Identifying influential observations in concurrent functional regression with weighted bootstrap

Ryan D. Pittman[a] and David B. Hitchcock[a]

[a] University of South Carolina Department of Statistics

ARTICLE HISTORY

Compiled January 17, 2022

Abstract

Metrics such as $DFBETAS$, $DFFITS$, and Cook's Distance are used in ordinary linear regression to assess the influence each individual observation has on the fitted model. We seek to quantify the influence of functional observations in the concurrent linear functional model in which both the predictor and response are functional observations. We present multiple influence measures that can be used to identify which functional observations are the most influential on the model. We provide a weighted bootstrapping with perturbations method to identify when these measures indicate an observation is significantly influential on the fitted regression model. We conclude by showing this method's validity using a simulation study and two real data examples that fit the model setting.

KEYWORDS

Functional data analysis; Influence measures; Concurrent model; Bootstrapping; River stage; Weather data

CONTACT R. D. Pittman. Email: rpittman@email.sc.edu

1. Introduction

In ordinary linear regression, it is common practice to assess the influence of the individual observations [1]. It is important to understand how trustworthy the predicted outcomes are and how influential each observation is on various results. This paper 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. These methods will then be applied to a river stage reconstruction and a coastal air and water temperature dataset, with computations carried out using our own R code in combination with functional regression estimation tools within the `fda` package [2] in `R` [3].

Functional linear regression model diagnostics have not been explored to the degree 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. Some prior work in this field has been done by Shen and Xu [4], who quantify the influence of many non-functional predictors on a functional response, and Chiou and Müller [5] who consider the case in which the response variable is functional but the predictor variables are either multivariate vectors or random functions. Chen, Huang, and Lin [6] build on Chiou and Müller's work and is similar to our study, in that both the response and predictors are functional observations. They calculate a version of functional Cook's distance and a likelihood distance, and present a small simulation study in which they intentionally insert outlying measurement points within a 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 a whole set of paired functional data $(X_i(t), Y_i(t)), i = 1, \ldots, N$. Febrero-Bande, Galeano, and González-Monteiga [7] 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 distinct from 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; we will also endeavor to approximate a null distribution using bootstrap. Throughout the rest of this paper, 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 a concurrent functional regression fitted model having one functional predictor and one corresponding functional response variable. Our usage of the concurrent functional model, along with a smoothing (using B-splines or Fourier bases) of the observed functional data, implicitly accounts for the nature of the functional observations, in particular the dependency of the measurements across time within the functional data. 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 assess the performance of the method, and we conclude with two different regression applications with real functional data.

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. It is common to use measures such as the diagonals of the hat matrix to measure an observation's leverage, and $DFBETAS$, $DFFITS$, and Cook's distance [8] to identify influential events by calculating each measure for every observation (formulas can be found in [9]). If any observation has a calculated influence measure above some corresponding accepted threshold then it should be investigated, and remedial measures may be necessary to produce reliable model results.

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 $\mathcal{T}$:

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

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

$$h_i(t) = i\text{th diagonal of the } N \times N \text{ matrix } \mathbf{H}_t \text{ where } \mathbf{H}_t = \mathbf{X}_t \left(\mathbf{X}_t^T \mathbf{X}_t\right)^{-1} \mathbf{X}_t^T \tag{2}$$

$$DFBETAS_{p,i}(t) = \frac{\hat{\beta}_p(t) - \hat{\beta}_{p(i)}(t)}{se(\hat{\beta}_{p(i)}(t))},\ \ p = 1, 2,\ \ i = 1, \ldots, N \tag{3}$$

$$DFFITS_i(t) = \frac{\hat{y}_i(t) - \hat{y}_{i(i)}(t)}{\sqrt{MSE_{(i)}(t)h_i(t)}},\ \ i = 1, \ldots, N \tag{4}$$

$$D_i(t) = \frac{\sum_{j=1}^{n} \left(\hat{y}_j(t) - \hat{y}_{j(i)}(t)\right)^2}{(k+1)MSE(t)},\ \ i = 1, \ldots, N \tag{5}$$

In Equation (2), $\mathbf{X}_t$ 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:

$$\overline{|DFBETAS_p|}_i = \frac{1}{n} \sum_{j \in \{1,\ldots,n\}} |DFBETAS_{p,i}(t_j)| \text{ for } i = 1, \ldots, N \tag{6}$$

$$\overline{|DFFITS|}_i = \frac{1}{n} \sum_{j \in \{1,\ldots,n\}} |DFFITS_i(t_j)| \text{ for } i = 1, \ldots, N \tag{7}$$

$$\overline{D}_i = \frac{1}{n} \sum_{j \in \{1,\ldots,n\}} D_i(t_j) \text{ for } i = 1, \ldots, N \tag{8}$$

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 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 apply these measures to another dataset that investigates the relationship between air and water temperatures at weather stations along the US coastline.

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 start by defining any particular measure of influence (averaged across time) generically as r_i, calculating it for each observation, and then using the following equation to translate the metric value for observation i into a selection probability θ_i:

$$\theta_i = \frac{(1/r_i)^\alpha}{\sum_i \left[(1/r_i)^\alpha\right]}, \quad \alpha \geq 0. \tag{9}$$

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 α exceeds 0.5, the selection probabilities of the more influential observations becomes too small and the resulting bootstrap samples consist mostly of the observations with minimal influence, leading to an unreliable resulting null distribution. 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 often enough to constitute 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 the following weighted bootstrapping with perturbation method. While this method can be implemented for any sample size, it is most useful in the small sample setting.

(1) Define r_i for each observation to be a particular influence measure of interest (namely, one of $\overline{|DFBETAS_p|}_i$, $\overline{|DFFITS|}_i$, or $\overline{D}_i$).

(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 event 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 $N \times B$ 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 distort the calculated metrics because any curve 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 process is the Ornstein-Uhlenbeck process approximated via the smoothed Euler-Maruyama method. The Ornstein-Uhlenbeck process, defined by Uhlenbeck and Ornstein [10], is x_t 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 process. The Euler-Maruyama approximation yields discrete values of this process and is applied to the functional response curves using:

$$\kappa_{n+1} = \kappa_n - \theta(\kappa_n)\delta t + \sigma Z\sqrt{\delta t} \tag{10}$$

where κ_0 is initialized by selecting a single value from a $N(0, \sigma^2)$ distribution and where Z is a random standard normal value.

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 $\{\text{range}[y_1(t)], \ldots, \text{range}[y_N(t)]\}$, where $\text{range}[y_j(t)] = \max_{\text{t}} y_j(t) - \min_{\text{t}} y_j(t)$.

(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 these 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 (9) 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 2 in Section 5.2 provides an example of the selection probabilities corresponding to various levels of α. 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 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 moderately influential, above the 95th percentile significantly influential, and above the 99th percentile as highly significantly influential and requiring further investigation.

4. Simulation study

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 where $\{1, 2, \ldots, 1000\}$ 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 values of 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 $\{-300, -200, -100, 100, 200, 300\}$.

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

By alternating the combination of parameter values used to generate the functional data, we produce curves that are similar and resemble the same underlying curve $m(t) = t/12$. An example of $N = 20$ $X(t)$ curves are shown in Figure 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.

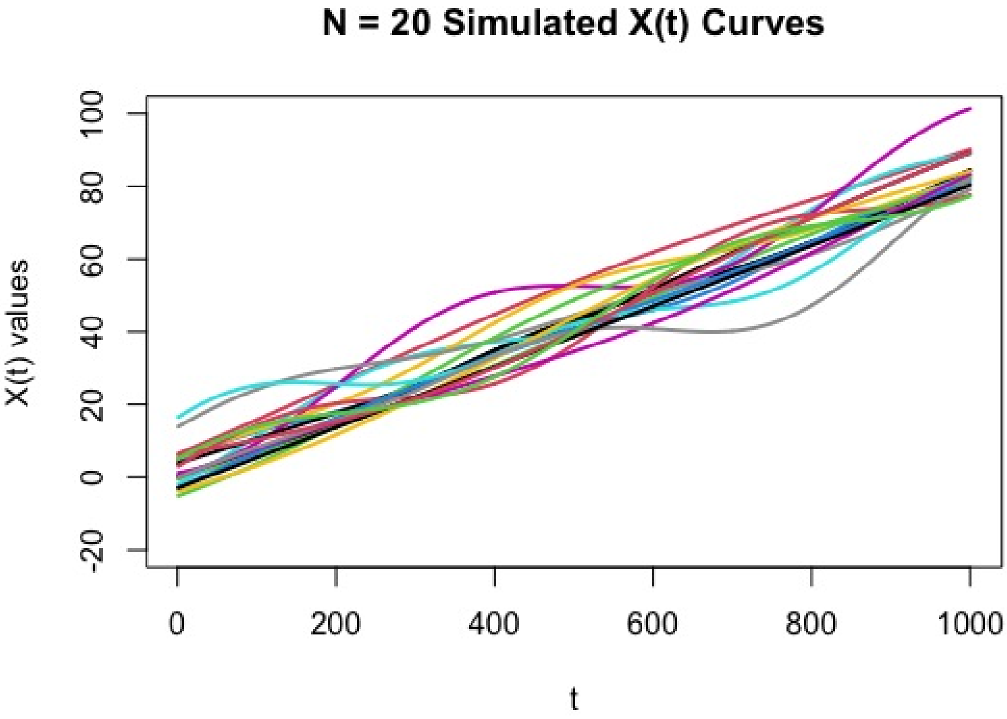

Figure 1. Example of $N = 20$ generated $X(t)$ curves using the described functional data generation method.

Next we set the functional slope and intercept functions to be:

$$\beta_0(t) = \cos(t/200) + 2 \tag{11}$$

$$\beta_1(t) = \sin(t/200) + 2 \tag{12}$$

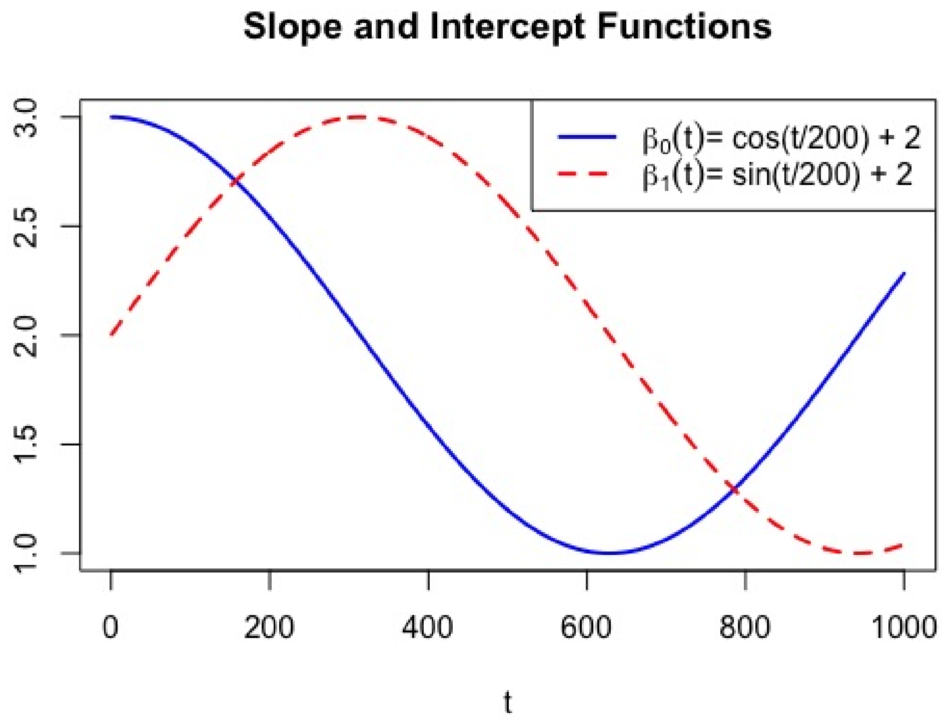

Figure 2. Defined functional intercept $\beta_0(t)$ (solid blue) and functional slope $\beta_1(t)$ (dashed red) used to generate response curves $Y_i(t)$ using $X_i(t)$.

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 $X(t)$ and $Y(t)$ curves. We do this by adding realizations of the Ornstein-Uhlenbeck process, approximated by the Euler-Maruyama method, to the mean response curves calculated using the generated predictor curves and the slope and intercept functions applied to Equation (1). An example of resulting set of simulated response curves is shown in Figure 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.

We intentionally contaminated the $\beta_1(t)$ function for one of the N observations

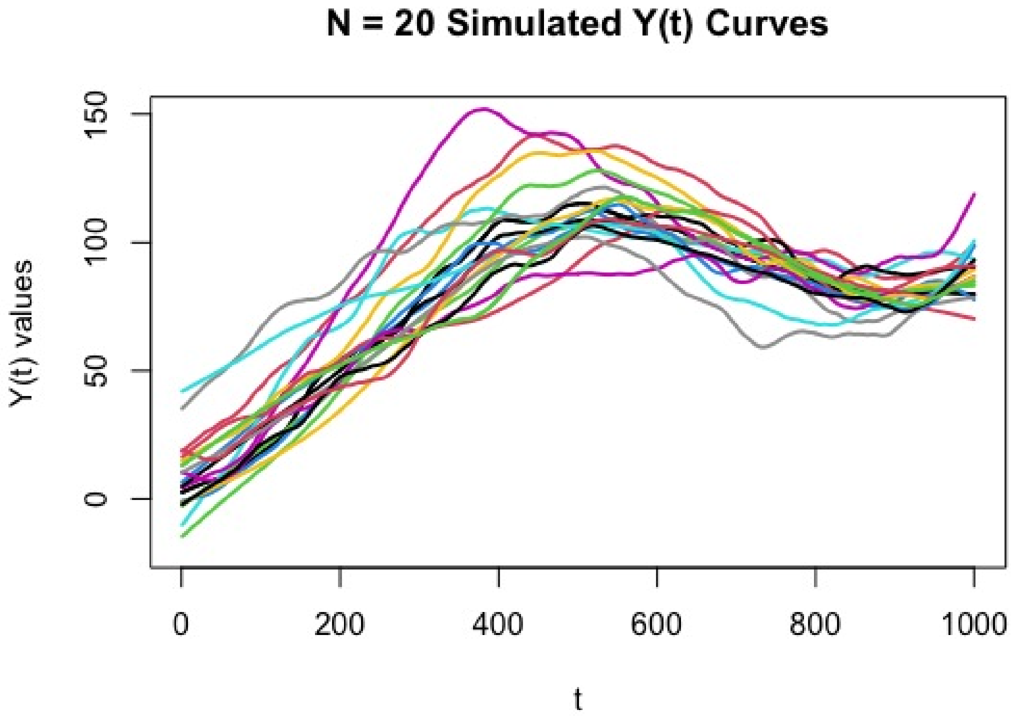

Figure 3. Example of $N = 20$ response ($Y(t)$) curves used in simulation with no contaminated observations ($\lambda = 1$).

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 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:

(1) Select λ.

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

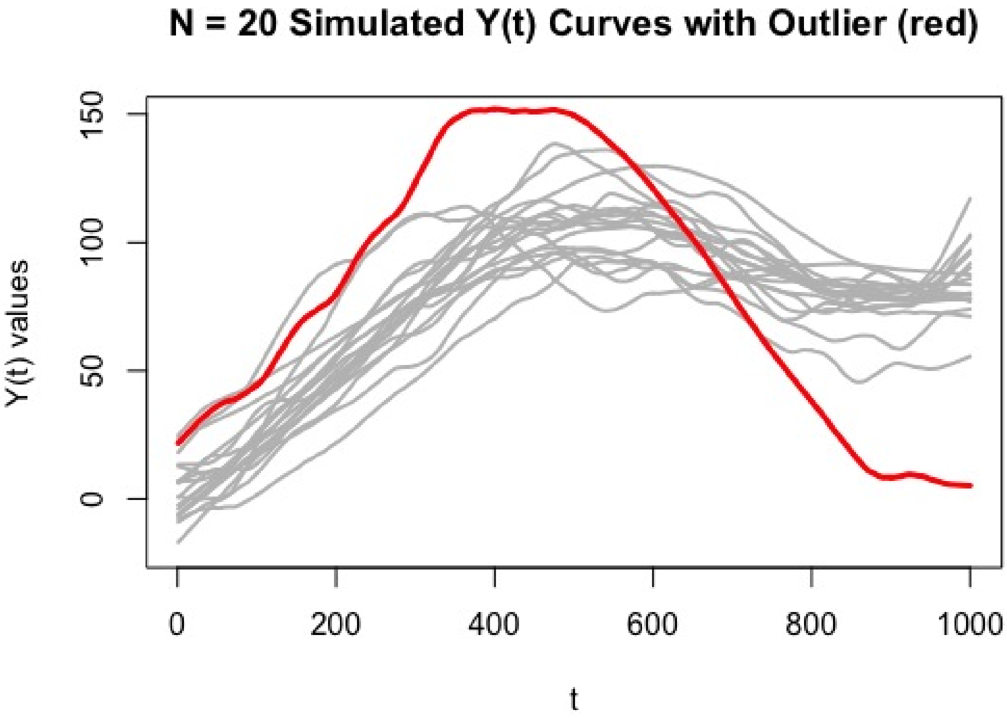

Figure 4. Example of $N = 20$ response ($Y(t)$) curves used in simulation with one outlier (red) using $\lambda = 2$.

(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) Select α and calculate the selection probabilities θ_i for each observation using Equation (9).

(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).

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

(7) Repeat this process 100 times for each combination of the 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.

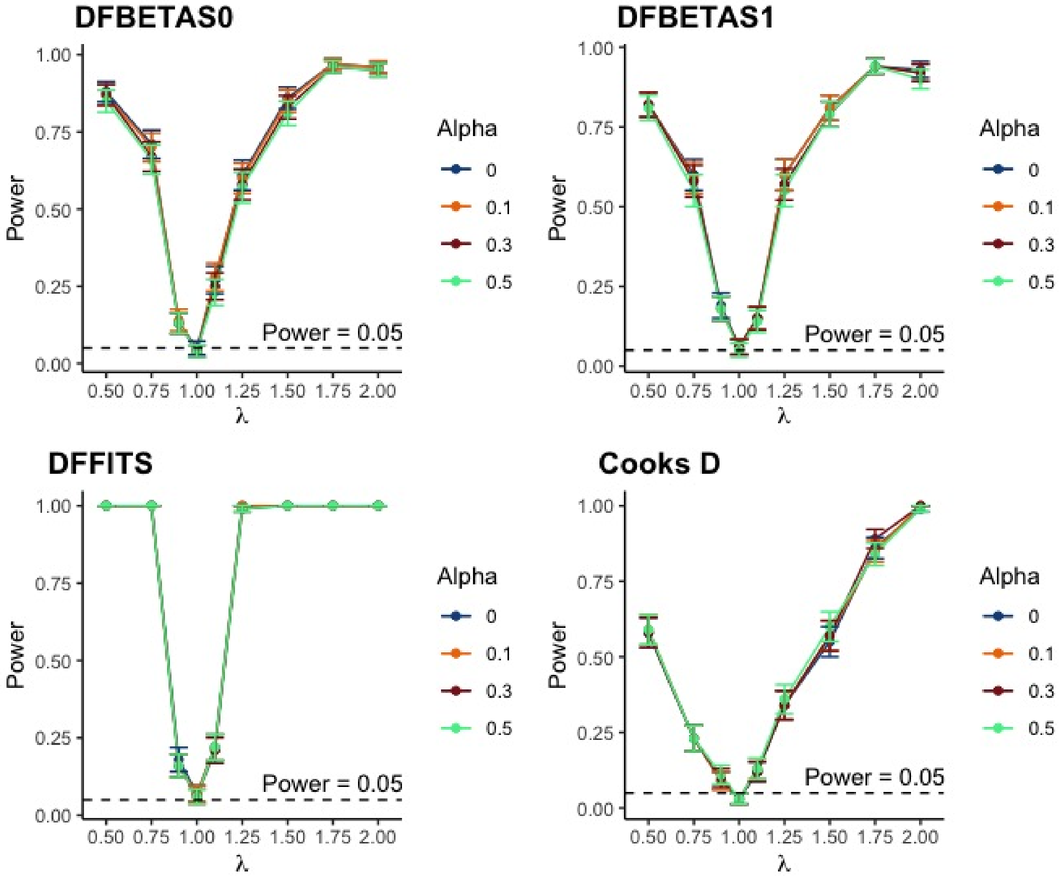

Figure 5. Power functions displaying the average proportion of contaminated observations above the 95th percentile for the four influence measures and different values of α (with error bars representing one standard error) for $N = 100$.

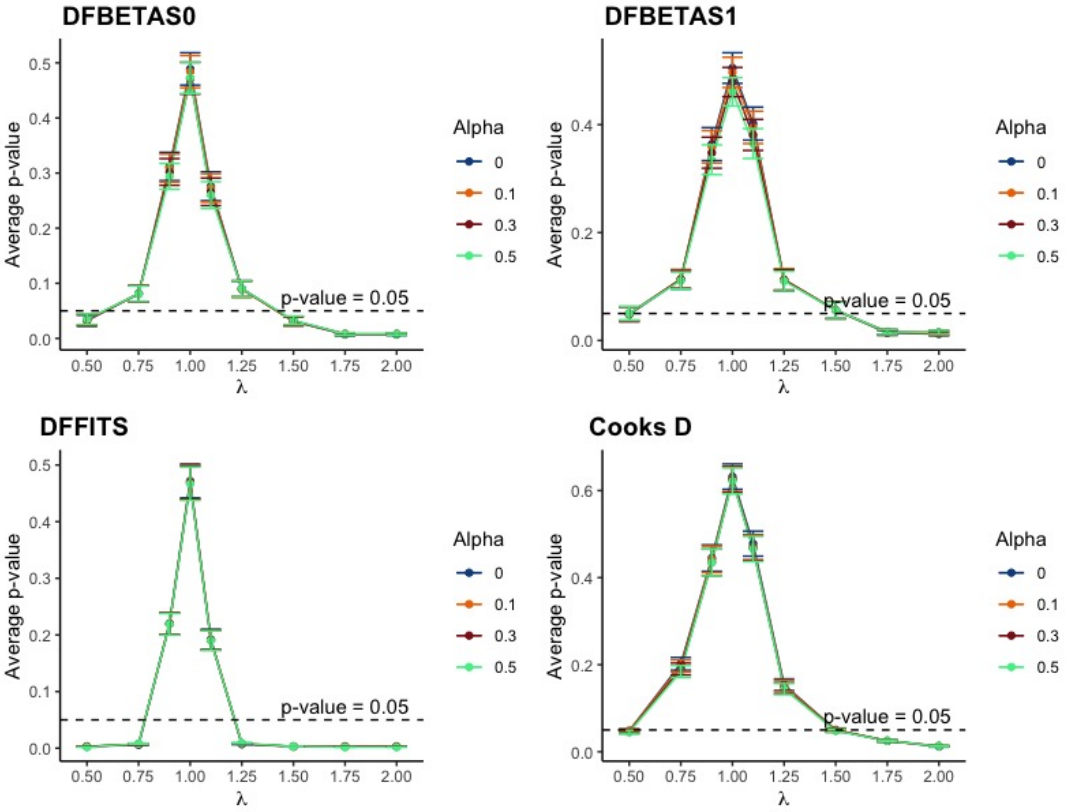

Figure 6. Average p-value (1− percentile within bootstrap distribution) of contaminated observations for the four influence measures and different values of α (with error bars representing one standard error) for $N = 100$.

Figure 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 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 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, indicating that the contaminated observation's influence measure is frequently significant. Figure 5 and Figure 6 also show that with a large sample size of $N = 100$, the effect of α is negligible.

Figure 7 and Figure 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 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 supplementary material 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

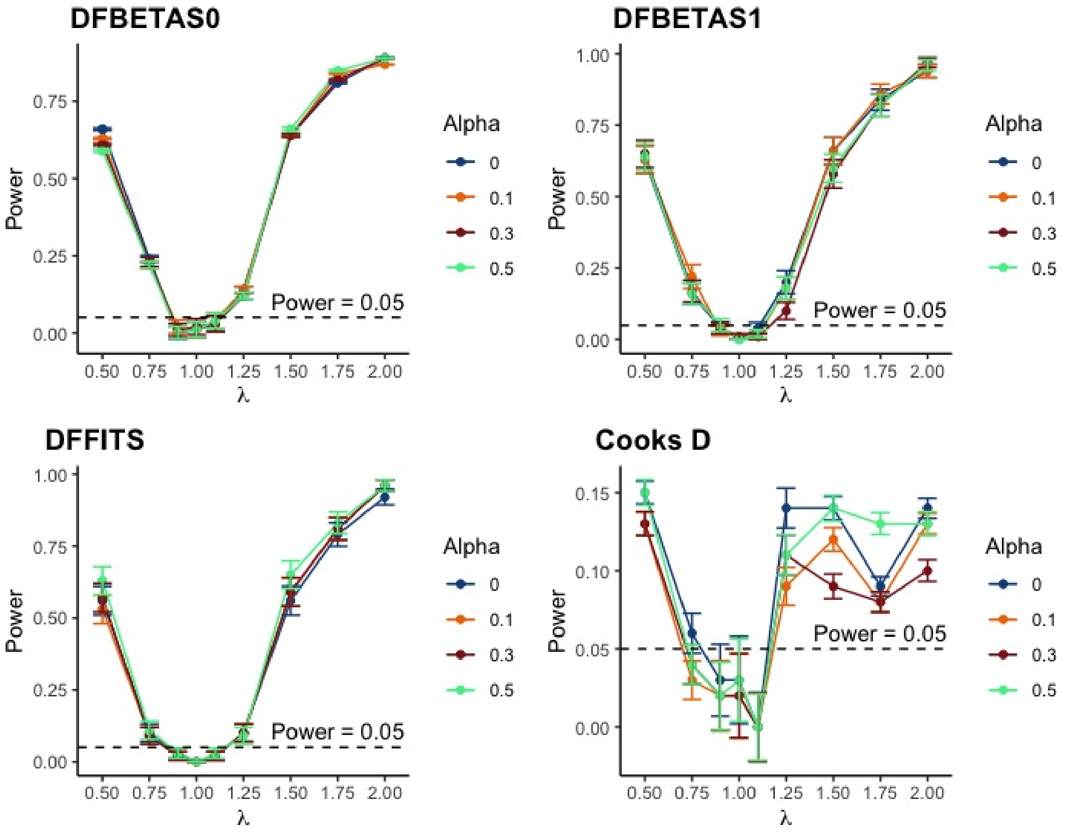

Figure 7. Power functions displaying the average proportion of contaminated observations above the 95th percentile for the four influence measures and different values of α (with error bars representing one standard error) for $N = 10$.

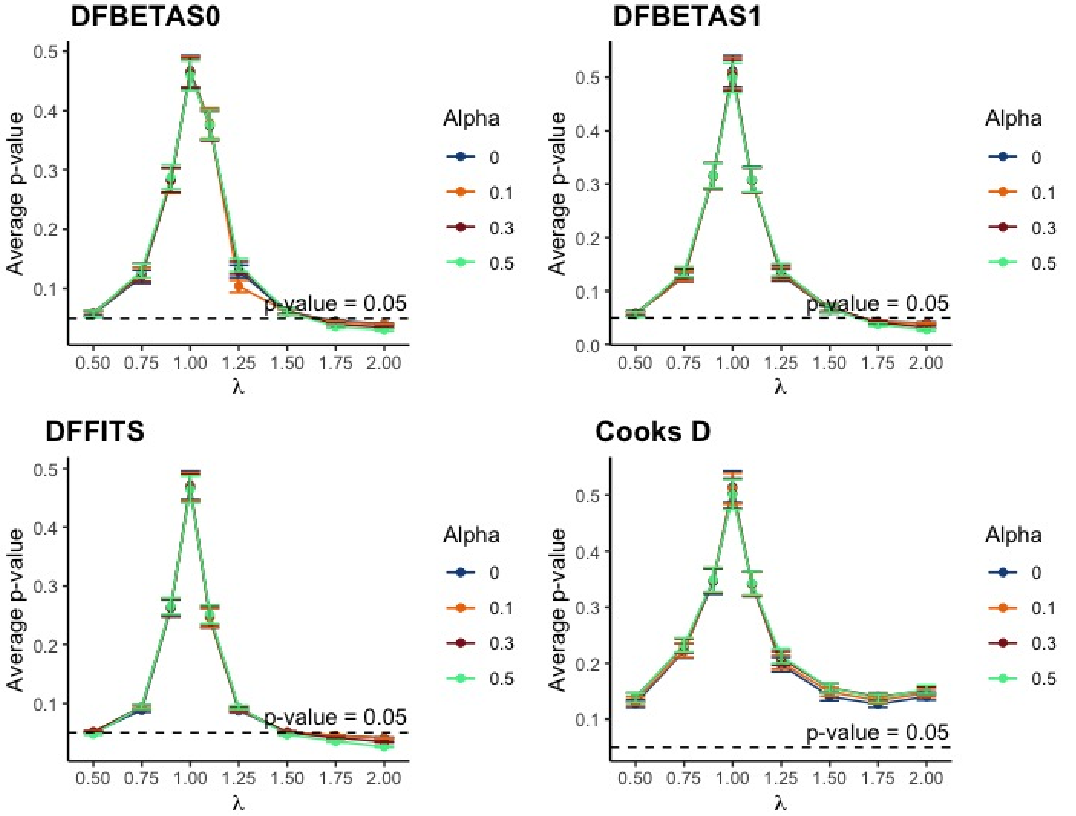

Figure 8. Average p-value (1− percentile within bootstrap distribution) of contaminated observations for the four influence measures and different values of α (with error bars representing one standard error) for $N = 10$.

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 percentiles 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 $\alpha = 0$.

5. Application: River stage data during flood events

5.1. Applying the functional influence detection to river stage data

Pittman, Hitchcock, and Grego analyzed river stages from two related gage locations at Congaree National Park near Columbia, South Carolina [11] . 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 [12] flowed over bank, through the floodplains, and into Cedar Creek [13]. 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 calculate for the river stage data is $DFBETAS_{p,i}(t)$ 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 $\beta_1(t)$ is the February 2020 flood event. Figure 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 supplementary material.

$DFBETAS_{p,i}(t)$ for the ten events is given in Figure 10. We see no obvious outlying event, and only a couple of the curves visually deviating far from the others. To determine which event has the most impact on the estimates $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$, values of $\overline{|DFBETAS_p|}_i$ of each event $i = 1, \ldots, 10$, are provided in Table 1. Most of the $DFBETAS_{p,i}(t)$ values remain within the standard threshold values in the non-functional scenario, indicating that the cutoffs used in ordinary linear regression may not be too different than those appropriate for the functional framework.

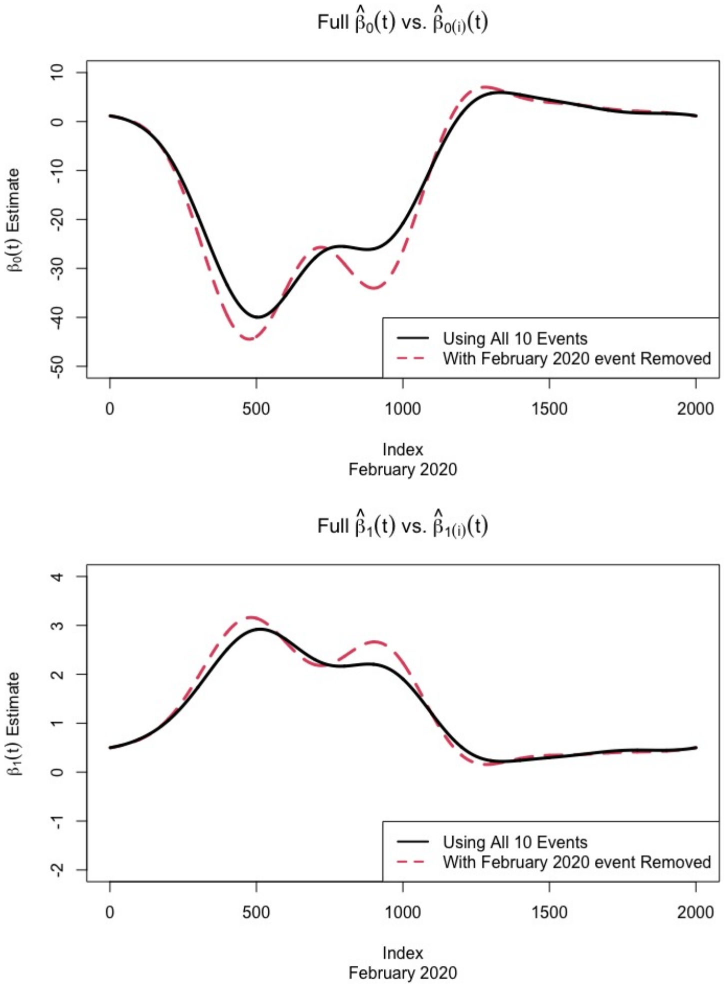

Figure 9. Comparison of $\hat{\beta}_0(t)$ and $\hat{\beta}_{0(i)}(t)$ (top) and $\hat{\beta}_1(t)$ and $\hat{\beta}_{1(i)}(t)$ (bottom) where the black solid curve represents the $\beta_p(t)$ estimate with all 10 historic flood events included and the red dashed curve is the estimate when the February 2020 event is removed.

The February 2020 and August 1995 flood events had the highest $\overline{|DFBETAS_p|}$, indicating 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 $i = 1, \dots, 10$, $DFFITS_i(t)$ measures 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 supplementary material, 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

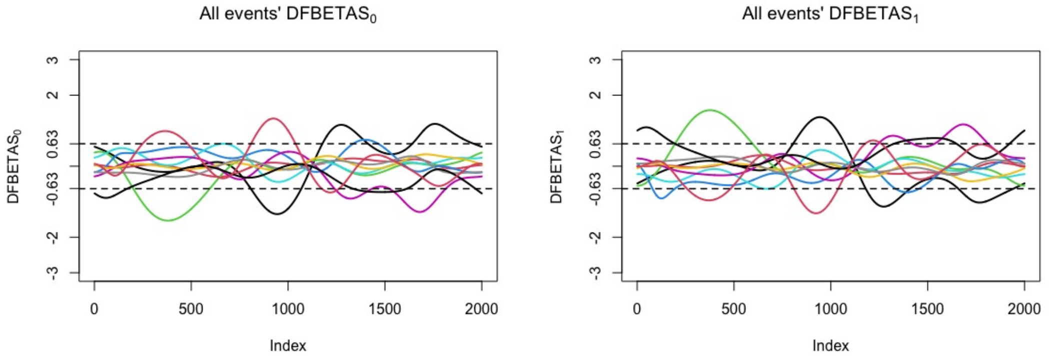

Figure 10. $DFBETAS_{(p)}(t)$ for all ten historic flood events (solid lines) with a reference of what may be considered large (dashed line) in non-functional linear regression $\pm 0.63 = \pm 2/\sqrt{N}$ for $N = 10$.

	Average Influence Measures			
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**

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

particularly flat, Figure 11 shows that $DFFITS_{10}(t)$ is the most sporadic and has the largest measurements. Table 1 provides $\overline{\lvert DFFITS \rvert}_i$ for each event (averaging across t).

Using only Table 1 without any other context, the February 2020 event had by far the highest $\overline{\lvert DFFITS \rvert}$, 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 supplementary material, but Figure 11 presents $DFFITS(t)$ for the August 1995 and February 2020 flood events, showing just how large the February 2020 event's $DFFITS(t)$ is. The large $\overline{\lvert DFFITS \rvert}$ for February 2020 is 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.

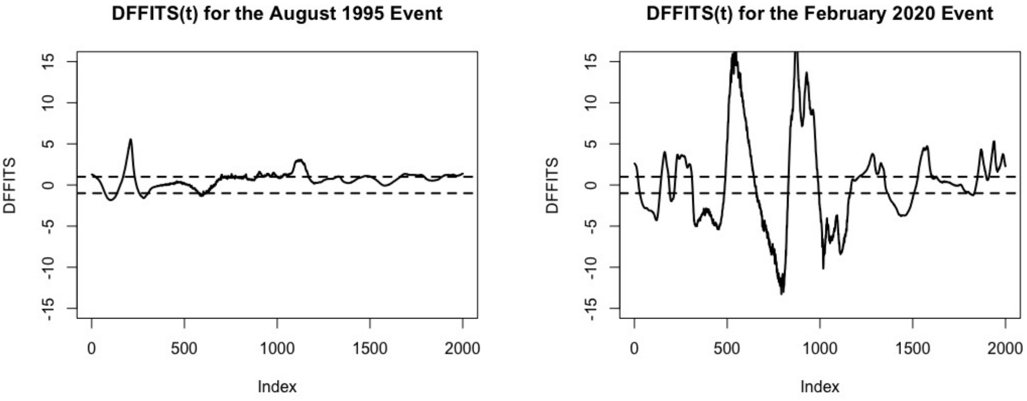

Figure 11. $DFFITS(t)$ for the August 1995 (left) and February 2020 (right) flood events (solid curve) as well as an informal cutoff line at ±1 (dashed lines).

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 plots of $D_i(t)$ are given in the supplementary material, but Figure 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 [9]. 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 1 confirms that the February 2020 flood event has the highest impact on the set of all fitted curves with $\overline{D} = 2.312$, with the next highest being only 0.436.

We calculated each of these functional influence metrics (namely functional versions of $DFBETAS_0$, $DFBETAS_1$, $DFFITS$, and Cook's distance) to determine which of the ten complete flood events used in the functional regression is the most influential. The values of each of these metrics all point to one main conclusion: The

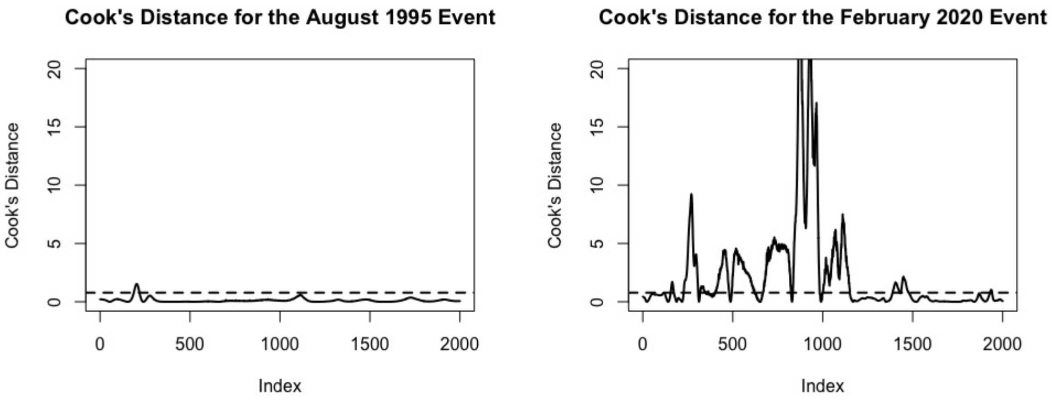

Figure 12. Cook's distance $D(t)$ for the 1995 (left) and 2020 (right) flood events, showing how influential the 2020 event is on the set of all fitted curves with a dashed line at the informal ordinary regression threshold $y = F(0.5, 2, 8) = 0.757$.

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 $\overline{|DFBETAS_1|}$. The diagnostic plots for the February 2020 event indicate 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.

5.2. Applying bootstrapping with perturbations method to river stage data

Since the average range for the ten Cedar Creek curves is 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 2).

From the table, we see the selection probability for the events with the largest $\overline{|DFFITS|}$ decreases as α is increased. For example, the February 2020 event has the

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

Table 2. The probability θ_i that each flood event is selected into the bootstrapped sample for the $\overline{|DFFITS|}$ measure using different choices of α.

largest $\overline{|DFFITS|}$, and its selection probability is about half as large when $\alpha = 0.5$ relative to when $\alpha = 0$ (equal selection probability), ensuring that event does not affect the bootstrapped percentiles unduly.

We repeat this process for each influence measure of interest, where the selection probabilities θ_i for each observation are calculated using the observed influence measure for observation i. 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.

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 shows that these averages fall 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 has the largest observed $\overline{|DFBETAS_1|}$, but it barely surpasses the 90th percentile when $\alpha = 0$ and is 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

$\overline{\vert DFBETAS_0\vert}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.535	0.570
95%	0.681	0.718
99%	0.972	1.067
Maximum observed value: 0.552 (Aug. 1995)		
$\overline{\vert DFBETAS_1\vert}$	$\alpha = 0$	$\alpha = 0.5$
90%	0.525	0.560
95%	0.652	0.694
99%	0.978	0.946
Maximum observed value: 0.527 (Aug. 1995)		
$\overline{\vert DFFITS\vert}$	$\alpha = 0$	$\alpha = 0.5$
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)		

Table 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.

that the observed $\overline{|DFFITS|}$ for the February 2020 event does have 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 falls 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 several potential reasons that the February 2020 flood event stands out as more influential than the others across many of these diagnostic measures. Of the ten events, the February 2020 flood event has the highest recorded Congaree River crest. The difference between the February 2020 Congaree crest and the next highest crest from March 2003 is greater than the difference between the March 2003 Congaree River 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.

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 [14]. 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 13, and each specific location is listed in the supplementary material.

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 14, with two specific air and water

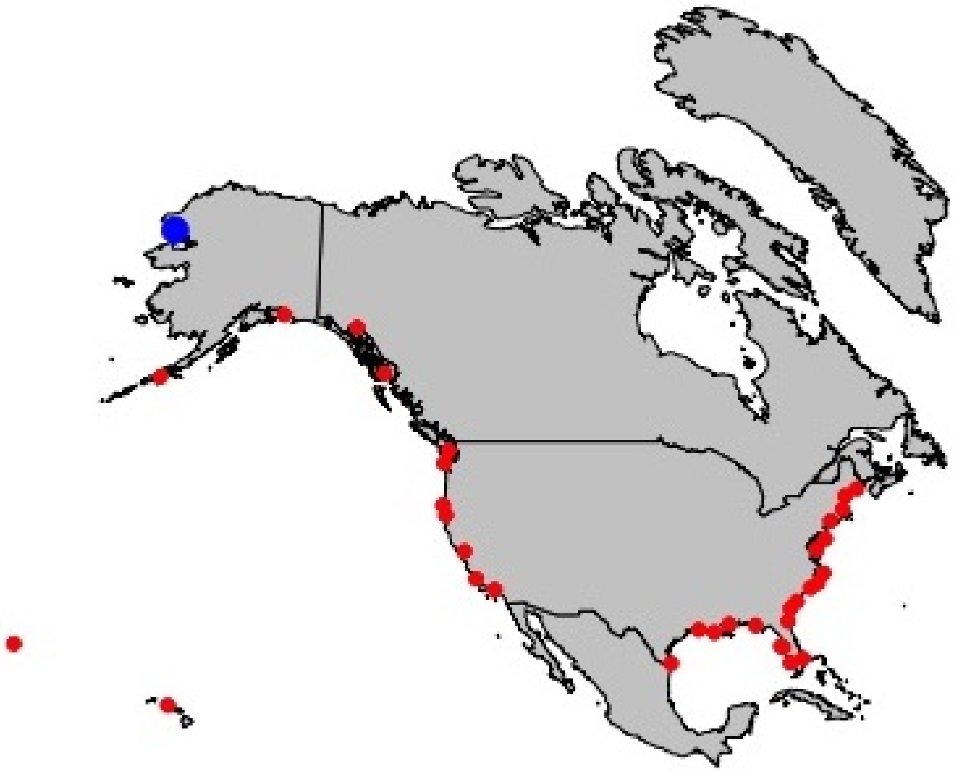

Figure 13. Exact location of each station used to create the functional regression model between air and water temperature. The map was created using the `mapproj` package in `R` [15].

temperature curves shown in Figure 15. Note that the low gold curve in Figure 14 is from Red Dock, Alaska, which is depicted with the blue dot in Figure 13.

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 (1) (estimates shown in Figure 16).

Note that the slope function $\hat{\beta}_1(t)$ is positive year-round. This indicates that no matter the time of year, as air temperature increases, water temperature also increases. This is intuitive, 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

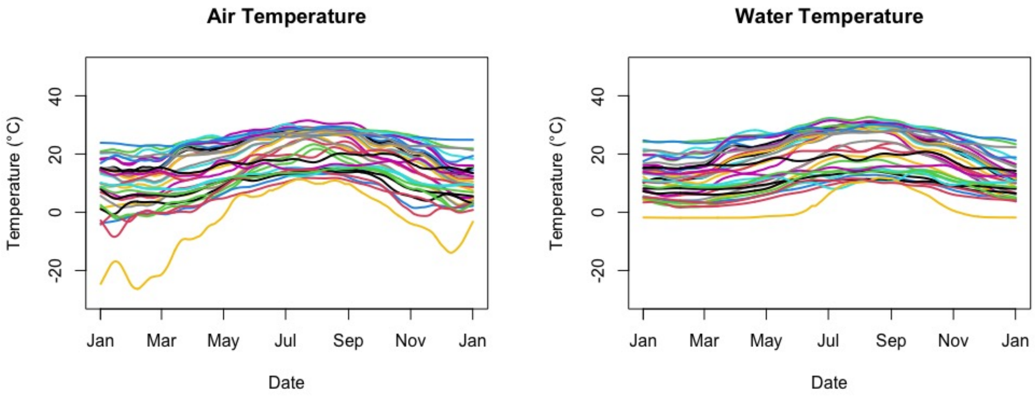

Figure 14. All 35 smoothed Air (left) and Water (right) temperatures used in the model.

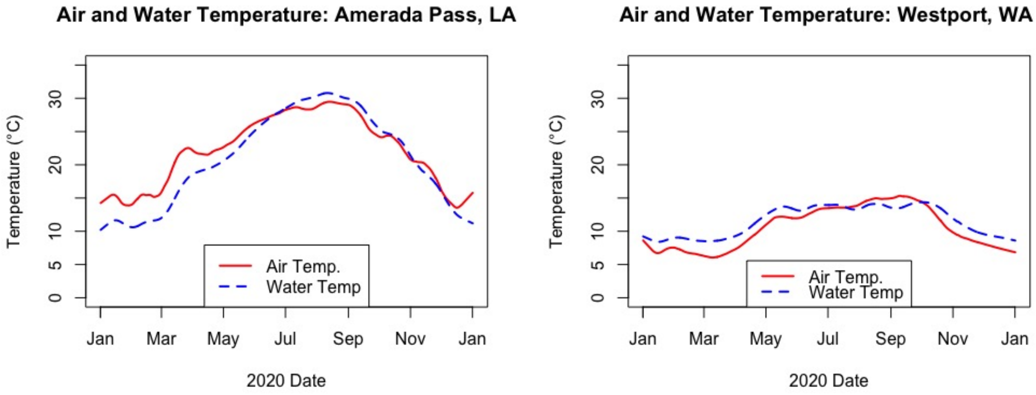

Figure 15. Air and Water Temperature for Amerada Pass, Louisiana (left) and Westport, Washington (right).

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 did confirm that a leave-one-out model does a good job of predicting the omitted functional response.

We calculated each functional influence metric ($\overline{|DFBETAS_0|}_i$, $\overline{|DFBETAS_1|}_i$, $\overline{|DFFITS|}_i$, and $\overline{D}_i$) for each of the 35 functional observations. The complete table of results can be found in the supplementary material. In general, the only observation that visually stands out as more influential than the rest is the aforementioned Red Dog Dock station in Alaska. All of its influence measures are at least twice as large as the

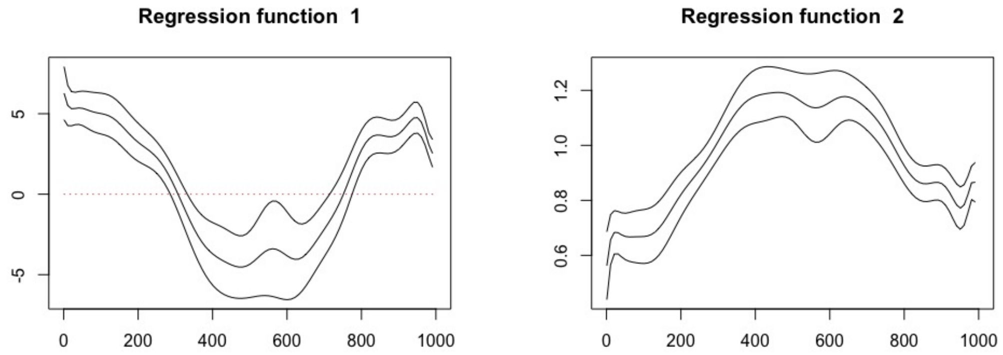

Figure 16. Estimated functional intercept $\hat{\beta}_0(t)$ (left) and estimated functional slope $\hat{\beta}_1(t)$ (right) and corresponding pointwise 95% confidence interval.

next highest, indicating that it is likely very influential on the model; our bootstrapping with perturbations method can confirm that this observation is 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 4 gives the resulting percentiles for each metric.

For every influence measure, Red Dog Dock (observation 31) is well above the 99th percentile, indicating that it is highly influential on the regression equation. This makes sense given how much lower the air temperature is at this location in the winter months compared to the rest of the observations while the water temperature is not as low, proportionally. Additionally, given how much larger the influence of this observation is compared to the rest, using the percentiles calculated using $\alpha = 0.5$ is most appropriate.

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

The Fernandina Beach location in Florida had the second highest

$\overline{\lvert DFBETAS_0 \rvert}$	$\alpha = 0$	$\alpha = 0.5$
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)		

Table 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.

$\overline{|DFBETAS_1|} = 0.286$. 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 is 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 has on the percentiles is nullified and the 99th percentile decreases 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 is above the 90th percentile even when the effect on the null distribution of the most influential observations is 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.

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.

References

[1] Belsley DA, Kuh E, Welsch RE. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Hoboken, New Jersey: Wiley Series in Probability and Mathematical Statistics; 1980.

[2] Ramsay JO, Graves S, Hooker G. fda: Functional data analysis; 2020. R package version 5.1.5.1; Available from: `https://CRAN.R-project.org/package=fda`.

[3] R Core Team. R: A language and environment for statistical computing. Vienna, Austria:

R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/.

[4] Shen Q, Xu H. Diagnostics for linear models with functional responses. Technometrics. 2007;49(1):26–33. Available from: https://doi.org/10.1198/004017006000000444.

[5] Chiou J, Müller H. Diagnostics for functional regression via residual processes. Computational Statistics and Data Analysis. 2007;51(10):4849–4863. Available from: https://www.sciencedirect.com/science/article/pii/S0167947306002465.

[6] Chen G, Huang C, Lin J. Statistical diagnostics for functional linear regression models with gaussian process errors. Communication on Applied Mathematics and Computation. 2014;28(1):118–126. Available from: https://www.researchgate.net/publication/260750220_Statistical_diagnostics_for_functional_linear_regression_models_with_Gaussian_process_errors.

[7] Febrero-Bande M, Galeano P, González-Manteiga W. Measures of influence for the functional linear model with scalar response. Journal of Multivariate Analysis. 2010; 101(2):327–339. Available from: https://www.sciencedirect.com/science/article/pii/S0047259X08002765.

[8] Cook RD. Detection of influential observation in linear regression. Technometrics. 1977; 19(1):15–18. Available from: https://doi.org/10.1080/00401706.1977.10489493.

[9] Kutner MH, Nachtsheim JC, Neter J, et al. Applied linear statistical models. 5th ed. New York: McGraw-Hill Irwin; 2005.

[10] Uhlenbeck GE, Ornstein LS. On the theory of the brownian motion. Physical Review. 1930 Sep;36:823–841. Available from: https://link.aps.org/doi/10.1103/PhysRev.36.823.

[11] Pittman RD, Hitchcock DB, Grego JM. Concurrent functional regression to reconstruct river stage data during flood events. Environmental and Ecological Statistics. 2021; 28:219–237.

[12] United States Geological Survey. Usgs 02169625 congaree river at congaree np near gadsden, sc ; 2020. https://waterdata.usgs.gov/sc/nwis/uv?site_no=02169625.

[13] United States Geological Survey. Usgs 02169672 cedar creek at congaree np near gadsden, sc ; 2020. https://waterdata.usgs.gov/sc/nwis/uv?site_no=02169672.

[14] National Oceanic and Atmospheric Administration. National data buoy center ; 2021. https://www.ndbc.noaa.gov/obs.shtml.

[15] McIlroy D, Brownrigg R, Minka TP, et al. mapproj: Map projections; 2020. R package

version 1.2.7; Available from: https://CRAN.R-project.org/package=mapproj.