Appendix 1

Report on Flow vs. Escapement Model and Environmental Data

Report on Flow vs. Escapement Model and Environmental Data

Gary Lorden, Ph.D.
Jay Bartroff, Ph.D.
Lordenstats

December 6, 2010

1 The Proposed Simple Regression Model of Escapement on Flow

The proposed simple regression model of SJR escapement on flow has a number of weaknesses. The following four subsections describe weaknesses our analyses have uncovered.

1.1 Evidence Against the Relationship Inferred from the Model Fit

To assess the quality and efficacy of a simple linear regression model of escapement vs. flow, we first performed statistical calculations similar to the ones done by F&G and FISHBIO on the available escapement and flow data from the period 1953-2009. Figures 1-3 show the data, model fit, residuals, and quantile-quantile (Q-Q) plots. Rudimentary straight-line modeling of this kind has been proposed as a useful description of a relationship governing these variables.

If there were such a simple relationship between these variables, that relationship should appear consistently when one partitions the 57 data points into subsets. We have examined two natural ways of doing this, breaking up the data into groups according to time periods and according to magnitudes of flow. In both cases, the results were inconsistent, calling into question the validity of the proposed simple relationship.

Figure 4 shows the same data and straight-line fit in black as in the first plot, Figure 1, but here the 1999-2009 data are shown in red, along with a red line fitted to those data by the same linear regression method. The 1999-2009 data actually has a slight *negative* correlation between escapement and flow, and hence the red line has a negative slope. Since these data for the last 11 years constitute the most recent data, it would seem that they provide an important check on the potential value of the proposed linear model in predicting a relationship between flow and escapement in future years. It has been brought to our attention that this 1999-2009 period is in fact one in which a new program of water resource management has been in effect.

Figure 5 shows the data and fitted lines when the flow range is broken into 1, 2, 3, or 4 bins of equal sizes. The fitted lines (and hence, the correlation estimates) vary from bin to bin, indicating that there is not a linear relationship that holds over the entire range of flow values. Note that one of the fits in the fourth row even has negative slope. These simple data summaries contradict a major conclusion of Newman's (2008, p. 75) hierarchical Bayesian model, which concluded that there was

a strong positive correlation between escapement and flow over all flow ranges. All of these fits suffer from low R^2 values: The ten plots in Figure 5 have R^2 values in the range [.0043, .41].

Additional doubts about the validity and value of the linear-fit model arose when we noticed that a small number of data points overly influence and inflate the linear relationship between escapement and flow. It is well known that simple linear regression is highly non-robust and can easily be "fooled" by a small number of data points. (See also the discussion of outliers in Section 1.2). When a small number of data points are overly influential, one would expect to see inconsistencies between linear fits made using random subsets of the data points, since the highly influential points will affect some fits and not others. This behavior is observed in Figure 6, where the data were divided into four subsets of equal size at random; each row represents an independent realization of this process. Note that the model fits vary widely and a negative correlation is even found in one subset in the first and fourth realizations.

1.2 Violations of Model Assumptions

Returning to Figures 1-3, there are several fundamental assumptions of the regression model that seem to be violated by the data.

The model assumes that the observations of the y variable, here escapement, is normally distributed. When this holds, the shape of points in the scatterplot is roughly "football" shaped along the fit line, which is not evident in Figure 1. Another standard way to assess this normality assumption is to examine the Q-Q plot of the residuals, which compares their distribution to the assumed normal distribution. If normality holds, then the points in the Q-Q plot should lie close to the dotted line in the third plot. The fact that they are not close in Figure 3 is evidence of non-normality.

Another assumption of the model is that observations of the y variable are subject to random variations whose scale is constant and which average out to zero. When this holds, the residual plot should appear as roughly a uniform cloud of points, symmetric around the horizontal dotted line. That is not the case in Figure 2, which on the contrary indicates both a bias (non-zero average) and a non-constant scale of variations. Moreover, the numbered points in Figures 2 and 3 are outliers – points that represent deviations from the linear model that are too large to be consistent with that model.

Finally, we note that the model fit in Figure 1 has an R^2 value of .27. R^2, the coefficient of determination, is the square of the correlation coefficient and thus takes values between 0 and 1. R^2 is a measure of goodness of fit of the model and, more specifically, is the fraction of the variation in the y data that is considered to be "explained by the linear fit" on the x data. A value of .27 is generally considered quite low and indicates that this proposed model does not capture a meaningful relationship between the two variables.

1.3 Lack of Predictive Power

As would be expected from its poor fit of the available data, particularly in the most recent time period, the linear model seems to have very little predictive power. A standard way to assess the usefulness of a fitted model is to calculate and examine so-called "prediction intervals" computed from it. These are confidence intervals for future observations, calculated so that they should be correct at some confidence level. Table 1 contains prediction intervals calculated at the 95% confidence level from the linear model fit to the pre-1999 data and compared with the actual 1999-

2009 data. These prediction intervals are extremely wide – too wide to have any useful predictive power. For example, with the exception of one year, the upper prediction for each year is larger than any escapement measurement made in the entire 1952-2009 data set. (The largest escapement measurement was 80,000 while the 2004 upper prediction was 79,324). In spite of their extreme width, the prediction intervals for two years – 2007 and 2008 – do not in fact contain the actual escapements observed in those years. Figure 7 contains a graphical representation of the prediction intervals and the actual 1999-2009 observations.

Table 1: Predictions and 95% confidence prediction intervals for 1999-2009. Values in bold are violated by observed escapements.

Year	Vernalis flow (avg. over daily values 2.5 yrs prior)	Escapement	Predicted Escapement	Lower Prediction	Upper Prediction
2007	10597	1241	20382	**1961**	211882
2009	2829	1323	8244	811	83776
2008	25545	2229	37252	**3424**	405258
2006	2476	4169	7524	740	76522
2005	2707	6376	7999	787	81304
2004	2611	10319	7802	767	79324
2003	3185	11144	8941	880	90837
1999	4575	17347	11460	1126	116672
2002	4811	25666	11862	1164	120839
2001	5364	26659	12781	1253	130412
2000	18665	39447	30043	2814	320704

1.4 Inferential Problems

Because linear regression analysis is so widely used, a number of mistakes and fallacies that occur frequently in their interpretation are well known. Two that are relevant here are the Ecological Fallacy and the Correlation/Causation Fallacy.

The Ecological Fallacy refers to making inferences at the individual level based on regression analysis performed at a subgroup level. This typically occurs when data are averaged or combined over a subgroup before fitting a regression model. This can lead to fallacious conclusions because averaging reduces variation and therefore can falsely inflate the strength of linear relationship, or make one appear when in fact there is a more complex relationship– or no relationship at all. The current proposed model is in danger of this because the flow data are averaged over two months before performing the regression fit, a very crude form of data reduction in this setting that suppresses a large source of natural variability. The proposers of this model have the responsibility to show that the variation lost in averaging does not affect the inferred relationship.

Another relevant fallacy is the Correlation/Causation Fallacy, in which an estimated correlation in a regression analysis is mistaken for causation– i.e. that the variables have a genuine cause-and-effect relationship. Although a robust model fit can indicate a possibility of causation, that is not the case for the sort of linear model proposed between flow and escapement, which is highly non-robust in light of the inconsistencies cited in Section 1.1 and the violations of model assumptions

cited in Section 1.2. The proposers have not shown that the estimated correlation corresponds with a causal relationship.

2 Environmental data

Figures 8-13 contain boxplots of the available environmental data, before any averaging occurs. Figure 14 contains scatterplots of this data, on the log scale, after being averaged over the period April 15 - June 15 for each year of available data. In these scatterplots the escapement data were paired with the corresponding variable from two years prior. The temperature data in Figure 14 is hourly, back to 1999, and was obtained from the California Department of Water Resources webpage.

Other than Vernalis flow, there is an overall scarcity of environmental data available, and what exists is further compressed by the yearly averaging. We suspect that this is one reason for the focus on Vernalis flow by F&G as an "explanatory" variable for escapement. For example, it is clear that water temperature may have a large affect on escapement. However, hourly water temperature data is available only back to 1999. After averaging and matching up with escapement data two years later, this results in only nine data points corresponding with the 2001-2009 escapement data. This is a small amount of data to develop any sort of meaningful model. Note also that even if other environmental variables had more data available, any model that includes temperature would be restricted to using only these nine years.

We fit a multiple regression model of y = SJR escapement (on the logarithmic scale) on the variables

$$\begin{aligned} x_1 &= \text{Vernalis temperature} \\ x_2 &= \text{Mossdale dissolved oxygen} \\ x_3 &= \text{Mossdale temperature} \\ x_4 &= \text{CVP exports} \\ x_5 &= \text{SWP exports,} \end{aligned}$$

all on the logarithmic scale, depicted in Figure 14. Quadratic terms were included for the Vernalis and Mossdale temperature variables since it is expected that extreme temperatures, both low and high, tend to reduce escapement. The least squares model fit is given by

$$y = -14092.5 + 777.7x_1 - 113.0x_1^2 + 14.2x_2 + 5909.3x_3 - 681.9x_3^2 - 4.2x_4 + 4.6x_5,$$

and has an R^2 value of .6. Though the small number of data points likely causes this R^2 value to be somewhat inflated, this result suggests that one might be able to model escapement in a statistically useful way using multiple variables in addition to flow.

3 F&G's "Plug and Play" Model Components

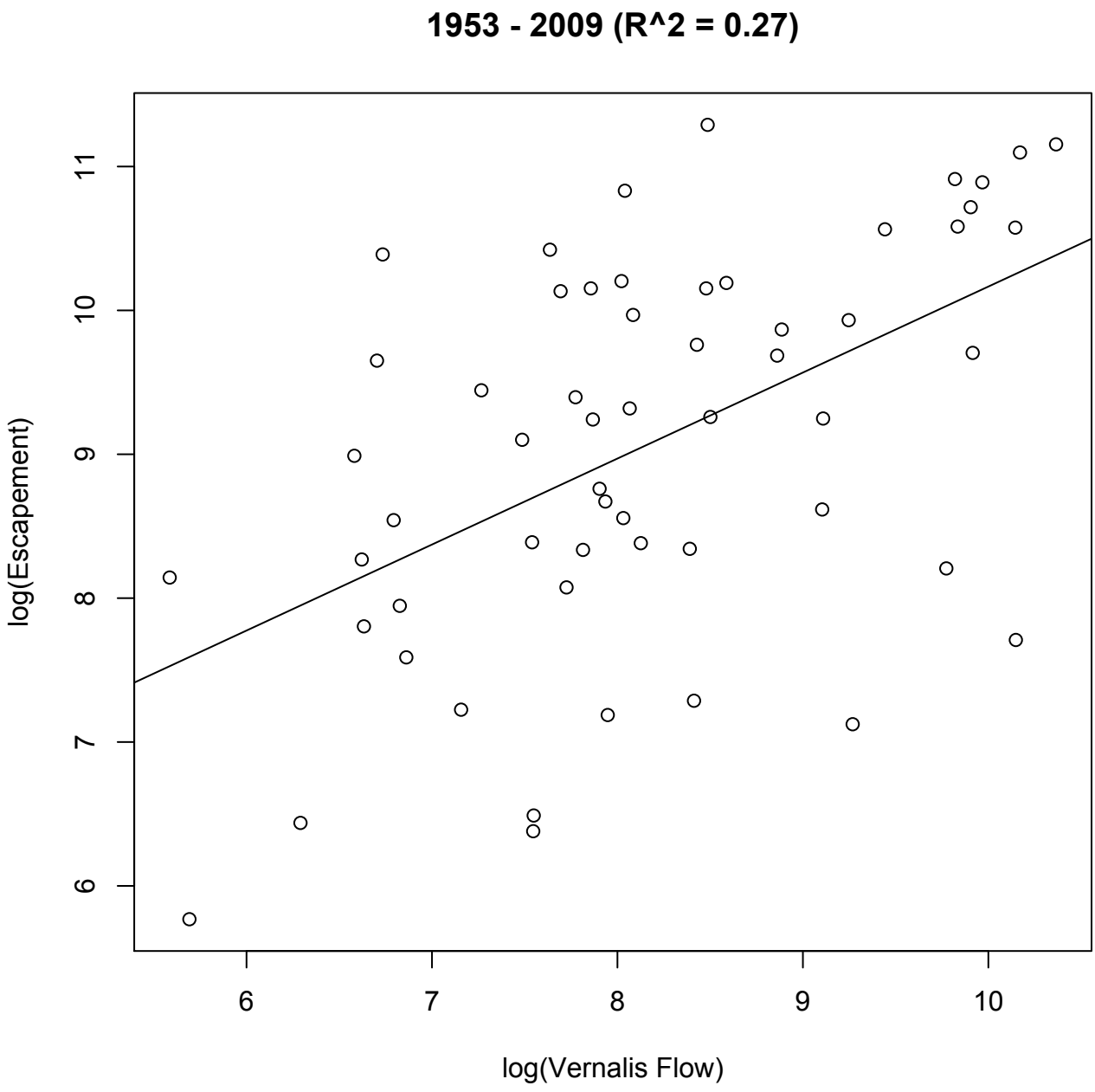

Figure 1: Data and linear model fit for flow versus escapement data, 1953-2009, on the logarithmic scale.

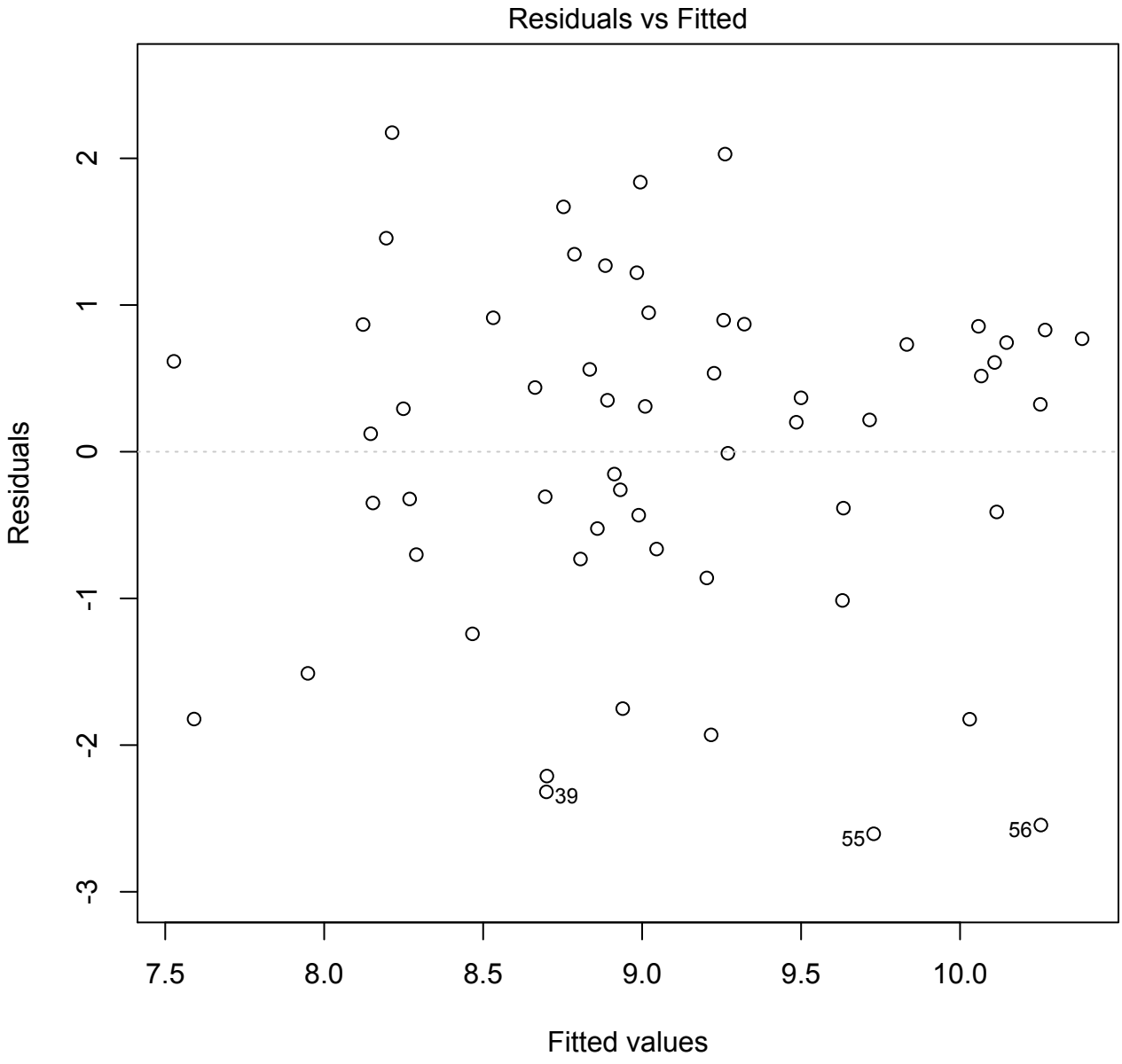

Figure 2: Residuals for flow versus escapement data, 1953-2009, on the logarithmic scale.

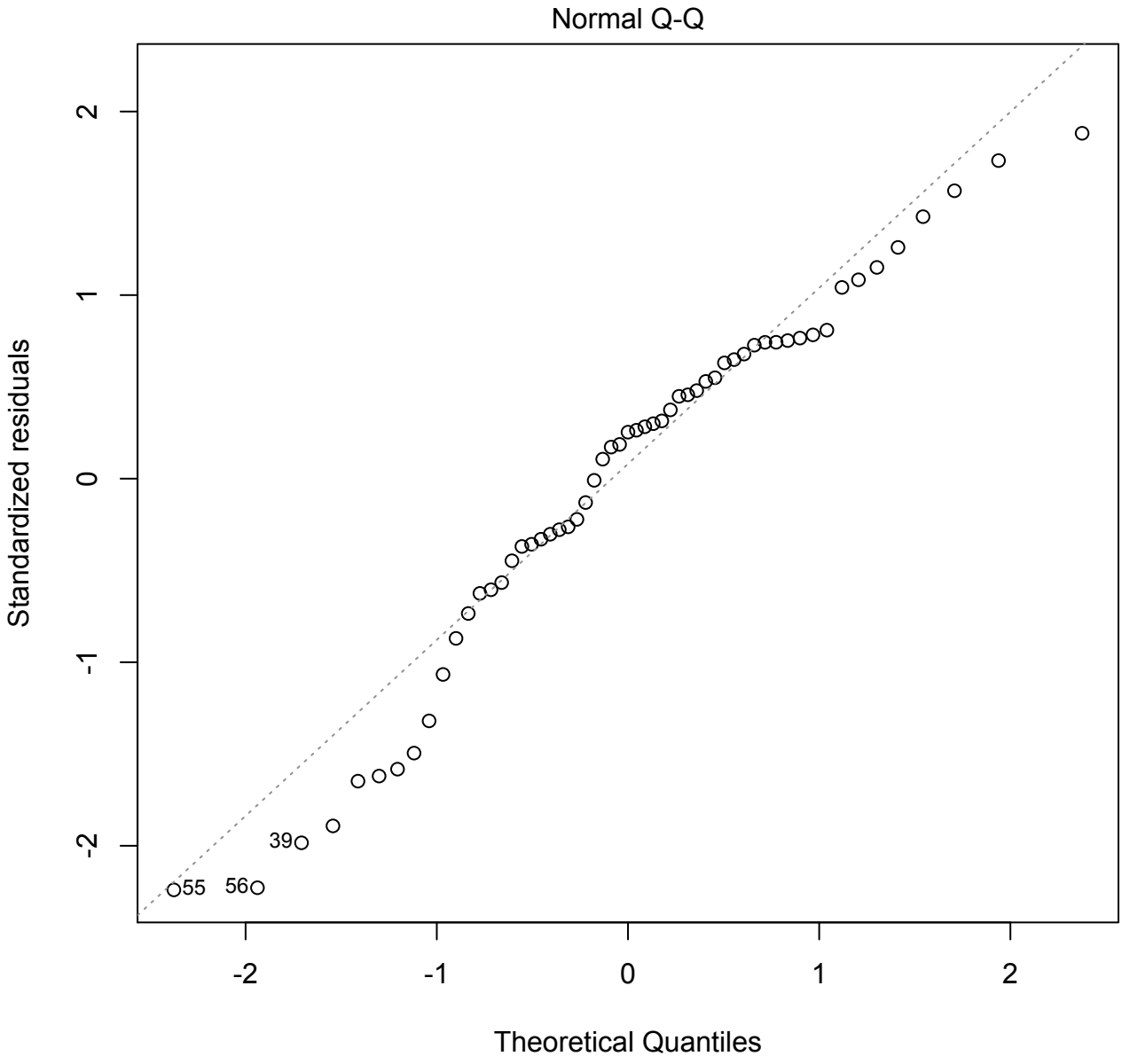

Figure 3: Q-Q plot for flow versus escapement data, 1953-2009, on the logarithmic scale.

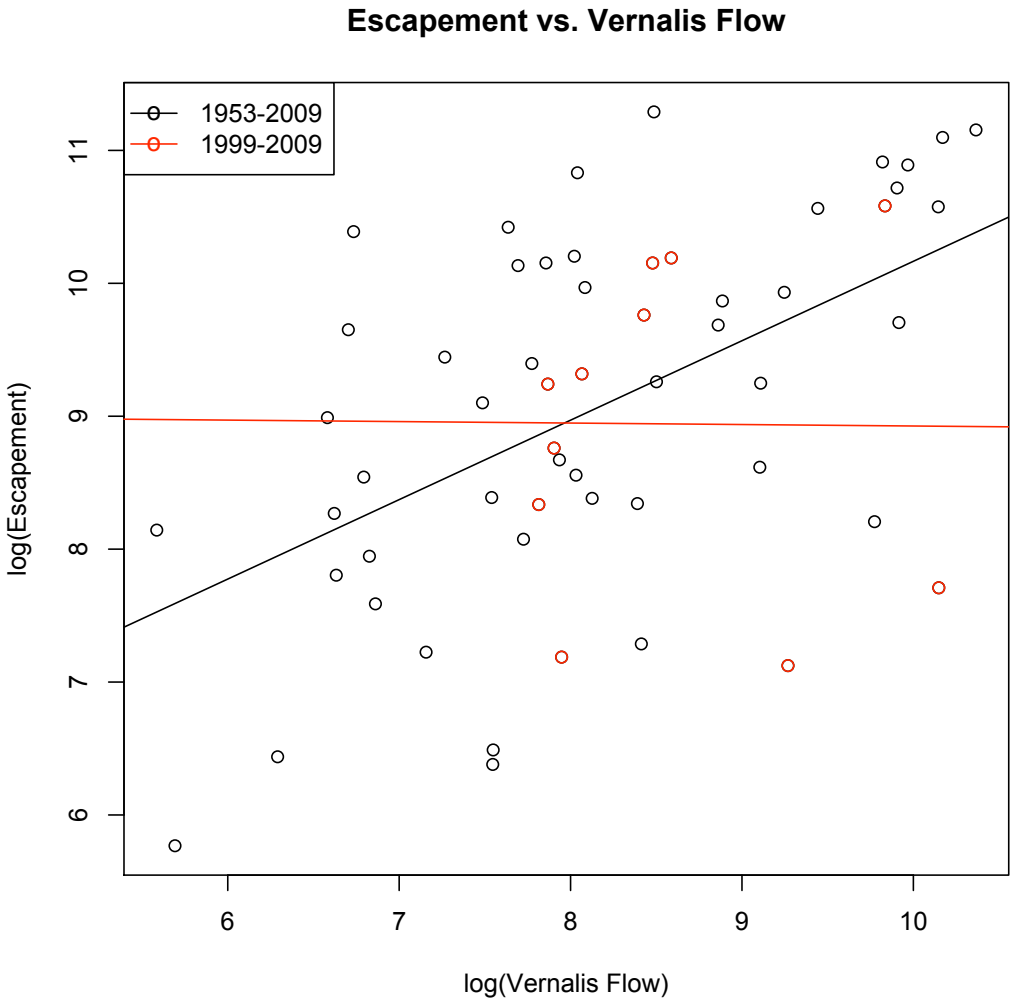

Figure 4: Data and linear model fits for 1953-2009 and 1999-2009 data.

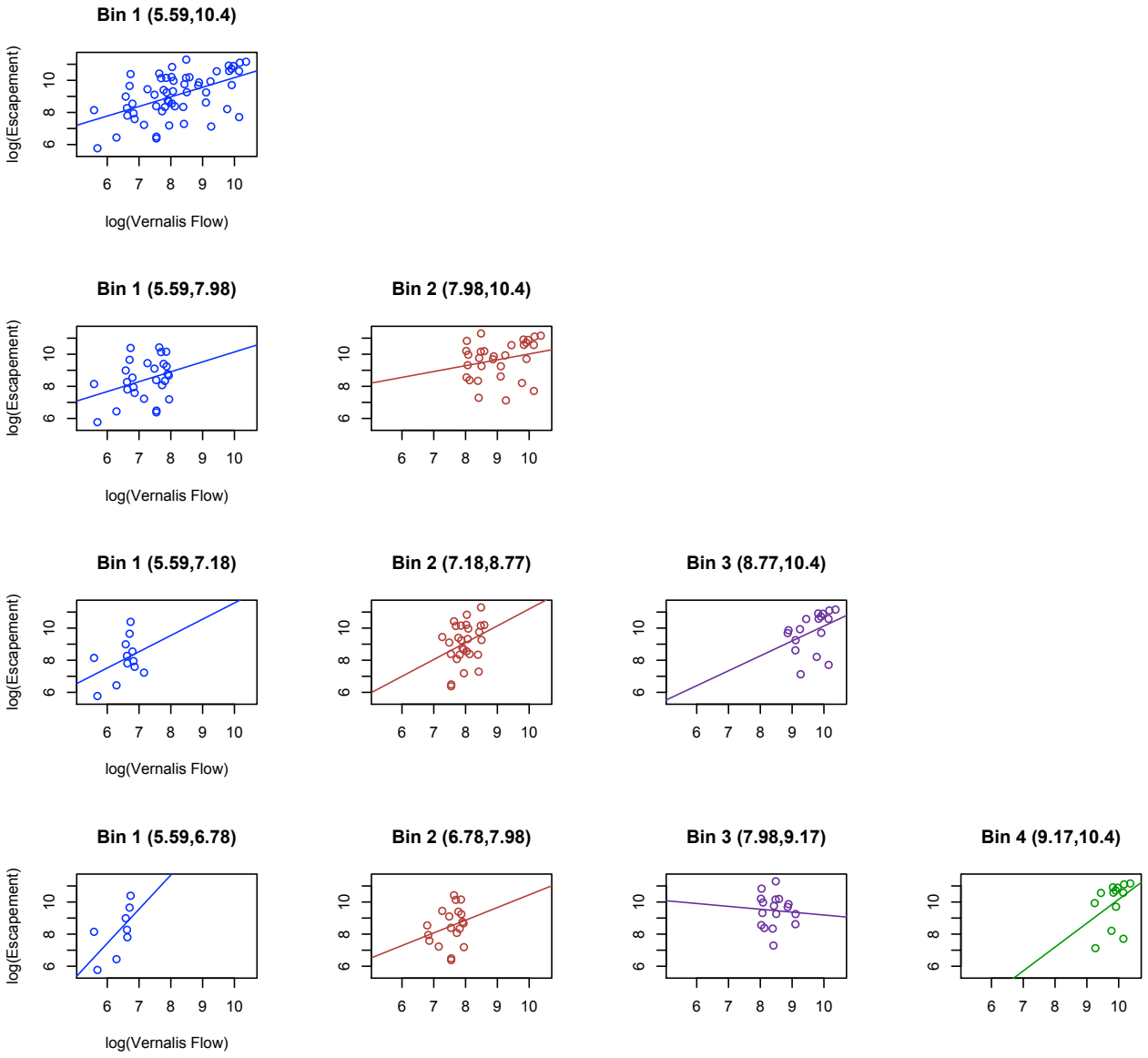

Figure 5: Data and linear model fits for 1953-2009 data when flow range is divided into 1-4 equally sized bins (rows 1-4). The R^2 values for these ten fits are all low, in the range $[.0043, .41]$.

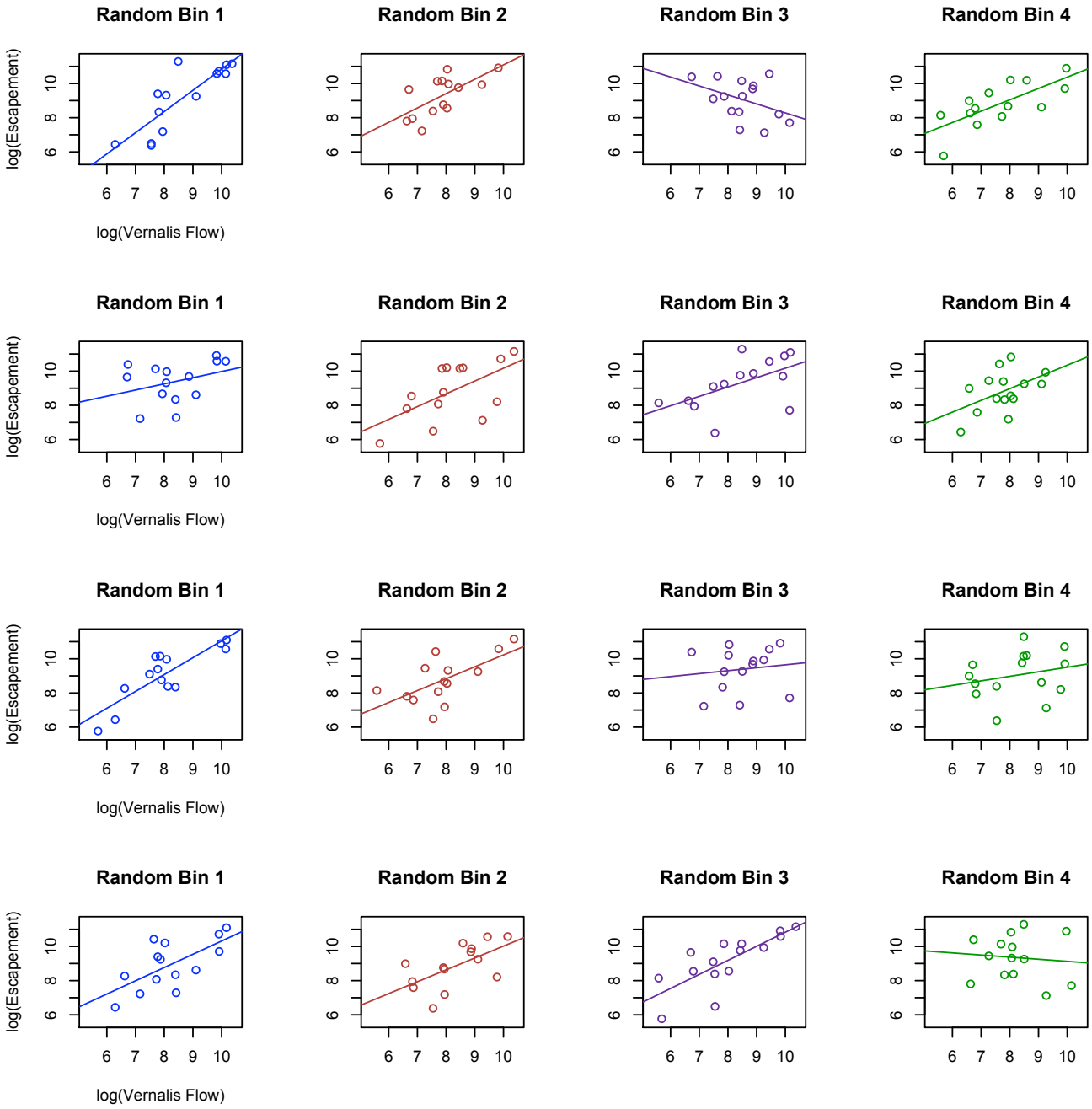

Figure 6: Data and linear model fits for 1953 - 2009 data divided into four subsets at random. Each row is an independent realization.

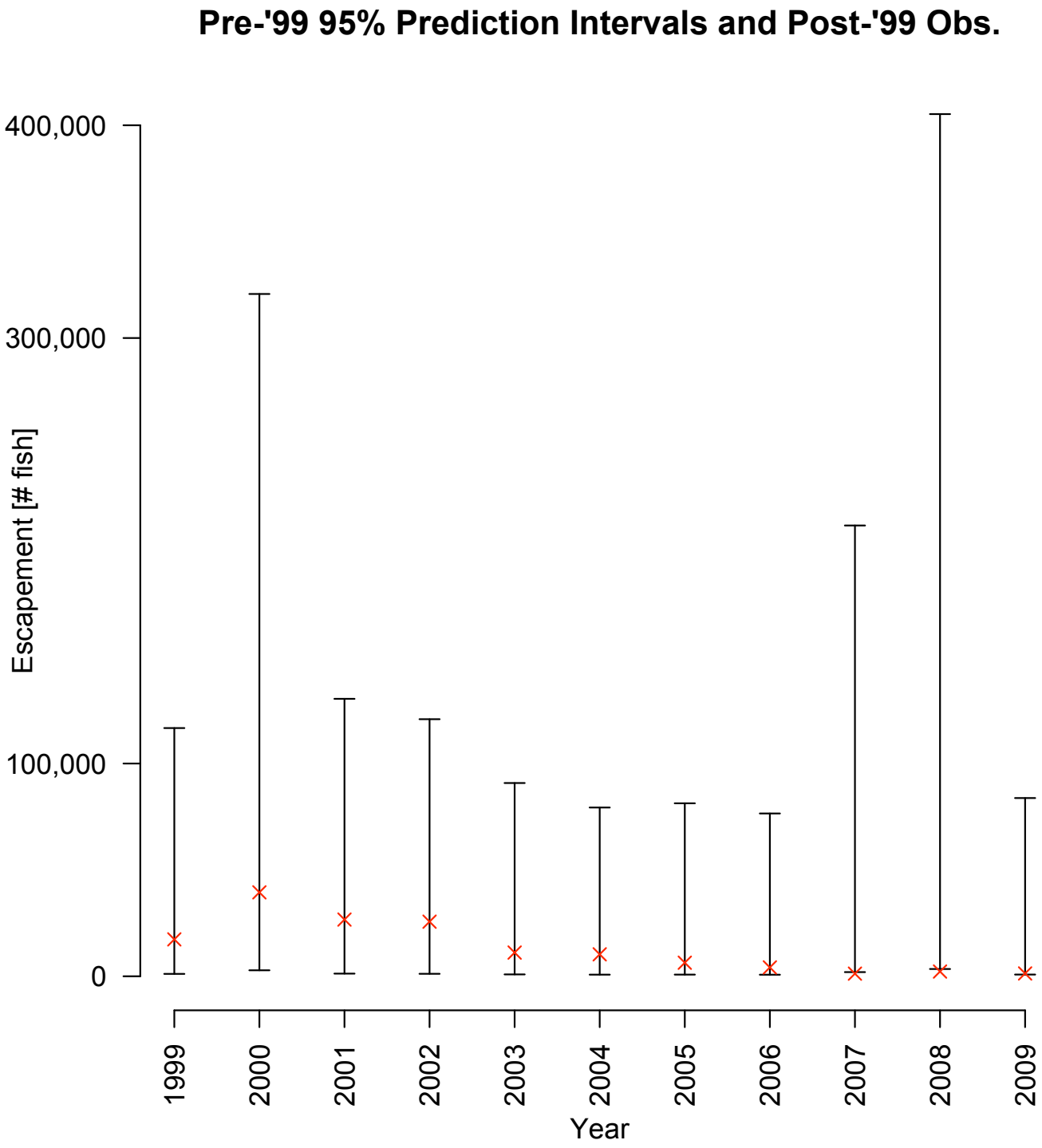

Figure 7: 95% confidence prediction intervals from 1952-1998 model for the 1999-2009 data.

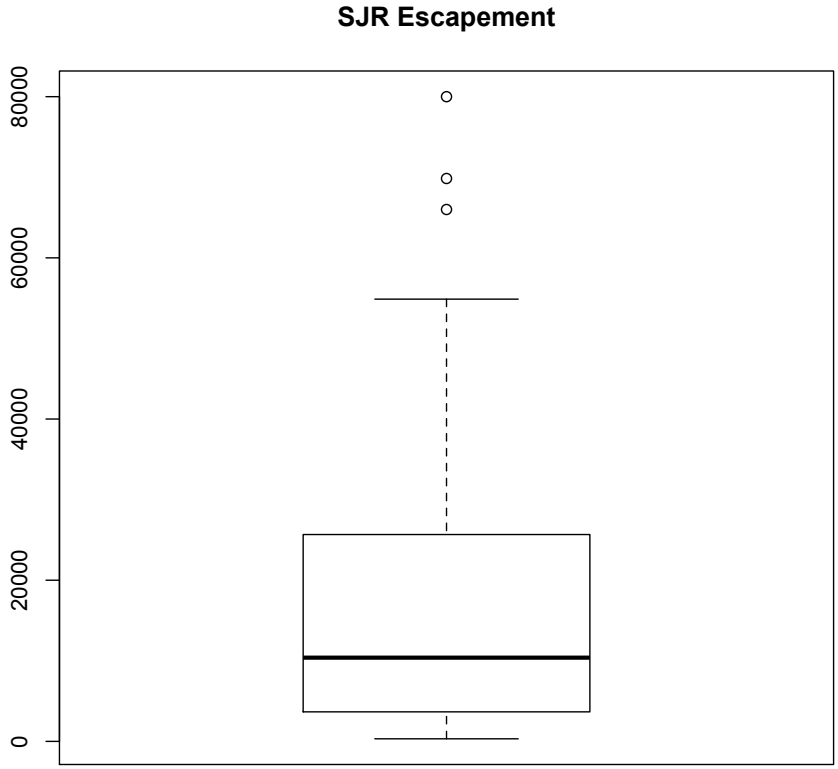

Figure 8: Boxplot of SJR escapement data, 1952-2009.

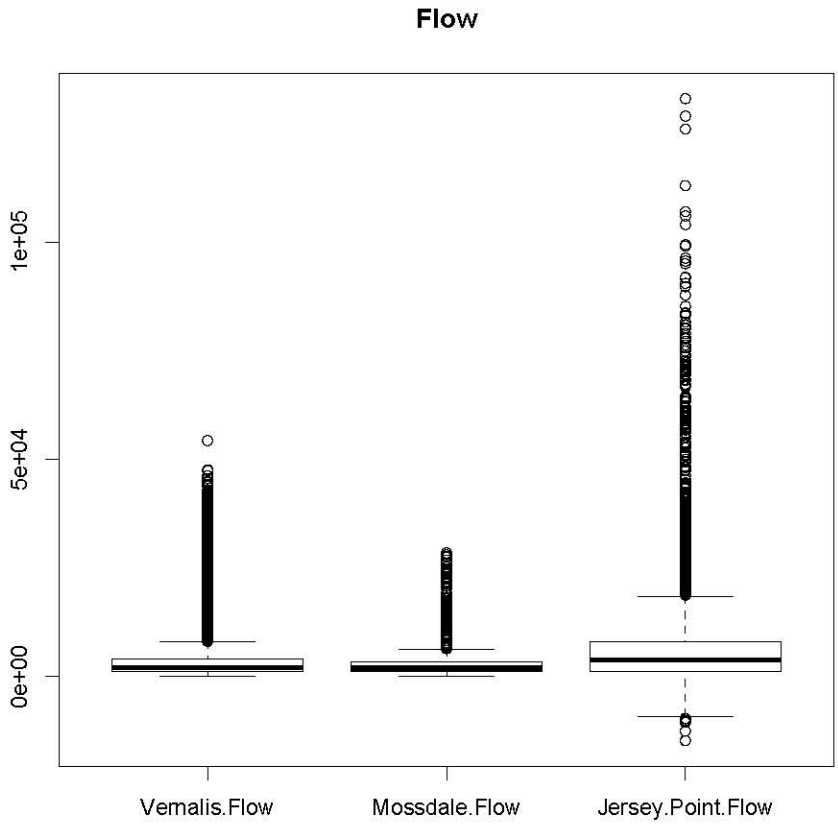

Figure 9: Boxplots of daily flow data.

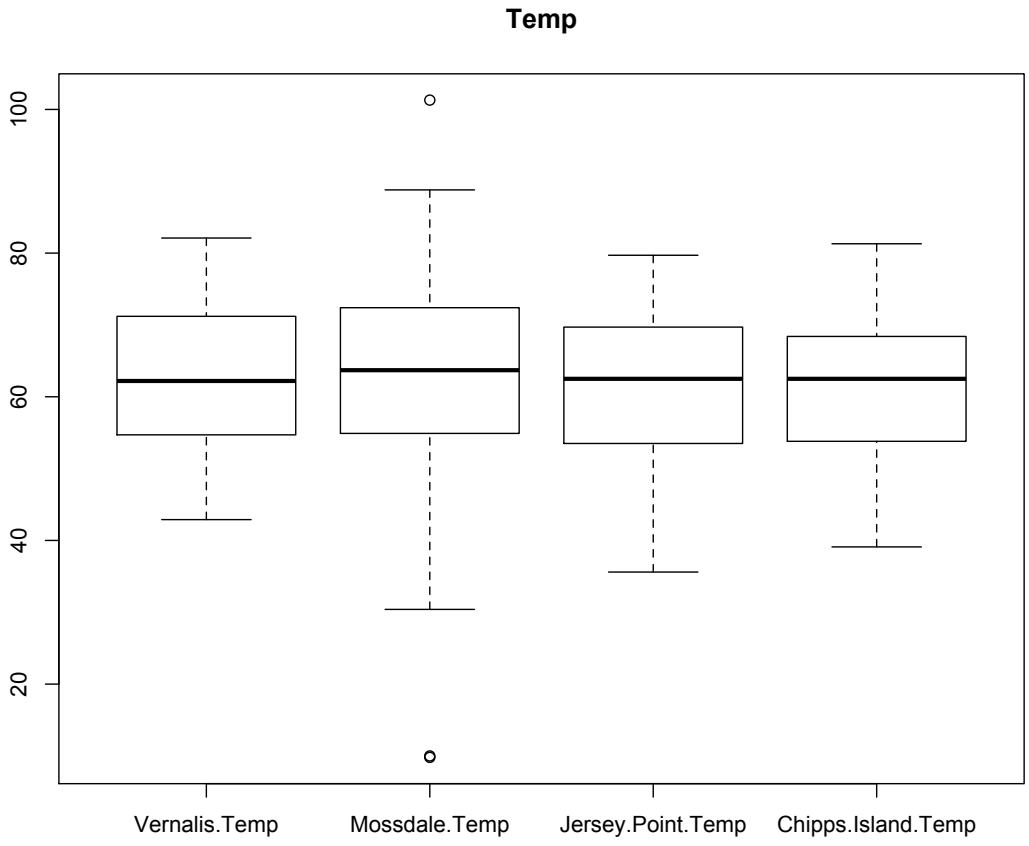

Figure 10: Boxplot of daily temperature data.

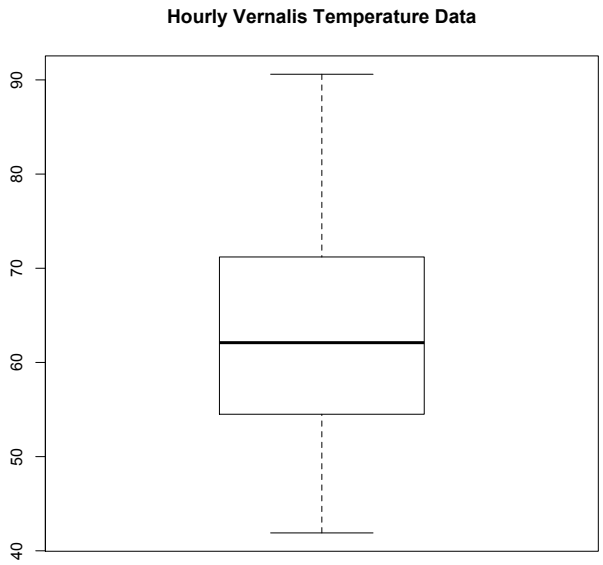

Figure 11: Boxplot of hourly Vernalis temperature data.

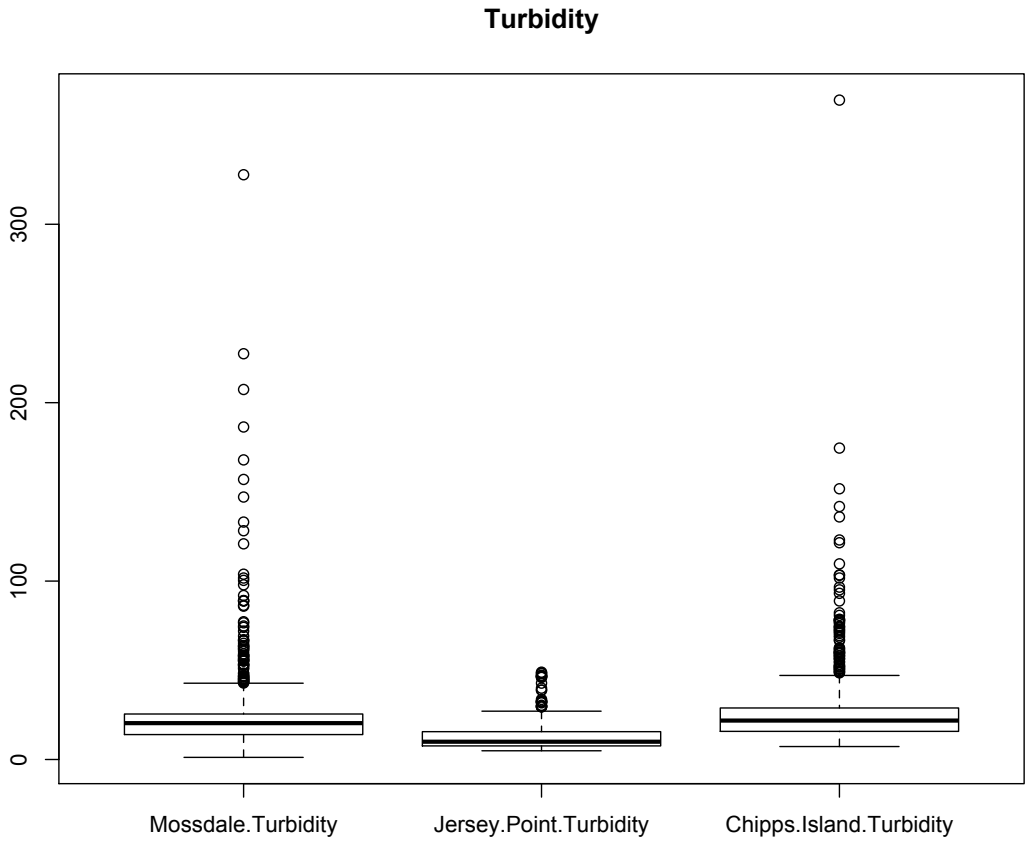

Figure 12: Boxplot of turbidity data.

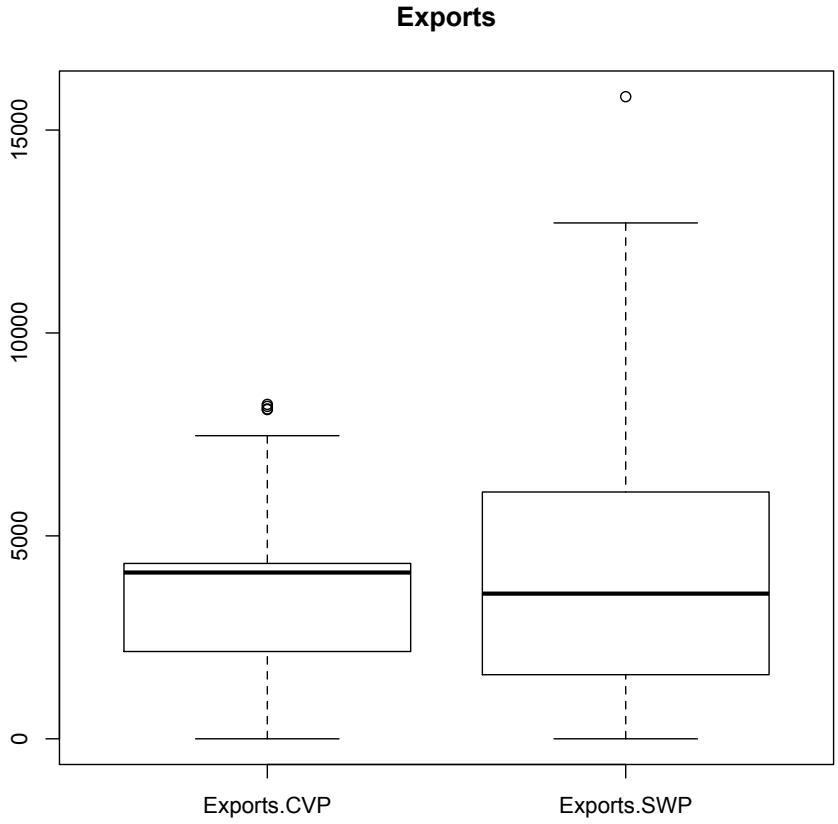

Figure 13: Boxplots of exports data.

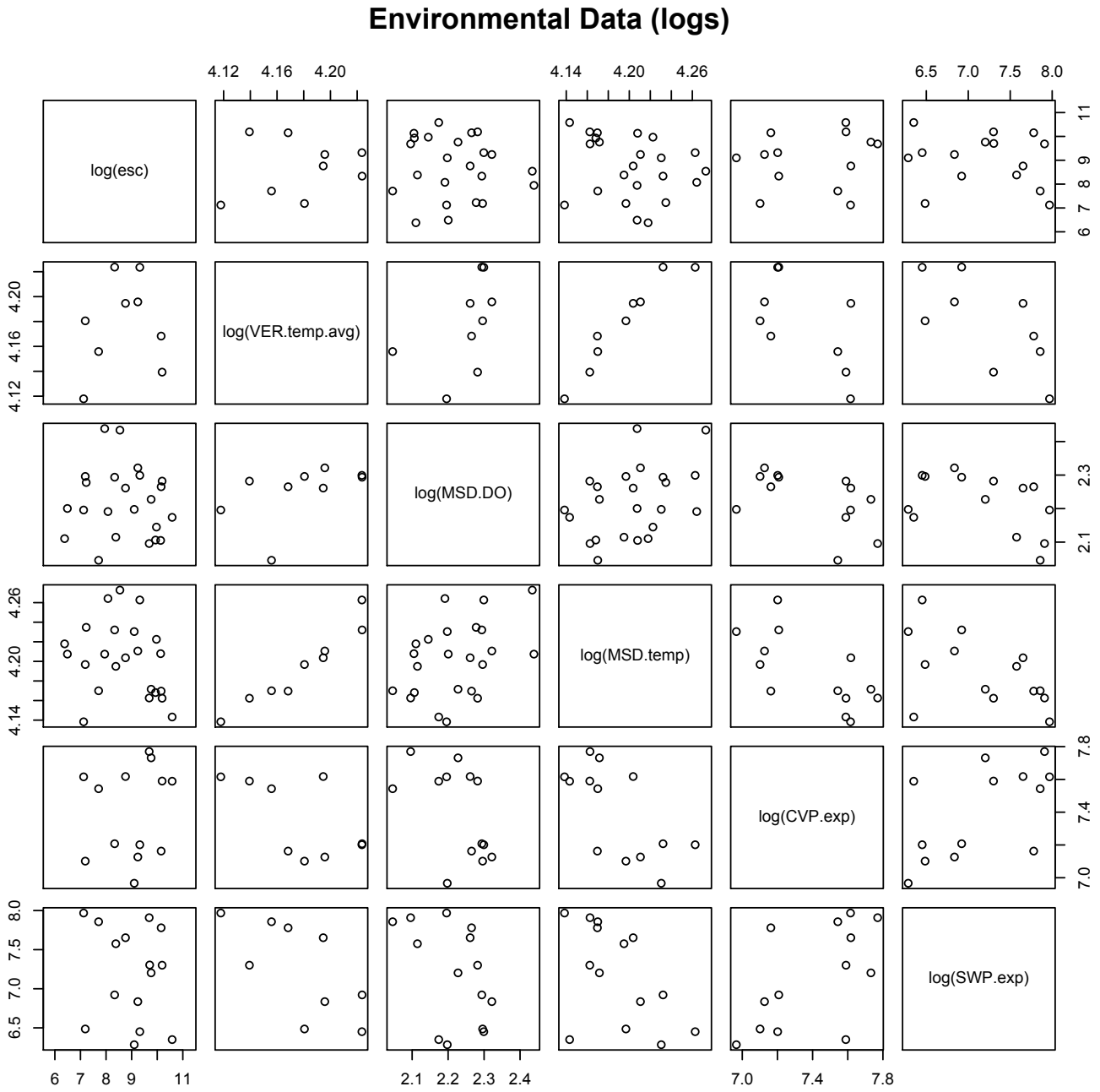

Figure 14: Scatterplots of SJR escapement and April 15 - June 15 averages for other environmental data, on the log scale.