we require that the slope also be continuous, then we have a single polynomial across all breakpoints and we no longer have a piecewise polynomial.

In general a piecewise polynomial function will have one or more derivative with discontinuities at the interior breakpoints. Special names are applied to some piecewise polynomials. If all but the highest potentially non-zero derivative for the polynomials being used is continuous, the piecewise polynomial is called a polynomial spline. If the polynomial in each panel is of the first degree, that is, linear, it is called a linear spline. The first derivative is discontinuous at each breakpoint in this case and the first derivative is the highest potentially non-zero derivative for a first degree polynomial. If the polynomial in each panel is cubic, then the third derivative is discontinuous at each interior breakpoint, and the resulting piecewise polynomial is called a cubic spline. Again the third derivative is the highest potentially non-zero derivative of a third degree polynomial. See De Boor(1987) for a more extensive treatment of piecewise polynomial approximations.

We will be using linear splines for all of the functions to be found here. A linear spline is both simple and versatile and can give a close fit to data and other functions. We do not need a continuous first derivative in our analysis so this function will be able to reflect local variation without suffering global shifts in value like a cubic or quartic polynomial applied to the entire range.

The simplest way to represent a linear spline for mathematical fitting is to define it in terms of its basis functions. That is if $f(x)$ is a linear spline we represent it as follows:

$$f(x) = a_1 H_{1,m}(x) + a_2 H_{2,m}(x) + \cdots + a_m H_{m,m}(x) \qquad \text{(A-1)}$$

where a_i for $i = 1, \ldots, m$ are coefficients to be determined and $H_{i,n}$ for $i = 1, \ldots, m$ are the basis functions for a linear spline. Implicit in the basis functions and the linear spline are the locations of the breakpoints. The basis functions are defined once the breakpoints are selected. Let x_i for $i = 1, \ldots, n$ be the break points. Then

$$H_{1,m}(x) = \begin{cases} \dfrac{x_2 - x}{x_2 - x_1}, & \text{if } x_1 \le x \le x_2; \\ 0, & \text{otherwise;} \end{cases} \qquad \text{(A-2.1)}$$

$$H_{i,n}(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & \text{if } x_{i-1} \le x \le x_i; \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & \text{if } x_i \le x \le x_{i+1}; \\ 0, & \text{otherwise;} \end{cases} \qquad \text{(A-2.2)}$$

$$H_{m,m}(x) = \begin{cases} \dfrac{x - x_{m-1}}{x_m - x_{m-1}}, & \text{if } x_{m-1} \le x \le x_m; \\ 0, & \text{otherwise;} \end{cases} \tag{A-2.3}$$

The H-functions are called "hat" functions, as Figure A-1 shows for the breakpoint sequence of $0, 1.0, 1.5, 4.0$ Notice that each interior breakpoint has a complete "hat" but the initial and final breakpoints only have "one-half" hat! These functions are called basis functions because all possible linear-splines with a given breakpoint sequence can be represented by the proper choice of coefficients in Eq. A-1. Another convenience of linear spline functions is that the coefficients in Eq. A-1 are the same as the spline function value at that breakpoint. This makes defining an interpolating linear spline simple. Select the breakpoints, then find the function values at the breakpoints, and then use them as the coefficients in Eq. A-1 to interpolate values of the function.

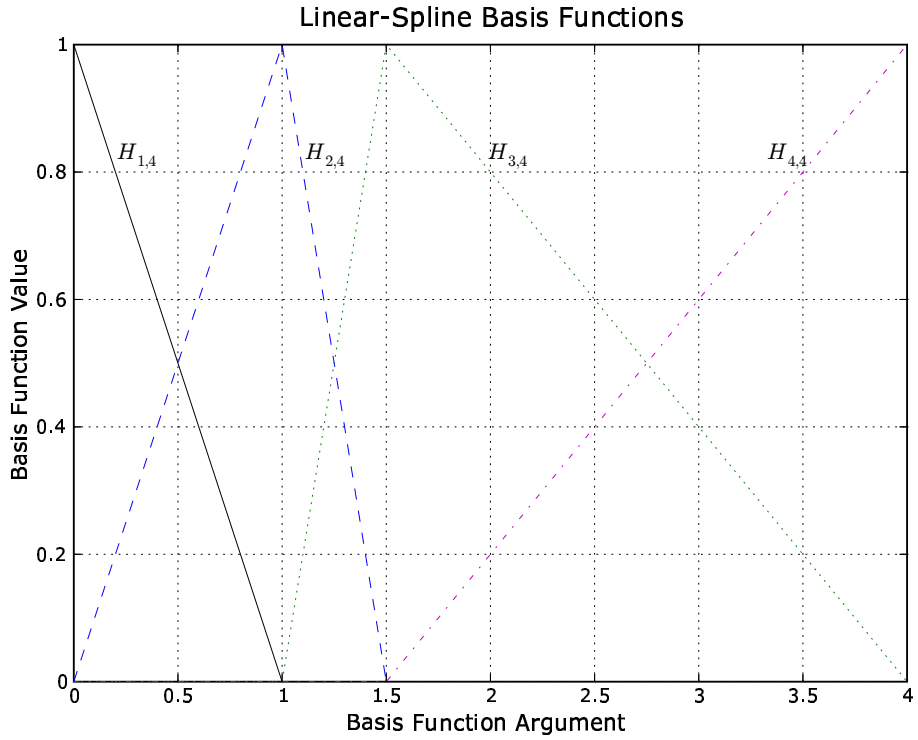

Figure A-1 Example Linear-Spline Basis Functions

Least-Square Fitting of Linear Splines

Linear splines are simple yet versatile in representing functions. Least-squares fitting is often the method of choice when defining function based on experimental or observational data. Using least squares with linear splines is no more complex than using polynomials and has advantages as well. When we fit a function to data we find that the fit is not exact. This comes about from several sources: the data may be subject to observational errors, the data may be a sample from some random process, or the function we are using may not be the exact function but only an approximation. Usually all three factors are present. If we have observations on pairs of values, say x_i, y_i for $i = 1, \ldots, n$ we can write an approximating linear spline as

$$y_i = \hat{a}_1 H_{1,m}(x_i) + \cdots + \hat{a}_j H_{j,m}(x_i) + \cdots + \hat{a}_m H_{m,m}(x_i) + e_i \tag{A-3}$$

where e_i = the residual error for the $i-$th data point. We use the hat symbol on the coefficients to indicate that we are finding estimates of some unknown true values.

The least-squares criterion requires that we minimize the sum of the squares of the residuals, that is to minimize $\sum_{i=1}^{n} e_i$. Let

$$F(\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_m) = \sum_{i=1}^{n} \left[y_i - \hat{a}_1 H_{1,m}(x_i) - \cdots - \hat{a}_j H_{j,m}(x_i) - \cdots - \hat{a}_m H_{m,m}(x_i) \right]^2 \tag{A-4}$$

Then to find the minimum value of this sum of squares we take the derivative of F with respect to each of the coefficients and set them to zero. This defines m equations with the m coefficients as the unknowns. We compute the derivative for a typical coefficient, $\hat{a}_j$ for $j = 1, \cdots, m$ and then use the same pattern for all the others.

$$\frac{\partial F}{\partial \hat{a}_i} = \sum_{i=1}^{n} 2 \left[y_i - \hat{a}_1 H_{1,m}(x_i) - \cdots - \hat{a}_j H_{j,m}(x_i) - \cdots - \hat{a}_m H_{m,m}(x_i) \right] H_{j,m}(x_i) = 0 \tag{A-5}$$

The factor of 2 drops out of the equation since the right-hand side is 0. Multiply each term in the square brackets by $H_{j,m}$ and apply the summation to each term to get

$$\sum_{i=1}^{n} y_i H_{j,m}(x_i) - \hat{a}_1 \sum_{i=1}^{n} H_{1,m} H_{j,m}(x_i) - \cdots - \hat{a}_j \sum_{i=1}^{n} H_{j,m}(x_i) H_{j,m}(x_i) - \cdots - \hat{a}_m \sum_{i=1}^{n} H_{m,m}(x_i) H_{j,m}(x_i) = 0 \tag{A-6}$$

To get the final form of the typical $j-$th equation, transpose the left-most summation to the right-hand side and multiply the equation by -1 to yield

$$\hat{a}_1 \sum_{i=1}^{n} H_{1,m}(x_i) H_{j,m}(x_i) + \cdots + \hat{a}_j \sum_{i=1}^{n} H_{j,m}(x_i) H_{j,m}(x_i) + \cdots + \hat{a}_m \sum_{i=1}^{n} H_{m,m}(x_i) H_{j,m}(x_i) = \sum_{i=1}^{n} y_i H_{j,m}(x_i) \tag{A-7}$$

This equation is then given once for each j to yield the m equations to solve for the m unknown coefficients. The notation in these equations is a bit dense so let us write out the terms in full for a three-panel linear spline. There will then be four breakpoints and four basis functions. Note that the breakpoints are implicit in the basis functions and are not denoted explicitly.

$$\hat{a}_1 \sum_{i=1}^{n} H_{1,4}(x_i)^2 + \hat{a}_2 \sum_{i=1}^{n} H_{2,4}(x_i) H_{1,4}(x_i) + \hat{a}_3 \sum_{i=1}^{n} H_{3,4}(x_i) H_{1,4}(x_i) + \hat{a}_4 \sum_{i=1}^{n} H_{4,4}(x_i) H_{1,4}(x_i) = \sum_{i=1}^{n} y_i H_{1,4}(x_i) \tag{A-8.1}$$

$$\hat{a}_1 \sum_{i=1}^{n} H_{1,4}(x_i) H_{2,4}(x_i) + \hat{a}_2 \sum_{i=1}^{n} H_{2,4}(x_i)^2 + \hat{a}_3 \sum_{i=1}^{n} H_{3,4}(x_i) H_{2,4}(x_i) + \hat{a}_4 \sum_{i=1}^{n} H_{4,4}(x_i) H_{2,4}(x_i) = \sum_{i=1}^{n} y_i H_{2,4}(x_i) \tag{A-8.2}$$

$$\hat{a}_1 \sum_{i=1}^{n} H_{1,4}(x_i) H_{3,4}(x_i) + \hat{a}_2 \sum_{i=1}^{n} H_{2,4}(x_i) H_{3,4}(x_i) + \hat{a}_3 \sum_{i=1}^{n} H_{3,4}(x_i)^2 + \hat{a}_4 \sum_{i=1}^{n} H_{4,4}(x_i) H_{3,4}(x_i) = \sum_{i=1}^{n} y_i H_{3,4}(x_i) \tag{A-8.3}$$

$$\hat{a}_1 \sum_{i=1}^{n} H_{1,4}(x_i) H_{4,4}(x_i) + \hat{a}_2 \sum_{i=1}^{n} H_{2,4}(x_i) H_{4,4}(x_i) + \hat{a}_3 \sum_{i=1}^{n} H_{3,4}(x_i) H_{4,4}(x_i) + \hat{a}_4 \sum_{i=1}^{n} H_{4,4}(x_i)^2 = \sum_{i=1}^{n} y_i H_{3,4}(x_i) \tag{A-8.4}$$

These equations have some zero terms. Each of the sums on the left-hand side involves the sum of the product of two basis functions. If the indices for the basis function differ by more than 1, then the product

of those two basis functions is 0! For example in the first of the four equations we have the sum of the product, $H_{3,4}(x_i)H_{1,4}(x_i)$. Look at Figure A-1 showing an example of the basis function for a three-panel linear spline. This figure applies to all three-panel linear splines by merely changing the breakpoint values. The pattern will remain the same. Clearly whenever $H_{3,4}(x_i) > 0$ we see that $H_{1,4}(x_i) = 0$. Pick any pair of basis functions whose indices differ by more than 1 and you will find that their product is always zero. Consequently the equations simplify to

$$\hat{a}_1 \sum_{i=1}^{n} H_{1,4}(x_i)^2 + \hat{a}_2 \sum_{i=1}^{n} H_{2,4}(x_i)H_{1,4}(x_i) = \sum_{i=1}^{n} y_i H_{1,4}(x_i) \tag{A-9.1}$$

$$\hat{a}_1 \sum_{i=1}^{n} H_{1,4}(x_i)H_{2,4}(x_i) + \hat{a}_2 \sum_{i=1}^{n} H_{2,4}(x_i)^2 + \hat{a}_3 \sum_{i=1}^{n} H_{3,4}(x_i)H_{2,4}(x_i) = \sum_{i=1}^{n} y_i H_{2,4}(x_i) \tag{A-9.2}$$

$$\hat{a}_2 \sum_{i=1}^{n} H_{2,4}(x_i)H_{3,4}(x_i) + \hat{a}_3 \sum_{i=1}^{n} H_{3,4}(x_i)^2 + \hat{a}_4 \sum_{i=1}^{n} H_{4,4}(x_i)H_{3,4}(x_i) = \sum_{i=1}^{n} y_i H_{3,4}(x_i) \tag{A-9.3}$$

$$\hat{a}_3 \sum_{i=1}^{n} H_{3,4}(x_i)H_{4,4}(x_i) + \hat{a}_4 \sum_{i=1}^{n} H_{4,4}(x_i)^2 = \sum_{i=1}^{n} y_i H_{3,4}(x_i) \tag{A-9.4}$$

The coefficient matrix for the linear system of equations is both symmetric and tri-diagonal. Thus we need only compute the elements on the diagonal and on the super diagonal. The solution for a tri-diagonal linear system is also simple and fast. However, the number of unknowns here is small, at most four, so that such nuances are of no significance in the current application. Larger applications can benefit from this feature however.

The solution of this system of equations provides the coefficients for the linear spline that best fits the data in the least-squares sense. Any statements about the coefficients and the goodness of fit of the function to the data can only be made if we are willing to make additional assumptions about the source of the data. We then assume that the e_i are random variables such that $E(e_i) = 0$, $E(e_i^2) = \sigma^2$, $E(e_i e_j) = cov(e_i E_j) = 0$ for $i \neq j$, and that the random variable follows the normal probability distribution. Here E is the expectation operator from probability theory, σ = standard deviation of the residuals, and *cov* gives the covariance of the residuals. This is the simplest possible set of assumptions to make. The set of assumptions applies in many cases and many users of least-squares methods are only aware of this simplest set of assumptions. However, there is a whole range of extensions to this simple set of assumptions that is useful and for which methods of analysis exist. We list two of them here to give a sample and in the next section we will give an overview of the principal results of interest from what is called generalized least squares analysis.

The assumption that the standard deviation is the same for all residual values is called the assumption of homescedasticity. There are cases in which such an assumption does not make sense. For example, the value being predicted might be one for which negative values make no sense. This is the case for each of the three functions we will be using here: a negative value for a peak flow is meaningless. However, the homoscedastic assumption implies that the residual about the best-fit line is the same for flows near zero as is for the maximum flow in the sample of points. The normal distribution is unbounded and consequently the assumption can assign a significant probability to there being a negative peak flow when the flows are small. This is an acute problem for the overflow at Everson Main Street. This peak flow is zero more often than it is non-zero. We cannot eliminate the implication of negative flows if we retain the assumption that the residuals are normally distributed. Changing that assumptions results in there being no established methods for the analysis. Thus we retain the assumption that the residuals are normally distributed but we

now assume that they are heteroscedastic, that is, their standard deviation may vary with respect to one of the independent variables.

Another deviation from the simplest set of assumptions that is often useful is that the residuals have a non-zero correlation, that is, the residuals are auto-correlated. This is not used here but it could become an issue when least-squares fitting is used with time series data that is observed with far greater frequency than a series of annual flood peaks.

Least-Squares Overview

Least-squares computations using linear splines is essentially the same as for the usual multiple regression analysis. Thus from this point forward we will use the more compact matrix notation to present the results that we will use later in the analysis of the peak-flow data. In the interest of brevity, derivations will not be given. Johnston(1972) is the basis for the results given here. We will also list results from that source for heteroscedastic residuals.

All of the results are presented using matrix notation. The book by Johnston contains a chapter on matrix algebra that gives all the results needed for those not familiar with this notation. Familiarity with that notation is needed to make best use of the following results and is well worth the investment of time.

Let $\mathbf{X}$ = the $(n \times m)$ matrix that contains the linear-spline basis function values for each flood peak being used as the predictor (the independent variable). Then

$$\mathbf{X} = \begin{bmatrix} H_{1,m}(x_1) & H2,m(x_1) & \dots & H_{m,m}(x_1) \\ H_{1,m}(x_2) & H2,m(x_2) & \dots & H_{m,m}(x_2) \\ \vdots & \vdots & & \vdots \\ H_{1,m}(x_n) & H2,m(x_n) & \cdots & H_{m,m}(x_n) \end{bmatrix} \qquad \textbf{(A-10)}$$

Let $\mathbf{y}$ = the $(n \times 1)$ column vector of the flood peak being predicted (the dependent variable), let $\hat{\mathbf{a}}$ = the $(m \times 1)$ column vector of the coefficients being sought, and let $\mathbf{u}$ = the $(n \times 1)$ column vector of the residual or disturbance term. Then we have

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \hat{\mathbf{a}} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_m \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \qquad \textbf{(A-13)}$$

The solution for the coefficient vector is then

$$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad \textbf{(A-14)}$$

and the variance matrix for the coefficients (which includes the covariances in the off diagonal elements) is

$$var(\hat{\mathbf{a}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \qquad \textbf{(A-15)}$$

and we also find that

$$E(\hat{\mathbf{a}}) = \mathbf{a} \qquad \textbf{(A-16)}$$

That is, the estimators are unbiased. The prime ($'$) symbol on a matrix denotes the transpose of that matrix.

An unbiased estimator for the variance is

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}})'((\mathbf{y} - \mathbf{X}\hat{\mathbf{a}})}{n - m} \qquad \text{(A-17)}$$

The above results apply to ordinary least squares (OLS) wherein the disturbance terms is taken to be homoscedastic and uncorrelated. When we introduce heteroscedasticity we need generalized least squares (GLS).

For generalized least squares assume that

$$E(\mathbf{uu}') = \sigma^2\mathbf{\Omega} \tag{A-18}$$

whereas for OLS we assumed that

$$E(\mathbf{uu}') = \sigma^2\mathbf{I} \tag{A-19}$$

Here $\mathbf{I}$ is the identity matrix. The diagonal elements in $\mathbf{\Omega}$ may differ from 1 and the off diagonal elements may differ from zero.

The above results for OLS become as follows for GLS:

$$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y} \tag{A-20}$$

$$var(\hat{\mathbf{a}}) = \sigma^2(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1} \tag{A-21}$$

$$E(\hat{\mathbf{a}}) = \mathbf{a} \tag{A-22}$$

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}})'\mathbf{\Omega}^{-1}((\mathbf{y} - \mathbf{X}\hat{\mathbf{a}})}{n - m} \tag{A-23}$$

We assume that the contents of $\mathbf{\Omega}$ are known, assumed, or estimated.

For the assumption of heteroscedasticity, we take $\mathbf{\Omega}$ to be

$$E(\mathbf{uu}') = \sigma^2\mathbf{\Omega} = \begin{bmatrix} 1/\lambda_1 & 0 & \dots & 0 \\ 0 & 1/\lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1/\lambda_n \end{bmatrix} \tag{A24}$$

where the λ's are assumed known positive numbers but as above, σ^2 is unknown. Thus to include a heteroscadastic disturbance term we need to estimate a function that relates λ to the predictor variable. How this is done will be specific to each application. In general however, to preserve physical meaning of positive values as they approach zero, it must be the case that the variance of the disturbance term approaches zero as the predictor variable approaches zero. The approach to zero must be such that the standard deviation is only a fraction of the mean value, that is the value from the least-squares fit, so that the probability of a negative value is small. Having a standard deviation that is about $1/4$ or less of the mean value will make the probability of a negative value suitably small.

APPENDIX B: References

De Boor, C., 1978, **A Practical Guide to Splines**, Applied Mathematical Sciences 27, Springer-Verlag, New York, 392 pp.

Efron, B. and R. J. Tibshirani, 1993, **An Introduction to the Bootstrap**, Monographs on Statistics and Probability 57, Chapman and Hall, New York, 436 pp.

Johnston, J., 1972, **Econometric Methods**, McGraw-Hill, New York, 437 pp.

Lundgren, B. W., 1968, **Statistical Theory**, The MacMillan Company, New York, 521 pp.