Fall 2021

Using Concurrent Functional Regression to Reconstruct River Stage Data During Flood Events and Identify Influential Functional Measurements

Ryan Pittman

Using Concurrent Functional Regression to Reconstruct River Stage Data during Flood Events and Identify Influential Functional Measurements

by

Ryan Pittman

Bachelor of Science
Anderson University (SC) 2017

Master of Science
University of South Carolina 2019

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2022

Accepted by:

David B. Hitchcock, Major Professor

John M. Grego, Major Professor

Karl B. Gregory, Committee Member

S. Scott Sutton, Committee Member

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

ABSTRACT

On October 4, 2015, the Cedar Creek gage at Congaree National Park stopped reporting stages, and the readings did not resume until approximately two weeks later because of record-breaking rainfall that led to some of the worst flooding in South Carolina history. Our goal is to reconstruct the Cedar Creek stage during this missing two-week window. The Congaree River gage in Congaree National Park remained functioning throughout the October 2015 flood, when the stage reached its maximum recorded crest. The stages from the two gages are directly related during floods as water travels through the local spillways and flood planes to connect the two locations. We introduce a new method called Landmark Aligned L_1 (LAL_1) distance to objectively determine the start and end points of each of the 10 flood events in the sample and then use these events to reconstruct the missing Cedar Creek stage. This alignment substantially improves the accuracy of the reconstruction and reduces the related prediction interval for the target event. We treat the stage as functional data and use a concurrent functional model to establish the relationship between the two locations during each timepoint of prior flood events. Once this relationship is found, the known Congaree stage from October 2015 is used to reconstruct the missing Cedar Creek stage during the 2015 flood. The results show that the novel LAL_1 distance data selection method is effective, and that there is a strong functional relationship between the two locations. Based on our reconstruction, we estimate that the crest of Cedar Creek reached a historic high in October 2015, with stages exceeding 17 feet, compared to a previous high of just over 16 feet. Furthermore, the next aim of this project is to determine which of the functional observations are most influential to the fitted concurrent model and

reconstruction/prediction. We modify preexisting linear regression measures of influence ($DFFITS$, $DFBETAS$, Cook's Distance) and create two additional metrics (Δ and AIP) to measure the sensitivity of the reconstruction and the impact of the known prior flood events. These functional measures can be used independently or in conjunction to identify the functional observations with the largest influence. Lastly, we introduce a weighted bootstrapping (with perturbations) method to approximate a null distribution for each influence measure to assess the significance level of the influence for each observation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

On October 3-4, 2015, Columbia, South Carolina and other areas of the state experienced record-breaking rainfall. Over that two-day period the Columbia Metro Airport saw 10.28 inches of rain, far exceeding the previous two-day record of 7.69 inches set in 1949 (National Weather Service, 2015). The result of this record rainfall was some of the most severe flooding in South Carolina history, leading to about $12 billion in damages across the state (Burris, 2015). Among the casualties of the storm was the water gage (United States Geological Survey, 2020b) that measured the Cedar Creek stage, in Richland County, South Carolina. At 11:00 PM, on October 4, the gage stopped reporting stages, and the readings did not recommence until they sporadically appeared, beginning approximately two weeks later (see Figure 1.1). The initial goal of this project is to reconstruct the Cedar Creek stage during the two-week window when the river stage was not recorded. Stage is the water level above an arbitrarily chosen reference datum, typically measured in feet (United States Geological Survey, 2019b). Gage heights can be used for a variety of reasons: "flood prediction, water management and allocation, engineering design, research operation of locks and dams, and recreation safety and enjoyment" (United States Geological Survey, 2019a). In this case, knowing the height at the Cedar Creek gage allows us to see how that portion of the river was behaving during the peak of this catastrophic flood.

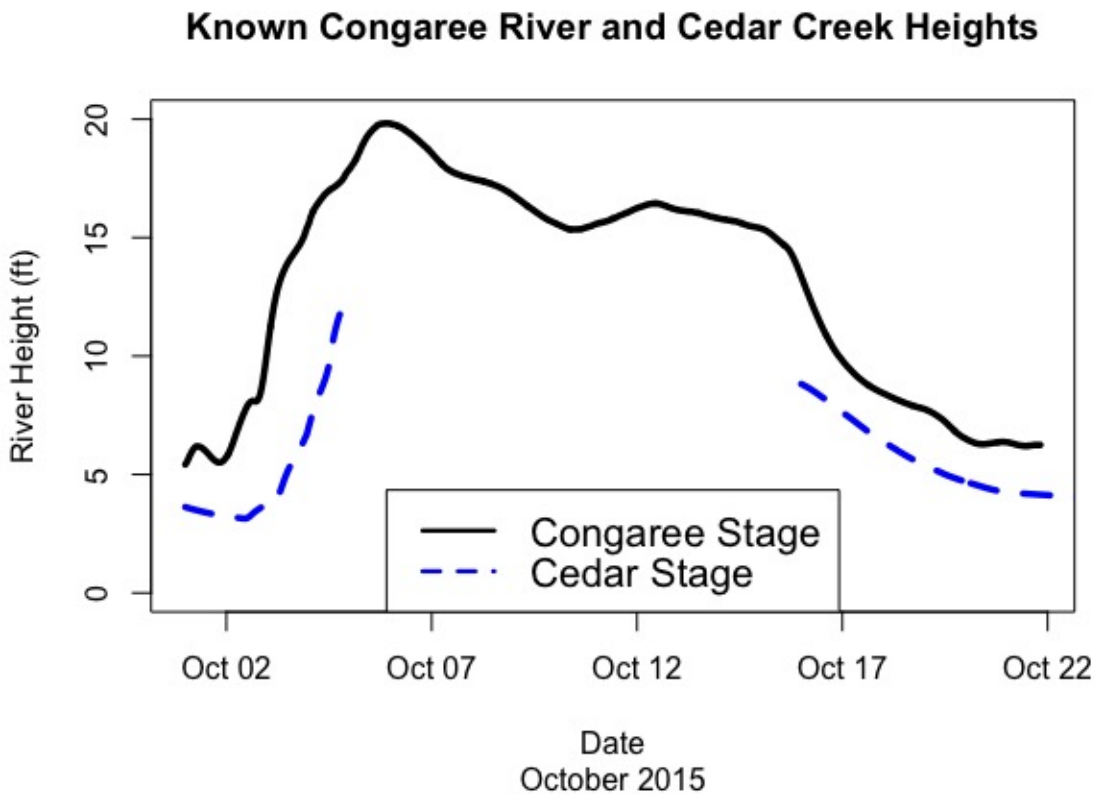


Figure 1.1. Observed river stages for the Congaree River (Solid Line) and Cedar Creek (Dashed Line) during the major October 2015 flood event in Columbia SC: Note the missing portion of the Cedar Creek height.

Our approach is to use the observed heights at a gage in the Congaree River to reconstruct the river height at the missing gage location. The Congaree River gage at Congaree National Park (United States Geological Survey, 2020a) remained functioning throughout the October 2015 flood. This gage is located a few miles west of the Cedar Creek gage. Figure 1.2 highlights the location for each gage (National Park Service, 2019).

During a flood, the Congaree River flows overbank and moves through the local natural floodplain channels, through the wetlands, into Cedar Creek. Therefore, if a functional relationship between river stages can be established for other similar floods in the past, then the missing river stage at Cedar Creek can be reconstructed using the known Congaree River heights.

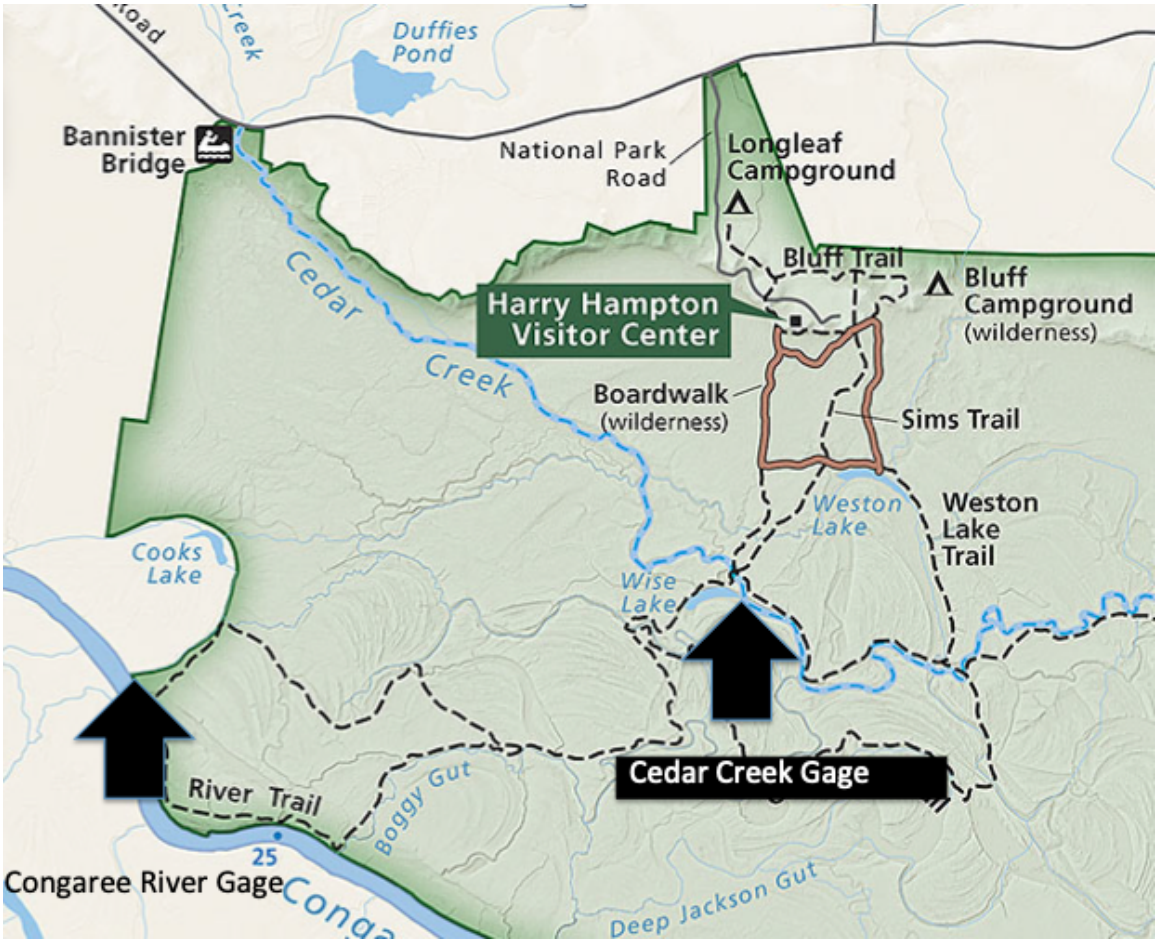


Figure 1.2. Map of Congaree National Park along with the approximate location of both of the gages used in this study.

Once we have implemented our novel historical curve selection procedure, we will employ functional data analysis (FDA), which is appropriate when the variables can naturally be viewed as smooth curves or functions. "FDA can ... be thought of as the statistical analysis of samples of curves" (Kokoszka and Reimherr, 2017). Therefore, FDA can be applied to the river height data in order to establish the relationship between the Congaree River gage values and Cedar Creek gage values to obtain the missing river stage function.

The employment of functional regression to handle data that is best treated as functional data rather than discrete observations is becoming more common in a variety of fields. Authors such as Ramsay and Silverman (2005), Kokoszka and Reimherr (2017),

and Ramsay, Hooker, and Graves (2009) present numerous techniques used to analyze functional data. The functional regression model was implemented by Das et al. (2019) to create a method that improves the accuracy of total hemoglobin (SpHb) monitors; it is a noninvasive hemoglobin monitoring tool that aids in creating better critical care protocols in trauma care. Zhang, Clayton, and Townsend (2011) used functional concurrent linear regression for spatial images. They related information from a set of spatial images to study forest nitrogen cycling. Wang et al. (2019) take a more robust approach to functional regression to forecast wind speed using multiple functional variables as inputs. FDA was also used by Ferraty, Rabhi and Vieu (2005) to regress scalar response variables on an explanatory variable that should be treated as functional in order to obtain conditional quantiles during an El Niño event in 1998. Ramsay et al. (2009) took hip and knee angle data from a joint rotation study conducted by Olshen et al. (1989) and used FDA to establish the relationship between hip and knee angles for children at corresponding time points as they walk.

Moreover, FDA has been used to describe river data similar to ours. Masselot et al. (2016) used functional regression to forecast streamflow. Streamflow is naturally a continuous variable with respect to time, as are the meteorological variables which influence it, and thus functional regression models can be created to forecast streamflow. In particular, Masselot et al. were interested in forecasting autumn streamflow and used meteorological data such as precipitation curves. Their results indicated that functional linear models perform better than neural networks when predicting the shape of hydrographs. Chebana, Dabo-Niang, and Ouarda (2012) analyzed stream flow as functional data, using data from hydrographs to adapt a model to deal with floods and droughts. While applying their techniques to data obtained from Magpie Lake in Quebec, Canada, they concluded that FDA can safely be applied to floods as it performs a single analysis on the whole data, not several univariate or multivariate analyses. They do not create models for predictive or reconstructive purposes, but they do recognize that as a poten-

tial future study, indicating that FDA is a reasonable approach for reconstructing flood curves. Our study will use functional regression to analyze floods; however, instead of using stream flow, we use river stages as our variables.

In addition to applying our reconstruction method to the Cedar Creek stage during the October 2015 flood, we will create and apply several measures that will help to better understand the results of the reconstruction. Some of the measures will determine which of the historic flood events were the most influential in the reconstruction of the missing Cedar Creek stage. Many of these will be modeled after the traditional linear regression measures of influence and leverage, including $DFFIT(s)$, $DFBETA(s)$, Cook's Distance, and the "Hat" matrix, but these will be extended to be applicable to the functional regression model (Belsley, Kuh, and Welsch, 1980). Additionally, we will propose two new metrics, Δ and AIP, that also take into account the predictor curve for the out-of-sample observation that needs to be reconstructed/predicted instead of solely focusing on the complete historic events. After describing these new influence measures in detail, we describe a weighted bootstrapping (with perturbations) method to approximate a null distribution for each measure within a given dataset to establish a significance level associated with each measure for each observation. We will also discuss identifying the most influential portions of the functional observation on the prediction. This could help determine where the landmark alignment should be focused in the future and where a slight change in the predictor curve could have a massive impact on the response curve.

1.2 Structure of the Dissertation

The dissertation is structured as follows. In Chapter 2, we present our work published in *Environmental and Ecological Statistics*, on reconstructing the missing 2015 Cedar Creek stage data (Pittman, Hitchcock, and Grego, 2021). First we will discuss river data collection and a novel landmark alignment and selection process to determine the be-

ginning and end of observed flood events. Then we describe how to apply the concurrent functional model to our data to establish the relationship between the Congaree River stages and Cedar Creek stages, using this model to reconstruct the missing Cedar Creek Stage. In Chapter 3, we propose an applicable extension to this study, presenting numerous influence measures to use with the concurrent functional model to identify potentially significant observations. In Chapter 4, we introduce new influence measures also applicable to the concurrent functional model but where the focus is determining the influence each observation has on the prediction/reconstruction of an out-of-sample observation's functional response curve.

CHAPTER 2

RECONSTRUCT RIVER STAGE DATA DURING FLOOD EVENTS

2.1 INTRODUCTION

This Chapter will focus on the reconstruction of the October 2015 Cedar Creek heights using the fully functional concurrent model. Usually, the initial step in functional data analysis is to express the data through basis expansion

$$X_i(t) \approx \sum_{m=1}^{M} c_{im} B_m(t), \quad 1 \le i \le N \tag{2.1}$$

where $B_m(t)$, $m = 1,\ldots,M$ are a standard collection of basis functions such as spline, wavelets or cosine and sine functions and M is the number of basis functions used, with c_{im} being the corresponding coefficient. Also, i is the index for a specific curve, while N is the total number of curves (Kokoszka and Reimherr, 2017). Essentially, these M basis functions are created to replace the raw measurements for numerous practical purposes. Note that expressing the data using covariance eigenfunctions is an alternative method that can be used to expand the functional predictor (Baíllo and Grané, 2009). When the sets of timepoints at which the data are collected differ among subjects, basis expansion puts all of the curves into a common domain, making them easier to compare and analyze. Additionally, M will almost always be smaller than the number of observed timepoints, so basis expansion acts as a type of data reduction, where for each i, the specific X_i curve is represented by the column vector $\mathbf{c}_i = [c_{i1}, c_{i2}, \ldots, c_{iM}]^T$, of dimension M. In this study, we will allow our functional data to be expressed via Fourier basis functions and use an objective method to determine how many of them should be used to represent the data.

2.2 Data Collection and Landmark Aligned Selection

2.2.1 Locating flood events

Functional regression models are used to predict or explain a functional response $Y(t)$ using a functional predictor $X(t)$. One type of functional regression model is the concurrent model. The equation for this model is:

$$Y_i(t) = \beta_0(t) + \beta_1(t)X_i(t) + \epsilon_i(t), \ i = 1, \ldots, N \tag{2.2}$$

where the set of discretely measured functional observations can be written in matrix form as

$$\mathbf{X} = \begin{bmatrix} X_1(t_1) & \ldots & X_N(t_1) \\ X_1(t_2) & \ddots & \vdots \\ \vdots & \vdots & \vdots \\ X_1(t_n) & \ldots & X_N(t_n) \end{bmatrix} \tag{2.3}$$

and

$$\mathbf{Y} = \begin{bmatrix} Y_1(t_1) & \ldots & Y_N(t_1) \\ Y_1(t_2) & \ddots & \vdots \\ \vdots & \vdots & \vdots \\ Y_1(t_n) & \ldots & Y_N(t_n) \end{bmatrix} \tag{2.4}$$

In our case study, the goal is to find the relationship between the heights of the Congaree River and Cedar Creek during previous flood events and then to use the known Congaree River heights during the October 2015 flood to reconstruct the corresponding Cedar Creek stage function. In order to establish the relationship between gage values, we collect data from prior flood events for both the Congaree River gage values and Cedar Creek gage values. Nearly complete stage records from January 1, 1995 to April 30, 2020 were made available to us from members of the U.S. Geological Survey-Water Resources Division. These data can be found on `https://github.com/rpittman188/fdaconcur/tree/master/fdaconcur/R`. According to the National Oceanic and Atmospheric Administration (NOAA), the Congaree River at Congaree National Park is at

a moderate flood stage when it reaches 18 feet or more (National Weather Service and NOAA, 2019). Historical data shows that this threshold has been met only eight times, with a maximum height of 19.83 feet which happens to be during our flood of interest in October 2015. Another of the events, on January 1, 2016, does not have available corresponding Cedar Creek heights, and thus cannot be used in a regression model, leaving six usable events remaining. In order to include more historic floods, we loosened the cutoff to a crest of 17.85 feet, allowing us to use four more flood events. Further reducing the cutoff below 17.85 results in more incomplete and unavailable data and would permit events that may not be be considered true flood events. The list of historic crests for the Congaree River at Congaree National Park is in Table 2.1.

Table 2.1. Historic Congaree River Crests

Rank	Max Stage (ft.)	Date of Crest	Start Date and Time	End Date and Time
1	19.83*	10/05/2015	10/01/15 00:00	10/21/15 19:45
2	19.54	02/10/2020	01/31/20 11:30	03/13/20 13:00
3	18.65	03/23/2003	03/20/03 12:45	04/01/03 12:45
4	18.28*	01/01/2016	Not Used	Not Used
5	18.27	08/31/1995	08/26/95 12:00	09/07/95 00:00
6	18.20	02/06/1998	02/02/98 10:00	02/18/98 02:00
7	18.16	05/09/2013	05/05/13 03:30	05/15/13 22:30
8	18.16	09/11/2004	09/07/04 08:45	09/18/04 00:45
9	17.95	03/05/2007	02/28/07 18:15	03/16/07 23:30
10	17.90	11/18/2018	11/17/18 22:45	11/22/18 07:30
11	17.85	02/08/2010	01/24/10 05:00	02/06/10 00:00
12	17.85	05/25/2003	05/21/03 23:15	06/03/03 00:15

* Indicates no available corresponding Cedar Creek heights Maximum historic crests for the Congaree River gage at Congaree National Park, 1995-2020 and complete dates of observed flood records used in concurrent model

2.2.2 LANDMARK ALIGNED DATA SELECTION

After determining the dates of the peaks of interest, we need an objective method for selecting each flood event's starting and ending point. In the concurrent model, the selected flood events should be aligned as closely as possible, which will enable a more ac-

curate prediction and narrower prediction interval for the predicted October 2015 Cedar Creek curve. Since our particular goal is to use the Congaree stage to reconstruct the Cedar Creek stage during the October 2015 flood, the curves for the past events used in the model should resemble this October 2015 event as closely as possible. Additionally, since the stages for these two locations are more strongly related when the Congaree stage is high (when the river overflows across the floodplains into Cedar Creek), we place more emphasis on aligning the curves at the higher stages of the events. This motivates our novel Landmark Aligned L_1 distance (LAL_1) approach.

Landmark Aligned L_1 distance is based on traditional L_1 distance between two curves:

$$d_1 = \int |a(t) - b(t)| dt \tag{2.5}$$

which we estimate via trapezoidal approximation, using the function `trapz` in the `pracma` package in R (Borchers, 2019). Here t is the index of the flood, which for our discretely observed data, ranges over the number of measurement points of the target event's curve $b(t)$, and $a(t)$ represents one of the selected raw curves that needs to be aligned with the target event's curve.

A method of flood event definition that simply uses L_1 distance is given in the appendix; however, this L_1 distance-based method is inadequate for selecting start and end times of some of the events that have multiple peaks.

Our new LAL_1 approach places more weight on aligning the highest sections of the stage curves. This selection method starts with a single untrimmed flood event, and systematically trims the raw event to define the starting and ending points of each complete flood event (denoted, say $X(t)$) in order to minimize the LAL_1 distance between each event and the target event of interest (October 2015), according to the following criterion:

$$LAL_1 = \int |X(t) - X^*(t)|[X^*(t)^2] dt \tag{2.6}$$

Here, $X^*(t)$ is the October 2015 Congaree River height ranging from October 1, 0:00 to October 21, 19:45. The discretely measured observations are spaced 15 minutes apart,

leading to 2000 total observations. By multiplying the absolute difference by the square of the Congaree stage at each t before approximating the integral, the LAL_1 distance is heavily influenced by the distance between $X(t)$ and $X^*(t)$ when $X^*(t)$ is at its highest points. As a result, the selected $X(t)$ curve that minimizes this LAL_1 distance will resemble the target $X^*(t)$ curve at the higher sections of $X^*(t)$ much better than had we chosen the start and end points using standard unweighted L_1 distance.

Note that it is rather common in functional data analyses to pre-smooth the observed curves before the analysis (Ramsay and Silverman, 2005). This is done to eliminate roughness in the curves arising from natural variability or measurement error. In our case, we choose to perform the LAL_1 alignment on the raw stage curves. This is justified in our situation since the highly frequent (every 15 minutes) measurements of the stage result in a function that, viewed on the scale of the multi-day flood event, is already intrinsically smooth. Furthermore, there are reasons to treat the measurement error as negligible in our case: The USGS requires that "stage accuracy requirements are stringent"; the overall accuracy of stage data established for USGS gaging stations is the greater of 0.01 foot or 0.2 percent of the effective stage (Sauer and Turnipseed, 2010). While measurement error could be an issue in small, urbanized watersheds with a high percentage of impervious cover, the Congaree watershed is large, while Cedar Creek's watershed is rural with little impervious cover, and the gage is located in the interior of a floodplain where any measurement issues caused by local runoff have long since been attenuated farther upstream.

2.2.3 Applying landmark aligned data selection

We now describe our user-created `LaL1.align` R function (available at `https://github.com/rpittman188/fdaconcur/tree/master/fdaconcur/R`) to define the start and end times of our flood events. In our case study, there are 10 usable historical flood events. For each event, the date of the Congaree River crest is known. We begin with an exces-

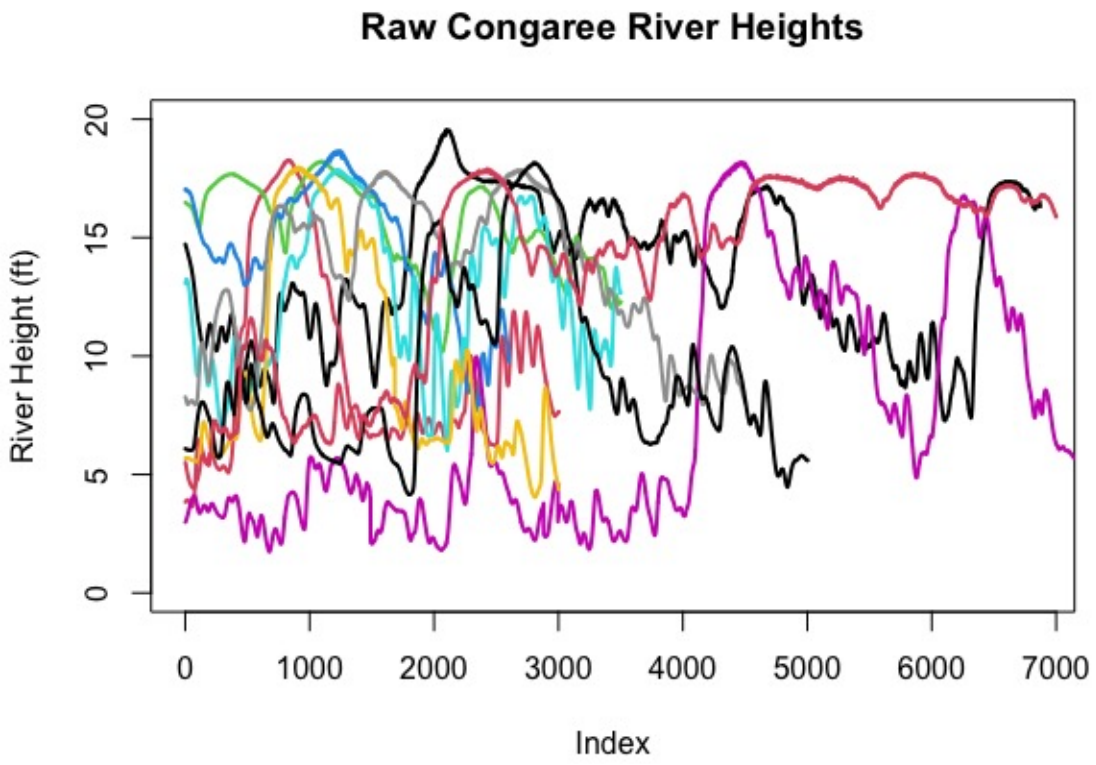


Figure 2.1. Raw Congaree River stage curves for all ten of the available flood events prior to using the selection method.

sively long timeframe of stage measurements before and after the crest of each flood event. We alternately remove one point from the beginning of the raw event and then from the end; which of these "trims" is used is based on which produces a smaller LAL_1 distance between the trimmed curve and the target (after interpolating to make the resulting vector the same length as the target vector). This process of trimming from either the beginning or the end of the event's curve repeats until it has trimmed the entire vector for the event in question. Then the pair of beginning and ending indices that had yielded the lowest LAL_1 distance from the target event is selected, which defines an event that best resembles the target October 2015 event.

We now illustrate the effect of the algorithm to define our flood events' start and end times. The raw Congaree River stage curves for the 10 full flood events are shown

in Figure 2.1. They are very dissimilar, with different patterns, maximum heights, and lengths. These raw events are not suitable for the concurrent model. In contrast, Figure 2.2 displays the 10 Congaree River stage curves after defining the start and end times of each flood event based on the LAL_1 alignment approach. The similarity among the curves that arise from this careful definition of the flood event timeframes will allow a much better reconstruction of the October 2015 Cedar Creek curve via the concurrent functional regression model. Once the dates and times of the best starting and ending points of each event are established based on the Congaree heights, the corresponding Cedar Creek stage height is observed from that start time until that end time, as seen in Figure 2.3 for the February 2020 event.

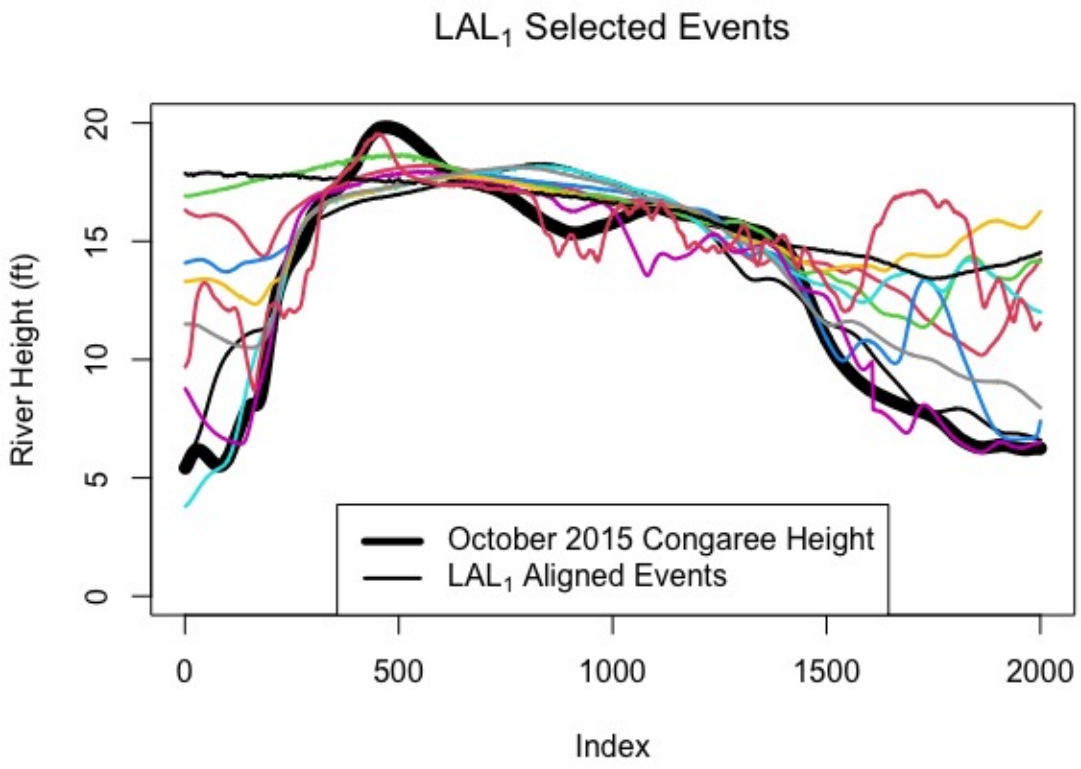


Figure 2.2. All 10 LAL_1 selected Congaree River curves (Colored Lines) aligned with the target October 2015 Congaree River event (Black Line).

In order to implement the concurrent model, the discretized curves for all flood

events must be the same length as each other and as the target event (the October 2015 event). In practice, we will use interpolation within each curve to attain a common set of measurement points across the set of curves. Since in reality, the flood events all have different durations in terms of real clock time, we will define the "timepoints" of our adjusted flood event curves in terms of fractions of the flood event duration. This is a common approach in alignment and registration of functional data (see, for example, the "time-warping" approach of Kokoszka and Reimherr (2017)), and it does not hinder the analysis of the relationship between the Congaree River curves and the corresponding Cedar Creek stage curves. Finding the best way to adjust for the variation in the durations of the functional observations is one of the major contributions of this approach.

Again, since the Congaree River and Cedar Creek are most closely related when the Congaree River is at its highest stage, the curves' differences in Figure 2.2 towards the beginnings and ends of the events are not troubling. In other data scenarios where every section of the event is equally relevant, the start and end times could be selected using standard L_1 distance methods (such an alternate approach is implemented for these data in the appendix).

The complete starting and ending points of these ten events are found in Table 2.1. These ten "complete" flood events make up the dataset that we use to establish the functional regression relationship between the gage heights. We note that the untrimmed February 2010 event was quite sporadic, having three local maxima in a very short period of time. The crest of the trimmed flood event that was selected by our method is not the global maximum, but is only 0.08 feet less than the highest peak. Also, for the November 2018 event, the flood event defined based on the true minimum LAL_1 distance is only five days long. We note that uniquely for this event, other choices of starting and ending points led to a very similar LAL_1 distance between it and the October 2015 Congaree stages. While visually the other selection options looked more like

a full flood event, we found that replacing the five-day definition of this flood with a lengthier event definition had virtually no impact on the final results; therefore, for the purposes of this study, we chose to use the shorter November 2018 defined event that truly minimized the LAL_1 distance.

Once the start and end dates for the flood events were found, we input the Congaree River stage values into the **X** matrix in Equation (2.3) and the corresponding Cedar Creek curves into the **Y** matrix in Equation (2.4), in order to fit the concurrent model. There is a visually clear association between the two curves, as seen in Figure 2.3, which shows the Congaree River stage values and Cedar Creek stage values for the February 2020 event, and the notable association between the curves in this plot is evident in all ten flood events.

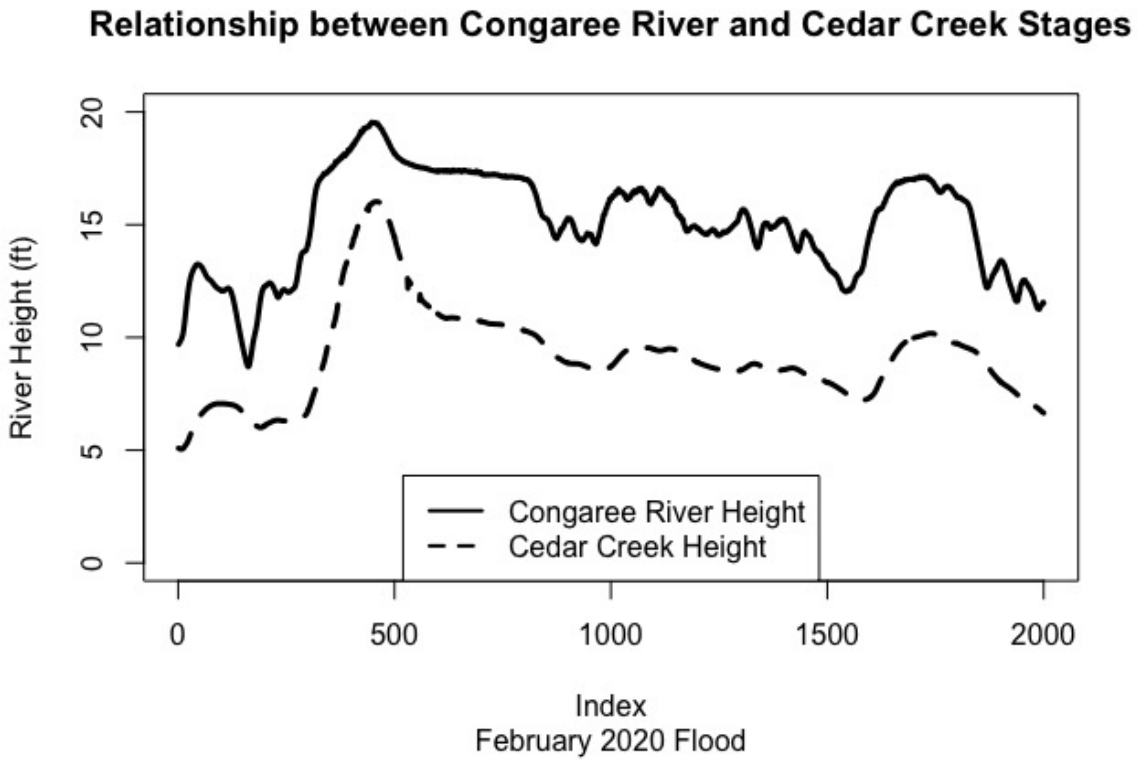


Figure 2.3. Full, known stages for Congaree River (Solid Line) and Cedar Creek (Dashed Line) during the February 2020 flood event

We briefly note that the dataset required that three feet be added to Cedar Creek stage values prior to October 1, 1998, because of a change in the Cedar Creek gage's measurement baseline on that date, as evidenced by an abrupt shift in gage height from 1.44 feet to 4.44 feet on October 1, 1998 (the start of the new water year). These ten "complete" flood events make up the datasets that we use to establish the relationship between the gage heights.

2.3 Implementing FDA on the Gage Height Data Using the `fRegress` function

We employ the `fRegress` function from the `fda` package (Ramsay, Graves, and Hooker, 2020) to fit the concurrent model in `R` (R Core Team, 2020). This function can be applied to a scalar dependent variable model or the concurrent functional dependent variable model, the latter of which applies to our case study.

In this model, the value of the response curve $Y(t)$ depends on the value of the regressor curve at the same time t (hence the name concurrent). In order to fit the concurrent model using `fRegress`, the vectors representing the discretized functional observations for all ten flood events must be the same length, as previously stated. The operation of interpolation to attain a common set of measurement points across the flood events has a similar effect as time warping, (Ramsay, Hooker, and Graves, 2009), in that chronological time is adjusted across the sampled curves to yield a time domain more convenient for the functional data analysis. Since the goal is to establish a relationship between the Congaree River and Cedar Creek at all the regions of the flood events' domains, as long as the floods' interpolated functional observations are aligned well, the concurrent model is appropriate to use.

2.3.1 PARAMETER SELECTION FOR FUNCTIONAL REGRESSION

Once the datasets have the same number of timepoints, the functional data analysis can be implemented using the `fRegress` function. Obtaining estimates for the regression coefficient functions $\beta_0(t)$ and $\beta_1(t)$ from Equation (2.2) is a necessary first step, and we will use these estimates to reconstruct the missing October 2015 values for the Cedar Creek gage (and obtain prediction intervals). To estimate $\beta_0(t)$ and $\beta_1(t)$, we must select an appropriate smoothing parameter. Since the data are collected at discrete points, the smoothing operation is the first step in converting the discretized functional data stored in $\mathbf{X}$ and $\mathbf{Y}$ into functional objects. The smoothing parameter (denoted by λ) measures the trade off between fit to the data and the variability of the smooth curve (Ramsay and Silverman, 2005). If the chosen λ is too small or too large, the smoothed curves will not represent the data well; therefore, selecting the correct value of λ is an important step in converting the raw discrete data to a functional object and estimating $\beta_0(t)$ and $\beta_1(t)$. To select the proper value of λ, Ramsay et al. (Ramsay, Hooker, and Graves, 2009) suggest generalized cross-validation (GCV), originally developed by Craven and Wahba (Craven and Wahba, 1978/79). The best choice for λ is the value that minimizes

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right) \tag{2.7}$$

Ramsay et al. (Ramsay, Hooker, and Graves, 2009) also provide `R` code to produce a plot over a grid of $\log_{10}(\lambda)$ to identify the value of λ that minimizes $GCV(\lambda)$.

Additionally, we must select the optimal number of Fourier basis functions to best represent the data as shown in Equation (2.1). Since our main goal is to use the concurrent model for prediction, we used an L_2-distance leave-one-out cross-validation to determine the number of Fourier basis functions that minimizes the L_2-distance (averaged over all flood events) between the true response curve and the same event's predicted (in a leave-one-out manner) response curve. Each distance is calculated by using

a trapezoidal approximation of

$$d_2^{(cv)} = N^{-1} \sum_{i=1}^{N} \int (Y_i(t) - \hat{Y}_{i(i)}(t))^2 dt \tag{2.8}$$

where $Y_i(t)$ is the true i-th response curve and $\hat{Y}_{i(i)}(t)$ is the predicted response function for the i-th event (predicted with a functional regression model fitted using all the events except the i-th event).

Once we have selected the smoothing parameter and an appropriate number of Fourier basis functions to use, we can fit the concurrent model to the river height data and obtain estimates $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ using the `fRegress` function. Additionally, we obtain pointwise 95% confidence intervals for $\beta_0(t)$ and $\beta_1(t)$. The `fRegress` function also produces estimates of the residual covariances and confidence limits for both $\beta_0(t)$ and $\beta_1(t)$. These $\beta_0(t)$ and $\beta_1(t)$ estimates can then be used to reconstruct the October 2015 Cedar Creek stage using the known October 2015 Congaree River stage using Equation (2.9):

$$\hat{Y}_i(t) = \hat{\beta}_0(t) + \hat{\beta}_1(t) X_i(t), \ i = 1, \ldots, N \tag{2.9}$$

2.4 Auxiliary Functions

We now describe several `R` functions created to quickly calculate quantities described in the prior sections. The functions in their entirety are available via `https://github.com/rpittman188/fdaconcur/tree/master/fdaconcur/R`.

2.4.1 `LaL1.align` function

The `LaL1.align` function takes the target curve of interest and an additional event of interest and determines the optimal beginning and ending points of the trimmed event that minimize the Landmark Aligned L_1 (LAL_1) distance between that curve and the target event. It then returns a vector of the trimmed additional event that is the same length as the main curve. For maximum performance, input the timeframe of the secondary event to be much wider than needed, with roughly equal-sized tails on each side

of the expected relevant portion of that event, and allow the algorithm to narrow the timeframe down to the most significant portion of the secondary event based on the target event.

2.4.2 PredictFRegressNormTest FUNCTION

The `PredictFRegressNormTest` function takes a matrix of discretized explanatory functional variables along with a corresponding response matrix to estimate the slope and intercept curves in the concurrent model. Additionally, the function allows the user to choose the number of Fourier basis functions and to specify the smoothing parameter λ. Most importantly, we can also include an additional predictor vector (for a new functional observation) that the function will use to create a predicted response curve for that new functional observation and a 95% prediction interval that is calculated using parametric bootstrapping. The construction of the interval using the parametric bootstrapping method is described in the next section.

2.4.3 L2Error.fRegress FUNCTION

The `L2Error.fRegress function` calculates the L_2 distance d_2 when the user inputs a predictor matrix $\mathbf{X}$, response matrix $\mathbf{Y}$, a new predictor vector, and the corresponding true response vector. This function fits the concurrent model to get a predicted response and then calculates the L_2 distance between the predicted responses and the true responses at each time point, using trapezoidal approximation to calculate the distance over all time points and to ensure that the data are treated as continuous rather than discrete. This function is used in conjunction with the following `L2bestEst` function.

2.4.4 L2bestEst FUNCTION

The `L2Error.fRegress` function also allows the user to specify the basis type and number of basis functions M (See Equation (2.1)). The `L2bestEst` function is used to choose

the optimal number of basis functions by finding the number that yields the smallest average L_2 distance across all of the events (this is $d_2^{(cv)}$ from Equation (2.8)). This function takes as its input **X** and **Y**. During each pass through a loop, one column (corresponding to one flood event) at a time is left out and the concurrent model is fit with the remaining columns. The L_2 distance is calculated for each leave-one-column-out analysis. The average of these distances is called `average.L2diff` in the function. This entire process is repeated for a specified set of choices for M, which the user provides. Once the process is repeated for each value of M, the `L2bestEst` function returns the value of the smallest average L_2 distance as well as the value of M that yields this optimal value. Once the best M has been found, the `PredictFRegressNormTest` function can be used to obtain predictions for the concurrent functional regression model.

2.5 Parametric Bootstrapping for Prediction Intervals

The following steps show how we use parametric bootstrapping in the `PredictFRegressNormTest` function to obtain 95% pointwise prediction intervals for predicted response curves. The general idea is to generate $\beta_0^*(t)$ and $\beta_1^*(t)$ 1000 times for every timepoint as well as 1000 $\epsilon^*(t)$'s for each timepoint. Then, using the equation $Y^*(t) = \beta_0^*(t) + \beta_1^*(t)X(t) + \epsilon^*(t)$, 1000 $Y^*(t)$ values are found, and the prediction interval is found by taking the 2.5% and 97.5% quantiles of the $Y^*(t)$ values, for each t.

1. Use the `fRegress` function to find $\hat{y}_i(t)$, then plug that estimate into the formula for $MSE(t) = \frac{\sum_{i=1}^{n}(y_i(t)-\hat{y}_i(t))^2}{n-2}$ where, in our case study, $n = 10$ since there are ten complete flood events.

2. Generate 1000 $\epsilon^*(t)$ from a $N(0, MSE(t))$ distribution, for each t.

3. Use the standard error outputted from the `fRegress` function to estimate the variances of $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ at each t.

4. Estimate the covariance of $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ pointwise for each t as in simple linear regression, where $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\overline{X} Var\left(\hat{\beta}_1\right)$.

5. Create a 2×2 variance-covariance matrix for every timepoint by combining the results in steps 3 and 4.

6. Using the `mvrnorm` function from the `MASS` package (Venables and Ripley, 2002), generate 1000 dependent $\beta_0^*(t)$ and $\beta_1^*(t)$ values for each timepoint, generated from a bivariate normal distribution with mean vector containing the point estimates $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ obtained from the `fRegress` output, and variance-covariance matrix created in step 5.

7. With 1000 $\beta_0^*(t)$, $\beta_1^*(t)$, and $\epsilon^*(t)$ generated, calculate 1000 estimates for the stage of Cedar Creek, $Y^*(t)$, for each t.

8. Sort the 1000 $Y^*(t)$'s at each t and take the 2.5 and 97.5 percentiles at each of these timepoints to get a pointwise 95% prediction interval.

In order for the `mvrnorm` function to work in step 6, every 2×2 variance-covariance matrix must be positive definite. In some cases (including at a small portion of the river stage data), the natural noise in the data requires the matrix to be slightly modified to become positive definite. Using the function `make.positive.definite` from the `corpcor` package (Schafer et al., 2017), we can slightly adjust the variance-covariance matrices to correct this problem. In our data, roughly 10% of the timepoints needed to be corrected, and upon further examination, there is nearly no difference between the numerically non-positive definite matrices compared to their corrected positive definite versions.

To check the assumption of normal errors implicit in our parametric bootstrap approach, we examined normal Q-Q plots of the residuals at each of the 2000 time points, and tested the residuals for normality using a Shapiro-Wilk test at each of these times.

Of the 2000 Shapiro-Wilk tests, only 85 produced a p-value less than 0.05, 4.25% of the tests, indicating the tests do not detect much departure from error normality overall. The individual Q-Q plots did not show much marked departure from normality either. Additionally, there is no clear pattern between the Shapiro-Wilk test p-values and the regions of the flood event, and the 2000 p-values are evenly distributed between 0 and 1. This information indicates that using multivariate normal parametric bootstrapping is an acceptable method for producing prediction intervals for the October 2015 flood stage reconstruction.

2.6 Applying Method to River Gage Height Data

Using the `R` functions previously described, a functional regression model can be established to relate the stage functions at the two locations, and then we can reconstruct the stage function for the flood event in which the Cedar Creek gage failed in October 2015. Recall that there are ten flood events for which both the Congaree River and Cedar Creek gage have complete data, which we will use to determine the proper number of basis functions in the regression model relating the two gage height functions.

The results of the process outlined by Ramsay et al. (Ramsay, Hooker, and Graves, 2009) show that changing the smoothing parameter λ for this problem does not have a strong impact on the resulting estimates. For our data, the smoothing parameter can take on a wide range of values (roughly 10^{-10} to 10^{10}) without affecting the results: The slope and intercept plots look exactly the same using any values in this range. With this in mind, we use $\lambda = 10^{-1}$ for the remainder of the study. We also determine the optimal number of Fourier basis functions using the aforementioned `L2bestEst` function. After comparing the average error for a wide grid of basis values of Fourier basis, the smallest error occurs with $M = 11$ Fourier basis functions. Therefore, the rest of the analysis will be done using 11 Fourier basis functions.

Now, using the optimized basis type and number, we produce estimates for $\beta_0(t)$ and $\beta_1(t)$, whose graphs are shown in Figure 2.4 and Figure 2.5. Regression function 1 represents the estimated intercept function $\hat{\beta}_0(t)$ throughout the flood event, and Regression function 2 is the slope function $\hat{\beta}_1(t)$. This is the default output from the `plotbeta` command from the `fda` package.

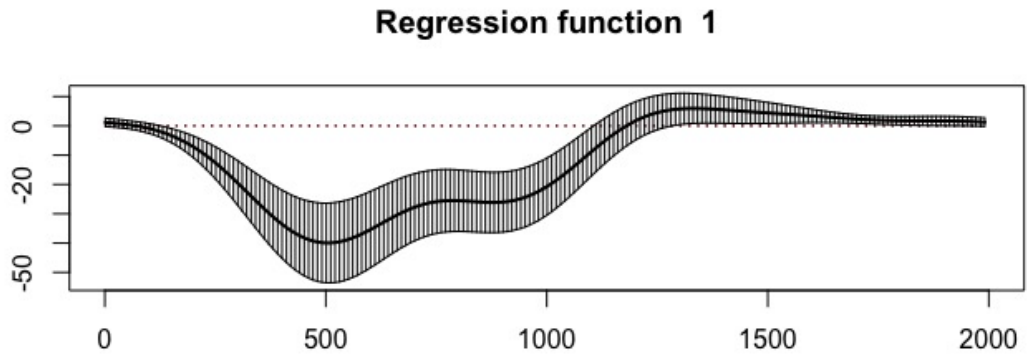


Figure 2.4. $\beta_0(t)$ (Regression Function 1 = Intercept) estimate using optimized LAL_1 distance selected data, optimized number of Fourier basis functions and pointwise 95% confidence limits.

Both the $\hat{\beta}_0(t)$ and the $\hat{\beta}_1(t)$ attain their largest magnitude at the peak portion of the flood event (around the time labeled 500). This could be because of the transition in Cedar Creek's flow from a base flow, at the lower stages, to a flow that is dominated by the flooding from the rising Congaree River. The key takeaway from these graphs is that all of the values in the $\hat{\beta}_1(t)$ (Regression Function 2) graph are positive. This indicates that no matter the time within the flood event, when the stage of the Congaree River increases, so does the predicted stage of Cedar Creek. Another observation is that near the peak of the flood event, an increase in the stage of the Congaree River causes a substantially greater increase in the predicted stage of Cedar Creek. This is consistent with the known relationship between these two locations, as the Congaree River only

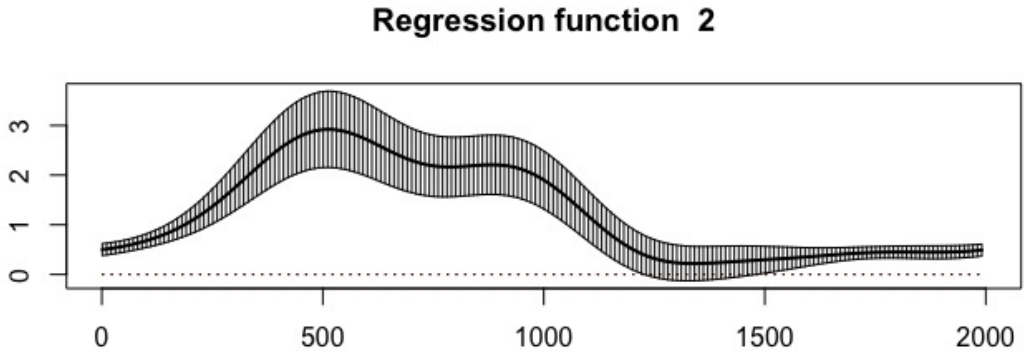


Figure 2.5. $\beta_1(t)$ (Regression Function 2 = Slope) estimate using optimized LAL_1 distance selected data, optimized number of Fourier basis functions and pointwise 95% confidence limits.

feeds into Cedar Creek once it gets high enough to flow through the floodplains in the national park (see Figure 1.2).

The key is that for each specific flood, the relationship between the Congaree River and Cedar Creek stages follows a similar pattern, and that pattern is what the concurrent functional model captures. The model establishes a relationship between the two river stages at each portion of the flood event that can then be used to reconstruct the Cedar Creek (response) gage height based on the time within the flood event and the height of the Congaree River at that point. Figure 2.3 gives an example (for the February 2020 event) of the strong association between the respective stages of the two locations, which gives credence to the appropriateness of the concurrent regression model for these data.

2.6.2 Application: Reconstructing Cedar Creek stage for October 2015 flood event

Once the relationship between the two locations during a flood event has been established, the $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ estimates as well as the known 2015 Congaree River stage can be plugged into Equation (2.9), the concurrent model, to reconstruct the Cedar Creek

stage during this flood event.

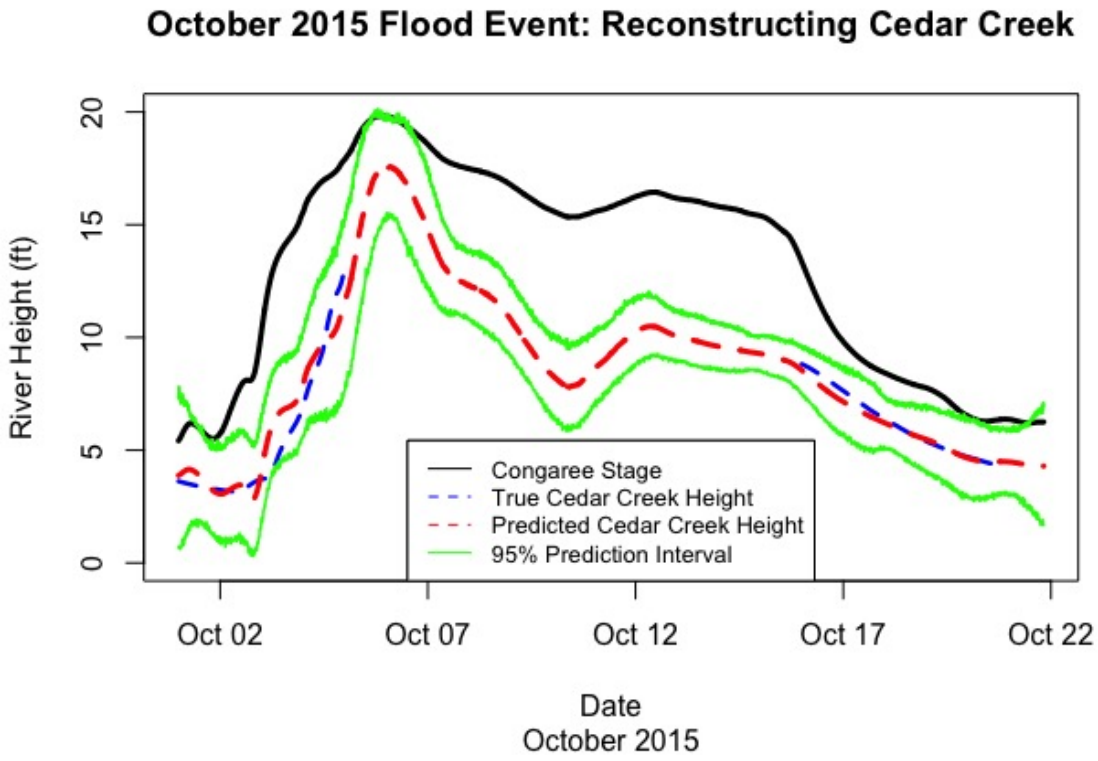


Figure 2.6. Reconstructed Cedar Creek stage (Red Dashed) for October 2015 flood event when the gage fails, accompanied by 95% pointwise confidence intervals (Solid Green) and available true gage heights for Cedar Creek during the flood event (Blue Dashed).

The graph in Figure 2.6 shows the resulting full October 2015 Cedar Creek stage prediction and estimates how high Cedar Creek rose once the gage stopped producing data. The prediction follows the available Cedar Creek data at the beginning and end of the flood event (dotted-dashed curve) quite well despite the fact that the available stages were not used in the reconstruction. The 95% prediction interval obtained from the aforementioned parametric bootstrapping is also very encouraging, as it is relatively the same width all the way through the flood event, most notably at the crest of the event. The predicted maximum Cedar Creek stage is 17.59 feet. Since the focal point of the selection of the flood event timeframes was to correctly capture the behavior at the peak, it is appropriate to investigate the validity of this predicted maximum.

The highest Cedar Creek stage on record is 16.02 feet during the February 2020 flood. The second highest recorded Congaree River stage occurred during the February 2020 flood, with the highest crest occurring during the October 2015 flood of interest, so it makes sense that the October 2015 Cedar Creek prediction would yield a maximum value higher than 16.02 feet. While 17.59 feet might seem a little bit higher than expected, note that the October 2015 flood is unique. The Congaree River experienced at least a 25-year flood in October 2015 and all its tributaries flowing through Congaree National Park recorded historically high flows. On top of that, local dams failed, exacerbating already extreme flood conditions, leading to much of the damage and destruction discussed in the introduction. As a result, a predicted maximum height of 17.59 feet is very reasonable for this historic flood event. That, along with how well the model reconstructs the known portions of the 2015 Cedar Creek stages, is further confirmation of the validity of the results and therefore the method as a whole.

2.7 Discussion

Overall, the results of our method are promising. The LAL_1 difference method used to select the start and end times of our flood events performs well and leads to a reliable reconstruction of the missing 2015 Cedar Creek stage. It is important to note that in some classical functional data sets that arise from planned experiments, such as the hip and knee angle data of Olshen et al. (Olshen et al., 1989), the start and end times of each functional observation are known, being decided by the experimenter. However, in certain observational data sets such as our river stage data, the functional observations are sections of longer time series of data, and the start and end times of the functions are not obvious. Our investigation has shown that selecting the start and end times of the functions (i.e., defining the timeframes of the flood events, in our data example) has a sizable impact on the quality of the regression results. In particular, for the functional regression problem, selecting the start and end points of the observed functions so that they

resemble (in whatever aspect is most relevant) the explanatory function corresponding to the unknown response function to be predicted is crucial.

This suggests that in other situations where the explanatory and response variables can be treated as concurrently related functional data objects, not only can the functional regression produce estimates for the $\beta_0(t)$ and $\beta_1(t)$ curves, but our method will also do well at reconstructing missing response data as long as the timeframes defining the explanatory curves have been appropriately selected. We note that implementation of functional data analysis for prediction (or reconstruction) of unknown response curves is something that has rarely been done in the statistical literature; many previous uses of functional regression have primarily focused on explaining the association between two functional data processes, rather than primarily aiming to use an observed explanatory function to predict an unobserved response function. This fact makes this study an innovative application of functional data analysis.

CHAPTER 3

FUNCTIONAL REGRESSION INFLUENTIAL MEASURES ON MODEL FIT

3.1 INTRODUCTION

In ordinary linear regression, it is common practice to assess the influence of the individual observations (Belsley, Kuh, and Welsch, 1980). It is important to understand how trustworthy the predicted outcomes are and how influential each observation is on various results. This chapter will extend ordinary linear regression influence diagnostics to the fully functional linear regression model framework. We will present additional diagnostic tools that can be used jointly to identify functional observations with large influence on the model and the resulting predictions. Some properties of the method will be investigated using a simulation study. Then these methods will be applied to a river stage reconstruction and a coastal air and water temperature dataset.

It is important to note that functional linear regression model diagnostics have not been explored at the level of their non-functional ordinary linear regression counterparts. This is partially because FDA is still a rapidly growing field of statistics, but also because regression with functional data adds an extra dimension that can make quantifying influence more challenging. Shen and Xu (2007) develop model selection and diagnostic tools for functional regression with a functional response variable and many non-functional predictors. They use the L_2 norm of the residuals to compute studentized residuals for each observation and the typical leverage value from the diagonal h_{ii} of the Hat matrix to calculate a functional Cook's Distance D_i, inspired by the or-

dinary regression measure of influence of Cook (1977). Chiou and Müller (2007) also study the case where the response variable is functional but where the predictor variables are either multivariate vectors or random functions. They use functional principal component analysis (FPCA) to determine whether the residuals are dependent on the covariate (in a single covariate setting). In this case, any functional linear model will not need an intercept term since all components will pass through the origin. Chiou and Müller treat the $n \times n$ Hat matrix as a function of t, which is how we will treat most of our diagnostic measures. Chen, Huang, and Lin (2014) build on Chiou and Müller's work and is similar to our study, in that both the response and predictors are functional observations. They also take discretized observations and describe them using basis functions, transforming them into functions of time. They calculate a version of functional Cook's distance and a likelihood distance, finishing with a small simulation study in which they intentionally insert outlying measurement points within their single functional object and then confirm that their method identifies such points as influential. We are concerned with identifying an entire influential functional observation from our set of historic flood events. Chen et al. (2014) primarily focus on deleting a specific point within the domain of the single set of curves and refitting the model. Febrero-Bande, Galeano, and González-Monteiga (2010) build on Chiou and Müller's work, focusing on finding influential observations when there are functional predictors and a scalar response. While this framework is also dissimilar to our study which has functional responses and predictors, they propose a bootstrap with a smoothing method to approximate an underlying null distribution of each of their metrics to establish estimated quantiles of their metrics to determine each observation's influence. Throughout the rest of this chapter, we will build on ideas from several of the aforementioned studies to establish a method for determining which functional observations are the most influential on the concurrent functional regression fitted model when there is one set of functional predictors and one set of corresponding functional response variables.

We present multiple new functional influence measures and describe a novel weighted bootstrapping with perturbations approach for determining the significance of those measures. Then we provide a simulation study to test the performance of the method, concluding with two independent applications.

3.2 Influence Measures in the Functional Framework

Simple linear regression relates one predictor vector $\mathbf{X}$ and one corresponding response vector $\mathbf{Y}$ via the fitted equation, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$. Our interest is in how influential each single observation is in estimating that relationship. The most common leverage diagnostic (needed for the influence measures) uses the $n \times n$ Hat matrix

$$\mathbf{H}_{n\times n} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{3.1}$$

where n is the number of observations. Each element in the diagonal h_{ii} is the leverage of the ith observation. Typically, any leverage over $2p/n$ is considered high, given that the mean leverage value is p/n, though some guidelines indicate that anything over 0.5 is high leverage and anything between 0.2 and 0.5 is moderate leverage (Kutner et al., 2005). A high leverage point does not necessarily need to be removed from the study, but it should be investigated further to determine whether it unduly affects the linear model.

Another common way to determine an observation's effect on the model is to measure its influence on the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. This can be accomplished by removing one observation i at a time and computing $DFBETAS_p$, defined for the p-th regression coefficient as

$$DFBETAS_{p,i} = \frac{\hat{\beta}_p - \hat{\beta}_{p(i)}}{se(\hat{\beta}_{p(i)})} \tag{3.2}$$

where $\hat{\beta}_p$ is the estimate of β_p using all of the observations, $\hat{\beta}_{p(i)}$ is the estimate of β_p based on a data set with the ith observation removed, and $se(\hat{\beta}_{p(i)})$ is the standard error of the estimate when the ith event is excluded. Typically, any $DFBETAS_p$ more extreme

than $\pm 2/\sqrt{n}$ is considered influential in estimating β_p and should be investigated further.

Another metric used to measure influence is $DFFITS$ which measures how the ith predicted value changes when that observation is removed. The formula is

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)} h_i}} \tag{3.3}$$

where $\hat{y}_i$ is the predicted value for the ith observation with the full data used to make the prediction, and $\hat{y}_{i(i)}$ is the prediction for that same ith observation using β estimates computed using the data without that observation included. The value of h_i is the ith diagonal element of the Hat matrix, and $MSE_{(i)}$ is the mean squared error when the ith observation is removed. Generally, any observation with $|DFFITS_i| > 2/\sqrt{p/n}$ is considered influential (Kutner et al., 2005).

Cook's distance is the last common measure of influence we will discuss for the ith observation and is defined as:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{(k+1)MSE}. \tag{3.4}$$

Here $\hat{y}_j$ is the predicted value of y_j from the fit with all observations included and $\hat{y}_{j(i)}$ is the predicted value of y_j from the fit when the ith event is excluded. In the denominator, k is the number of covariates included in the model and MSE is the ordinary mean squared error with all of the observations included. There is not a standard criterion for what constitutes a significantly large value, but it is recommended that any observation with a Cook's Distance much higher than the rest be investigated. Some guidelines say to investigate values over $4/n$, some say 0.5 and others recommend 1. Some literature even uses the F distribution to establish a cutoff. Any value above the 50th percentile of an F distribution with $(k+1, n-k-1)$ degrees of freedom is considered influential (Kutner et al., 2005).

Using the ordinary regression formulas as a starting point, we applied these metrics to the concurrent functional regression model that relates a set of functional predictors

$X_i(t), i = 1,\ldots,N$ to a corresponding set of functional responses $Y_i(t)$ at each time t in the domain.

$$Y_i(t) = \beta_0(t) + \beta_1(t)X_i(t) + \epsilon_i(t),\ \ i = 1,\ldots,N \tag{3.5}$$

In the functional data framework, each influence measure is calculated at each location t, creating measures that are functions of t. The resulting formulas are:

$$h_i(t) = i\text{th diagonal of the } N \times N \text{ matrix } \mathbf{H}_t \text{ where } \mathbf{H}_t = \mathbf{X}_t\left(\mathbf{X}_t^T\mathbf{X}_t\right)^{-1}\mathbf{X}_t^T \tag{3.6}$$

$$DFBETAS_{p,i}(t) = \frac{\hat{\beta}_p(t) - \hat{\beta}_{p(i)}(t)}{se(\hat{\beta}_{p(i)}(t))},\ \ p = 1,2,\ \ i = 1,\ldots,N \tag{3.7}$$

$$DFFITS_i(t) = \frac{\hat{y}_i(t) - \hat{y}_{i(i)}(t)}{\sqrt{MSE_{(i)}(t)h_i(t)}},\ \ i = 1,\ldots,N \tag{3.8}$$

$$D_i(t) = \frac{\sum_{j=1}^{n}\left(\hat{y}_j(t) - \hat{y}_{j(i)}(t)\right)^2}{(k+1)MSE(t)},\ \ i = 1,\ldots,N \tag{3.9}$$

In Equation (3.6), $\mathbf{X}_t$ is a $2 \times N$ design matrix defined at time t. $\hat{\beta}_p(t)$ is the coefficient estimate (for $p = 0,1$) using all N observations in the calculation, and $\hat{\beta}_{p(i)}(t)$ is the coefficient estimate (for $p = 0,1$) when observation i is left out. Similarly, $\hat{y}_i(t)$ is the predicted response curve for observation i with all N observations and $\hat{y}_{i(i)}(t)$ is that same prediction when observation i is excluded. $\hat{y}_{j(i)}(t)$ is the predicted response value for observation j when observation i is withheld. Note that the values of $\hat{\beta}_p(t)$, $se(\hat{\beta}_p(t))$, $MSE_{(i)}(t)$, $\hat{y}_i(t)$, etc., are calculated with the concurrent functional regression model, creating measures that are time dependent. Taking the mean (across the n timepoints) of the absolute values of the metric for each observation gives a single convenient measure of influence for that observation. Therefore, we define the following:

$$\begin{aligned}
\overline{|DFBETAS_p|}_i &= \frac{1}{n}\sum_{j\in\{1,\ldots,n\}} |DFBETAS_{p,i}(t_j)| \text{ for } i = 1,\ldots,N \\
\overline{|DFFITS|}_i &= \frac{1}{n}\sum_{j\in\{1,\ldots,n\}} |DFFITS_i(t_j)| \text{ for } i = 1,\ldots,N \\
\overline{D}_i &= \frac{1}{n}\sum_{j\in\{1,\ldots,n\}} D_i(t_j) \text{ for } i = 1,\ldots,N
\end{aligned} \tag{3.10}$$

Even in the non-functional regression scenario, an easily defined threshold to determine if the measure is "large" is not readily agreed upon and is often ad hoc. In functional regression, those informal cutoffs may be even less appropriate. Therefore, we will use a weighted bootstrapping approach (with perturbations) on each functional metric to determine how large a metric's value must be to label an observation as influential. We provide a simulation study in which we evaluate the performance of each functional influence measure. We apply these metrics in the context of river stage data during floods. Lastly, we will briefly apply these measures to another dataset that investigates the relationship between air and water temperatures at weather stations along the US coastline.

3.3 Bootstrapping to Approximate a Null Distribution of Influential Measures

In the functional regression framework, we now propose a formal test to determine whether the larger values of these regression diagnostic metrics are statistically significantly large. In order to discern this, we repeatedly resample the functional data, calculate the influence measure of interest for each resampled data set, and compare these calculated values to the metrics from the observed data. We refer to our approach as "weighted bootstrapping with perturbations." We use weighted bootstrapping to create a distribution of metric values that serves as a null distribution, i.e., a distribution for the metric under the condition that there is no especially influential curve. To accomplish this, when selecting our bootstrap sample we propose to sample the apparently less influential observations from our observed curves more often than the apparently most influential observations. We define any particular measure of influence generically as r_i, calculate it for each observation, and then use the following equation to translate the

metric value for observation i into a selection probability θ_i:

$$\theta_i = \frac{(1/r_i)^\alpha}{\sum_i [(1/r_i)^\alpha]}, \quad \alpha \geq 0. \tag{3.11}$$

Note that $\alpha = 0$ corresponds to equal selection probabilities for each observation. In general α should not exceed 0.5 and is most crucial when N is small. When N is small and one observation from the sample has an extreme (high or low) average measure of influence compared to the rest, it is possible that in a certain bootstrap iteration that observation will be selected to compose most of that iteration's sample, unless it has a small selection probability. This results in misleading and sometimes incalculable influence measures for that sample. To correct for this, we provide following weighted bootstrapping with perturbation method. While this method can be implemented for any sample size, it is most applicable in the small sample setting.

1. Define r_i for each observation as the influence measure of interest.

2. Select an appropriate value of α (or allow a range of choices) and calculate θ_i for $i = 1, \ldots, N$.

3. Sample N observations with replacement from the original set of data, where the ith observation has probability θ_i of being selected.

4. Apply independent realizations of a perturbation process to each sampled response curve. For our perturbations, we use the Ornstein-Uhlenbeck process, approximated discretely using the Euler-Maruyama method (more details below). Each bootstrap sample then consists of N functional pairs
   $\{(X_1^*(t), Y_1^*(t)), \ldots, (X_N^*(t), Y_N^*(t))\}$.

5. Using these new pairs of functional data, fit the concurrent functional regression model and calculate the same measure of influence, for each observation $i = 1, \ldots, N$.

6. Repeat Steps 3-5 for the desired number of bootstrap iterations (B) to obtain NB values of the metric, which approximate a null distribution for that influence measure.

7. The original metric from the observed dataset can be compared to percentiles from the respective bootstrap distribution to determine whether the largest values identified in the original data analysis are significantly large relative to the null distribution.

Having identical observations selected repeatedly in a given bootstrap sample could skew the calculated metrics because any of the selected observations sampled only once might be deemed "influential" simply because it differed from the other observations. To avoid this, we added small perturbations to the sampled response curves to ensure that no two sampled observations are identical, without obscuring the underlying relationship between the predictor and response curves. Our perturbation approach is the Ornstein-Uhlenbeck process approximated via the smoothed Euler-Maruyama method. The Ornstein-Uhlenbeck process was defined by Uhlenbeck and Ornstein (1930), where the process x_t is defined by the stochastic differential equation $dx_t = \theta(\mu - x_t)dt + \sigma dW_t$, where $\theta > 0$ is the drift parameter that pulls the process back to its mean μ and $\sigma > 0$ is the standard deviation of the error added to the process. The value W_t represents the Wiener stochastic process. The Euler-Maruyama approximation yields discrete values of this process using the following discrete stochastic process:

$$\kappa_{n+1} = \kappa_n + \theta(\mu - \kappa_n)\Delta t + \sigma \Delta W_n \tag{3.12}$$

where κ_0 is initialized by selecting a single value from a $N(0, \sigma^2)$ distribution. The Wiener process (Brownian motion) values ΔW_n are independent identically distributed increments distributed normally with mean 0 and variance Δt (Resnick, 1992). i.e. $W_{t_{n+1}} - W_{t_n} = \Delta W_n \sim N(0, \Delta t) = \sqrt{\Delta t} N(0,1)$. To obtain a mean-zero perturbation process, we

set $\mu = 0$. Therefore, the perturbations added to the response curves are of the form:

$$\kappa_{n+1} = \kappa_n - \theta(\kappa_n)\delta t + \sigma Z\sqrt{\delta t} \tag{3.13}$$

where Z is a random value from the standard normal distribution. While there is no general rule of thumb for choosing θ and σ, we recommend that the drift parameter θ should range from 0.5 to 1 and σ should be chosen based on the values that are being perturbed. Since θ is responsible for pulling the process back towards the mean, if it is too small then the perturbed curve becomes too different from the original curve. The value of σ should be selected based on the range of the functional observations being perturbed using the following method:

1. Calculate γ = mean of {range$[y_1(t)]$, ..., range$[y_N(t)]$},
   where range$[y_j(t)]$ = $\max_t y_j(t)$ - $\min_t y_j(t)$.
2. Set $\gamma_l = \gamma/3$ and $\gamma_u = \gamma/2$.

The value of σ can reasonably be between γ_l and γ_u. Any combination of θ and σ following this criteria appropriately adds enough variation to the underlying curves without extensively altering them. For each bootstrap iteration we randomly select θ from $Uniform(0.5, 1)$ and σ from $Uniform(\gamma_l, \gamma_u)$.

The ideal value of α in Equation (3.11) will vary based on the observed measures from the initial dataset. In general, we recommend using $\alpha = 0.5$ when N is small or when on of the observed measures is noticeably larger or smaller than the rest. If values of the metric have little variability, the bootstrapped percentiles will be similar regardless of $\alpha \in (0, 0.5)$; however, when the observed influence measures are more spread out or one observation's influence measure is much larger than the rest, using $\alpha = 0.5$ dampens the effect that the observation has on the approximated percentiles, resulting in percentiles that better resemble a null distribution. This allows truly significant influential observations to be flagged rather than be dominated by the values for the most

influential observations. For large sample sizes, an observation with a large influence measure has less impact on the approximate null distribution as it is less likely to be sampled in a given iteration regardless of the value of α compared to when sample size is small; therefore, using $\alpha = 0$ in large sample scenarios is appropriate. Table 3.2 provides an example of the selection probabilities when letting α vary. If the sample size is moderate, or it is unclear if the largest influence measure is too much larger than the next highest, we recommend performing the weighted bootstrap analysis on the data using both $\alpha = 0$ and $\alpha = 0.5$ independently and comparing the resulting percentiles to see the effect of the more influential observations.

After performing this bootstrapping method, we recommend marking the 90th, 95th, and 99th percentiles. The percentiles can then be used in to identify the significantly influential functional observations from the initial dataset by comparing the observed measures to the resulting percentiles. We define a value above the 90th percentile as as moderately influential, above the 95th percentile significantly influential, and above the 99th percentile also as significantly influential and need to investigate further.

3.4 Simulation Study

We investigate the performance of our method at identifying influential observations in a simulation study. For this example, we generate as simulated predictor functions N independent $X(t)$ curves where $t \in \{1, 2, \ldots, 1000\}$ using the following formula:

$$X(t) = (t/12)[a_s \sin[(1/k_s)(t - d_s)] + c_s][a_c \cos[(1/k_c)(t - d_c)] + c_c]$$

where each of the N curves are generated by randomly selecting the parameters within the equation.

- a_s, a_c, c_s and c_c are independently sampled from the list $\{-3, -2, -1, 0, 1, 2, 3\}$.
- k_s and k_c are sampled from the list $\{-300, -200, -100, 100, 200, 300\}$.

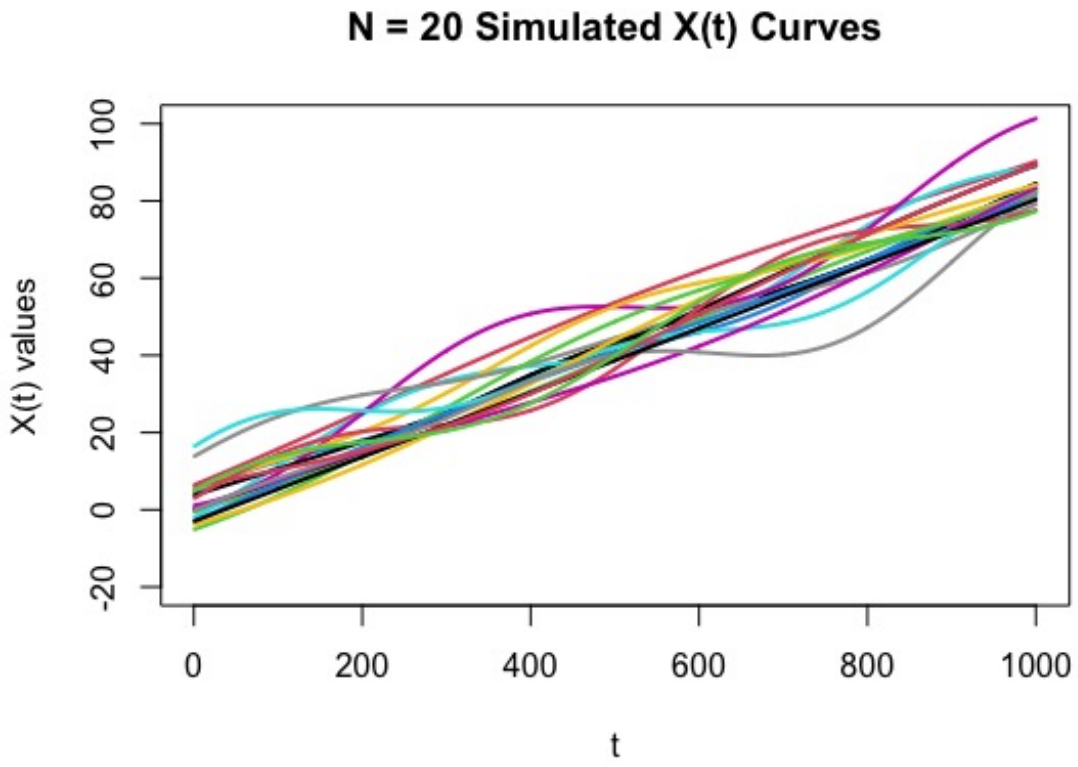


Figure 3.1. Example of $N = 20$ generated $X(t)$ curves using the described functional data generation method.

- d_s and d_c are sampled from the list $\{-100, -50, 0, 50, 100\}$.

By alternating the combination of parameters used to generate the functional data, we produce curves that are similar and follow a the same underlying curve $m(t) = t/12$. An example of $N = 20$ $X(t)$ curves are shown in Figure 3.1. Note that the simulation results are not changed if the parameter's ranges are expanded as long as they are the same for all N curves. Next we set the functional slope and intercept functions to be:

$$\beta_0(t) = \cos(t/200) + 2$$

$$\beta_1(t) = \sin(t/200) + 2$$

We dampen the relationship between the predictor and response curves by generating noise functions $\epsilon_i(t)$ to slightly distort the functional relationship between each pair of

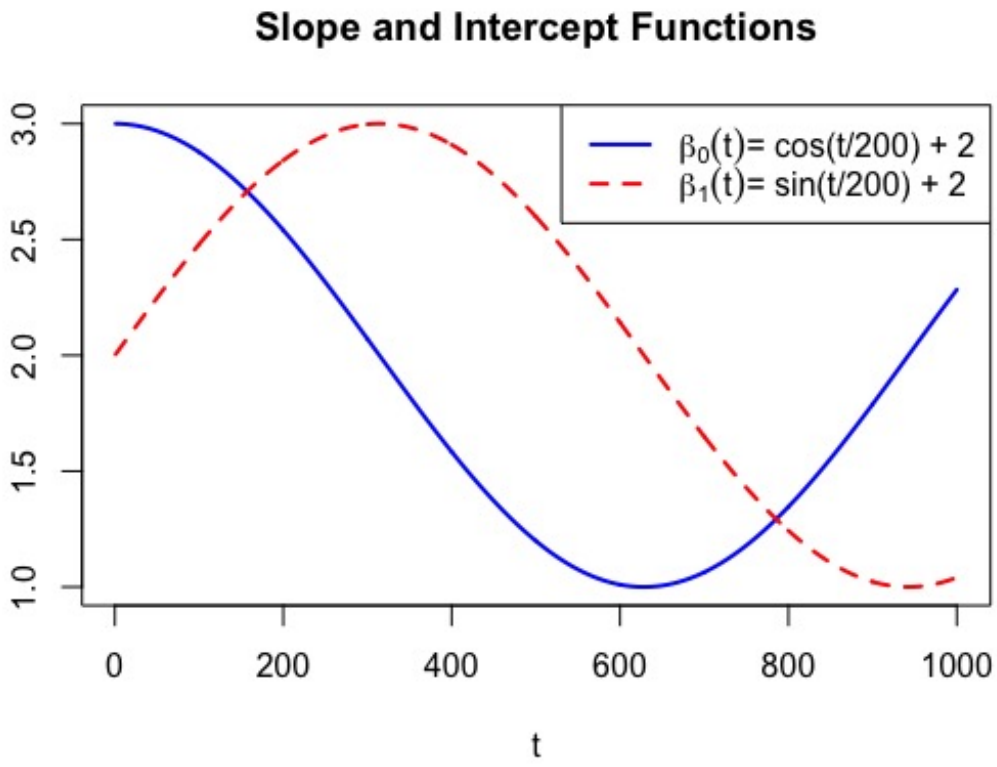


Figure 3.2. Defined functional intercept $\beta_0(t)$ (solid blue) and functional slope $\beta_1(t)$ (dashed red) used to generate response curves $Y_i(t)$ using $X_i(t)$.

$X(t)$ and $Y(t)$ curves. We do this by adding realizations of the Ornstein-Uhlenbeck process, approximated using the Euler-Maruyama method, to the mean response curves calculated using the generated predictor curves and the slope and intercept functions using Equation (3.5). An example of resulting set of simulated response curves is shown in Figure 3.3.

As a preliminary check that these generated data followed our functional linear model, before introducing any contamination, we fit the model for each of 100 generated data sets and verified the estimates of $\beta_0(t)$ and $\beta_1(t)$ resembled the true functional slope and intercept on average. However, all further analysis was done on simulated data with contamination, as we described next.

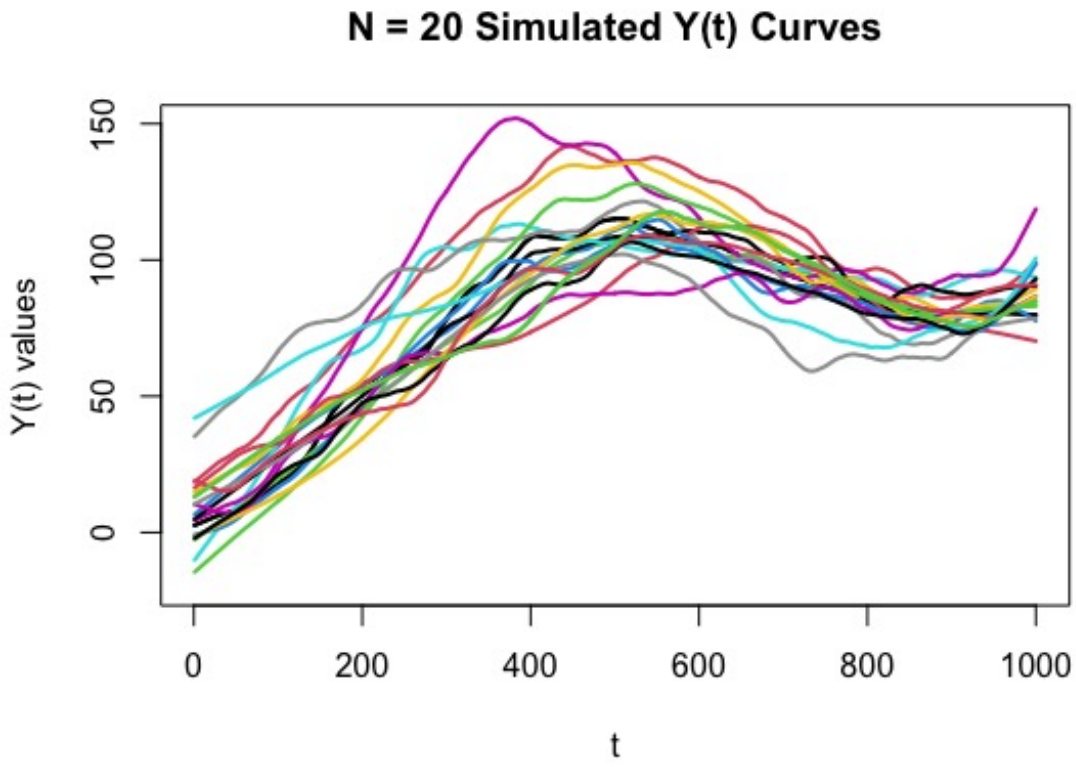


Figure 3.3. Example of N = 20 response ($Y(t)$) curves used in simulation with no contaminated observations ($\lambda = 1$).

We intentionally contaminated the $\beta_1(t)$ function for one of the N observations and see how often our method identifies the contaminated observation as influential. For this contaminated observation, we let $\beta_1(t) = \lambda \times \sin(t/200) + 2$ for some $\lambda > 0$. Clearly, $\lambda = 1$ represents the control case in which the contaminated observation is generated the same way as the others. In this simulation, we set λ at the levels $\{0.5, 0.75, 0.9, 1, 1.1, 1.25, 1.5, 1.75, 2.0\}$ and examine the performance of our approach to detect influential curves in the functional regression model. Figure 3.4 gives an example of $N = 20$ response curves with the contaminated curve generated using $\lambda = 2$.

We also investigate the effect of varying α when $N = 100$, $N = 50$, $N = 20$, and $N = 10$ using the following method:

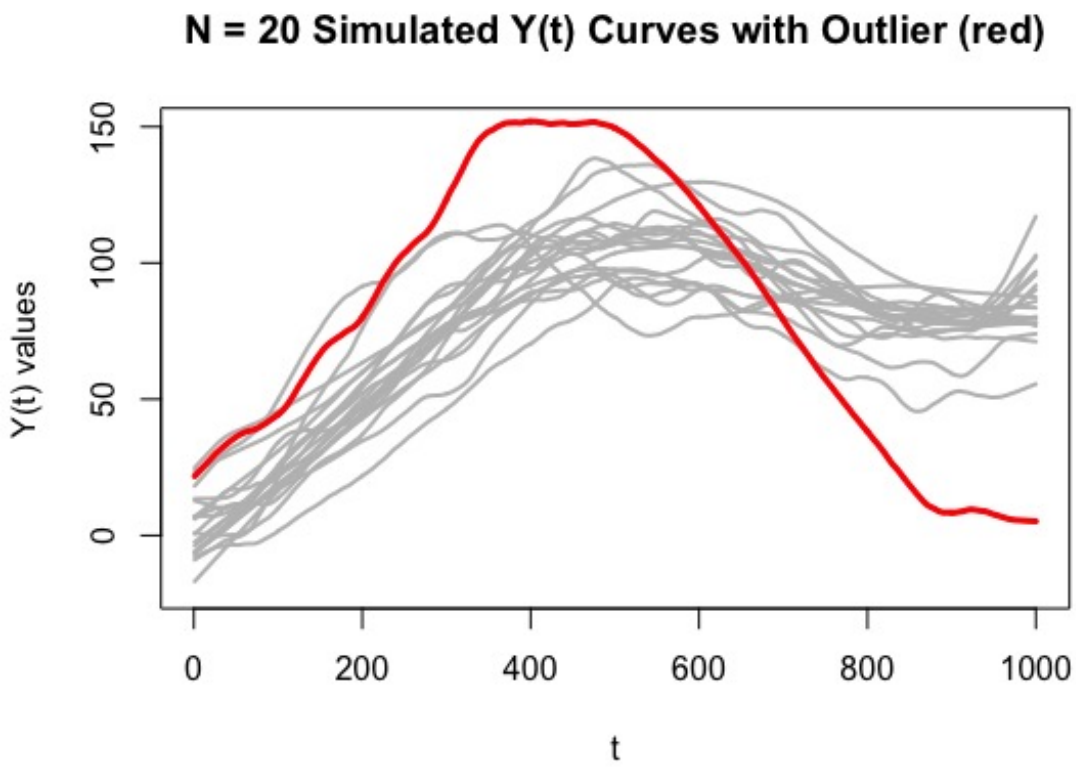


Figure 3.4. Example of N = 20 response ($Y(t)$) curves used in simulation with one contaminated observation (red) using $\lambda = 2$.

1. Select λ.

2. Generate N sets of $\{X_i(t), Y_i(t)\}$ curves with one $Y_i(t)$ curve contaminated using λ.

3. Calculate the functional influence measure ($\overline{|DFBETAS_0|}_i$, $\overline{|DFBETAS_1|}_i$, $\overline{|DFFITS|}_i$, or $\overline{D}_i$) for $i = 1, \ldots, N$.

4. For each influence measure separately, select α and calculate the selection probabilities θ_i for each observation using Equation (3.11).

5. Perform $B = 100$ bootstrap iterations, sampling the N observations with replacement, calculating the influence measure for each observation in each iteration (yielding NB values of the measure).

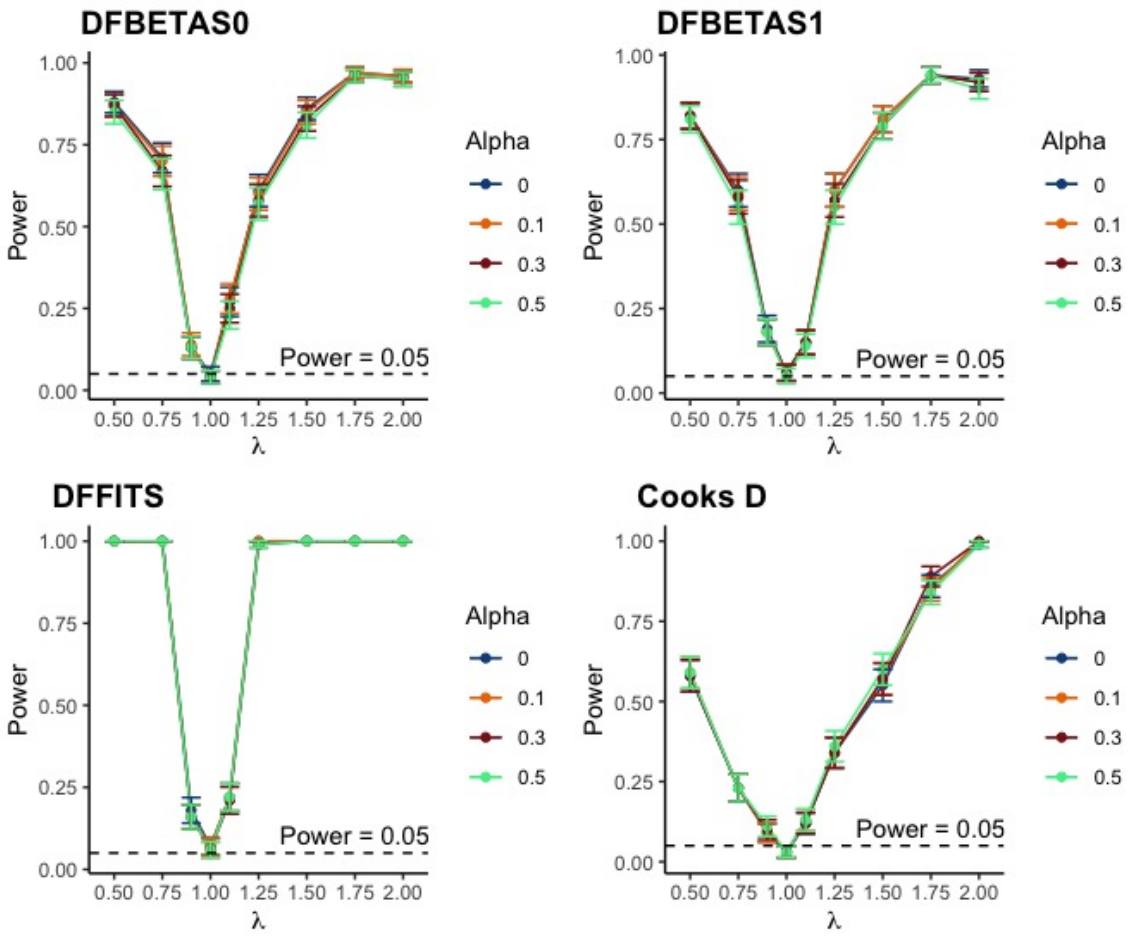


Figure 3.5. Power functions displaying the average proportion of contaminated observations above the 95th percentile for the four influence measures and different values of α (with error bars representing one standard error) for $N = 100$.

6. Determine the percentile relative to this bootstrap distribution of the originally contaminated observation's influence measure, tracking indicate whether it is above the 95th percentile.

7. Repeat this process 100 times for each combination of desired influence measure, λ, and α.

Note that for each data generation, the bootstrapping process was executed using each choice of α on the same generated data.

Figure 3.5 shows the average proportion of contaminated observations that are above the 95th percentile for each influence measure when $N = 100$. This is analogous to the

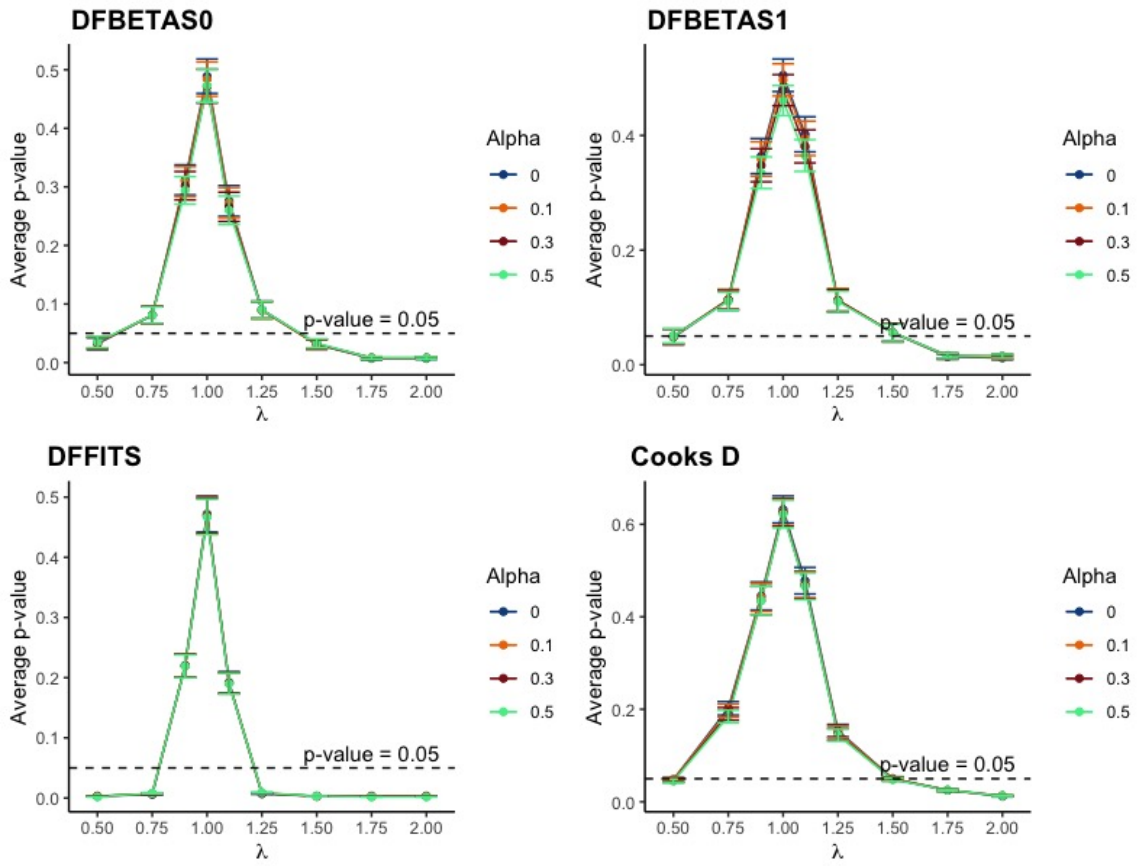


Figure 3.6. Average p-value (1− percentile within bootstrap distribution) of contaminated observations for the four influence measures and different values of α (with error bars representing one standard error) for $N = 100$.

power of the procedure at an implied significance level of 0.05. As λ moves away from 1, the proportion of contaminated observations flagged increases for each measure. This correctly indicates that when an observation is more extreme, it is flagged as influential more often. When using $|\overline{DFFITS}|$, the contamination need not be especially extreme for this value to be consistently above the 95th percentile, whereas when using the functional Cook's distance, the contamination must be more extreme for $\overline{D}$ to be flagged on average.

Figure 3.6 provides additional results from the same simulation. Here we plot the average p-value, which is 1 minus the average percentile within the bootstrap distribution of the contaminated observation. As λ moves away from 1, the p-value decreases, indi-

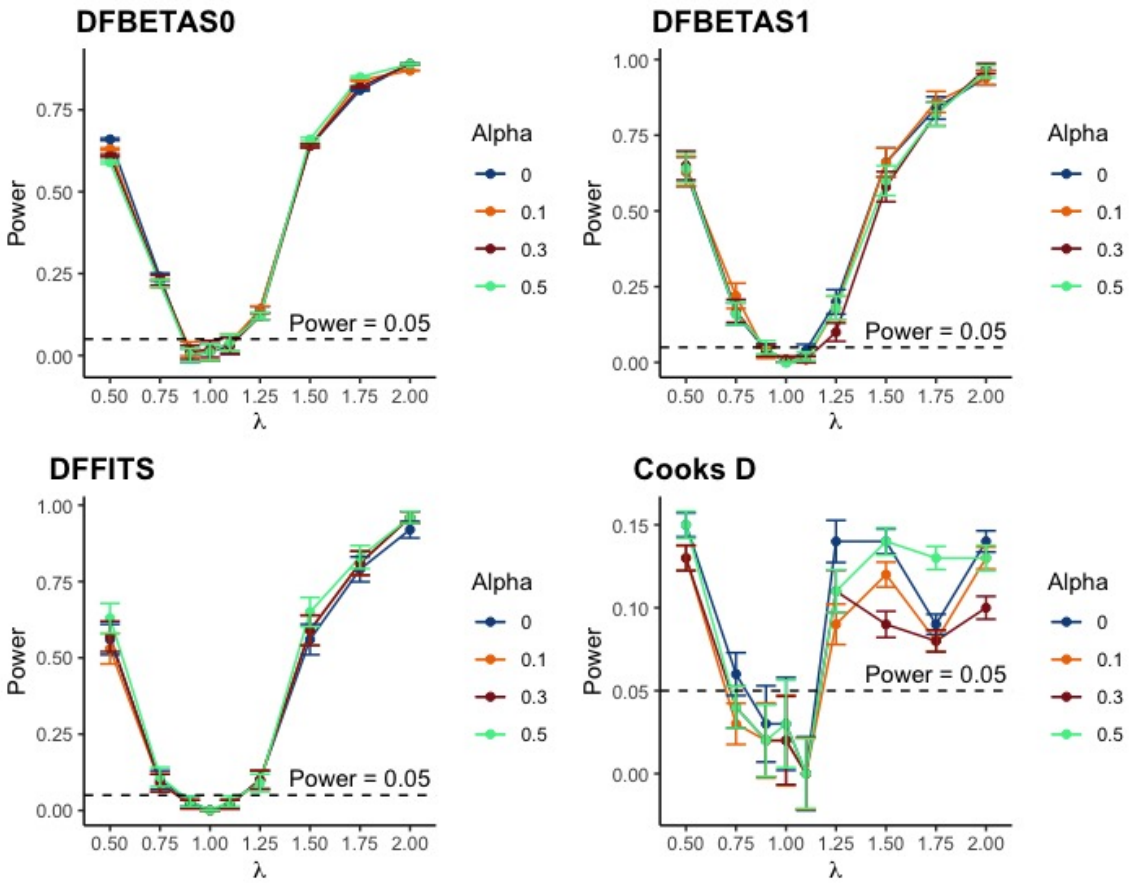


Figure 3.7. Power functions displaying the average proportion of contaminated observations above the 95th percentile for the four influence measures and different values of α (with error bars representing one standard error) for $N = 10$.

cating that the contaminated observation's influence measure is frequently significant. Figure 3.5 and Figure 3.6 also show that with a large sample size of $N = 100$, the effect of α is negligible.

Figure 3.7 and Figure 3.8 show plots of the power and average p-value when $N = 10$. As λ moves further from 1 the bootstrap method detects the contaminated observation more often. Note that with a small sample size, setting $\alpha = 0.5$ slightly increases the power and reduces the average p-value (especially with $\overline{|DFFITS|}$) by better dampening the effect the contaminated observation has on the bootstrap null distribution. When $N = 10$, using the functional Cook's distance, the bootstrap method almost never marks the contaminated observation as influential, within the range of λ we used. Given these

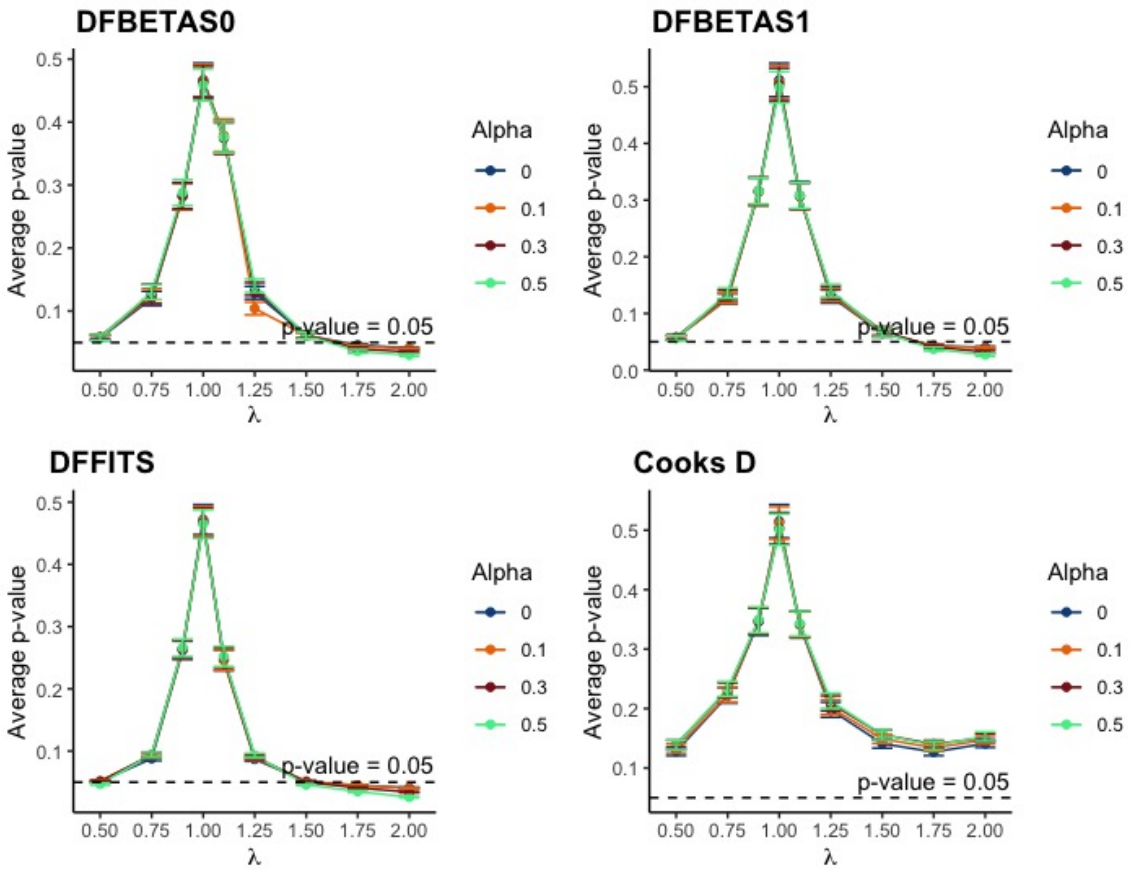


Figure 3.8. Average p-value (1− percentile within bootstrap distribution) of contaminated observations for the four influence measures and different values of α (with error bars representing one standard error) for $N = 10$.

results, in a real data application of this method, if an observation is influential based on the bootstrap approach with Cook's distance, then it is likely that the observation is strongly influential on the functional model. Similar plots when $N = 50$ and $N = 20$ are provided in the appendix and show analogous patterns to the sample sizes discussed above.

Overall, we recommend approximating a null distribution for each of the four measures to evaluate the overall influence of each observation. When the sample size N is large, using $\alpha = 0$ is recommended given the minor differences in p-value and power. When the sample size is small, we recommend performing the bootstrapping method with $\alpha = 0.5$. If no measure is substantially larger than the rest, then the sets of per-

centiles will be similar regardless of the choice of α; however, if one observation is extremely influential, then it will generally inflate the higher percentiles when $\alpha = 0$.

3.5 Real Data Application Using River Stages

3.5.1 Applying the functional influence detection to river stage data

Pittman, Hitchcock, and Grego (2021) analyzed river stages from two related gage locations at Congaree National Park near Columbia, South Carolina. A novel landmark alignment technique was used to determine objectively the optimal start and end points of ten flood events in which the Congaree River United States Geological Survey (2020a) flowed over bank, through the floodplains, and into Cedar Creek United States Geological Survey (2020b). This resulted in 10 historic flood events that could be directly used in the concurrent functional model. The purpose of using functional regression was to relate the Congaree River stage to the Cedar Creek stage during flood events. Then this relationship could be used to reconstruct the Cedar Creek stage during a major flood event in October 2015 when the Cedar Creek gage went offline but the Congaree River gage remained functional.

The first measure of influence we calculated was $DFBETAS_{p,i}(t)$ where $p = 0$ represents the intercept function and $p = 1$ the slope function. To calculate $DFBETAS_{p,i}(t)$, an entire flood event was removed and the coefficient functions re-estimated. One of the events with the most influence on the estimation of $\beta_0(t)$ and $\beta_1(t)$ was the February 2020 flood event. Figure 3.9 shows the difference between $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ using all ten events (black curve) and with the February 2020 event removed (red curve). The distance between these curves at each point is the numerator of the $DFBETAS_{p,i}(t)$ formula. Analogous plots for the remaining nine events are shown in the appendix.

$DFBETAS_{p,i}(t)$ for the ten events is given in Figure 3.10. We see no obvious outlying event, and only a couple of the curves visually deviating far from the others. To determine which event had the most impact on the estimates $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$, values of

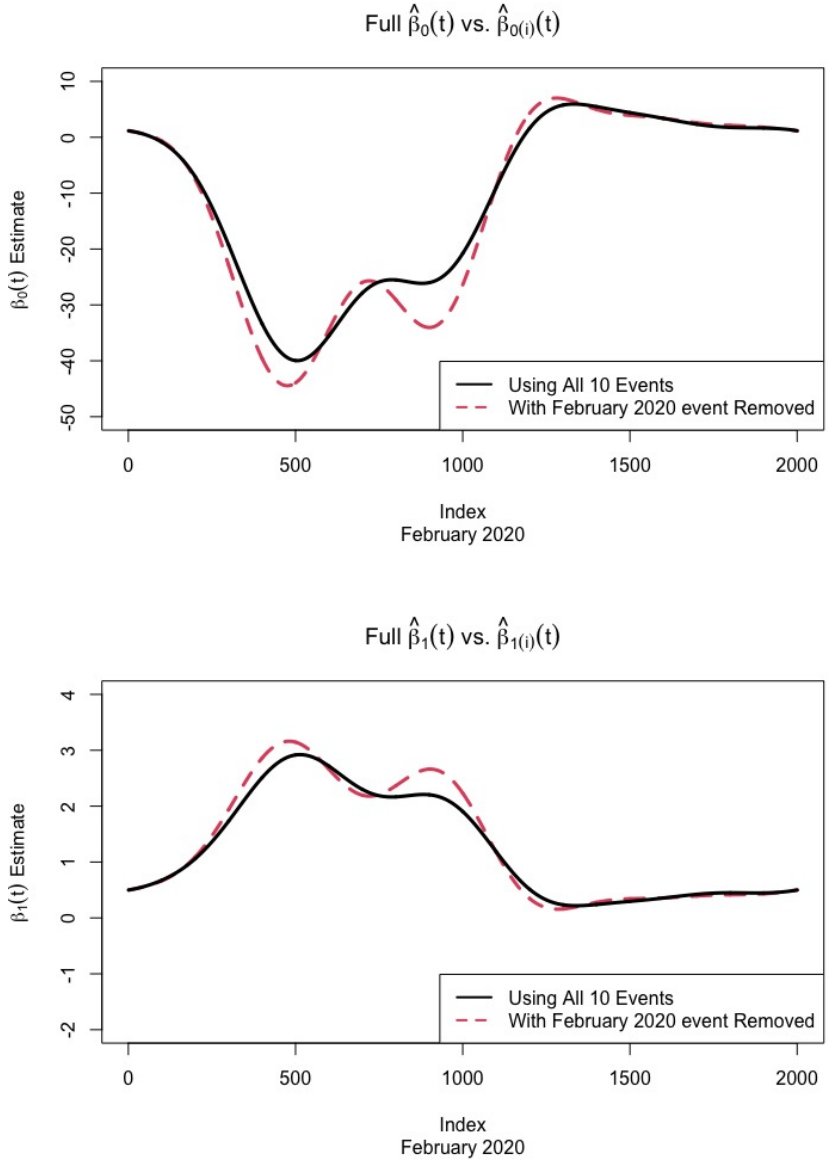


Figure 3.9. Comparison of $\hat{\beta}_0(t)$ and $\hat{\beta}_{0(i)}(t)$ (top) and $\hat{\beta}_1(t)$ and $\hat{\beta}_{1(i)}(t)$ (bottom) where the black curve represents the $\beta_p(t)$ estimate with all 10 historic flood events included and the red curve is the estimate when the February 2020 event is removed.

$\overline{|DFBETAS_p|}_i$ of each event $i = 1,\ldots,10$, are provided in Table 3.1. Most of the $DFBETAS_{p,i}(t)$ values remained within the standard threshold values in the nonfunctional scenario, indicating that the cutoffs used in ordinary linear regression may not be too different than those appropriate for the functional framework.

The February 2020 and August 1995 flood events had the highest $\overline{|DFBETAS_p|}$, in-

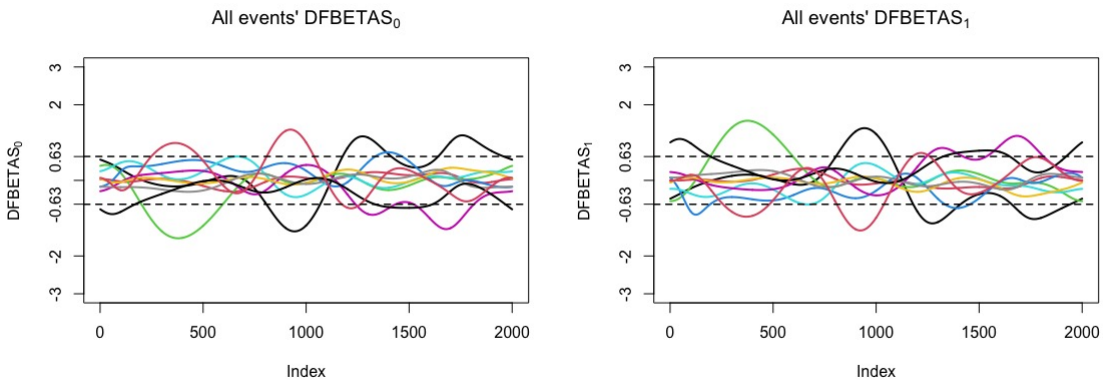


Figure 3.10. $DFBETAS_{(p)}(t)$ for all ten historic flood events (solid lines) with a reference of what may be considered large (dashed line) in non-functional linear regression $\pm 0.63 = \pm 2/\sqrt{N}$ for $N = 10$.

Table 3.1. Mean of each influence measure across t for each of the events $i = 1,\ldots,10$ with the highest values in bold.

Event	$\overline{\lvert DFBETAS_0 \rvert}$	$\overline{\lvert DFBETAS_1 \rvert}$	$\overline{\lvert DFFITS \rvert}$	$\overline{D}$
August 1995	**0.552**	**0.527**	0.927	0.1223
February 1998	0.104	0.113	0.827	0.085
March 2003	0.361	0.387	0.898	0.103
May 2003	0.303	0.339	1.797	0.258
Sept. 2004	0.232	0.246	1.801	0.235
March 2007	0.421	0.396	1.712	0.436
February 2010	0.109	0.122	0.853	0.068
May 2013	0.151	0.132	1.079	0.137
November 2018	0.356	0.410	0.748	0.069
February 2020	0.444	0.445	**4.062**	**2.312**

dicating that these events have the most influence on the $\beta_0(t)$ and $\beta_1(t)$ estimates in the concurrent model. The informal cutoff used in ordinary linear regression is $2/\sqrt{N} = 2/\sqrt{10} = 0.632$. While this value should not be unthinkingly applied in the functional framework, it gives us a decent starting point.

For $i = 1,\ldots,10$, $DFFITS_i(t)$ measured the effect of event i on the predicted value of the response for event i at each t. The fitted curves $\hat{Y}_i(t)$ and $\hat{Y}_{i(i)}(t)$, based on the regression's fit with and without event i are given in the appendix, along with each calculated $DFFITS_i(t)$. The most notable difference in fitted curves is in the tenth event

(February 2020). While none of the $DFFITS_i(t)$ curves are particularly flat, Figure 3.11 shows that $DFFITS_{10}(t)$ is the most sporadic and has the largest measurements. Table 3.1 provides $\overline{|DFFITS|}_i$ for each event (averaging across t).

Using only Table 3.1 without any other context, the February 2020 event had by far the highest $\overline{|DFFITS|}$, indicating that this event has the most influence on the fitted functional regression equation. All $DFFITS_i(t)$ graphs, $i = 1,\ldots,10$, are shown in the appendix, but Figure 3.11 presents $DFFITS(t)$ for the August 1995 and February 2020 flood events, showing just how large the February 2020 event's $DFFITS(t)$ was. The large $\overline{|DFFITS|}$ for February 2020 was not merely the result of a single extreme spike but rather a truly significant impact throughout the domain of the event, in contrast to the August 1995 event, which has a small spike at the beginning of of its domain but overall is not especially influential on the fitted model. Both the table and the graphs elucidate that based on the $DFFITS$ influence measure the February 2020 event is the most influential event in the functional regression on these river stage curves.

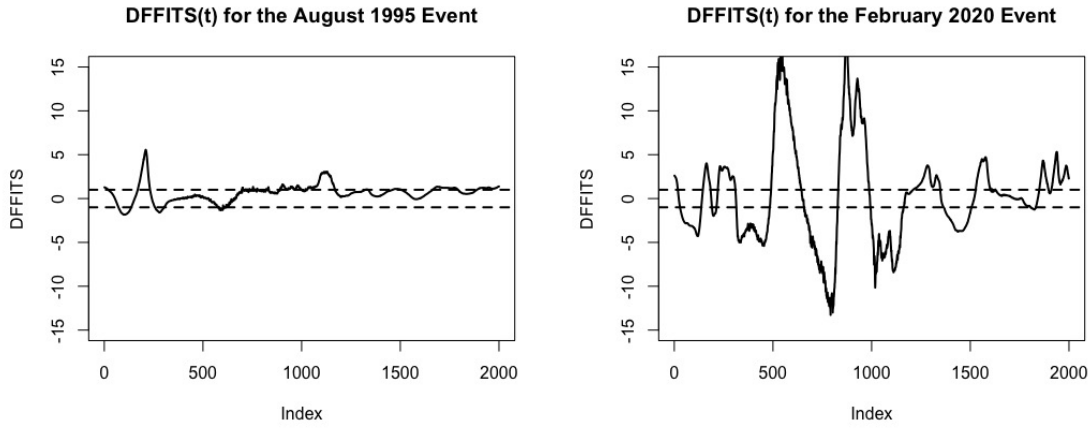


Figure 3.11. $DFFITS(t)$ for the August 1995 (left) and February 2020 (right) flood events (solid curve) as well as an informal cutoff line at ± 1 (dashed lines)

Next, we conducted a similar analysis to assess a functional version of Cook's distance $D_i(t)$, which measures each event's influence on the set of all fitted curves. All ten

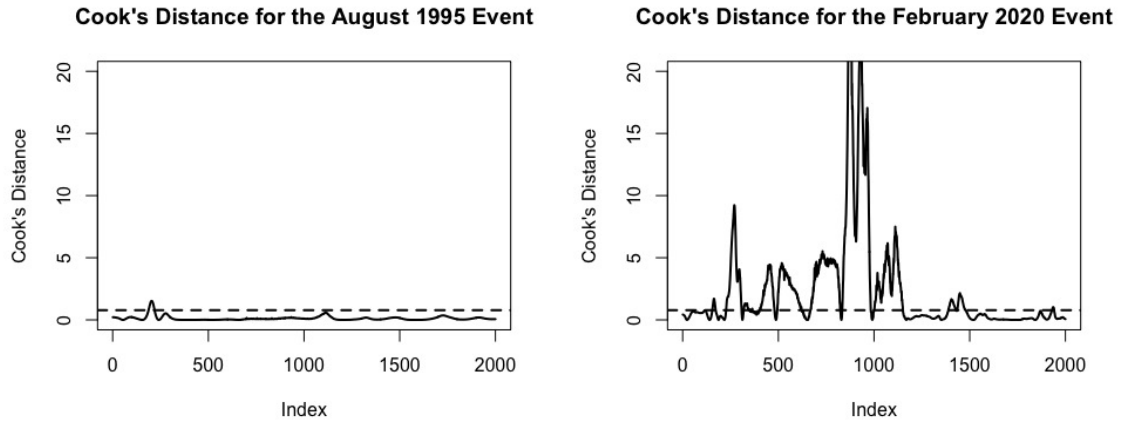


Figure 3.12. Cook's distance $D(t)$ for the 1995 (left) and 2020 (right) flood events, showing how influential the 2020 event is on the set of all fitted curves.

plots of $D_i(t)$ are given in the appendix, but Figure 3.12 shows the measure for the August 1995 (left) and February 2020 (right) events along with a dashed line at $y = 0.757 = F(0.5, 2, 8)$, a customary indicator of a potentially large Cook's distance (Kutner et al., 2005). The plot for the August 1995 event shows that it is generally not influential on the functional regression equation, but the February 2020 event shows by far the highest Cook's distance values of all the events, indicating that this event has the most impact on the set of all fitted curves.

With a large number of functional observations, looking through each observation's $D_i(t)$ graph is not feasible, so examining $\overline{D}_i$, for $i = 1, \ldots, N$, helps quickly locate the most influential events. Table 3.1 confirms that the February 2020 flood event had the highest impact on the set of all fitted curves with $\overline{D} = 2.312$, with the next highest being only 0.436.

The primary function of these metrics is to determine which of ten complete events used in the functional regression is the most influential. We recommend using several diagnostics metrics to determine whether a functional observation is influential in the functional linear model. The values of each of these metrics all point to one

Table 3.2. The probability θ_i that each flood event is selected in the bootstrapped sample for the $\overline{|DFFITS|}$ measure using different choices of α.

Event	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$
August 1995	0.1	0.103	0.108	0.112
February 1998	0.1	0.104	0.111	0.118
March 2003	0.1	0.103	0.109	0.114
May 2003	0.1	0.096	0.088	0.080
September 2004	0.1	0.096	0.088	0.080
March 2007	0.1	0.1	0.099	0.098
February 2010	0.1	0.104	0.110	0.117
May 2013	0.1	0.175	0.103	0.104
November 2018	0.1	0.105	0.115	0.124
February 2020	0.1	0.089	0.069	0.053

main conclusion: The February 2020 flood event had the most influence on the regression model used to reconstruct the October 2015 Cedar Creek curve. It had the largest $\overline{|DFFITS|}$, the highest $\overline{D}$ by a significant amount and the second highest $\overline{|DFBETAS_0|}$ and $\overline{|DFBETAS_1|}$. Additionally, the diagnostic plots for the February 2020 event indicated that the higher average values are not the result of a single spike at only one portion of the event but rather a result of the event truly being more influential over the entire domain.

3.5.2 Applying bootstrapping with perturbations method to river stage data

Since the average range for the ten Cedar Creek curves was 8.413, we generated values of σ from $Uniform(3,5)$, and generated θ from $Uniform(0.5,1)$. We performed $B = 500$ iterations of this bootstrapping with perturbation (generating new values of σ and θ each time), giving us $N = 10$ of each metric for each bootstrap sample for a total of 5000 realizations of each statistic. The empirical distribution of these 5000 realizations approximated the null distribution of each metric. For example, to approximate the null distribution of $\overline{|DFFITS|}$, we let $r_i = \overline{|DFFITS|}_i$, for $i = 1,\ldots,N$, when calculating θ_i (given in Table 3.2).

Table 3.3. The bootstrapped 90th, 95th and 99th percentiles for each influence measure from the approximate null distribution ($N = 10$ and $B = 500$) along with the maximum observed measure from the river stage data.

$\overline{\lvert DFBETAS_0 \rvert}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.535	0.570
95%	0.681	0.718
99%	0.972	1.067
Maximum observed value: 0.552 (Aug. 1995)		
$\overline{\lvert DFBETAS_1 \rvert}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.525	0.560
95%	0.652	0.694
99%	0.978	0.946
Maximum observed value: 0.527 (Aug. 1995)		
$\overline{\lvert DFFITS \rvert}$	$\alpha = 0$	$\alpha = 0.5$
90%	1.991	1.767
95%	2.563	2.238
99%	3.384	3.429
Maximum observed value: 4.062 (Feb. 2020)		
$\overline{D}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.699	0.408
95%	1.346	0.682
99%	2.650	2.417
Maximum observed value: 2.312 (Feb. 2020)		

From the table, we see the selection probability for the events with the largest $\overline{\lvert DFFITS \rvert}$ decreases as α is increased. For example, the February 2020 event had the largest $\overline{\lvert DFFITS \rvert}$, and its selection probability was about half as large when $\alpha = 0.5$ relative to when $\alpha = 0$ (equal selection probability), ensuring that event did not affect the bootstrapped percentiles unduly.

We repeated this process for each influence measure of interest, where the selection probabilities θ_i for each observation were calculated using the observed influence measure for the ith observation. The resulting 90th, 95th, and 99th percentiles from each measurement's approximate null distribution, along with the maximum observed value for each metric, are given in Table 3.3.

The August 1995 flood event had the largest influence on the fitted regression coefficients. Its $\overline{|DFBETAS_0|} = 0.552$ and $\overline{|DFBETAS_1|} = 0.527$. Table 3.3 shows that these averages fell slightly above the 90th percentile of the approximate null distribution of $\overline{|DFBETAS_0|}$ for $\alpha = 0$ but slightly below that percentile in the approximated distribution when $\alpha = 0.5$. This indicates that while this observation does have the highest influence on the functional intercept estimate, it is not significantly large. The same conclusion holds true for the influence on the functional slope estimate, measured by $\overline{|DFBETAS_1|}$. The August 1995 event had the largest observed $\overline{|DFBETAS_1|}$, but it barely surpassed the 90th percentile when $\alpha = 0$ and was below the 90th percentile when using $\alpha = 0.5$ which is the recommended value since the sample size is small.

The February 2020 flood event had the largest $\overline{|DFFITS|}$ by a wide margin. The observed value for the February 2020 event's $\overline{|DFFITS|}$ was 4.062, which far exceeded the approximate null distribution's 99th percentile for either α, indicating that the observed $\overline{|DFFITS|}$ for the February 2020 event had a significant impact on the regression model's prediction of its response. Evidence of its influence is strengthened by the approximate null distribution of $\overline{D}_i$, measuring how much all the fitted values change when the ith observation is deleted. The February 2020 event's $\overline{D} = 2.312$, which fell beyond the null distribution's 95th percentile for every α. Clearly the February 2020 flood event had a significant impact on the fitted functional regression model results and should be further investigated.

There are many potential reasons that the February 2020 flood event stands out as more influential than the others using these diagnostic measures. Of the ten events, the February 2020 flood event had the highest recorded Congaree River crest. The difference between this crest and the next highest crest (March 2003) was greater than the difference between the March 2003 Congaree crest and the lowest crest of any event in the sample (May 2003). This large difference in stage crest could be one factor that leads to the February 2020 flood event standing out as influential in the fitted model.

3.6 Application: Air and Water Temperature Along the United States Coastlines

At any given time of year, the air and water temperature at a specific location are strongly related. In this section, we quantify this relationship across the year 2020 using 35 United States coastline stations that record the local air and water temperature in six-minute intervals throughout the year, for a total of 87,600 potential measurement time points. We obtained the data from the National Data Buoy Center (National Oceanic and Atmospheric Administration, 2021). To be eligible for our sample, stations needed to have at least roughly 90% non-missing values for each of air and water temperatures over the 87,600 timepoints in 2020. We first preprocessed these data and then fit the concurrent functional model to establish a general relationship between air and water temperature across 2020. We then used our functional influence detection procedure to identify locations with the most influence on the model estimates, perhaps due to having a significantly different air and water temperature relationship compared to other locations.

These 35 locations are located all around the United States coastline, including East Coast, West Coast, Gulf of Mexico, Alaskan coastline, and Hawaii. The station locations are displayed in Figure 3.13, and each specific location is listed in the appendix.

For each set of temperature curves, there is a lot of day-to-day variability, there are a handful of missing temperature readings, and the records are generally recorded every six minutes, leading to datasets with over 80,000 records. Therefore, before the regression, we used linear interpolation to fill in any missing records, then smoothed out the daily variation to focus on the yearly trends. Lastly, while preserving the underlying relationship between air and water temperature throughout the year, we resized the length of each smoothed discretized curve to 1000 equally-spaced observations across the year to speed up the functional calculations. The resulting smoothed air and water temperature curves can be found in Figure 3.14, with two specific air and water temperature

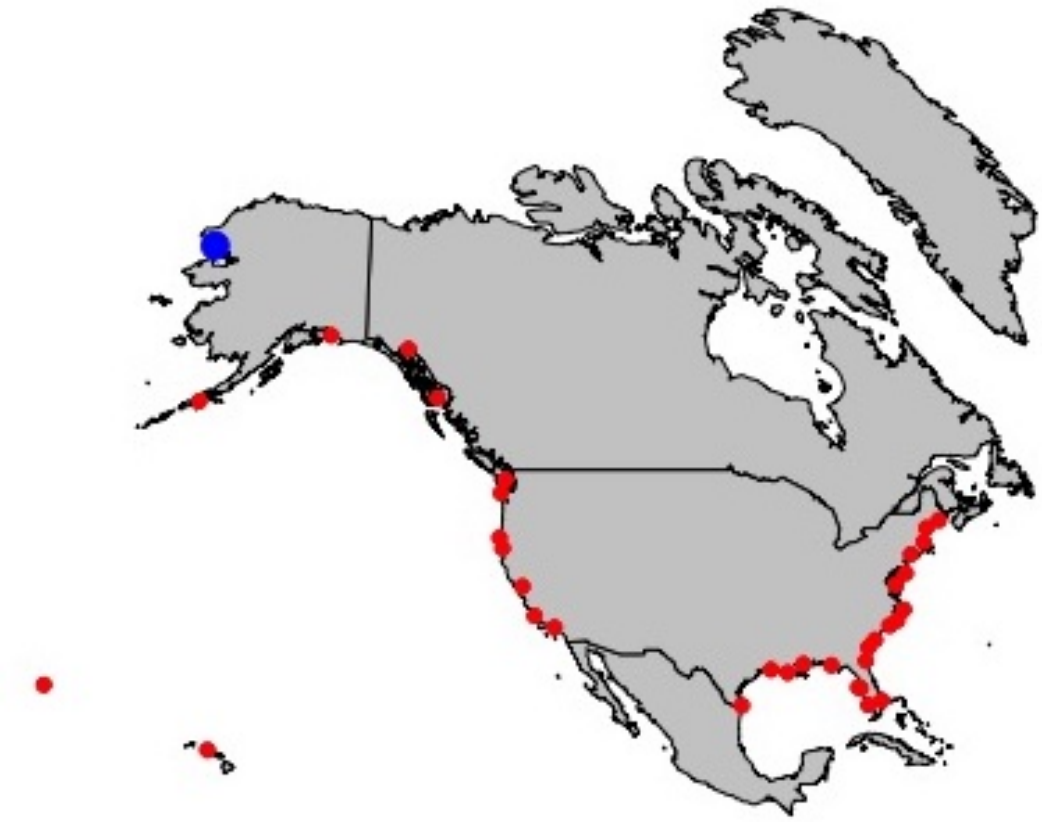


Figure 3.13. Exact location of each station used to create the functional regression model between air and water temperature. The map was create using the `mapproj` package in `R` (McIlroy et al., 2020).

curves shown in Figure 3.15. Note that the low gold curve in Figure 3.14 is from Red Dog Dock, Alaska, which is depicted with the blue dot in Figure 3.13.

3.6.1 Applying the time-dependent influence measures to air and water temperature data

We represented each of these 35 pairs of functional observations using 21 B-spline basis functions. We then estimated $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ from functional regression equation (3.5) (estimates shown in Figure 3.16).

Note that the slope was positive year-round. This indicates that no matter the time of year, as air temperature increases, water temperature also increases. This is intuitive,

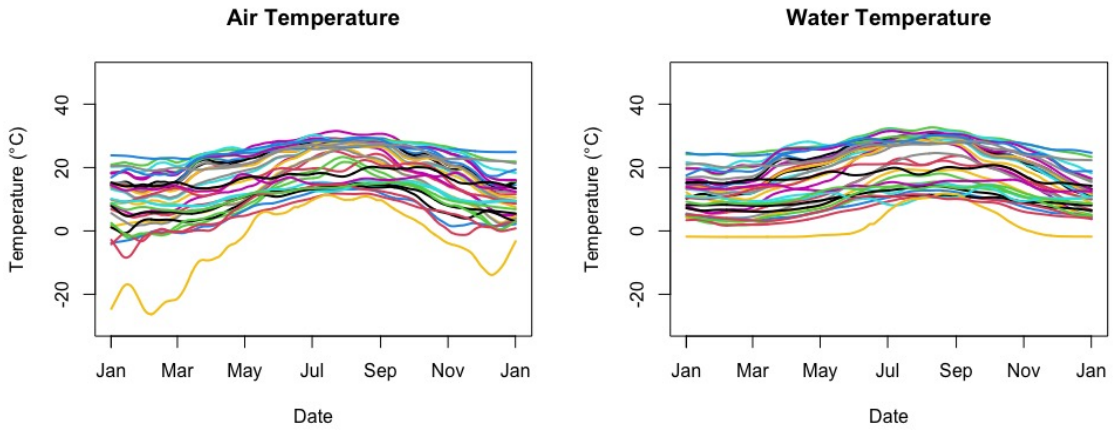


Figure 3.14. All 35 smoothed Air (left) and Water (right) temperatures used in the model.

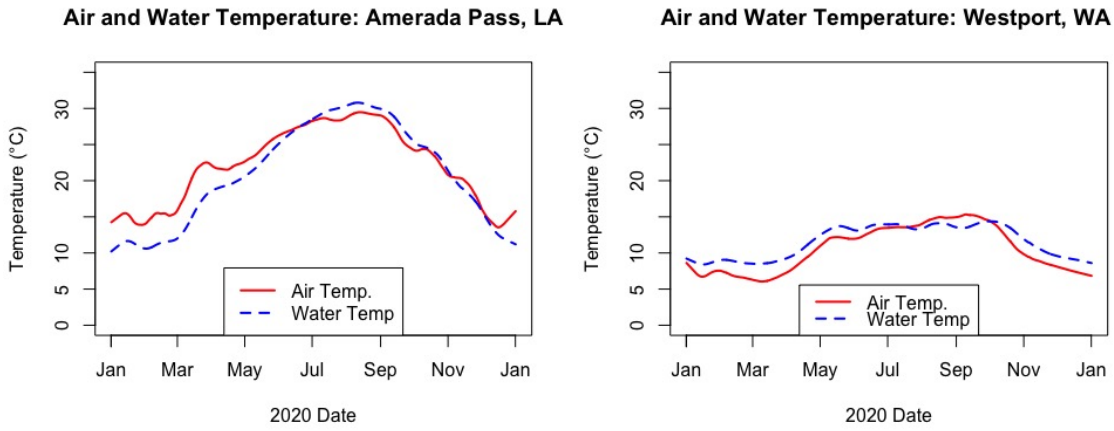


Figure 3.15. Air and Water Temperature for Amerada Pass, Louisiana (left) and Westport, Washington (right).

but we also see that the strength of this relationship is not constant throughout the year. During the summer months, an increase in air temperature results in a larger increase in water temperature than in the winter months on average.

The main purpose of this example is to apply our influence measures analysis on a real dataset rather than to predict missing water temperatures using their corresponding air temperatures; however, prior to our investigation of influence, we confirmed that a leave-one-out model did a good job of predicting the omitted functional response.

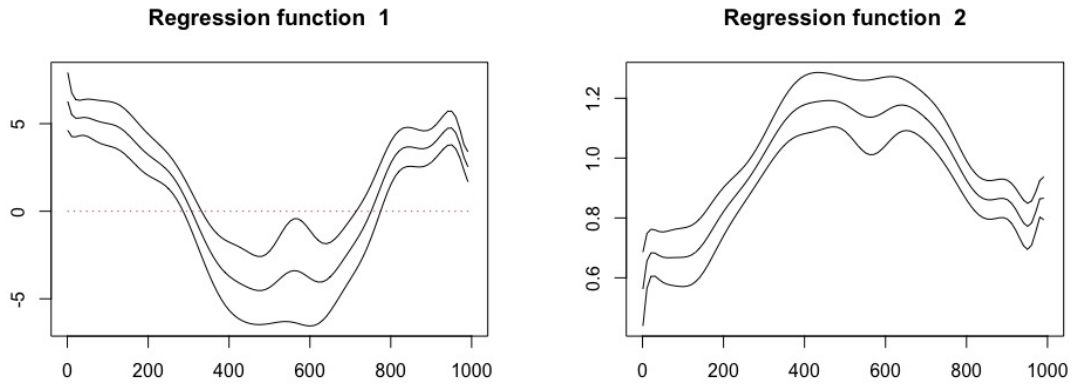


Figure 3.16. Estimated functional intercept $\hat{\beta}_0(t)$ (left) and estimated functional slope $\hat{\beta}_1(t)$ (right) and corresponding 95 percent confidence interval.

We calculated each influence measure ($\overline{|DFBETAS_0|}_i$, $\overline{|DFBETAS_1|}_i$, $\overline{|DFFITS|}_i$, and $\overline{D}_i$) for each of the 35 functional observations. The complete table of results can be found in the appendix. In general, the only observation that visually stood out as more influential than the rest was the aforementioned Red Dog Dock station in Alaska. All of its influence measures were at least twice as large as the next highest, indicating that it was likely very influential on the model; our bootstrapping with perturbations method can confirm that this observation was influential and can evaluate the potential influence of the other observations.

The average range of water temperatures (γ) is approximately 13.2 across the stations, so we generated σ from $Uniform(4.4, 6.6)$ in our perturbation method. We performed our method for each metric with $B = 100$ bootstrap iterations and used $\alpha = 0$ and $\alpha = 0.5$. Table 3.4 gives the resulting percentiles for each metric.

For every influence measure, Red Dog Dock's measure (observation 31) was well above the 99th percentile, indicating that it was highly influential on the regression equation. This makes sense given how much lower the air temperature was at this location in the winter months compared to the rest of the observations while the water

Table 3.4. The bootstrapped 90th, 95th and 99th percentiles for each influence measure from the approximate null distribution along with the maximum observed measure from the air and water temperature data.

$\overline{\lvert DFBETAS_0 \rvert}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.216	0.263
95%	0.284	0.322
99%	0.678	0.497
Maximum observed value: 1.418 (Red Dog Dock)		
$\overline{\lvert DFBETAS_1 \rvert}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.214	0.230
95%	0.274	0.288
99%	1.027	0.437
Maximum observed value: 1.426 (Red Dog Dock)		
$\overline{\lvert DFFITS \rvert}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.795	0.861
95%	0.876	0.960
99%	1.240	1.200
Maximum observed value: 1.863 (Red Dog Dock)		
$\overline{D}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.022	0.026
95%	0.027	0.031
99%	0.359	0.043
Maximum observed value: 0.623 (Red Dog Dock)		

temperature was not as low, proportionally. Additionally, given how much larger the influence of this observation was compared to the rest, using the percentiles calculated using $\alpha = 0.5$ is most appropriate.

While the other Alaskan stations had moderate influence, the next highest $\overline{|DFBETAS_0|} = 0.255$ for the Port Orford station. If the unweighted sampling probabilities were used ($\alpha = 0$), this station's $\overline{|DFBETAS_0|}$ was above the 90th percentile; however, if the effect of the most influential observations was dampened ($\alpha = 0.5$), we conclude that this event does not have a significant impact on the functional intercept estimate.

The Fernandina Beach location in Florida had the second highest $\overline{|DFBETAS_1|} = 0.286$. Based on the null distribution with $\alpha = 0.5$ this value was well above the 90th and near the 95th percentile, indicating that it also has a notable influence on the slope estimate.

Atlantic City, NJ had the second largest $\overline{|DFFITS|} = 0.957$. Similarly, when using $\alpha = 0.5$, this station was easily above the 90th and near the 95th percentile, indicating that it also has a noteworthy influence on the fitted values from the model.

The functional Cook's distance had different results than the others. The largest observed $\overline{D} = 0.623$ (Red Dog Dock), and the second largest was 0.026, so that intuitively Red Dog Dock is an influential observation. Note that when using a positive α so that the more influential observations are being sampled with a low probability, the effect Red Dog Dock itself had on the percentiles was nullified and the 99th percentile decreased to within the range of the rest of the observed $\overline{D}_i$ measures. This shows the benefit of the weighted sampling, because with $\alpha > 0$ there is more support that the observation with $\overline{D} = 0.026$ (Atlantic City) is a moderately influential observation, as it was above the 90th percentile even when the effect on the null distribution of the most influential observations was nullified. Examining these measures collectively, it is clear that the Red Dog Dock location has a substantially large amount of influence on the functional regression model, which makes sense given the observed air temperature curve and the location of the station. Additionally, our method identifies the Atlantic City observation as also influential on the functional regression model. In a complete functional regression analysis of these data, we recommend investigating these observations more closely for possible removal.

3.7 Conclusion

Our method successfully offers a practical way of identifying influential functional observations in the concurrent model. By formulating the ordinary regression influence metrics as a function of time and then averaging them across t for each observation, we successfully detect the observations with the most influence on the estimates and predictions from the model. Additionally, simulation shows that our bootstrapping with perturbations approach performs well in identifying the most influential observations as significant. In both the river stage example and the air and water temperature example, we sensibly identify certain observations as more influential than the rest, and then the bootstrap method confirms their influence is significantly large, further illustrating that our method is appropriate to identify influential functional observations in the concurrent model.

3.8 Additional Uses for the Concurrent Model

Cross validation is a common model checking method in which the data are split into training data used to fit the model and testing data, used to assess the validity of the model using observed but held-out data not used in the fitted model. When comparing the known response values from the test data to the corresponding predictions of the test data observations using the model, major discrepancies could signify that the model is not appropriate. A similar method can measure the appropriateness of the fully functional linear model on two sets of concurrent curves.

Even though each of the previous flood events was selected to emulate the October 2015 Congaree River curve, performing a leave-one-out cross validation-like analysis is helpful to see if the concurrent model is even appropriate for predicting river stage data. In this case, one entire flood event is left out and then that response curve is predicted using the same method used to predict the October 2015 Cedar Creek stage, and then the

prediction is compared to the known, true stage during that event. This method is how we discovered the abrupt 3-foot jump in the Cedar Creek stage record in 1998 that was mentioned in Section 2.2.3. The actual stages for each of the events prior to this jump were systematically lower than the predicted stage for the event using the concurrent model and the remaining known events.

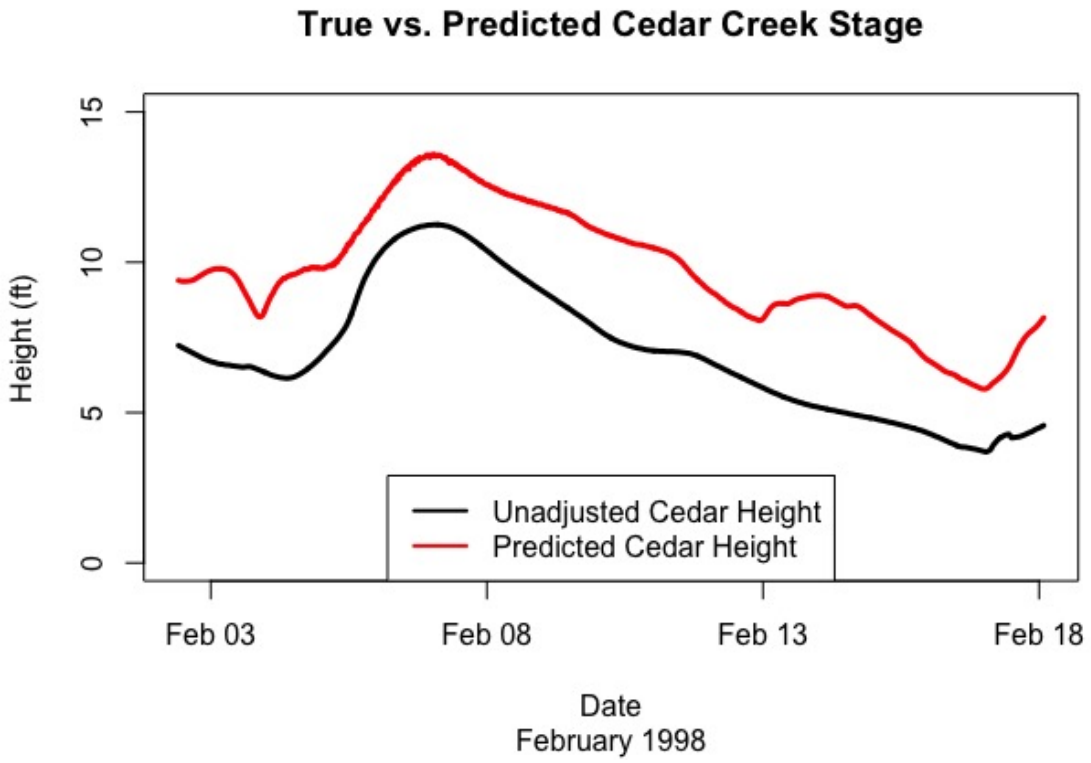


Figure 3.17. True and predicted Cedar Creek height during the 1998 flood before adding 3 feet to the true value.

Figure 3.17 shows that the predicted value for the February 1998 flood using the remaining nine events (even the 1995 Cedar Creek unadjusted stage) was significantly higher than the true stage during this event. The post-1998 events are predicted relatively well even with the inaccurate Cedar Creek levels. This indicated that the pre-1998 flood events needed investigating, and we discovered that we needed to add three feet to all Cedar Creek stages prior to October 1, 1998, which was the beginning of the new water year and marked a shift in baseline measurements at the Cedar Creek gage loca-

tion. Figure 3.18 shows how accurately the concurrent model predicts this flood once the 3-foot adjustment was made.

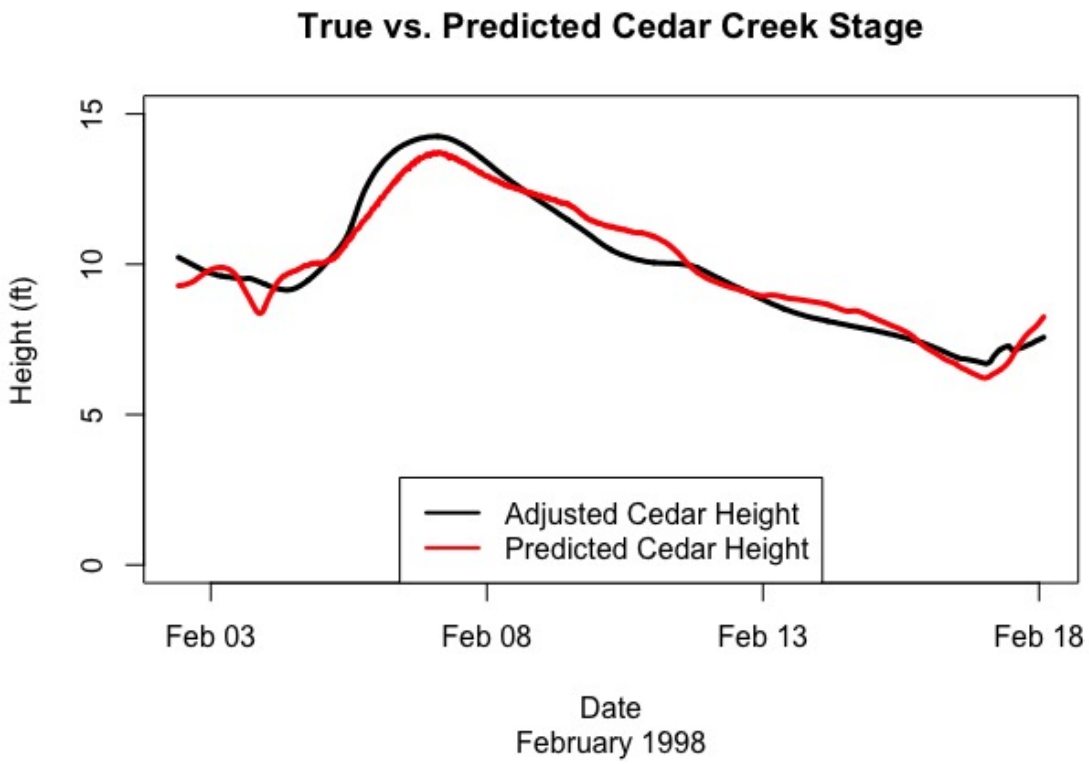


Figure 3.18. True and predicted Cedar Creek height during the February 1998 flood after adding 3 feet to the Cedar Creek data prior to the baseline shift.

Once that adjustment was completed, we repeated the leave-one out cross-validation method and the overall results support that using the concurrent model for this type of flood data is an appropriate method for prediction/reconstruction. Figure 3.19 shows one example of the cross-validation results. The appendix includes CV graphs for all ten events. Overall, the predicted stage is consistent with the true height, further supporting this process.

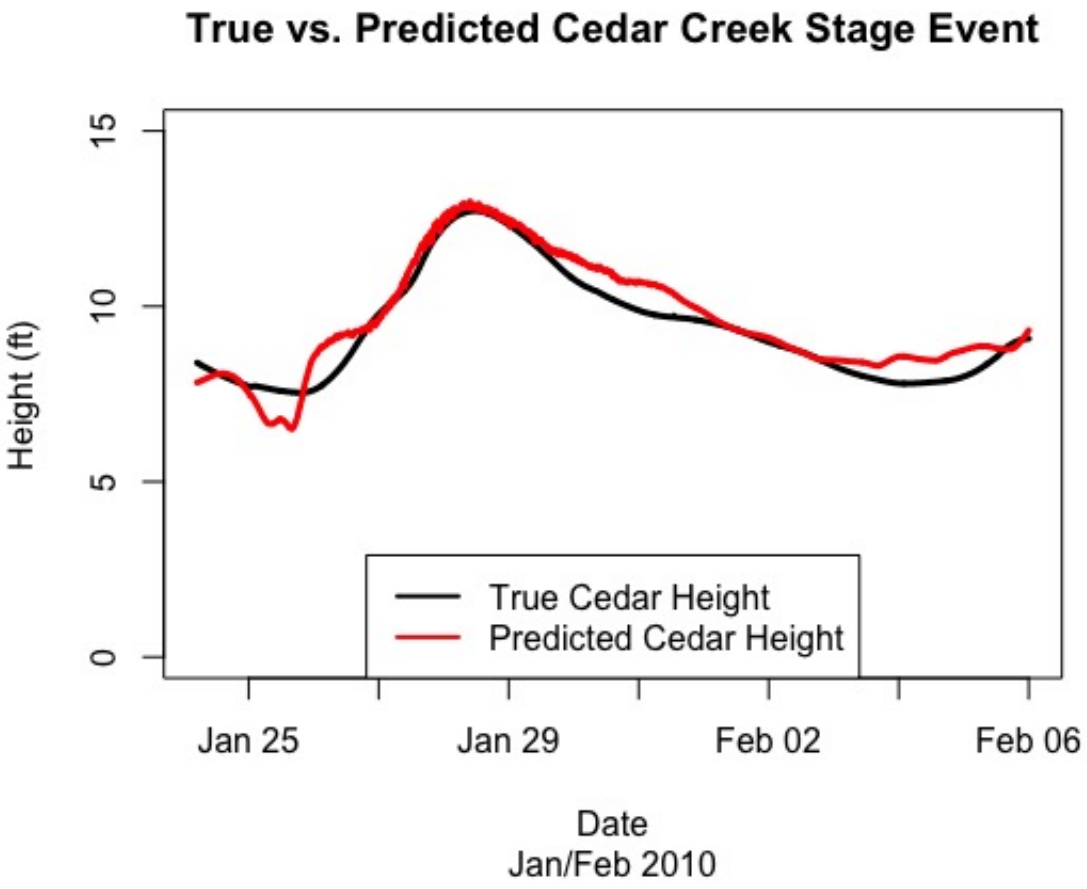


Figure 3.19. True and predicted Cedar Creek height during the February 2010 flood event

Recognizing the discrepancy in the Cedar Creek data was a very important step in our functional data analysis. Without making the 3-foot adjustment, the reconstructed October 2015 Cedar Creek stage would not be accurate. Therefore, while plotting the cross-validated curve alongside the observed curve is not an advanced idea, it is a useful and recommended procedure when using any form of functional data analysis for reconstructive purposes.

CHAPTER 4

FUNCTIONAL REGRESSION MEASURES OF INFLUENCE ON PREDICTION OF AN OUT-OF-SAMPLE OBSERVATION

4.1 INTRODUCTION

In traditional regression, it is common to measure the influence an observation has on a fitted model. In this chapter, we present two new measures of influence Δ_i and Accumulated Influence Percentiles (AIP) to use with the concurrent functional regression model that calculate how the prediction of an out-of-sample response curve of interest, say $\hat{Y}^{new}(t)$, changes when the concurrent functional model is fit with and without each functional observation in turn. This is different than the traditional influence measures because it examines how the ith observation influences the prediction of a new target response curve $\hat{Y}^{new}(t)$ when the new out-of-sample observation is separate from those used to fit the model. These measures are useful because an observation may not be influential when fitting regression models using the preexisting data, but when considering the model's prediction of an out-of-sample observation, an observation could be influential based on the corresponding predictor $X^{new}(t)$ functional observation. In practice, knowing which observations have the most influence on a specific new prediction and whether or not that observation's measure is significant is critical in constructing reliable predictions of new response curves in the functional regression setting. Our measures produce a single Δ_i and AIP_i for the ith functional observation $i = 1, \ldots, N$, where the larger values indicate that the observation has higher influence when predicting an external observation's response curve. While Δ_i is a metric that indicates whether

or not an entire observation is influential in the prediction by measuring the magnitude of the change, AIP_i is similar but takes into account the duration of the influence spanning throughout the functional observation. Furthermore, this study presents a similar method to further determine whether a large influence measure value is caused by only a small portion of the functional observation, rather than across the entire observation. After describing both measures in detail, including a weighted bootstrap method to determine whether an observation's measure is significantly large, we present a simulation study supporting the effectiveness of each measure. We then apply our proposed measures to two real datasets.

4.1.1 Influence Measure: Δ_i

The concurrent functional regression model is:

$$Y_i(t) = \beta_0(t) + \beta_1(t)X_i(t) + \epsilon_i(t),\ i = 1, \ldots, N \tag{4.1}$$

where $\beta_0(t)$ is the functional intercept and $\beta_1(t)$ is the functional slope that relates predictor ($X_i(t)$) and response ($Y_i(t)$) observations at each t. Estimates of $\beta_0(t)$ and $\beta_1(t)$ are computed using all N pairs of functional data represented using a set of basis functions, and the resulting $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ are used in the fitted functional regression model along with the external observation's known predictor curve $X^{new}(t)$ to calculate $\hat{Y}^{new}(t)$. The fitted model can be obtained with the `fda` package (Ramsay, Graves, and Hooker, 2020) in R (R Core Team, 2020). The formula for Δ_i is

$$\Delta_i = \sqrt{\int (\hat{Y}^{new}(t) - \hat{Y}^{new}_{(i)}(t))^2 dt}, \tag{4.2}$$

which we approximate using trapezoidal approximation via the `trapz` function in the `pracma` package (Borchers, 2019). $\hat{Y}^{new}(t)$ is the predicted value for a new response curve using all N sets of $X_i(t)$ and $Y_i(t)$ observations. $\hat{Y}^{new}_{(i)}(t)$ is that same prediction from a model fit with observation i withheld. The observation with the highest Δ_i value

has the most overall influence on the prediction of $\hat{Y}^{new}(t)$ and should be further investigated, especially if this value is significantly larger than the next highest Δ_i value. In section 4.2, we describe a weighted bootstrapping approach to approximate the null distribution of Δ for a particular out-of-sample target observation, which can be used to determine whether any observed Δ_i measures are significantly large.

4.1.2 Influence Measure: Accumulated Influence Percentiles (AIP_i)

The previously discussed measure calculates the overall influence each observation has on an out-of-sample prediction. Accumulated Influence Percentiles (AIP) also calculates the difference in the prediction of an external observation (with and without each observation), but accounts for whether that difference is entirely at a single portion of the functional observation or spread throughout it. Additionally, AIP combines the magnitude of the differences with the duration of the differences, forming a single measure of influence for each of the N observations. The formula for AIP_i is:

$$AIP_i = \int_0^1 \phi_p \left(\left| \hat{Y}^{new}(\mathbf{t}) - \hat{Y}^{new}_{(i)}(\mathbf{t}) \right| \right) dp, \ \mathbf{t} = (t_1, \ldots, t_n) \tag{4.3}$$

where $\phi_p(\mathbf{x})$ is the pth percentile of the values in the vector $\mathbf{x}$. If an observation has a large Δ, but it is caused by a large difference at a single portion of the prediction, then the resulting percentile curve of absolute differences will be relatively flat across $p \in [0,1]$ and increase sharply at the higher percentiles. Taking the area under the curve minimizes the effect of a single large difference and balances the magnitude of the difference and the duration of the difference.

4.2 Approximating a null distribution of Δ and AIP

To determine the significance of the observed influence measures, we perform a weighted bootstrapping method that approximates a null distribution of both Δ and AIP, i.e. a distribution for the metric under the condition that there is no especially influential

curve. To accomplish this, when selecting our bootstrap sample we propose to sample the apparently less influential observations from our observed curves more often than the apparently most influential observations. We calculate the influence measure Δ_i or AIP_i (generally denoted as r_i) for each observation, and then use the following equation to translate the metric value for observation i into a selection probability θ_i:

$$\theta_i = \frac{(1/r_i)^\alpha}{\sum_i [(1/r_i)^\alpha]}, \quad \alpha \geq 0. \tag{4.4}$$

Note that $\alpha = 0$ corresponds to equal selection probabilities for each observation, and in general α should not exceed 0.5 and is most crucial when N is small. Once determining the influence measure of interest (Δ or AIP), we provide the following weighted bootstrap method to approximate a null distribution of the measure:

1. Calculate r_i for each observation.

2. Select an appropriate value of α (or allow a range of choices) and calculate θ_i for $i = 1, \ldots, N$.

3. Sample N observations with replacement from the original set of data, where the ith observation has probability θ_i of being selected.

4. Each bootstrap sample then consists of N functional pairs $\{(X_1^*(t), Y_1^*(t)), \ldots, (X_N^*(t), Y_N^*(t))\}$.

5. Using these new pairs of functional data, fit the concurrent functional regression model and calculate r_i for each observation $i = 1, \ldots, N$.

6. Repeat Steps 3-5 for the desired number of bootstrap iterations (B) to obtain NB values of the measure, which approximate a null distribution for that influence measure.

7. The original metric from the observed dataset can be compared to percentiles from the respective bootstrap distribution to determine whether the largest val-

ues identified in the original data analysis are significantly large relative to the null distribution.

The ideal value of α in Equation (4.4) varies based on the observed measures from the initial dataset. In general, we recommend using $\alpha = 0.5$ when N is small or when one of the observed measures is noticeably larger or smaller than the rest. If values of the metric have little variability, the bootstrapped percentiles will be similar regardless of $\alpha \in (0, 0.5)$; however, when the observed influence measures are more spread out or one observation's influence measure is much larger than the rest, using $\alpha = 0.5$ dampens the effect that the observation has on the bootstrap sample, resulting a bootstrap sample that better resembles a null distribution. This allows truly significant influential observations to be flagged rather than be dominated by the values for the most influential observations. For large sample sizes, an observation with a large influence measure has less impact on the approximate null distribution since it is less likely to be sampled in a given iteration (regardless of the value of α) compared to when sample size is small; therefore, using $\alpha = 0$ in large sample scenarios is appropriate. If the sample size is moderate, or it is unclear whether the largest influence measure is much larger than the next highest, we recommend using both $\alpha = 0$ and $\alpha = 0.5$ separately and comparing the resulting percentiles to see the effect of the more influential observations.

After performing this bootstrapping method, the 90th, 95th, and 99th percentiles can then be used to identify the significantly influential functional observations by comparing the observed measures from the original dataset to those percentiles.

4.3 Simulation Study

To elucidate the effectiveness of these new influence measures and the bootstrap method, we perform a simulation study in which we generate a concurrent set of functional predictor and response observations and intentionally contaminate one of the response curves. After generating an a additional predictor curve which corresponds to a hypo-

thetical out-of-sample target observation, we calculate for each observation the measure of influence on the prediction of the response curve of the target observation. Once we have the influence measure for each observation, we perform the bootstrap method to determine where the observed influence measure of the contaminated value falls with respect to the null distribution.

We investigate the performance of our method in identifying influential observations in a simulation study. For this example, we generate as simulated predictor functions N independent $X(t)$ curves over a grid of values $t \in \{1, 2, \ldots, 1000\}$ using the following formula:

$$X(t) = (t/12)[a_s \sin[(1/k_s)(t - d_s)] + c_s][a_c \cos[(1/k_c)(t - d_c)] + c_c]$$

where each of the N curves is generated by randomly selecting the parameters within the equation as follows:

- a_s, a_c, c_s and c_c are independently sampled from the list $\{-3, -2, -1, 0, 1, 2, 3\}$.

- k_s and k_c are sampled from the list $\{-300, -200, -100, 100, 200, 300\}$.

- d_s and d_c are sampled from the list $\{-100, -50, 0, 50, 100\}$.

By alternating the combination of parameters used to generate the functional data, we produce curves that are similar and follow the same underlying signal curve $m(t) = t/12$. An example of $N = 20$ such $X(t)$ curves is shown in Figure 4.1. Note that the simulation results are not changed if the parameters' ranges are expanded as long as they are the same for all N curves. We set the functional slope and intercept functions to be:

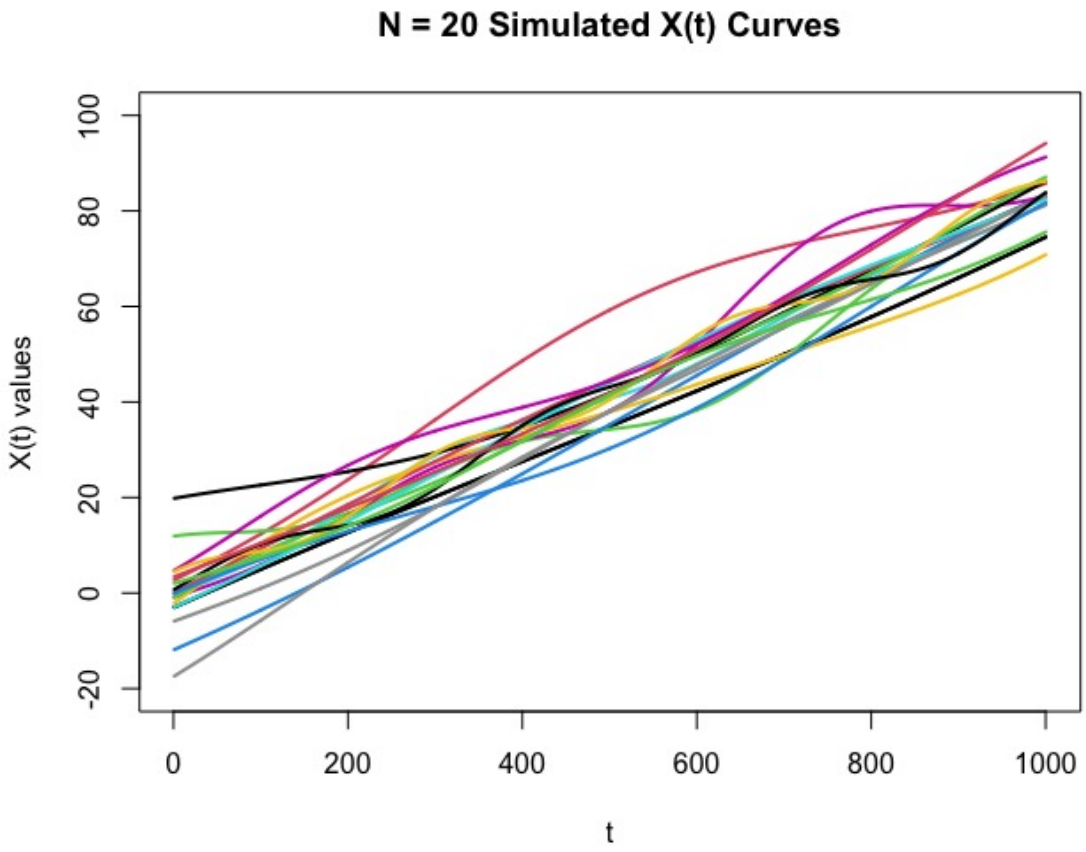


Figure 4.1. Example of $N = 20$ generated $X(t)$ curves using the described functional data generation method.

$$\beta_0(t) = \cos(t/200) + 2$$

$$\beta_1(t) = \sin(t/200) + 2$$

We dampen the relationship between the predictor and response curves by generating noise functions $\epsilon_i(t)$ to slightly distort the functional relationship between each pair of simulated $X(t)$ and $Y(t)$ curves: We add realizations of the Ornstein-Uhlenbeck process, approximated using the Euler-Maruyama method, to the mean response curves $Y_i(t) = \beta_0(t) + \beta_1(t)X_i(t),\ i = 1,\ldots,N$. An example of a resulting set of 20 simulated response curves is shown in Figure 4.3.

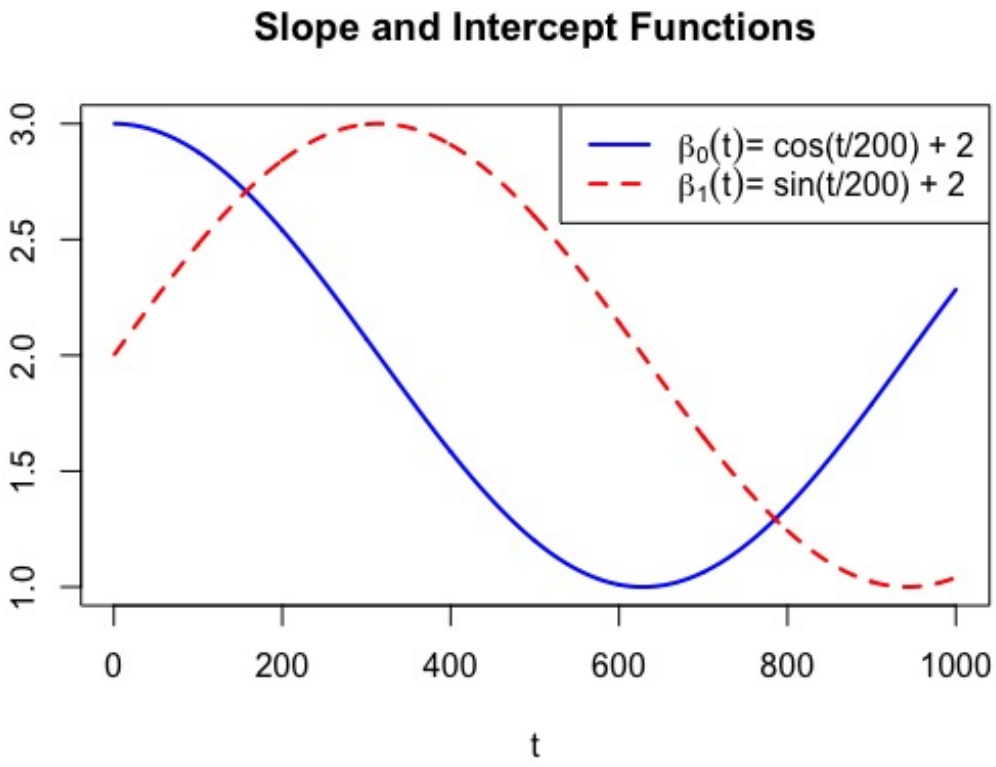


Figure 4.2. Defined functional intercept $\beta_0(t)$ (solid blue) and functional slope $\beta_1(t)$ (dashed red) used to generate response curves $Y_i(t)$ using $X_i(t)$.

As an initial check that these generated data follow our concurrent functional linear model, before introducing any contamination, we generated 100 independent data sets, fit the model for each, and verified the estimates of $\beta_0(t)$ and $\beta_1(t)$ resembled the true functional slope and intercept on average. However, all further analysis was done on simulated data with contamination, as we described next.

We then intentionally contaminate the $\beta_1(t)$ function for one of the N observations and see how often our method identifies the contaminated observation as influential. For this contaminated observation, we let $\beta_1(t) = \lambda \times \sin(t/200) + 2$ for some $\lambda > 0$. Clearly, $\lambda = 1$ represents the control case in which the contaminated observation is generated the same way as the others. In this simulation, we set $\lambda \in \{0.25,\ 0.5,\ 0.75,\ 0.9,\ 1,\ 1.1,\ 1.25,\ 1.5,\ 1.75,\ 2\}$. Figure 4.4 gives an example of $N = 20$ response curves with the

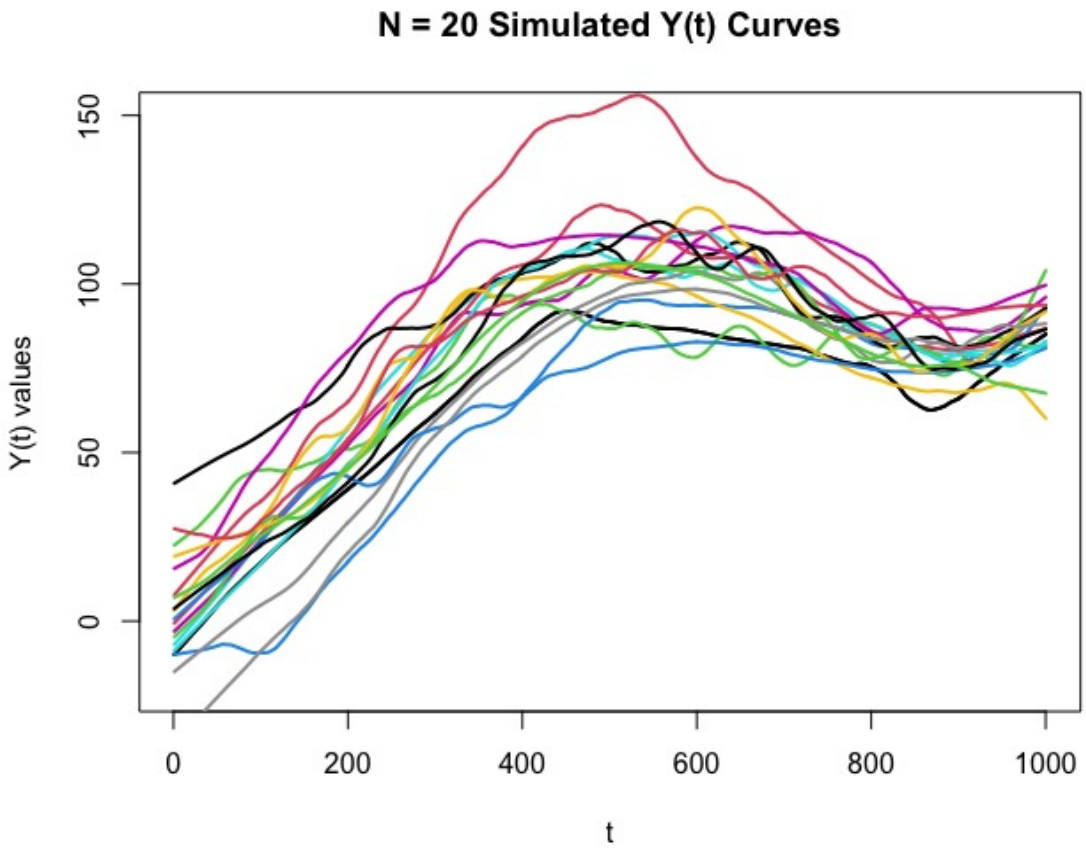


Figure 4.3. Example of N = 20 response ($Y(t)$) curves used in simulation with no contaminated observations ($\lambda = 1$).

contaminated curve generated using $\lambda = 2$.

We implement the following algorithm for combinations of: $N = 100$, $N = 50$, $N = 20$, and $N = 10$; $\lambda \in \{0.25, 0.5, 0.75, 0.9, 1, 1.1, 1.25, 1.5, 1.75, 2\}$; and for $\alpha = 0, 0.5$.

1. Select λ.

2. Generate N sets of $\{X_i(t), Y_i(t)\}$ curves with one $Y_i(t)$ curve contaminated using λ.

3. Generate the predictor curve of an out-of-sample target observation.

4. For $i = 1, \ldots, N$, calculate the measure of influence on the prediction of the response curve corresponding to the target observation.

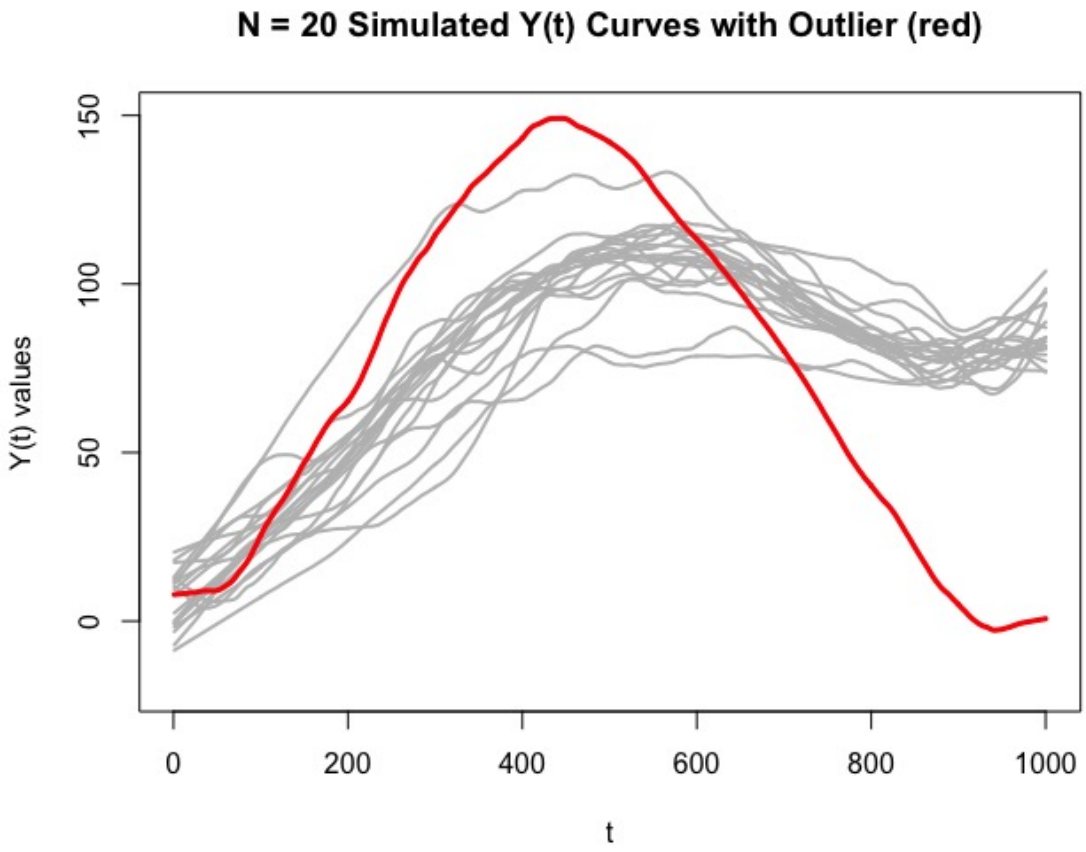


Figure 4.4. Example of N = 20 response ($Y(t)$) curves used in simulation with one outlier (red) using $\lambda = 2$.

5. Select α and calculate the selection probabilities θ_i for each observation using Equation (4.4).

6. Perform $B = 100$ bootstrap iterations, sampling the N observations with replacement, calculating the influence measure for each observation in each iteration (yielding NB values of the measure).

7. Determine the percentile relative to this bootstrap distribution of the originally contaminated observation's influence measure, indicate whether it is above the 95th percentile.

8. Repeat 100 times for each combination of desired influence measure, N, λ, and α.

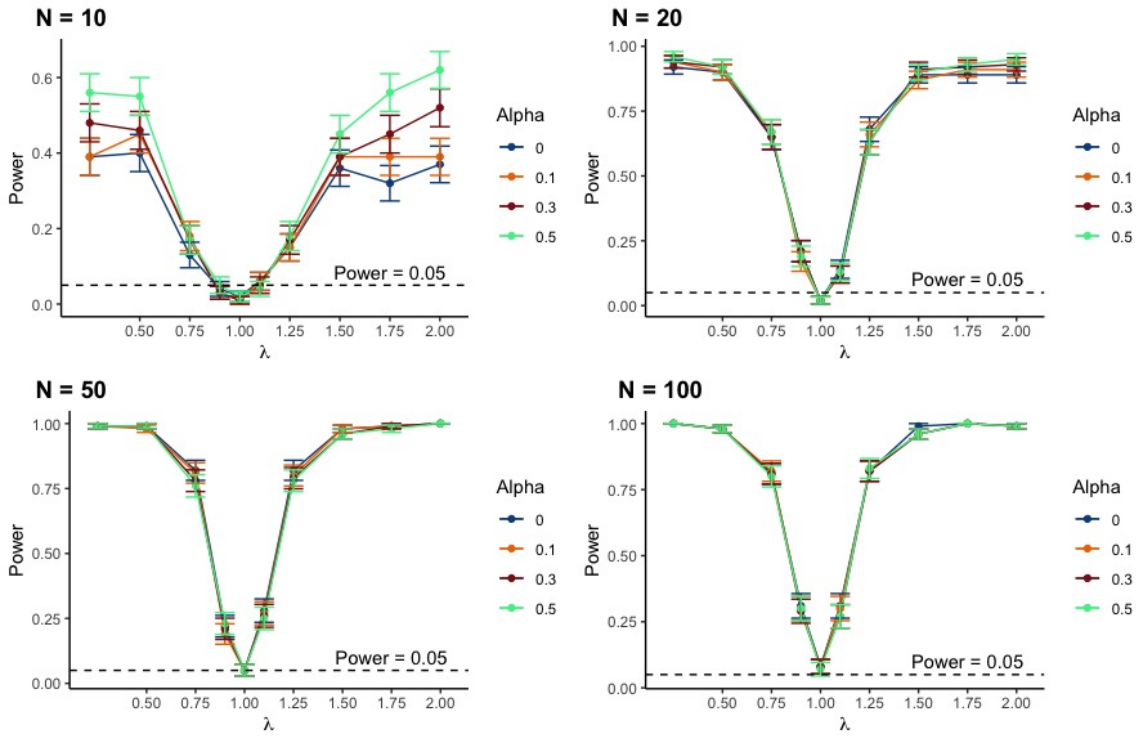


Figure 4.5. Power functions displaying the average proportion of contaminated observations above the 95th percentile from the approximate null distribution of Δ for different values of α (with error bars representing one standard error) for different sample sizes N.

Note that for each data generation, the bootstrapping process is executed using each choice of α on the same generated data.

4.3.1 Simulation study for Δ

Next, we applied this simulation method to our new influence measure, Δ, to test whether this measure successfully identifies the contaminated observation as influential, on average.

Figure 4.5 shows the average proportion of contaminated observations that were above the 95th percentile for Δ_i for $N \in \{10, 20, 50, 100\}$. This is analogous to the power of the procedure. When λ moved away from 1, the proportion of contaminated observations flagged increased. This correctly indicates that when an observation is more extreme, it is flagged as influential more often. Additionally, when N is small, this test

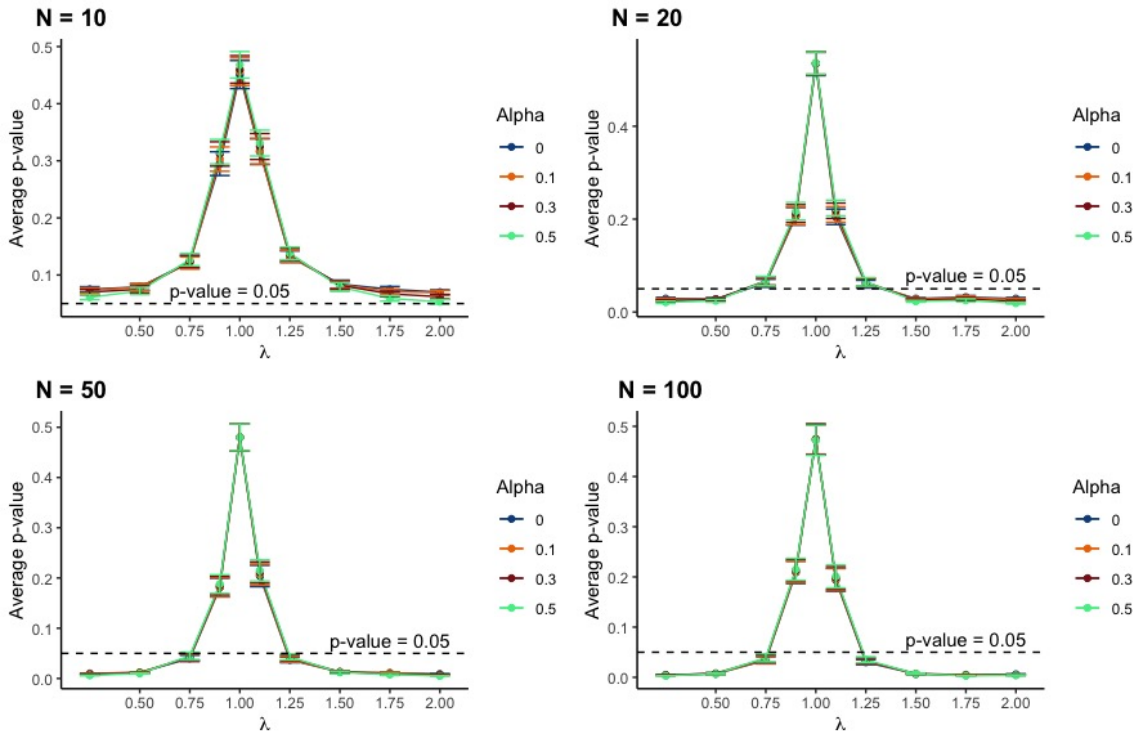


Figure 4.6. Average p-value (1− percentile within bootstrap distribution) of contaminated observations for different values of α (with error bars representing one standard error) for different sample sizes N for Δ.

had a substantially higher power when $\alpha = 0.5$, indicating that choosing $\alpha > 0$ is desirable for small N.

Furthermore, Figure 4.6 provides additional results from the same simulation. Here we plot the average p-value, which is 1 minus the average percentile within the bootstrap distribution of the contaminated observation. When λ moved away from 1, the p-value decreased, indicating that the contaminated observation's influence measure was frequently significant. These results consistently indicate that our method successfully identifies the observations that are truly influential on an out-of-sample prediction and that with a small sample size, using $\alpha = 0.5$ improves the strength of the method.

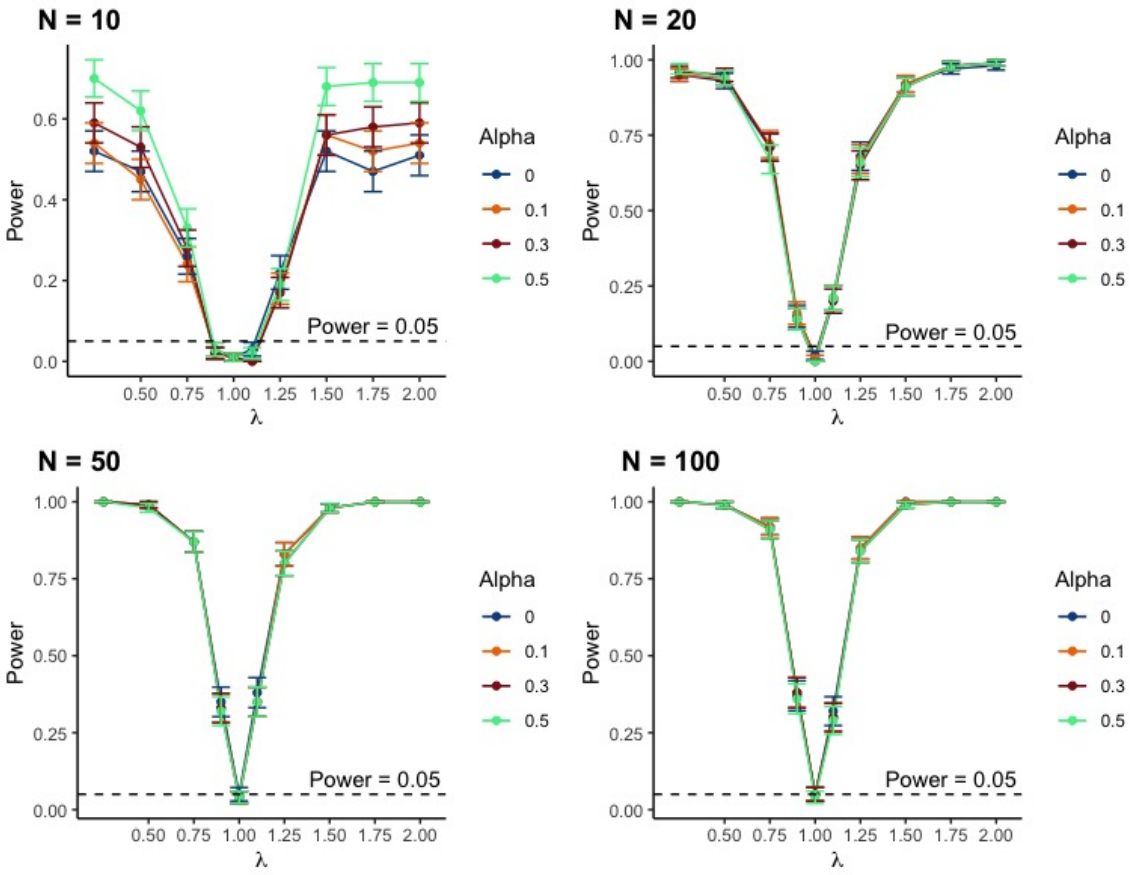


Figure 4.7. Power functions displaying the average proportion of contaminated observations above the 95th percentile from the approximate null distribution of AIP for different values of α (with error bars representing one standard error) for different sample size N.

4.3.2 *AIP* Simulation Study

We performed a simulation to establish the validity of AIP in identifying influential observation in outside prediction. Figure 4.7 and Figure 4.8 show the average estimated power and p-value of the measure.

Figure 4.7 shows the average proportion (i.e. power) of contaminated observations that were above the 95th percentile of the approximate null distribution of AIP for $N \in \{10, 20, 50, 100\}$. When λ moved away from 1, the proportion of contaminated observations flagged increased. This appropriately indicates that when an observation is more extreme, it is flagged as influential more often. Additionally, when N is small, this

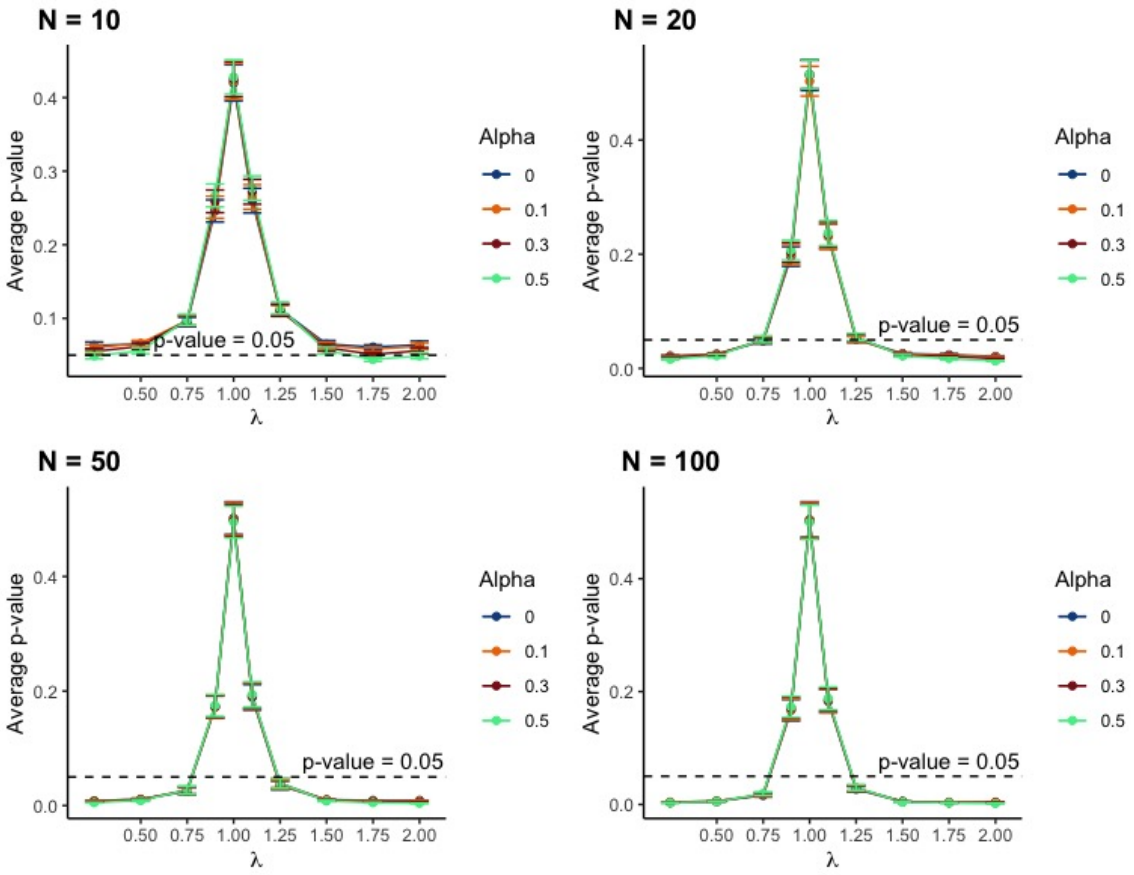


Figure 4.8. Average p-value (1− percentile within bootstrap distribution) of contaminated observations for different values of α (with error bars representing one standard error) for different sample sizes N for AIP.

test had a substantially higher power when $\alpha = 0.5$, again indicating that setting $\alpha > 0$ was useful in this case.

Figure 4.8 provides plots of the average p-value. When λ moved away from 1, the p-value decreased, indicating that the contaminated observation's AIP was often significantly large. These results consistently indicate that our new influence measure, AIP, successfully identifies the observations that are truly influential on an out-of-sample prediction.

4.4 Application: River Stage Data During Flood Events

4.4.1 Analysis of influence using Δ

We applied these measures to river stage data study by Pittman, Hitchcock, and Grego (2021), who looked at river stages from two related gage locations at Congaree National Park near Columbia, South Carolina. They used a landmark alignment technique to objectively determine the optimal start and end points of ten flood events in which the Congaree River (United States Geological Survey, 2020a) flowed over-bank, through the floodplains, and into Cedar Creek (United States Geological Survey, 2020b). This resulted in 10 historic flood events that could be used directly in the concurrent functional model. The purpose of using functional regression was to relate the Congaree River stage to the Cedar Creek stage during flood events. Then this relationship could be used to reconstruct the Cedar Creek stage during a major flood event in October 2015 when this gage went offline, but the Congaree River gage remained functioning. The Δ influence measure helps to determine which of the 10 historic flood events' influence on the eventual reconstruction of the October 2015 Cedar Creek stage and whether any of these events' influences are significantly larger than expected.

We first used Equation (4.1) to reconstruct the missing Cedar Creek stage during the October 2015 flood. We then, repeated the reconstructions, leaving out each one of the functional observations in turn. Figure 4.9 provides an example of the difference between the full reconstruction $\hat{Y}^{new}(t)$ and the reconstruction with the February 2020 event withheld $\hat{Y}^{new}_{(10)}(t)$.

Δ was calculated as the L_2 distance between the two curves using Equation (4.2). The resulting Δ_i for each flood event i is given in Table 4.1 and plots of the October 2015 Cedar Creek reconstruction with and without each event can be found in the appendix.

The March 2003 and February 2020 flood events stand out from the rest of the events when comparing the reconstruction of the October 2015 Cedar Creek curve with and

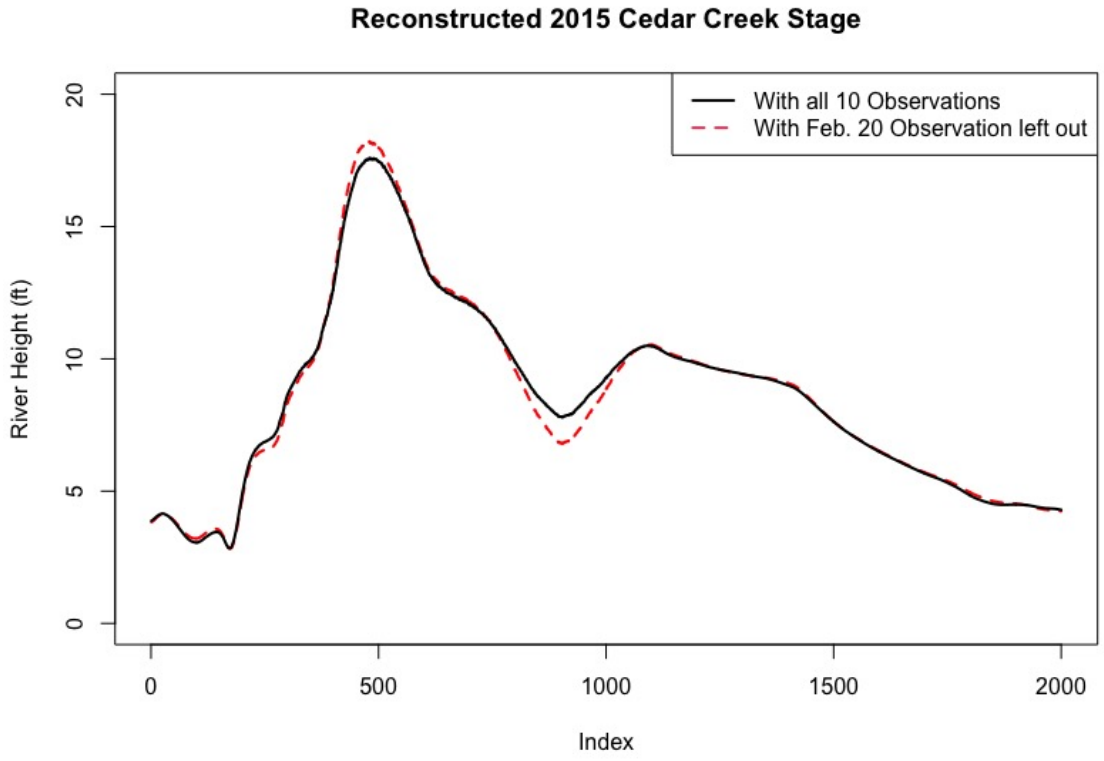


Figure 4.9. The reconstructed October 2015 Cedar Creek stage with all 10 observations (solid black) and with the February 2020 observation withheld (dashed red).

without these individual events. Their Δ_i values were about twice as large as the next highest values, indicating that these two events have the greatest impact on the target event's reconstruction.

With a small sample size ($N = 10$), $\alpha = 0.5$ was used to approximate the null distribution. The March 2003 event had the highest Δ (14.165), but this was not as large as the 90th percentile of 15.070. This indicates that even though there were a couple of observations with a larger Δ compared to the others, none of the ten prior flood events have a significantly large impact on the the reconstructed October 2015 Cedar Creek stage.

4.4.2 Analysis of influence using AIP

In this section we identified potentially influential functional observations using AIP using the river stage data. For this measure, we took the absolute difference between

Table 4.1. The L_2 distance (Δ_i) between the 2015 Cedar Creek reconstructions with all 10 observations included and with each observation individually withheld.

Δ Analysis	
Event	Δ
August 1995	8.706
February 1998	2.559
March 2003	**14.165**
May 2003	6.684
September 2004	6.285
March 2007	8.628
February 2010	2.118
May 2013	3.204
November 2018	5.580
February 2020	13.377

Table 4.2. Δ percentiles of each influential measurement from 5000 bootstrapped river stage observations along with the observed maximum of each metric in the river stage data context. Note that $\alpha = 0.5$ is most appropriate to use given the small sample size $N = 10$.

Bootstrap approximated null distribution's percentiles		
Percentile	$\alpha = 0$	$\alpha = 0.5$
90%	13.253	15.070
95%	18.078	19.730
99%	27.942	27.586
Max Obs. 14.165 (March 2003)		

$\hat{Y}^{new}(t)$ and $\hat{Y}^{new}_{(i)}(t)$ (see, e.g., difference between curves in Figure 4.9). All 10 observed absolute difference functions are shown in Figure 4.10.

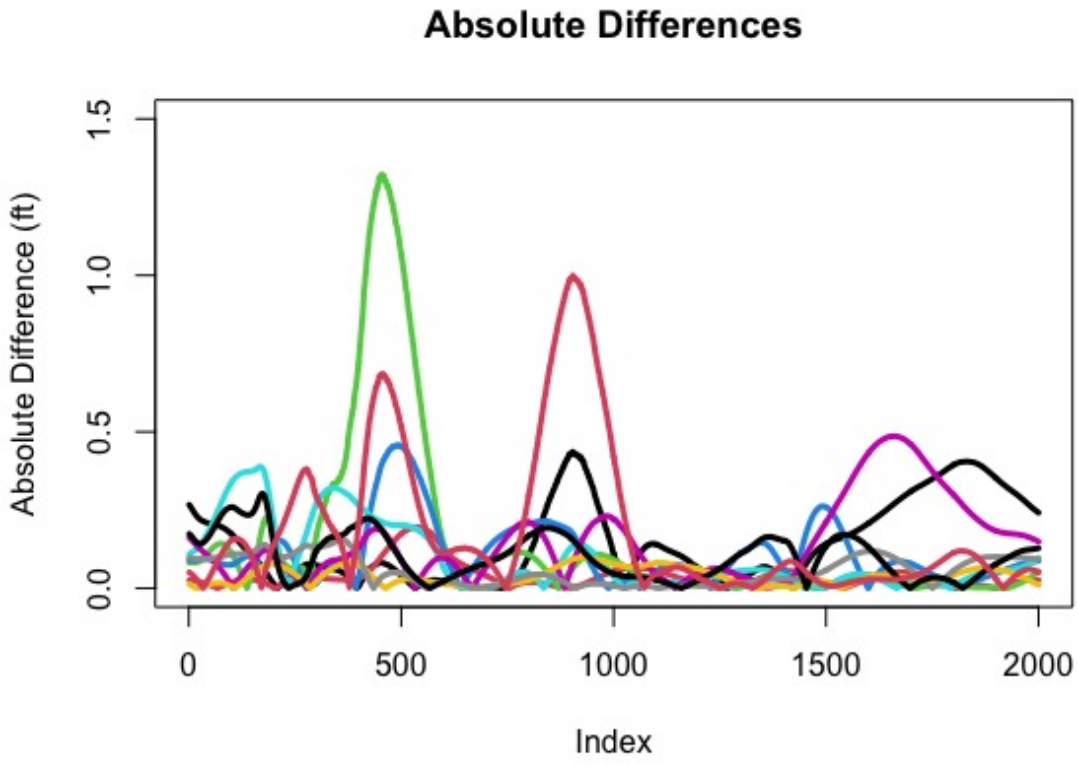


Figure 4.10. All 10 absolute difference curves between the October 2015 reconstruction using all 10 observations and when each observation was removed. The green curve is from the March 2003 event and the pronounced red curve is from the February 2020 event.

After finding the absolute differences, we calculated the percentiles of each vector of absolute difference independently for each curve as plotted in Figure 4.11. This identified the observations whose influence persisted over a long duration.

At the lower percentiles, Figure 4.11 shows that there are no pronounced differences across curves; however, the red curve representing the February 2020 flood event begins a noticeable increase around its 80th percentile, and the green curve depicting the March 2003 event begins a sharp increase shortly thereafter. The area under each curve indicates which of these two observations has the most overall impact on the October 2015 Cedar Creek reconstruction based on the magnitude and duration of the difference. Table 4.3 provides these AIP values for each flood event.

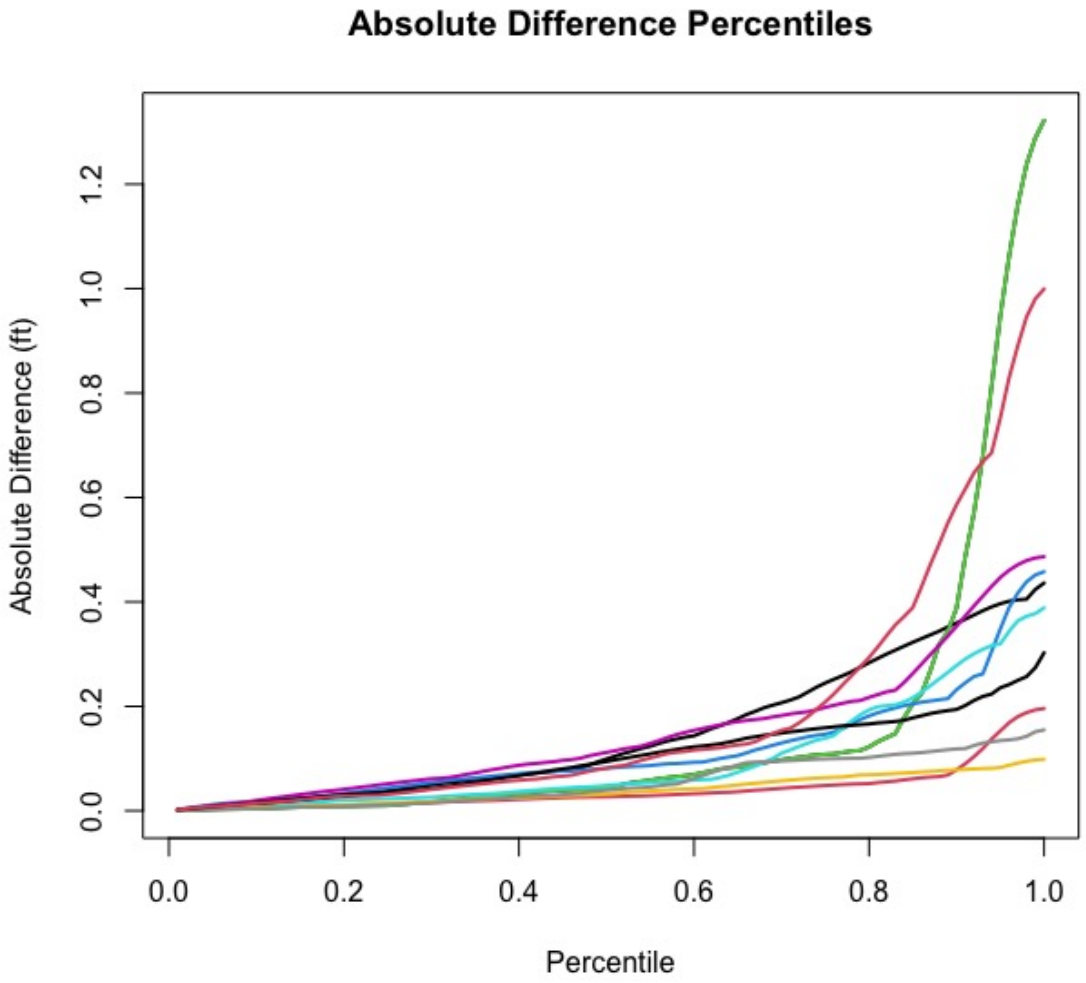


Figure 4.11. All 10 observation's percentiles of absolute differences between the October 2015 reconstruction using all 10 observations and when each observation was removed. The green curve is still the March 2003 event and the pronounced red curve the February 2020 event.

These results show that the February 2020 flood event was the most influential when we account for the duration of the influence. Therefore, we applied our weighted bootstrap method to approximate a null distribution of AIP within the context of the river stage data, to determine whether this observation was significantly influential on the reconstruction..

The bootstrapped percentiles given in Table 4.4 indicate that in terms of AIP, none of the prior flood events were significantly influential on the reconstruction of the Oc-

Table 4.3. The area under each curve in Figure 4.11 (AIP).

AIP Analysis	
Event	AIP
August 1995	14.615
February 1998	3.909
March 2003	14.668
May 2003	11.146
September 2004	9.475
March 2007	14.689
February 2010	3.868
May 2013	5.495
November 2018	10.201
February 2020	**18.200**

Table 4.4. AIP percentiles from each 5000 bootstrapped river stage observations along with the observed maximum of each metric in the river stage data context. Note that $\alpha = 0.5$ is most appropriate to use given the small sample size $N = 10$.

Bootstrap approximated null distribution's percentiles for AIP		
Percentile	$\alpha = 0$	$\alpha = 0.5$
90%	18.168	19.668
95%	22.949	25.282
99%	33.374	36.747
Max Obs. 18.2 (Feb. 2020)		

tober 2015 Cedar Creek stage. These results have several explanations. The first reason could be that Pittman, Hitchcock, and Grego, 2021 intentionally aligned the predictor curves to best resemble the out-of-sample October 2015 Congaree River curve, specifically to ensure that none of the 10 events carried too much weight and the reconstruction was accurate. These results suggest that the landmark alignment method was successful. Additionally, the simulation study in the previous section indicated that when $N = 10$, it takes a very influential observation to surpass the 95th percentile, since the average power was around 0.60 when λ was far from 1, so it is possible that the insignificant influence measure was also a result of the small sample size.

If one of the observations was flagged as significantly influential using one or both of the aforementioned measures, a natural goal would be to identify which portion of that functional observation had the most impact. Merely looking at the plot of just 2 reconstructed curves may not be enough since it is unclear how much of a difference in the raw context is a lot. We answered this question by taking the raw differences (or absolute difference) between $\hat{Y}^{new}(t_j)$ and $\hat{Y}^{new}_{(i)}(t_j)$, for $i = 1,\dots N$ and $t_1,\dots,t_n$. This produced $N \times n$ differences and indicated the specific portions of each observation that had the most impact on the reconstruction. After standardizing the $N \times n$ differences (each observation having n difference), we plot them for each observation and identify the portion of the functional observation that is large (> 2), specifically focusing on the observations that were flagged using the Δ or AIP influence measures.

We propose a simple method of judging the standardized pointwise differences between the full reconstruction $\hat{Y}^{new}(t)$ and the same reconstruction with each event removed, $\hat{Y}^{new}_{(i)}(t)$ shown in Figure 4.12. Using these river stage data, the standardized measures given in Figure 4.12 are not particularly needed; however, if it were not as clear as to see why the March 2003 event had the largest Δ, looking at the standardized differences shows how large each difference is relative to the values of all other differences. Typically, values beyond the dashed lines at ± 2 are notable. The green curve depicting the March 2003 event has a pronounced difference in the reconstruction at index = 500, indicating that the large Δ and drastic spike in Figure 4.11 is heavily influenced by this portion of the March 2003 flood event. The red curve (February 2020 event) yields the largest negative difference at time index 500 and the largest positive difference at index 1000. This suggests that its large Δ is a result of influence at multiple portions of the time domain rather than only one spike like the March 2003 event.

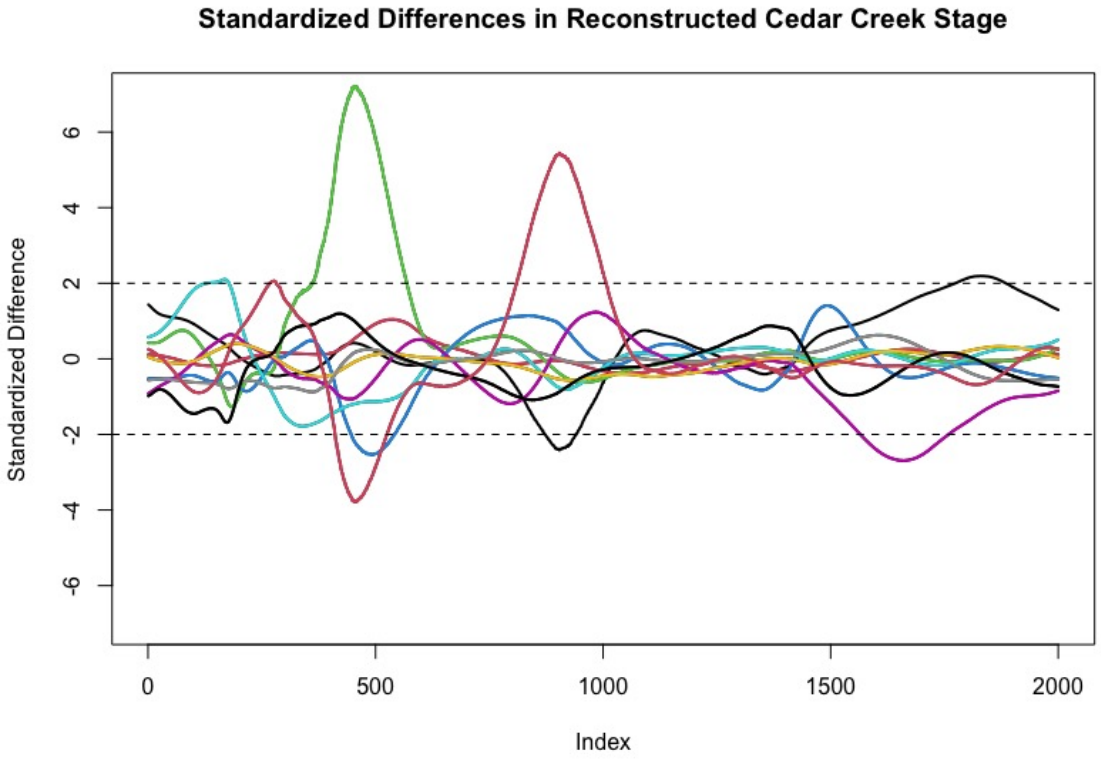


Figure 4.12. The standardized difference between the reconstructed October 2015 Cedar Creek stages with all 10 observations and the reconstruction with observation i withheld with dashed lines at ± 2.

4.6 Application: Air and Water Temperature

4.6.1 Analysis of influence using Δ

We calculated our influence measures and applied our method on an air and water temperature dataset coming from 35 weather stations along the US coastline in 2020. We obtained the data from the National Data Buoy Center (National Oceanic and Atmospheric Administration, 2021). These 35 stations are located all around the United States coastline, including East Coast, West Coast, Gulf of Mexico, the Alaskan coastline, and Hawaii (map of specific locations provided in the appendix).

Each station's data contained roughly 87,600 temperature measurements in 6-minute intervals across 2020. To be eligible for inclusion, the station's air and water tempera-

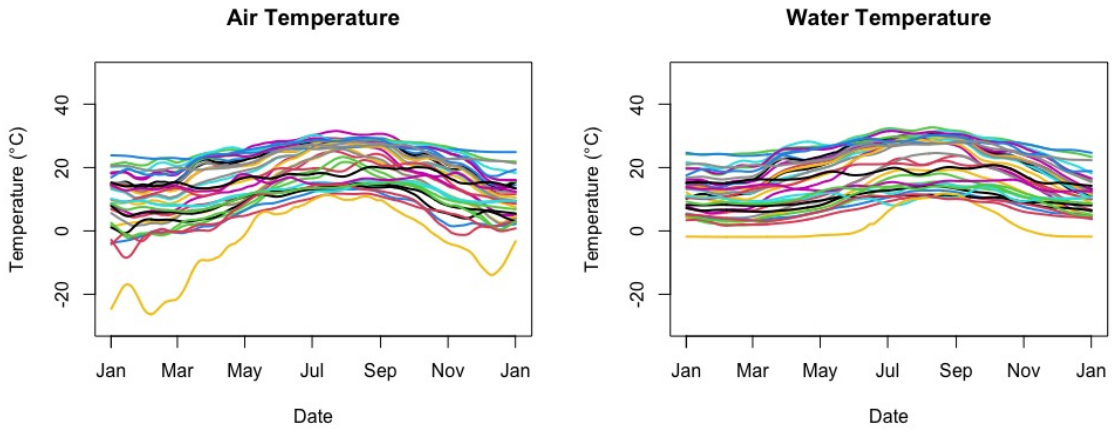


Figure 4.13. All 35 smoothed air (left) and water (right) temperatures used in the model.

tures had to be at least 90% non-missing. Then we preprocessed the data, which included linear interpolation to fill in any missing record and pre-smoothing each observation to remove the day-to-day variability and focus on the yearly trends. Lastly, without disrupting the underlying relationship between the air and water temperature curves, we resized the length of each smoothed discretized curve to 1000 equally-spaced observations across the year to speed up the functional calculations. Figure 4.13 provides all 35 air and water temperatures used in the model which we represented using 21 B-spline basis functions. Note that the low gold curve in Figure 4.13 is from Red Dog Dock, Alaska, which is the most northern station used in the sample.

We then used air temperature to predict concurrent water temperature throughout the calendar year. This analysis differed from the previous river stage example because in this study, we had 35 complete air and water temperature functional observations, and we introduced 5 out-of-sample observations that had available air temperature data but incomplete or completely missing water temperatures. Using the 35 complete air and water temperature observations, we estimated $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ in the concurrent model (Equation (4.1)), predicted the five missing stations' water temperature functions using each the known air temperature functions at those stations, and calculated each of the 35 observations' $\Delta_i, i = 1,\ldots,35$ for each of the five out-of-sample stations. Note

Table 4.5. Δ_i for all 35 observations for all five locations with missing water temperatures along with the average for each of them.

Column	Location	Adak Island AK	Kahului HI	Prudhoe Bay AK	Rockport TX	Ship John Shoal NJ	Average
1	Amerada Pass, LA	1.12	5.07	5.25	4.42	1.64	3.50
2	Atlantic City, NJ	1.85	1.05	3.58	1.29	1.74	1.90
3	Bar Harbor, ME	4.38	1.54	7.66	0.99	2.26	3.37
4	Bay Waveland Yacht Club, MS	0.92	2.40	2.49	2.28	0.94	1.81
5	Beaufort, NC	0.45	1.09	0.80	1.20	0.85	0.88
6	Bishops Head, MD	2.02	1.03	3.78	1.37	1.89	2.02
7	Boston, MA	3.79	2.83	7.60	3.32	3.87	4.28
8	Bridgeport CT	2.84	1.08	5.78	1.21	2.24	2.63
9	Calcasieu Pass, LA	0.45	1.33	1.48	1.28	0.72	1.05
10	Charleston Cooper River Entrance, SC	0.36	1.02	0.92	1.04	0.63	0.79
11	Clearwater Beach, FL	1.49	3.88	4.87	3.96	1.96	3.23
12	Cordova, AK	3.80	0.82	6.44	0.69	1.36	2.62
13	Crescent City, CA	4.93	0.60	6.37	0.82	1.38	2.82
14	Fernandina Beach, FL	3.10	5.16	7.71	5.74	2.06	4.76
15	Fort Pulaski, GA	0.39	1.10	0.96	1.09	0.64	0.84
16	Johnny Mercer Pier Wrightsville Beach, NC	0.69	1.29	0.87	1.21	0.91	0.99
17	Ketchikan, AK	3.62	0.38	5.20	0.56	1.14	2.18
18	King Cove, AK	4.75	0.80	8.51	1.01	1.90	3.39
19	Lake Worth Pier, FL	0.90	6.81	10.99	4.57	0.84	4.82
20	Mokuoloe, HI	0.84	5.22	8.97	3.56	0.78	3.87
21	Naples, FL	0.83	3.30	5.02	2.68	1.00	2.57
22	Old Port Tampa, FL	0.97	3.31	4.30	3.06	1.42	2.62
23	Oregon Inlet Marina, NC	0.71	0.88	0.82	0.96	0.86	0.85
24	Panama City Beach, FL	0.74	2.79	1.98	2.38	1.27	1.83
25	Port Angeles, WA	4.24	0.44	5.92	0.46	1.44	2.50
26	Port Chicago, CA	1.92	1.38	2.20	1.22	1.21	1.59
27	Portland, ME	2.80	1.39	4.74	1.23	1.81	2.39
28	Port Isabel, TX	0.42	1.87	2.60	1.40	0.56	1.37
29	Port Orford, OR	7.15	1.22	9.26	0.99	2.09	4.14
30	Port San Luis, CA	2.37	2.70	2.69	2.42	1.86	2.41
31	Red Dog Dock, AK	**26.88**	**24.66**	**101.58**	**9.82**	**16.90**	**35.97**
32	Sand Island, Midway Islands	0.31	2.76	4.08	1.95	0.56	1.93
33	Santa Monica Pier, CA	2.14	1.41	2.66	1.08	1.02	1.66
34	Skagway, AK	5.12	2.31	11.71	0.92	2.69	4.55
35	Westport, WA	4.18	0.45	5.65	0.63	1.41	2.47

that using the leave-one-out method to predict the water temperature for a withheld observation successfully captured the truth. Using $\hat{Y}^{new}(t)$ and $\hat{Y}^{new}_{(i)}(t)$ we calculated Δ_i for $i = 1,\ldots,35$ for each of the five target observations. The complete results are given in Table 4.5.

There are several main conclusions from Table 4.5. The first is that Red Dog Dock has the highest Δ for each of the five out-of-sample observations and was so much larger than the rest that there is little doubt that this observation has massive influence on the

predicted values. This station corresponds to the gold curve in Figure 4.13 which has a visually low air temperature during the winter months compared to the other 34 locations, whereas the corresponding water temperature is only slightly lower than the rest. Note that which specific out-of-sample observation was considered played a large role in the magnitude of Δ, indicating that using a general threshold for judging Δ values is not appropriate. Like Red Dog Dock, Prudhoe Bay in Alaska had a low air temperature with a moderate water temperature, and the values of the Δ_i were large for all 35 locations corresponding to the Prudhoe Bay prediction. Estimating a null distribution via a weighted bootstrap is a better method of determining which observations have a significant influence on predicting an outside observation.

4.6.2 Bootstrap distribution of Δ in temperature data

With a sample size $N = 35$, using either $\alpha = 0$ or $\alpha = 0.5$ in our weighted bootstrap method to approximate a null distribution may be appropriate; however, given the magnitude of Δ for the Red Dog Dock station for each target observation, it was best to focus on the percentiles using $\alpha = 0.5$ since it dampened the effect of that observation on the approximate null distribution. This made it easier to determine if any other stations had a significant impact on the prediction.

After applying the weighted bootstrap method (with $B = 100$) independently for each of the five target external observations, we estimated a null distribution of Δ. Table 4.6 gives the resulting 90th, 95th, and 99th percentiles.

The clearest conclusion from Table 4.6 is that regardless of the target observation, Red Dog Dock (observation 31) had a significantly large Δ, falling above the 99th percentile for all five external target observations. Note that different functional observations in the sample provide the second largest observed Δ across the five target observations, but in some cases, that second observation is also highly influential on the predicted water temperature curve of the out-of-sample location. In the Ship John Shoal

Table 4.6. Percentiles from an approximate null distribution of Δ for each of the five target observations and the highest and 2nd highest observed Δ for each observation.

		Adak Island	**Kahului**	**Prudhoe Bay**	**Rockport**	**Ship John Shoal**
$\alpha = 0$	90%	5.217	4.908	9.626	3.939	2.522
	95%	7.021	6.169	12.432	4.912	3.773
	99%	22.406	15.239	88.636	8.073	8.499
$\alpha = 0.5$	90%	7.356	4.799	12.635	3.742	2.520
	95%	10.43	6.867	16.676	4.99	3.564
	99%	20.90	10.230	30.643	9.663	7.082
	Max Obs.	**26.88 (31)**	**24.66 (31)**	**101.58 (31)**	**9.82 (31)**	**16.90 (31)**
	2nd Highest	7.15 (29)	6.81 (19)	11.7 (34)	5.7 (14)	3.87 (7)

location, the Boston observation's Δ was above the 95th percentile, indicating that it had a significant amount of influence on the Ship John Shoal prediction. In the Kahului location, the Lake Worth Pier observation's Δ was slightly below the 95th percentile, indicating that it had a moderate influence on the Kahului prediction. In the Rockport location, the Fernandina Beach observation's Δ was well above the 95th percentile, showing that it has significant influence on the water temperature prediction at the Rockport station.

4.6.3 Analysis of influence using AIP

Lastly, we calculated AIP_i of each of the 35 sampled observations for each of the five target. We first calculated the predicted water temperature for each of the five target locations independently, using all 35 observations, and then with each of the 35 observations sequentially removed. Then we took the percentiles of the absolute differences as described in Section 4.1.2. The plot of the percentiles of absolute differences for one of the out-of-sample locations, Adak Island, is given in Figure 4.14.

The other four target observations' absolute difference percentiles plots are very similar. The area under each curve yields the AIC of each of the 35 observations for each target observation (provided in Table 4.7).

As with the Δ influence measure, Red Dog Dock had the largest AIP for all five target observations. This suggests that the observation has a large impact on the water

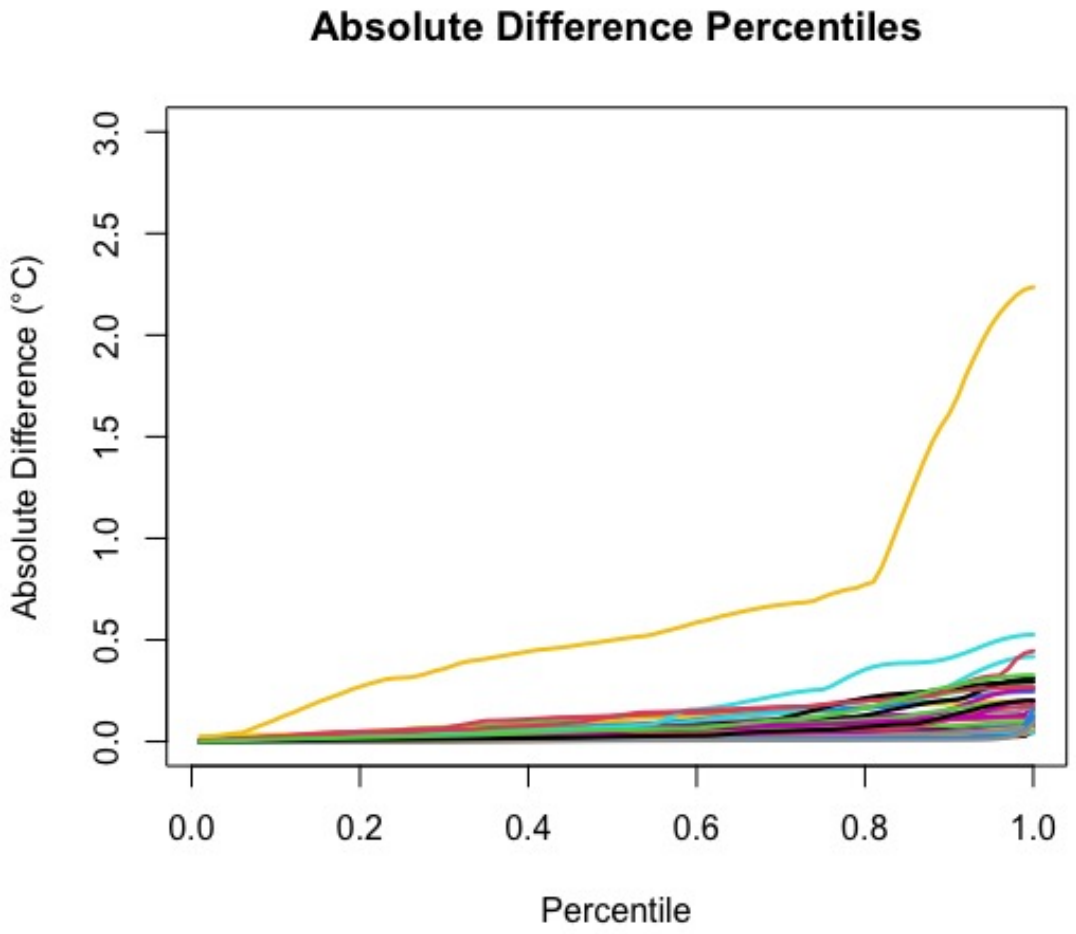


Figure 4.14. All 35 observations' percentiles of absolute differences between the Adak Island water temperature prediction using all 35 observations and with each observation removed in turn. The gold curve is the Red Dog Dock observation's results.

temperature prediction at each of the five target locations. We performed our weighted bootstrap algorithm to these data to confirm the formal significance of its influence and investigate the potential influence of other observations.

4.6.4 Approximating null distribution of AIP with temperature data

We carried out the weighted bootstrap method independently to approximate the null distribution of AIP for each target observation. Given the large magnitude of Red Dog Dock's AIP compared to the others', it was most appropriate to use $\alpha = 0.5$. Table 4.8

Table 4.7. AIP influence measure for all 35 stations for all five target observations with the largest measures in bold.

Location	Adak Island AK	Kahului HI	Prudhoe Bay AK	Rockport TX	Ship John Shoal NJ	Average
Amerada Pass, LA	2.691	12.507	12.393	11.197	4.337	8.625
Atlantic City, NJ	4.580	2.357	7.956	2.959	4.39	4.450
Bar Harbor, ME	11.571	3.330	20.557	1.703	6.025	8.637
Bay Waveland Yacht Club, MS	2.006	6.069	6.494	5.960	2.468	4.599
Beaufort, NC	1.215	3.108	1.776	3.336	2.387	2.364
Bishops Head, MD	4.840	2.837	8.434	3.798	5.017	4.985
Boston, MA	10.576	5.396	19.072	6.755	10.575	10.475
Bridgeport CT	7.779	2.716	14.442	3.239	6.390	6.913
Calcasieu Pass, LA	1.150	3.387	3.387	3.379	1.710	2.603
Charleston, Cooper River Entrance, SC	0.855	2.686	2.132	2.733	1.578	1.997
Clearwater Beach, FL	3.759	11.894	14.430	11.933	5.030	9.409
Cordova, AK	9.749	1.781	16.936	1.626	3.358	6.690
Crescent City, CA	11.086	1.528	14.973	2.079	3.517	6.637
Fernandina Beach, FL	7.099	14.335	20.354	15.423	5.297	12.502
Fort Pulaski, GA	0.890	2.655	2.011	2.635	1.566	1.951
Johnny Mercer Pier, Wrightsville Beach, NC	1.339	2.969	1.346	2.878	2.096	2.126
Ketchikan, AK	8.490	0.880	13.351	1.188	2.923	5.366
King Cove, AK	12.829	2.065	23.300	2.423	4.689	9.061
Lake Worth Pier, FL	2.158	13.924	21.453	9.937	1.955	9.885
Mokuoloe, HI	2.122	11.666	17.873	8.847	2.049	8.512
Naples, FL	2.270	8.629	12.354	7.648	2.498	6.680
Old Port Tampa, FL	2.571	9.698	11.877	9.225	3.741	7.423
Oregon Inlet Marina, NC	1.573	2.234	1.934	2.424	2.176	2.068
Panama City Beach, FL	1.766	7.395	5.315	6.714	3.521	4.942
Port Angeles, WA	9.653	0.940	14.168	1.058	3.471	5.858
Port Chicago, CA	4.385	3.777	4.588	3.382	3.346	3.896
Portland, ME	7.217	2.675	12.470	1.941	4.565	5.773
Port Isabel, TX	0.973	4.397	5.524	3.550	1.436	3.176
Port Orford, OR	15.891	3.095	21.146	2.499	5.366	9.599
Port San Luis, CA	5.801	5.838	6.074	5.118	4.287	5.424
Red Dog Dock, AK	**64.813**	**43.227**	**190.866**	**22.505**	**32.393**	**70.761**
Sand Is., Midway Islands	0.550	5.876	7.659	4.566	1.460	4.022
Santa Monica Pier	4.433	3.696	5.674	2.817	2.704	3.865
Skagway, AK	12.935	3.829	25.831	1.946	5.663	10.041
Westport, WA	9.240	1.091	13.133	1.461	3.386	5.662

provides the approximate null distribution's percentiles using $\alpha = 0$ and $\alpha = 0.5$, respectively, along with the largest and second largest observed AIP.

When using $\alpha = 0.5$, for each target observation, Red Dog Dock's AIP was well above the 95th percentile, and for every location except Rockport it was well above the 99th percentile. This indicates that the Red Dog Dock observation was significantly influential on water temperature prediction regardless of the target observation. Moreover, since it was substantially above the 99th percentile for most of the target locations, this observation should be investigated further for possible removal, as it may have incorrectly distorted the prediction of water temperatures. When we dampened the effect of

Table 4.8. Approximate null distribution percentiles of AIP for each target observation for $\alpha = 0$ and $\alpha = 0.5$ along with the largest (Red Dog Dock) and 2nd largest observed AIP given.

		Adak Island	**Kahului**	**Prudhoe Bay**	**Rockport**	**Ship John Shoal**
$\alpha = 0$	90%	12.866	11.987	24.210	10.333	6.377
	95%	17.620	15.346	31.015	13.004	9.36
	99%	50.205	24.650	109.713	19.895	25.681
$\alpha = 0.5$	90%	18.410	11.313	30.349	9.922	6.293
	95%	25.979	16.06	40.190	14.57	8.396
	99%	44.651	26.519	80.285	24.423	16.181
	Max Obs.	**64.81**	**43.226**	**190.866**	**22.505**	**32.393**
	2nd Highest	15.89 (29)	14.33 (14)	25.83 (34)	15.42 (14)	10.57 (7)

the Red Dog Dock location by using $\alpha = 0.5$, the Fernandina Beach AIP was above the 90th percentile for predicting Kahului and the 95th percentile for predicting Rockport, indicating that it also had significant influence on the prediction of water temperatures at these locations. The Boston station's AIP was above the 95th percentile for predicting the Ship John Shoal water temperature, suggesting that it (along with Red Dog Dock) was influential.

4.7 Conclusion

One application of functional data analysis is to use a concurrent functional relationship between sets of observations to predict an out-of-sample (target) observation's response. Our new measures of influence, Δ and AIP, offer a pragmatic way to spot functional observations that have a large impact on specific predictions of target response curves. Additionally, simulation shows that our weighted bootstrapping approach performs well in identifying whether the most influential observations truly have a large impact on the prediction. In both the river stage and air and water temperature examples, we sensibly identify certain observations as more influential than the rest, and then the bootstrap method confirms their influence is significantly large, further illustrating that our method satisfactorily identifies functional observations in the concurrent model that influence the prediction of an external response curve.

CHAPTER 5

CLOSING REMARKS

The inspiration for this study began on October 3, 2015, when massive rainfall in Columbia, South Carolina resulted in devastating damages. The Cedar Creek gage located in Congaree National Park was damaged during the storm. Our initial goal of the project was to determine how high Cedar Creek was throughout the duration of the flood event when its readings went offline.

During flood events, the Congaree River stage and the Cedar Creek stage have a strong functional relationship. Therefore, we identified 10 historic flood events and applied our innovative landmark aligned data selection method to determine the optimal start and end points of these historic events to best align them with our target October 2015 flood event, with emphasis on aligning the events at higher river stages. Using these 10 sets of functional observations, we utilized the concurrent functional regression model to estimate the relationship between the Congaree River and Cedar Creek at each time within of a flood event. We used this model, along with the observed Congaree River stage during the October 2015 flood, to reconstruct the Cedar Creek stage at the corresponding time. Our conclusion was that during the October 2015 flood event Cedar Creek reached a record height of 17.59 feet, which is appropriate given that the Congaree River also reached record heights during this flood. Moreover, our reconstructed Cedar Creek stage resembles the small portion of available Cedar Creek heights and our 95% prediction interval is narrow, indicating that our reconstruction was reliable.

Our next goal was to evaluate the influence each of the 10 events had on the functional model and on the reconstructed river stage. Therefore, we extended several mea-

sures used in ordinary regression ($DFBETAS$, $DFFITS$, and Cook's Distance) to the functional framework. The ordinary regression influence measures produce a single number for each observation. Given that we had another dimension, time, to take into account, we calculated each ordinary measure pointwise across the functional event to produce an influence measure for each of the N observations that was a function of t, e.g. $DFBETAS_{p,i}(t)$. Then, we took the mean of the absolute influence measure across the functional observation, resulting for each observation in a single number for each influence measure. Each measure gauged the influence a functional observation has on the regression equation. Since we had no general guidelines about what constitutes a "large" value for these new measures, we proposed a new weighted bootstrapping with perturbations method to approximate a null distribution of each measure given a dataset. The observed influence measures could then be compared to the percentiles of the distribution to determine whether they were significantly influential.

Since our initial question predominantly pertained to reconstructing/predicting an out-of-sample target flood event, we also created measures of the impact each observation had on an external functional prediction of the concurrent model. Therefore, we introduced two new measures, Δ and AIP, that each yield a single influence measure for each observation. These measures fit the concurrent model and predict the out-of-sample observation with the full data and without each functional observation and then compare the differences. The Δ_i influence measure is the L_2 distance between the full prediction and the prediction with the ith event withheld, representing a total difference across the functional observation. On the other hand, AIP takes the area under the curve of percentiles of the discretized differences. It accounts for both the magnitude of the differences and the length of the regions where these differences are notably large. These measures can be used together to accurately assess which observations have the greatest impact on an out-of-sample prediction. Again, since there is no general cutoff defining what constitutes a "large" measure, we used a weighted

bootstrapping approach to approximate a null distribution for each measure relative to a particular out-of-sample target observation. Our observed influence measures can then be compared to this null distribution to determine significance.

Many aspects of this research are relevant to other functional regression applications. We provide several new functional influence measures and a weighted bootstrapping method to assess significance levels. However, future studies may be able to enhance these measures with an alternative approach for transforming the pointwise measures of influence into a single measure, aside from calculating the average across the observation. Improvement of the bootstrap sampling method to enhance the approximation of the null distribution of each measure may be possible. Moreover, in addition to creating new functional influence measures, we presented a selection and alignment method to objectively determine the domains defining our sample of functional data. We expanded upon the usual use of the concurrent model and used it for prediction. Then we presented numerous functional influence measures to assess the usability of the collected data and determine which functional observations had the most impact on the regression model and out-of-sample prediction.

BIBLIOGRAPHY

Baíllo, A. and A. Grané (2009). "Local linear regression for functional predictor and scalar response". In: *Journal of Multivariate Analysis* 100, pp. 102–111.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken, New Jersey: Wiley Series in Probability and Mathematical Statistics.

Borchers, H. W. (2019). *pracma: Practical Numerical Math Functions*. R package version 2.2.9. URL: https://CRAN.R-project.org/package=pracma.

Burris, R. (Dec. 1, 2015). "SC floods' damage: $12 billion, economists say". In: *The State*. URL: https://www.thestate.com/news/local/article47471060.html (visited on 06/16/2019).

Chebana, F., S. Dabo-Niang, and T. B. M. J. Ouarda (2012). "Exploratory Functional Flood Frequency Analysis and Outlier Detection". In: *Water Resources Research* 48(4), pp. 43–60.

Chen, G., C. Huang, and J. Lin (2014). "Statistical diagnostics for functional linear regression models with Gaussian process errors". In: *Communication on Applied Mathematics and Computation* 28.1, pp. 118–126. ISSN: 0167-9473. DOI: 10.3969/j.issn.1006-6330.2014.01.015. URL: https://www.researchgate.net/publication/260750220_Statistical_diagnostics_for_functional_linear_regression_models_with_Gaussian_process_errors.

Chiou, J. and H. Müller (2007). "Diagnostics for functional regression via residual processes". In: *Computational Statistics and Data Analysis* 51.10, pp. 4849–4863. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2006.07.042. URL: https://www.sciencedirect.com/science/article/pii/S0167947306002465.

Cook, R. D. (1977). "Detection of influential observation in linear regression". In: *Technometrics* 19.1, pp. 15–18. DOI: 10.1080/00401706.1977.10489493. URL: https://doi.org/10.1080/00401706.1977.10489493.

Craven, P. and G Wahba (1978/79). “Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation”. In: *Numerische Mathematik* 31, 377–404.

Das, D., K. S. Pasupathy, N. N. Haddad, M. S. Hallbeck, M. D. Zielinski, and M. Y. Sir (2019). “Improving Accuracy of Noninvasive Hemoglobin Monitors: A Functional Regression Model for Streaming SpHb Data”. In: *IEEE Transactions on Biomedical Engineering* 66, 759–767.

Febrero-Bande, M., P. Galeano, and W. González-Manteiga (2010). “Measures of influence for the functional linear model with scalar response”. In: *Journal of Multivariate Analysis* 101.2, pp. 327–339. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2008.12.011. URL: https://www.sciencedirect.com/science/article/pii/S0047259X08002765.

Ferraty, F., A. Rabhi, and P. Vieu (2005). “Conditional Quantiles for Dependent Functional Data with Application to the Climatic “El Niño" Phenomenon”. In: *Sankhyā: The Indian Journal of Statistics (2003-2007)* 67.2, pp. 378–398. ISSN: 09727671. URL: http://www.jstor.org/stable/25053438.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. Boca Raton, Florida: CRC Press.

Kutner, M. H., J. C. Nachtsheim, J. Neter, and W. Li (2005). *Applied Linear Statistical Models*. 5th. New York: McGraw-Hill Irwin.

Masselot, P., S. Dabo-Niang, F. Chebana, and T. B. M. J. Ouarda (2016). “Streamflow Forecasting Using Functional Regression”. In: *Journal of Hydrology* 538, pp. 754–766.

McIlroy, D., R. Brownrigg, T. P. Minka, R. Bivand, and Alex Deckmyn (2020). *mapproj: Map Projections*. R package version 1.2.7. URL: https://CRAN.R-project.org/package=mapproj.

National Oceanic and Atmospheric Administration (2021). *National Data Buoy Center*. https://www.ndbc.noaa.gov/obs.shtml. (Visited on 10/26/2021).

National Park Service (2019). *Maps and Brochures*. URL: https://www.nps.gov/cong/planyourvisit/maps.htm (visited on 06/17/2019).

National Weather Service (2015). *Historic October 1st to 5th, 2015 South Carolina Flooding Event*. URL: https://www.weather.gov/cae/HistoricFloodingOct2015.html (visited on 06/17/2019).

National Weather Service and NOAA (2019). *National Weather Service Advanced Hydrologic Prediction Service*. URL: https://water.weather.gov/ahps2/hydrograph.php?wfo=cae&gage=gads1 (visited on 06/17/2019).

Olshen, R. A., E. N. Biden, M. P. Wyatt, and D. H. Sutherland (1989). "Gait Analysis and the Bootstrap". In: *The Annals of Statistics* 17(4), pp. 1419–1440.

Pittman, R. D., D. B. Hitchcock, and J. M. Grego (2021). "Concurrent functional regression to reconstruct river stage data during flood events". In: *Environmental and Ecological Statistics* 28, pp. 219–237.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ramsay, J. and B. W. Silverman (2005). *Functional Data Analysis*. New York: Springer.

Ramsay, J. O., S. Graves, and G. Hooker (2020). *fda: Functional Data Analysis*. R package version 5.1.5.1. URL: https://CRAN.R-project.org/package=fda.

Ramsay, J. O., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. New York: Springer.

Resnick, S. I. (1992). *Adventures in Stochastic Processes*. Boston: Birkhauser Verlag. ISBN: 0817635912.

Sauer, V. B. and D. P. Turnipseed (2010). *Stage measurement at gaging stations: U.S. Geological Survey Techniques and Methods book 3*. Reston, Virginia. Chap. A7.

Schafer, J., R. Opgen-Rhein, V. Zuber, M. Ahdesmaki, A. P. D. Silva, and K. Strimmer. (2017). *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*. R package version 1.6.9. URL: https://CRAN.R-project.org/package=corpcor.

Shen, Q. and H. Xu (2007). "Diagnostics for linear models with functional Responses". In: *Technometrics* 49.1, pp. 26–33. DOI: 10.1198/004017006000000444. URL: https://doi.org/10.1198/004017006000000444.

Uhlenbeck, G. E. and L. S. Ornstein (1930). "On the theory of the Brownian motion". In: *Phys. Rev.* 36 (5), pp. 823–841. DOI: 10.1103/PhysRev.36.823. URL: https://link.aps.org/doi/10.1103/PhysRev.36.823.

United States Geological Survey (2019a). *How Streamflow is Measured*. URL: https://www.usgs.gov/special-topic/water-science-school/science/how-streamflow-measured?qt-science_center_objects=0#qt-science_center_objects (visited on 06/17/2019).

United States Geological Survey (2019b). *Water Questions & Answers What does the term river stage mean?* URL: https://water.usgs.gov/edu/qa-measure-streamstage.html (visited on 06/17/2019).

United States Geological Survey (2020a). *USGS 02169625 Congaree River at Congaree NP near Gadsden, SC.* https://waterdata.usgs.gov/sc/nwis/uv?site_no=02169625. (Visited on 06/23/2020).

United States Geological Survey (2020b). *USGS 02169672 Cedar Creek at Congaree NP near Gadsden, SC.* https://waterdata.usgs.gov/sc/nwis/uv?site_no=02169672. (Visited on 06/23/2020).

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: http://www.stats.ox.ac.uk/pub/MASS4.

Wang, Y., H. Wang, D. Srinivasan, and Q. Hu (2019). "Robust Functional Regression for Wind Speed Forecasting Based on Sparse Bayesian Learning". In: *Renewable Energy* 132(C), pp. 43–60.

Zhang, J., M. K. Clayton, and P. A. Townsend (2011). "Functional Concurrent Linear Regression Model for Spatial Images". In: *Journal of Agricultural, Biological, and Environmental Statistics* 16, pp. 105–130.

APPENDIX A

A.1 CHAPTER 2 ADDITIONAL NOTES

A.1.1 AN ALTERNATE APPROACH: USING L_1 DISTANCE TO SELECT START AND END TIMES OF FLOOD EVENTS

The goal is to determine the optimal beginning and ending points of the ten complete raw flood events. Here, instead of using LAL_1 distance between the trimmed curve and the October 2015 target event, we use L_1 distance as we alternately remove one point from the beginning of the raw event and then from the end. Then we determine which of these "trims" is used is based on which produces a smaller L_1 distance between the trimmed curve and the target. The process of trimming one observation from the beginning or end of the raw flood event is repeated until the combination of starting and ending points that yields the smallest L_1 distance is found.

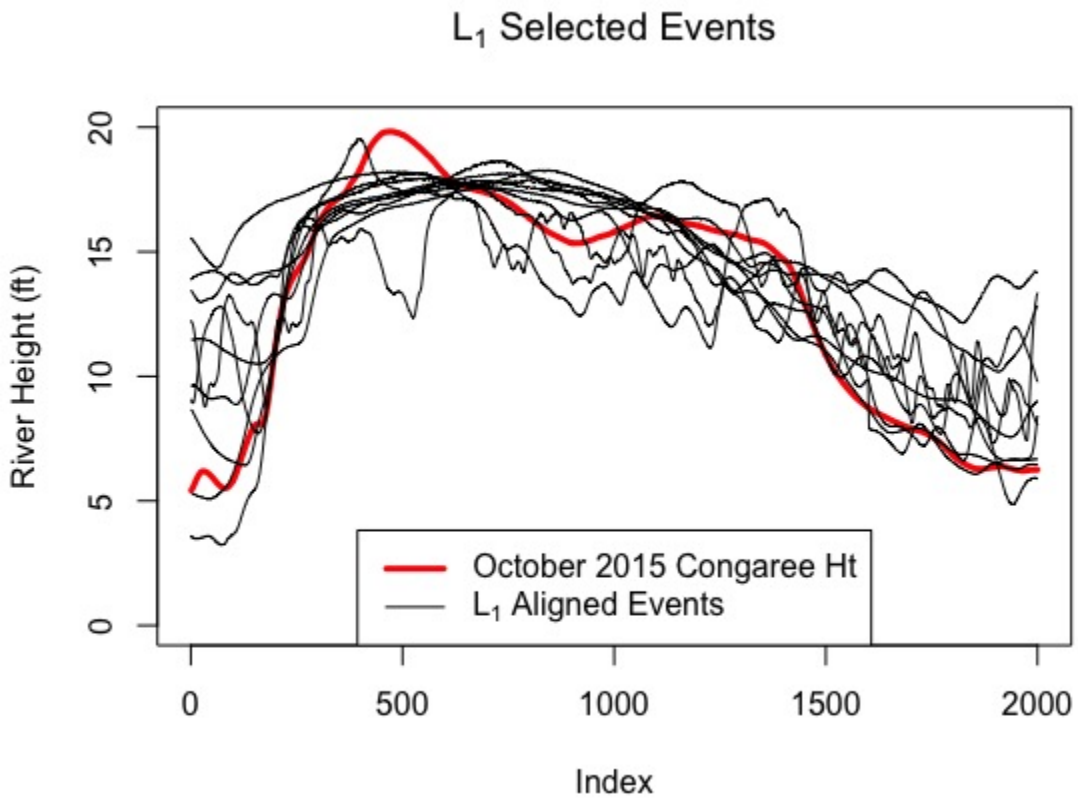


Figure A.1. All 10 L_1 distance selected Congaree River curves aligned with the target October 2015 Congaree River event.

We see here that there are some clear differences between the October 2015 Congaree curve and some of the other events. Comparing this to the LAL_1 selected events, the LAL_1 method does a better job overall of aligning the other Congaree River flood events to the highest portions of the October 2015 event.

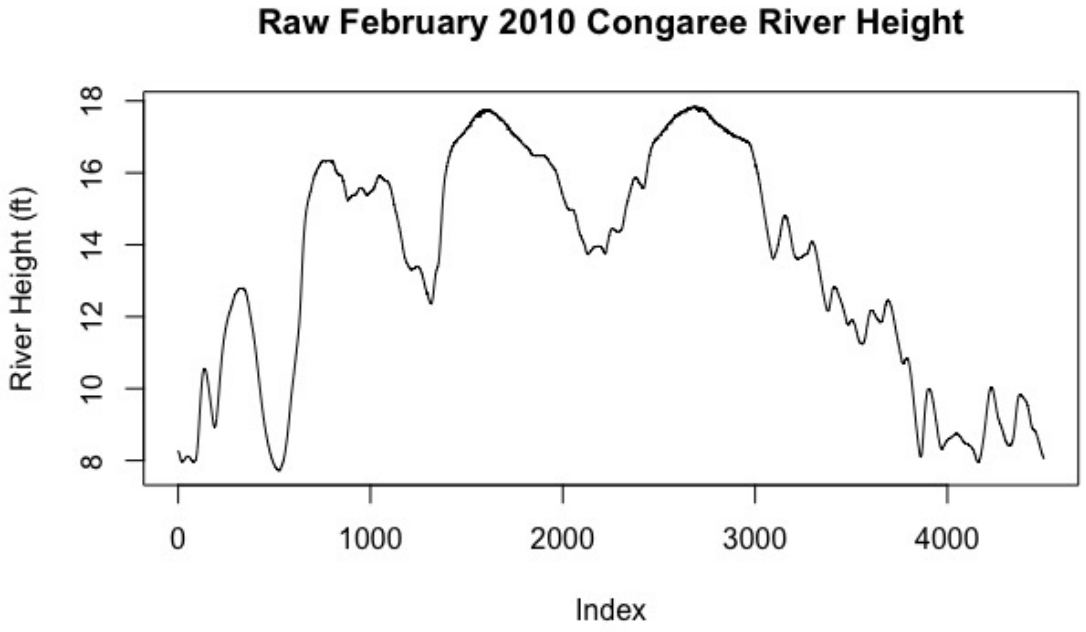


Figure A.2. Raw February 2010 Congaree River heights before selection.

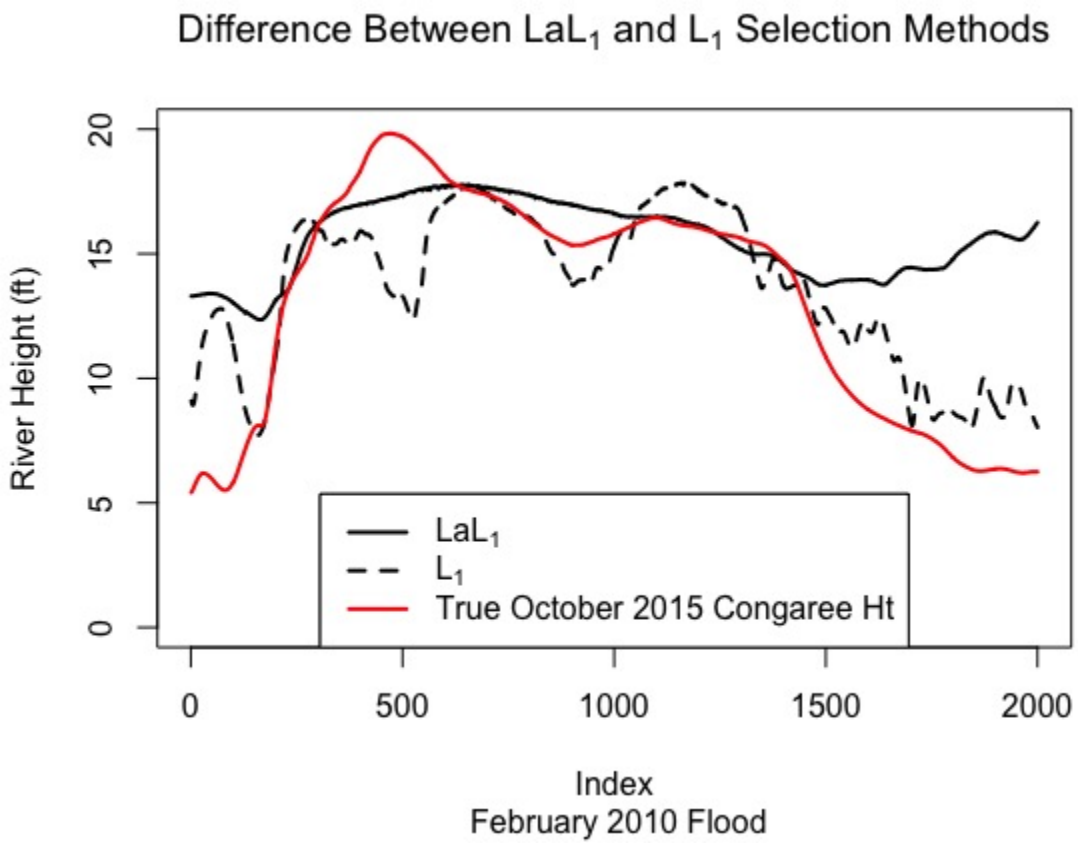


Figure A.3. Difference in the selected curve for February 2010 using the L_1 difference selection method vs the LAL_1 difference selection method.

The main reason for the big differences is that the raw data has multiple peaks. Most of the flood events have one main peak, so the LAL_1 selection results are very similar to the L_1 results. In this case, the L_1 distance selection method does a poor job of aligning the peak with that of the October 2015 event. This event is one of the key reasons why for this study we use LAL_1 instead of L_1. The LAL_1 distance selection method focuses on aligning the new curve with the peak of the October 2015 Congaree River curve, and any difference around the peak will be magnified and adjusted for much better than the L_1 distance selection method. While the L_1 distance selection method might do a better job of resembling the target curve overall, the focal point of this study is to discover how high Cedar Creek rose during the October 2015 flood, we are not as concerned with the Creek's stage once the water levels returned to normal; therefore, the LAL_1 distance

selection method can get away with the poor match post-peak, as long as it performs better at the peak.

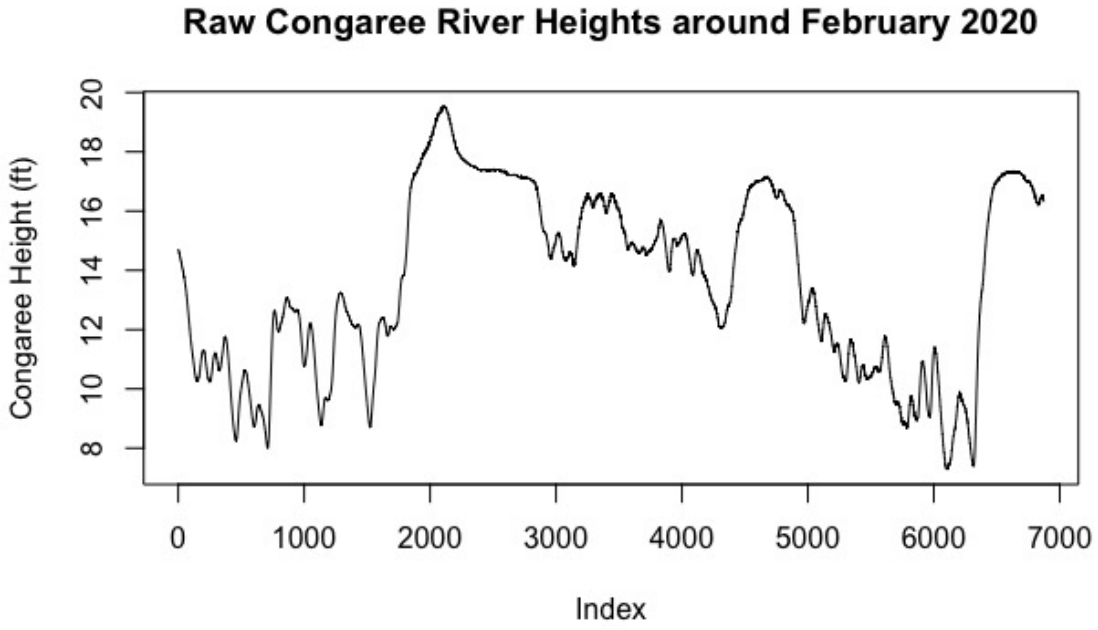


Figure A.4. Raw February 2020 Congaree River heights before selection.

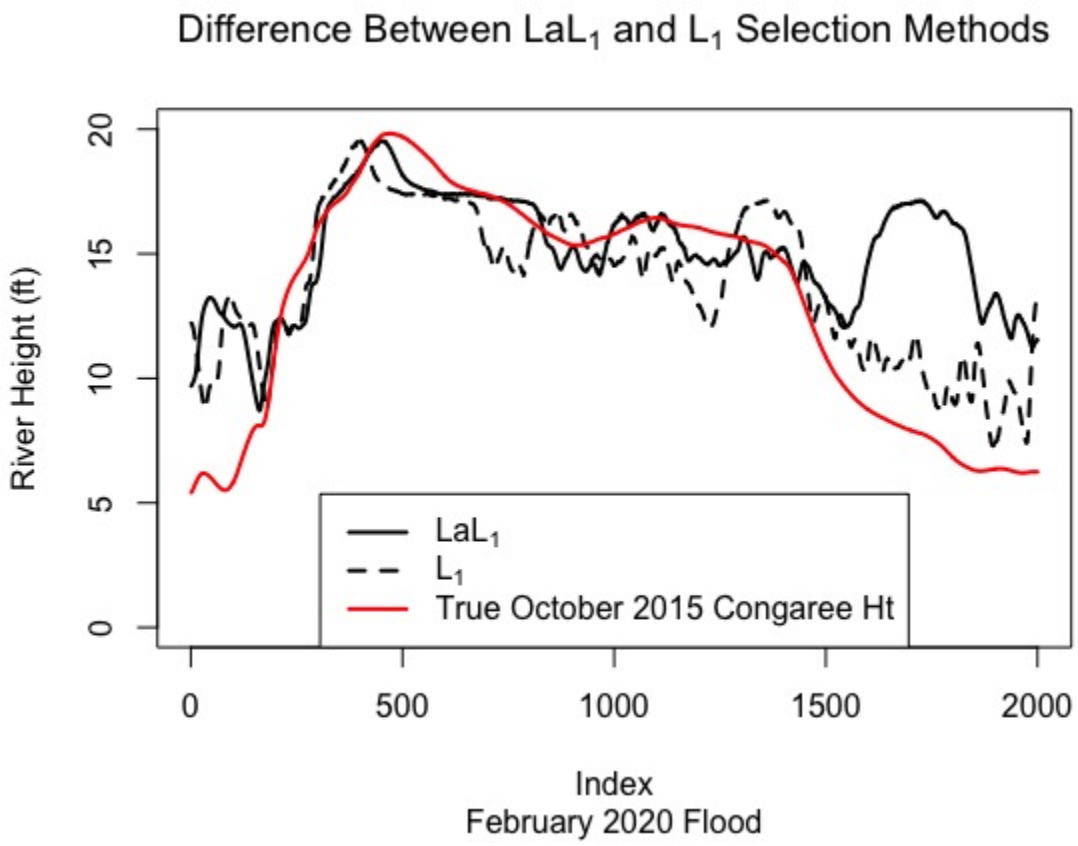


Figure A.5. Difference in the Selected Curve for February 2020 using the L_1 difference selection method vs the LAL_1 difference selection method.

That small difference at the peak between the LAL_1 and L_1 is very significant in reconstructing the targeted October 2015 Cedar Creek. Since LAL_1 selected data is so close to the true October 2015 Congaree curve when it is aligned, the model formed based on the flood events selected via LAL_1 better reflects the relationship between the Congaree River and Cedar Creek at this point of the flood event. This information is critical in getting an accurate reconstruction.

A.1.3 SELECTING THE SMOOTHING PARAMETER λ

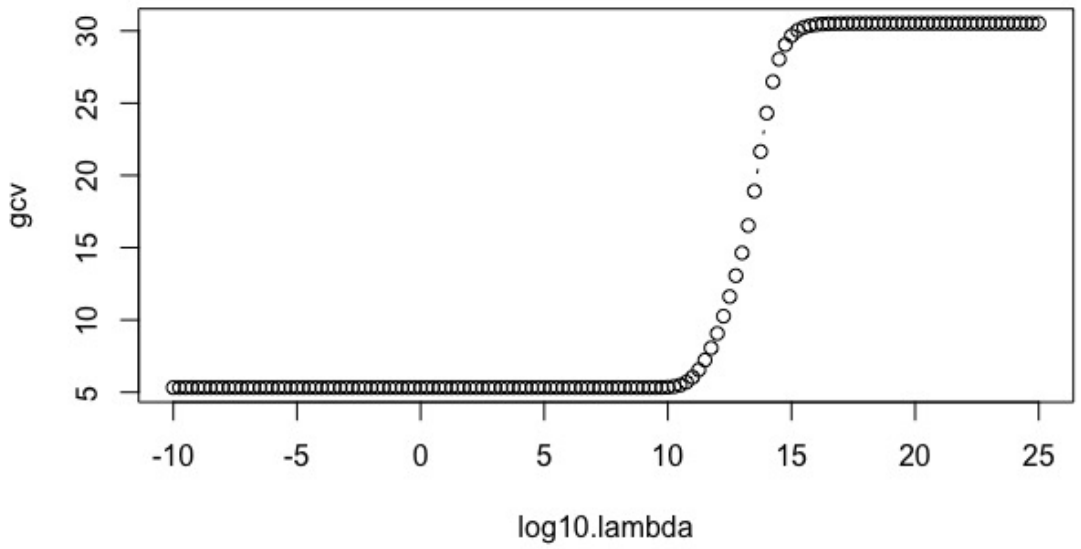


Figure A.6. Using GCV to select the optimal lambda to use to describe the river heights data.

Since the goal is to select the λ that minimizes GCV, there is a wide range of appropriate choices, ranging from 10^{-10} to 10^{10}. We chose to use 10^{-1} for simplicity. Note that choosing any λ in this range yields the same final results.

A.1.4 CODE FOR FINDING SMOOTHING PARAMETER

```
choosing_lambda<-function(dataX, dataY, nFourrierBasis = 15){
  require(fda)
n=nrow(dataX)

gaittime <- seq(1:n)
gaitrange <- c(0,n)
gaitfine = seq(0,n,1)
```

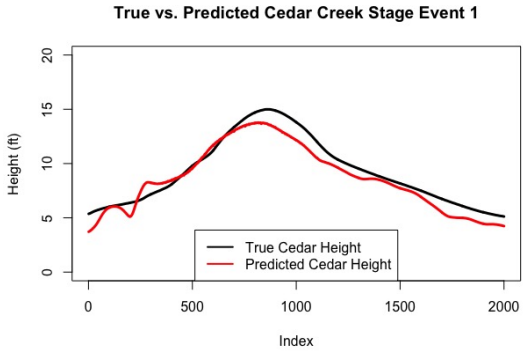
```
harmaccelLfd1220 <- vec2Lfd(c(0, (2*pi/n)^2, 0),
rangeval=gaitrange)
gaitbasis <- create.fourier.basis(gaitrange,
nbasis=nFourrierBasis) #original 15

mygaitExp <- array(NA, dim = c(n,ncol(dataX),2))
mygaitExp[1:n, ,] <- seq(1:n)
for(i in 1:ncol(dataX)){
mygaitExp[,i, 1] <- dataX[,i]
mygaitExp[,i, 2] <- dataY[,i]
}

#Begin here assuming the mygaitExp is complete.
#This part helps choose a negative lambda Shows that really
anything below 6 is fine
gaitLoglam = seq(-10,25,0.25)
nglam    = length(gaitLoglam)
gaitSmoothStats = array(NA, dim=c(nglam,
3),dimnames=list(gaitLoglam, c("log10.lambda", "df", "gcv") ) )
gaitSmoothStats[, 1] = gaitLoglam
for (ilam in 1:nglam) {
  gaitSmooth = smooth.basisPar(gaittime, mygaitExp, gaitbasis,
                               Lfdobj=harmaccelLfd1220,
                               lambda=10^gaitLoglam[ilam])
  gaitSmoothStats[ilam, "df"]  = gaitSmooth$df
  gaitSmoothStats[ilam, "gcv"] = sum(gaitSmooth$gcv)}
# note: gcv is a matrix in this case
```

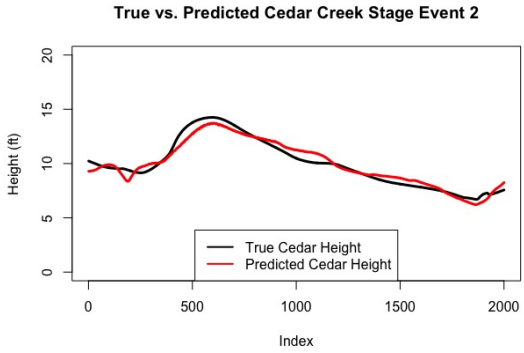
```
gaitSmoothStats
plot(gaitSmoothStats[, c(1, 3)], type='b')
#This is GCV want this minimized
}
```

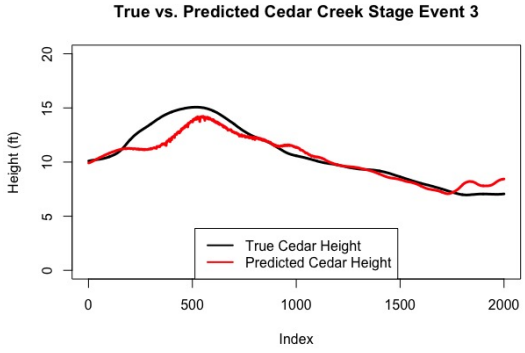
A.2 Chapter 3 Additional Notes

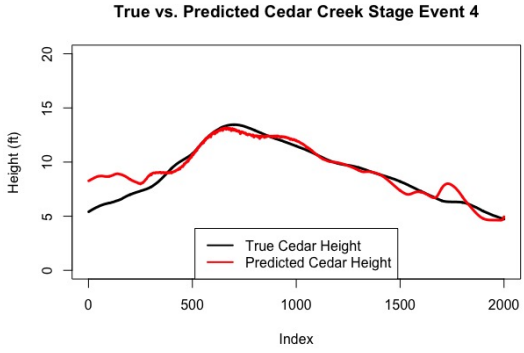
A.2.1 Cross Validation

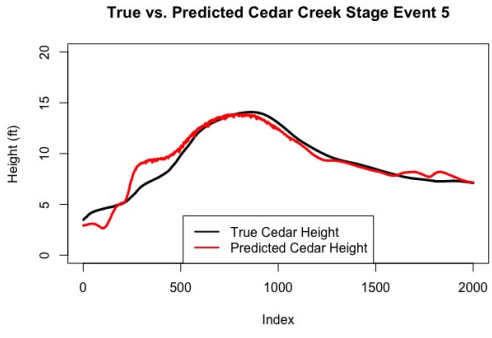
Below gives the plots of all 10 event's Cedar Creek curve along with the reconstructed Cedar Creek curve when that event is withheld from the model. The key takeaway is that once the 3-foot adjustment is made for the events prior to 1998, there is no clear discrepancy between the reconstructed observation and the true observation, indicating that the concurrent functional model is appropriate to use with the river stage data.

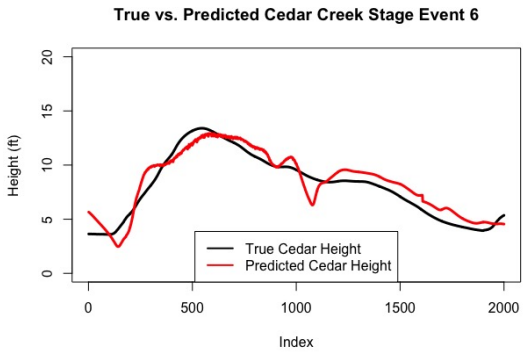

True vs. Predicted Cedar Creek Stage Event 1
Height (ft)
0
5
10
15
20
0
500
1000
1500
2000
Index
True Cedar Height
Predicted Cedar Height


True vs. Predicted Cedar Creek Stage Event 2
Height (ft)
0
5
10
15
20
0
500
1000
1500
2000
Index
True Cedar Height
Predicted Cedar Height


True vs. Predicted Cedar Creek Stage Event 3
Height (ft)
0
5
10
15
20
0
500
1000
1500
2000
Index
True Cedar Height
Predicted Cedar Height


True vs. Predicted Cedar Creek Stage Event 4
Height (ft)
0
5
10
15
20
0
500
1000
1500
2000
Index
True Cedar Height
Predicted Cedar Height


True vs. Predicted Cedar Creek Stage Event 5
Height (ft)
0
5
10
15
20
0
500
1000
1500
2000
Index
True Cedar Height
Predicted Cedar Height


True vs. Predicted Cedar Creek Stage Event 6
Height (ft)
0
5
10
15
20
0
500
1000
1500
2000
Index
True Cedar Height
Predicted Cedar Height

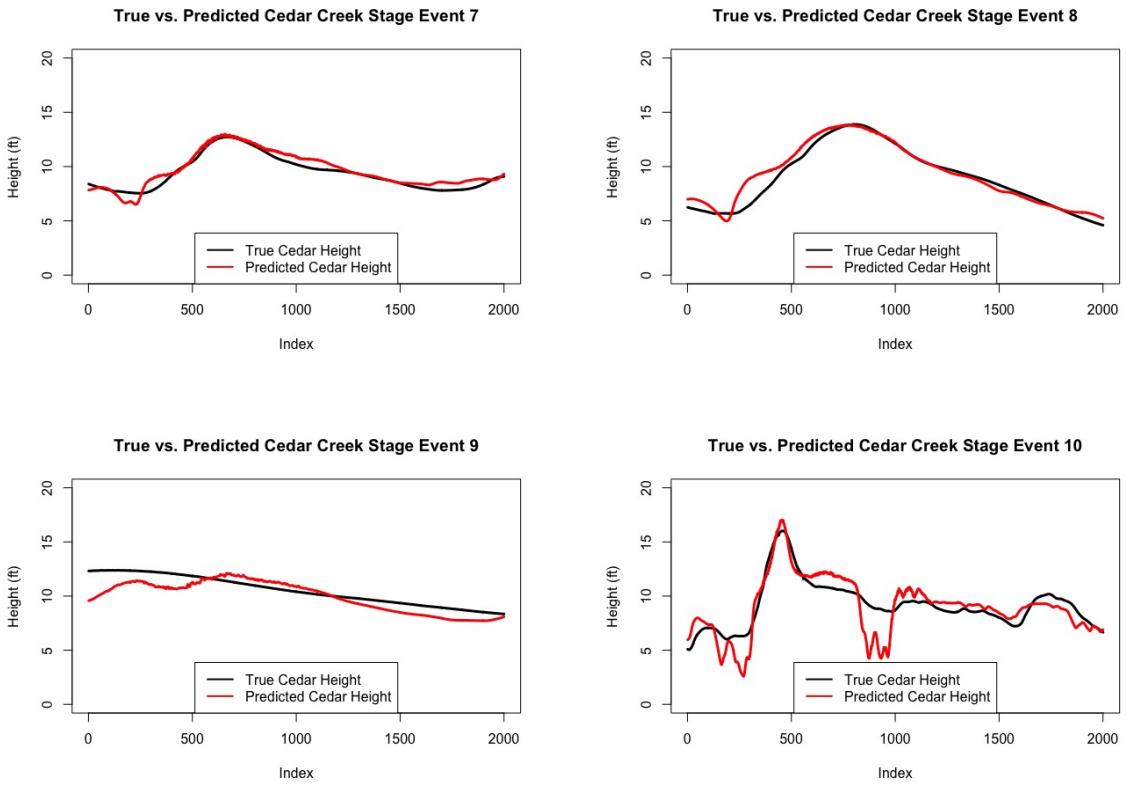
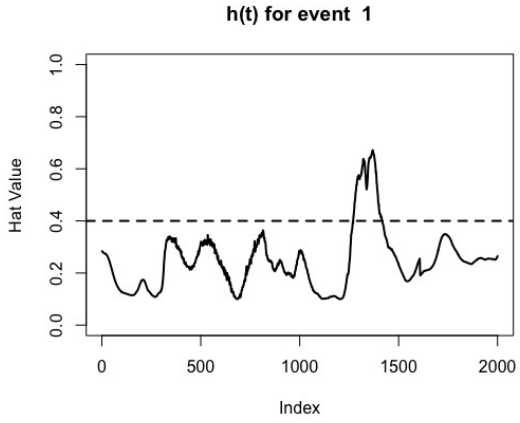
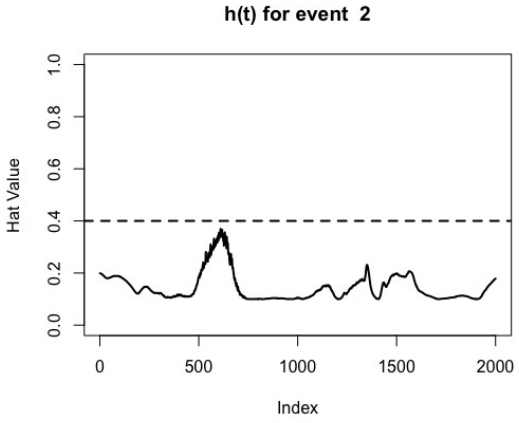


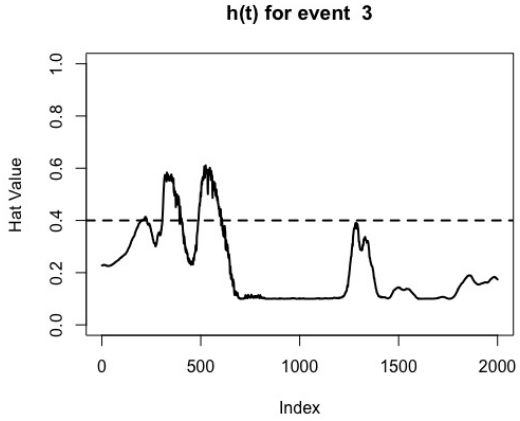
Figure A.8. True Cedar Creek height vs. the reconstructed Cedar Creek height with that event removed from the model.

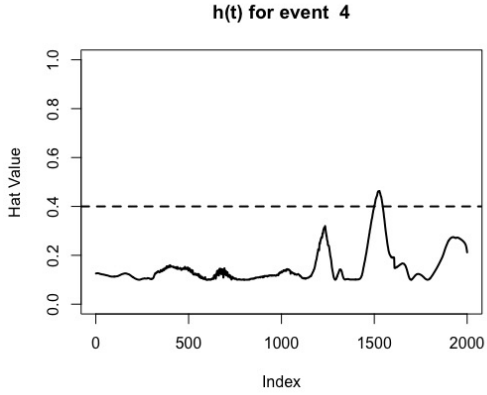
A.2.2 Hat Matrix diagonal at each t $h_i(t)$

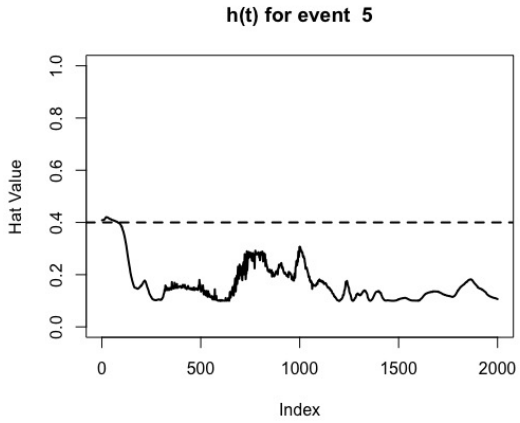
Below gives the leverage measurements for all 10 flood events via the diagonal of the hat matrix at each timepoint. Any observation with a high number of values across time should be noted as events with potentially high leverage.

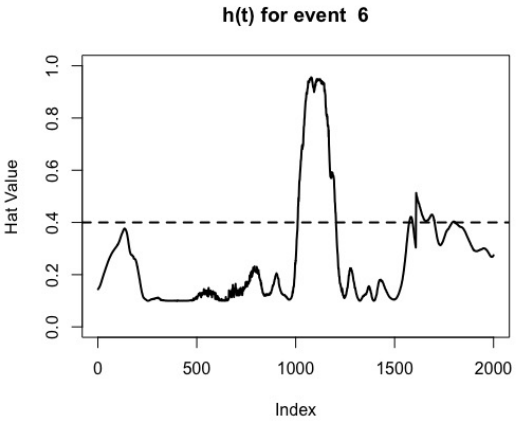

h(t) for event 1
Hat Value
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Index


h(t) for event 2
Hat Value
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Index


h(t) for event 3
Hat Value
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Index


h(t) for event 4
Hat Value
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Index


h(t) for event 5
Hat Value
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Index


h(t) for event 6
Hat Value
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Index

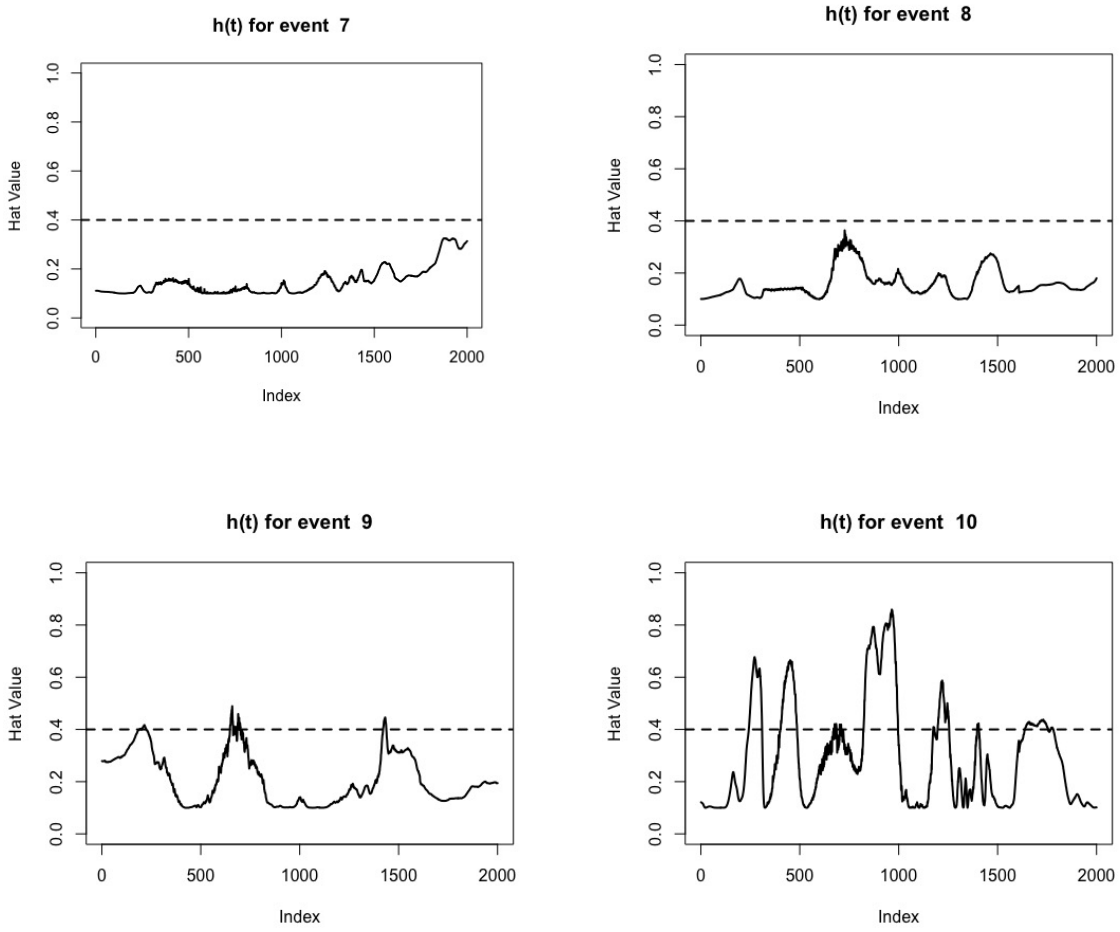


Figure A.10. Hat Matrix diagonal $h_{ii}(t)$ for all ten events (solid line) and the informal cutoff of 0.4 (dashed line).

A.2.3 PLOTS OF $\beta_p(t)$ VS. $\beta_{p(i)}(t)$

Below shows the difference between $\beta_0(t)$ vs. $\beta_{0(i)}(t)$ and $\beta_1(t)$ vs. $\beta_{1(i)}(t)$ when event i is removed.

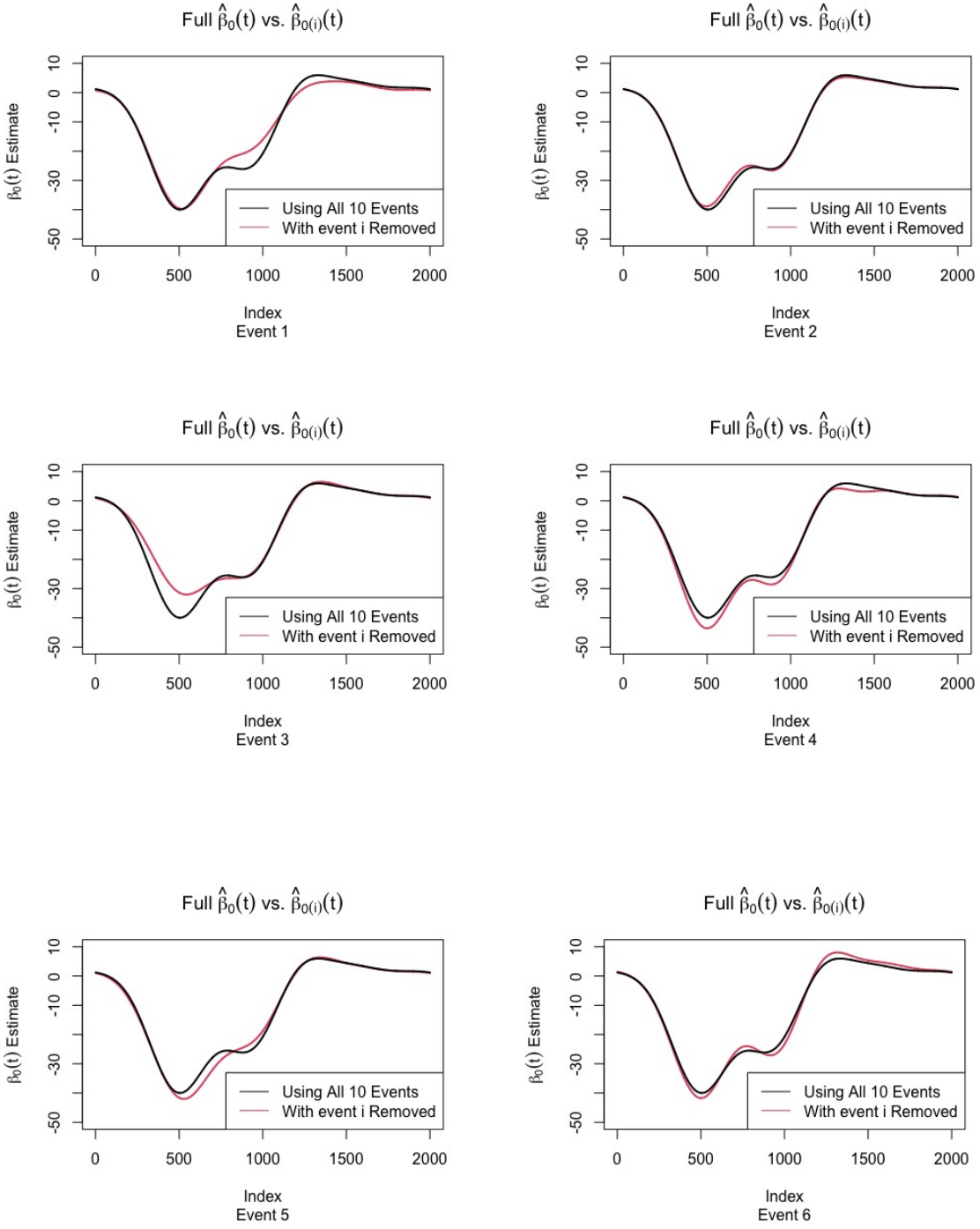


Figure A.11. The difference between the estimate for $\beta_0(t)$ vs. $\beta_{0(i)}(t)$ when event i is removed.

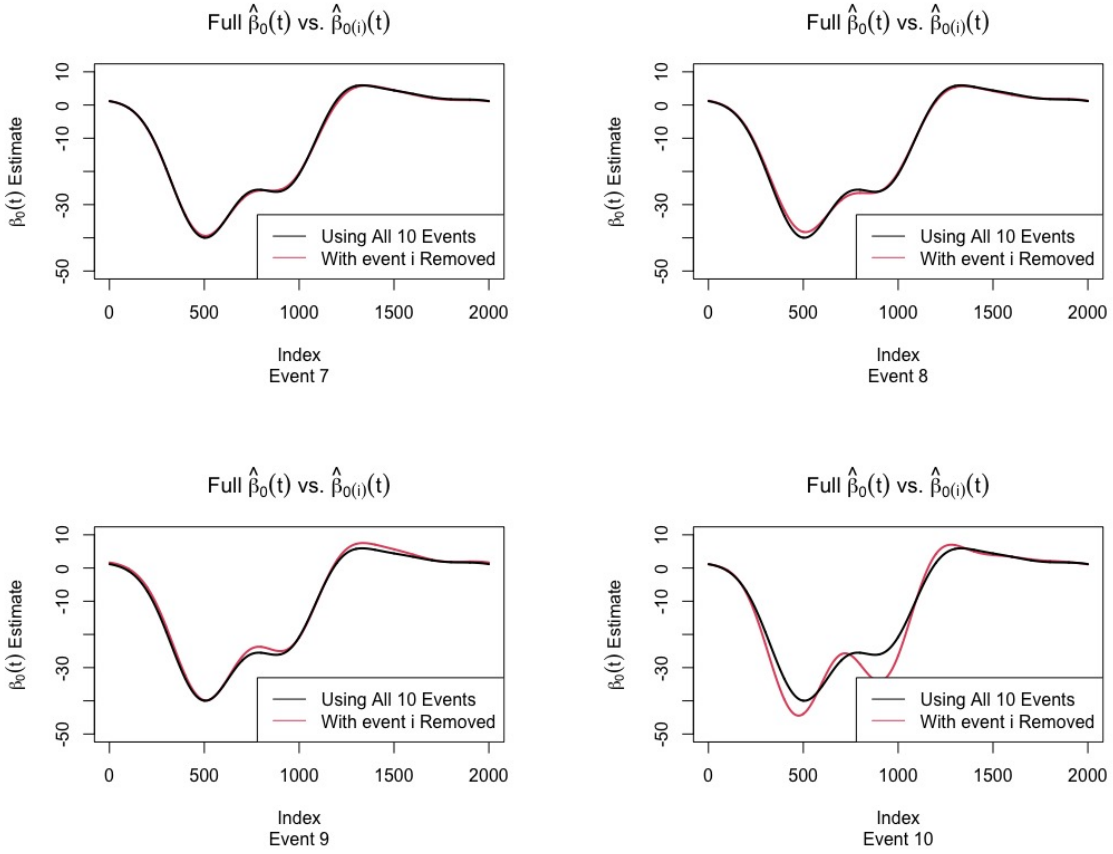


Figure A.12. The difference between the estimate for $\beta_0(t)$ vs. $\beta_{0(i)}(t)$ when event i is removed.

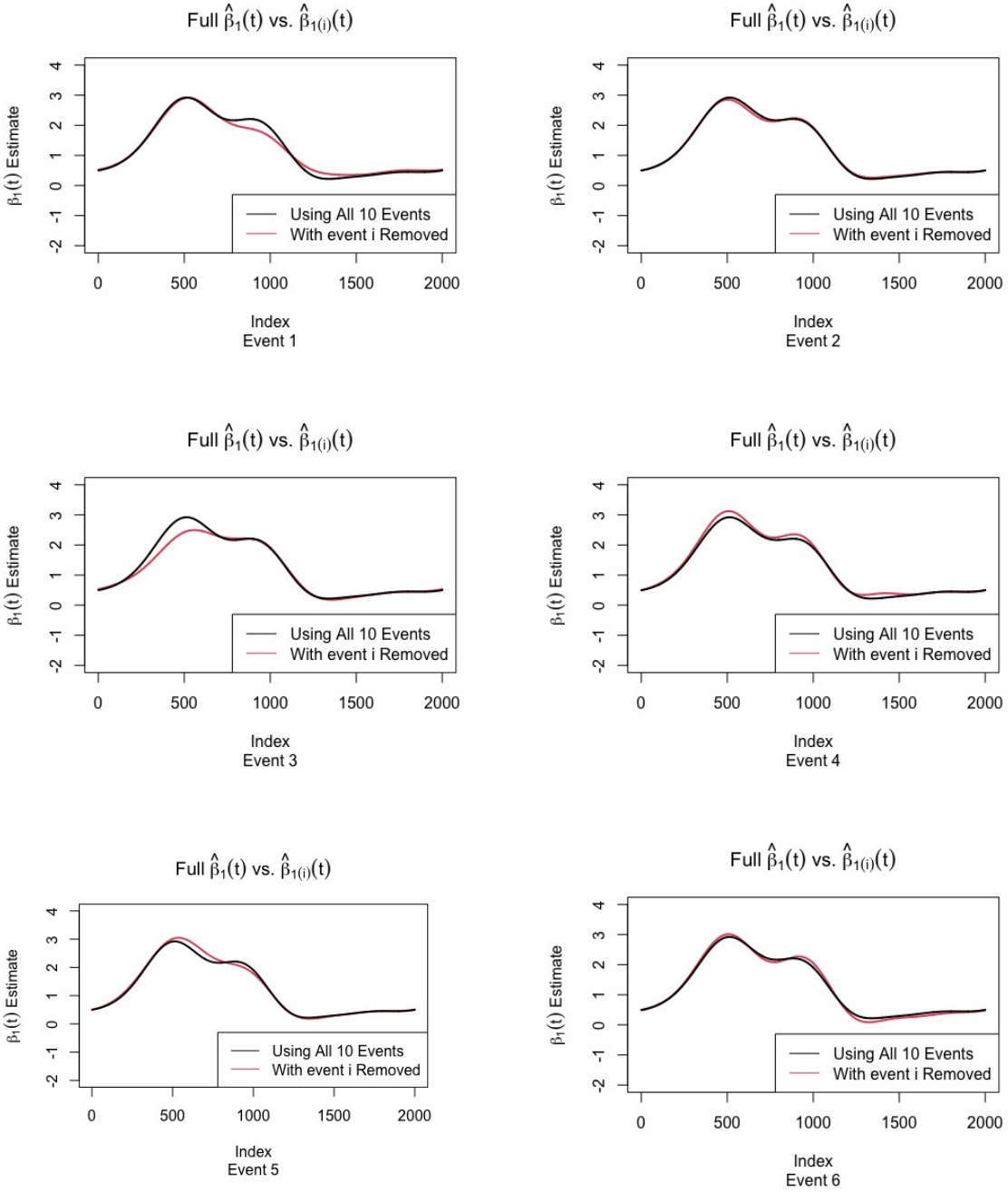


Figure A.13. The difference between the estimate for $\beta_0(t)$ vs. $\beta_{0(i)}(t)$ and $\beta_1(t)$ vs. $\beta_{1(i)}(t)$ when event i is removed.

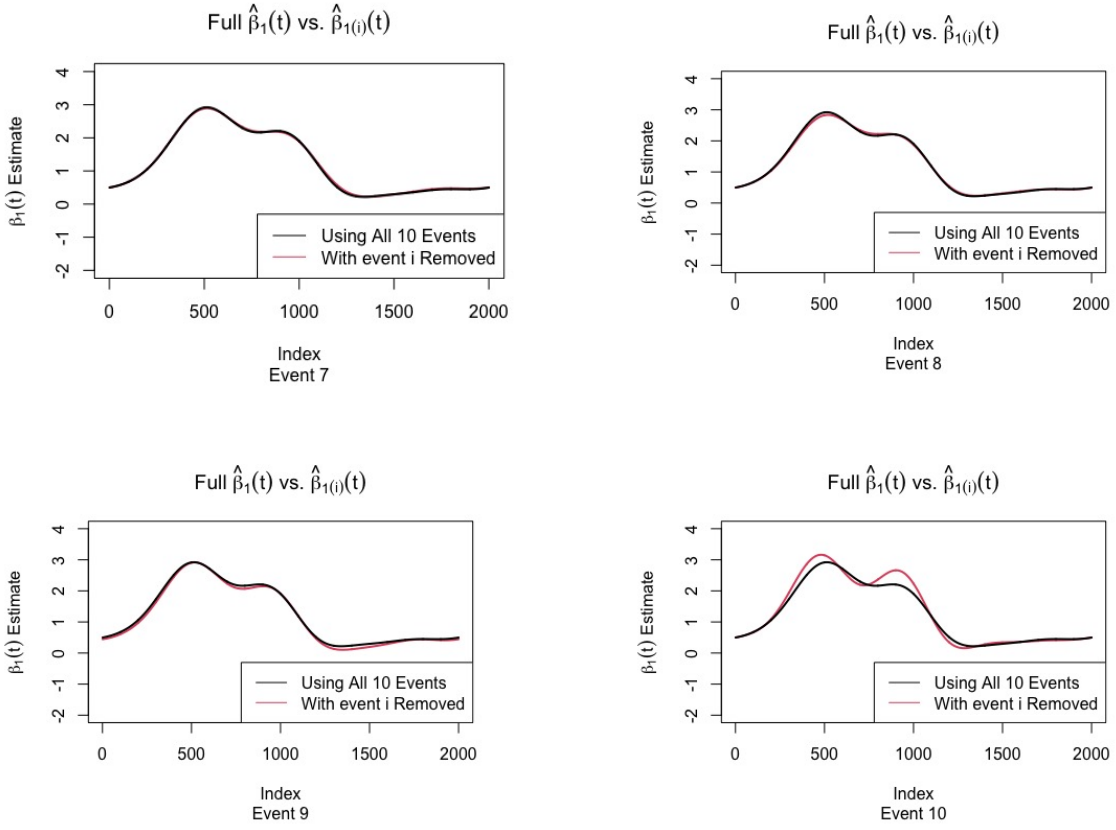


Figure A.14. The difference between the estimate for $\beta_1(t)$ vs. $\beta_{1(i)}(t)$ when event i is removed

A.2.4 ALL $DFFITS(t)$ FOR EACH FLOOD EVENT

Below shows all $DFFITS(t)$ results for the 10 flood events. We are mostly interested in identifying the events that are much larger than the others overall or have the largest spikes at a certain portion of the observation.

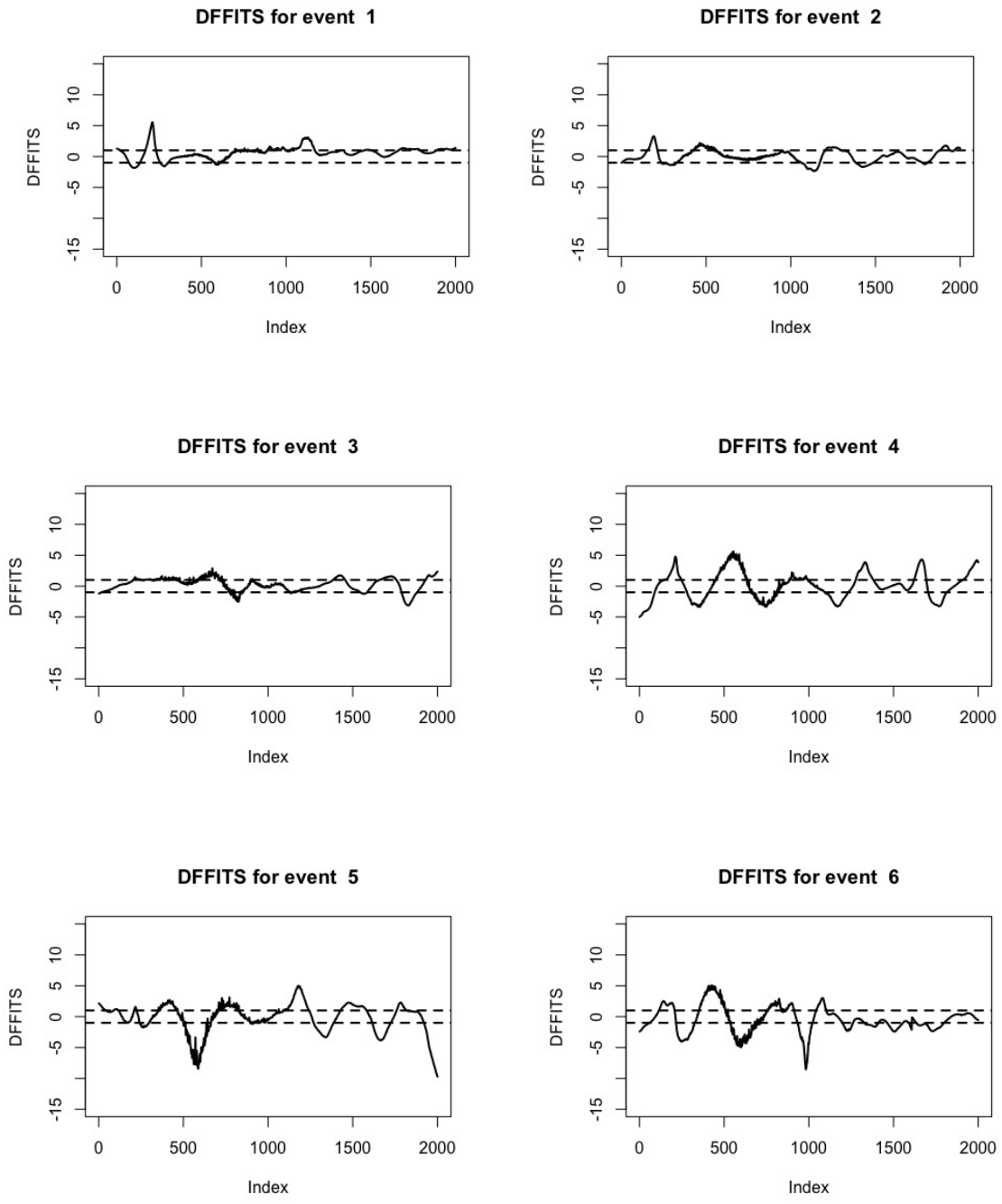


Figure A.15. DFFITS(t) for all ten events (solid line) and the informal cutoff of ±1 (dashed line)

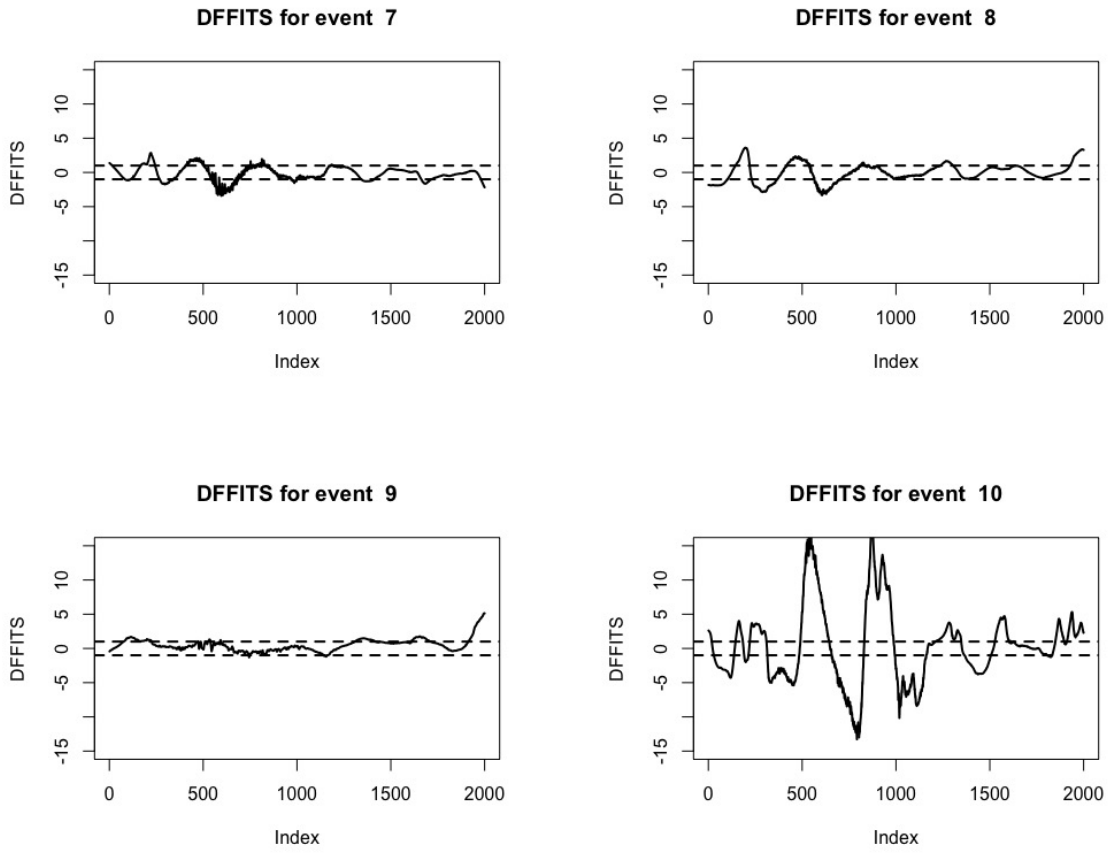


Figure A.16. DFFITS(t) for all ten events (solid line) and the informal cutoff of ± 1 (dashed line)

A.2.5 ALL PLOTS FOR COOK'S DISTANCE $D_i(t)$ FOR EACH EVENT

Below provides all 10 event's Cook's Distance $D_i(t)$ for each t. Here we are looking for any events that are consistently above the informal cutoff line and events that are regularly higher than the others or show large spikes. Note that event 6 and 10 stand out as potentially influential.

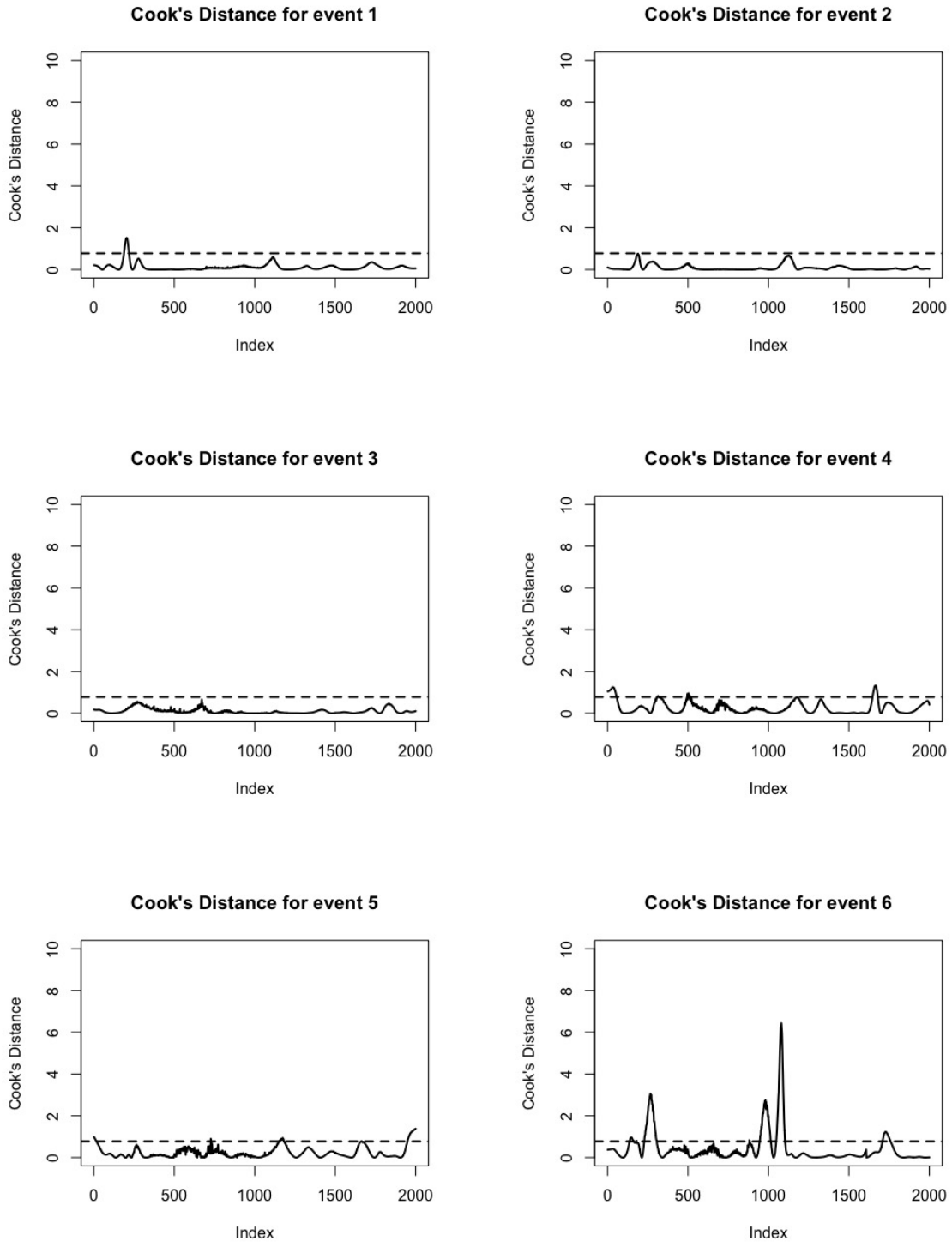


Figure A.17. Cook's distance (solid line) across t for each of the ten events with indication of significance (dashed line)

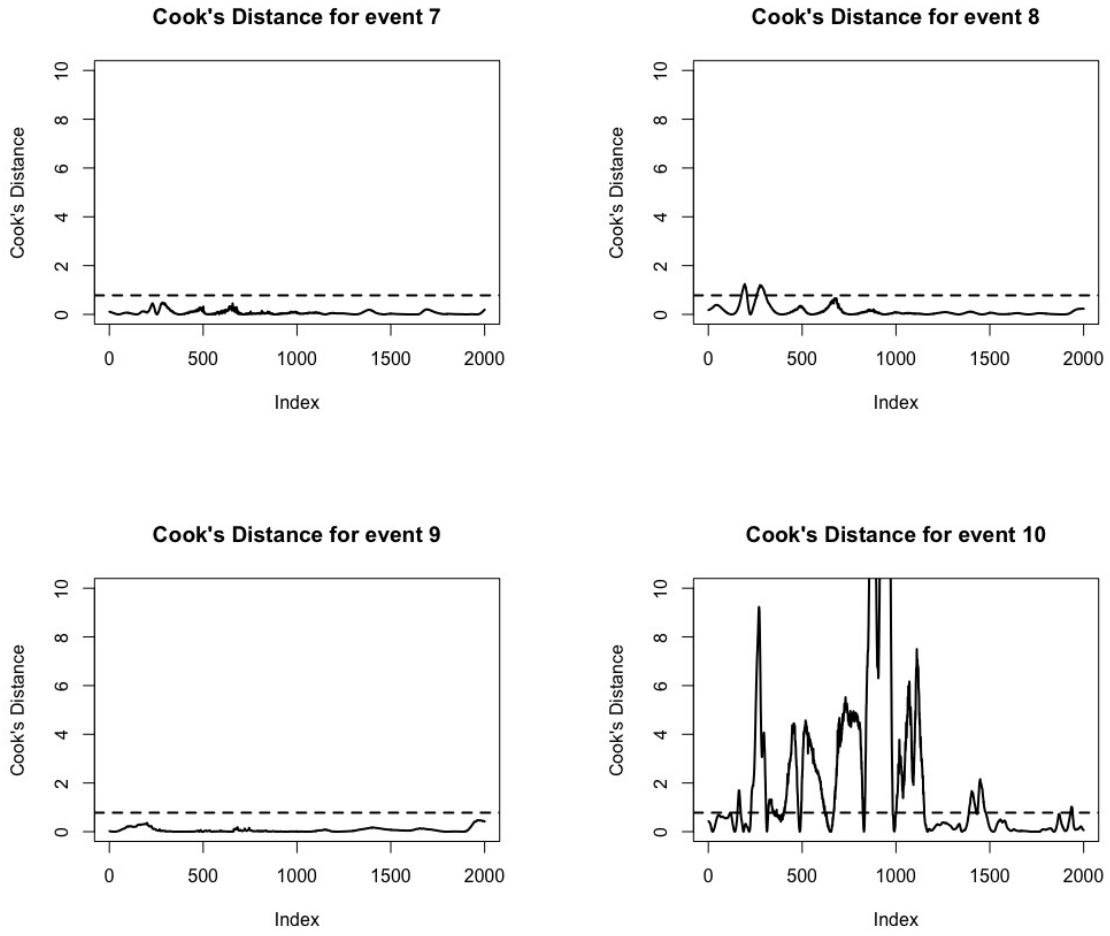


Figure A.18. Cook's distance (solid line) across t for each of the ten events with indication of significance (dashed line).

A.3 Chapter 4 Additional Notes

A.3.1 Difference between $\hat{Y}^{new}(t)$ and $\hat{Y}^{new}_{(i)}(t)$ for each event

Each event's Δp is calculated by looking at the squared area between the October 2015 Cedar Creek reconstructed stage with all 10 events used for reconstruction compared to the reconstruction with event i left out. The observation with the greatest difference between the two curves is the observation that has the greatest overall impact on the October 2015 Cedar Creek reconstruction.

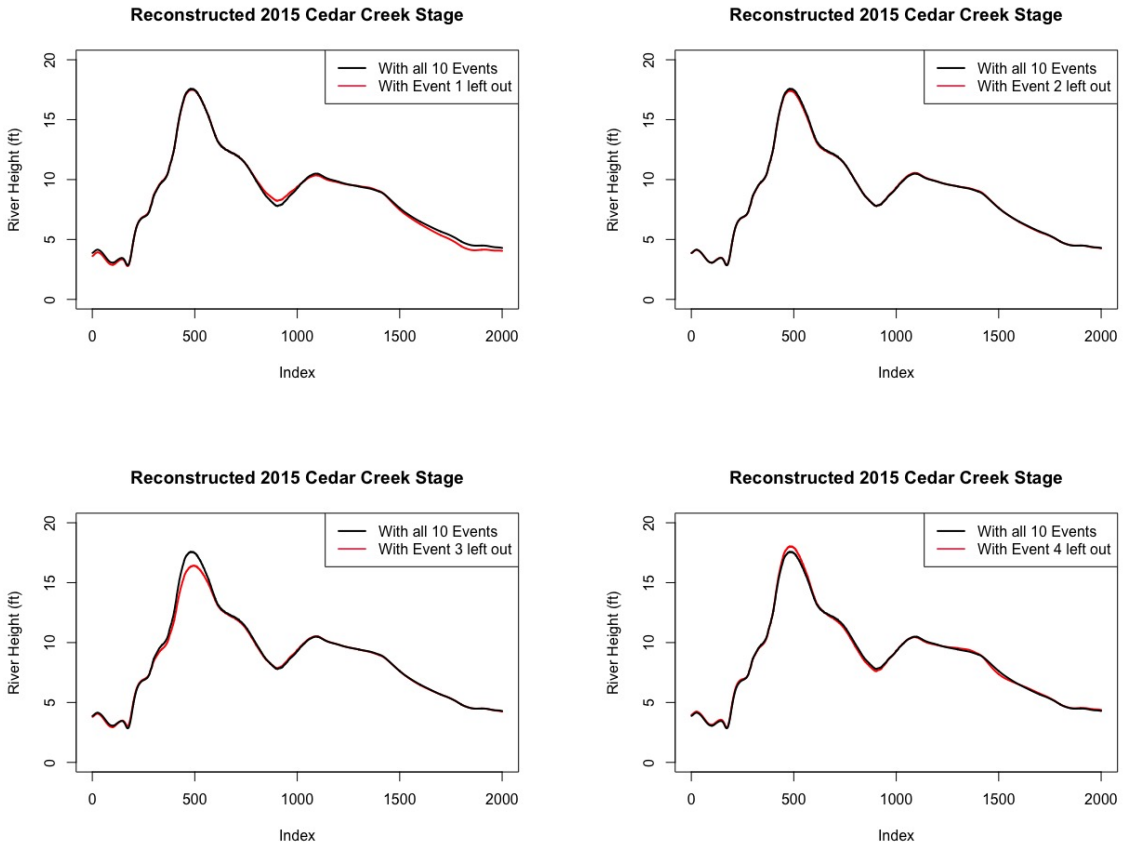


Figure A.19. October 2015 Cedar Creek reconstruction with all events (black line) and with event i withheld from the reconstruction (red line).

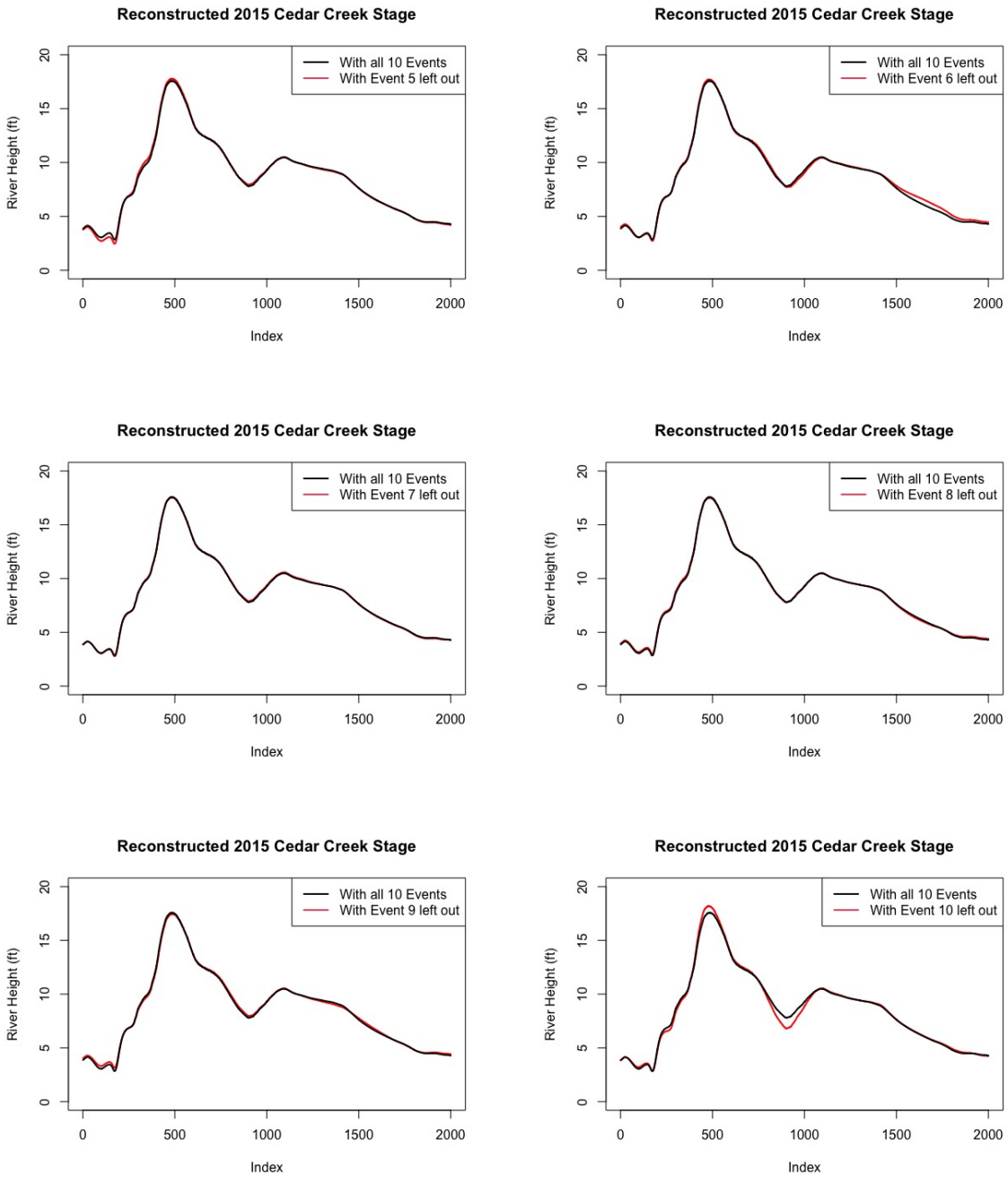


Figure A.20. October 2015 Cedar Creek reconstruction with all events (black line) and with event i withheld from the reconstruction (red line).

A.3.2 TIME MANAGEMENT R CODE

The function `julsecondsymd` is needed when the date variable is formatted yyyy-mm-dd and `julsecondsmdy` function is when the data variable is in mm/dd/yy format.

```
julsecondsymd<-function(dataset, origin="1970-01-01"){
  julsecvec <-rep(NA,nrow(dataset))
  data.time<-dataset[,"Time"]
  for(i in 1:nrow(dataset)){
    julsecvec[i]<-julian(as.Date(dataset[i,"Date"],"%Y-%m-%d"))[1]
    *24*60*60 +
    as.numeric(as.period(hms(data.time[i]), unit = "sec"))
  }
  return(julsecvec)
}


julsecondsmdy<-function(dataset, origin="1970-01-01"){
  julsecvec <-rep(NA,nrow(dataset))
  data.time<-dataset[,"Time"]
  for(i in 1:nrow(dataset)){
    julsecvec[i]<-julian(as.Date(dataset[i,"Date"],"%m/%d/%y"))[1]
    *24*60*60 + as.numeric(as.period(hm(data.time[i]), unit = "sec"))
  }
  return(julsecvec)
}
jul2dt<- function(juliansec){
  as.POSIXlt(juliansec,origin="1970-01-01", tz="GMT")
  #Timezone does not matter in my case GMT makes it work
}
```

A.3.3 Additional Simulation Results

Below shows additional average power and p-value for the simulation study where $N = 20$ and $N = 50$ where $B = 100$ and 100 repetitions.

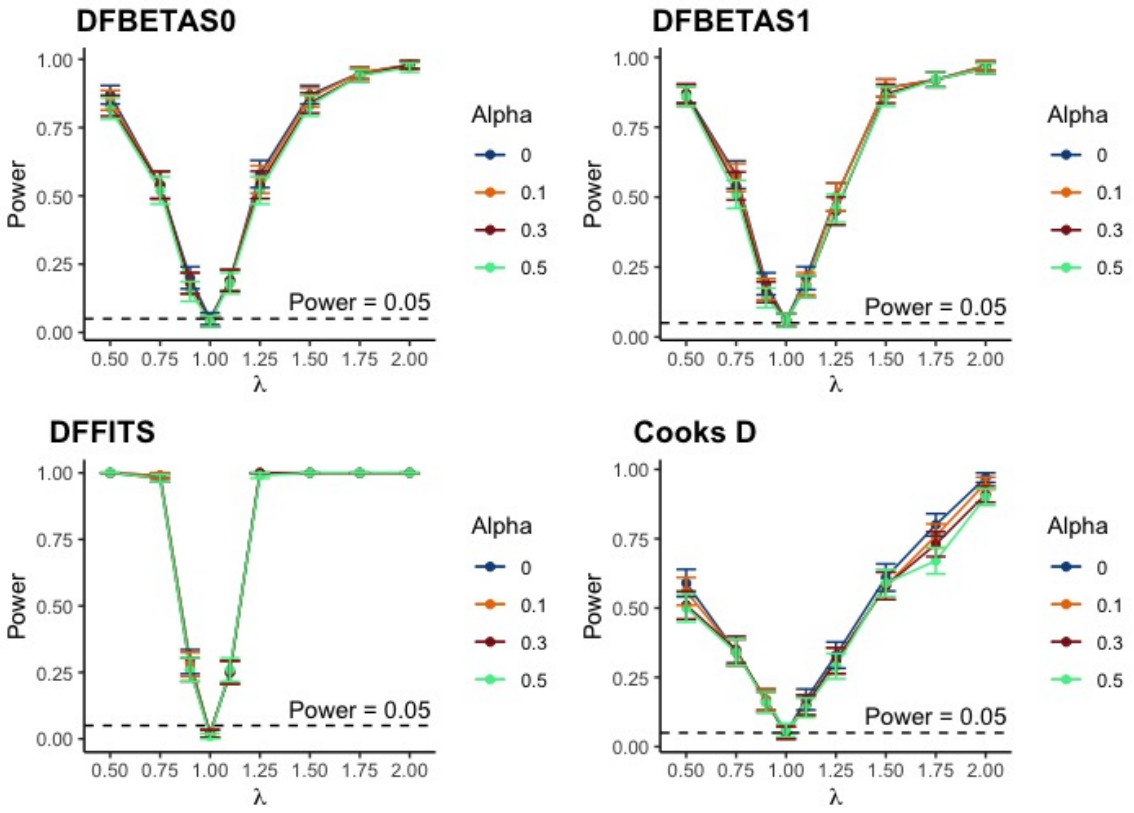


Figure A.21. Average proportion of adjusted observations above the 95th percentile for the four influence measures and different values of α with standard error for $N = 50$.

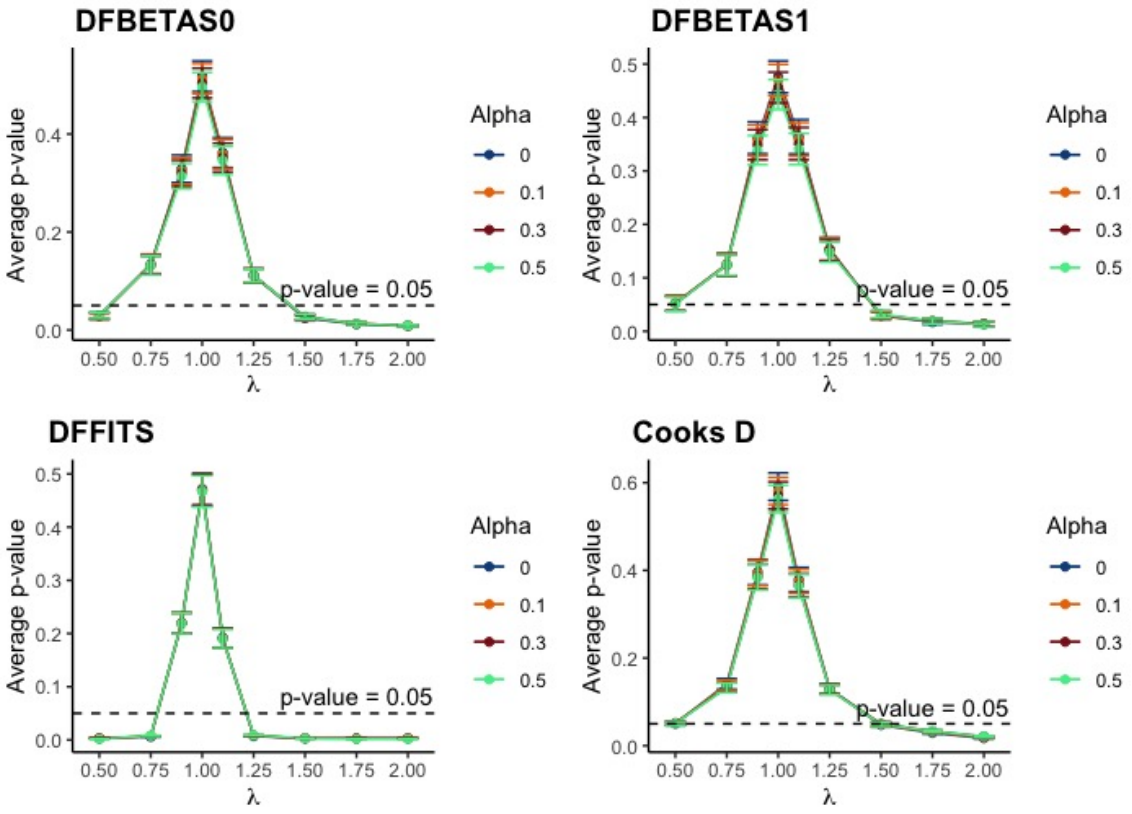


Figure A.22. Average p-value (1-percentile) of adjusted observations for the four influence measures and different values of α with standard error for $N = 50$.

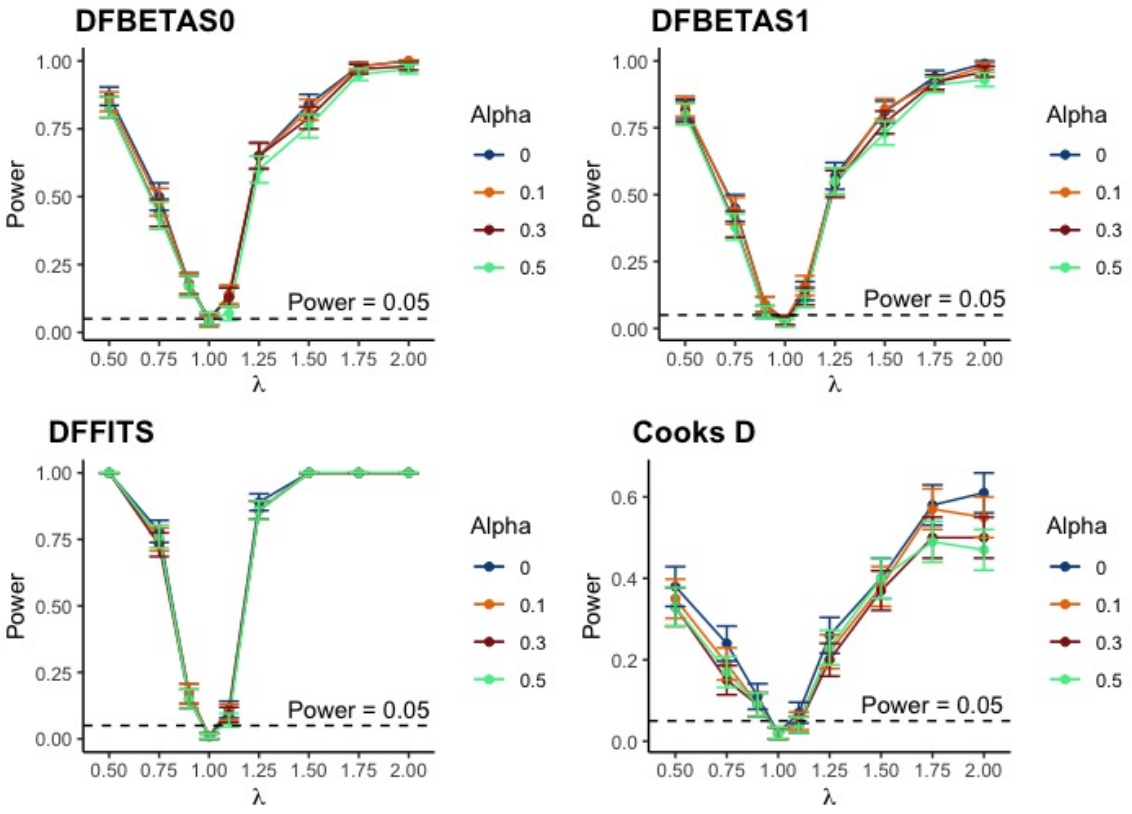


Figure A.23. Average proportion of adjusted observations above the 95th percentile for the four influence measures and different values of α with standard error for $N = 20$.

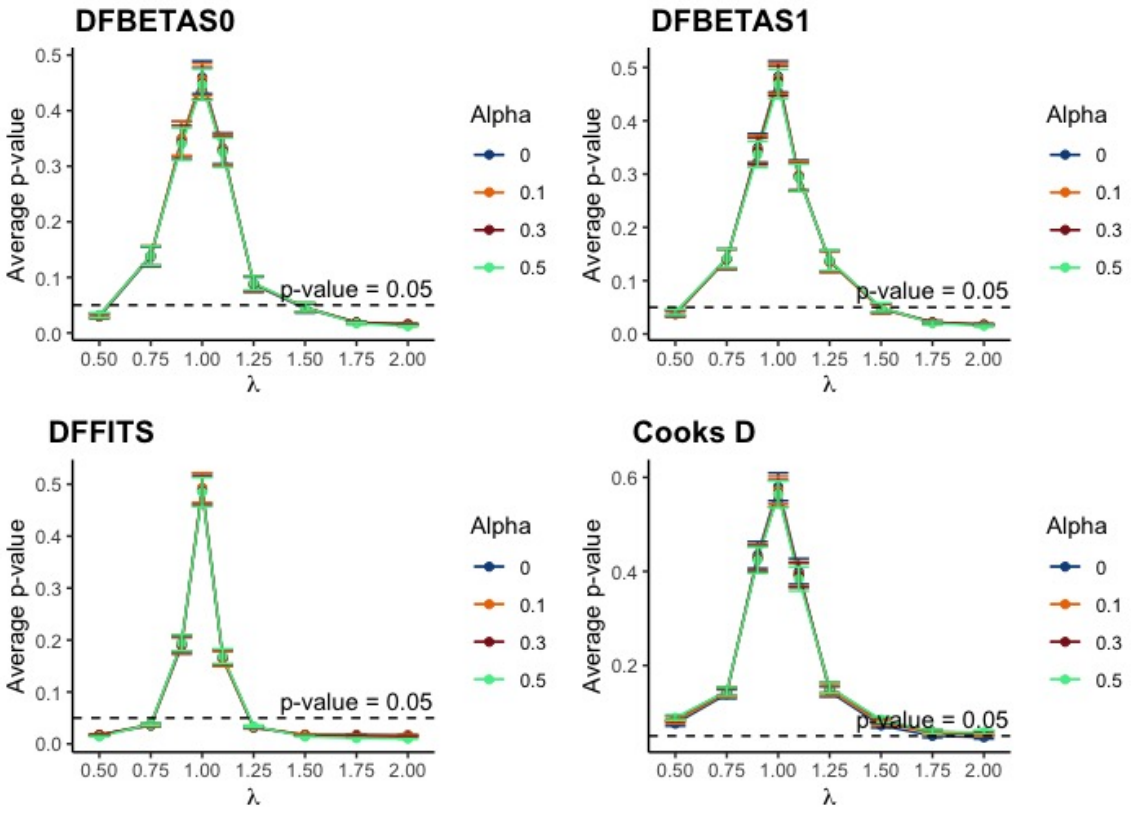


Figure A.24. Average p-value (1-percentile) of adjusted observations for the four influence measures and different values of α with standard error for $N = 20$.

A.4 Locations of Each Weather Station

Table A.1. Name and location of each weather station used in air and water temperature example.

Weather Stations			
Column Number	Location	Station ID	Coordinates
1.	Amerada Pass, LA	AMRL1 - 8764227 - LAWMA	29.450 N 91.338 W
2.	Atlantic City, NJ	ACYN4 - 8534720	39.357 N 74.418 W
3.	Bar Harbor, ME	ATGM1	44.392 N, 68.204 W
4.	Bay Waveland Yacht Club, MS	WYCM6 - 8747437	30.326 N 89.326 W
5.	Beaufort, NC	BFTN7 - 8656483	34.717 N 76.671 W
6.	Bishops Head, MD	BISM2 - 8571421	38.220 N 76.039 W
7.	Boston, MA	BHBM3	42.355 N 71.050 W
8.	Bridgeport CT	BRHC3	41.174 N 73.181 W
9.	Calcasieu Pass, LA	CAPL1 - 8768094	29.768 N 93.343 W
10.	Charleston, Cooper River Entrance, SC	CHTS1 - 8665530	32.781 N 79.924 W
11.	Clearwater Beach, FL	CWBF1 - 8726724	27.978 N 82.832 W
12.	Cordova, AK	CRVA2 - 9454050	60.558 N 145.752 W
13.	Crescent City, CA	CECC1 - 9419750	41.746 N 124.184 W
14.	Fernandina Beach, FL	FRDF1 - 8720030	30.675 N 81.465 W
15.	Fort Pulaski, GA	FPKG1 - 8670870	32.035 N 80.903 W
16.	Johnny Mercer Pier, Wrightsville Beach, NC	JMPN7 - 8658163	34.213 N 77.786 W
17.	Ketchikan, AK	KECA2 - 9450460	55.331 N 131.625 W
18.	King Cove, AK	KGCA2 - 9459881	55.062 N 162.327 W
19.	Lake Worth Pier, FL	LKWF1 - 8722670	26.613 N 80.034 W
20.	Mokuoloe, HI	MOKH1 - 1612480	21.433 157.790
21.	Naples, FL	NPSF1 - 8725110	26.132 N 81.807 W
22.	Old Port Tampa, FL	OPTF1 - 8726607	27.858 N 82.553 W
23.	Oregon Inlet Marina, NC	ORIN7 - 8652587	35.796 N 75.548 W
24.	Panama City Beach, FL	PCBF1 - 8729210	30.213 N 85.880 W
25.	Port Angeles, WA	PTAW1 - 9444090	48.125 N 123.441 W
26.	Port Chicago, CA	PCOC1 - 9415144	38.056 N 122.039 W
27.	Portland, ME	CASM1 - 8418150	43.656 N 70.246 W
28.	Port Isabel, TX	PTIT2 - 8779770	26.061 97.215
29.	Port Orford, OR	PORO3 - 9431647	42.739 N 124.498 W
30.	Port San Luis, CA	PSLC1 - 9412110	35.169 N 120.754 W
31.	Red Dog Dock, AK	RDDA2 - 9491094	67.575 N 164.067 W
32.	Sand Island, Midway Islands	SNDP5 1619910	28.215 N 177.361 W
33.	Santa Monica Pier	ICAC1 - 9410840	34.008 N 118.500 W
34.	Skagway, AK	SKTA2 - 9452400	59.450 N 135.327 W
35.	Westport, WA	WPTW1 - 9441102	46.904 N 124.105 W

Table A.2. All observed influence measures for Air and Water temperature example

Column	Location	$\overline{\lvert DFBETAS_0 \rvert}$	$\overline{\lvert DFBETAS_1 \rvert}$	$\bar{D}$	$\overline{\lvert DFFITS \rvert}$
1	Amerada Pass, LA	0.063	0.168	0.025	0.828
2	Atlantic City, NJ	0.074	0.046	0.026	0.957
3	Bar Harbor, ME	0.179	0.132	0.009	0.467
4	Bay Waveland Yacht Club, MS	0.051	0.101	0.025	0.781
5	Beaufort, NC	0.023	0.036	0.017	0.777
6	Bishops Head, MD	0.092	0.052	0.023	0.915
7	Boston, MA	0.184	0.132	0.008	0.525
8	Bridgeport CT	0.14	0.086	0.014	0.653
9	Calcasieu Pass, LA	0.024	0.046	0.023	0.797
10	Charleston, Cooper River Entrance, SC	0.018	0.035	0.025	0.859
11	Clearwater Beach, FL	0.098	0.199	0.012	0.509
12	Cordova, AK	0.167	0.144	0.008	0.327
13	Crescent City, CA	0.163	0.118	0.007	0.367
14	Fernandina Beach, FL	0.172	0.286	0.023	0.689
15	Fort Pulaski, GA	0.017	0.032	0.025	0.904
16	Johnny Mercer Pier, Wrightsville Beach, NC	0.024	0.029	0.023	0.887
17	Ketchikan, AK	0.134	0.106	0.014	0.509
18	King Cove, AK	0.213	0.182	0.011	0.398
19	Lake Worth Pier, FL	0.071	0.187	0.009	0.377
20	Mokuoloe, HI	0.067	0.158	0.004	0.25
21	Naples, FL	0.061	0.132	0.01	0.403
22	Old Port Tampa, FL	0.068	0.15	0.012	0.495
23	Oregon Inlet Marina, NC	0.026	0.015	0.012	0.685
24	Panama City Beach, FL	0.04	0.091	0.018	0.756
25	Port Angeles, WA	0.162	0.125	0.004	0.271
26	Port Chicago, CA	0.057	0.029	0.016	0.698
27	Portland, ME	0.107	0.072	0.007	0.417
28	Port Isabel, TX	0.025	0.06	0.021	0.621
29	Port Orford, OR	0.255	0.18	0.008	0.493
30	Port San Luis, CA	0.085	0.07	0.021	0.815
31	Red Dog Dock, AK	1.418	1.426	0.623	1.863
32	Sand Island, Midway Islands	0.02	0.071	0.002	0.223
33	Santa Monica Pier	0.057	0.044	0.009	0.533
34	Skagway, AK	0.238	0.206	0.011	0.37
35	Westport, WA	0.139	0.102	0.003	0.249