Optimizing Social Climate Action: AI Assisted Reconstruction of Riverbeds for Global Climate Risk Management

Michelle Weng
Geothara Team 1A
Fall 2023, AI Studio Project

Business Context

The following is an ongoing project by Solomon Vimal and Geothara, a Cornell Tech-based startup company. Solomon founded this startup to leverage data science and satellite data to develop software that helps clients better utilize these innovative advances for applications in Water, Energy, and Climate (WEC) Security. In collaboration with the Break Through Tech AI program, Team 1A's project focused on predicting river cross sections using various supervised learning regression machine learning models.

The significance of river cross-section prediction lies in its crucial applications within flood modeling and flood insurance, especially in the broader context of climate change, a pressing issue in today's world. Climate change remains one of the most pressing issues facing our society in the 21st century. If left unaddressed, it can lead to numerous consequences for our planet, including hotter temperatures, more severe storms, species loss, and increased health risks. The impact of climate change on our river systems is especially critical because flooding can damage people's ability to earn a living. In the United States alone, there are approximately 2.7 million 2-km river segments, but only about 0.3% of these segments have gauges to measure their flow. While satellite technology has long been used to measure river widths and approximate flow rates, its use is limited with complex-shaped riverbeds.

To address various climate risks associated with water and energy, it is necessary to generate more sophisticated computer models, and key to this is data on river cross-section measurements. These measurements enable us to better understand climate risks associated with water flow and hydropower generation. The focus of the Break Through Tech AI project is to utilize the large dataset of river measurements gathered by several US state and federal initiatives in recent years. Each state have Flood Risk Information Systems (FRIS) databases that contain HEC-RAS models, which are hydraulic models used for water flow and floodplain determination. With the help of machine learning techniques, the goal is to better understand riverbeds globally, even in regions where data is scarce. For this project, we specifically looked at the states of Alabama and North Carolina. By doing so, we can address various climate risks, such as floods, droughts, energy reliability, and environmental conservation, more effectively.

**The Break Through Tech group Geothara 1A was among two undergraduate student groups assigned to assist in this project from August to December 2023. For the Geothara 1A team, our efforts were focused largely on the data generation portion of the project along with some machine learning experimentation. Our journey and methodologies are detailed below.

Data

The following two links are the geometry (.g0) data files for the two states. For time purposes, our team looked only at the Alabama and North Carolina datasets, but it is of interest to note that the company aims to compile a clean, nation-wide dataset with river cross section information.

Alabama:
https://drive.google.com/file/d/1gZn7UcLSTJ5vuR_TgF9b8J2ciYrmTgOW/view?usp=drive_link
North Carolina:
https://drive.google.com/file/d/1ykvOsiXZP6FDQXg8zu56IE6II9Q5arD8/view?usp=drive_link

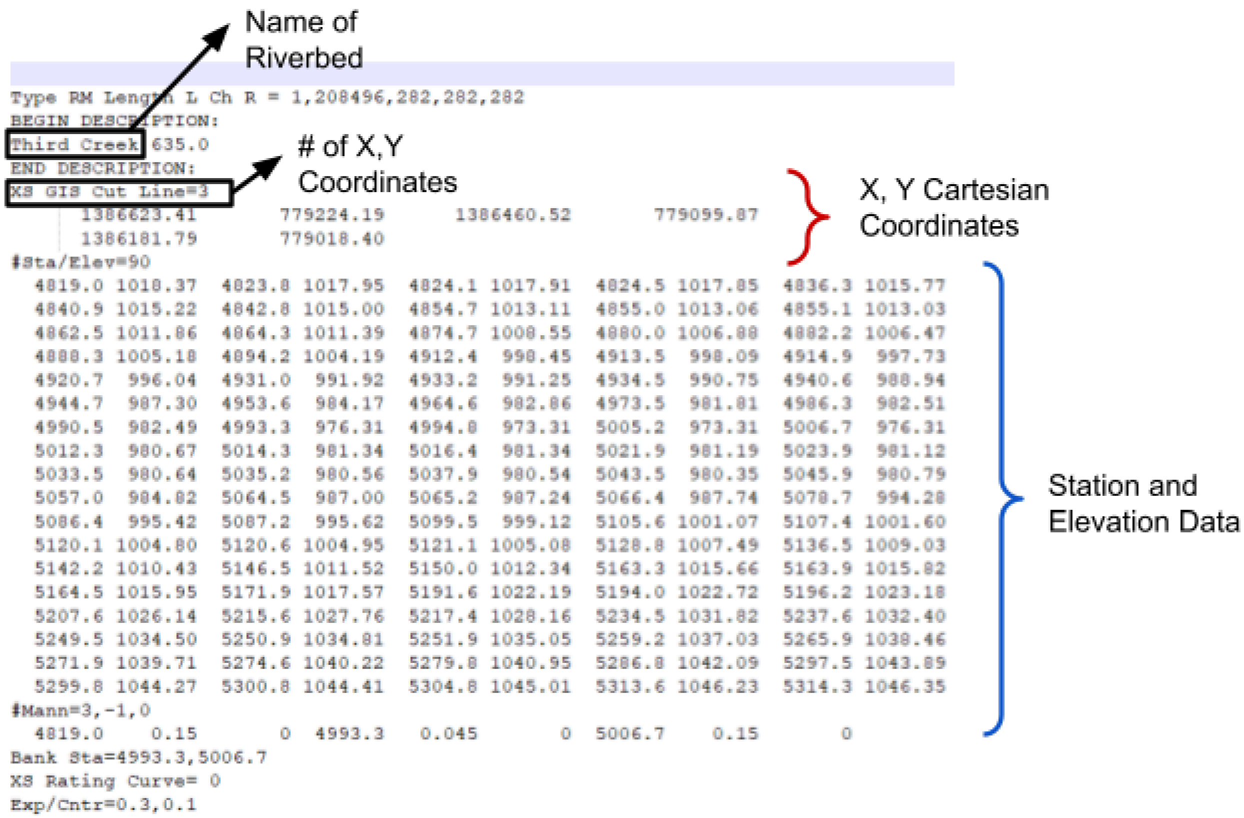
G01 or geometry files contain geometric data of the river system being analyzed. For river cross sections, this includes river depth, length, etc.They mostly belong to HEC-RAS, a downloadable software. These files are text files, so they contain words and text and can be viewed with any text editor such as Windows Editor, Nano for Linux, and TextEdit for macOS.

Data Understanding and Preparation

For data preparation, the main task was to parse through the raw data. The raw data consisted of separate files for each river for each county for each state. Each file for a specific river would contain many tables in its file, which represent the cross sections of the river.

(X,Y) and (Station, Elevation)
We needed to parse the necessary data and ignore unnecessary portions. The raw river cross section data is essentially made up of two types of coordinates: X and Y cartesian coordinates; Station # and Elevation

coordinates. To make the data into a more utilizable format, the extracted points were made into list objects, making them easily accessible through indexing. The first index allows access to a specific river cross section. A second index allows access to either the x,y vertices of the cut or the station/elevation points. An example of a specific cross section in a G01 file may be seen in Figure 1.




```
Type RM Length L Ch R = 1,208496,282,282,282
BEGIN DESCRIPTION:
Third Creek 635.0
END DESCRIPTION:
XS GIS Cut Line=3
      1386623.41       779224.19       1386460.52       779055.87
      1386181.79       779018.40
#Sta/Elev=90
  4819.0 1018.37  4823.8 1017.95  4824.1 1017.91  4824.5 1017.85  4836.3 1015.77
  4840.9 1015.22  4842.8 1015.00  4854.7 1013.11  4855.0 1013.06  4855.1 1013.03
  4862.5 1011.86  4864.3 1011.39  4874.7 1008.55  4880.0 1006.88  4882.2 1006.47
  4888.3 1005.18  4894.2 1004.19  4912.4  998.45  4913.5  998.09  4914.9  997.73
  4920.7  996.04  4931.0  991.92  4933.2  991.25  4934.5  990.75  4940.6  988.94
  4944.7  987.30  4953.6  984.17  4964.6  982.86  4973.5  981.81  4986.3  982.51
  4990.5  982.49  4993.3  976.31  4994.8  973.31  5005.2  973.31  5006.7  976.31
  5012.3  980.67  5014.3  981.34  5016.4  981.34  5021.9  981.19  5023.9  981.12
  5033.5  980.64  5035.2  980.56  5037.9  980.54  5043.5  980.35  5045.9  980.79
  5057.0  984.82  5064.5  987.00  5065.2  987.24  5066.4  987.74  5078.7  994.28
  5086.4  995.42  5087.2  995.62  5099.5  999.12  5105.6 1001.07  5107.4 1001.60
  5120.1 1004.80  5120.6 1004.95  5121.1 1005.08  5128.8 1007.49  5136.5 1009.03
  5142.2 1010.43  5146.5 1011.52  5150.0 1012.34  5163.3 1015.66  5163.9 1015.82
  5164.5 1015.95  5171.9 1017.57  5191.6 1022.19  5194.0 1022.72  5196.2 1023.18
  5207.6 1026.14  5215.6 1027.76  5217.4 1028.16  5234.5 1031.82  5237.6 1032.40
  5249.5 1034.50  5250.9 1034.81  5251.9 1035.05  5259.2 1037.03  5265.9 1038.46
  5271.9 1039.71  5274.6 1040.22  5279.8 1040.95  5286.8 1042.09  5297.5 1043.89
  5299.8 1044.27  5300.8 1044.41  5304.8 1045.01  5313.6 1046.23  5314.3 1046.35
#Mann=3,-1,0
  4819.0    0.15       0  4993.3   0.045       0  5006.7    0.15       0
Bank Sta=4993.3,5006.7
XS Rating Curve= 0
Exp/Cntr=0.3,0.1
```

Figure 1.

Let us say the following is a cross section within an arbitrary River A .g0 data file (Figure 2). This is a simple example to illustrate the data preparation process, and it is similar to the real example in Figure 1 since the cross section is formed by 3 X,Y coordinates.

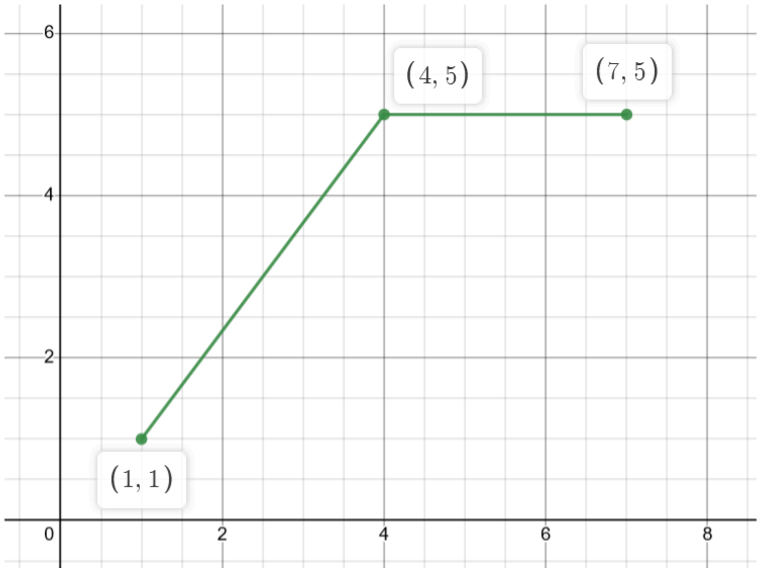


Figure 2.

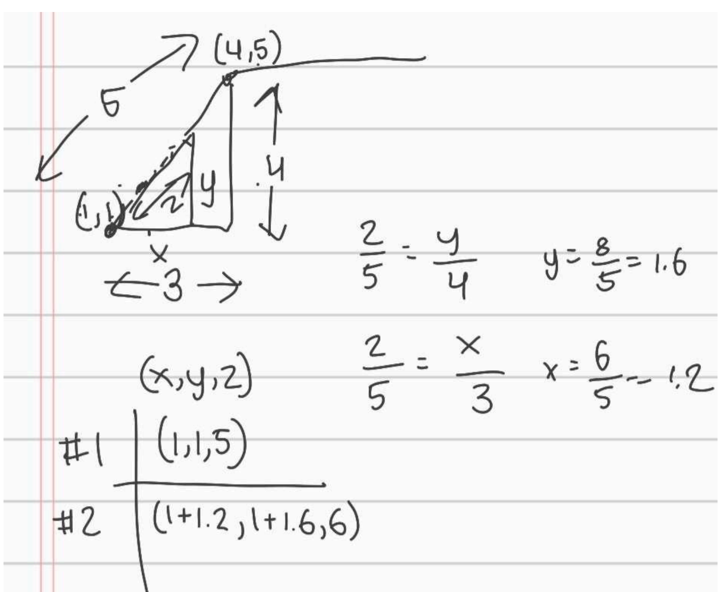
As seen in the figure here, the points in green are joined together by straight lines to represent a particular cross section for our hypothetical river. It is made up of three coordinates, so two “cut lines”. The first cut is from (1,1) to (4,5) and the second cut is from (4,5) to (7,5). In the real scenario, the number of cuts that make up a cross section is indicated by the string XS GIS Cut Line= (Refer again to Figure 1).

Now, let’s also say that the station elevation pairs are given as a list from the parsing task: [[15, 5], [17, 6], [19, 7], [21, 8], [23, 8]].

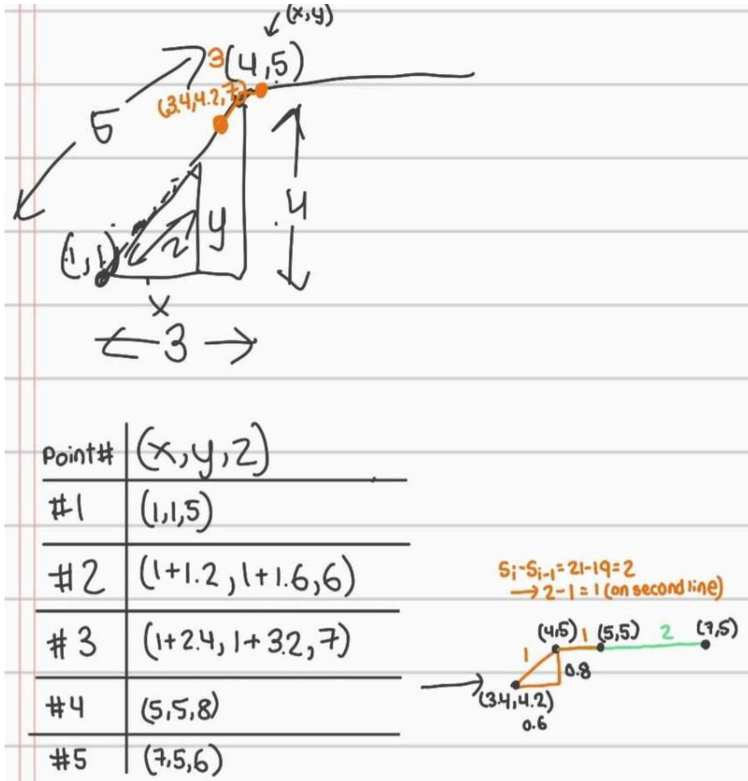
In each station elevation pair, the station basically indicates how much distance you walk along the cross section and the elevation value indicates the depth (z in xyz) at that point.

So assuming that the start of our cross section is the point (1,1) (Note (7,5) is also viable), our first [x,y,z] list will be: [1, 1, 5]

By using triangle ratios or calculating the percentage of the way along the cut, the consecutive [x,y,z] points were calculated. We know from mathematics that for a linear line on the x-y plane, you can calculate hypotenuse from the x and y coordinates. Also, as you can see, the next station and elevation



pair is [17,6]. 17-15=**2** which means **we move 2 units along the cross section. This means we're at ⅖ (40%) along the first cut line.** See pictures of the work if visualization is needed. I used triangle ratios and got 1.2 for X and 1.6 for Y. By adding it to the previous X,Y coordinate, we get [2.2, 2.6, 6].



You can definitely work the rest of the example easily by yourself if you'd like, but it is better to let the computer do it since we figured out the mechanism. See code:

```
simple_ex = [[[1,1],[4,5],[7,5]],[[15, 5], [17, 6], [19, 7], [21, 8], [23, 8]]]
simple = Task_2(simple_ex)
print(simple)
```

Output:

```
[(1.0, 1.0, 5), (2.2, 2.6, 6), (3.4, 4.2, 7), (5.0, 5.0, 8), (7, 5, 8)]
```

As demonstrated, we verified our code could work by using this simple example before using it on an actual cross section.

The last task was repeating this task for all cross sections for all rivers in a state for both states. For visualizing in QGIS, the [x,y,z] structures were converted into shape (.shp) files using the python libraries Fiona and Shapely.

Entire code for data prep can be seen in our colab:
https://colab.research.google.com/drive/1hx9fK3m1FoIG2bnPPH-NKaA49xYadvQc?usp=drive_link

Visualization

Figure 3.

Figure 3 shows the QGIS generated image of the North Carolina river cross sections led by Vladislav Vostrikov (Hunter College) and my efforts in data prepration, with support from Ryan Chen and Sabreen Shigri (both from Stony Brook). The cross sections show the shape of the rivers when viewed in birds eye. Having generated a uniform data representation for all the G01 files, the next modeling and validation phase of the project was largely led by Carla Sanchez and Vladislav Vostrikov (both from Hunter College).

Modeling and Validation

Our project focused on predicting river cross sections using Random Forest, Decision Tree, and KNN regression models. Before running each model, we evaluated the best hyperparameter settings using GridSearchCV. To simplify the model, solely x and z coordinates were used, resulting in plots looking like the one below in Figure 4. In this graph, the horizontal axis depicts the x coordinates after normalization and the vertical axis represents elevation.

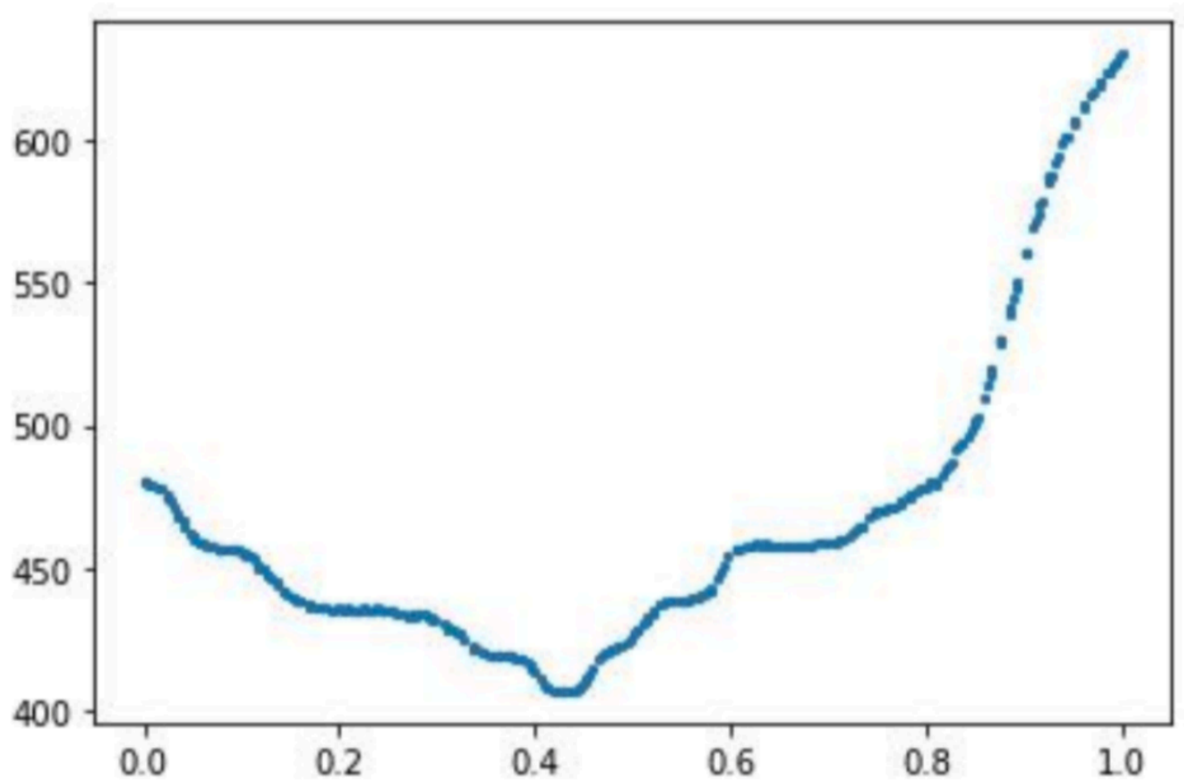


Figure 4.

To create our river cross-section models, the team experimented with different-order polynomials. However, higher-degree polynomials perform poorly when it comes to generalization, an indication of overfitting. Second-degree polynomials seemed to underfit the data. Thus, our final solution was to carry out a combination of third-order polynomials. When combining two of them, we obtained accuracy scores ranging from -0.802 to 0.986. The fittings can be seen in Figure 5.

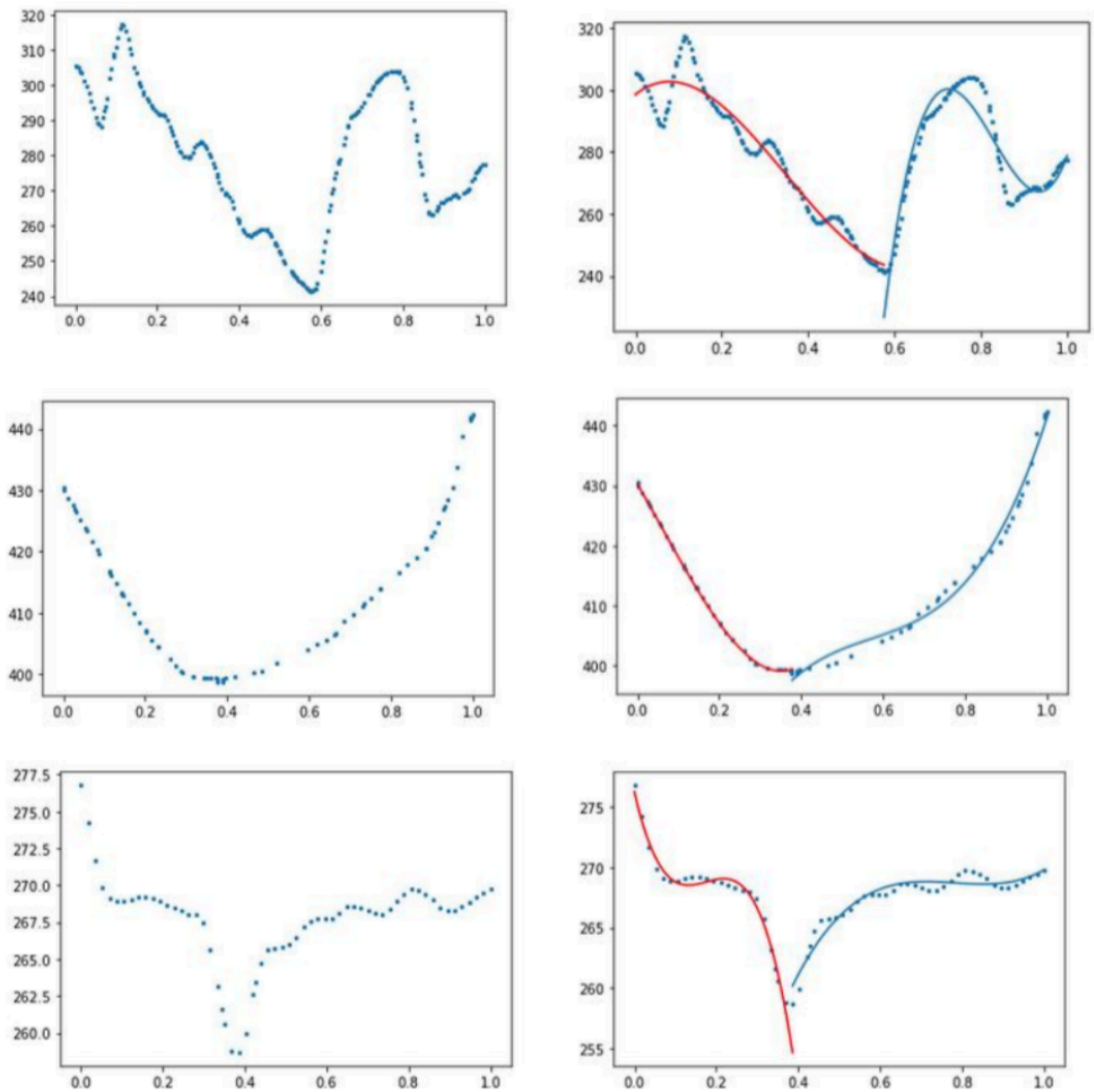


Figure 5.

Lessons Learned

There were many lessons learned in this project. First, the importance of our project in climate change solutions. The future directions for the project will be used in a very relevant problem for this generation. I also learned about the HEC-RAS and QGIS softwares. In HEC-RAS, I could open .g0 files to analyze/make sense of the raw data. At the same time, I can see what G01 files looked like in the Text Editor. In QGIS, I was able to open .shp files to see the direct results of the compilation of [x,y,z] cross section points using our Python data preparation code.

At the same time, there were also many questions. For instance, about confluence points, which occurs when two or more flowing bodies of water join together to form a single channel. My team mapped the confluence points, which are important points in river flow modeling, but we did not end up getting too deep into this aspect. It would be interesting to predict the flow of rivers because floods not only depend on the shape and depth of the rivers but also other geometric factors like confluence points and hydraulic factors like velocity.

Regardless, the data our team generated for North Carolina and Alabana was a good starting point for analysis for a complex problem. Our Challange Advisor, CEO Solomon Vimal, is also continuing to work on this project and finding potential partnerships in the water, energy, and climate security.

Acknowledgements

Thank you to CEO Solomon Vimal, the Break Through Tech AI program, and members of my team Carla Sanchez, Ryan Chen, Sabreen Shigri, and Vlad Vostrikov for a fun and educative journey. This project wouldn’t have been possible without their collaboration and support.

If you would like more information about the Break Through Tech AI program, you can visit breakthroughtech.org.