STATISTICAL MODELLING OF TEMPORARY STREAM FLOW IN CANADIAN PRAIRIE PROVINCES

by

Hua Zheng

B.Sc., University of Science & Technology of China, 2007

A Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Sciences

SIMON FRASER UNIVERSITY
Summer 2012

APPROVAL

Name: Hua Zheng

Degree: Master of Science

Title of Thesis: Statistical Modelling of Temporary Stream Flow in Canadian Prairie Provinces

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Charmaine Dean, Adjunct Professor, Statistics
Simon Fraser University
Senior Supervisor

Dr. Joan Hu, Professor, Statistics
Simon Fraser University
Supervisor

Dr. Melody Ghahramani,
Associate Professor
University of Winnipeg
External Examiner

Date Approved: 20 Aug 2012

Partial Copyright Licence

Abstract

Accurate forecasting of streamflow is of vital importance in semi-arid regions in order to meet the needs of humans, such as agriculture, and for wildlife. It is also of considerable interest for predicting streamflow for ungauged basins and for detecting change due to landuse or climate variations. Daily streamflows in semi-arid and arid regions are characterized by zero-inflation, seasonality, autoregression and extreme events such as floods and droughts. Analyses at the level of daily data for intermittent streams are problematic because of the preponderance of zero flows. Basic modelling approaches are often inappropriate when many zero flow events are present; approaches need to be modified to allow greater flexibility in incorporating zeros than is possible with traditional methods. This project discusses the utility of spline compartment models for analysis of data from intermittent streams, whereby the log-odds of the probability of a non-zero flow day, as well as the logarithm of non-zero flow rate can be studied. These models permit handling of large numbers of zero-flow days; the use of splines and other smoothers have the benefit that they permit a wide range of distributional shapes to be fitted. The models are illustrated for ten streams in the Canadian Prairie Provinces.

Acknowledgments

First I would like to express my deepest gratitude and appreciation to my senior supervisor, Dr. Charmaine Dean for the opportunity to be her student for the past three years. Without her patience, guidance and encouragement, I don't think I would have reached this far. I appreciate her endless support and generous help in all aspects of my completing this degree.

Many thanks to Paul Whitfield for all his advice and help on the hydrology aspects of this project. I am grateful to Paul for introducing me to such an interesting project. I also would like to thank Melody Ghahramani for her advice and suggestions related to this project.

Thanks to Dr. Melody Ghahramani, Dr. Joan Hu and Dr. Richard Lockhart for reading the project and serving as my committee members. Special thanks should be given to my fellow graduates in the department, Yu Xia, Crystal Li, Huijing Wang, Jing Cai, Harlan Campbell, Alisha Albert-Green, Darby Thompson, Cindy Feng, Elizabeth Juarez-Colunga, Ryan Leikevitz, Harsha Perera, Ricky Tang for their friendship and help. I also want to thank all the faculty and staff members in Statistics and Acturial Science Department for providing us such a family-like environment.

Finally, I want to thank my family for their support, love and understanding!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Hydrologists have developed an impressive array of tools for the analysis of flow data. Flood and drought frequency analyses are one focus when investigating flows; these estimate the level of the T-year event based on a probability distribution model postulated for annual extremes. While an analysis based on annual maxima/minima avoids the need to model seasonal variation and serial correlation, as is present in daily data, such an approach requires long time series. As well, these tools are generally restricted to the analysis and interpretation of continuous data, using a variety of continuous distributions, such as the lognormal. Methods for the hydrological analysis of flow records that contain many days with no flow, i.e. statistically are 'zero-rich', are relatively rare. In this case, simple modifications, such as adding a small amount to the zero values, are inappropriate; the data are really of mixed types: discrete (the zeros) and continuous (positive flow) and models should reflect these two compartments of the data. Importantly, since temporary streams (intermittent, ephemeral, and episodic) are very sensitive to changes in landuse and climate (Buttle et al.), robust, flexible, methods for estimating hydrological attributes and the detection of change (climate, landuse, etc.) are required, and should be considered in the specification of the models.

Cataldo et al. (2004) considered how simple regression equations could be used for predicting flows or flow losses as a function of covariates such as antecedent conditions and basin properties. Regression equations involve highly non-linear relationships between rainfall and runoff, and there are generally few available data for deriving and testing these relationships. This is particularly true for the Canadian Prairies where flow and precipitation networks have developed independent of each other. Cataldo et al. (2004) noted that while

such models are straightforward to implement, they lack direct connections to the specific physical processes governing transmission losses, making them difficult to apply at sites other than the one(s) for which they were developed. In particular, while high streamflow periods may be accurately predicted by such regression on convective rainfall events in the summer, this would not be the case necessarily for the springtime, when flow is driven by snowmelt. As well, Cataldo et al. (2004) and many others use log-transformed data, and assume normality; this may well be suitable for many perennial streams, but not for intermittent streams because of the large numbers of days with zero flow. When working with log-transformed data containing many zeros, the common practice of adding a small amount to the zeros before taking logarithms is inappropriate; developing a modeling a modeling strategy that acknowledges the zeros is more useful and can also provide additional information on when the zeros tend to occur, the seasonality of zeros, etc. Since rainfall that may generate flow is stochastic, models that link precipitation and the resulting flow response have often been used in simulations.

Lee (1975) used a Weibull-based approach to describe intermittent and ephemeral flow events, focusing on attributes that are important to temporary streams: flow duration, inter-event time, and total flow volume; however flow in ephemeral streams was synthesized only at a monthly level (Lee, 1975). Similarly, Srikanthan and McMahon (1980) considered several stochastic procedures for generating monthly flows in ephemeral streams. The major problem they encountered was modelling of zero flows. Chebaane et al. (1995) presented a series of models with the capability of handling the high percentage of zero flows in intermittent streams, reproducing the percentage of zero flows in each month, the monthly mean and variance, and the month-to-month correlation of the intermittent flows. However, monthly models are overly smooth in comparison to the time step of the actual processes.

Aksoy and Bayazit (2000) instead explored generating synthetic daily flows for intermittent streams. Their model considers: (1) days on which flow occurs; (2) days on which a flow increment occurs, estimated using a three-state Markov Chain; (3) magnitude of the increment, assuming a gamma distribution; and (4) the magnitude of the flow decrement when flow is reduced, using an exponential equation. Averages from ten simulations of the fitted model preserved the short-term characteristics of the data such as hydrograph shape and peak, in addition to long-term attributes (mean, variance, covariance and zero flow percentage). However, parameters were again estimated on a monthly basis and, even so, there were a large number of model parameters to be estimated (over 100). Furthermore,

low flow events were not well estimated and the models were developed for an area where rainfall generates streamflow rather than an area that exhibits spring snowmelt.

One aspect regarding Prairie streams is that they are frequently excluded from studies, particularly those that consider climate (e.g. Whitfield and Cannon, 2000) and land-use impacts because these gauges are seasonally operated and subject to anthropogenic influence and/or regulation. Monk et al. (2011) suggest that streams from the Prairie region constitute a major gap in their and other trend studies, as Prairie streams may represent a distinct hydrologic regime. Similarly, the exclusion of these streams from many studies hinders efforts to reasonably make predictions for ungauged basins.

The objective of the present study is to develop a flexible statistical conceptual approach that could be built upon to support in depth modelling of flows in temporary streams with an aim that this approach can be used in the future to validate process models. In much of the Canadian Prairies there are two distinct processes that may generate streamflow; in the spring, much of the flow may arise from the regional melting of snow, while in the summer, convective rainstorms may produce flow in a localized area. Both of these processes have the potential to be hydrologically important with respect to floods, and water availability. Similarly, the absence of flow is of considerable interest.

The aim here is to (i) assess the evidence in the data for the main runoff generation processes for modelling Prairie (intermittent) flow, (ii) develop a model which accommodates the mixed nature of the data, especially the large numbers of daily zero flows, (iii) illustrate the utility of smoothers as robust, flexible models, and, (iv) apply the methods for ten case studies. A multistate model is used in an exploratory manner to study the 'state' of flow and the transition probabilities to an alternate state, for instance the probabilities that a day with zero flow is followed by a day with or without flow, or, when flow exists, that it increases or decreases. The main contribution of this project is that flow volume, or discharge (m^3/sec), is modelled through a two-compartment approach which accounts for zero-flows in a logistic framework while modelling positive flows using a logarithmic distribution model, with both of these compartments estimated and fitted separately for snowmelt and summer convective periods. Though the work presented here is in the realm of exploratory data analysis, it can be used as a foundation for more sophisticated modelling as discussed in the final chapter of the project. Such models would provide quantitative methods for both prediction of flows in ungauged basins and in detecting change due to changes in landuse or climate.

The project is organized as follows. We present the results of our exploratory analyses in Chapter 2. Chapter 3 reviews the theory of Generalized Additive Models (GAMs) and presents the GAMs to be considered in our stream flow analyses. In Chapter 4, we present results from fitting our GAMs and compare the fits of the various models considered. We summarize and discuss our future work in Chapter 5.

Chapter 2

Exploratory Data Analysis

2.1 Flow data

Using Environment Canada's Data Explorer, it was determined that there are 85 seasonal and continuously measured stations in Manitoba, Saskatchewan, and Alberta where the reported daily flow includes zero flows. A number of these sites are members of the Canadian Reference Hydrologic Basin Network (RHBN) that are considered suitable for studies of climate and hydrology relationships (Brimley et al., 1999); however additional sites were also considered. From this list, all stations with less than 20 years of record were removed. Many of the remaining streams might contain relatively small regulation structures, the effects of which are discussed by McGee et al. This list of 62 stations was further investigated; stations which were considered regulated, stations where data quality was considered less than good, and stations not within the Prairie ecosystem, were removed. This resulted in 24 unregulated stations, with 20 or more years of record and containing periods of observed zero-flow. The data for these stations were extracted from the 2010 version of HYDAT (Environment Canada, 2011). In the present work, the ten stations listed in Table 2.1 and displayed in Figure 2.1 were chosen to be a meaningful illustrative sample of these types of Prairie streams. These data, and the models we develop, are for the period from March to October as this is the period sampled for seasonal hydrometric stations. The data are derived from a continuous recording of the hydrometric response termed stage; this represents water level. A rating curve is used to convert the stage values to daily mean flow. At gauging stations, the stream discharge is measured across the stream channel with a flow meter. The relationship for stream stage vs. stream discharge is unique. A rating

curve is a graphical representation which defines this relationship. If the stage of the river is measured, then discharge is calculated by means of the rating curve. Through this process the continuous record of water level can be converted to mean daily flow in m3/s.

Station Number	Station Name	Province	Latitude	Longitude	Drainage Area (km^2)
05AF010	MANYBERRIES CREEK AT BRODIN'S FARM	AB	49°21′2″ N	110°43′3″ W	338
05AH001	BOXELDER CREEK NEAR WALSH	SK	49°57′42″ N	109°59′21″ W	321
05JF008	FAHLMAN CREEK NEAR DAVIN	SK	50°22′10″ N	104°11′30″ W	15
05JL002	INDIANHEAD CREEK NEAR INDIAN HEAD	SK	50°38′42″ N	103°36′14″ W	327
05NA003	LONG CREEK AT WESTERN CROSSING OF INTERNATIONAL BOUNDARY	SK	49°0′1″ N	103°21′8″ W	3210
05NF002	ANTLER RIVER NEAR MELITA	MB	49°3′26″ N	101°2′57″ W	3220
05NF008	GRAHAM CREEK NEAR MELITA	MB	49°15′45″ N	100°59′53″ W	741
05NG007	PLUM CREEK NEAR SOURIS	MB	49°37′33″ N	100°18′12″ W	5420
11AB075	LYONS CREEK AT INTERNATIONAL BOUNDARY	SK	49°0′17″ N	109°13′48″ W	174
11AE009	LYONS CREEK AT INTERNATIONAL BOUNDARY	AB	48°58′10″ N	106°50′20″ W	837

Table 2.1: Attributes of Water Survey of Canada hydrometric stations considered in this study.

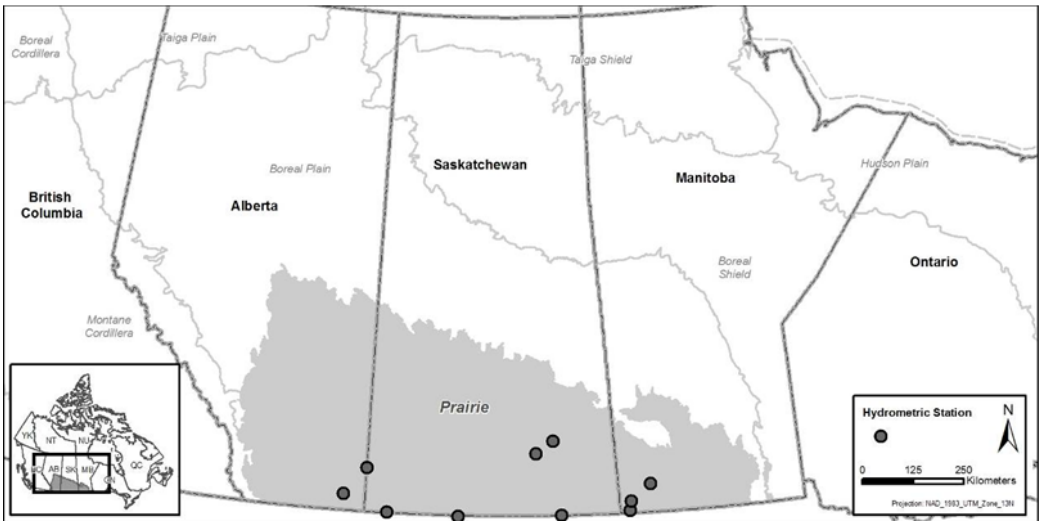

Figure 2.1: The location of the ten hydrometric stations used in this study

2.2 Summary of Ten Illustrative Stations

The total number of observations at each of the ten stations, along with the fraction of no flow days is presented in Table 2.2. Note that in Table 2.2, the sum of missing and non-missing values calculated from the second and third column is 12784, namely, the total number of days from year 1959 to 1993. Almost all the missing records are from January to February and November to December as can be seen from columns 3 and 4 in Table 2.2. The time period from March to October has been chosen for study since it represents the timing of meaningful water flow activities for each year. We also examine the percentage of non-flow days over the full year as well as for the March to April and May to October periods which reflect the two distinct hydrological processes mentioned earlier (snow melt/summer convection). Table 2.2 shows that stations $05AF010$ and $11AE009$ have far fewer non-flow days than the rest of the stations, which have about 50% or more non-flow days. Except for station $05AF010$, the period May to October tends to have more non-flow days compared with March to April. The minimum flow is 0 m^3/s and the maximum varies from 4 m^3/s to 152 m^3/s over all the stations. Due to some extreme flood events, the distribution of flow is right skewed as evidenced by the difference between the median and mean flow rates for each station.

Station Number	Number of Observations	Number of Missing Values Over Full Year	Number of Missing Values (Jan-Feb) & (Nov-Dec)	% Observations being zero Full Year	% Observations being zero (March-April)	% Observations being zero (May-Oct)	Observed Median (m^3/s)	Observed Mean (m^3/s)	Observed Max (m^3/s)
05AF010	8701	4083	4077	5.7%	8.5%	4.1%	0.08	0.68	124
05AH001	9166	3618	3613	78.7%	55.5%	94.3%	0	0.10	14
05JF008	8649	4135	4128	71.8%	47.5%	86.8%	0	0.02	4
05JL002	8594	4190	4181	89.5%	75.8%	97.8%	0	0.16	35
05NA003	12784	0	0	61.6%	34.1%	60.7%	0	0.88	123
05NF002	10566	2219	2214	55.3%	34.0%	57.3%	0	0.84	118
05NF008	8612	4172	4164	70.9%	56.0%	80.1%	0	0.15	35
05NG007	11346	1438	1436	48.6%	33.5%	49.7%	0	0.78	152
11AB075	8653	4131	4123	88.7%	74.9%	97.1%	0	0.07	19
11AE009	10501	2283	2280	16.9%	7.9%	18.7%	0.05	0.65	98

Table 2.2: Summary statistics for intermittent flow at 10 stations used in this study; the time period of observation is 1959-1993 inclusive.

Plots of the median daily flow, aggregated over years, for each of the ten streams, are presented in Figure 2.2. The zero-heavy nature of intermittent flow is evident from Table 2.2 and Figure 2.2; as well, peak flow periods in the spring are also evident. There seem to exist seasonal patterns for these ten stations.

2.3 Flow Duration Curve

Flow Duration Curves (FDCs) are hydrological curves that represent the relationship between the magnitude and frequency of daily flow for a particular basin, and provide an estimate of the percentage of time the flow equaled or exceeded specific values over the period under study. In addition, the variability in flow is visually illustrated via an FDC curve. From a statistical point of view, an FDC is the complement of the cumulative distribution function (cdf) of daily flow, with Q as the daily flow and p as its corresponding exceedance probability. The FDC plots Q_p, the p-th quantile of daily flow versus exceedance probability p, where p is given by $p = 1 - P(Q \leq Q_p)$; $P(A)$ refers to the probability of the event A. The sharp decline on the left of the flow duration curves as evidenced in Figure 2.3 for the ten selected streams reflects extreme events; while the zero inflation is observed in the right tail.

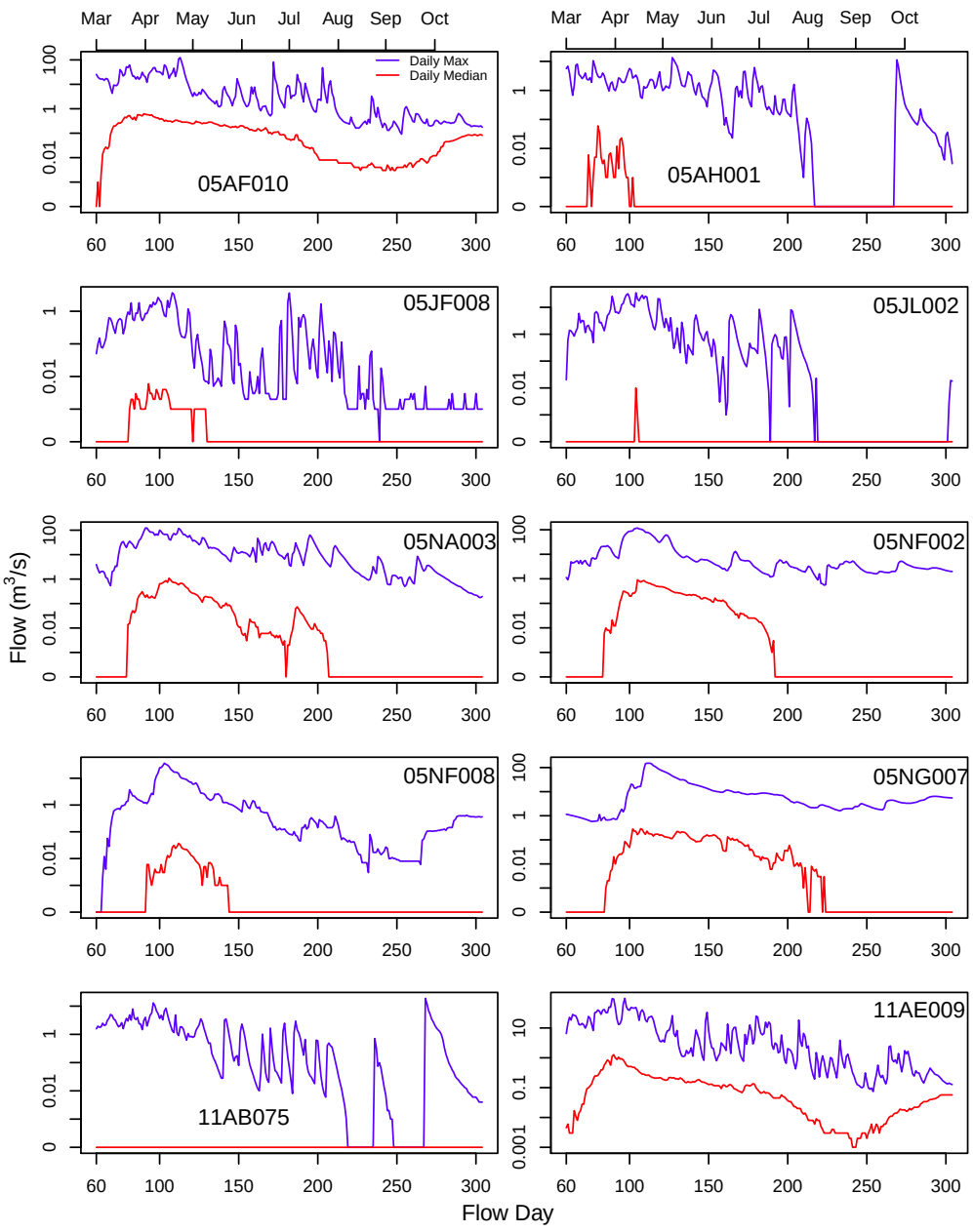

Figure 2.2: Maximum and median daily flow for 1959-1993 at the ten stations considered in this study.

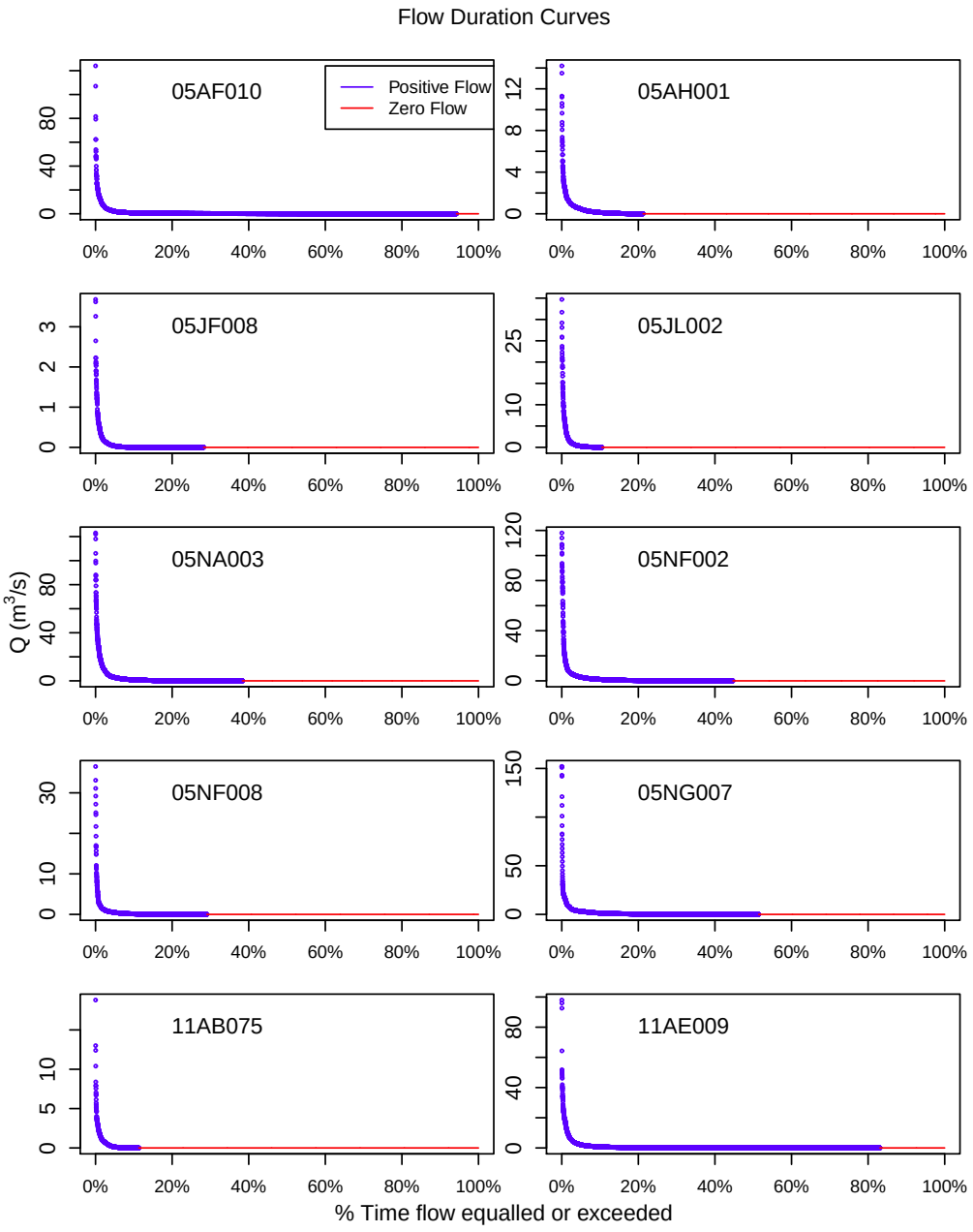

Figure 2.3: Flow duration curves for the ten hydrometric stations listed in Table 2.1. Note that the scale is different for each station.

2.4 Flow Transition Matrix

Markov modelling of day-to-day transitions in the state of a stream permits an assessment of how flow on one day affects flow on the next. In analogy with Aksoy and Bayazit (2000a), flow is regarded as being in one of the following four states: increasing (+), decreasing (−), (non-zero) constant (c), or zero (0). If Q_t is the flow on the t-th day and $X_t = Q_t - Q_{t-1}$, the increment over the flow for the previous day, flow is increasing if $X_t > 0$, decreasing if $X_t < 0$, and constant if $X_t = 0$ (zero-flow, or constant non-zero). For each of the ten streams, a transition probability matrix P, was estimated for transitions between the four states. The estimate of P_{ij} is calculated as the proportion of times a stream made a transition from state i to j, whenever it was in state i. Each entry in the matrix labeled P_{ij}, is the probability that flow will be in state j on day t, given that flow was in state i on day $t-1$, with $i, j = +, -, c$, or 0. The entries of the matrix represent the following transitions

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{0+} & P_{0-} & P_{0c} \\ P_{+0} & P_{++} & P_{+-} & P_{+c} \\ P_{-0} & P_{-+} & P_{--} & P_{-c} \\ P_{c0} & P_{c+} & P_{c-} & P_{cc} \end{bmatrix} \tag{2.1}$$

Note the presence of structural zeros, values of P_{ij} which are zero by definition: for P_{+0} (a transition to zero from + cannot be immediate but needs to be preceded by state −), P_{0-}, and P_{0c} (by definition).

The main feature we have observed is that the diagonal entries tend to be much larger than the off-diagonal entries, demonstrating that flows tend to be very similar from day to day, particularly so for the zero-flow state. Table 2.3 provides the diagonal entries (representing the highest transition probabilities) of the estimated transition probability matrix for each station, illustrating the consistency of this transitional pattern, while Table 2.4 provides the overall transition matrix for all the stations combined.

2.5 Case Study: Long Creek at Western Crossing of International Boundary, Saskatchewan

In what follows, the salient features of intermittent flow that particularly challenge the relevance of simple models are illustrated using the daily flows of the station *Long Creek*

Station Number	Zero followed by zero(P_{00})	An increase followed by an increase(P_{++})	An increase followed by a decrease(P_{--})	A constant non-zero flow followed by the same (P_{cc})
05AF010	0.92	0.51	0.71	0.57
05AH001	0.99	0.57	0.77	0.62
05JF008	0.97	0.36	0.49	0.78
05JL002	0.99	0.50	0.79	0.10
05NA003	0.99	0.56	0 .80	0.56
05NF002	0.99	0.60	0.79	0.65
05NF008	0.99	0.48	0.63	0.50
05NG007	0.98	0.61	0.75	0.58
11AB075	0.99	0.52	0.80	0.23
11AE009	0.95	0.52	0.71	0.45

Table 2.3: Selected daily transition probabilities, these are the main diagonals of P and in each case represent the largest estimated transition probabilities amongst the elements of P.

	0	+	-	c
0	0.99	0.01	0	0
+	0	0.54	0.35	0.11
-	0.03	0.13	0.73	0.11
c	0.04	0.16	0 .23	0.58

Table 2.4: Overall estimated average daily transition probability matrix for all ten hydrometric stations. For example, across all ten stations the proportion of zero flow to zero flow transitions is 0.99.

at Western Crossing of International Boundary in Saskatchewan. This station is hereafter referred to as Long Creek. There are no missing records for this station in the period considered.

Figure 2.4 displays the time series plot for stream flow rate for Long Creek from 1959 to 1993.

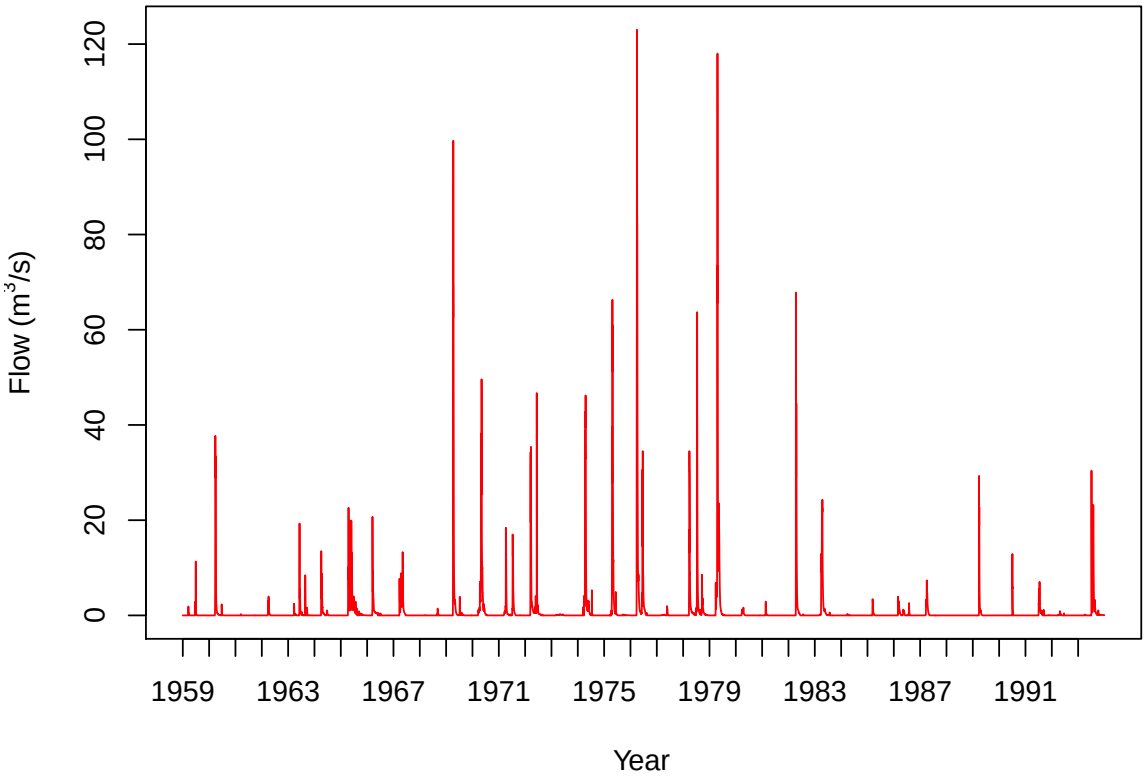

Figure 2.4: Time Series Plot of Flow Rate for Long Creek Station ($05NA003$).

The maximum flow occurs in 1976, and is 123 m^3/s. Figure 2.5 displays the percentage of flow days, days on which the flow rate was non-zero, by day of year for Long Creek. The seasonality of flow days can be directly observed; at the beginning and end of the year the flow rate is zero, as expected, with the peak percentage of flow days around mid-April, and a smaller peak evidenced around July. Figure 2.6 presents the annual percentage of flow days by year; no clear trends are apparent. Some years stand out as having large percentages of flow days, namely, 1965, 1978 and 1986.

Figures 2.7 and 2.8 plot the mean and median flow rates by day of year (Figure 2.7) and year (Figure 2.8). Figure 2.7 reflects approximately the same timing of peaks as seen in Figure 2.5. We also note that in 1976, because of a large flow episode, the mean flow rate in substantially higher than other years. As well, in 1978 there were three large flood events which occurred over very long durations, including one event which lasted over two months; this accounts for the very large median value in that year.

For Long Creek the estimate of the Markov transition matrix modelling day-to-day transitions (P; see Equation 2.1) is

	0	+	-	c
0	0.99	0.01	0	0
+	0	0.56	0.38	0.06
-	0.02	0.11	0.80	0.07
c	0.05	0.14	0 .24	0.56 .

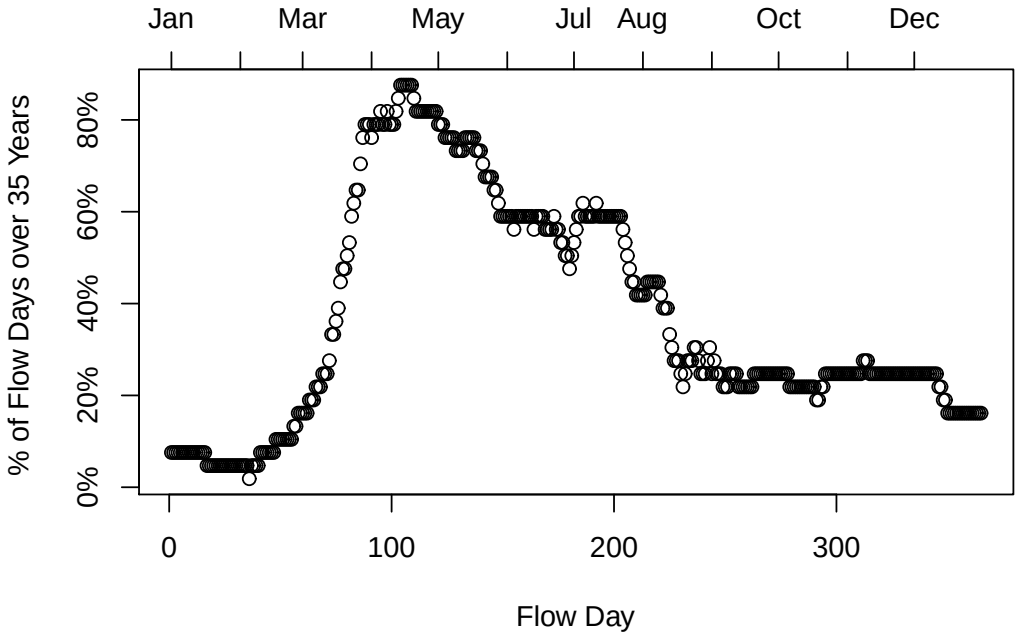

Figure 2.5: Percentage of Flow Days by Day of Year for Long Creek Station ($05NA003$).

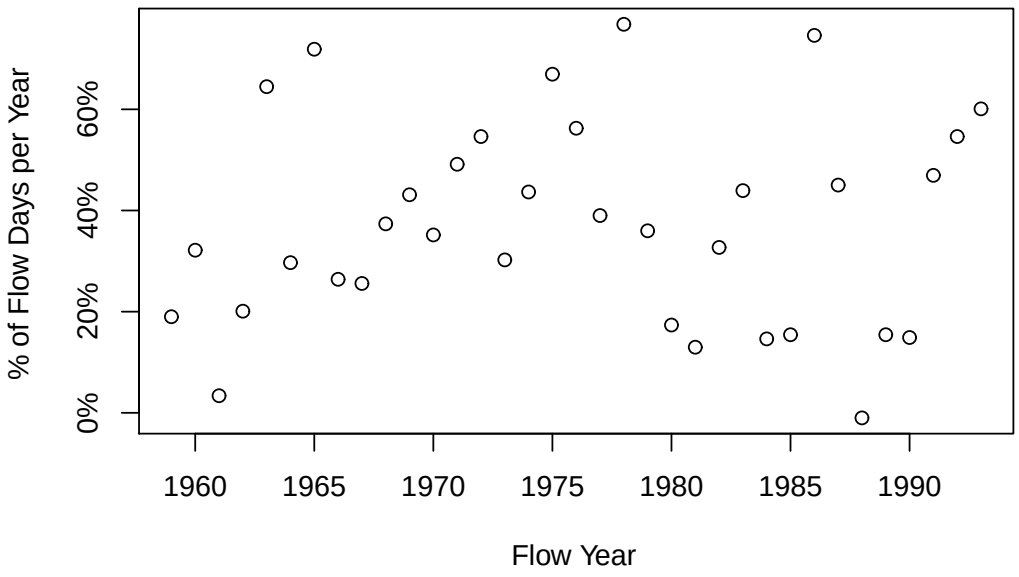

Figure 2.6: Percentage of Flow Days by Year for Long Creek Station ($05NA003$).

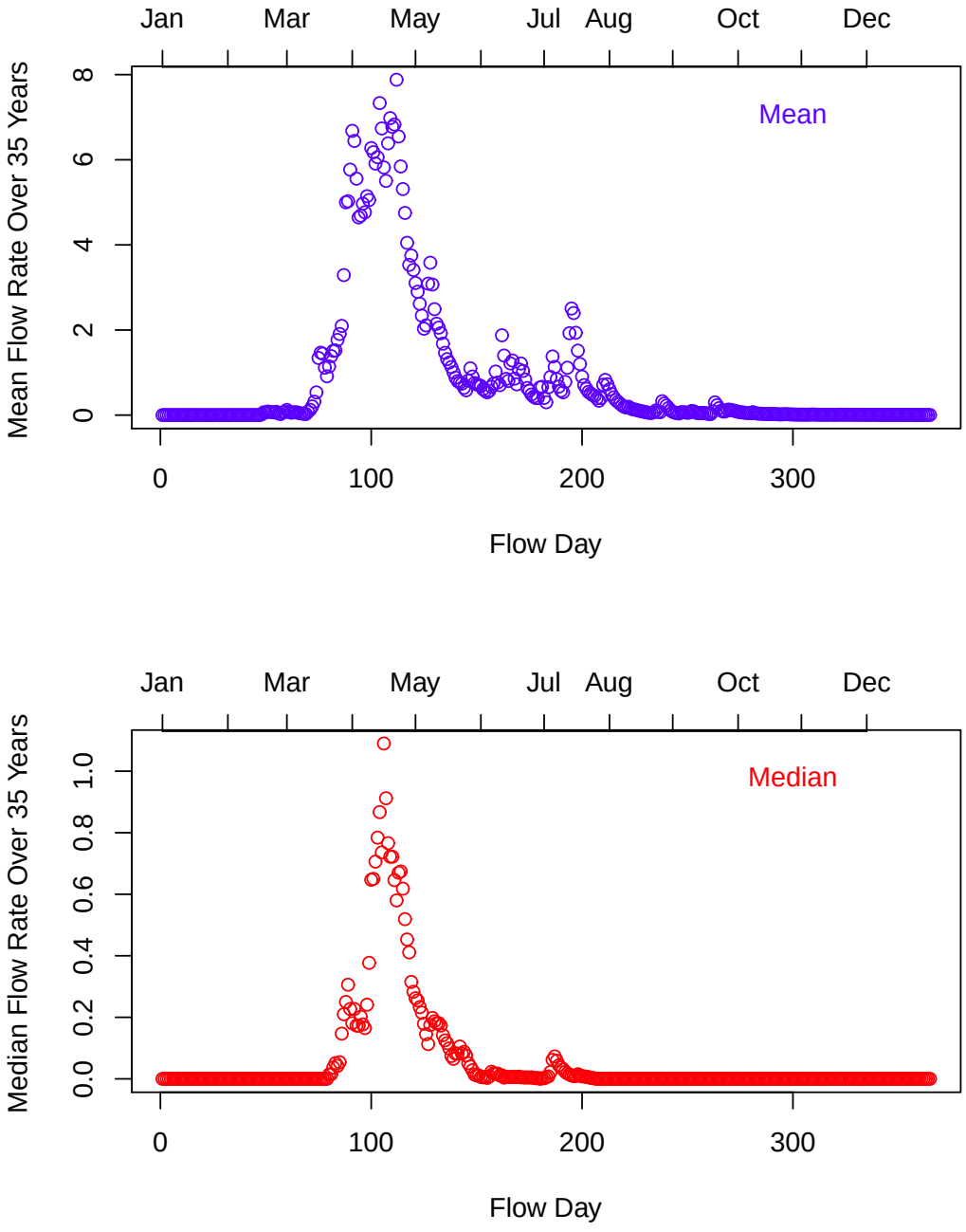

Figure 2.7: Mean/Median Flow Rate by Day of Year for Long Creek Station (05NA003) Over 35 Years.

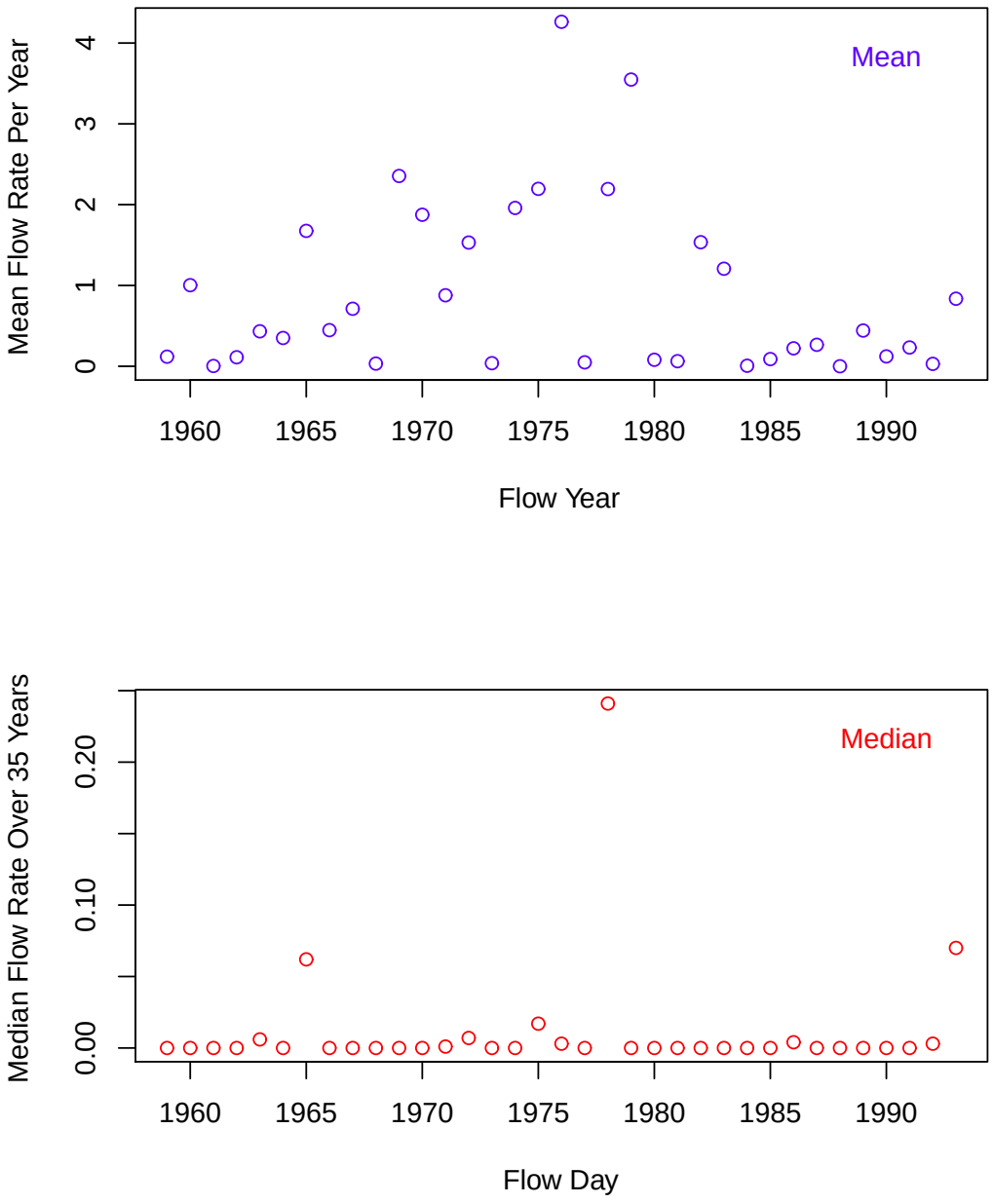

Figure 2.8: Mean/Median Flow Rate by Year for Long Creek Station (05NA003) Over 35 Years.

2.6 Summary of Exploratory Analysis

The statistical features associated with intermittent flow are illustrated through exploratory analyses for the set of ten seasonally observed (March to October) Prairie streams. The time series plots of flow by day (Figure 2.2) and the Flow Duration Curves (FDCs, Figure 2.3) exhibit (i) seasonality, (ii) extreme events (floods), (iii) zero-inflation (sequences of no flow days), and these features need to be captured in our exploratory models. Our statistical models incorporate possible day-to-day linkage in flow through the use of smoothers. Smoothers allow for continuous changes in probabilities of flow and of mean flow rates; they support similarity of values for these elements of the model within small time changes. This is important in analyses involving daily data of stream flow. In the next chapter, we will illustrate the use of smoothers by introducing generalized additive models as well as penalized spline smoothing approaches and then later employ these methods for our study.

Chapter 3

Statistical Models

In this chapter, we review generalized additive models and penalized spline smoothing approaches. We start by introducing to generalized additive models, then describe the details of spline smoothing. Furthermore, we illustrate the use of thin-plate splines, the choice of basis functions to be used for this project.

3.1 Generalized Additive Models

Generalized Additive Models proposed by Hastie & Tibshirani (1990), are flexible regression models that extend the Generalized Linear Model (GLM) by incorporating nonparametric smooth functions into covariate effects. The general framework of a GAM for modeling the mean of a response variable can be stated as follows:

$$g(\mu_i) = \widetilde{\mathbf{X}}_i\boldsymbol{\beta} + \mathbf{m}(\mathbf{x}_i) \tag{3.1}$$

where $g(\cdot)$ is termed the link function, $\mu_i = E(Y_i)$, Y_i is the response variable, which follows some exponential family distribution with probability density function of the form:

$$f_\omega(y) = exp[y\omega - b(\omega)/a(\phi) + c(y,\phi)];$$

here b, a and c are known functions, ϕ is the 'scale' parameter and ω is known as the 'canonical parameter' of the distribution. A row of model matrix, $\widetilde{\mathbf{X}}_i$, corresponds to strictly parametric model components with regression effect β, while $\mathbf{x}_i$ contributes to the nonparametric term as described below. The term $\mathbf{m}(\cdot)$ is a sum of smooth functions of covariates

which has the following structure:

$$\mathbf{m}(\mathbf{x}_i) = m_1(x_{1i}) + m_2(x_{2i}) + m_3(\mathbf{x}_{3i}) + \cdots \tag{3.2}$$

Here the $m_j(\cdot)$ are smooth functions of the covariates, x_j, some of which may be multidimensional, as for example, $\mathbf{x}_{3i}$ as the representation above. If $\mathbf{x}_{3i} = (x_{3i1}, x_{3i2})$, $m_3(\mathbf{x}_{3i})$ would be a bivariate smoother.

GAMs permit the conditional mean of the response to be dependent on a sum of smooth functions of covariates and yield a more flexible model specification than generalized linear models (GLM). Well-designed GAMs can therefore provide the potential for better fits to data than GLMs. The smooth functions are linear combinations of what is known as basis functions.

For univariate smoothers, these smooth functions could be piecewise polynomial functions such as cubic regression splines. A cubic regression spline is a curve constructed from segments of cubic polynomials joined together so that the curve is continuous up to the second derivative. The points at which they are joined are termed the knots of the spline. Usually, the univariate smoothers take the form:

$$m_j(x_j) = \sum_{k=1}^{K} b_{jk}(x_j)\theta_{jk}, \tag{3.3}$$

where $b_{jk}(x_j)$ is the basis function of the jth smoother (for the jth covariate). Note that $j = 1, 2, \cdots, J$ and K equals the sum of the number of inner knots and the number of degrees of freedom for GAMs. Here θ_{jk} are the corresponding coefficients for the basis function $b_{jk}(x_j)$.

Overfitting can be a problem if K is large. However, penalized spline smoothing techniques have been introduced to overcome this problem. The idea is that GAMs can be estimated by penalized likelihood maximization by introducing penalties for overly wiggly estimates of the $m_j(\cdot)$ terms, given that the size of the basis dimension is fixed and slightly larger than it is believed could reasonably be necessary.

By letting $\mathbf{m}_{\mathbf{j}i} = m_j(x_{ji})$ and $\boldsymbol{\vartheta}_j = [\theta_{j1}, \theta_{j2}, \cdots, \theta_{jK}]^T$, equation (3.2) can be rewritten as follows:

$$\begin{aligned} \mathbf{m}(x_i) &= \mathbf{m}_{\mathbf{1}i} + \mathbf{m}_{\mathbf{2}i} + \mathbf{m}_{\mathbf{3}i} + \cdots + \mathbf{m}_{\mathbf{J}i} \\ &= \overline{\mathbf{X}}_{\mathbf{1}}\boldsymbol{\vartheta}_{\mathbf{1}} + \overline{\mathbf{X}}_{\mathbf{2}}\boldsymbol{\vartheta}_{\mathbf{2}} + \overline{\mathbf{X}}_{\mathbf{3}}\boldsymbol{\vartheta}_{\mathbf{3}} + \cdots + \overline{\mathbf{X}}_{\mathbf{J}}\boldsymbol{\vartheta}_{\mathbf{J}} \end{aligned} \tag{3.4}$$

where

$$\overline{\mathbf{X}}_{\mathbf{j},ik} = b_{jk}(x_{ji}).$$

Note that x_j could be either a scalar or a vector quantity.

We reparameterize by new parameters $\boldsymbol{\theta}_j$ such that $\boldsymbol{\vartheta}_j = \mathbf{S}\boldsymbol{\theta}_j$ to meet the centering constraints:

$$\mathbf{1^T}\overline{\mathbf{X}}_j\boldsymbol{\vartheta}_j = 0,$$

by finding a matrix $\mathbf{S}$ with $K-1$ orthogonal columns and

$$\mathbf{1^T}\overline{\mathbf{X}}_\mathbf{j}\mathbf{S} = 0.$$

Now $\mathbf{m}_{\mathbf{j}i} = \overline{\mathbf{X}}_\mathbf{j}\mathbf{S}\boldsymbol{\theta}_j$ and we define $\mathbf{X}_j = \overline{\mathbf{X}}_\mathbf{j}\mathbf{S}$. Model (3.1) can be reformulated into the form of a GLM:

$$g(\mu_i) = \mathbf{X}\boldsymbol{\theta} \tag{3.5}$$

where $\mathbf{X} = [\widetilde{\mathbf{X}}, \mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_J}]$ and $\boldsymbol{\theta}^T = [\boldsymbol{\beta}^T, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \cdots, \boldsymbol{\theta}_J^T]$. Because of the connection with GLMs, the likelihood, $L(\theta)$, becomes conceptually straightforward. The penalized likelihood then takes the form:

$$l_p(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - \frac{1}{2}\sum_j \lambda_j \boldsymbol{\theta}^T Z_j \boldsymbol{\theta}. \tag{3.6}$$

The penalty term $\boldsymbol{\theta}^T Z_j \boldsymbol{\theta}$ is derived from the integrated square of second derivatives of the smooth functions $m_j(x)$:

$$\int [m_j''(x)]^2 dx$$

and penalizes models that are too "wiggly". The trade off between model fit and model smoothness is controlled by the smoothing parameters λ_j. If $\lambda_j = 0$, the estimates obtained by maximizing the likelihood would be un-penalized regression spline estimates. However, if $\lambda_j \to \infty$, the resulting estimate of the smoother would be a straight line.

To estimate $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$, the penalized spline smoothing approach proceeds in the following way. For a fixed $\boldsymbol{\lambda}$, problem (3.6) can be solved by minimizing penalized iteratively re-weighted least squares(P-IRLS), which is similar to the GLM iterative estimation procedure. There are two common approaches for selecting the smoothing parameter $\boldsymbol{\lambda}$: if the scale parameter σ is known, the smoothing parameter can be determined by minimizing the Un-Biased Risk Estimator(UBRE; Craven and Wahba, 1979), also called Mallows' C_p (Mallows, 1973):

$$V_\mu(\boldsymbol{\lambda}) = \|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2/n - \sigma^2 + 2tr(\mathbf{A})\sigma^2/n.$$

Here $\mathbf{A} = \mathbf{X}(\mathbf{X^T W X} + \mathbf{Z})^{-1}\mathbf{X^T W}$ is the influence matrix of the GAM; $\mathbf{W}$ is a diagonal weight matrix with $W_{ii}^{-1} = V(\mu_i)g'(\mu_i)$ with $V(\mu) = b''(\theta)/\omega$; $tr(A)$ is the trace of A, and $\mathbf{Z} = \sum_j \lambda_j Z_j$. If the scale parameter is unknown, Generalized Cross Validation (GCV; Wood, 2006; Wood and Augustin, 2002) can be used to obtain an estimate which minimizes the mean square prediction error, that is, the average squared error in predicting a new observation y using the fitted model:

$$V_g = \frac{nD(\hat{\boldsymbol{\theta}})}{[n - tr(A)]^2}.$$

Here, $D(\hat{\boldsymbol{\theta}}) = 2[l_p(\hat{\boldsymbol{\theta}}_{max}) - l_p(\hat{\boldsymbol{\theta}})]\phi$ is the model deviance, $\hat{\boldsymbol{\theta}}_{max}$ is the maximum likelihood estimate (m.l.e) of the saturated model which estimates values for each y by the observed value while $\hat{\boldsymbol{\theta}}$ is the m.l.e of the model of interest. The trace $tr(A)$ plays an important role in determining the smoothing parameter and has been defined as the effective degrees of freedom in measuring the flexibility of the fitted model. As the smoothing parameter varies from zero to infinity, $tr(A)$ ranges from the number of parameters less the number of constraints to rank $(\sum_j Z_j)$ minus the number of constraints. The discussion above considers GAMs broadly. Section 3.2 discusses a specific form which we apply to the stream flow data.

3.2 Thin-Plate Splines

In this section, we focus on the thin-plate spline regression. Thin plate splines, introduced by Duchon(1977), provide a general solution to the problem of estimating a smooth function of multiple predictor variables.

Given n observations $(y_i, \mathbf{x_i})$, consider the general problem of estimating the smooth function $\mathbf{f}(\mathbf{x})$ using the following model:

$$y_i = f(\mathbf{x_i}) + \epsilon_i$$

where ϵ_i is a random error term and $\mathbf{x}$ is a vector of length d.

To estimate $\mathbf{f}$, thin-plate spline smoothing finds the function $\hat{\mathbf{g}}$ which satisfies:

$$\min_g \|\mathbf{y} - \mathbf{g}\| + \lambda J_{md}(\mathbf{g}), \tag{3.7}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and $\mathbf{g} = [g(\mathbf{x_1}), g(\mathbf{x_1}), \ldots, g(\mathbf{x_1})]^T$. As in Section 3.2, $J_{hd}(g)$ is introduced to penalize the 'wiggliness'of g and is defined as follows:

$$J_{md} = \int \cdots \int_{\Re^d} \sum_{\tau_1 + \cdots + \tau_d = h} \frac{m!}{\tau_1! \ldots \tau_d!} \left(\frac{\partial^h g}{\partial x_1^{\tau_1} \ldots \partial x_d^{\tau_d}} \right)^2 dx_1 \ldots dx_d \tag{3.8}$$

The above minimization problem (3.7) can be optimized by the following solution given $2h > d$:

$$\hat{g}(\mathbf{x}) = \sum_{i=1}^{n} \kappa_i \varsigma_{hd}(\|\mathbf{x} - \mathbf{x_i}\|) + \sum_{j=1}^{H} \alpha_j \psi_j(\mathbf{x}), \tag{3.9}$$

where $\boldsymbol{\kappa}$ and $\boldsymbol{\alpha}$ are parameters to be estimated and in particular, $\boldsymbol{\kappa}$ has to satisfy the linear constraints that $\mathbf{S^T}\boldsymbol{\kappa} = \mathbf{0}$. Here $S_{ij} = \psi_j(\mathbf{x}_i)$ and n is the number of knots. There exist $H = \binom{h+d+1}{d}$ polynomials $\psi_i(\cdot)$ whose degrees are less than h; these are linearly independent polynomials spanning the space of polynomials in $\Re^d$. We also have

$$\varsigma_{hd}(r) = \begin{cases} \frac{(-1)^{h+1+d/2}}{2^{2h-1}\pi^{d/2}(h-1)!(h-d/2)!} r^{2h-d} \log(r) & d \quad \text{even} \\ \\ \frac{\Gamma(d/2-h)}{2^{2h}\pi^{d/2}(h-1)!} r^{2h-d} & d \quad \text{odd} \end{cases}$$

Let $\mathbf{E}$ be $E_{ij} = \varsigma_{hd}(\|\mathbf{x_i} - \mathbf{x_j}\|)$; the thin plate spline fitting problem becomes:

$$\begin{aligned} &minimize \quad \|\mathbf{y} - \mathbf{E}\boldsymbol{\kappa} - \mathbf{S}\boldsymbol{\alpha}\| + \lambda \boldsymbol{\kappa}^T \mathbf{E} \boldsymbol{\kappa}, \\ &subject\ to \quad \mathbf{S^T}\boldsymbol{\kappa} = \mathbf{0}, \end{aligned} \tag{3.10}$$

with respect to $\boldsymbol{\kappa}$ and $\boldsymbol{\alpha}$.

3.3 Application to Our Study

Our hydrologic analysis is built to handle a zero-rich streamflow series. Our exploratory model incorporates two compartments for seasonality and annual trend: one for zero flow days, the second for flow rate, given flow is occurring. These two compartments resonate with the dual nature of the data: mixed between discrete (zeros) and continuous (flows, given presence of flow). As well, smoothers in the form of Generalized Additive Models (GAMs) for each of these compartments are used to provide local (temporal) continuity in the absence of flow as well as local continuity in flow rates, given presence. It may be that there will be residual autocorrelation to consider, as discussed in the final section.

Different models for snowmelt periods (March-April) and convective precipitation periods (May-October) were investigated as these are based on different hydrological processes. The term snowmelt period here refers to the period where snowmelt is generally the dominant process producing streamflow. Snowmelt periods will differ by year but March to April was used as a fair approximation in this study; extensions to the model which estimate the timing of snowmelt are discussed in the last section of the study.

Hence, our basic exploratory models for intermittent flow incorporate (i) two compartments to model presence/absence of flow, and flow rates, given flow (ii) smoothers, for each of these compartments, to build day-to-day and year-to-year connections, and because these are flexible modelling forms, (iii) possibly separate models for snowmelt and convective periods. Three types of smoothers are considered for modelling the log-odds of the probability of a non-zero flow day and the logarithm of (non-zero) flow rate; these three smoothers reflect different degrees of complexity.

Model 1:

$$\text{logit}(P(Z_{tk} = 1)) = S_0(t, k) \tag{3.11a}$$

$$\log(Q_{tk}) = f_0(t, k) + \epsilon_{tk}, \text{ when } Q_{tk} > 0 \tag{3.11b}$$

where Q_{tk} represent the flow on the t-th day in the k-th year, $Z_{tk} = 1$ if the flow of the t-th day in the k-th year is non-zero, $Z_{tk} = 0$, otherwise, logit(.) refers to the log odds function, $S_0(t, k)$ and $f_0(t, k)$ are two-dimensional spline smoothers in day and year, while ϵ_{tk} are independent and normally distributed.

Model 2:

$$\text{logit}(P(Z_{tk} = 1)) = S_{01}(t) + S_{02}(k) \tag{3.12a}$$

$$\log(Q_{tk}) = f_{01}(t) + f_{02}(k) + \epsilon_{tk}, \text{ when } Q_{tk} > 0 \tag{3.12b}$$

where $S_{01}(t)$ and $f_{01}(t)$ are one-dimensional spline smoothers in day of year, $S_{02}(k)$ and $f_{02}(k)$ are spline smoothers in year, while ϵ_{tk} are independent and normally distributed. Model (2) uses an additive framework to incorporate annual effects and is more parsimonious than Model (1).

Model 3:

$$\text{logit}(P(Z_{tk} = 1)) = (\beta_0 + \beta_1 t)\sin(2\pi t/365) + (\beta_3 + \beta_4 t)\cos(2\pi t/365) + S_0(k) \tag{3.13a}$$

$$\log(Q_{tk}|Q_{tk} > 0) = (\beta_4+\beta_5 t)\sin(2\pi t/365)+(\beta_6+\beta_7 t)\cos(2\pi t/365)+f_0(k)+\epsilon_{tk}, \text{ when } Q_{tk} > 0 \tag{3.13b}$$

where $S_0(k)$ and $f_0(k)$ are again one-dimensional spline smoothers in year, and ϵ_{tk} are independent and normally distributed. This is a simple smoother that incorporates harmonic terms which are commonly used to model seasonality. The coefficients of the harmonic terms linearly change with time, which permits the seasonal effects to change with time.

The number of parameters required for estimation of the log-odds compartment and for the logarithm compartment for each of Models 1, 2 and 3 is 30, 19 and 14, respectively. In the subsequent discussion we focus on comparisons between Model 2 and Model 3. However, remarks relating to Model 1 are also provided for comparison. Comparisons were made using some typical model selection criteria for GAMs, namely the minimum Unbiased Risk Estimator (UBRE; Craven and Wahba, 1979) score, the GCV score and the percent deviance explained by the model.

Chapter 4

Modeling Results

4.1 Model Comparison

Our preliminary models based on smoothers fitted to Long Creek data yielded reasonable fits. Based on the exploratory analyses, a preliminary model for flow should incorporate annual trends and seasonality. If annual trends are weak, the seasonality compartment may suffice. An additive model with these terms instead of the full interaction spline smoothers in Model 1, with suitable model reductions as needed, may also be considered. Both Models 2 and 3 seem suitable in this context. Model 3 is proposed as an initial starting point for future model development. Our justification of the choice of models is based on model selection criteria, goodness of fit plots and adopting the principal of parsimony. These models are discussed below.

The fitted models are illustrated through (a) a plot of the estimated probability of a non-zero flow day, against day and year, (b) the mean log-flow, for a positive flow day. The contour plots illustrate specific probability or mean values for the day, year combinations and the timing for which these values are highest. Results for Long Creek (Figures 4.1 and 4.2) reveal a peak flow day around day 120 and seasonality in the probability of a non-zero flow day since trends in these probabilities tend to be (generally) similar across years. There are subtle trends across years - the contours are not perfectly aligned vertically. Figure 4.3 displays the smoother terms for the two compartments from Models 2 and 3. The log flow compartment seems to peak around day 104 while the logit compartment peaks at about day 107. Note that the shapes of the seasonality curves for both models compartments (logit/log flow) are very similar. The annual trend of the logit compartment is slightly

different from the annual trend of the log flow compartment. The logit compartment seems to reach a maximum in 1976. However, for the log flow compartment, the annual trends have peaks around 1968 and 1978.

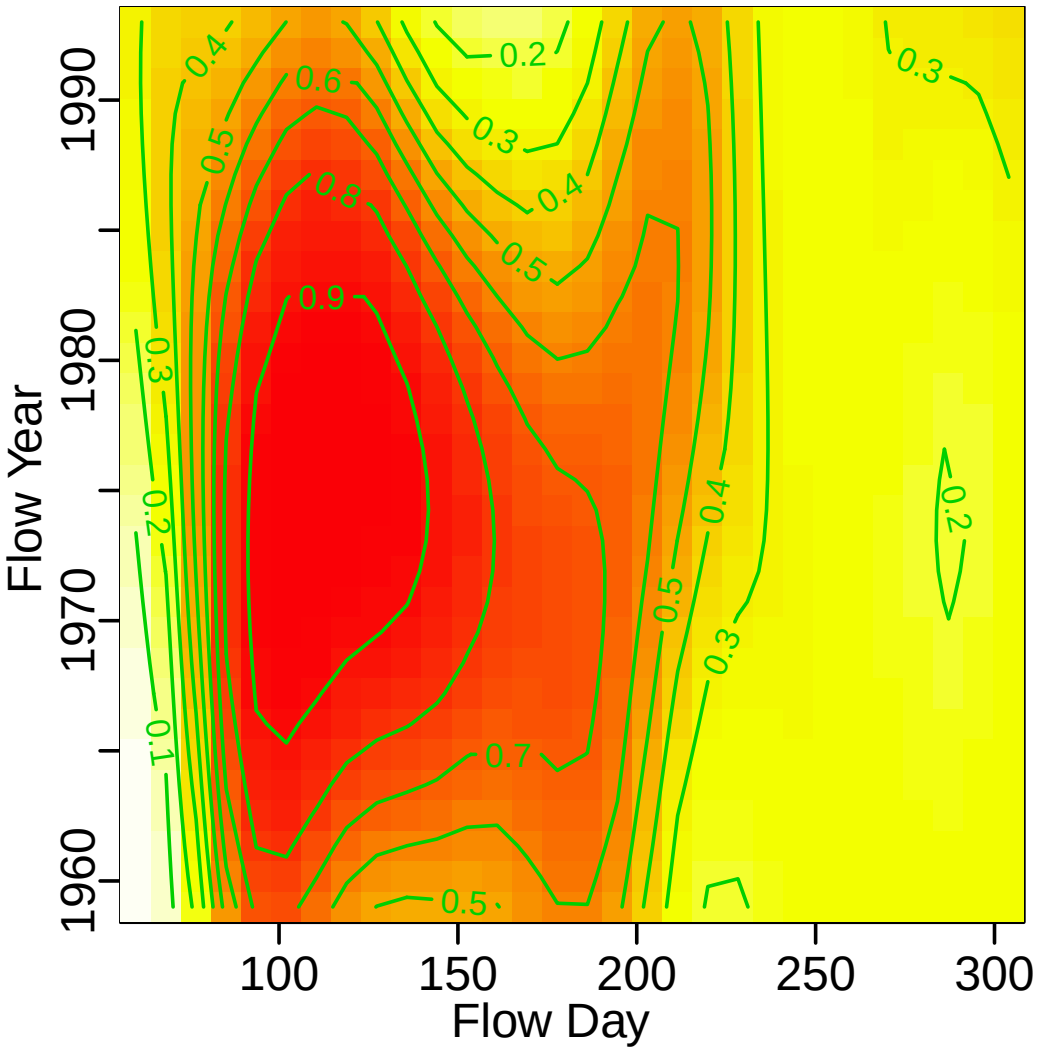

Figure 4.1: Contour plot of the estimated probability of non-zero flow by day and year for Long Creek based on Model 1 (Equation 3.11a).

For the logit compartment, the UBRE scores for Models 1, 2 and 3 are 0.111, 0.084, and 0.102, respectively. Since smaller values are preferred, Models 2 and 3 perform slightly better. For the logarithmic compartment, the GCV scores for Models 1, 2 and 3 are 1.032,

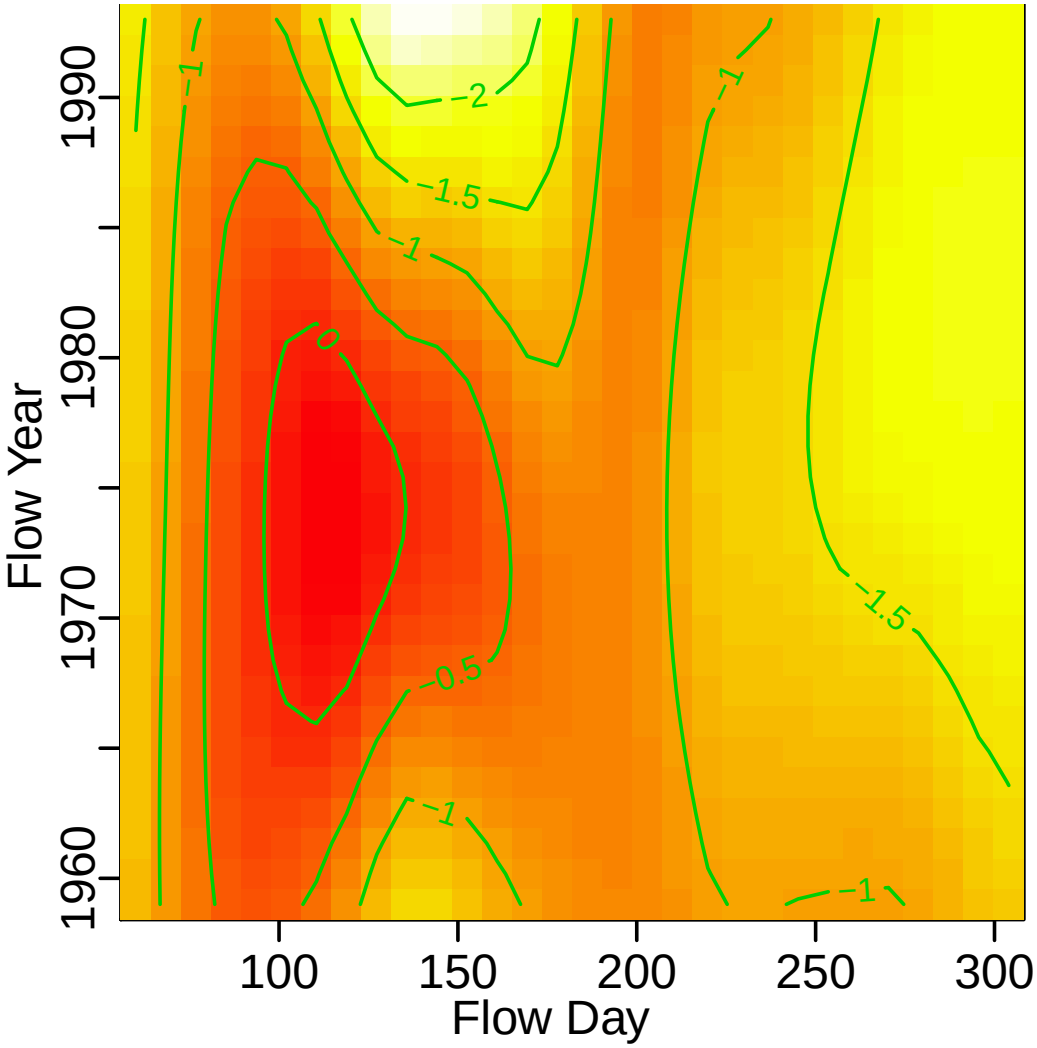

Figure 4.2: Contour plot of the logarithm of mean flow on non zero flow days by day and year for Long Creek based on Model 1 (Equation 3.11b)

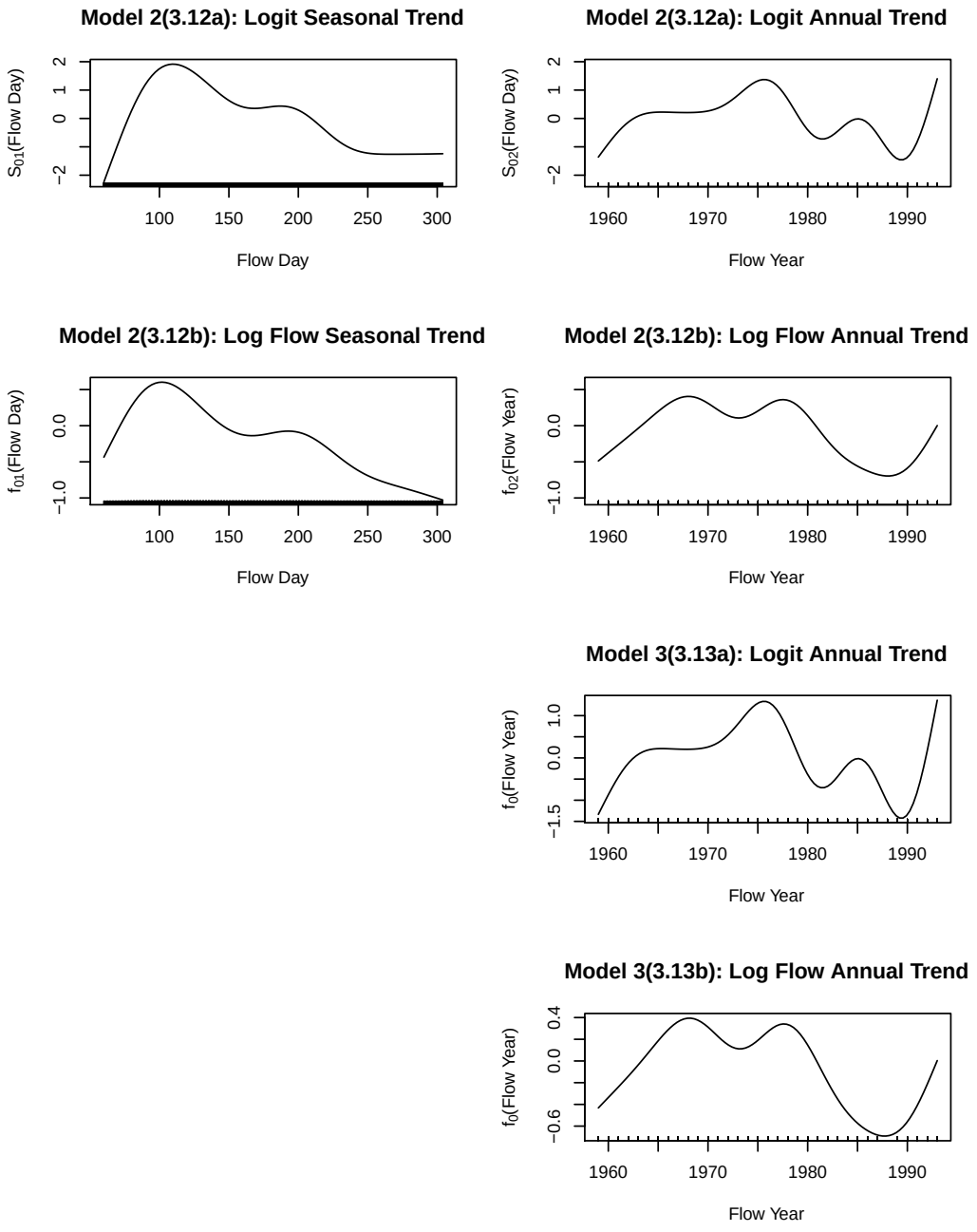

Figure 4.3: Smoother terms for Model 2 and 3; The top two panels are the estimated seasonal trend and annual trend of the logit compartment of Model 2; the two panels in second row are the estimated seasonal trend and annual trend of the log flow compartment of Model 2; the bottom two panels show the estimated annual trend of the logit and log flow compartments of Model 3.

1.068, and 1.087; once again this suggests a close fit between Models 2 and 3 with Model 1 performing slightly better. In addition, the percentage of deviance explained (larger values preferred) for each of these Models is 20.4 (Model 1), 22.1 (Model 2) and 20.8 (Model 3), yielding a slight preference for Model 2.

As interest centres on describing snowmelt and convective precipitation periods, further goodness of fit assessments for each model were carried out subseasonally. Figure 4.4 compares the fraction of observed flow days over 11-day windows with their expected values for all models. The choice of 11 days for this comparison approximates the scale of climatological processes reasonably well (see for example, Le Barbé and Lebel, 1997; Dissing and Wendler, 1998; Sparks and Menzel 2002; and Romolo et al., 2006a, 2006b). On the basis of the plots in Figure 4.4, subseasonal Models 2 and 3 are preferred as they have far fewer parameters than Model 1 and expected counts closely match observed counts for these models. Total log-flows and their expected values over 11-day windows are provided in Figure4.5. Here again subseasonal Models 2 and 3 provide reasonable fits.

A contour plot showing the fitted values for the estimated probabilities for non-zero flow based on Model 3, for the full year is provided in Figure 4.6; corresponding plots relating to separate models fitted to data reflecting (i) the snowmelt and (ii) the convective period are provided in Figures 4.7 and 4.8 respectively. Corresponding plots based on Model 2 are very similar. Generally, high probabilities of flow were observed around days 120 and 175 (April 29 and June 23, respectively), though this pattern was more consistent over years for day 120. There are also some years for which higher probabilities of flow persist over the full year, for example 1978 and 1986. The convective period was far more variable in terms of probability of flow, than the period of snowmelt.

The same conclusions can be drawn for the remaining nine stations considered in this study based upon the goodness of fit plots for Model 3 (Figures 4.9 and 4.10). We note that Model 1 fits the data reasonably well. However, when considering subseasonal approaches, as scientifically relevant in the hydrology, Models 2 and 3 seem very useful as they fit well and require far fewer parameters.

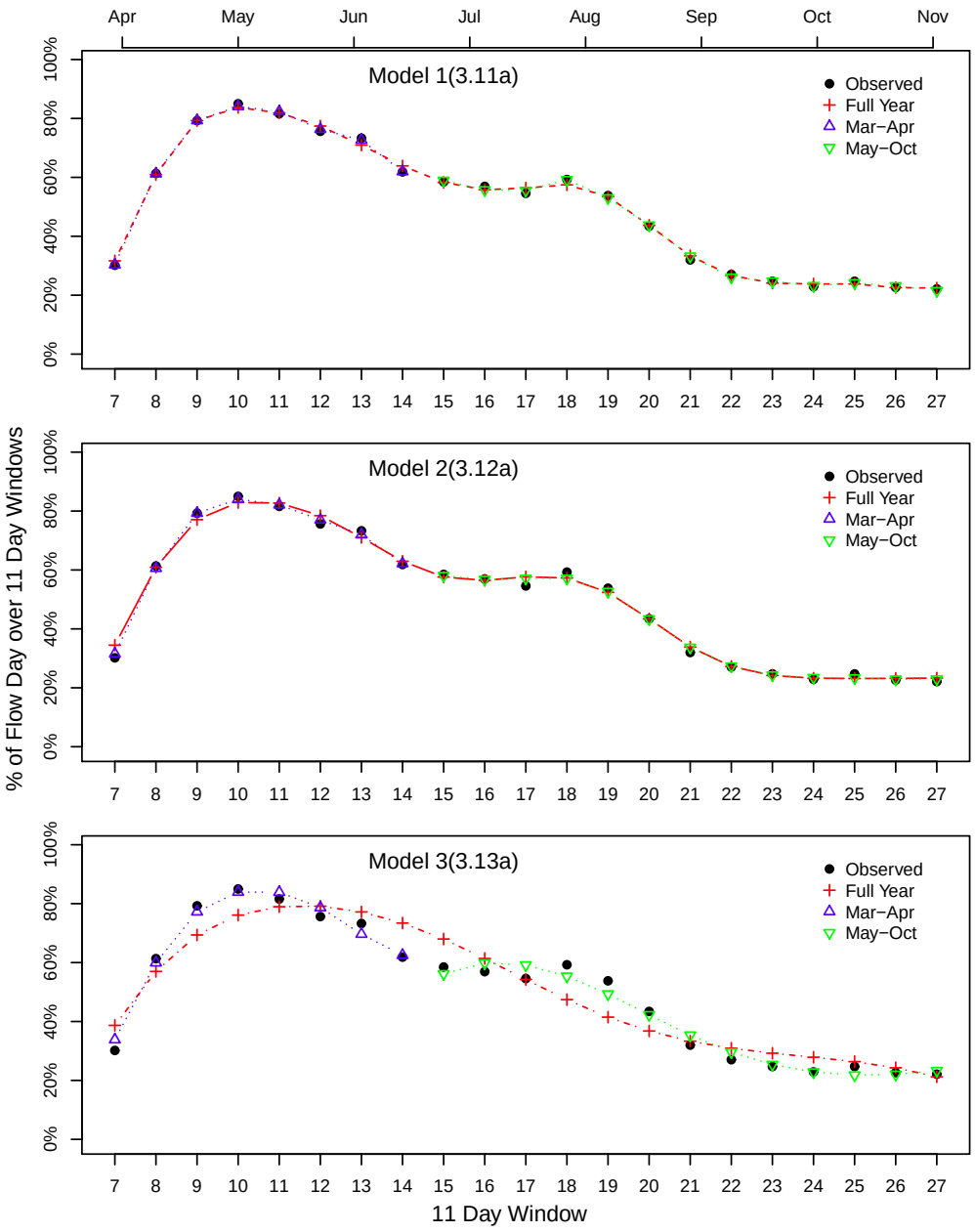

Figure 4.4: Observed and expected fraction of non-zero flow days over 11-day windows for all models for full year and subseasonal analyses (Long Creek) for Model 1 (Equation 3.11a), Model 2 (Equation 3.12a), and Model 3 (Equation 3.13a).

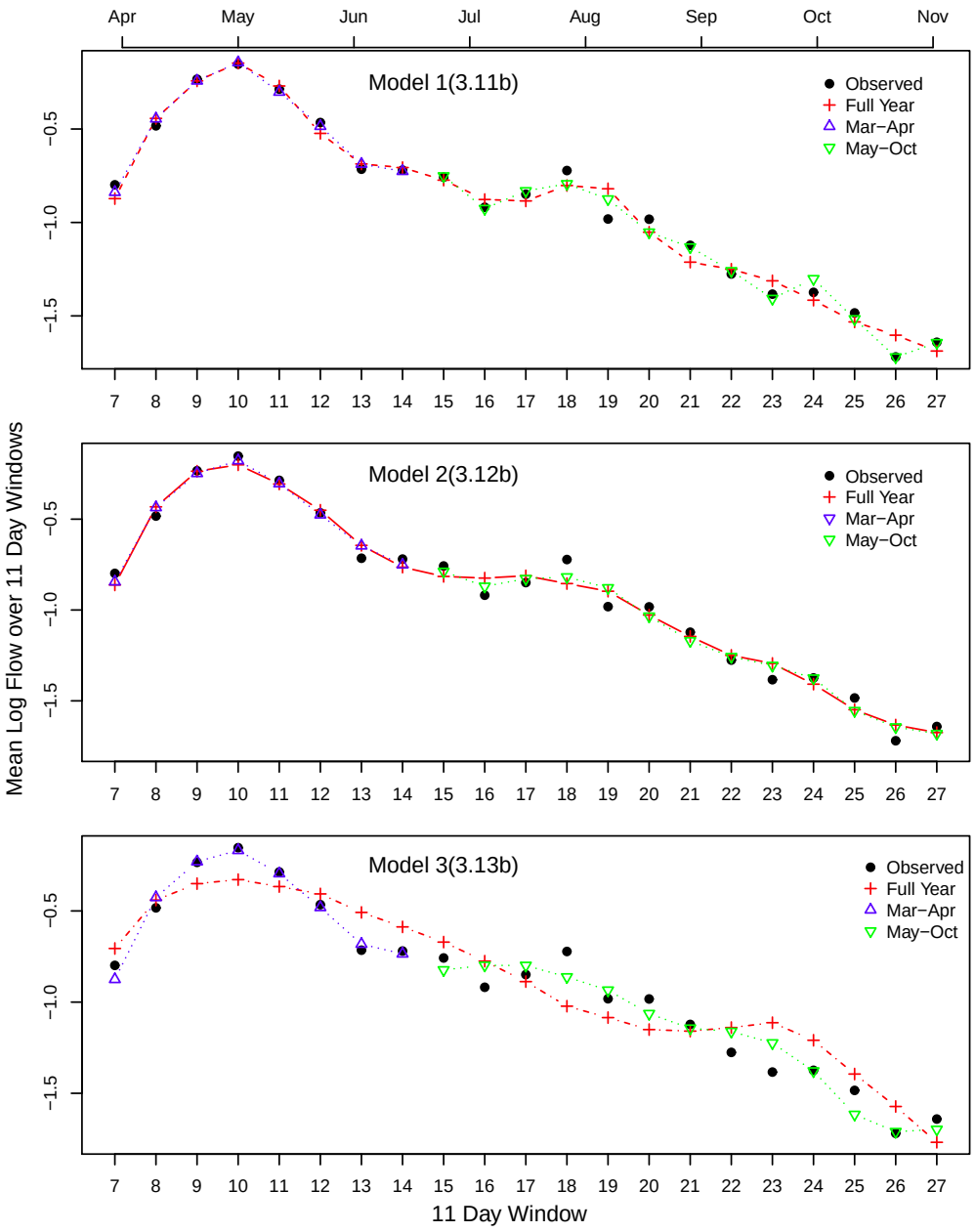

Figure 4.5: Observed and expected log flow over 11-day windows based for all models for full year and subseasonal analyses (Long Creek) for Model 1 (Equation 3.11b), Model 2 (Equation 3.12b), and Model 3 (Equation 3.13b).

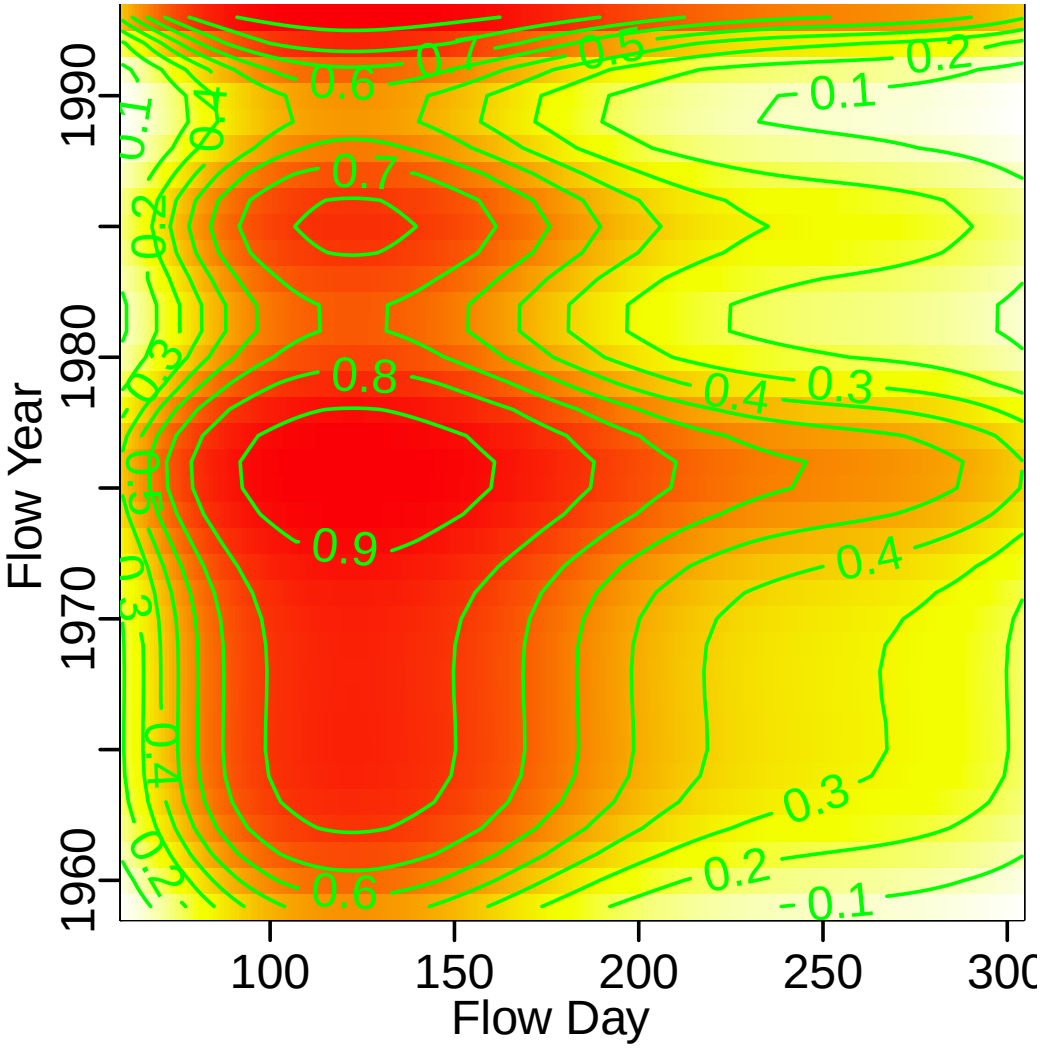

Figure 4.6: Contour plot for estimated probability of non-zero flow by day and year for Long Creek based on the full year using Model 3 (Equation 3.13a).

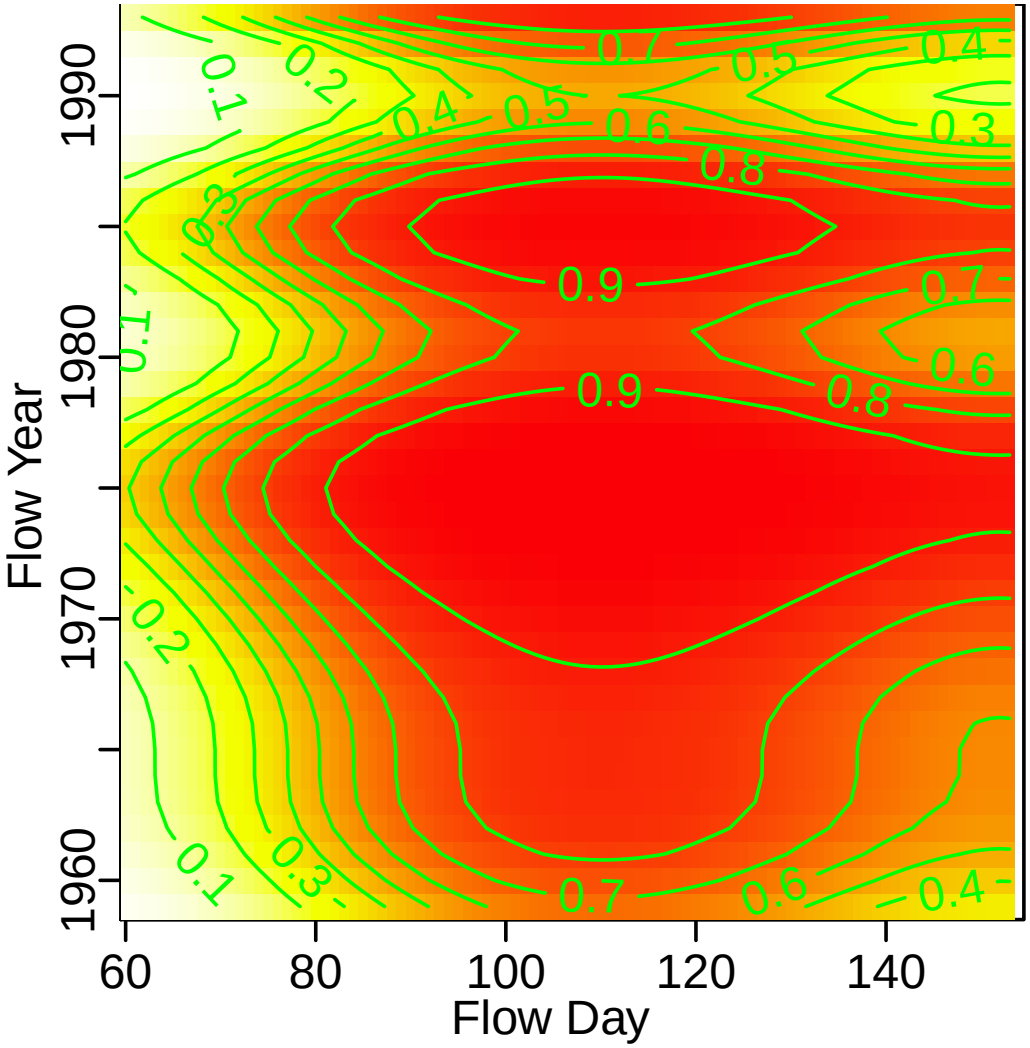

Figure 4.7: Contour plot for estimated probability of non-zero flow by day and year for Long Creek based on March-April subseasonal approach using Model (Equation 3.13a).

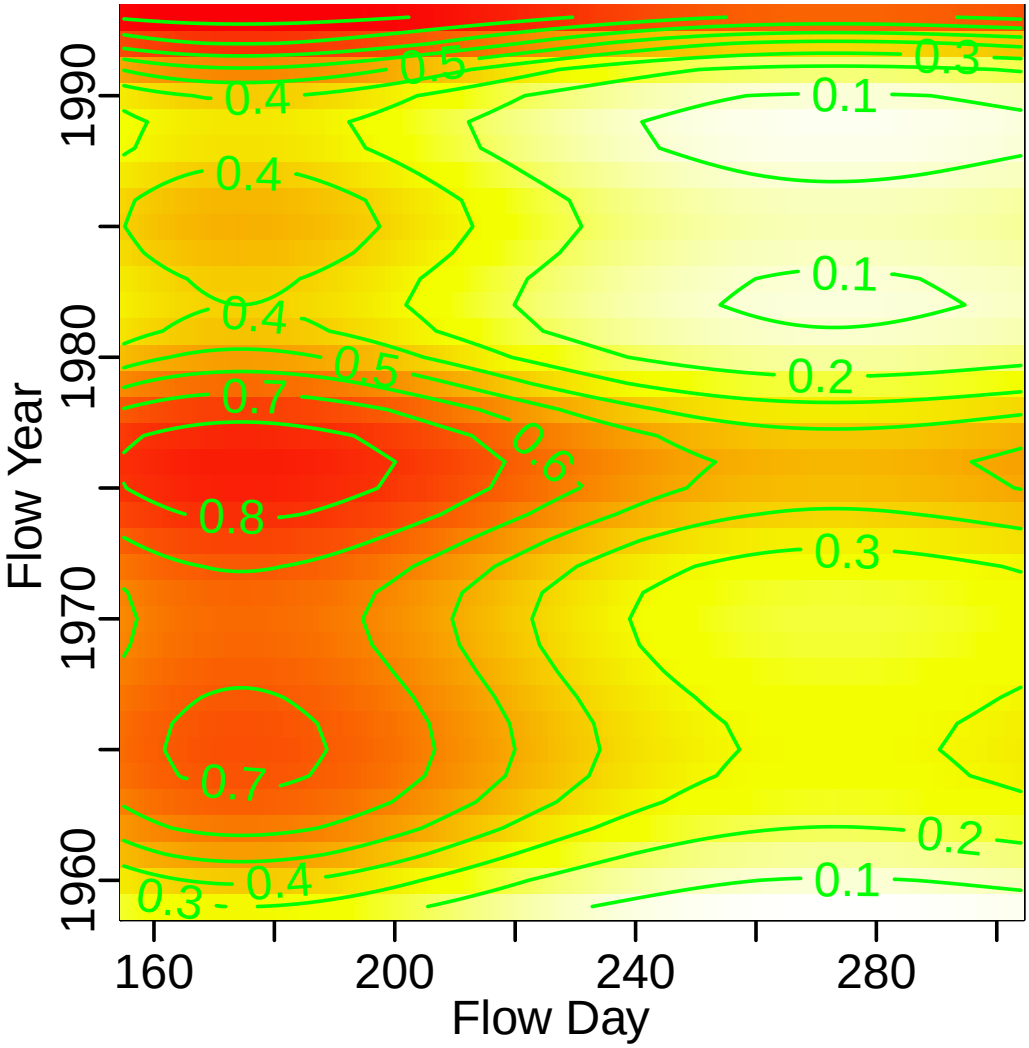

Figure 4.8: Contour plot for estimated probability of non-zero flow by day and year for Long Creek based on the May-October subseasonal approach using Model 3 (Equation 3.13a).

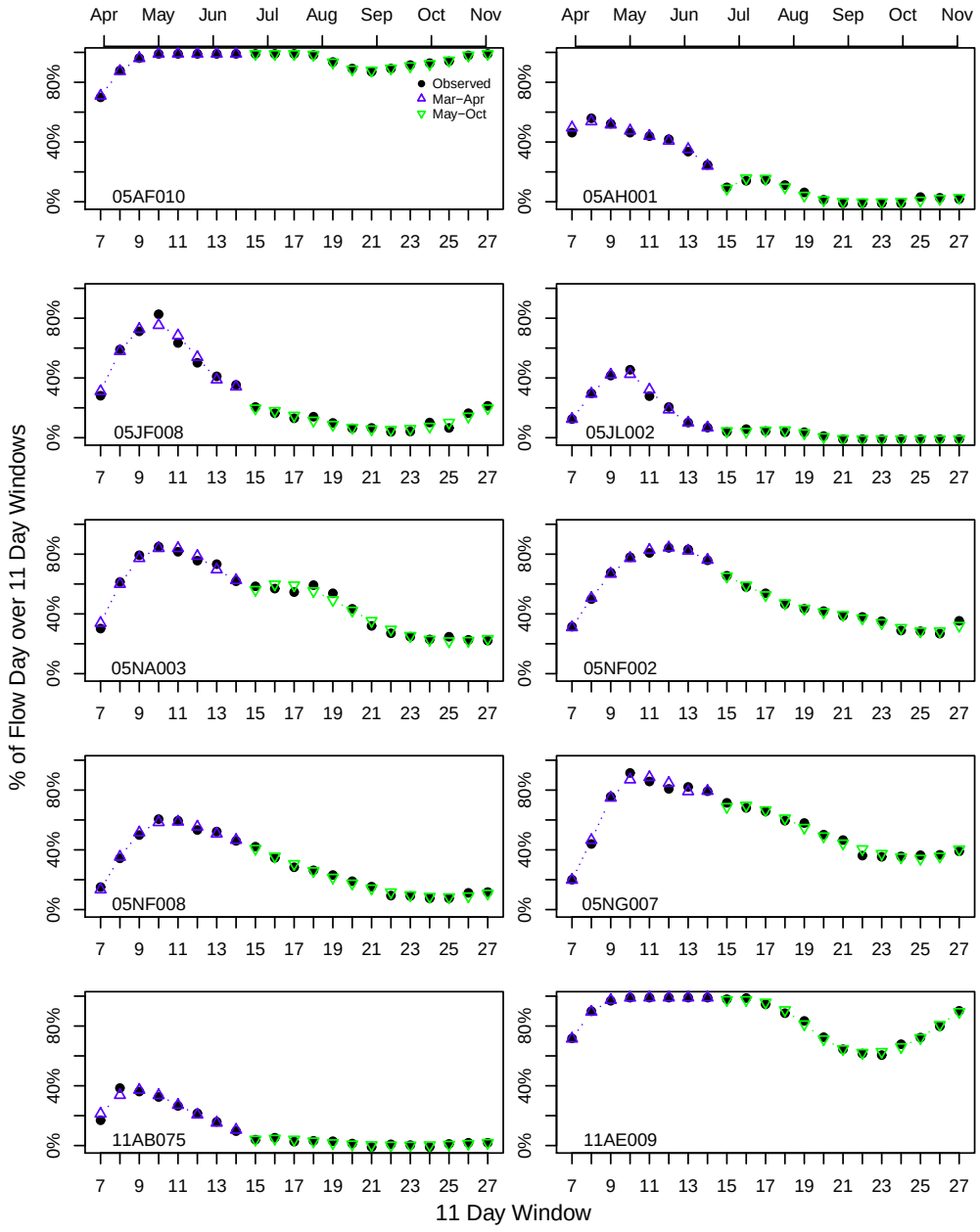

Figure 4.9: Observed and expected fraction of non-zero flow days over 11-day windows using Model 3 (Equation 3.13a) for all ten hydrometric stations

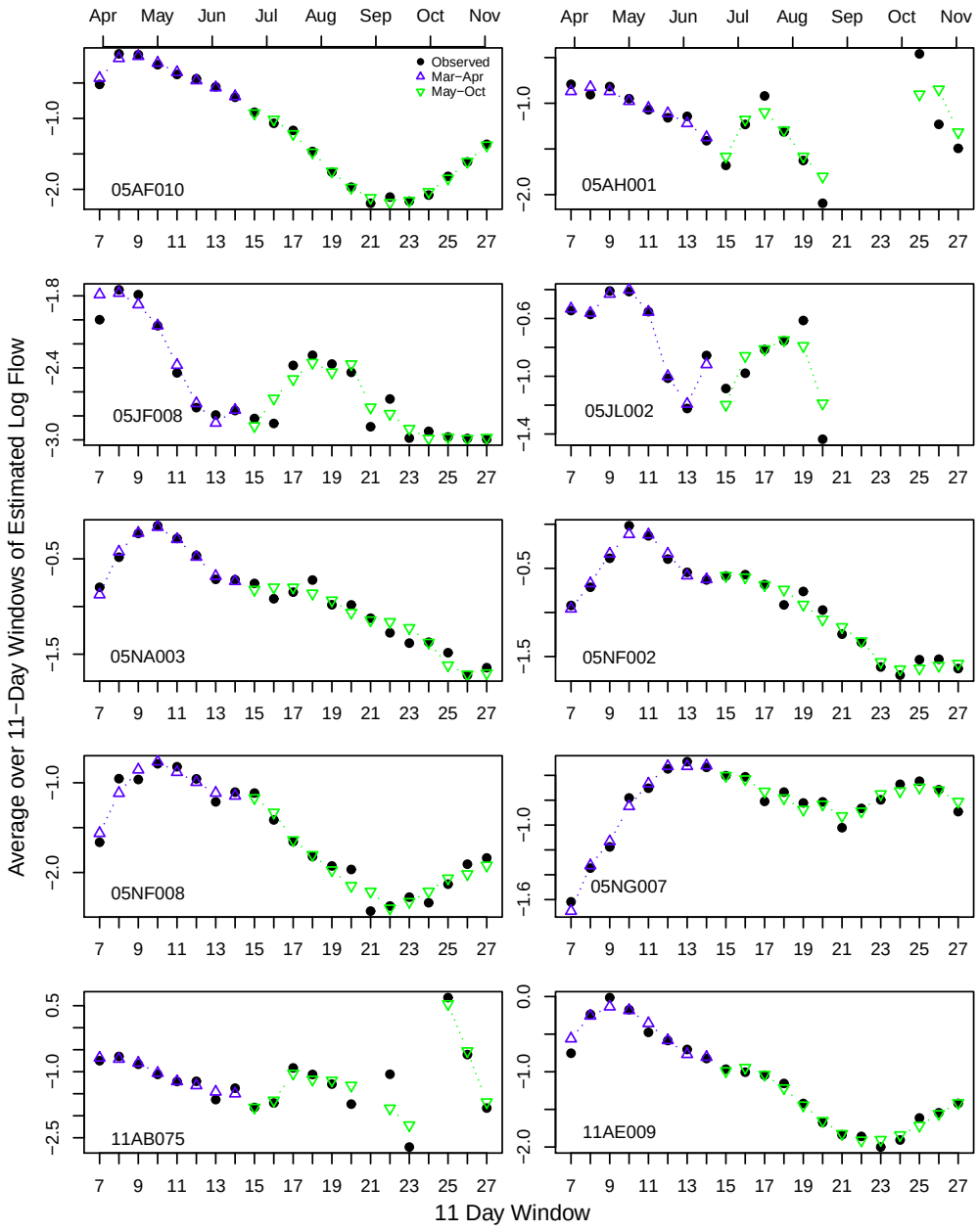

Figure 4.10: Observed and expected log of mean flow over 11-day windows based on Model 3 (Equation 3.13b) for all ten hydrometric stations.

Chapter 5

Discussion and Future Work

Interest in developing robust statistical models for seasonally observed temporary rivers focuses on being able to assess how or if these models alter our current understanding of the hydrological attributes. Indicators of low and zero flows is one such area as drought is of considerable interest; however, it would is desired to consider other areas of hydrological interest: return periods of floods; detecting and/or prediction change due to land-use or climate variations and for predictions in ungauged basins. These include the time between extreme events; duration of extreme events; total flow volume in an event; determining what sorts of climate variations would/could introduce specific vulnerabilities in the system.

The statistical models developed herein allow for a trend in the mean of the response and for seasonality in the response. Goodness-of-fit statistics revealed fair predictions. This preliminary model is proposed as the basic framework to build future extensions. In particular, an extension that incorporates autoregression in the residuals will be developed. Furthermore, as two significant peaks were observed for the log-odds of non-zero flow data and the logarithm of a non-zero flow rate at approximately the same time of the year, the model for a non-zero flow day should be linked with the model for the logarithm of a non-zero flow rate via a common random effect that is autoregressive in nature, to more tightly connect these two parts of the model. These extensions should lead to gains in precision in the analysis. With these extensions incorporated, simulations and other methods for evaluating hypotheses relating to the underlying hydrological processes will be considered in future work.

Another approach would be to use transitional models and incorporate smoothers into the transition rate. Future work will also compare transitional and marginal models though

we note that these models are conceptually different. The approach could also be extended to handle the joint modelling of several streams. Cigizoglu et al. (2002) characterized the joint distribution of flow events at two neighbouring river sites in arid and semi-arid regions using a combination of point process models for flow episodes separated by periods of zero flow, the first to model the clustering of these two compartments, and the second to model explicitly the clustering of events in periods of flow episodes. Bivariate modelling was necessitated by the fact that spate irrigation schemes, that is schemes that take advantage of sudden rushes of water, often exploit the runoff from a number of adjacent river channel systems which is then diverted into a single crop growing area. Wang and Robertson (2011) also used a joint probability modelling approach for forecasting flows at multiple sites in temporary streams by treating zero flow occurrences as censored data. Adapting these methodologies in the context of our spline compartment model may be appropriate for forecasting intermittent flow based on more than one site.

Statistical models for seasonally observed data that is zero-rich that are sufficiently robust are necessary for analysis and interpretation of the type of data available for most Prairie streams. Such models will greatly improve our ability to predict stream flow in ungauged basins in the Canadian Prairies. Further, these models can also provide the basis for detecting changes related to climate or landuse changes. The model development presented here captures seasonality, extreme events, sequences of no flow days, and regular seasonal patterns of spring snowmelt and summer convective rainfall. These models provide the basis upon which more complex models that include autocorrelation and joint probability modelling can be developed.

Bibliography

[1] H. Aksoy and M. Bayazit. A daily intermittent flow simulator. *Turkish Journal of Engineering*, 2000.

[2] H. Aksoy and M. Bayazit. A model for daily flows of intermittent streams. *Hydrological Processes*, 2000.

[3] B. Brimley, J-F. Cantin, D. Harvey, M. Kowalchuk, P. Marsh, T.M.B.J. Ouarda, B. Phinney, P. Pilon, M. Renouf, B. Tassone, R. Wedel, and T. Yuzyk. Establishment of the reference hydrometric basin network (RHBN) for Canada. *Environment Canada*, page 41, 1999.

[4] J.M. Buttle, S. Boon, D.L. Peters, C. Spence, H.J. van Meerveld, and P.H. Whitfield. An overview of temporary stream hydrology in Canada. To be published in Canadian Water Resources Journal, 2012.

[5] Environment Canada. HYDAT Surface water and sediment data. *Ottawa: Environment Canada, Water Survey of Canada (2011)*, 61:161–175, 2011.

[6] J. Cataldo, C. Behr, F. Montalto, and R.J. Pierce. A Summary of Published Reports of Transmission Losses in Ephemeral Streams in the U.S. *A Report to the National Center for Housing and the Environment, Wetland Science Applications Inc., Poolesville, MD*, page 42, 2004.

[7] M. Chebaane, J.D. Salas, and D.C. Boes. Product periodic autoregressive processes for modeling intermittent monthly flows. *Water Resources Research*, 31:1513–1518, 1995.

[8] H.K. Cigizoglu, P.T. Adamson, and A.V. Metcalfe. Stochastic modelling of ephemeral flow. *Hydrological Processes*, 16:1451–1465, 2002.

[9] P. Craven and G. Wahba. Smoothing Noisy Data With Spline Functions. *Numerische Mathematik*, 31:377–403, 1979.

[10] D. Dissing and G. Wendler. Solar radiation climatology of Alaska. *Theoretical and Applied Climatology*, 61:161–175, 1998.

[11] T. Hastie and R Tibshirani. Generalized Additive Models. *London: Chapman and Hall*, 1990.

[12] L. Le Barbé and T. Lebel. Rainfall climatology of the HAPEX-Sahel region during the years 1950-1990. *Journal of Hydrology*, 188-189:43–47, 1997.

[13] S Lee. Stochastic generation of synthetic streamflow sequences in ephemeral streams. *de l'Association Internationale des Sciences Hydrologiques Symposium de Tokyo*, 117:691–701, 1975.

[14] C. L. Mallows. Some comments on Cp. *Technometrics*, 15:661–675, 1973.

[15] D. McGee, S. Boon, and H.J. van Meerveld. Impacts of rural water diversions on Prairie streamflow. To be published in Canadian Water Resources Journal, 2012.

[16] W.A. Monk, D.L. Peters, R.A. Curry, and D.J. Baird. Quantifying trends in indicator hydroecological variables for regime-based groups of Canadian rivers. *Hydrological Processes*, 25:3086–3100, 2006.

[17] L. Romolo, T.D. Prowse, D. Blair, B.R. Bonsal, and L.W. Martz. The synoptic climate controls on hydrology in the upper reaches of thepeace river basin. Part I: snow accumulation. *Hydrological Processes*, 20(19):4097–4111, 2006.

[18] L. Romolo, T.D. Prowse, D. Blair, B.R. Bonsal, and L.W. Martz. The synoptic climate controls on hydrology in the upper reaches of thepeace river basin. Part II: snow ablation. *Hydrological Processes*, 20(19):4113–4129, 2006.

[19] T.H. Sparks and A. Menzel. Observed changes in seasons: an overview. *International Journal of Climatology*, 22:1715–1725, 2002.

[20] R. Srikanthan and T.A. McMahon. Stochastic generation of monthly flows for ephemeral streams. *Journal of Hydrology*, 47:19–40, 1980.

[21] Q.J. Wang and D.E. Robertson. Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research*, 47:W02546, doi:1029/2010WR009333., 2011.

[22] P.H. Whitfield and A. Cannon. Recent Variations in Climate and Hydrology in Canada. *Canadian Water Resources Journal*, 25:19–65, 2000.

[23] S Wood. *Generalized Additive Models: An Introduction with R.* Boca Raton, FL: Chapman-Hall, 2010.

[24] S Wood. mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL. *Available at http://cran.r-project.org/package=mgcv*, 2010.

[25] S. N. Wood and N. H. Augustin. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157(2-3):157–177, 2007.