Concurrent Functional Regression to Reconstruct River Stage Data during Flood Events

Ryan Pittman

July 29, 2021

Dr. David Hitchcock Dr. John Grego

Dept. of Statistics

University of South Carolina

216 LeConte College

1523 Greene Street

Columbia, SC 29063

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 2019). 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 02169672) 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 1a). The 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 (USGS 2019a). 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" (USGS 2019b). 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 GOES AROUND HERE]

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 (USGS 02169625) remained functioning throughout the October 2015 flood. This gage is located a few miles west of the Cedar Creek gage. Figure 2 highlights the location for each gage (National Park Service 2019).

[FIGURE 2 GOES AROUND HERE]

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.

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. (2018) 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 streamflow 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 potential future study, indicating that FDA is a reasonable approach for predicting flood curves. Our study will use functional regression to analyze floods; however, instead of using streamflow, we use river stages as our variables.

Usually, the initial step in functional data analysis is to express the data through basis expansion

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

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. 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 Data Collection and Landmark Aligned Selection

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:

$$Y_i(t) = \beta_0(t) + \beta_1(t)X_i(t) + \epsilon_i(t),\ i = 1, \ldots, N \tag{2}$$

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

$$\mathbf{X} = \begin{bmatrix} X_1(t_1) & \ldots & X_N(t_1) \\ X_1(t_2) & \ddots & \vdots \\ \vdots & \vdots & \vdots \\ X_1(t_n) & \ldots & X_N(t_n) \end{bmatrix} \tag{3}$$

and

$$\mathbf{Y} = \begin{bmatrix} Y_1(t_1) & \ldots & Y_N(t_1) \\ Y_1(t_2) & \ddots & \vdots \\ \vdots & \vdots & \vdots \\ Y_1(t_n) & \ldots & Y_N(t_n) \end{bmatrix} \tag{4}$$

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 September 30, 2019 were made available to us from members of the U.S. Geological Survey. These data can be found on [link redacted because of double blinding]. 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 (US Department of Commerce and NOAA). 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 1.

[TABLE 1 GOES AROUND HERE]

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 accurate 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:

$$d_1 = \int |a(t) - b(t)| dt \tag{5}$$

which we estimate via trapezoidal approximation, using the function `trapz` in the `pracma` package (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 described in the supplemental material; 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:

$$LAL_1 = \int |X(t) - X^*(t)| [X^*(t)^2] dt \tag{6}$$

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.

2.3 Applying Landmark Aligned Data Selection

We now describe our user-created `LaL1.align` R function (available at [link redacted because of double blinding]) 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 excessively 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 3a. They are very dissimilar, with different patterns, maximum heights, and lengths. These raw events are not suitable for the concurrent model. In contrast, Figure 3b 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 1b for the February 2020 event.

[FIGURE 3 GOES AROUND HERE]

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 3b 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 supplementary material).

The complete starting and ending points of these ten events are found in Table 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 $\mathbf{X}$ matrix in Equation (3) and the corresponding Cedar Creek curves into the $\mathbf{Y}$ matrix in Equation (4), in order to fit the concurrent model. There is a visually clear association between the two curves, as seen in Figure 1b, 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.

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.

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

We employ the `fRegress` function from the `fda` package (Ramsay et al. 2018) to fit the concurrent model in `R` (R Core Team 2019). 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 et al. 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.

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) 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 tradeoff 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. (2009) suggest generalized cross-validation (GCV), originally developed by Craven and Wahba (1979). The best choice for λ is the value that minimizes

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right) \tag{7}$$

Ramsay et al. (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 (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

$$d_2^{(cv)} = N^{-1} \sum_{i=1}^{N} \int (Y_i(t) - \hat{Y}_{i(i)}(t))^2 dt \tag{8}$$

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 (9):

$$\hat{Y}_i(t) = \hat{\beta}_0(t) + \hat{\beta}_1(t) X_i(t),\ i = 1, \ldots, N \tag{9}$$

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 [link redacted because of double blinding].

4.1 LaL1.align

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.

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.

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 conjuction with the following `L2bestEst` function.

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 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 (8)). This function takes as its input $\mathbf{X}$ and $\mathbf{Y}$. 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.

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 bootstapping is an acceptable method for producing prediction intervals for the October 2015 flood stage reconstruction.

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. (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. The code used to find λ can be found on [link redacted because of double blinding], and the resulting graph is available in the supplementary material.

Next, we 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.

6.1 Putting it all together: Producing Final Predictions

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 4a and Figure 4b. 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 4 GOES HERE]

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 feeds into Cedar Creek once it gets high enough to flow through the floodplains in the national park (see Figure 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 1b 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.

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), the concurrent model, to reconstruct the Cedar Creek stage during this flood event.

[FIGURE 5 GOES HERE]

The graph in Figure 5 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 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 occured during the February 2020 flood, with the highest crest occuring 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 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.

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. (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.

References

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

[2] Burris, R. "SC Floods' Damage: $12 Billion, Economists Say." *The State*, 1 Dec. 2015, `https://www.thestate.com/news/local/article47471060.html`, Accessed June 16, 2019.

[3] Chebana, F., S. Dabo-Niang, and T. B.M.J. Ouarda. 2012. "Exploratory Functional Flood Frequency Analysis and Outlier Detection." *Water Resources Research* 48(4). Retrieved June 17, 2019 (`https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1029/2011WR011040`).

[4] 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." *Numerische Mathematik* 31:377–404.

[5] 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'.' *IEEE Transactions on Biomedical Engineering* 66(3):759–67.

[6] Ferraty, F., A. Rabhi, and P. Vieu. 2005. "Conditional Quantiles for Dependent Functional Data with Application to the Climatic "El Niño" Phenomenon." *Sankhya: The Indian Journal of Statistics* 67(2):378-98.

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

[8] Masselot, P., S. Dabo-Niang, F. Chebana, and T. B.M.J. Ouarda. 2016. "Streamflow Forecasting Using Functional Regression." *Journal of Hydrology* 538:754–66.

[9] National Park Service, `https://www.nps.gov/cong/planyourvisit/maps.htm`, Accessed June 17, 2019.

[10] National Weather Service, *Historic October 1st to 5th, 2015 South Carolina Flooding Event*, `https://www.weather.gov/cae/HistoricFloodingOct2015.html`, Accessed June 17, 2019.

[11] Olshen, R. A., E. N. Biden, M. P. Wyatt, and D. H. Sutherland. 1989. "Gait Analysis and the Bootstrap." *The Annals of Statistics* 17(4):1419–40.

[12] R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

[13] Ramsay, J. and B. W. Silverman. 2005. *Functional Data Analysis*, New York: Springer.

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

[15] Ramsay, J. O., H. Wickham, S. Graves and G. Hooker. 2018. fda: Functional Data Analysis. R package version 2.4.8. `https://CRAN.R-project.org/package=fda`

[16] 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. `https://CRAN.R-project.org/package=corpcor`

[17] US Department of Commerce, and Noaa. *National Weather Service Advanced Hydrologic Prediction Service.* Advanced Hydrologic Prediction Service, `water.weather.gov/ahps2/hydrograph.php?wfo=cae&amp;gage=gads1`.

[18] USGS. *Water Questions & Answers What does the term river stage mean?*, `https://water.usgs.gov/edu/qa-measure-streamstage.html`, Accessed June 17, 2019a

[19] USGS. *How Streamflow is Measured*, `https://www.usgs.gov/special-topic/water-science-school/science/how-streamflow-measured?qt-science_center_objects=0#qt-science_center_objects` Accessed June 17, 2019b.

[20] USGS. *USGS 02169625 CONGAREE RIVER AT CONGAREE NP NEAR GADSDEN, SC.* `https://waterdata.usgs.gov/sc/nwis/uv?site_no=02169625`, Accessed June 23, 2020.

[21] USGS. *USGS 02169672 CEDAR CREEK AT CONGAREE NP NEAR GADSDEN, SC.* `https://waterdata.usgs.gov/sc/nwis/uv?site_no=02169672`, Accessed June 23, 2020.

[22] Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*, Springer, Fourth Edition, New York.

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

[24] Zhang, J., M. K. Clayton, and P. A. Townsend. 2011. "Functional Concurrent Linear Regression Model for Spatial Images." *Journal of Agricultural, Biological, and Environmental Statistics* 16(1):105-30.

8 Figures and Graphs

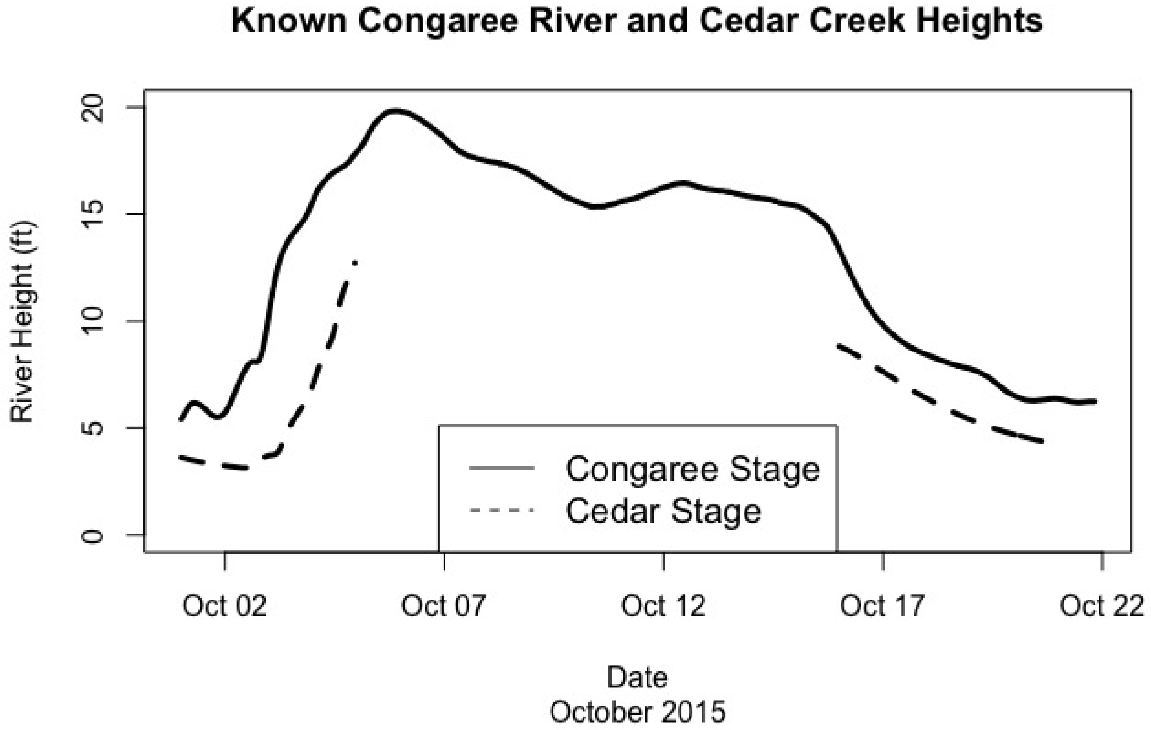
Table 1: Historic Congaree River Crests

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

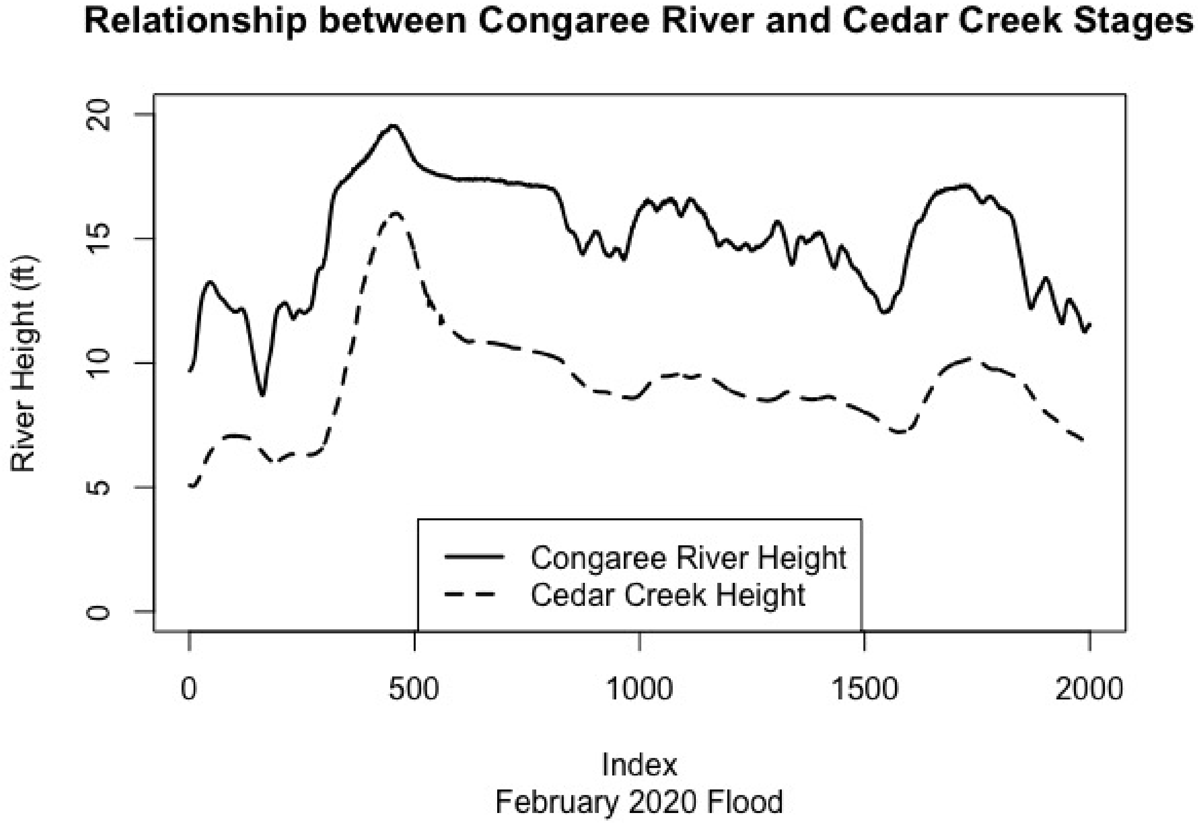
Maximum historic crests for the Congaree River gage at Congaree National Park, 1995-2020 and complete dates of observed flood records used in concurrent model

* Indicates no available corresponding Cedar Creek heights

Figure 1

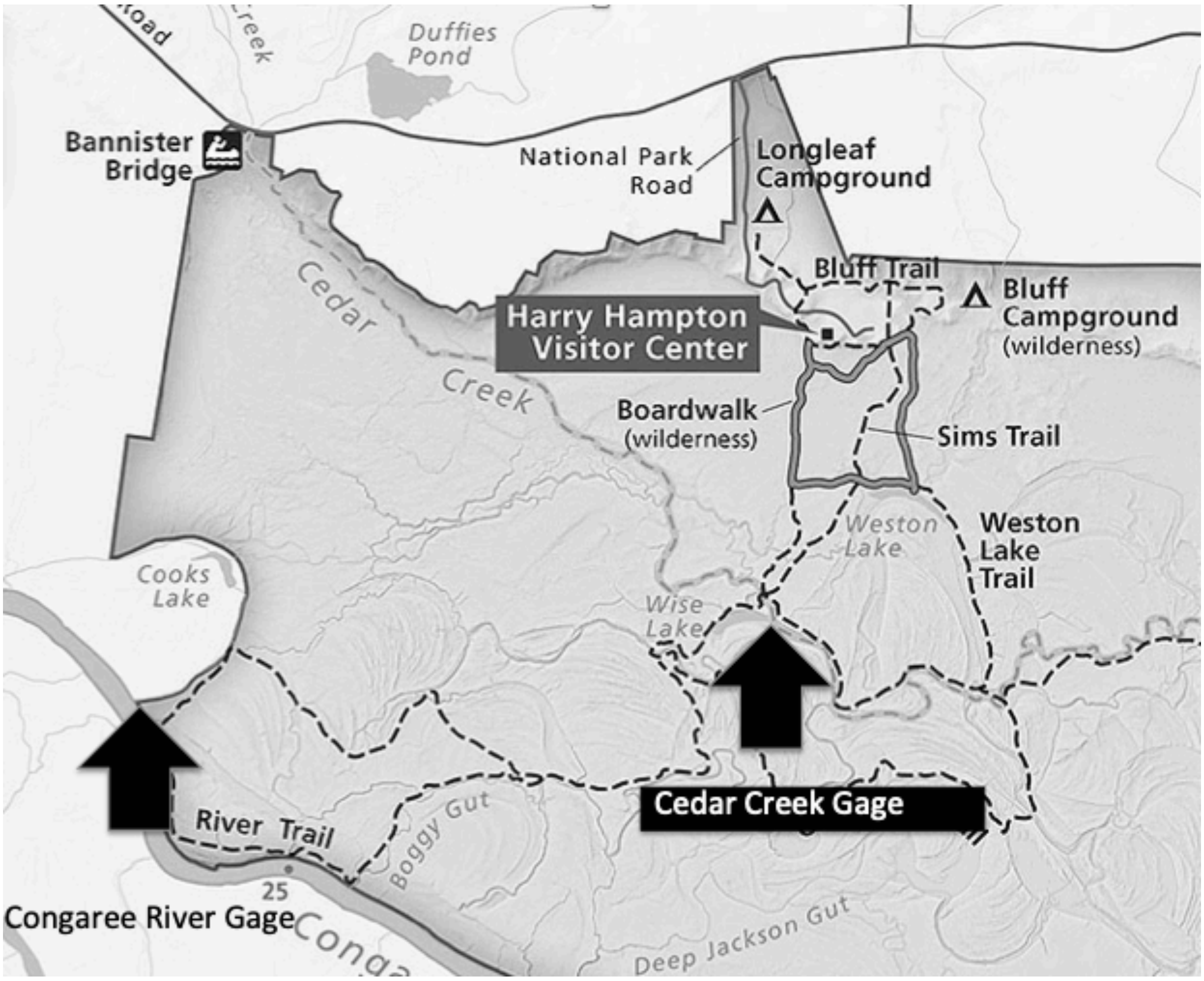


(a) Observed river stages for the Congaree River and Cedar Creek during the major October 2015 flood event in Columbia SC: Note the missing portion of the Cedar Creek height



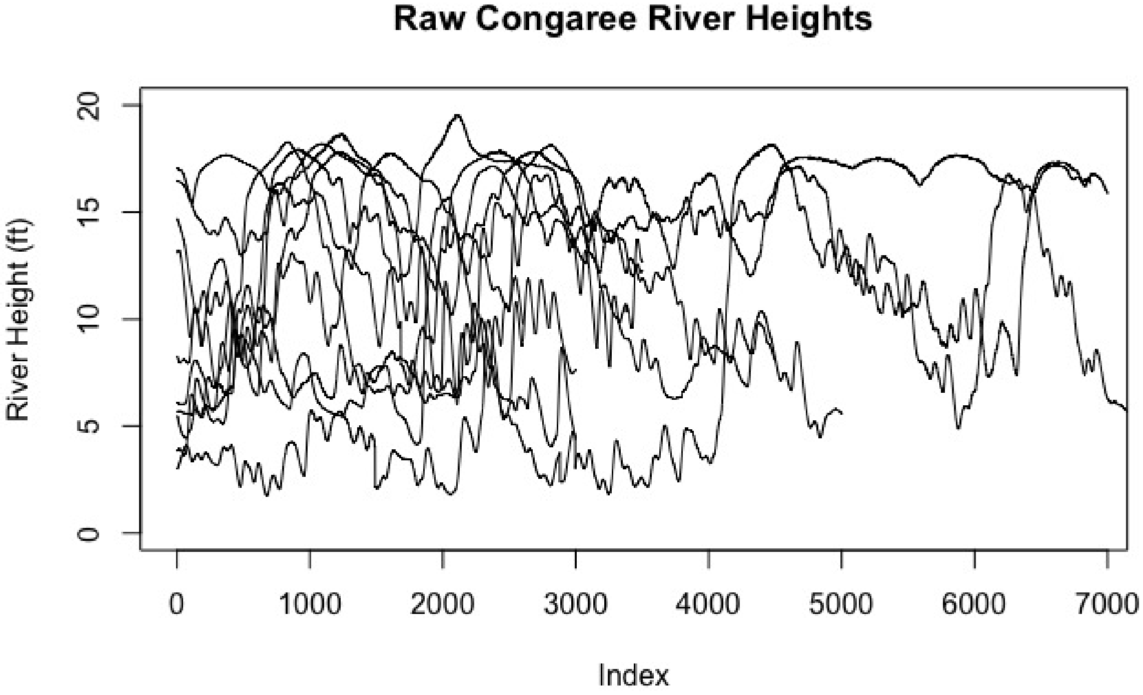
(b) Full, known stages for Congaree River and Cedar Creek during the February 2020 flood event

Figure 2

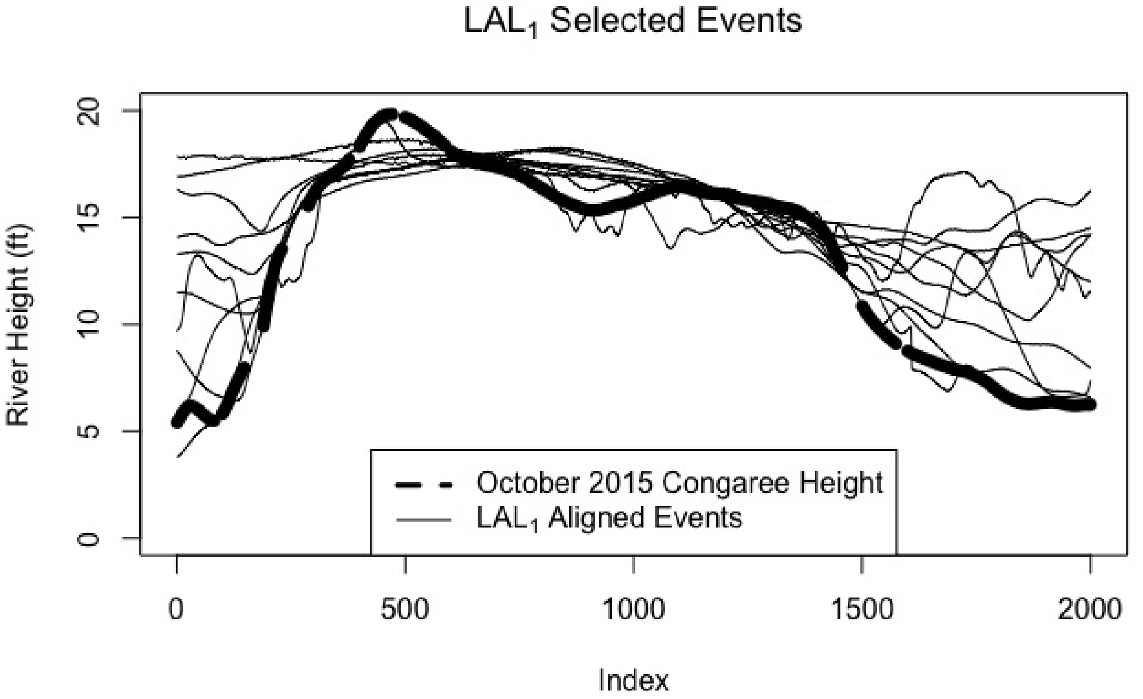


Map of Congaree National Park along with the approximate location of both of the gages used in this study

Figure 3

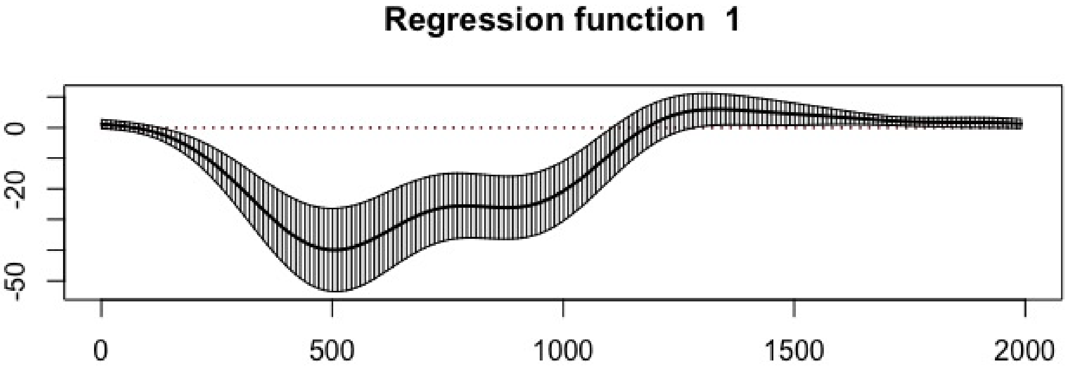


(a) Raw Congaree curves for all ten of the available flood events prior to using the selection method

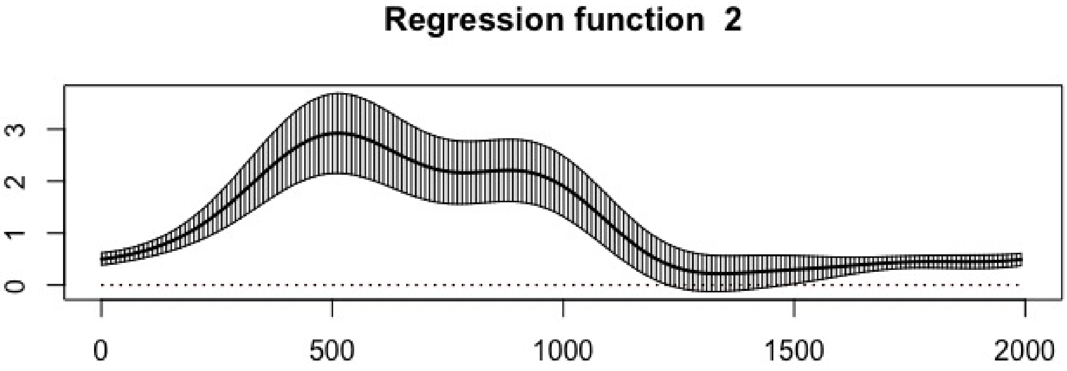


(b) All 10 LAL_1 selected Congaree River curves aligned with the target October 2015 Congaree River event

Figure 4

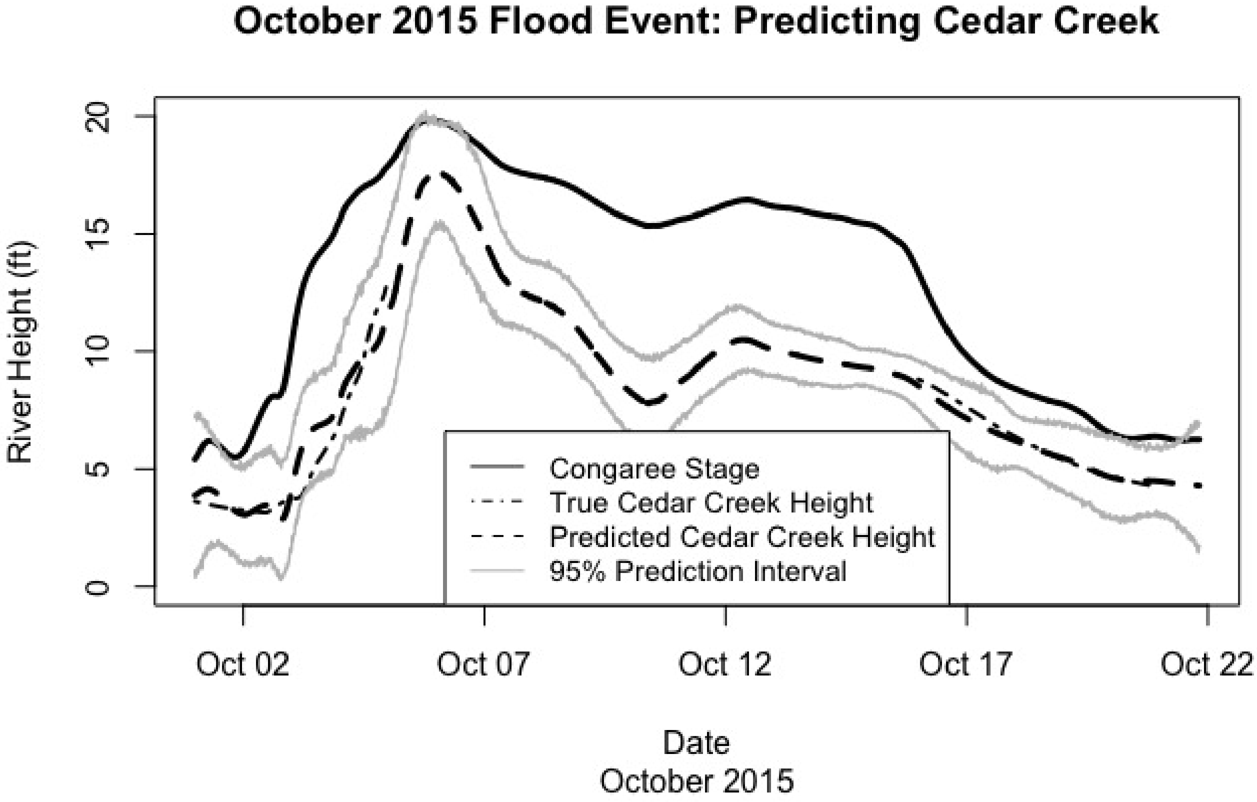


(a) $\beta_0(t)$ (Regression Function 1 = Intercept) Estimate using optimized LAL_1 distance selected data, optimized number of Fourier basis functions and pointwise 95% confidence limits



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

Figure 5



Predicted Cedar Creek stage for October 2015 flood event when the gage fails, accompanied by 95% pointwise confidence intervals and available true gage heights for Cedar Creek during the flood event.