watervue

NYU TANDON SCHOOL OF ENGINEERING

Water Quality Forecasting Models

Capstone Final Presentation

Anna Cui | Yifan Wu | Juanita Jacobraja | Keya Keya | Priya Kundu

05/06/2024

Presentation Agenda

Introduction

Objective, Goals and Deliverables

Data Pre-processing

Handling Missing values,

Trimming outliers,

Preparing the dataset for Model Training

Exploratory Data Analysis

Understanding each variable distribution and it's multicollinearity

Model Exploration & Selection

Researching and picking the best model for each water testing parameter based on its location

Model Performance

Analyzing each model's performance based on it;s accuracy.

Model Comparison

Comparing each model to understand which model suits best for the respective parameters

Improvements and Suggestions

Our comments as a team to make the project outcome even better.

Meet the Team

Anna Cui

Project Manager

Dual major in Managerial Economics and Spanish

Keya Keya

Data Analyst

B.E. Electronics & Telecommunications

Yifan Wu

Data Analyst

Actuarial Science

Juanita Jacobraja

ML Engineer

B.E. Computer Science

Priya Kundu

ML Engineer

B.E. Mechanical Engineering

Objective

Develop a robust machine learning model for precise water quality forecasting in Biscayne Bay, Florida.

Short-Term Goals

Provide stakeholders with actionable insights on the water data.

Forecast Water Testing Parameters and use clustering techniques to predict water quality.

Gain practical experience machine learning techniques and project management skills as a team

Deliverables

Data Acquisition and Preprocessing

Model Development and Training

Model Validation

Objective

Develop a robust machine learning model for precise water quality forecasting in Biscayne Bay, Florida.

Short-Term Goals

Provide stakeholders with actionable insights on the water data.

Forecast Water Testing Parameters and use clustering techniques to predict water quality.

Gain practical experience machine learning techniques and project management skills as a team

Deliverables

Data Acquisition and Preprocessing

Model Development and Training

Model Validation

Objective

Develop a robust machine learning model for precise water quality forecasting in Biscayne Bay, Florida.

Short-Term Goals

Provide stakeholders with actionable insights on the water data.

Forecast Water Testing Parameters and use clustering techniques to predict water quality.

Gain practical experience machine learning techniques and project management skills as a team

Deliverables

Data Acquisition and Preprocessing

Model Development and Training

Model Validation

Initial Challenges and Hypothesis

Limited initial knowledge on water quality/metrics; extensive research on each water quality parameters.

Choosing the best-fit datasets for our project.

Uncertainty about final result expectations, milestones, as well as the structuring of the entire project.

Efforts to cluster data and apply drinkable water standards (found from online sources)

Challenges

Hypothesis

There could be significant correlation between the provided parameters. These correlations can be used to predict one parameter based on others, improving the efficiency.

Different machine learning algorithms will vary in their effectiveness at forecasting water quality parameters.

Seasonality in data significantly influence water quality parameters.

Dataset Introduction

	Site #	Location	Sample Date	Analysis code	Analysis	DQC	Result	Units	Detection Limit
0	1.0	#1 HILLSBORO CANAL US 1	2/22/06	Chl a	Chlorophyll a	NaN	12.100	mg/m3	0.955
1	1.0	#1 HILLSBORO CANAL US 1	2/22/06	Conductivity	Specific Conductance	NaN	31300.000	umho/cm	1.000
2	1.0	#1 HILLSBORO CANAL US 1	2/22/06	DO	Dissolved Oxygen	NaN	6.980	mg/L	0.050
3	1.0	#1 HILLSBORO CANAL US 1	2/22/06	Sal	Salinity	NaN	19.400	ppt	0.025
4	1.0	#1 HILLSBORO CANAL US 1	2/22/06	TN	Total Nitrogen	NaN	0.830	mg/L	0.107
...	...	...	...	...	...	...	...	...	...
19395	129.0	#129 Taft Street	11/15/21	Sal	Salinity	NaN	1.230	ppt	0.025
19396	129.0	#129 Taft Street	11/15/21	TP	Total Phosphorus	I	0.035	mg/L	0.013
19397	129.0	#129 Taft Street	11/15/21	Turbidity	Turbidity	NaN	1.400	NTU	0.080
19398	129.0	#129 Taft Street	11/15/21	TN	Total Nitrogen	NaN	0.821	mg/L	0.049
19399	129.0	#129 Taft Street	11/15/21	Chl a	Chlorophyll a	NaN	11.300	ug/L	0.125

19400 rows × 9 columns

The dataset is primarily numerical and categorical, with continuous data.

There are 60 unique locations in the dataset and 7 different water testing parameters.

Some columns like DQC have a significant number of missing values, which needs to be preprocessed to use.

Data Pre-processing

Data Structuring

	Sample Date	Location	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
0	1/30/12	#1 HILLSBORO CANAL US 1	2.77	6.110	24.80	39000.0	0.3600	0.0770	1.70
1	1/30/12	#11 MIDDLE RIVER NW 21ST AVE	5.29	5.195	5.39	9485.0	0.7025	0.0345	1.00
2	1/30/12	#110 POMPANO CANAL AT DIXIE AN	5.53	7.130	0.20	424.0	1.1600	0.0330	2.00
3	1/30/12	#111 S. FORK MID R. @ N.E. 15	NaN	6.170	20.80	33300.0	0.5460	0.0490	1.00
4	1/30/12	#112 N. FORK MID R. @ N.E. 16	NaN	5.980	21.20	34000.0	0.4840	0.0530	0.90
...	...	...	...	...	...	...	...	...	...
3031	9/9/19	#2 HILLSBORO LOCK	15.00	6.250	0.29	598.0	1.1763	0.0810	1.80
3032	9/9/19	#3 HILLSBORO CANAL US 441	9.77	2.120	0.29	605.0	1.3565	0.0680	1.40
3033	9/9/19	#4 HILLSBORO CANAL SE GROWERS	12.50	4.190	0.34	716.0	1.5725	0.0250	4.50
3034	9/9/19	#8 POMPANO CANAL US 441	1.86	3.980	0.36	739.0	1.1845	-0.0010	0.60
3035	9/9/19	#89 NOB HILL RD POMPANO CANAL	1.78	3.450	0.36	743.0	1.1839	0.0000	0.45

3036 rows × 9 columns

01 Pivoting the Data Frame

02 Dropping Unnecessary Columns

03 Merging Similar Columns

Enhancing Data Quality

Removing Outliers utilizing the Interquartile Range (IQR)

```
Maximum value of Chlorophyll A: 99.2
Minimum value of Chlorophyll A: -0.048
Maximum value of Dissolved Oxygen: 61.5
Minimum value of Dissolved Oxygen: 0.32
Maximum value of Salinity: 72.4
Minimum value of Salinity: 0.0
Maximum value of Specific Conductance: 427119.0
Minimum value of Specific Conductance: 37.5
Maximum value of Total Nitrogen: 2.5149999999999997
Minimum value of Total Nitrogen: 0.0
Maximum value of Total Phosphorus: 0.649
Minimum value of Total Phosphorus: -0.147
Maximum value of Turbidity: 23.0
Minimum value of Turbidity: 0.0
```

Range of the Parameters Pre - Removal of Outliers

```
Maximum value of Chlorophyll A: 14.6
Minimum value of Chlorophyll A: -0.048
Maximum value of Dissolved Oxygen: 9.39
Minimum value of Dissolved Oxygen: 0.82
Maximum value of Salinity: 37.3
Minimum value of Salinity: 0.0
Maximum value of Specific Conductance: 58800.0
Minimum value of Specific Conductance: 37.5
Maximum value of Total Nitrogen: 2.38
Minimum value of Total Nitrogen: 0.0
Maximum value of Total Phosphorus: 0.147
Minimum value of Total Phosphorus: -0.047
Maximum value of Turbidity: 3.4
Minimum value of Turbidity: 0.0
```

Range of the Parameters Post - Removal of Outliers

Data Transformation

Logarithmic Transformation to reduce Skewness of 'Specific Conductance'

```
Maximum value of Chlorophyll A: 99.2
Minimum value of Chlorophyll A: -0.048
Maximum value of Dissolved Oxygen: 61.5
Minimum value of Dissolved Oxygen: 0.32
Maximum value of Salinity: 72.4
Minimum value of Salinity: 0.0
Maximum value of Specific Conductance: 427119.0
Minimum value of Specific Conductance: 37.5
Maximum value of Total Nitrogen: 2.5149999999999997
Minimum value of Total Nitrogen: 0.0
Maximum value of Total Phosphorus: 0.649
Minimum value of Total Phosphorus: -0.147
Maximum value of Turbidity: 23.0
Minimum value of Turbidity: 0.0
```

Range of the Parameter
Pre Log Transformation

```
Maximum value of Chlorophyll A: 14.6
Minimum value of Chlorophyll A: -0.048
Maximum value of Dissolved Oxygen: 9.39
Minimum value of Dissolved Oxygen: 0.82
Maximum value of Salinity: 37.3
Minimum value of Salinity: 0.0
Maximum value of Specific Conductance: 58800.0
Minimum value of Specific Conductance: 37.5
Maximum value of Total Nitrogen: 2.38
Minimum value of Total Nitrogen: 0.0
Maximum value of Total Phosphorus: 0.147
Minimum value of Total Phosphorus: -0.047
Maximum value of Turbidity: 3.4
Minimum value of Turbidity: 0.0
```

Range of the Parameter
Post Log Transformation

Final Dataset for Model Training and Testing

Data Resampling

	Location	Sample Date	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
0	#1 HILLSBORO CANAL US 1	2006-02-28	12.100000	6.980000	19.400000	10.351373	0.830000	0.086000	2.400000
1	#1 HILLSBORO CANAL US 1	2006-08-31	4.430000	5.540000	15.750000	10.165844	0.981000	0.109000	1.400000
2	#1 HILLSBORO CANAL US 1	2007-02-28	2.605000	6.195000	23.200000	10.501905	0.754000	0.083500	1.650000
3	#1 HILLSBORO CANAL US 1	2007-08-31	4.890000	4.730000	31.100000	10.774781	0.777000	0.094000	2.300000
4	#1 HILLSBORO CANAL US 1	2008-02-29	5.925000	5.770000	12.500000	9.893361	1.440000	0.102000	2.050000
...	...	...	...	...	...	...	...	...	...
1559	#89 NOB HILL RD POMPANO CANAL	2020-02-29	3.933333	5.703333	0.310000	6.451930	1.097133	0.014000	0.750000
1560	#89 NOB HILL RD POMPANO CANAL	2020-08-31	1.570000	7.760000	0.250000	6.265301	0.910000	0.011000	0.000000
1561	#89 NOB HILL RD POMPANO CANAL	2021-02-28	2.966667	5.833333	0.233333	6.153383	0.984033	0.003667	0.566667
1562	#89 NOB HILL RD POMPANO CANAL	2021-08-31	5.510000	5.370000	0.270000	6.324359	1.410500	0.049000	0.775000
1563	#89 NOB HILL RD POMPANO CANAL	2022-02-28	2.900000	4.940000	0.290000	6.388561	1.040000	0.008000	0.550000

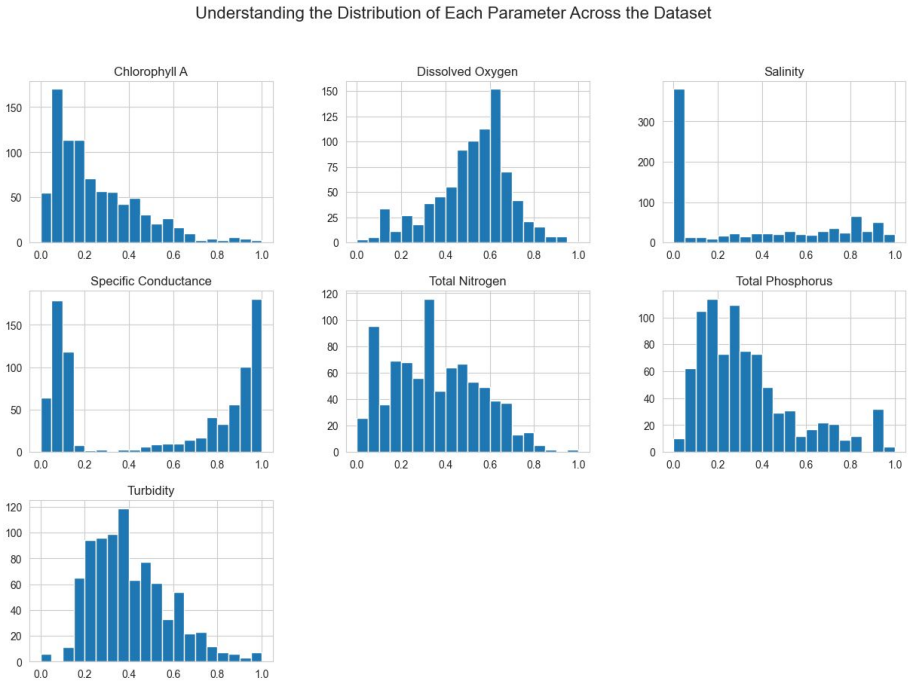
858 rows × 9 columns

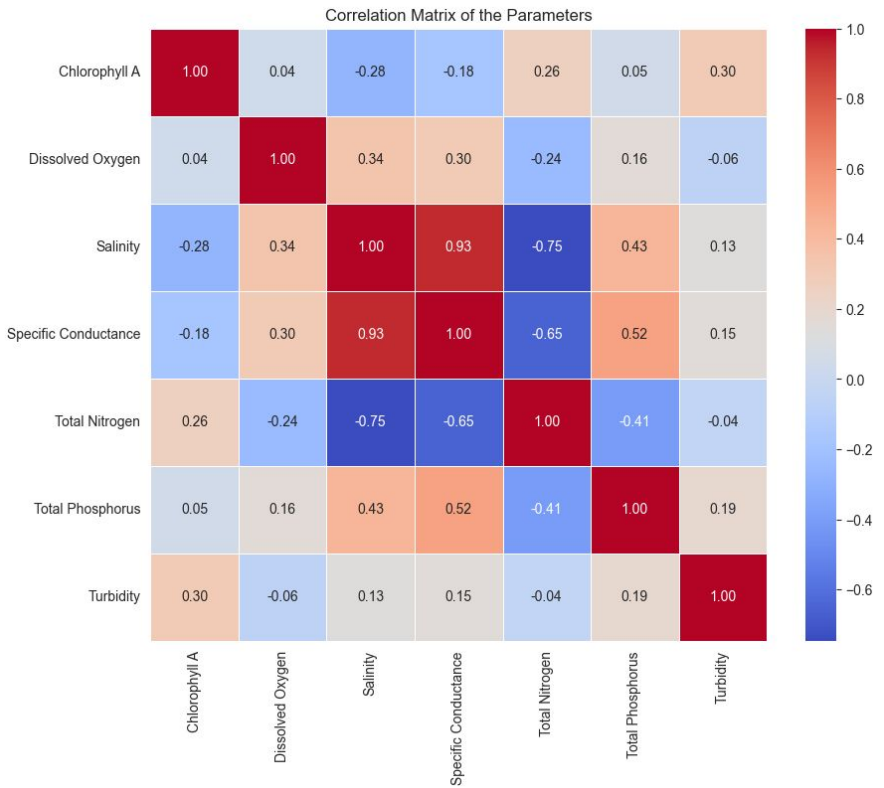
01 Resampled the data from quarterly data points to semi-annually

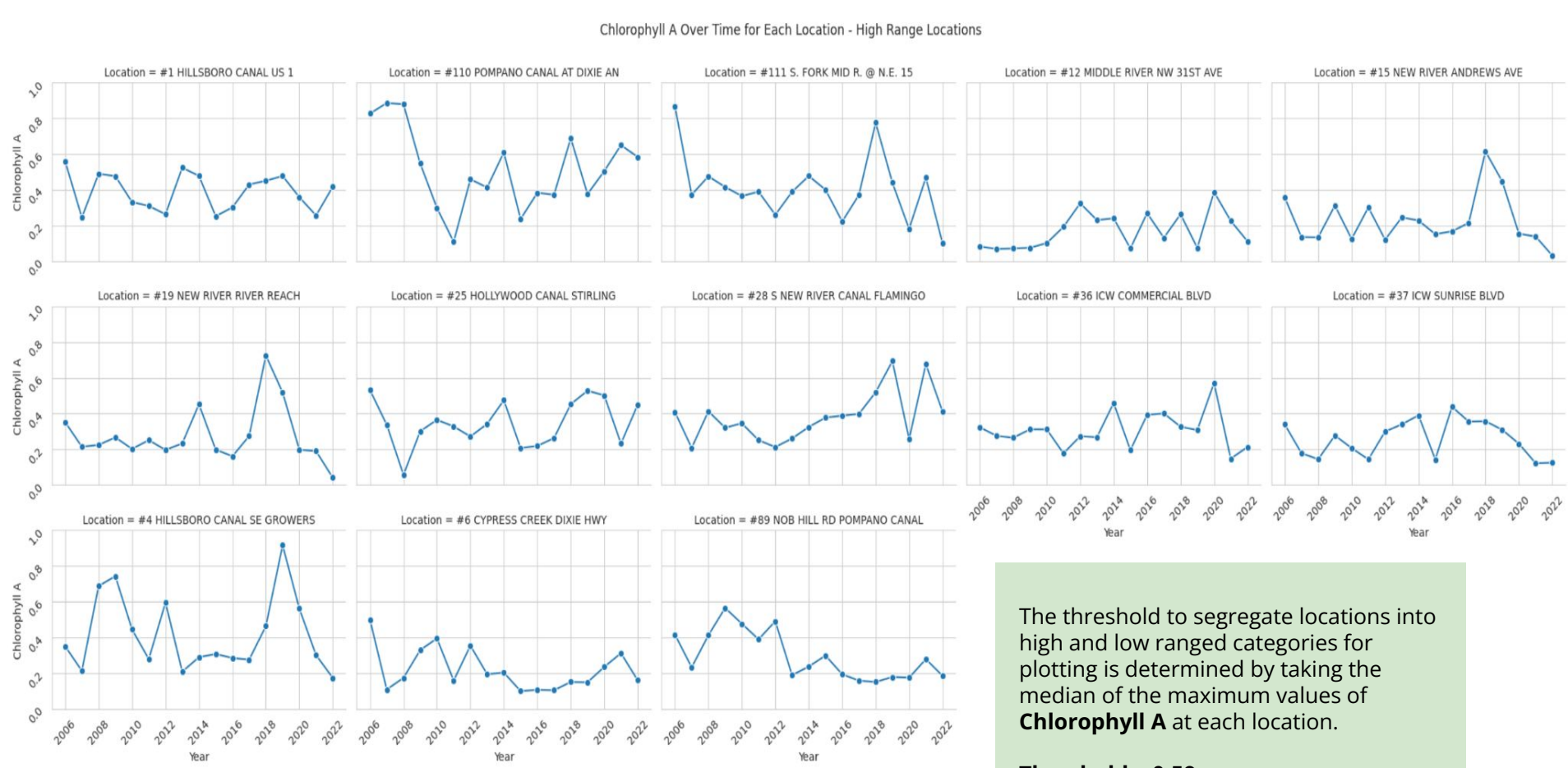
02 Further trimmed based on the Mode of the dataset

03 Interpolated for the missing values

Exploratory Data Analysis

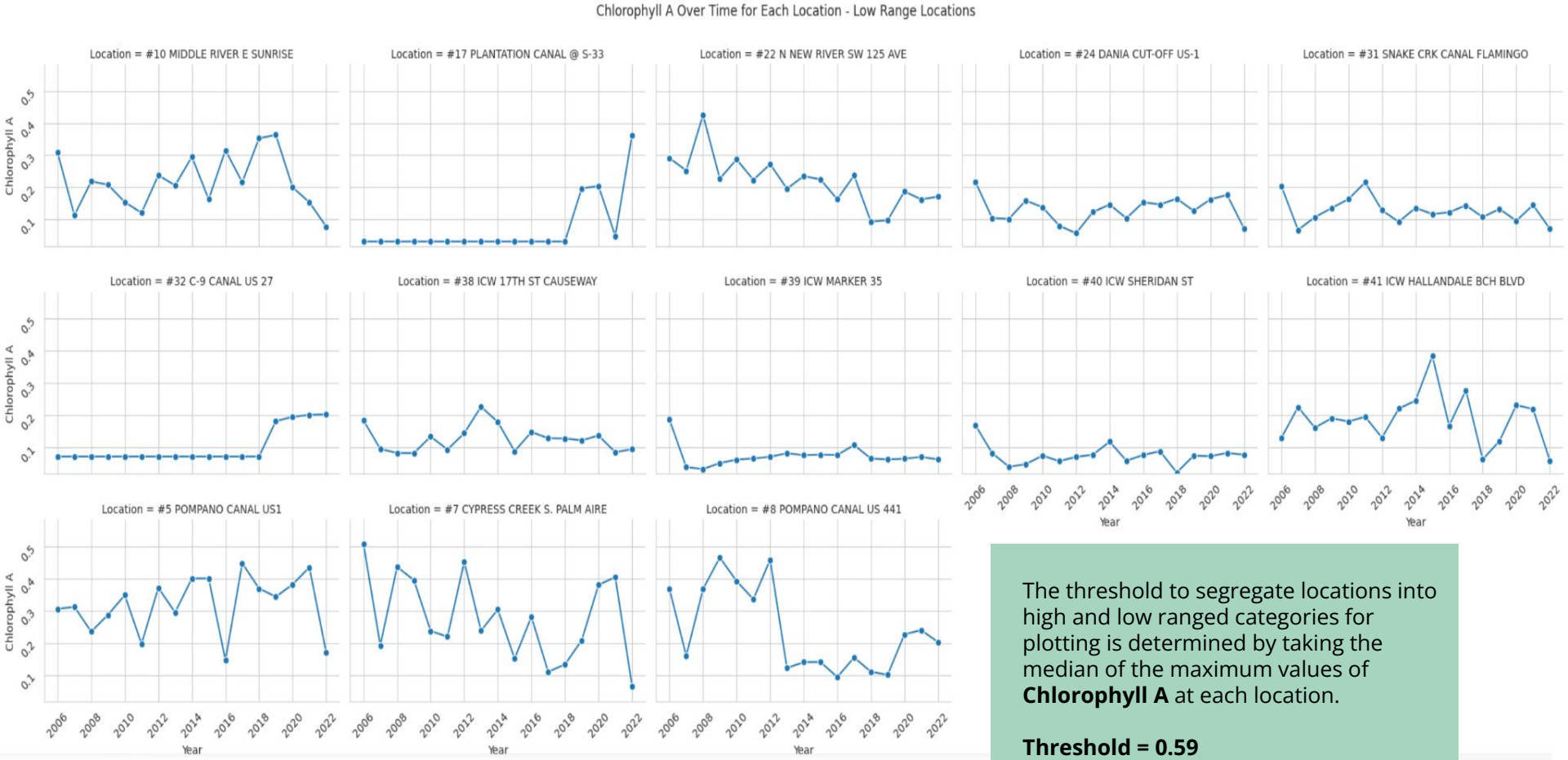






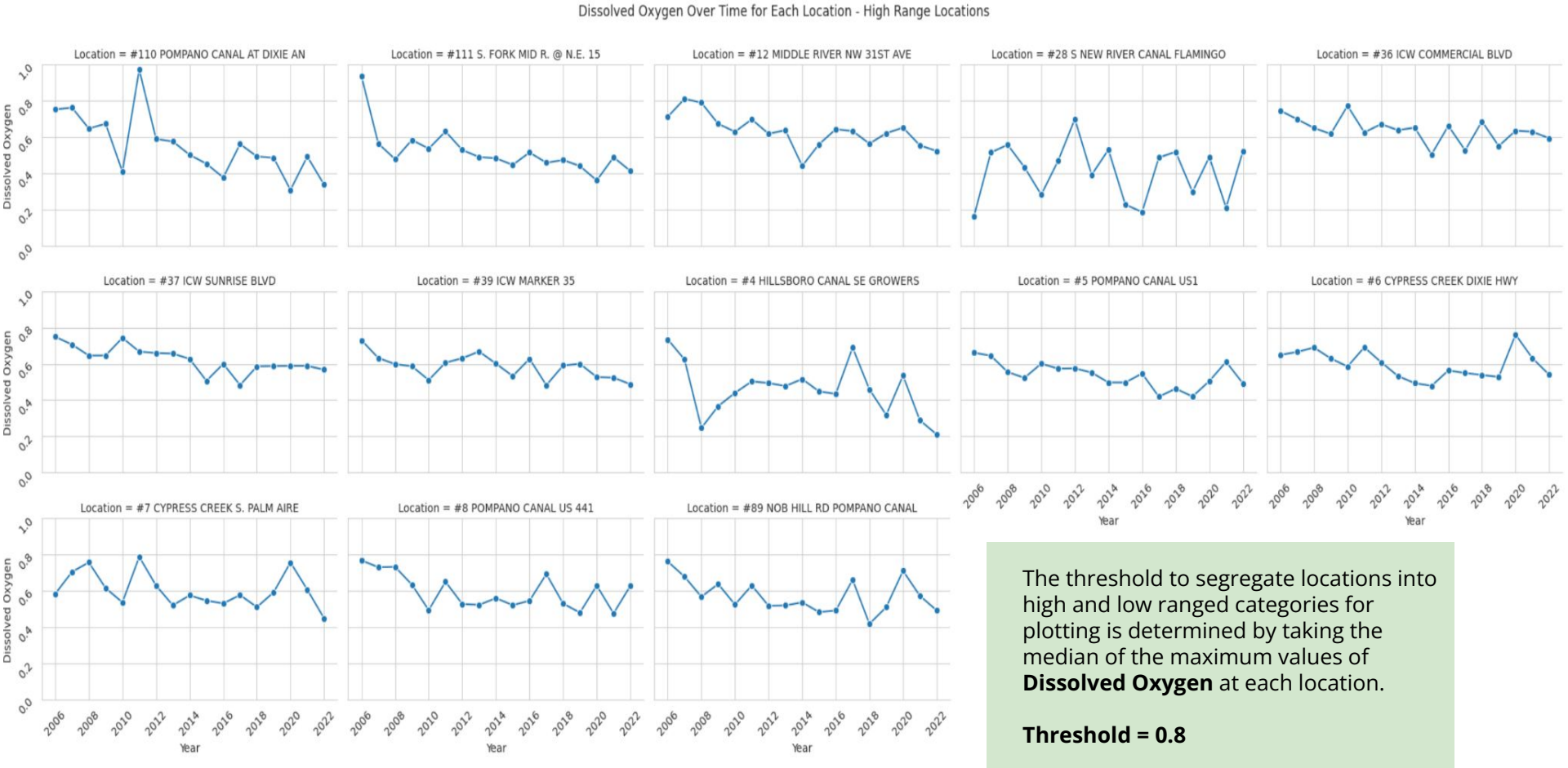
The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Chlorophyll A** at each location.

Threshold = 0.59



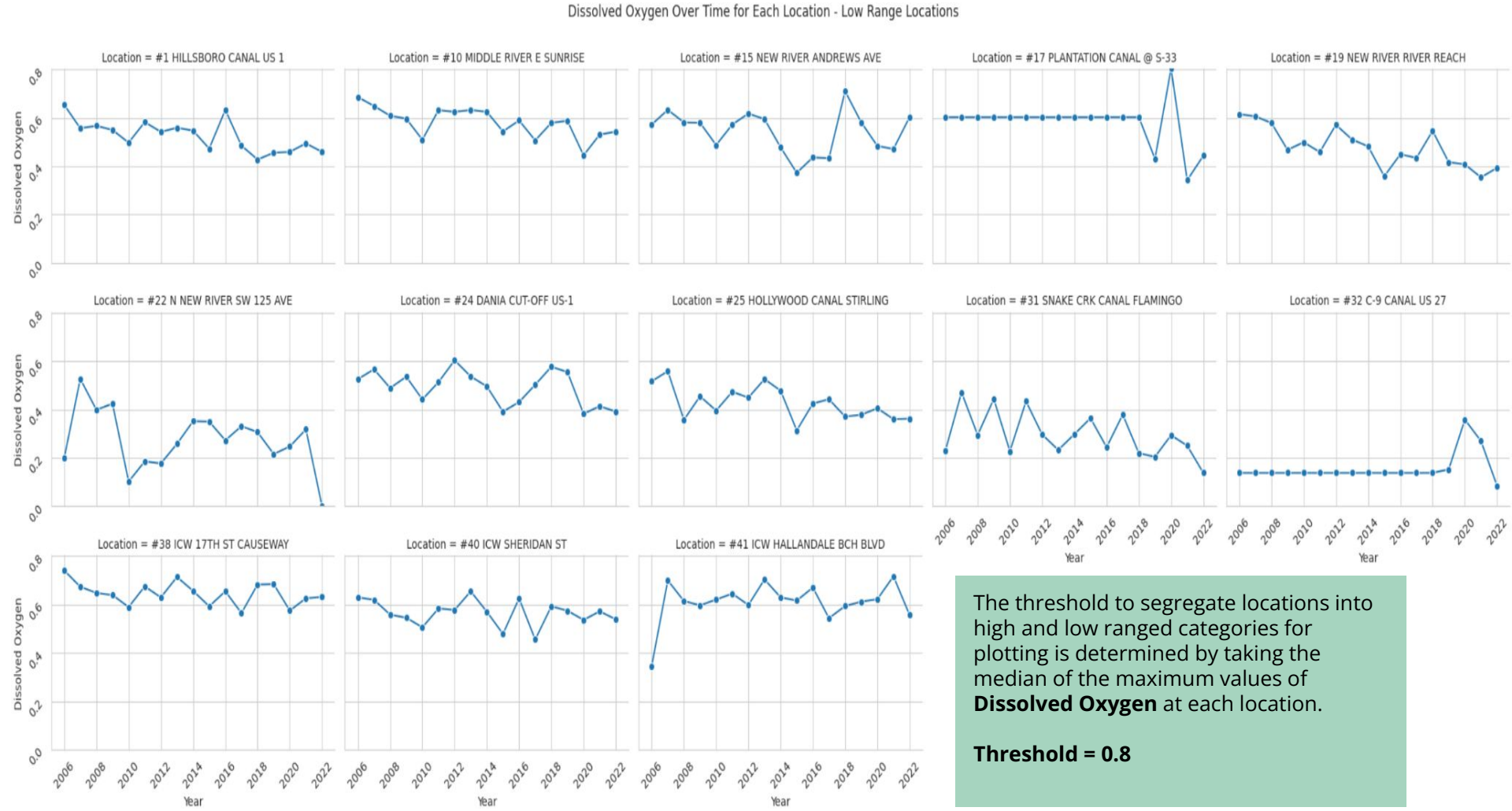
The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Chlorophyll A** at each location.

Threshold = 0.59



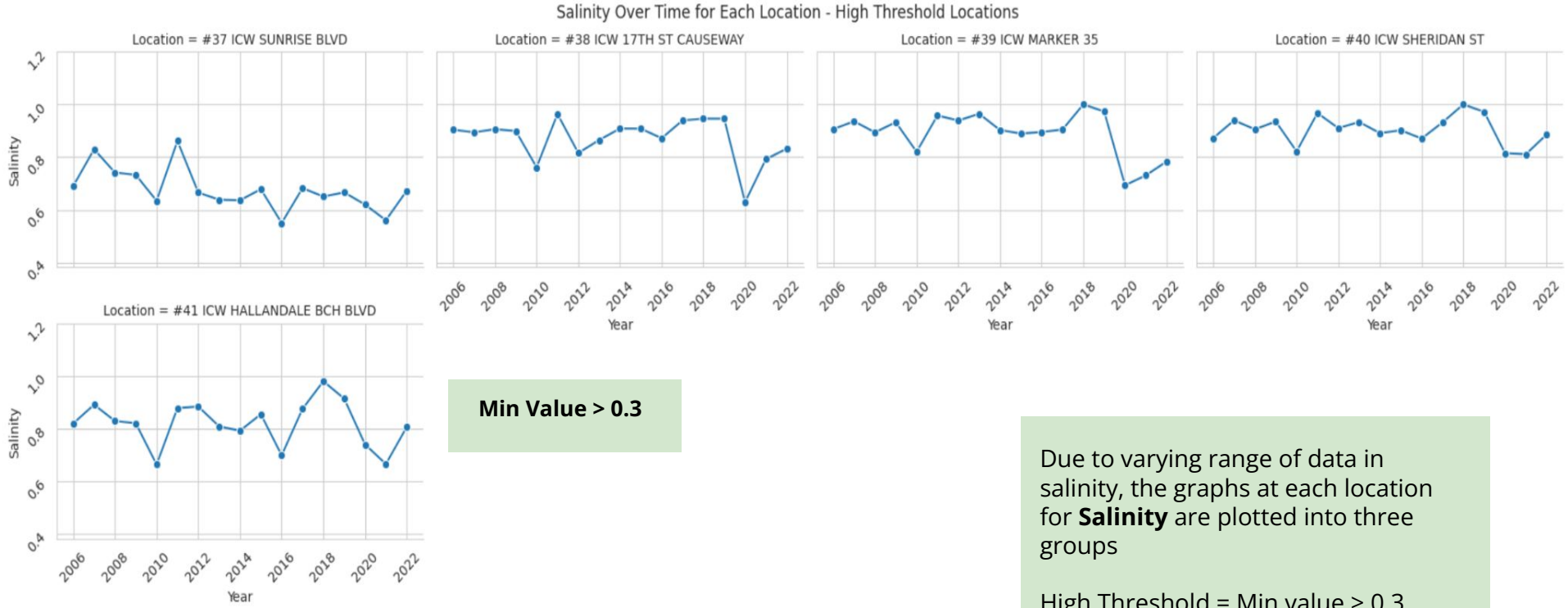
The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Dissolved Oxygen** at each location.

Threshold = 0.8



The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Dissolved Oxygen** at each location.

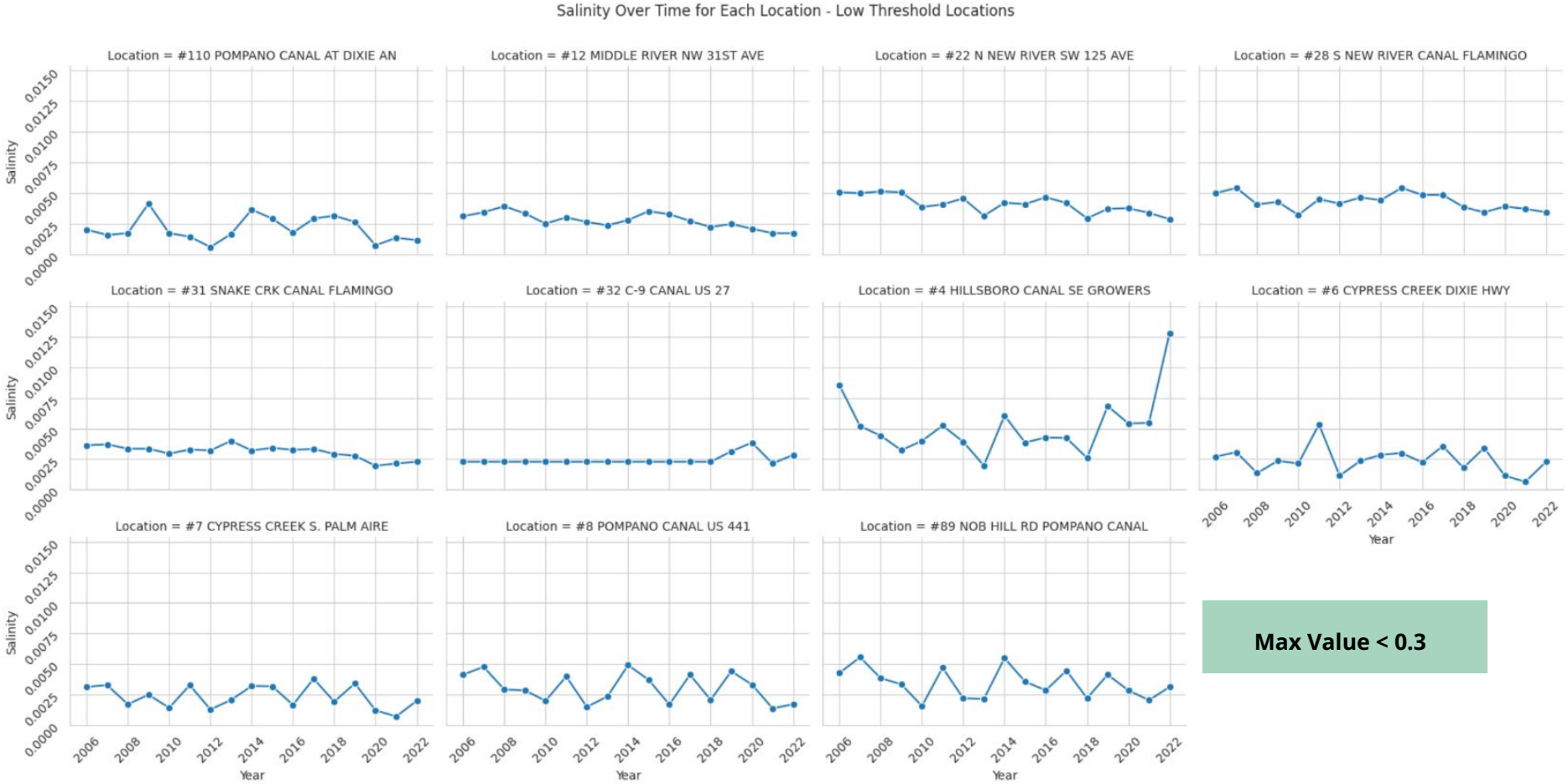
Threshold = 0.8



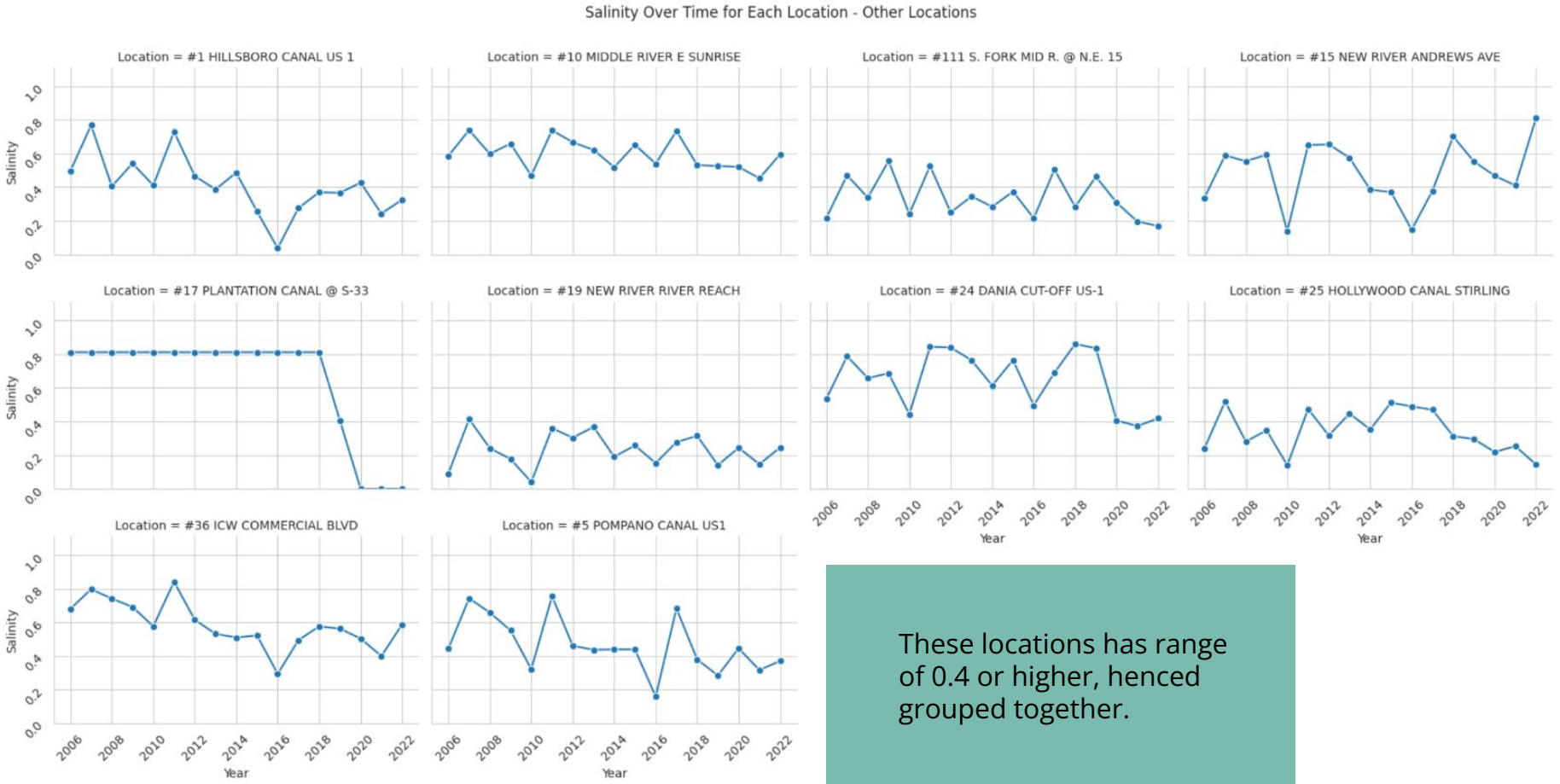
Min Value > 0.3

Due to varying range of data in salinity, the graphs at each location for **Salinity** are plotted into three groups

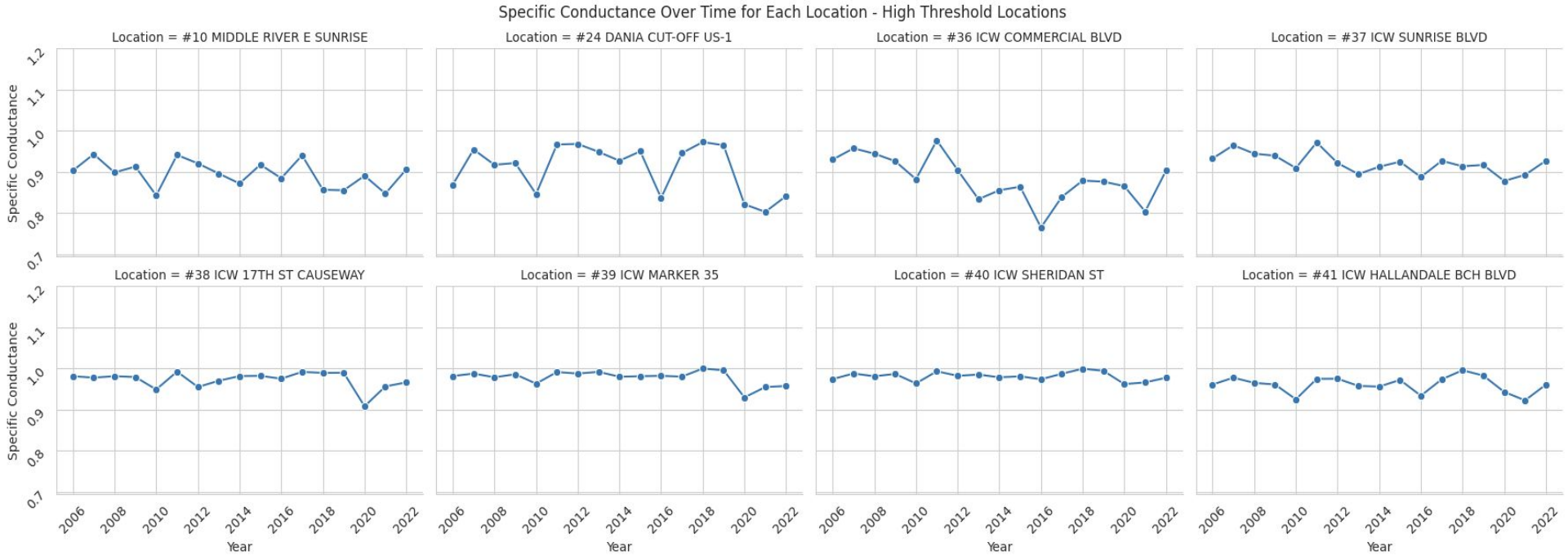
High Threshold = Min value > 0.3
Low Threshold = Max value <0.3
and others with varying range


Salinity Over Time for Each Location - Low Threshold Locations
Location = #110 POMPANO CANAL AT DIXIE AN
Location = #12 MIDDLE RIVER NW 31ST AVE
Location = #22 N NEW RIVER SW 125 AVE
Location = #28 S NEW RIVER CANAL FLAMINGO
Location = #31 SNAKE CRK CANAL FLAMINGO
Location = #32 C-9 CANAL US 27
Location = #4 HILLSBORO CANAL SE GROWERS
Location = #6 CYPRESS CREEK DIXIE HWY
Location = #7 CYPRESS CREEK S. PALM AIRE
Location = #8 POMPANO CANAL US 441
Location = #89 NOB HILL RD POMPANO CANAL
Salinity
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
2006
2008
2010
2012
2014
2016
2018
2020
2022
Year

Max Value < 0.3



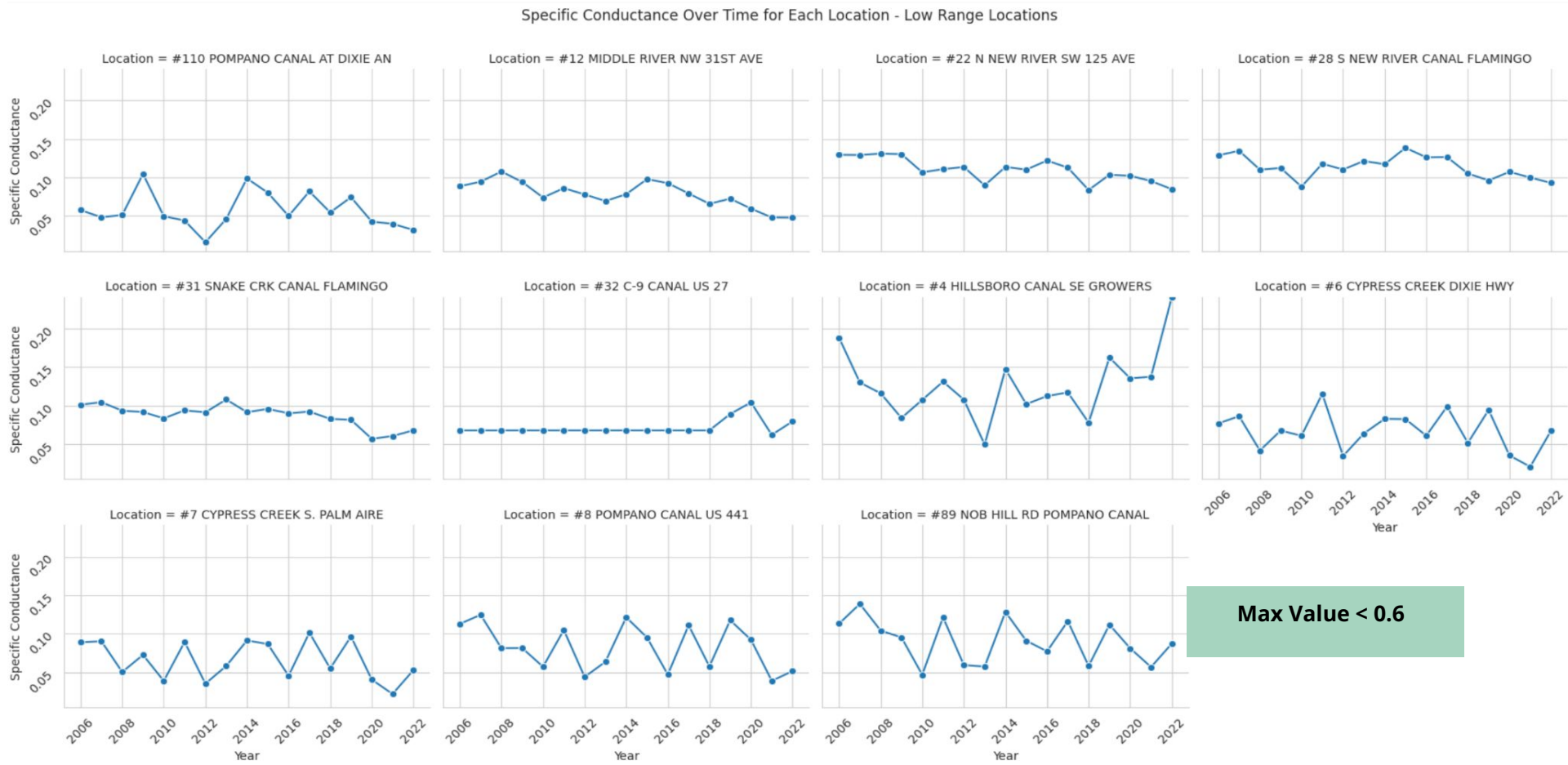
These locations has range of 0.4 or higher, henced grouped together.



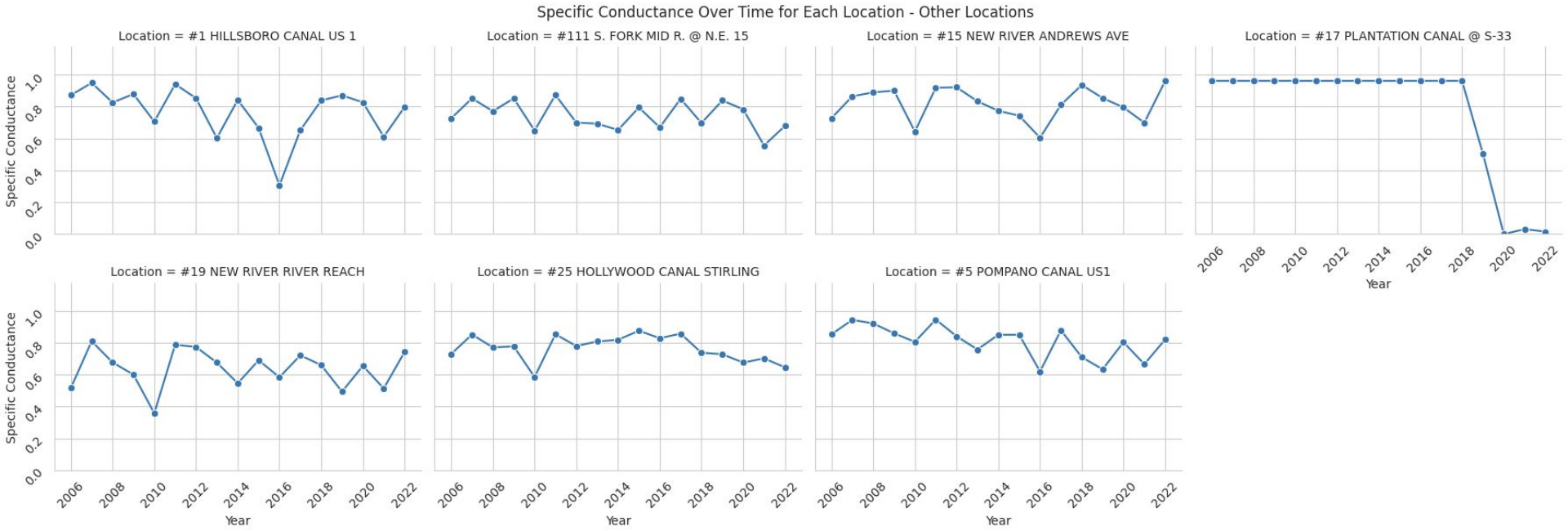
Due to varying range of data in salinity, the graphs at each location for **Specific Conductance** are plotted into three groups

High Threshold = Min value > 0.6
Low Threshold = Max value <0.6
and others with varying range

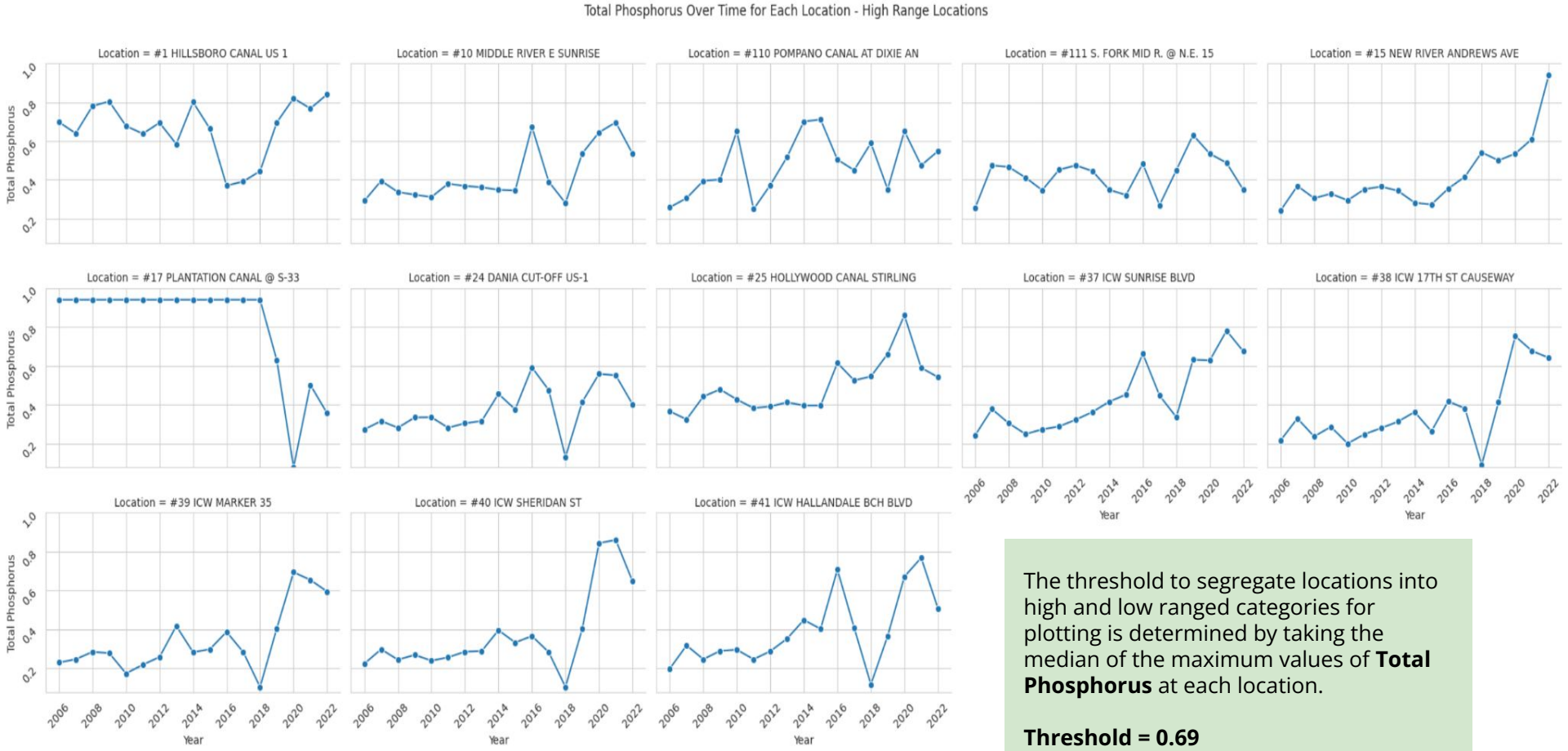
Min Value > 0.6


Specific Conductance Over Time for Each Location - Low Range Locations
Location = #110 POMPANO CANAL AT DIXIE AN
Location = #12 MIDDLE RIVER NW 31ST AVE
Location = #22 N NEW RIVER SW 125 AVE
Location = #28 S NEW RIVER CANAL FLAMINGO
Location = #31 SNAKE CRK CANAL FLAMINGO
Location = #32 C-9 CANAL US 27
Location = #4 HILLSBORO CANAL SE GROWERS
Location = #6 CYPRESS CREEK DIXIE HWY
Location = #7 CYPRESS CREEK S. PALM AIRE
Location = #8 POMPANO CANAL US 441
Location = #89 NOB HILL RD POMPANO CANAL
Specific Conductance
0.05
0.10
0.15
0.20
2006
2008
2010
2012
2014
2016
2018
2020
2022
Year

Max Value < 0.6

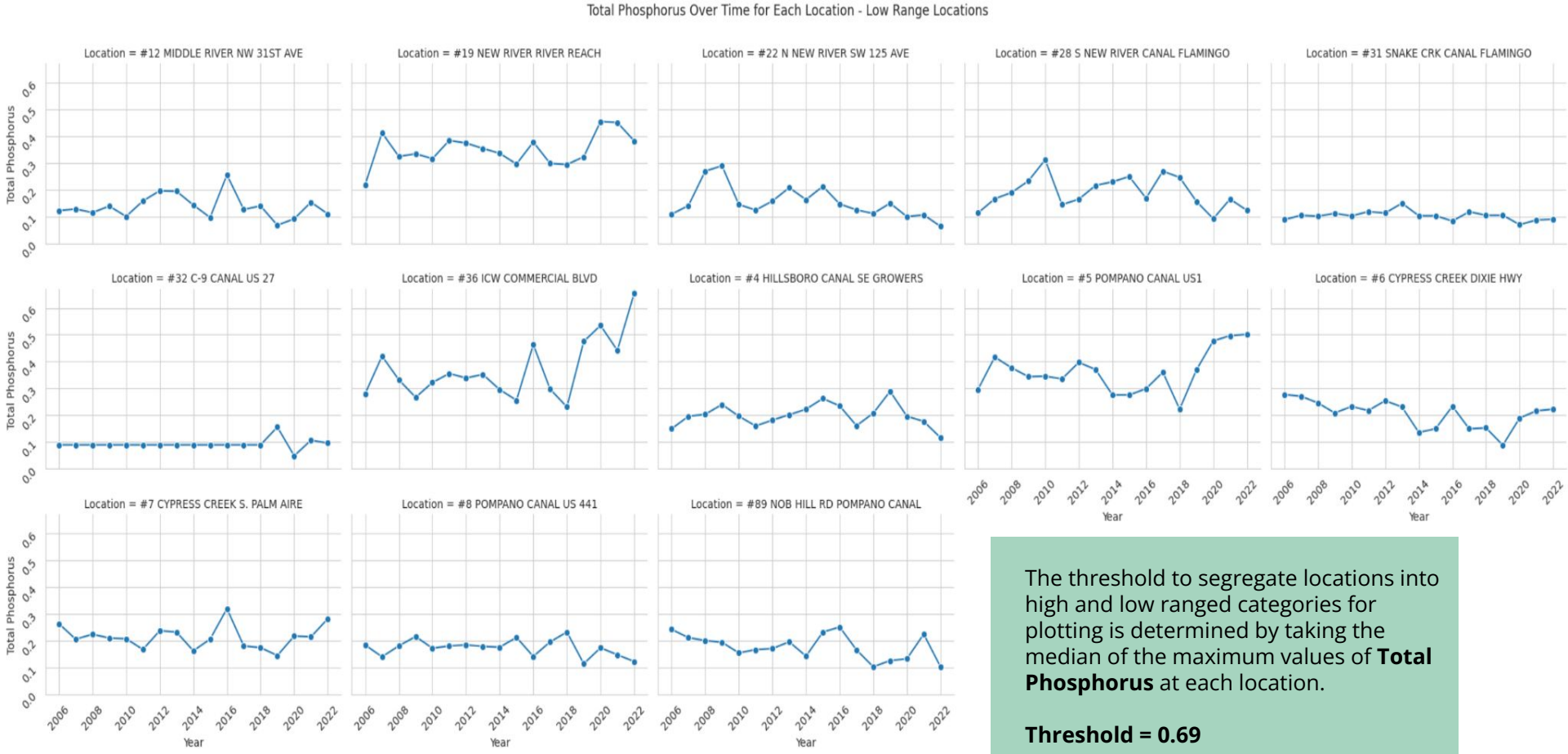


These locations has range of 0.4 or higher, henced grouped together.



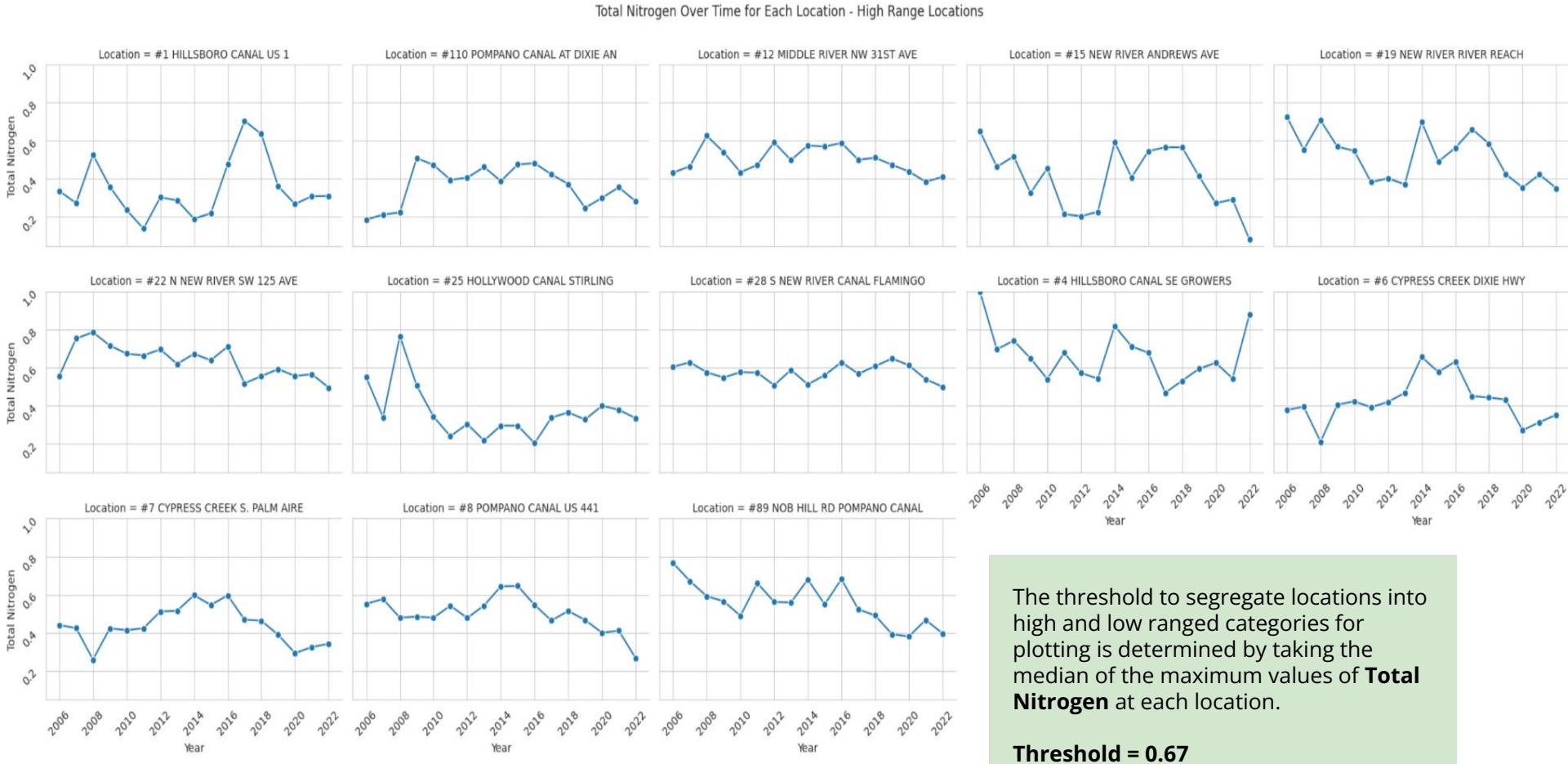
The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Total Phosphorus** at each location.

Threshold = 0.69



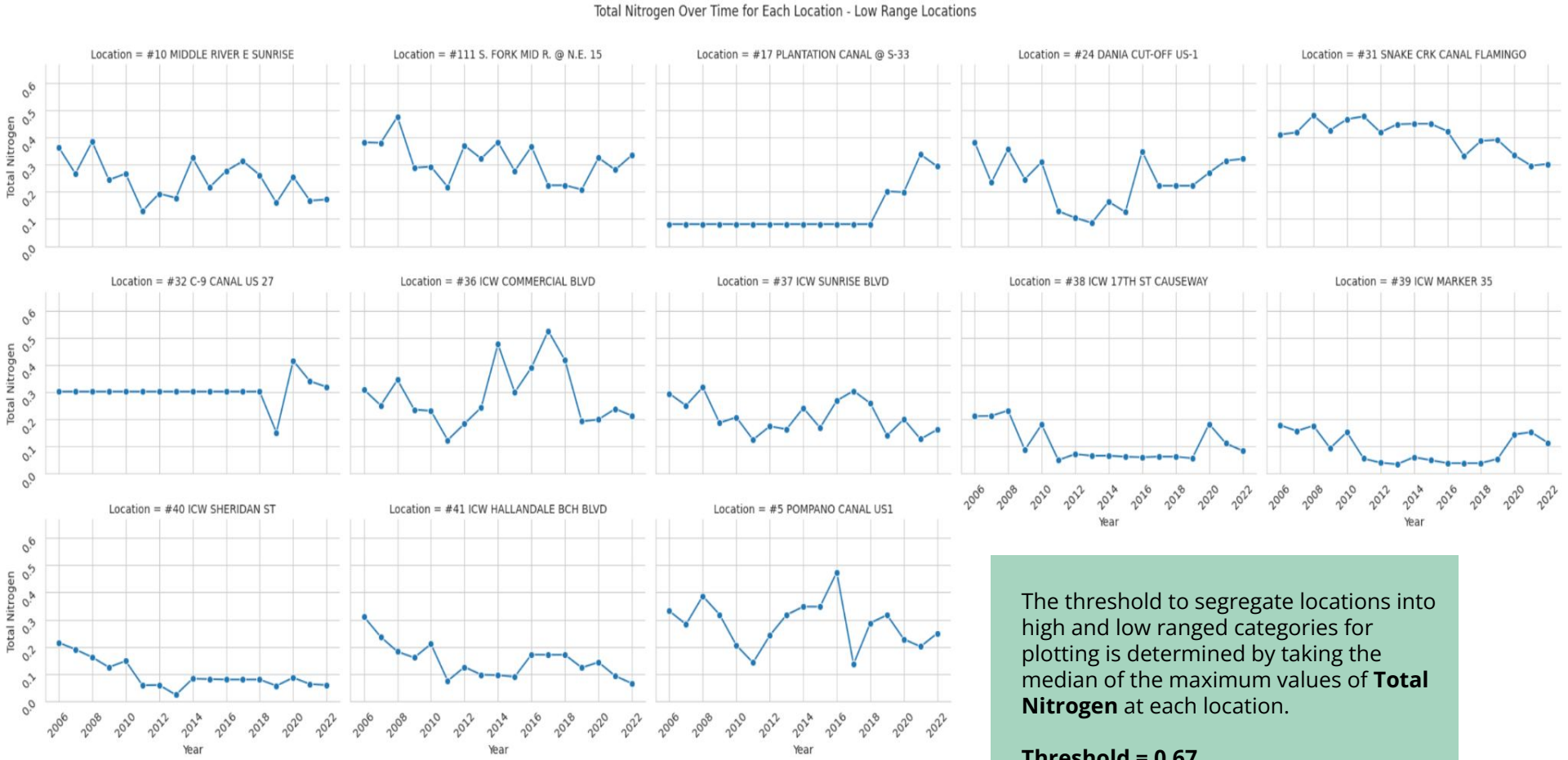
The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Total Phosphorus** at each location.

Threshold = 0.69



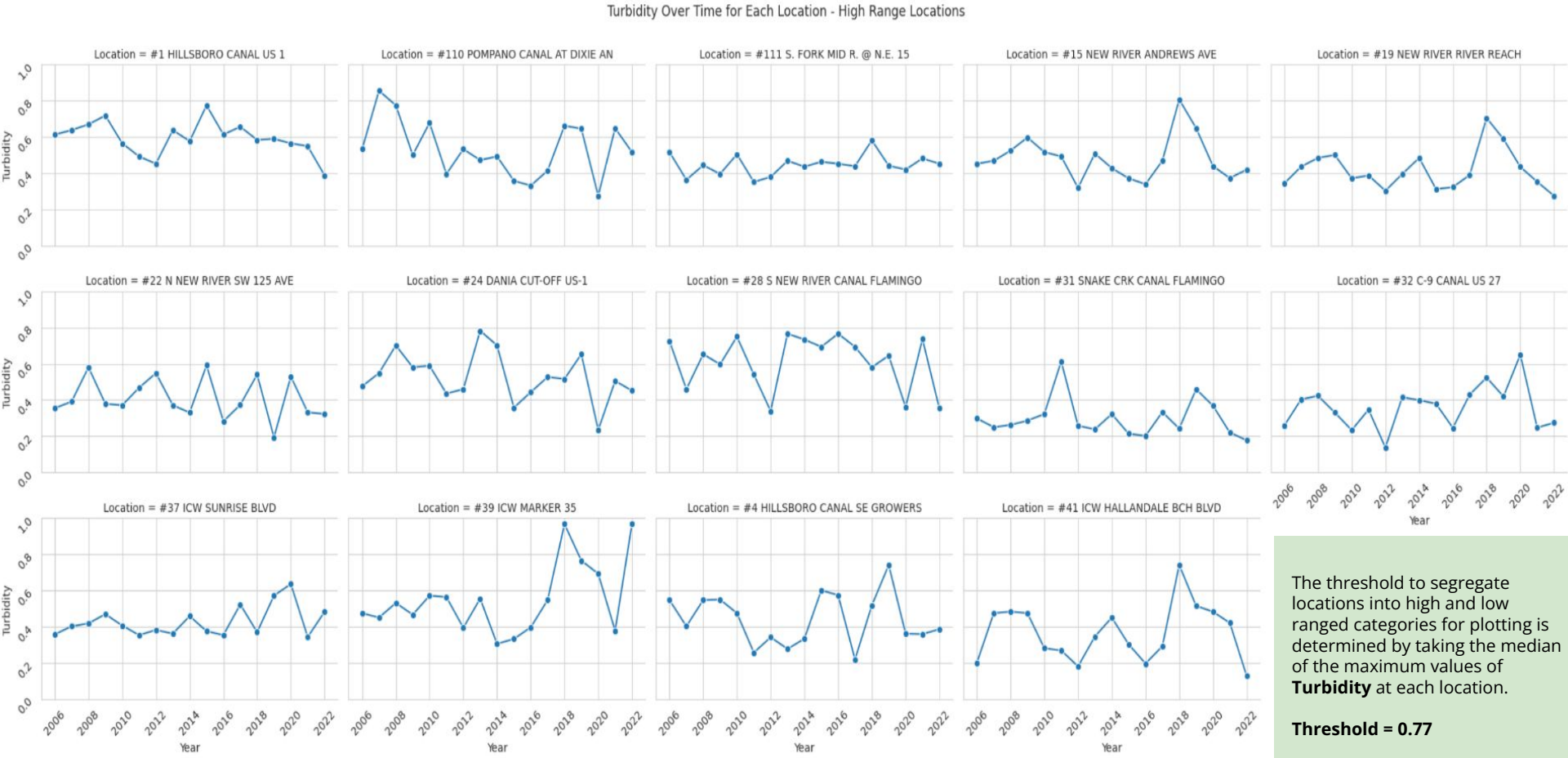
The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Total Nitrogen** at each location.

Threshold = 0.67



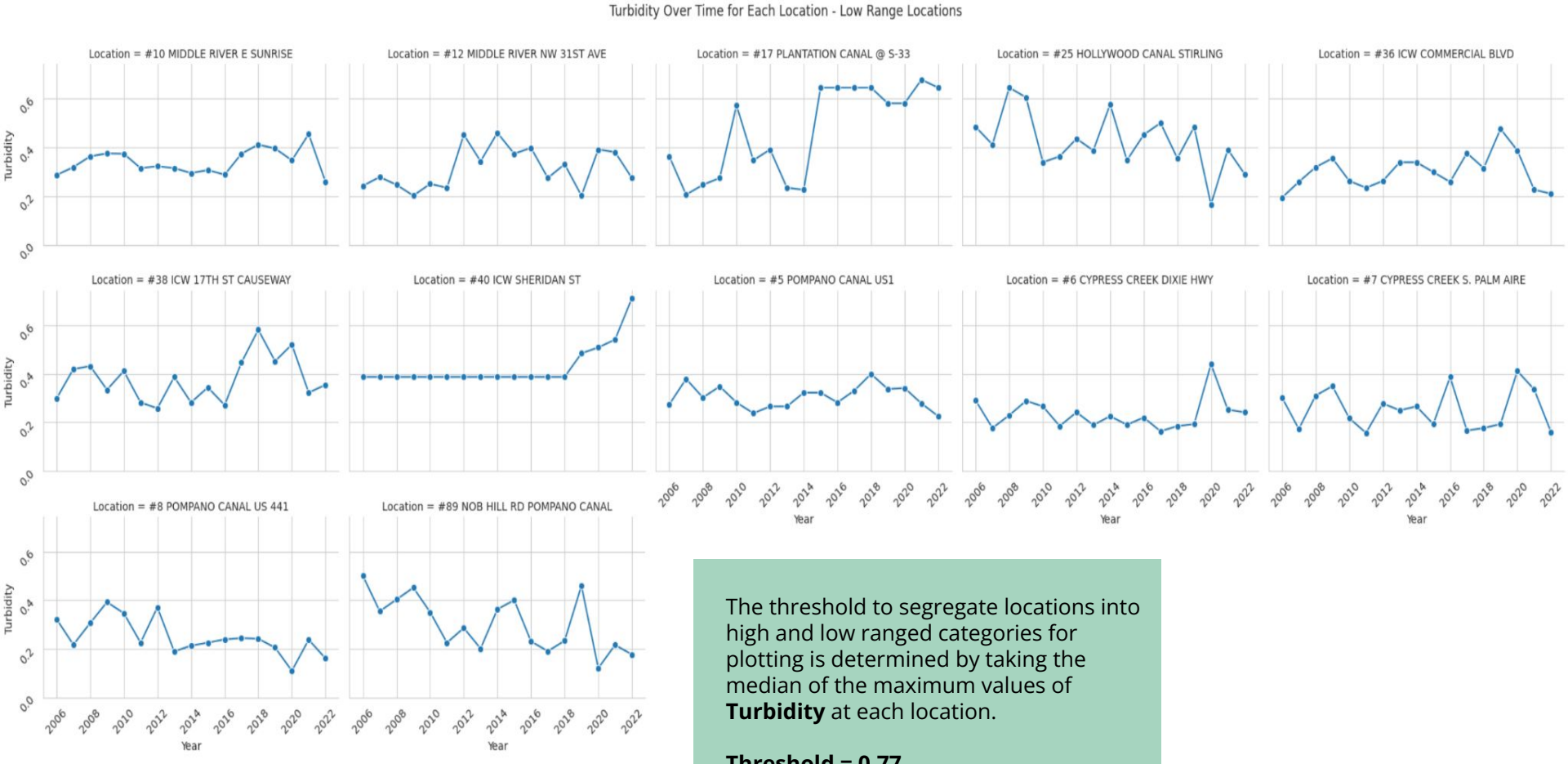
The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Total Nitrogen** at each location.

Threshold = 0.67



The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Turbidity** at each location.

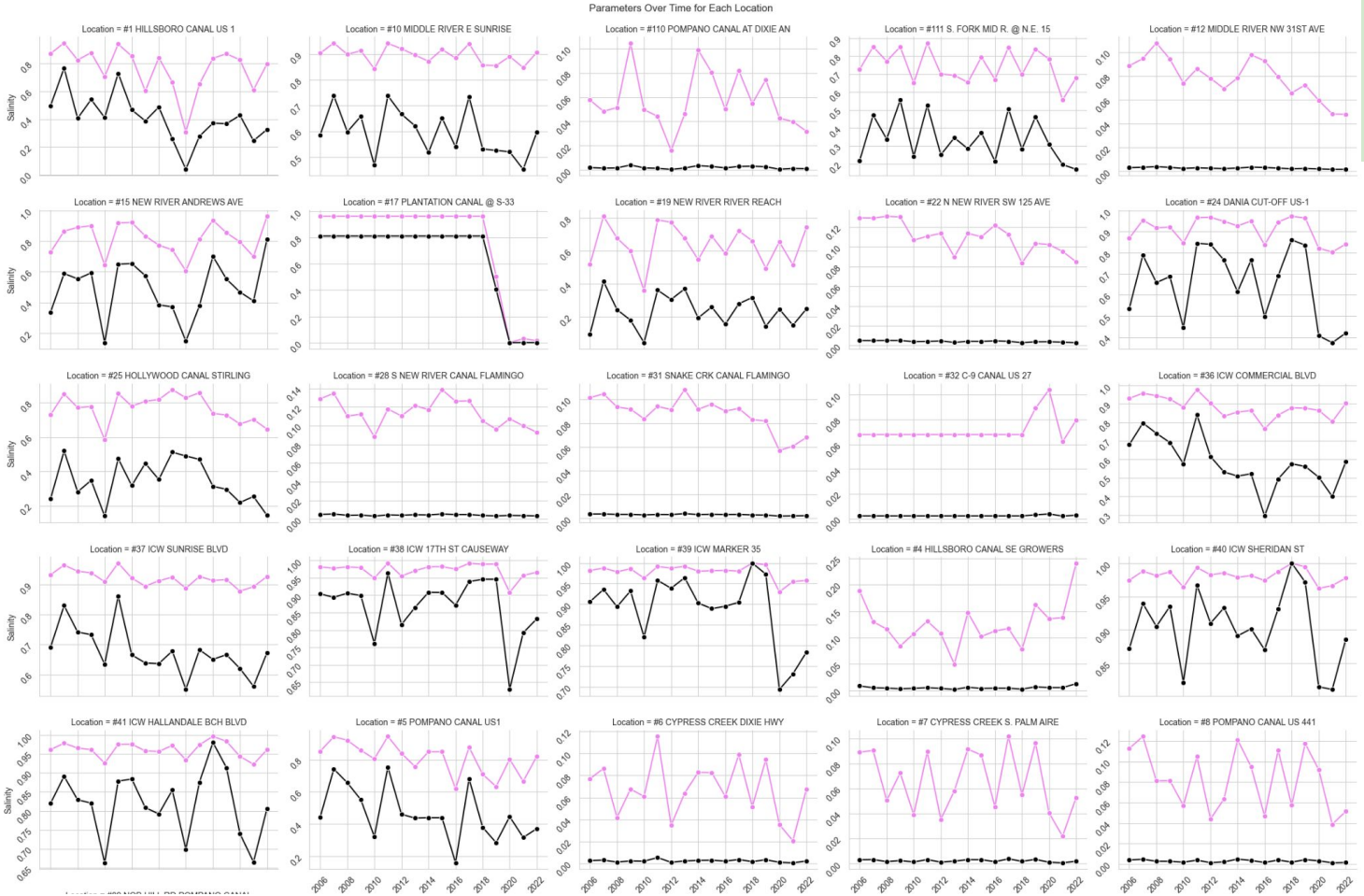
Threshold = 0.77



The threshold to segregate locations into high and low ranged categories for plotting is determined by taking the median of the maximum values of **Turbidity** at each location.

Threshold = 0.77

Correlation between Salinity and Specific Conductance



There is a strong **positive** relationship between salinity and specific conductance.

Correlation between Salinity and Total Nitrogen

There is a strong **negative** relationship between salinity and total nitrogen.

Our Approach

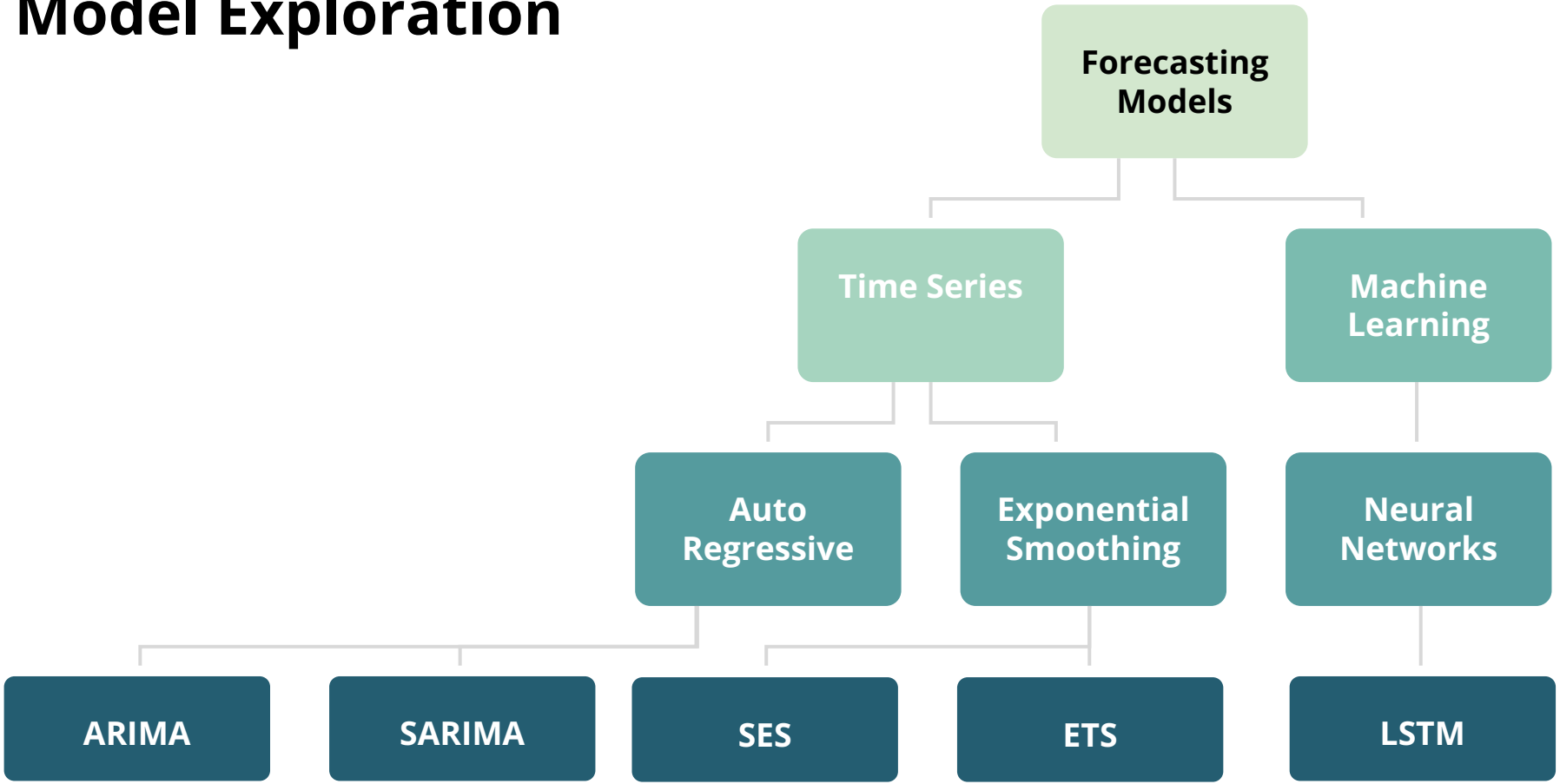
Researching forecasting models to determine the most suitable ones based on their strengths and limitations.

Evaluating model performance through plotting forecasting graphs, computing accuracies, and selecting the most optimal model.

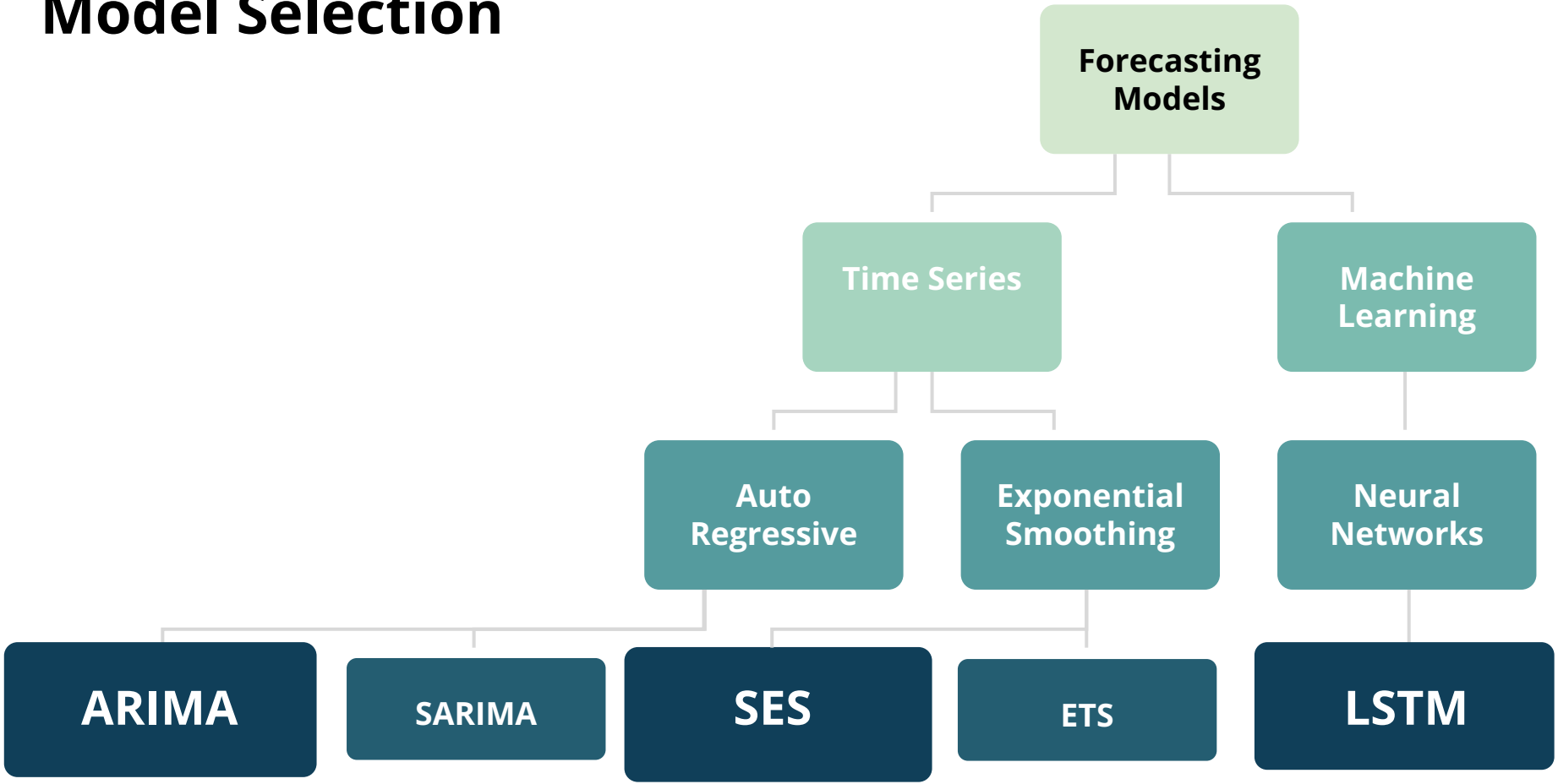
Conducting research on water quality categories and their corresponding parameters to gain insights into relevant factors.

Clustering the forecasted data into categories to predict the water quality based on it's Dissolved Oxygen Content.

Model Exploration



Model Selection



ARIMA (Autoregressive Integrated Moving Average)

Salinity

78%, 73% & 62%

On average for DO, Salinity and Total Phos

Graph shows which particular set of hyperparameters of the ARIMA model that gave the best accuracy for the respective locations

	Locations	Hyper Parameters	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
0	#1 HILLSBORO CANAL US 1	(1, 1, 1)	0.615309	0.958278	0.630289	0.319600	0.713442	0.811384	0.541153
1	#10 MIDDLE RIVER E SUNRISE	(1, 1, 2)	0.577258	0.945922	0.577939	0.837851	0.511905	0.244174	0.506421
2	#110 POMPANO CANAL AT DIXIE AN	(1, 0, 2)	0.679351	0.821200	0.997496	0.915108	0.756264	0.691273	0.193071
3	#111 S. FORK MID R. @ N.E. 15	(1, 0, 2)	0.160002	0.779321	0.350263	0.000000	0.571161	0.554009	0.754002
4	#12 MIDDLE RIVER NW 31ST AVE	(1, 1, 2)	0.590938	0.692889	0.998896	0.949559	0.797375	0.880907	0.635956
5	#15 NEW RIVER ANDREWS AVE	(1, 1, 2)	0.477718	0.783551	0.063842	0.249821	0.000000	0.306695	0.622171
6	#17 PLANTATION CANAL @ S-33	(1, 1, 2)	0.433728	0.544081	0.996704	0.897592	0.459240	0.000000	0.668987
7	#19 NEW RIVER RIVER REACH	(1, 1, 1)	0.392016	0.813423	0.575921	0.160165	0.462890	0.692295	0.590365
8	#22 N NEW RIVER SW 125 AVE	(1, 1, 1)	0.770236	0.662498	0.998240	0.961465	0.765539	0.847217	0.647648
9	#24 DANIA CUT-OFF US-1	(1, 1, 1)	0.803692	0.849058	0.000000	0.554302	0.561686	0.449060	0.810092
10	#25 HOLLYWOOD CANAL STIRLING	(1, 1, 2)	0.197860	0.770228	0.398214	0.502433	0.823984	0.119383	0.753700
11	#28 S NEW RIVER CANAL FLAMINGO	(1, 1, 2)	0.000000	0.653208	0.999275	0.973445	0.741624	0.824938	0.325621
12	#31 SNAKE CRK CANAL FLAMINGO	(1, 1, 1)	0.736587	0.780983	0.998025	0.928546	0.723304	0.891466	0.643240
13	#32 C-9 CANAL US 27	(1, 0, 2)	0.510375	0.000000	0.996395	0.918389	0.832662	0.880010	0.289753
14	#36 ICW COMMERCIAL BLVD	(1, 1, 1)	0.163357	0.966647	0.634508	0.792700	0.816777	0.527719	0.343031
15	#37 ICW SUNRISE BLVD	(1, 1, 2)	0.152149	0.980556	0.720593	0.931378	0.551390	0.269545	0.330255
16	#38 ICW 17TH ST CAUSEWAY	(1, 1, 1)	0.850319	0.984399	0.820355	0.967077	0.746818	0.740071	0.383638
17	#39 ICW MARKER 35	(1, 1, 1)	0.980262	0.891307	0.759936	0.968016	0.871902	0.672555	0.000000
18	#4 HILLSBORO CANAL SE GROWERS	(1, 0, 2)	0.208269	0.482041	0.984889	0.734858	0.381302	0.782173	0.751706
19	#40 ICW SHERIDAN ST	(1, 1, 2)	0.975861	0.975012	0.739819	0.952875	0.866164	0.519485	0.465635
20	#41 ICW HALLANDALE BCH BLVD	(1, 1, 1)	0.497815	0.810965	0.550009	0.871034	0.692308	0.437778	0.138402
21	#5 POMPANO CANAL US1	(1, 1, 1)	0.503494	0.761710	0.193020	0.202042	0.557065	0.627678	0.598871
22	#6 CYPRESS CREEK DIXIE HWY	(1, 0, 2)	0.607457	0.864985	0.992253	0.799904	0.803747	0.910813	0.718070
23	#7 CYPRESS CREEK S. PALM AIRE	(1, 1, 1)	0.286467	0.788156	0.994650	0.835139	0.894864	0.864145	0.756103
24	#8 POMPANO CANAL US 441	(1, 1, 2)	0.871339	0.736535	0.993687	0.816963	0.559662	0.853660	0.828449
25	#89 NOB HILL RD POMPANO CANAL	(1, 0, 2)	0.614432	0.889103	0.995206	0.867729	0.625870	0.642274	0.527332

Int

On an average across all the locations, ARIMA gave the best accuracy for Dissolved Oxygen, Salinity and Total Phosphorus

SES (Simple Exponential Smoothing)

Chlorophyll A

61% and 74%

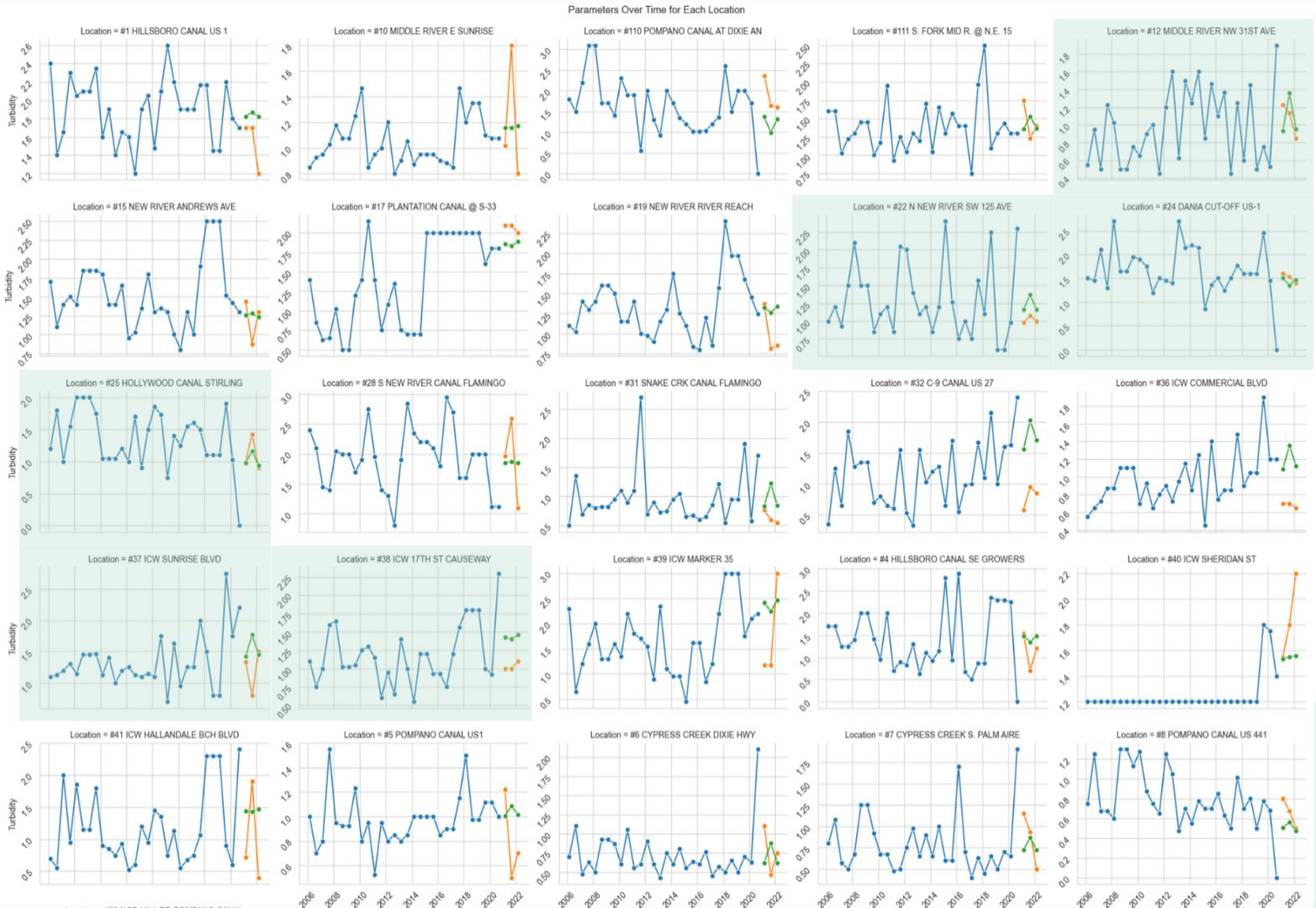
On average for Chlorophyll A and Specific Conductance

Graph shows which particular set of hyperparameters of the SES model that gave the best accuracy for the respective locations

SES (Simple Exponential Smoothing)

	Locations	Alpha	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
0	#1 HILLSBORO CANAL US 1	0.4	0.732455	0.898570	0.426741	0.389309	0.985889	0.840572	0.638950
1	#10 MIDDLE RIVER E SUNRISE	0.3	0.608503	0.908309	0.723417	0.897670	0.444506	0.000000	0.550351
2	#110 POMPANO CANAL AT DIXIE AN	0.3	0.598250	0.695327	0.998290	0.934680	0.674383	0.675148	0.254915
3	#111 S. FORK MID R. @ N.E. 15	0.1	0.463590	0.699853	0.261389	0.109638	0.439183	0.544568	0.772874
4	#12 MIDDLE RIVER NW 31ST AVE	0.6	0.254291	0.552112	0.999598	0.983967	0.780408	0.843465	0.632738
5	#15 NEW RIVER ANDREWS AVE	0.9	0.775197	0.608555	0.000000	0.275364	0.160778	0.024137	0.728827
6	#17 PLANTATION CANAL @ S-33	0.3	0.596886	0.000000	0.044772	0.000000	0.141862	0.687439	0.743859
7	#19 NEW RIVER RIVER REACH	0.5	0.544133	0.714239	0.571283	0.300581	0.506381	0.602085	0.491294
8	#22 N NEW RIVER SW 125 AVE	0.4	0.809626	0.432837	0.998291	0.969384	0.726942	0.762253	0.500798
9	#24 DANIA CUT-OFF US-1	0.5	0.835747	0.784662	0.257536	0.695855	0.423802	0.325388	0.331534
10	#25 HOLLYWOOD CANAL STIRLING	0.3	0.418185	0.653989	0.545828	0.660375	0.665071	0.319207	0.691474
11	#28 S NEW RIVER CANAL FLAMINGO	0.2	0.182308	0.362659	0.998564	0.961997	0.494817	0.723828	0.335805
12	#31 SNAKE CRK CANAL FLAMINGO	0.3	0.821087	0.661578	0.997895	0.935293	0.641842	0.856924	0.401670
13	#32 C-9 CANAL US 27	0.1	0.736311	0.557300	0.998290	0.952220	0.786017	0.831209	0.417674
14	#36 ICW COMMERCIAL BLVD	0.9	0.000000	0.868574	0.695373	0.832992	0.877745	0.594954	0.459917
15	#37 ICW SUNRISE BLVD	0.4	0.725255	0.943764	0.811346	0.928652	0.499175	0.101167	0.201676
16	#38 ICW 17TH ST CAUSEWAY	0.2	0.875203	0.985840	0.931596	0.980728	0.916096	0.263656	0.539145
17	#39 ICW MARKER 35	0.5	0.986553	0.823465	0.914082	0.961804	0.852658	0.886522	0.014359
18	#4 HILLSBORO CANAL SE GROWERS	0.5	0.159105	0.218628	0.984763	0.793355	0.000000	0.665875	0.642914
19	#40 ICW SHERIDAN ST	0.5	0.962312	0.912189	0.863613	0.982767	0.826931	0.539364	0.555442
20	#41 ICW HALLANDALE BCH BLVD	0.5	0.613816	0.667210	0.615565	0.906624	0.586863	0.150922	0.000000
21	#5 POMPANO CANAL US1	0.9	0.587615	0.701232	0.200194	0.257279	0.534973	0.455210	0.621162
22	#6 CYPRESS CREEK DIXIE HWY	0.6	0.693028	0.392288	0.996933	0.921561	0.805622	0.859478	0.191357
23	#7 CYPRESS CREEK S. PALM AIRE	0.5	0.399343	0.304084	0.997090	0.920114	0.842902	0.834040	0.495257
24	#8 POMPANO CANAL US 441	0.9	0.834263	0.261978	0.994437	0.854290	0.581287	0.791570	0.378711
25	#89 NOB HILL RD POMPANO CANAL	0.2	0.714157	0.788706	0.996332	0.906985	0.502884	0.461731	0.834051

On an average across all the locations, ARIMA gave the best accuracy for Chlorophyll A and Specific Conductance



LSTM (Long Short Term Memory)

Turbidity

60% and 67%

On average for Turbidity and Total Nitrogen

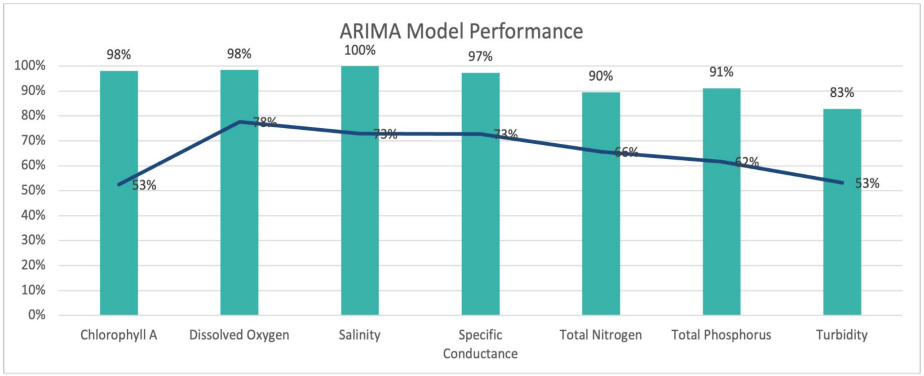
Graph shows which particular set of hyperparameters of the LSTM model that gave the best accuracy for the respective locations

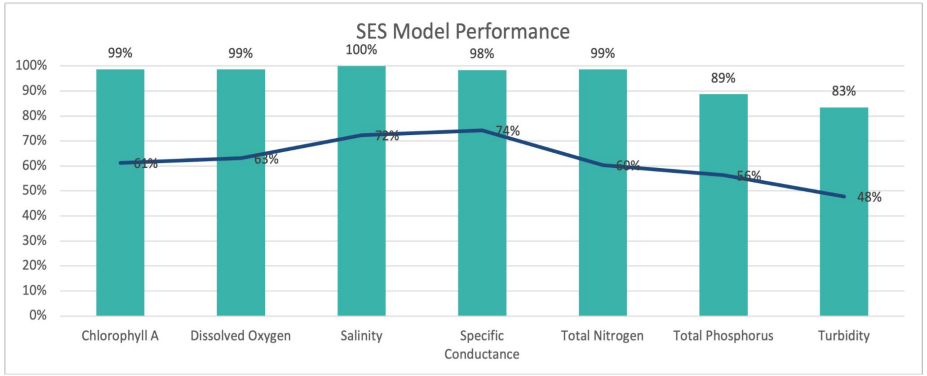
LSTM (Long Short Term Memory)

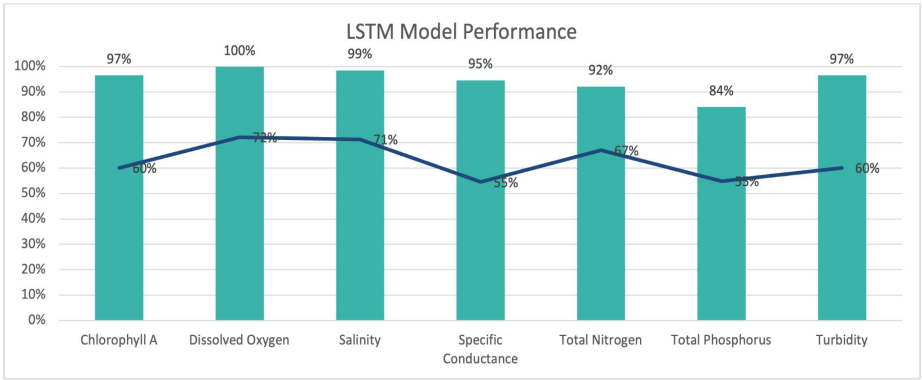
	Locations	Hyper Parameters	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
0	#1 HILLSBORO CANAL US 1	(64, relu)	0.746131	0.911175	0.542255	0.480485	0.937129	0.879653	0.570194
1	#10 MIDDLE RIVER E SUNRISE	(128, relu)	0.651612	0.926862	0.654565	0.885884	0.484212	0.416116	0.491947
2	#110 POMPANO CANAL AT DIXIE AN	(128, relu)	0.507180	0.641647	0.997984	0.912302	0.728476	0.913899	0.285190
3	#111 S. FORK MID R. @ N.E. 15	(64, relu)	0.437146	0.684312	0.278604	0.000000	0.453301	0.815937	0.745254
4	#12 MIDDLE RIVER NW 31ST AVE	(64, tanh)	0.852772	0.525952	0.997633	0.921578	0.336482	0.929061	0.742232
5	#15 NEW RIVER ANDREWS AVE	(64, relu)	0.612421	0.603855	0.037535	0.214929	0.000000	0.575114	0.549584
6	#17 PLANTATION CANAL @ S-33	(64, tanh)	0.518271	0.000000	0.750592	0.524140	0.392903	0.622836	0.658256
7	#19 NEW RIVER RIVER REACH	(64, tanh)	0.542802	0.609417	0.621468	0.234432	0.422708	0.814842	0.531703
8	#22 N NEW RIVER SW 125 AVE	(64, relu)	0.748946	0.362948	0.997674	0.946839	0.545782	0.871864	0.596310
9	#24 DANIA CUT-OFF US-1	(64, tanh)	0.846223	0.684151	0.000000	0.572020	0.359440	0.582018	0.853217
10	#25 HOLLYWOOD CANAL STIRLING	(128, relu)	0.603700	0.635467	0.435857	0.590904	0.664987	0.487247	0.707962
11	#28 S NEW RIVER CANAL FLAMINGO	(64, tanh)	0.000000	0.253643	0.998582	0.948766	0.676000	0.868542	0.263819
12	#31 SNAKE CRK CANAL FLAMINGO	(64, relu)	0.807646	0.533469	0.997710	0.920521	0.492612	0.935137	0.590215
13	#32 C-9 CANAL US 27	(64, relu)	0.724000	0.524320	0.986474	0.556125	0.889608	0.921844	0.187187
14	#36 ICW COMMERCIAL BLVD	(128, relu)	0.591903	0.921267	0.527910	0.795159	0.850300	0.618467	0.542188
15	#37 ICW SUNRISE BLVD	(64, tanh)	0.623039	0.942090	0.653748	0.906687	0.518595	0.436406	0.615457
16	#38 ICW 17TH ST CAUSEWAY	(64, tanh)	0.894359	0.983022	0.788587	0.909631	0.855682	0.000000	0.683768
17	#39 ICW MARKER 35	(64, relu)	0.981959	0.760684	0.773944	0.992651	0.791925	0.643656	0.000000
18	#4 HILLSBORO CANAL SE GROWERS	(128, relu)	0.345500	0.214114	0.983269	0.714956	0.092443	0.867196	0.652718
19	#40 ICW SHERIDAN ST	(64, relu)	0.963828	0.930230	0.737585	0.954589	0.849764	0.411580	0.357207
20	#41 ICW HALLANDALE BCH BLVD	(64, tanh)	0.610759	0.694954	0.635616	0.892517	0.651158	0.390703	0.190950
21	#5 POMPANO CANAL US1	(64, relu)	0.621212	0.632710	0.199365	0.230614	0.497207	0.693212	0.606824
22	#6 CYPRESS CREEK DIXIE HWY	(128, relu)	0.670970	0.792526	0.993887	0.827130	0.758244	0.940530	0.695569
23	#7 CYPRESS CREEK S. PALM AIRE	(64, tanh)	0.408712	0.577649	0.995277	0.860722	0.791215	0.915393	0.709355
24	#8 POMPANO CANAL US 441	(128, relu)	0.838808	0.424471	0.994762	0.865235	0.309305	0.889431	0.859026
25	#89 NOB HILL RD POMPANO CANAL	(64, tanh)	0.765450	0.736183	0.995923	0.900105	0.563838	0.790978	0.734563

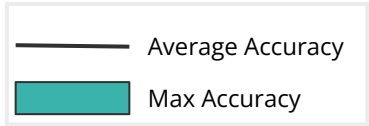
On an average across all the locations, LSTM gave the best accuracy for Turbidity and Total Nitrogen

Model Performance

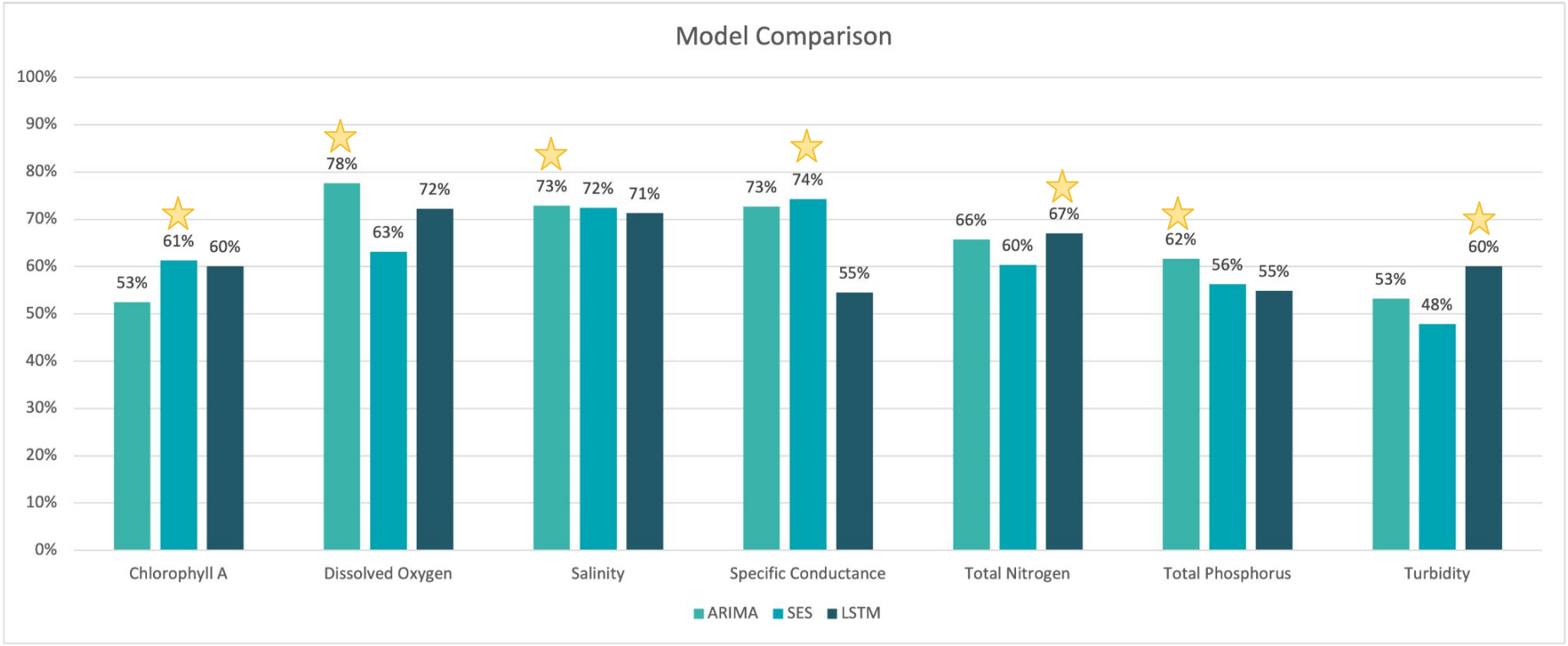




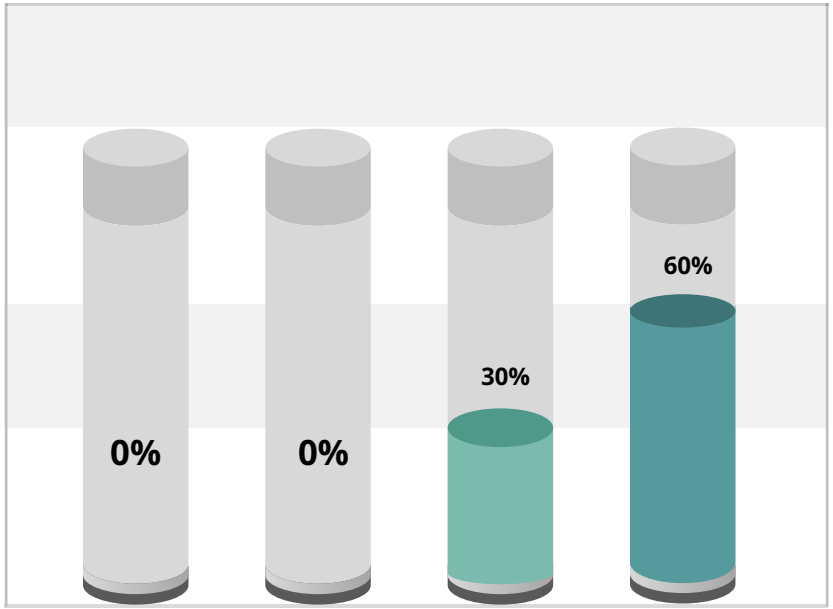


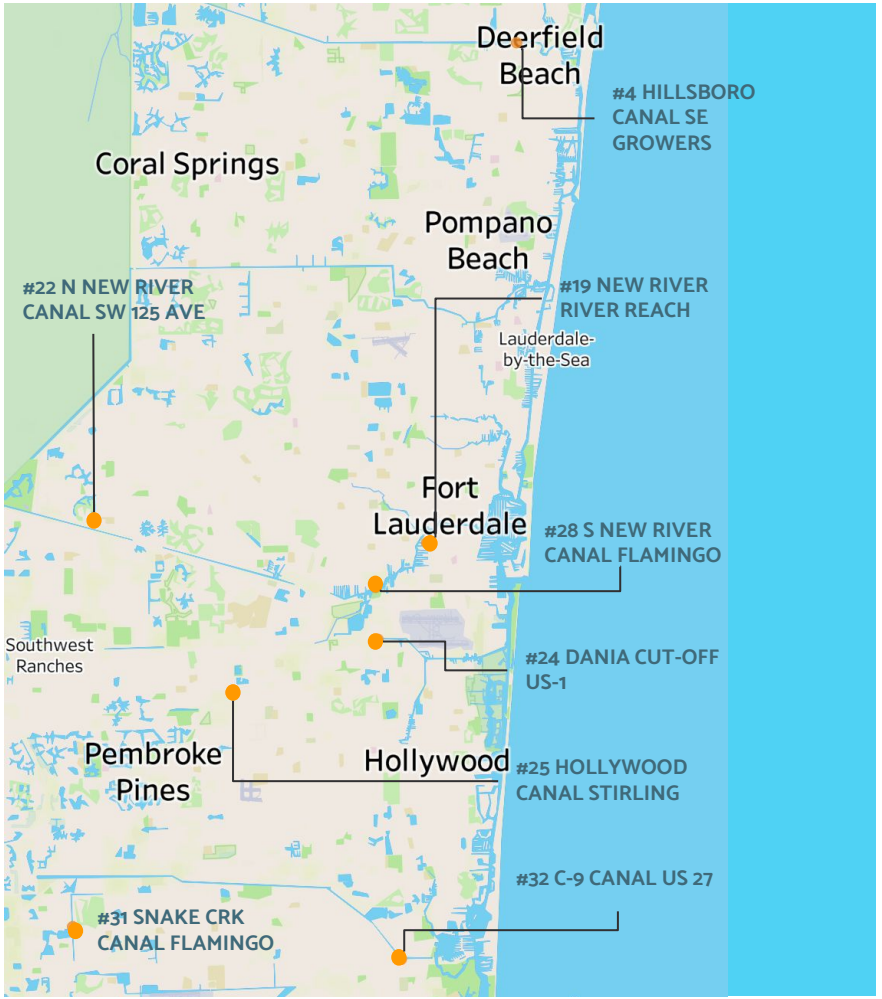


Model Comparison



Predicting Water Quality





LOCATIONS WITH BIOLOGICAL STRESS

Conclusion of Hypothesis

Comments on our Initial Hypothesis after modeling and results:

There could be significant correlation between the provided parameters. These correlations can be used to predict one parameter based on others, improving the efficiency.

Different machine learning algorithms will vary in their effectiveness at forecasting water quality parameters.

Seasonality in data significantly influence water quality parameters.

Improvements and Suggestions

Comments from the Team on the Project Outcome :

1

Exploring Additional Parameters:

There are other parameters we need to consider to enhance our understanding of the data, for example Tidal considerations, Rainfall, Temperature.

2

Lack of Data Consistency:

Based on our current approach, we believe that having biweekly or monthly data will given train the model better and we can expect better results.


NYU
TANDON SCHOOL
OF ENGINEERING
Thank You for
your time!