New England River Analysis Process Book

John Ryu, Nathaniel Wai, Charles Liu

Initial Ideas from our Project Proposal Stage (2/28/25):

Overview:

We want to figure out which factors are the greatest predictors of river discharge. These factors include precipitation or river discharge of previous weeks. Based on these predictions, we want to visualize which rivers are at the greatest risk of flood and also allow for users to explore each river's past history of precipitation and discharge and also allow the user insights into our prediction model such as feature importance and also the efficacy of our model for a given prediction.

Motivation:

We wanted to explore the applications of machine learning on environmental problems and we already had some data about river discharge. More specifically, discharge prediction can help predict extreme conditions which is useful for forecasting floods and drought-like conditions.

Related Work:

Charles' advisor (Dr. Randhir) suggested an ML model to predict discharge, using discharge from 1 week ago, 2 weeks ago, and precipitation today, 1 week ago and 2 weeks ago as inputs. We were curious to see the extent to which an ML model could predict discharge simply based on past observations and also learning how to visualize elements such as flood risk spatially with a map.

Questions:

We seek to answer the following questions with our data visualization:

- Which factors are greatest predictors of river discharge?
- What are the forecasted discharges of rivers in New England, given their predictors (such as precipitation from previous weeks and discharge from previous weeks) in the dataset? (The Outcome of the Model)
- Which rivers in New England are at risk of severe conditions (drought or flood)?
- How can we display each New England river's past precipitation and discharge?
- How can we evaluate our model and figure out which rivers it is the most effective at predicting discharge for?

Data:

Our sources included the U.S. Geological survey for historic gauge discharge data and PRISM for precipitation and temperature data. We directly downloaded csv data from these sites and combined these files together into two csv files, one for static data such as gage point latitude and longitude and one for dynamic data such as the amount of discharge at a gage point on a given date.

Ideas leading to our Milestone (4/13/25):

Overview:

At this stage of development, our general vision for our data visualization began to deviate slightly. We realized that our goals of implementing an ML model on top of designing and implementing our data visualizations may have been too ambitious and we decided instead to focus more on visualizing data trends rather than training a model to predict river discharge and visualizations for analyzing the quality of a river model. As a result, instead of relying on a model, users would be required to perform their own analysis, however we also began tailoring our visualizations to allow users to do so in an intuitive manner.

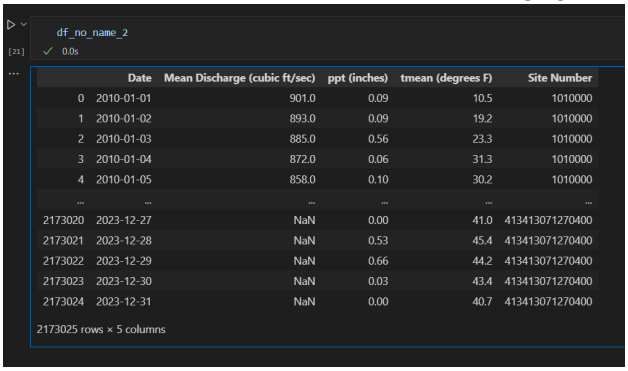
Motivation:

At this stage of development, our motivations changed alongside our initial visions for our data visualization. Instead of exploring the applications of machine learning on environmental problems, we wanted to learn the best types of visualizations to use to help users analyze individual gage points across different time periods.

Exploratory Data Analysis:

We looked through the dataset using a pandas dataframe in a python notebook. We learned that our dataset is too large to host on Supabase so we considered alternatives such as simply using python in our backend to store the data in a pandas dataframe and also preprocessing the dataset to remove unnecessary columns.

We further organized our data by allocating elevation, site name, latitude and longitude to a different table as these are all static values that correspond to a gage point.



```
df_no_name_2
[21] ✓ 0.0s
```

	Date	Mean Discharge (cubic ft/sec)	ppt (inches)	tmean (degrees F)	Site Number
0	2010-01-01	901.0	0.09	10.5	1010000
1	2010-01-02	893.0	0.09	19.2	1010000
2	2010-01-03	885.0	0.56	23.3	1010000
3	2010-01-04	872.0	0.06	31.3	1010000
4	2010-01-05	858.0	0.10	30.2	1010000
...	...	...	...	...	...
2173020	2023-12-27	NaN	0.00	41.0	413413071270400
2173021	2023-12-28	NaN	0.53	45.4	413413071270400
2173022	2023-12-29	NaN	0.66	44.2	413413071270400
2173023	2023-12-30	NaN	0.03	43.4	413413071270400
2173024	2023-12-31	NaN	0.00	40.7	413413071270400

2173025 rows × 5 columns

The above screenshot is taken from our python notebook, and showcases our dynamic table with unnecessary columns removed, cleaned up for demonstrative purposes using pandas library in python. We organized data into two different tables, one for static data such as gage point latitude and longitude and one for date-based data such as mean discharge.

Questions:

Of the questions we asked ourselves initially, we decided to appropriately limit the scope of our project and focus on only one of them:

- How can we display each New England river's past precipitation and discharge?

From there, we asked ourselves a few additional questions to inform our design process:

- Which visualizations would be the most useful for allowing users to analyze gage points?
- Which river analysis task can we most effectively support with our visualization?
- What type of data analysis should we encourage? Should we focus more on comparison across different gage points or individual analysis?

Design Evolution:

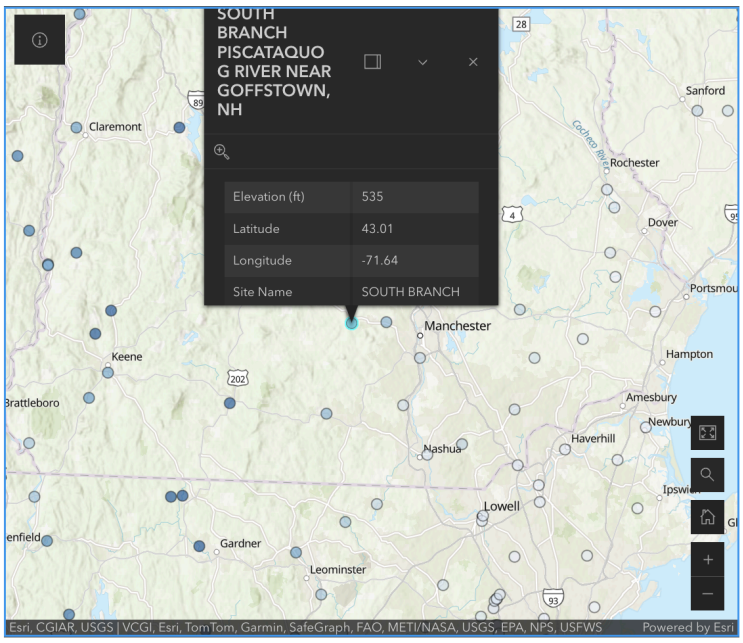
While developing our table component, we brainstormed the best ways to organize the table. We considered displaying every gage point by default then allowing users to add multiple filters, such as state and elevation. We ultimately decided it'd be best for us to organize gage points by state, and allow users to analyze states one at a time, which we felt provided the user with an adequate amount of control over their search process without being overbearing.

In terms of laying out our multi-view visualization, we settled on a layout of map in the top left, graph in the bottom left and table on the right, simply because we felt that the map and the graph were the shortest visualizations we would be implementing and it made the most sense to us to stack these two elements vertically to minimize whitespace.

Implementation:

In the following section, we will provide an overview of each of the interactive components we are developing for our data visualization:

Interactive Map:



This is our interactive map, which allows for users to explore the locations of gauge points spatially and click on them to gather basic information such as elevation, latitude, longitude, etc. Users can also search for specific locations and enter full screen mode for easier viewing.

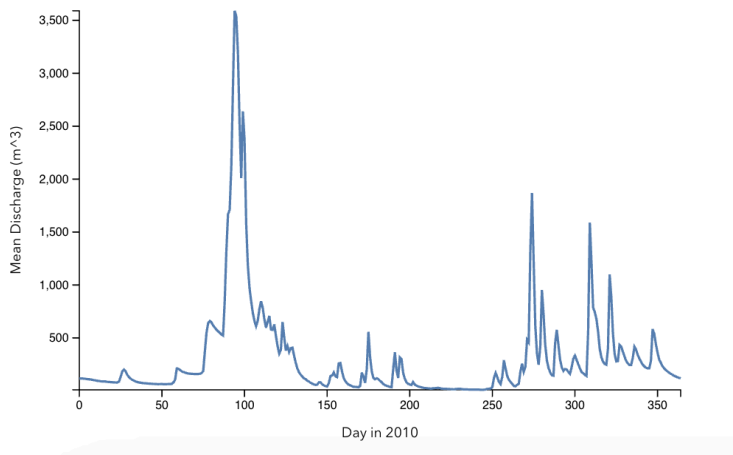
Table:

State: Massachusetts

site_number	state	latitude	longitude	elevation	site_name
1094400	Massachusetts	42.57620079	-71.7881285	518	NORTH NASHUA RIVER AT FITCHBURG, MA
1094500	Massachusetts	42.49506389	-71.7219333	348	NORTH NASHUA RIVER NEAR LEOMINSTER, MA
1095220	Massachusetts	42.41092507	-71.7911829	489	STILLWATER RIVER NEAR STERLING, MA
1095375	Massachusetts	42.37286967	-71.8281279	620	QUINAPOXET RIVER AT CANADA MILLS NEAR HOLDEN, MA
1095434	Massachusetts	42.36453675	-71.775349	509	GATES BROOK NEAR WEST BOYLSTON, MA
1095503	Massachusetts	42.41944444	-71.6661111	364	NASHUA RIVER, WATER STREET BRIDGE, AT CLINTON, MA
1096000	Massachusetts	42.63425619	-71.6578479	351	SQUANNACOOK RIVER NEAR WEST GROTON, MA
1096500	Massachusetts	42.66758945	-71.57506809	243	NASHUA RIVER AT EAST PEPPERELL, MA
1097000	Massachusetts	42.432038	-71.4497848	207	ASSABET RIVER AT MAYNARD, MA
1097300	Massachusetts	42.51259289	-71.40422829	207	NASHOBA BROOK NEAR ACTON, MA
1098500	Massachusetts	42.31514167	-71.3838083	167	COCHITUATE BK BL LAKE COCHITUATE AT FRAMINGHAM, MA
1098530	Massachusetts	42.3253732	-71.3975605	187	SUDBURY RIVER AT SAXONVILLE, MA

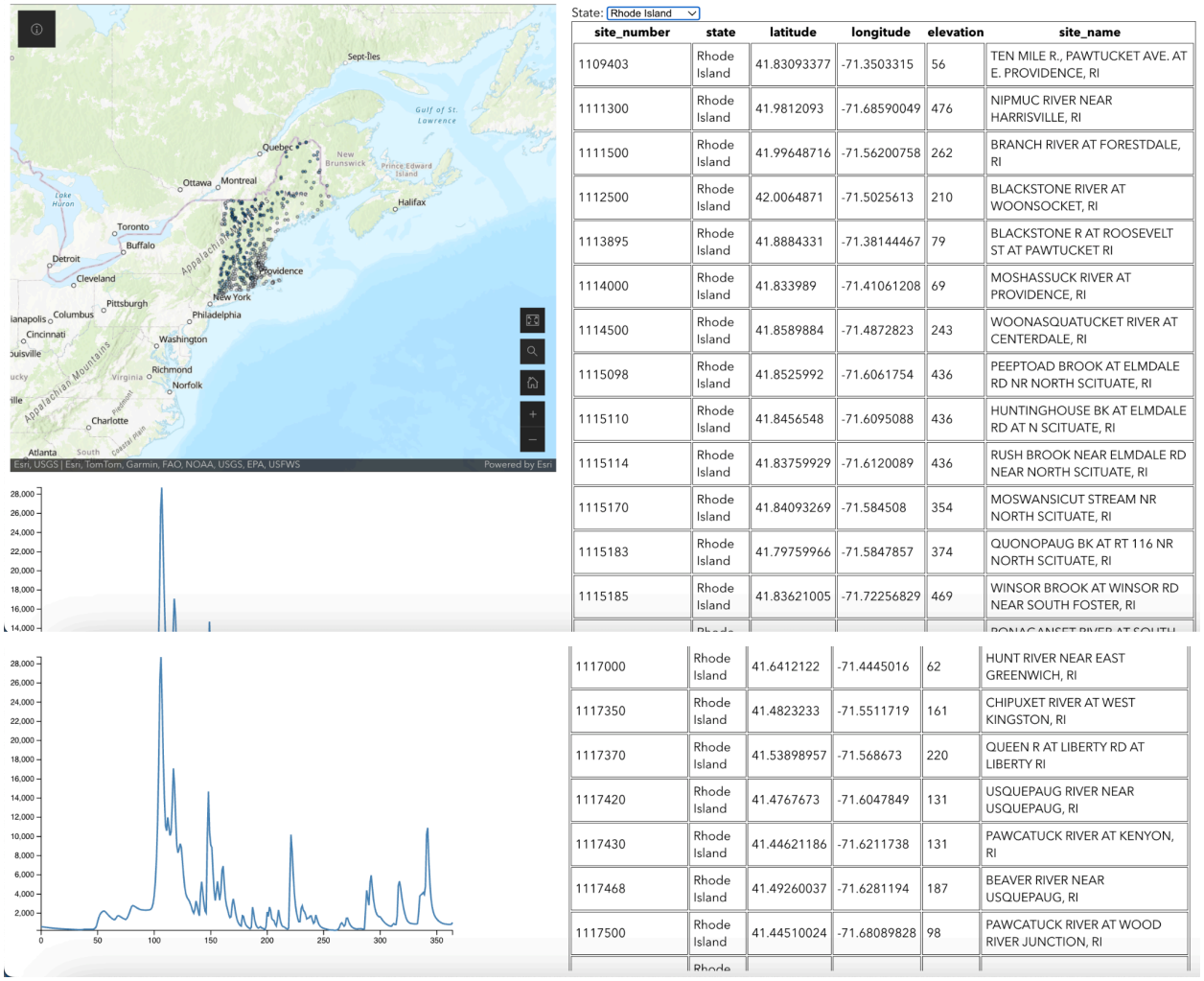
Our table allows users to search for all gauge points within the selected state, and view basic information such as site number, latitude, longitude, etc. This is meant to allow for a quicker but less intuitive method to gather information about a specific gauge point.

Discharge over time line graph:



Our line graph depicts the amount of mean discharge for dummy data over the course of 2010. This will later be linked with the table such that a gauge point can be selected and subsequently a date range can be selected instead of being hard-coded. This visualization is meant for analyzing rivers' patterns of discharge and will later be revamped to visualize other patterns such as precipitation over time.

Data visualization overview:



The above screenshots show our visualizations in their current state. In the top-left, we have an interactive map which allows users to see the locations of gauge points spatially and gather basic information, such as elevation, longitude and longitude, by clicking on certain gauge points. On the right we have a filter selector, which allows users to see all gauge points in a given state. On the bottom left we have a sample graph visualization precipitation at a predetermined gauge point at a specified date range.

In future iterations, we will link this graph to the table to allow for the user to select a specific gage point's data to visualize and we will also add a date range selector component onto the website.

Evaluation:

Since our visualizations are still in development, not much can be analyzed about the rivers in our website's current state. Our visualizations are effective for basic functions, however we still

need to build upon them more by adding linking and adding more interactive components such as allowing users to define a date range and a gage point to link the graph to.

Ideas leading to our Final Submission (5/12/25):

Overview:

We finalized our decision to focus on visualizing data trends rather than training a model to predict river discharge and providing visualizations that allow the users to explore individual gage points and analyze their precipitation and discharge over time.

Motivation:

We wanted to learn the best types of visualizations to use to help users analyze individual gage points across different time periods.

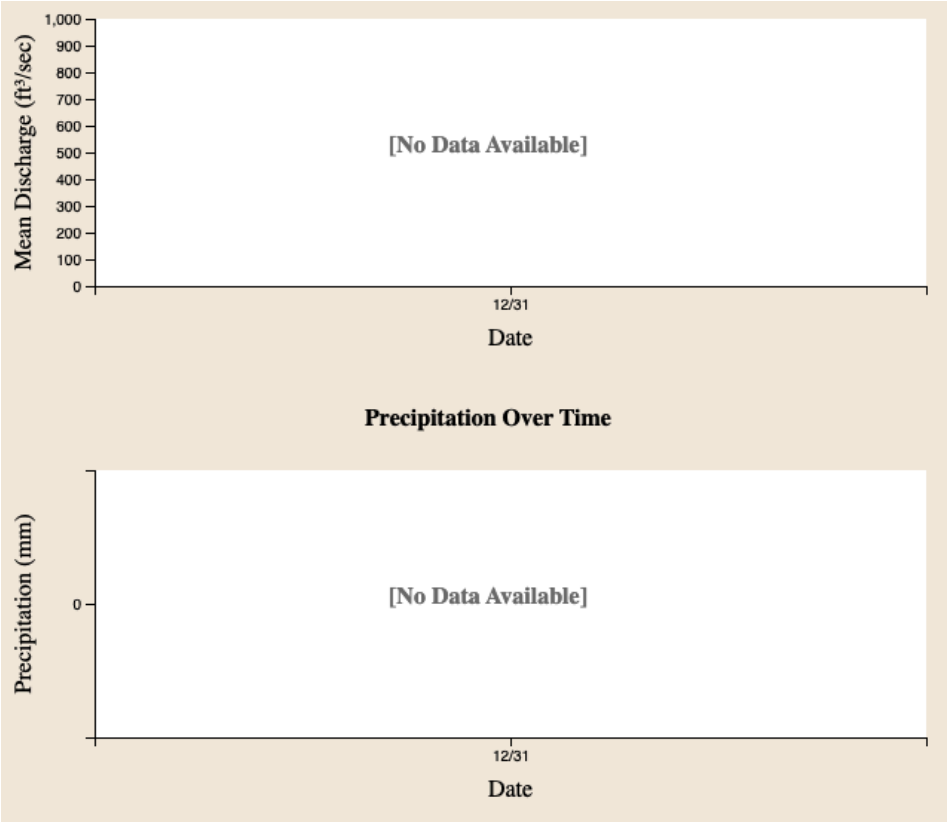
Related Work:

Questions:

We narrowed down the questions we identified leading up to the milestone to the following:

- Which visualizations would be the most useful for allowing users to analyze gage points?
- Which river analysis task can we most effectively support with our visualization?
- How can we offer multiple methods of exploring gage points in New England?

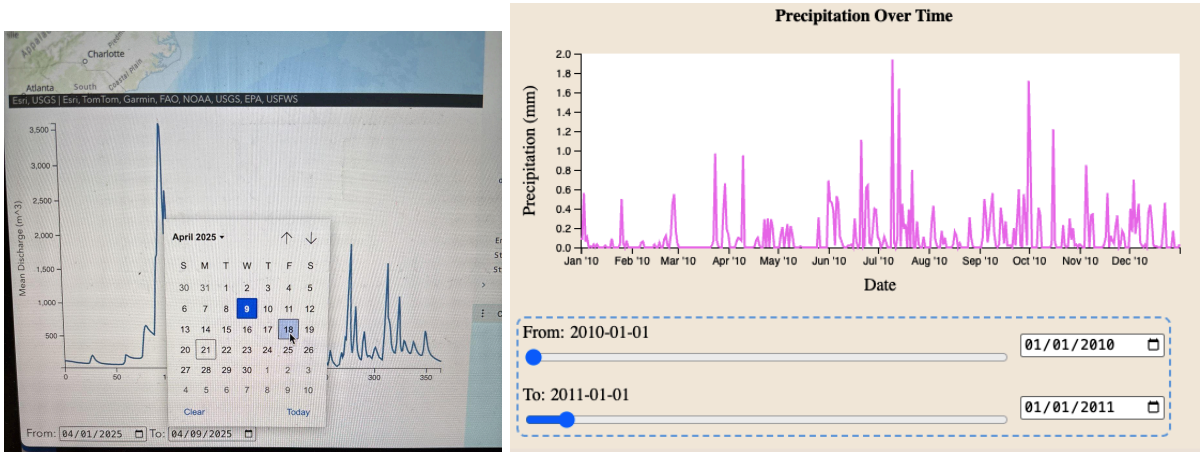
Exploratory Data Analysis:



While we didn't intentionally use our visualizations to understand our data better, we inadvertently ran into an error where NaN values would break our line graph visualizations. After further investigation, we found certain gage points in our dataset actually included only NaN values at every data point, and we performed further data cleaning by completely removing these gage points from our database.

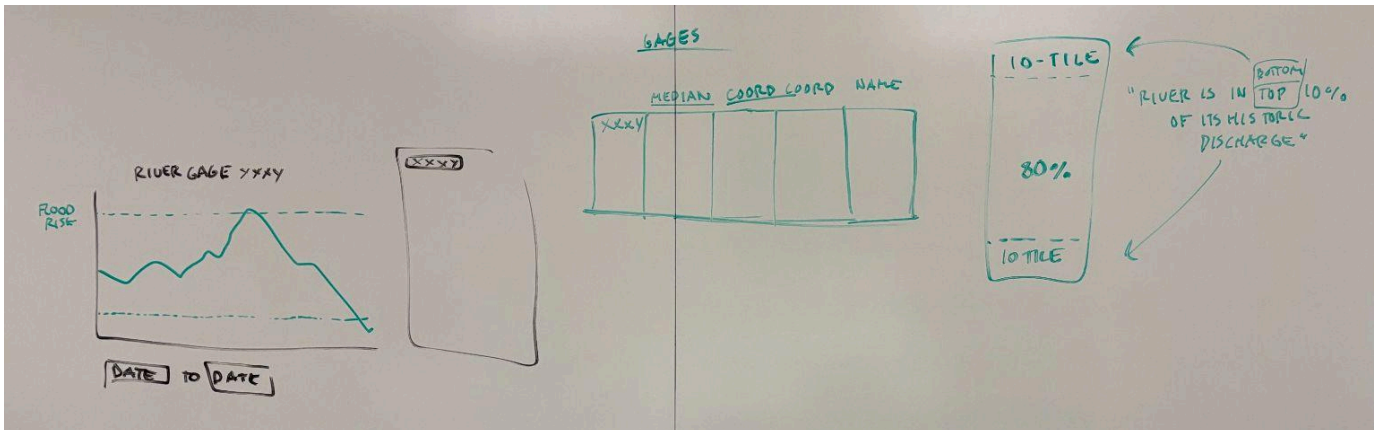
In cases where gage points contained only some NaN values, we decided to update our line graphs to handle these cases by simply displaying an empty chart with the informative text displayed above.

Design Evolution:



Old Date Range Selector (left) vs. Final Date Range Selector (right)

The figures above showcase the evolution of our date range selector component. Initially, we envisioned the component to simply provide two calendars from which you may select dates, however we felt that this design doesn't fully utilize design principles to illustrate the range of possible dates. In our redesign, we utilized the position on a common scale channel to represent the position of the current date within the full date range. Additionally, this visualization utilizes the channel of length to visualize the relative length of the currently selected date range in comparison to the full date range.



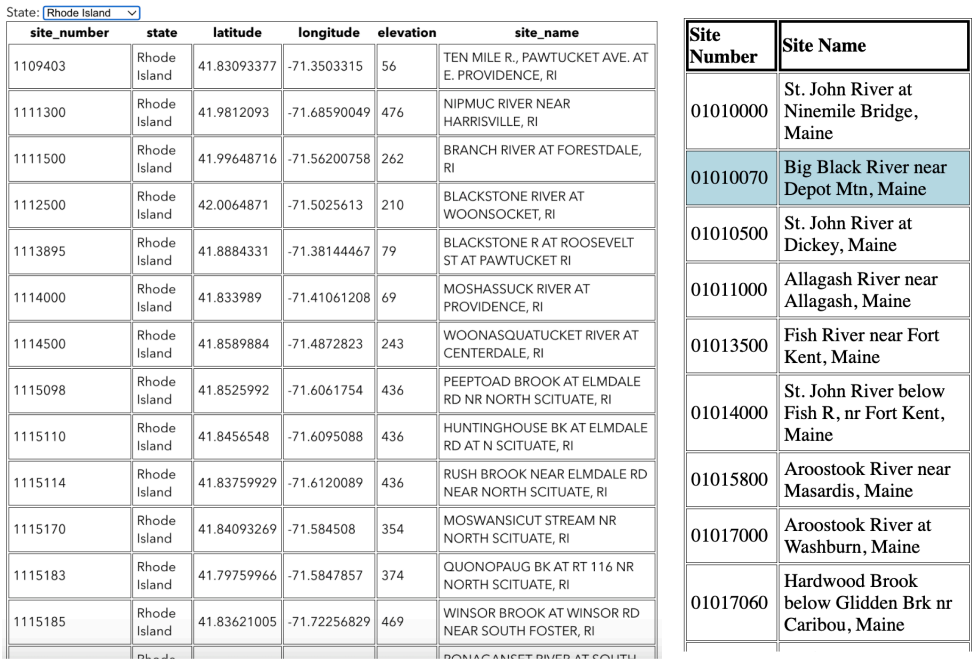
Visualization components brainstorm

The figures above illustrate some concepts of visualization idioms for presenting information about river gage discharge. The feature depicted on the leftmost graph was eventually integrated into our final visualization, and showcases horizontal lines which represent the level at which a river would flood and the level at which the river would be affected by droughting.

We realized that there were no readily available datasets with the drought levels and flood levels for the rivers in our current dataset, so we decided to adapt by simply displaying various discharge percentiles, taken from all of the recorded discharge values for a given gage point. At

one point in our development, we had considered allowing for analysis of gage points at a specific date and we developed the figure on the right to visualize the discharge of a gage point at a selected date in comparison to descriptive statistics of all of its recorded discharge datapoints. While this idea was dropped since we felt it didn't answer our main questions, the concept of visualizing percentile discharge data still appeared in our final iteration.

Lastly, the figure in the center displays our table component with additional columns for percentile discharge data. We ultimately decided this would result in our table taking too much space and felt it was unnecessary to display in the table, which we identified as mainly a tool for finding a specific gage point the user would like to analyze.



State: Rhode Island

site_number	state	latitude	longitude	elevation	site_name
1109403	Rhode Island	41.83093377	-71.3503315	56	TEN MILE R., PAWTUCKET AVE. AT E. PROVIDENCE, RI
1111300	Rhode Island	41.9812093	-71.68590049	476	NIPMUC RIVER NEAR HARRISVILLE, RI
1111500	Rhode Island	41.99648716	-71.56200758	262	BRANCH RIVER AT FORESTDALE, RI
1112500	Rhode Island	42.0064871	-71.5025613	210	BLACKSTONE RIVER AT WOONSOCKET, RI
1113895	Rhode Island	41.8884331	-71.38144467	79	BLACKSTONE R AT ROOSEVELT ST AT PAWTUCKET RI
1114000	Rhode Island	41.833989	-71.41061208	69	MOSHASSUCK RIVER AT PROVIDENCE, RI
1114500	Rhode Island	41.8589884	-71.4872823	243	WOONASQUATUCKET RIVER AT CENTERDALE, RI
1115098	Rhode Island	41.8525992	-71.6061754	436	PEEPTOAD BROOK AT ELMDALE RD NR NORTH SCITUATE, RI
1115110	Rhode Island	41.8456548	-71.6095088	436	HUNTINGHOUSE BK AT ELMDALE RD AT N SCITUATE, RI
1115114	Rhode Island	41.83759929	-71.6120089	436	RUSH BROOK NEAR ELMDALE RD NEAR NORTH SCITUATE, RI
1115170	Rhode Island	41.84093269	-71.584508	354	MOSWANSICUT STREAM NR NORTH SCITUATE, RI
1115183	Rhode Island	41.79759966	-71.5847857	374	QUONOPAUG BK AT RT 116 NR NORTH SCITUATE, RI
1115185	Rhode Island	41.83621005	-71.72256829	469	WINSOR BROOK AT WINSOR RD NEAR SOUTH FOSTER, RI

Site Number	Site Name
01010000	St. John River at Ninemile Bridge, Maine
01010070	Big Black River near Depot Mtn, Maine
01010500	St. John River at Dickey, Maine
01011000	Allagash River near Allagash, Maine
01013500	Fish River near Fort Kent, Maine
01014000	St. John River below Fish R, nr Fort Kent, Maine
01015800	Aroostook River near Masardis, Maine
01017000	Aroostook River at Washburn, Maine
01017060	Hardwood Brook below Glidden Brk nr Caribou, Maine

Milestone Table View (left) vs. Final Table View (right)

We decided to remove multiple columns from our table, as we realized that the table's main function is to identify gage points to learn more about, and that data such as latitude and longitude are unimportant for such tasks. This also provided the additional benefit of freeing up a lot of space on our webpage for other visualizations to occupy.

Selected Gage Point:

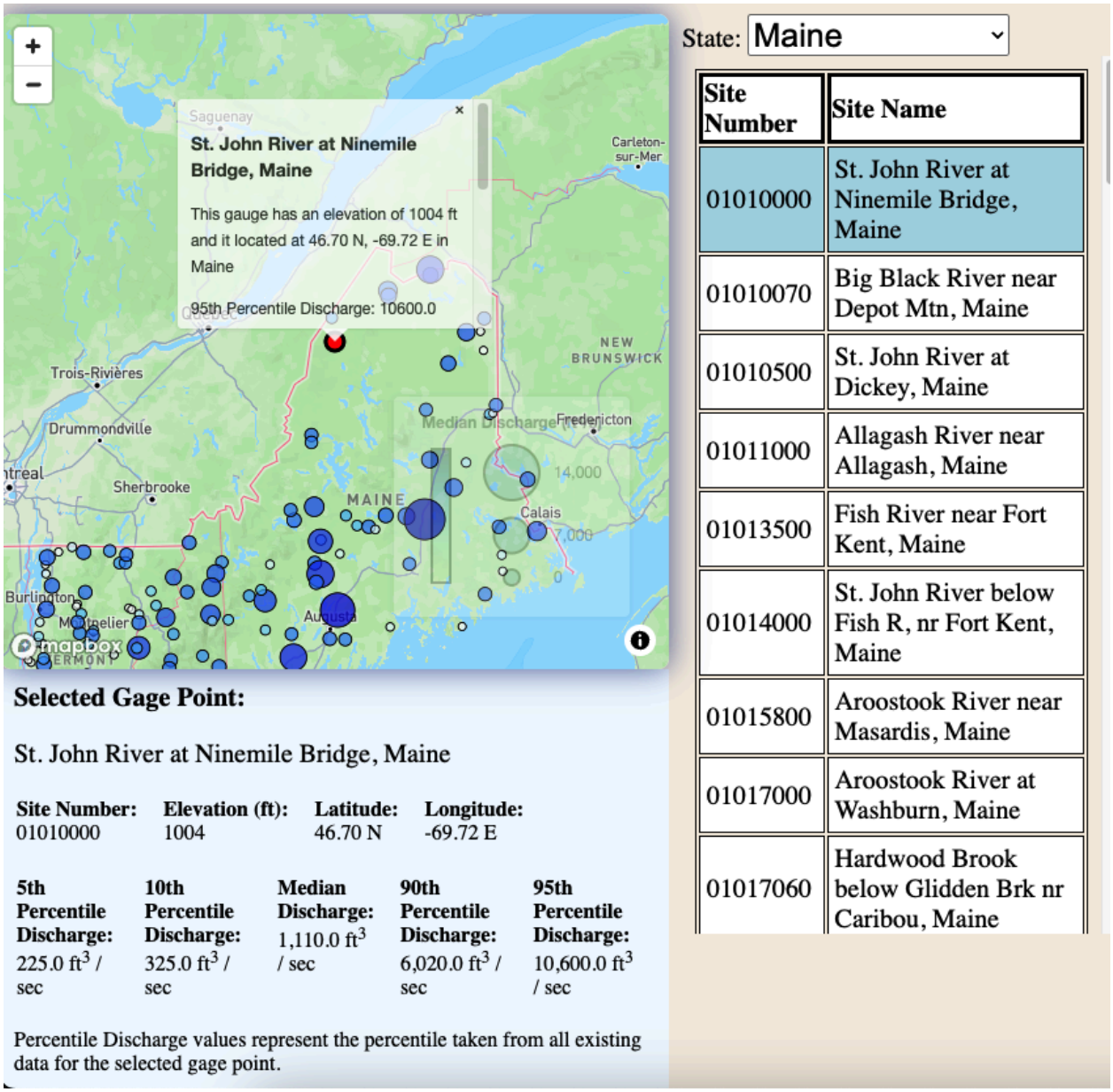
St. John River at Ninemile Bridge, Maine

Site Number:	Elevation (ft):	Latitude:	Longitude:
01010000	1004	46.70 N	-69.72 E

5th Percentile Discharge:	10th Percentile Discharge:	Median Discharge:	90th Percentile Discharge:	95th Percentile Discharge:
225.0 ft^3 / sec	325.0 ft^3 / sec	1,110.0 ft^3 / sec	6,020.0 ft^3 / sec	10,600.0 ft^3 / sec

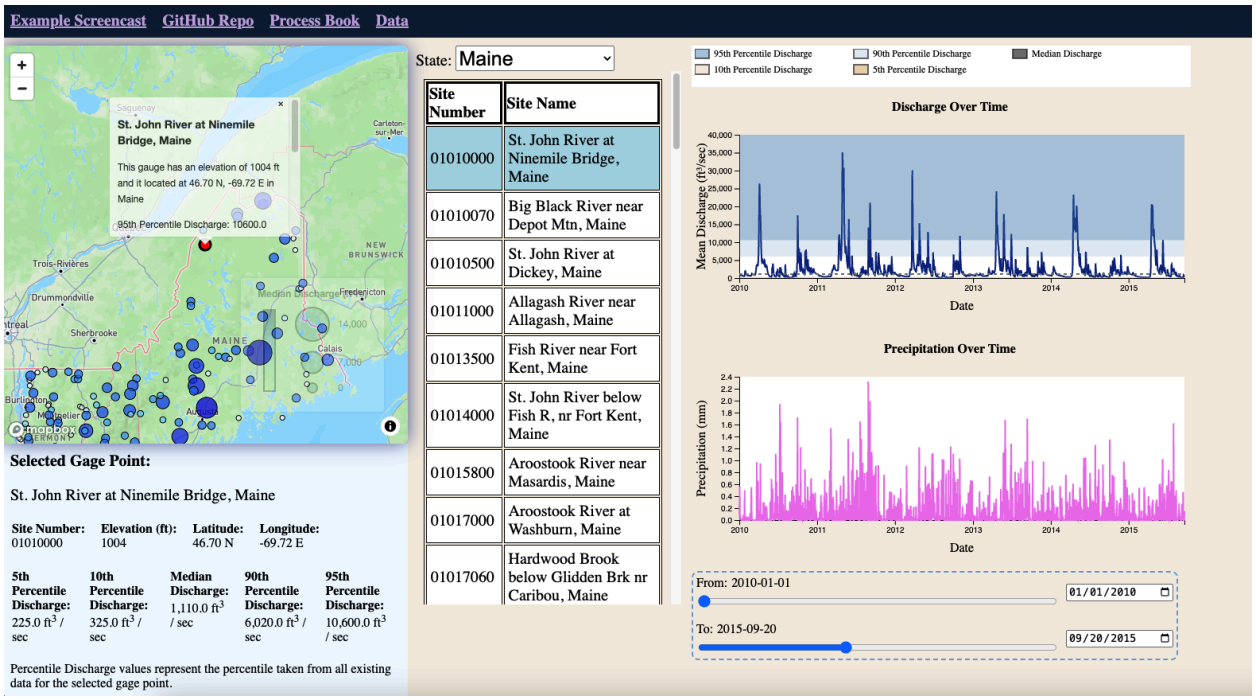
Percentile Discharge values represent the percentile taken from all existing data for the selected gage point.

Another visualization we decided to add to our webpage was the gage summary view, which provides all of the information for the selected gage point in a single location. We added this component as a result of percentile discharge data previously only being provided by map popups, which we felt was unnecessarily restrictive.
This feature was also developed to address the omission of certain columns in our table, which didn't have anywhere to be displayed before the development of the gage summary view.



Another feature we added was the linking of the map and table selectors. Selecting on the table will select the gage point on the map and focus on the location on the map and clicking a gage point on the map will select it on the table as well.

Implementation:



Our New England River Analysis Data Visualization in its final iteration.

Generally speaking, we have organized the display of the visualizations on this webpage with the exploratory visualization components on the left and the analytical visualization components on the right.

On the left, users are encouraged to browse gage points by location through the table or explore gages by their relative median discharges on the map. The table provides the user with organization based on state and while the map also displays gage points geographically, it mainly distinguishes between gage points by median discharge instead of by using text to describe them by name.

The map utilizes both 2D area and hue to represent the median discharge of a gage point, with larger and darker circles representing gage points with high median discharge and lighter and smaller circles representing gage points with low median discharge.

Users are encouraged to analyze individual gage points through the gage summary, which provides descriptive statistics which describe percentile data of all of a gage point's recorded data points.

On the right, users may observe specific gage points' relationship between discharge and precipitation. Additionally, they may hover over a specific point on each graph to view specific data values and their associated date. Our intention is for users to utilize these two graphs to

develop an understanding of the relationship between discharge and precipitation across multiple gage points and also various date ranges.

Evaluation:

Once we developed our visualizations, our group found it much easier to observe overall trends in individual rivers' discharge and precipitation over time. Additionally, our website provided us the benefit of being able to compare different gage points' precipitation and discharge over time. Previously, it would have been necessary to browse through gage points one at a time on the USGS website, but our visualization idioms allowed us to easily swap between gage points and compare them.

We believe we were able to solve all of the research questions we raised to ourselves after redefining the scope of our project while completing our milestone. First, through our iterative design process, we believe we found the best combination of visualizations to allow users easy exploration of gage points as well as analysis of gage points. Additionally, we believe we appropriately selected the main analysis task that our visualization caters to. Considering the size of our dataset, we felt that it was unlikely that many individuals would find much benefit to analysis of gage points at single dates. Finally, we believe we succeeded in providing multiple methods of gage point exploration, as both our table view and map view provide distinct, intuitive methods of locating gage points.

If we had more time, we would first like to spend more time adjusting the layout of our webpage, as we felt that some elements may be uncomfortably small for users. Originally, we wanted to add features such as sliding windows, which we believe would allow for larger display of components, however we left these features out of our webpage in its current state due to time constraints.

Additionally, a hovering feature which displays basic gage point info such as site number and site name could have been implemented for the map, to allow users who prefer to explore geographically some ability to explore gages by name.

While we decided to exclude it from our final submission of our webpage, we still believe there are many use cases where an ML model for predicting future river flooding or droughting could prove to be useful, as well as model analytic visualizations for understanding concepts such as feature importance.