Mini project 2: primary productivity in coastal waters

In this project you're again given a dataset and some questions. The data for this project come from the EPA's National Aquatic Resource Surveys, and in particular the National Coastal Condition Assessment (NCCA); broadly, you'll do an exploratory analysis of primary productivity in coastal waters.

By way of background, chlorophyll A is often used as a proxy for primary productivity in marine ecosystems; primary producers are important because they are at the base of the food web. Nitrogen and phosphorus are key nutrients that stimulate primary production.

In the data folder you'll find water chemistry data, site information, and metadata files. It might be helpful to keep the metadata files open when tidying up the data for analysis. It might also be helpful to keep in mind that these datasets contain a considerable amount of information, not all of which is relevant to answering the questions of interest. Notice that the questions pertain somewhat narrowly to just a few variables. It's recommended that you determine which variables might be useful and drop the rest.

As in the first mini project, there are accurate answers to each question that are mutually consistent with the data, but there aren't uniquely correct answers. You will likely notice that you have even more latitude in this project than in the first, as the questions are slightly broader. Since we've been emphasizing visual and exploratory techniques in class, you are encouraged (but not required) to support your answers with graphics.

The broader goal of these mini projects is to cultivate your problem-solving ability in an unstructured setting. Your work will be evaluated based on the following:

- choice of method(s) used to answer questions;
- clarity of presentation;
- code style and documentation.

Please write up your results separately from your codes; codes should be included at the end of the notebook.

Name: Preeti Kulkarni

Collaborators: TJ Sipin, Gian Taparan

Part 1: dataset

Merge the site information with the chemistry data and tidy it up. Determine which columns to keep based on what you use in answering the questions in part 2; then, print the first few rows here (but *do not include your codes used in tidying the data*) and write a brief description (1-2 paragraphs) of the dataset conveying what you take to be the key attributes. Direct your description to a reader unfamiliar with the data; ensure that in your data preview the columns are named intelligibly.

Suggestion: export your cleaned data as a separate `.csv` file and read that directly in below, as in: `pd.read_csv('YOUR DATA FILE').head()`.

In [5]:
```
# show a few rows of clean data
import pandas as pd
import numpy as np
import altair as alt
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 100)
```

In [47]:
```
data.head()
```

Out[47]:

	UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite	Total Nitrogen
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	0.000	3.34	0.014	0.028	0.014	0.40750
1	60	CA	7/1/2010	San Diego Bay	West	3.5	32.71424	-117.23527	0.010	2.45	0.020	0.026	0.010	0.23000
2	61	CA	7/1/2010	Mission Bay	West	2.2	32.78372	-117.22132	0.000	3.82	0.009	0.030	0.009	0.33625
3	62	CA	7/1/2010	San Diego Bay	West	9.5	32.72245	-117.20443	0.000	6.13	0.010	0.028	0.010	0.23875
4	63	NC	6/9/2010	White Oak River	Southeast	1.0	34.75098	-77.12117	0.002	9.79	0.030	0.043	0.028	0.63250

The key variables would be the UID, State, Ammonia, Total Nitrogen, Total Phosphorus, and Chlorophyll A. The dataset is over the year 2010, and takes several bodies of water in different regions to measure. Each observation includes date collected as well as longitude and latitude, as well as the water depth and Waterbody name. The Ammonia, Total Nitrogen, and Total Phosphorus are all nutrients that have a relationship with Chlorophyll A, representing the productivity. By using different Regions, we can identify how these relationships work.

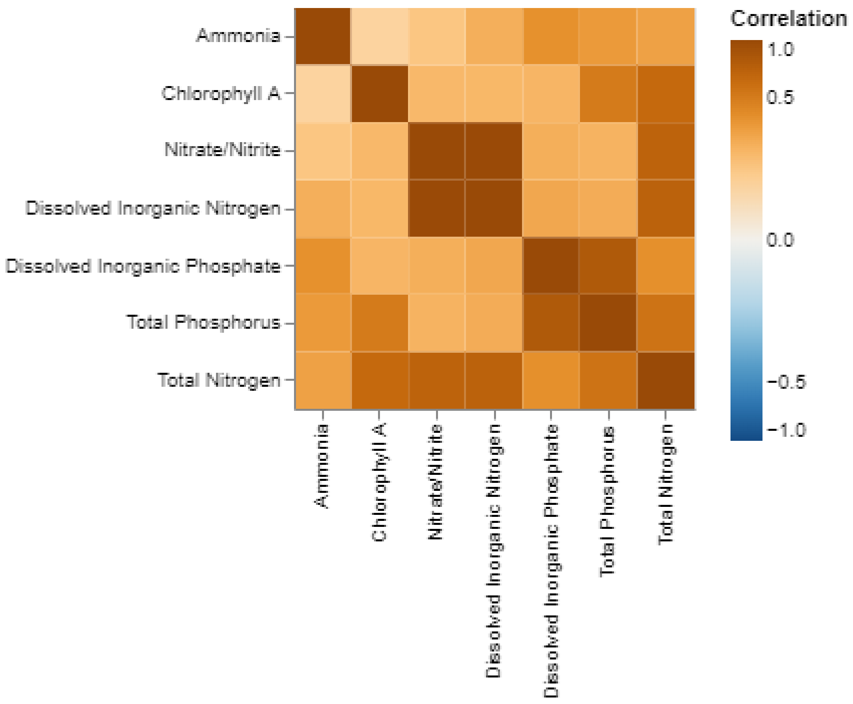
Part 2: exploratory analysis

Answer each question below and provide a visualization supporting your answer. A description and interpretation of the visualization should be offered.

Comment: you can either designate your plots in the codes section with clear names and reference them in your answers; or you can export your plots as image files and display them in markdown cells.

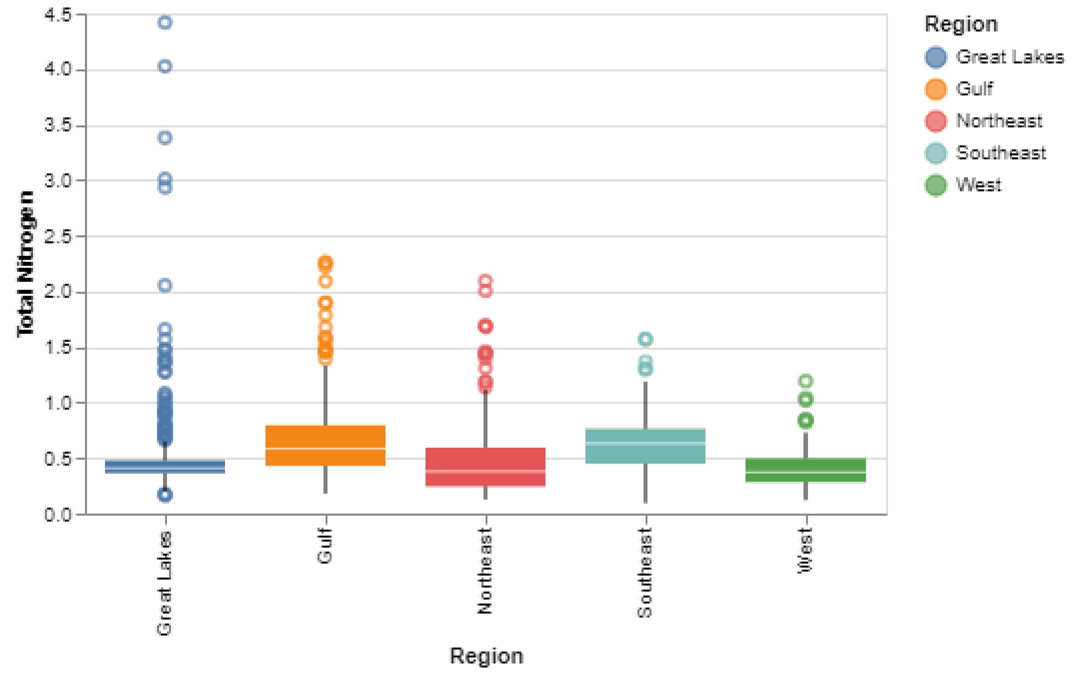
What is the apparent relationship between nutrient availability and productivity?

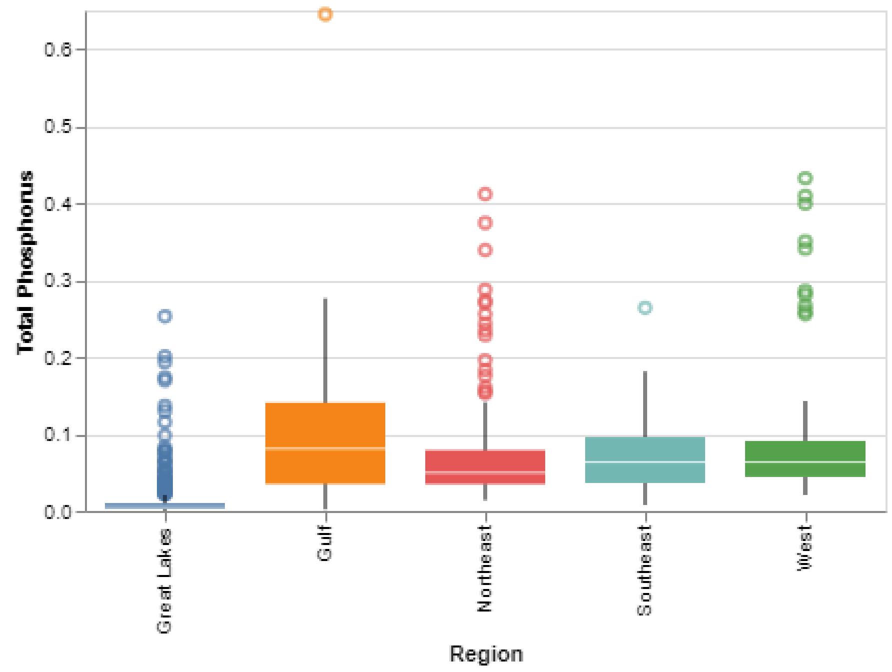
Comment: it's fine to examine each nutrient -- nitrogen and phosphorus -- separately, but do consider whether they might be related to each other.

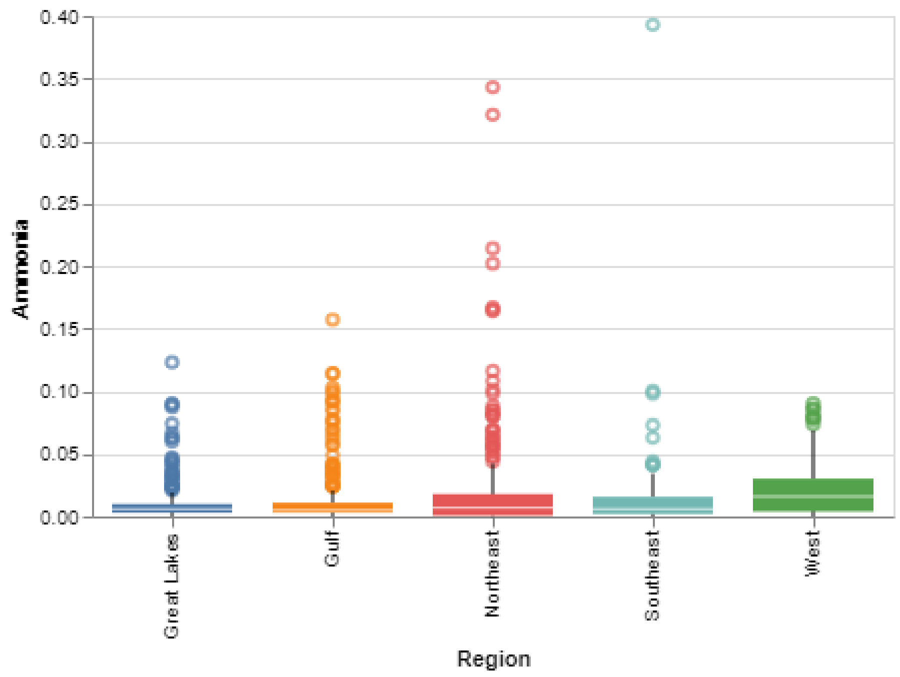


We can see a strong correlation between Chlorophyll A and the Total Nitrogen as well as Total Phosphorus and Chlorophyll A. There is also a relationship about Total Phosphate and Total Nitrogen. None of the variables have a negative correlation between each other, and all have some correlation between each other. There is a higher correlation between Total Phosphorus and Total Nitrogen with Chlorophyll A than it is with Ammonia. It seems that with more nutrients, as shown in the scatter panel, lead to more productivity.

Are there any notable differences in available nutrients among U.S. coastal regions?



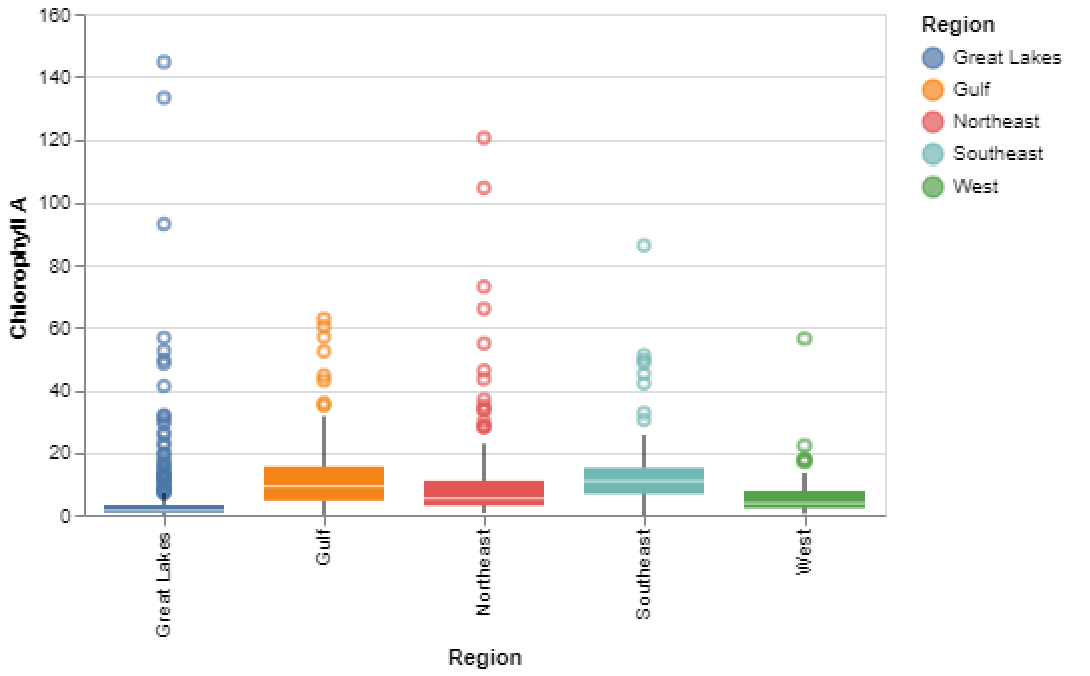




We can see that in the all regions there is a higher amount of Total Nitrogen than Phosphate. The Great Lake has the smallest median of Total Nitrogen, but the greatest number of outliers compared to the other regions. This might be because there is a buildup of Nitrogen, and unlike the West or Gulf, there is less flow of fresh water. The Gulf has the highest median of Total Nitrogen the next most outliers after the Gulf. The West has the highest Total Phosphorus levels, which may be due to the amount of Agricultural practices so close to the water. The Gulf has a concentrated amount of Phosphorus while other regions have many outliers. This is also the case with the Southeast in regards to Phosphorus, where there is only one outlier. The Northeast has the highest production of Ammonia and amount of outliers compared to other region. Generally, the West has a lower variablity of Ammonia due to the smaller amount of outliers and higher concentration.

Based on the 2010 data, does productivity seem to vary geographically in some way?

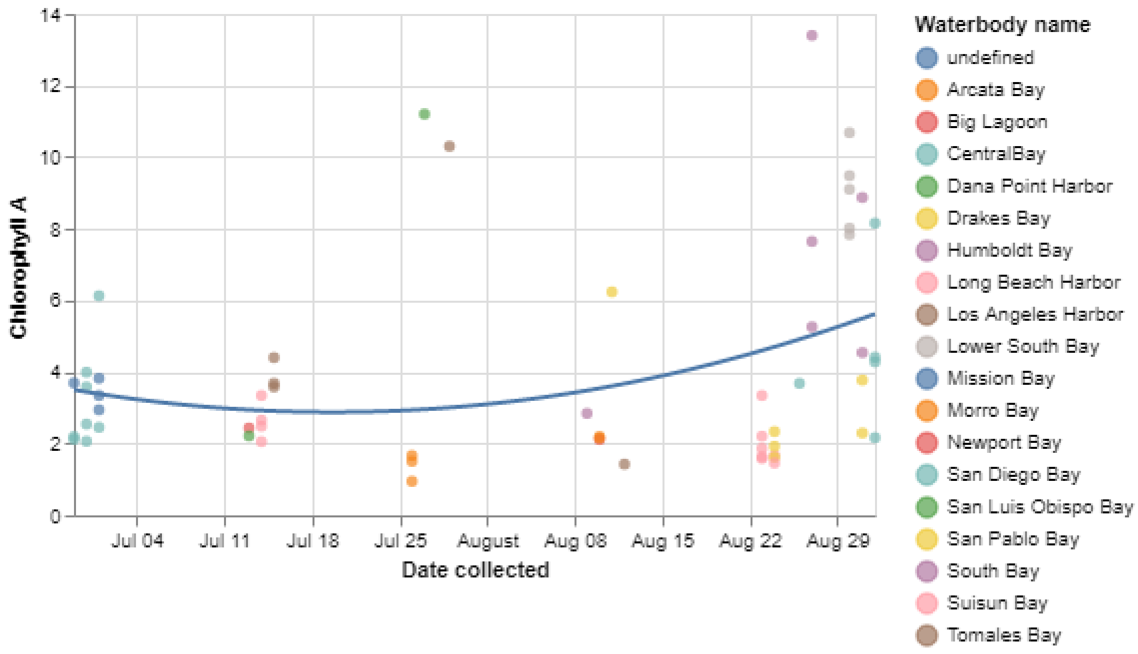
If so, explain how; If not, explain what options you considered and ruled out.



Based on the 2010 data, producitivity does vary by region. Geographically, there are different amounts of Ammonia, Nitrogen, and Phosphorus in different bodies of water. The amount of nutrients in the water could be due to positive or negative reinforcements in the society around them, or how well connected they are to a moving water source. We can see these trends following the graph, showing that productivity differs. If we look at the Northeast, we can see that there are consistently higher levels of Ammonia, Nitrogen, and Phosphorus compared to the other regions, perhaps the cause of the high productivity in this region. The West has a lower productivity of Chlorophyll, with a small group of outliers.

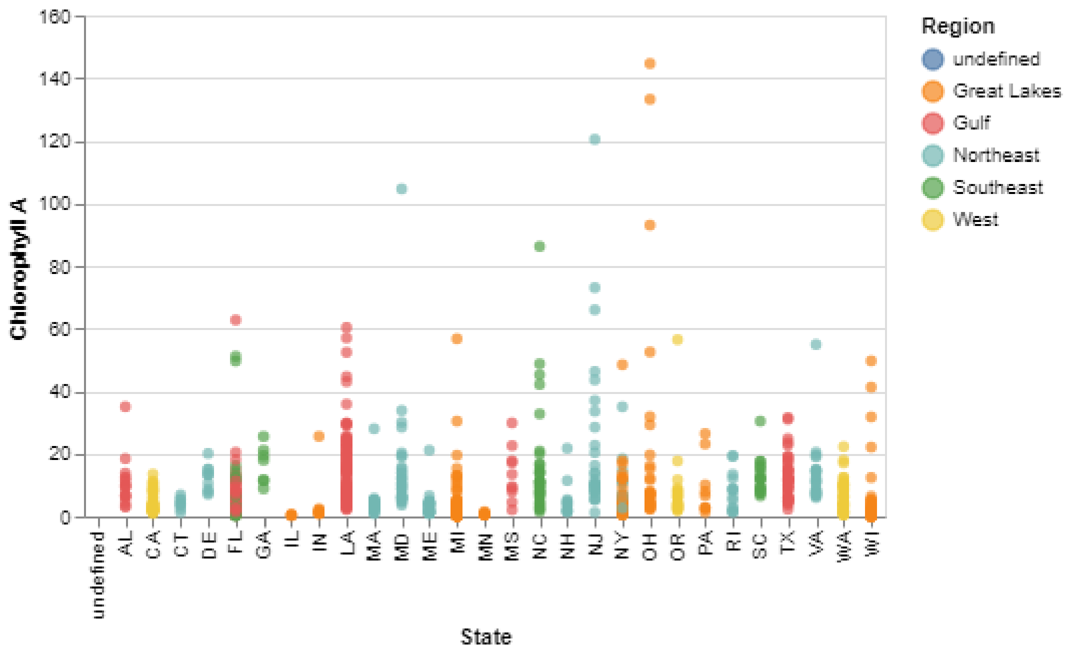
How does primary productivity in California coastal waters change seasonally in 2010, if at all?

Does your result make intuitive sense?



We can see that in California, there is a lot of variability between days. This does not allow us to accurately say if there is variability in CA in 2010 seasonally. It does look like there is a spike from July to August in some Waterbodys like CentralBay and San Diego Bay. Intuitively, this makes sense because these are the seasons to start the agricultural season. Thus, there will be more fertilizers and nutrients put into the water from the environment around it. Although we have a small sample size which also might prevent us from seeing seasonal trends clearly, the general trend is upwards and positive.

Pose and answer one additional question. How does productivity vary by state?



It seems as if Ohio has the highest productivity out of all the states. Initially, being from California I expected California to have the highest productibity because of how much farming happens. However, knowing that Ohio is a huge agricultural center, ie farming potatoes, this makes sense. There is a lot of untouched flat land which has great nutrient dense soil, which could add to the Chlorophyll amount. Illinois seems to be the smallest productivity in water, perhaps because there are no major bodies around to add nutrients, therefore increasing Chlorophyll A. We can also see that Illinois is in the Great Lake Region which we had previously discussed might not have as many nutrients because there is not a flowing water source, so water gets stagnant. California might also be particularly low because of the strict farming and pollution laws that do not exist in other states, which might be a factor in how many nutrients and what type of nutrients go into the soil.

Codes

Part 1: Tidy

In [7]:
```
import pandas as pd
import numpy as np
import altair as alt

ncca_raw = pd.read_csv('assessed_ncca2010_waterchem.csv')
ncca_sites = pd.read_csv('assessed_ncca2010_siteinfo.csv')
```

In [8]:
```
ncca_raw = pd.read_csv('assessed_ncca2010_waterchem.csv')
ncca_raw
ncca_sites = pd.read_csv('assessed_ncca2010_siteinfo.csv')
```

In [9]:
```
ncca_raw.head()
```

Out[9]:

	UID	SITE_ID	STATE	DATE_COL	BATCH_ID	PARAMETER	PARAMETER_NAME	RESULT	UNITS	MDL	MRL	PQL	DATE_ANALYZED	HOLD
0	59	NCCA10-1111	CA	7/1/2010	100714.1	NTL	Total Nitrogen	0.407500	mg N/L	0.0150	0.0300	NaN	7/14/2010	
1	59	NCCA10-1111	CA	7/1/2010	100708.1	NO3NO2	Nitrate/Nitrite	0.014000	mg N/L	0.0020	0.0040	NaN	7/8/2010	
2	59	NCCA10-1111	CA	7/1/2010	100708.1	SRP	Dissolved Inorganic Phosphate	0.028000	mg P/L	0.0027	0.0054	NaN	7/8/2010	
3	59	NCCA10-1111	CA	7/1/2010	IM_CALCULATED	DIN	Dissolved Inorganic Nitrogen	0.014000	mg N/L	NaN	NaN	NaN	NaN	
4	59	NCCA10-1111	CA	7/1/2010	100714.1	PTL	Total Phosphorus	0.061254	mg P/L	0.0012	0.0024	NaN	7/14/2010	

In [10]:
```
ncca_sites.head()
```

Out[10]:

	UID	SITE_ID	STATE	VISIT_NO	DATE_COL	WTBDY_NM	SITESAMP	INDEX_VISIT	EPA_REG	NCCR_REG	NCA_REGION	COUNTRY	PROVINCE	STAT
0	59	NCCA10-1111	CA	1.0	1-Jul-10	Mission Bay	Y	Y	9	West	West Coast	USA	Californian Province	
1	60	NCCA10-1119	CA	1.0	1-Jul-10	San Diego Bay	Y	Y	9	West	West Coast	USA	Californian Province	
2	61	NCCA10-1123	CA	1.0	1-Jul-10	Mission Bay	Y	Y	9	West	West Coast	USA	Californian Province	
3	62	NCCA10-1127	CA	1.0	1-Jul-10	San Diego Bay	Y	Y	9	West	West Coast	USA	Californian Province	
4	63	NCCA10-1133	NC	1.0	9-Jun-10	White Oak River	Y	Y	4	Southeast	East Coast	USA	Carolinian Province	

In [11]:
```
raw_vars = ['UID', 'STATE', 'DATE_COL',
 'PARAMETER_NAME', 'RESULT']
sites_vars = ['WTBDY_NM', 'NCCR_REG',
 'STATION_DEPTH', 'ALAT_DD',
 'ALON_DD']
vars_to_keep = raw_vars + sites_vars
```

In [12]:
```
df1 = pd.merge(ncca_sites, ncca_raw,
 how='right',
on = ['UID', 'SITE_ID', 'STATE',
 'DATE_COL']
 )
df1
```

Out[12]:

	UID	SITE_ID	STATE	VISIT_NO	DATE_COL	WTBDY_NM	SITESAMP	INDEX_VISIT	EPA_REG	NCCR_REG	NCA_REGION	COUNTRY	PROVINCE
0	59	NCCA10-1111	CA	NaN	7/1/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	59	NCCA10-1111	CA	NaN	7/1/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	59	NCCA10-1111	CA	NaN	7/1/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	59	NCCA10-1111	CA	NaN	7/1/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	59	NCCA10-1111	CA	NaN	7/1/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7871	16731	NCCA10-1108	CA	NaN	6/29/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7872	16731	NCCA10-1108	CA	NaN	6/29/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7873	16731	NCCA10-1108	CA	NaN	6/29/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7874	16731	NCCA10-1108	CA	NaN	6/29/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7875	16731	NCCA10-1108	CA	NaN	6/29/2010	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

7876 rows × 45 columns

In [13]:

```
#we can get rid of many NAN values here
df1a = pd.merge(ncca_raw, ncca_sites,
 how='right',
on = 'UID'
 )
df1a
```

Out[13]:

	UID	SITE_ID_x	STATE_x	DATE_COL_x	BATCH_ID	PARAMETER	PARAMETER_NAME	RESULT	UNITS	MDL	MRL	PQL	DATE_ANAL
0	59	NCCA10-1111	CA	7/1/2010	100714.1	NTL	Total Nitrogen	0.407500	mg N/L	0.0150	0.0300	NaN	7/14,
1	59	NCCA10-1111	CA	7/1/2010	100708.1	NO3NO2	Nitrate/Nitrite	0.014000	mg N/L	0.0020	0.0040	NaN	7/8,
2	59	NCCA10-1111	CA	7/1/2010	100708.1	SRP	Dissolved Inorganic Phosphate	0.028000	mg P/L	0.0027	0.0054	NaN	7/8,
3	59	NCCA10-1111	CA	7/1/2010	IM_CALCULATED	DIN	Dissolved Inorganic Nitrogen	0.014000	mg N/L	NaN	NaN	NaN	
4	59	NCCA10-1111	CA	7/1/2010	100714.1	PTL	Total Phosphorus	0.061254	mg P/L	0.0012	0.0024	NaN	7/14,
...	...	...	...	...	...	...	...	...	...	...	...	...	
7883	2010099	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7884	2010110	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7885	2010113	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7886	2010135	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7887	2010141	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

7888 rows × 48 columns

In [14]:

```
vars_to_keep_1a = ['UID', 'STATE_x', 'DATE_COL_x',
 'PARAMETER_NAME', 'RESULT','WTBDY_NM',
 'NCCR_REG', 'STATION_DEPTH', 'ALAT_DD',
 'ALON_DD']
```

In [17]:

```
#now we can use only the variables we want for 10 columns
df2 = df1a.loc[:,vars_to_keep_1a]
df2
```

Out[17]:

	UID	STATE_x	DATE_COL_x	PARAMETER_NAME	RESULT	WTBDY_NM	NCCR_REG	STATION_DEPTH	ALAT_DD	ALON_DD
0	59	CA	7/1/2010	Total Nitrogen	0.407500	Mission Bay	West	2.5	32.773610	-117.214710
1	59	CA	7/1/2010	Nitrate/Nitrite	0.014000	Mission Bay	West	2.5	32.773610	-117.214710
2	59	CA	7/1/2010	Dissolved Inorganic Phosphate	0.028000	Mission Bay	West	2.5	32.773610	-117.214710
3	59	CA	7/1/2010	Dissolved Inorganic Nitrogen	0.014000	Mission Bay	West	2.5	32.773610	-117.214710
4	59	CA	7/1/2010	Total Phosphorus	0.061254	Mission Bay	West	2.5	32.773610	-117.214710
...	...	...	...	...	...	...	...	...	...	...
7883	2010099	NaN	NaN	NaN	NaN	Lake Michigan	Great Lakes	NaN	45.845952	-86.751205
7884	2010110	NaN	NaN	NaN	NaN	Lake Michigan	Great Lakes	NaN	44.754051	-85.543548
7885	2010113	NaN	NaN	NaN	NaN	Fourleague Bay	Gulf	NaN	29.341875	-91.179798
7886	2010135	NaN	NaN	NaN	NaN	Hackberry Lake	Gulf	NaN	29.208959	-90.859280
7887	2010141	NaN	NaN	NaN	NaN	Lake Michigan	Great Lakes	NaN	44.777491	-85.616256

7888 rows × 10 columns

In [18]:

```
#now we can remove rows with missing values
df3 = df2[df2.STATE_x.notna()]
```

```
df3
df3.isna().sum()
```

Out[18]:

```
UID              0
STATE_x          0
DATE_COL_x       0
PARAMETER_NAME   0
RESULT           0
WTBDY_NM         0
NCCR_REG         0
STATION_DEPTH    0
ALAT_DD          0
ALON_DD          0
dtype: int64
```

In [19]:

```
#now remove more columns we do not need and use the name column
#as the obs column
df4 = df3.pivot(
 index = df3.drop(['PARAMETER_NAME', 'RESULT'], axis = 1).columns,
 columns = 'PARAMETER_NAME',
 values = 'RESULT'
).reset_index(
).rename_axis(
 columns = {'PARAMETER_NAME':''}
)
df4
```

Out[19]:

	UID	STATE_x	DATE_COL_x	WTBDY_NM	NCCR_REG	STATION_DEPTH	ALAT_DD	ALON_DD	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Di
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	0.000	3.34	0.014	0.028	
1	60	CA	7/1/2010	San Diego Bay	West	3.5	32.71424	-117.23527	0.010	2.45	0.020	0.026	
2	61	CA	7/1/2010	Mission Bay	West	2.2	32.78372	-117.22132	0.000	3.82	0.009	0.030	
3	62	CA	7/1/2010	San Diego Bay	West	9.5	32.72245	-117.20443	0.000	6.13	0.010	0.028	
4	63	NC	6/9/2010	White Oak River	Southeast	1.0	34.75098	-77.12117	0.002	9.79	0.030	0.043	
...	...	...	...	...	...	...	...	...	...	...	...	...	
1087	16727	MI	6/18/2010	Lake Michigan	Great Lakes	0.6	44.98607	-85.64046	0.003	0.75	0.260	0.007	
1088	16728	MI	6/25/2010	Lake Michigan	Great Lakes	2.3	44.94789	-85.94790	0.005	2.27	0.235	0.013	
1089	16729	MI	6/16/2010	Lake Michigan	Great Lakes	31.2	44.83721	-85.52862	0.010	1.11	0.250	0.004	
1090	16730	CA	6/29/2010	San Diego Bay	West	4.1	32.66443	-117.13879	0.017	2.11	0.028	0.034	
1091	16731	CA	6/29/2010	San Diego Bay	West	4.8	32.66243	-117.12712	0.016	2.19	0.028	0.033	

1092 rows × 23 columns

In [20]:

```
#now lets find the columns that have over 95% of not missing
#values
(df4.notna().sum()/len(df4)) > 0.95
```

Out[20]:

```
UID                             True
STATE_x                         True
DATE_COL_x                      True
WTBDY_NM                        True
NCCR_REG                        True
STATION_DEPTH                   True
ALAT_DD                         True
ALON_DD                         True
Ammonia                         True
Chlorophyll A                   True
Dissolved Inorganic Nitrogen    True
Dissolved Inorganic Phosphate   True
Dissolved Silica               False
Nitrate                        False
Nitrate/Nitrite                 True
Nitrite                        False
Nitrogen Particulate           False
Phosphorus Particulate         False
Total Dissolved Nitrogen       False
Total Dissolved Phosphorus     False
Total Kjeldahl Nitrogen        False
Total Nitrogen                  True
Total Phosphorus                True
dtype: bool
```

In [21]:

```
#now we can choose the columns that have over 90% of values
#that are not na
```

```
df5 = df4[df4.columns[(df4.notna().sum()/len(df4)) > 0.95]]
data = df5.rename(
 columns = {
 'STATE_x':'State',
 'DATE_COL_x':'Date collected',
 'WTBDY_NM':'Waterbody name',
 'NCCR_REG':'Region',
 'STATION_DEPTH':'Water depth (in meters)',
 'ALAT_DD':'Latitude',
 'ALON_DD':'Longitude'
 }
)
data
```

Out[21]:

	UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite	Nit
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	0.000	3.34	0.014	0.028	0.014	0.4
1	60	CA	7/1/2010	San Diego Bay	West	3.5	32.71424	-117.23527	0.010	2.45	0.020	0.026	0.010	0.2
2	61	CA	7/1/2010	Mission Bay	West	2.2	32.78372	-117.22132	0.000	3.82	0.009	0.030	0.009	0.3
3	62	CA	7/1/2010	San Diego Bay	West	9.5	32.72245	-117.20443	0.000	6.13	0.010	0.028	0.010	0.2
4	63	NC	6/9/2010	White Oak River	Southeast	1.0	34.75098	-77.12117	0.002	9.79	0.030	0.043	0.028	0.6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1087	16727	MI	6/18/2010	Lake Michigan	Great Lakes	0.6	44.98607	-85.64046	0.003	0.75	0.260	0.007	0.257	0.3
1088	16728	MI	6/25/2010	Lake Michigan	Great Lakes	2.3	44.94789	-85.94790	0.005	2.27	0.235	0.013	0.230	0.4
1089	16729	MI	6/16/2010	Lake Michigan	Great Lakes	31.2	44.83721	-85.52862	0.010	1.11	0.250	0.004	0.240	0.3
1090	16730	CA	6/29/2010	San Diego Bay	West	4.1	32.66443	-117.13879	0.017	2.11	0.028	0.034	0.011	0.2
1091	16731	CA	6/29/2010	San Diego Bay	West	4.8	32.66243	-117.12712	0.016	2.19	0.028	0.033	0.012	0.2

1092 rows × 15 columns

In [42]:
```
data_csv = data.to_csv('out', index=False)
```

2a.What is the apparent relationship between nutrient availability and productivity?

In [23]:
```
alt.data_transformers.disable_max_rows()
```

Out[23]:
```
DataTransformerRegistry.enable('default')
```

In [24]:
```
data
```

Out[24]:

	UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite	Nit
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	0.000	3.34	0.014	0.028	0.014	0.4
1	60	CA	7/1/2010	San Diego Bay	West	3.5	32.71424	-117.23527	0.010	2.45	0.020	0.026	0.010	0.2
2	61	CA	7/1/2010	Mission Bay	West	2.2	32.78372	-117.22132	0.000	3.82	0.009	0.030	0.009	0.3
3	62	CA	7/1/2010	San Diego Bay	West	9.5	32.72245	-117.20443	0.000	6.13	0.010	0.028	0.010	0.2
4	63	NC	6/9/2010	White Oak River	Southeast	1.0	34.75098	-77.12117	0.002	9.79	0.030	0.043	0.028	0.6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1087	16727	MI	6/18/2010	Lake Michigan	Great Lakes	0.6	44.98607	-85.64046	0.003	0.75	0.260	0.007	0.257	0.3
1088	16728	MI	6/25/2010	Lake Michigan	Great Lakes	2.3	44.94789	-85.94790	0.005	2.27	0.235	0.013	0.230	0.4
1089	16729	MI	6/16/2010	Lake Michigan	Great Lakes	31.2	44.83721	-85.52862	0.010	1.11	0.250	0.004	0.240	0.3
1090	16730	CA	6/29/2010	San Diego Bay	West	4.1	32.66443	-117.13879	0.017	2.11	0.028	0.034	0.011	0.2
1091	16731	CA	6/29/2010	San Diego Bay	West	4.8	32.66243	-117.12712	0.016	2.19	0.028	0.033	0.012	0.2

1092 rows × 15 columns

In [25]:
```
x_mx = data.iloc[:, 8:15]

# long form dataframe for plotting panel
scatter_df = x_mx.melt(
    var_name = 'row',
    value_name = 'row_index'
).join(
    pd.concat([x_mx, x_mx, x_mx, x_mx, x_mx, x_mx,x_mx, x_mx], axis = 0).reset_index(),
).drop(
    columns = 'index'
).melt(
    id_vars = ['row', 'row_index'],
    var_name = 'col',
    value_name = 'col_index'
)
scatter_df
```

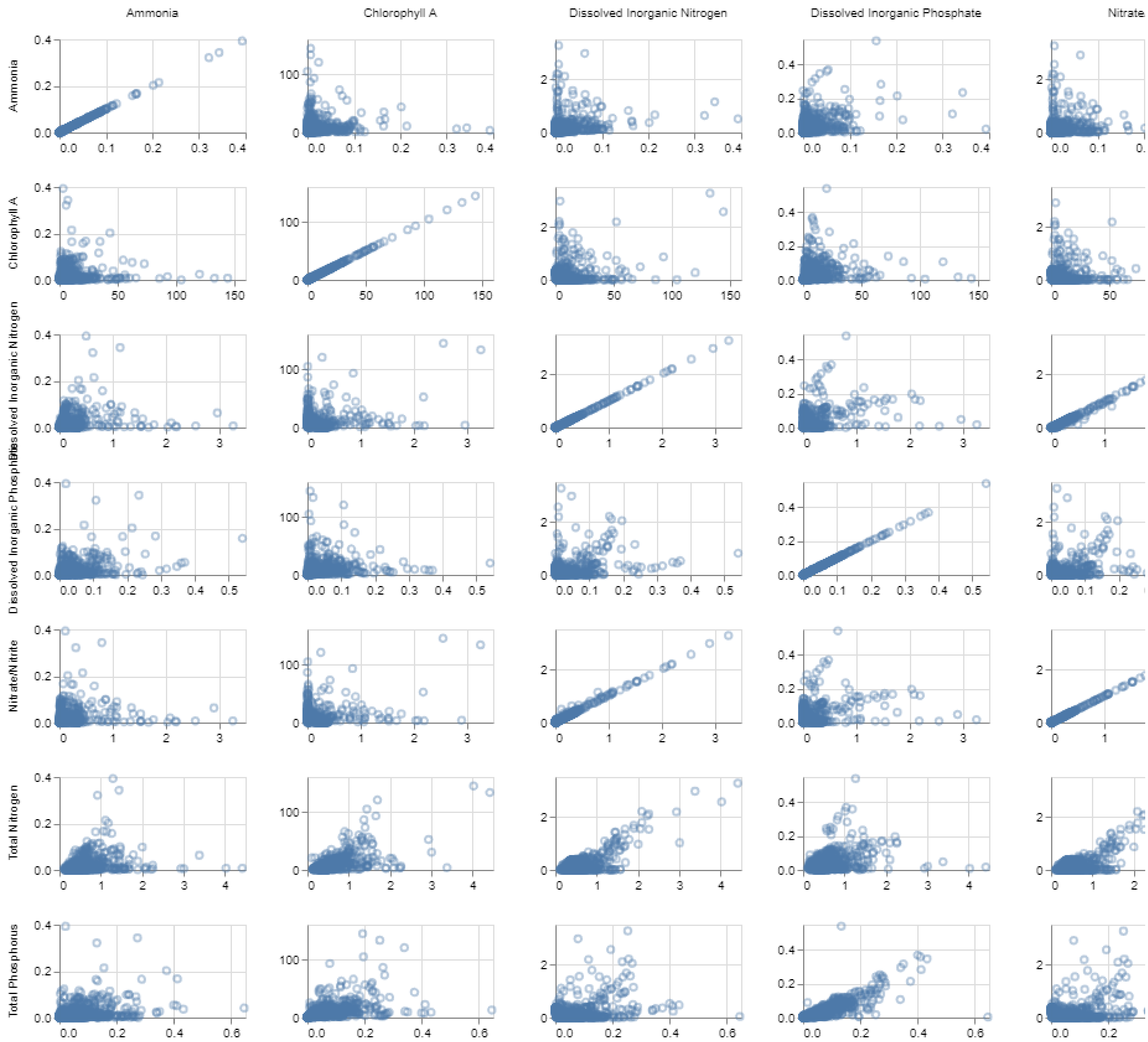
Out[25]:

	row	row_index	col	col_index
0	Ammonia	0.000000	Ammonia	0.000000
1	Ammonia	0.010000	Ammonia	0.010000
2	Ammonia	0.000000	Ammonia	0.000000
3	Ammonia	0.000000	Ammonia	0.000000
4	Ammonia	0.002000	Ammonia	0.002000
...	...	...	...	...
53503	Total Phosphorus	0.000000	Total Phosphorus	0.000000
53504	Total Phosphorus	0.006249	Total Phosphorus	0.006249
53505	Total Phosphorus	0.000000	Total Phosphorus	0.000000
53506	Total Phosphorus	0.044127	Total Phosphorus	0.044127
53507	Total Phosphorus	0.041821	Total Phosphorus	0.041821

53508 rows × 4 columns

In [26]:
```
scatter_panel = alt.Chart(scatter_df).mark_point(opacity = 0.4).encode(
 x = alt.X('row_index', scale = alt.Scale(zero = False), title = ''),
 y = alt.Y('col_index', scale = alt.Scale(zero = False), title = '')
).properties(
 width = 150,
 height = 75
).facet(
 column = alt.Column('col', title = ''),
 row = alt.Row('row', title = '')
).resolve_scale(x = 'independent', y = 'independent')
scatter_panel
```

Out[26]:



In [27]:
```
corr_mx=x_mx.corr()
corr_mx
```

Out[27]:

	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite	Total Nitrogen	Total Phosphorus
Ammonia	1.000000	0.076214	0.223906	0.373070	0.128686	0.288228	0.321642
Chlorophyll A	0.076214	1.000000	0.188035	0.196624	0.185112	0.641165	0.512931
Dissolved Inorganic Nitrogen	0.223906	0.188035	1.000000	0.258240	0.995142	0.716507	0.234987
Dissolved Inorganic Phosphate	0.373070	0.196624	0.258240	1.000000	0.224840	0.378746	0.807155
Nitrate/Nitrite	0.128686	0.185112	0.995142	0.224840	1.000000	0.700950	0.206868
Total Nitrogen	0.288228	0.641165	0.716507	0.378746	0.700950	1.000000	0.566093
Total Phosphorus	0.321642	0.512931	0.234987	0.807155	0.206868	0.566093	1.000000

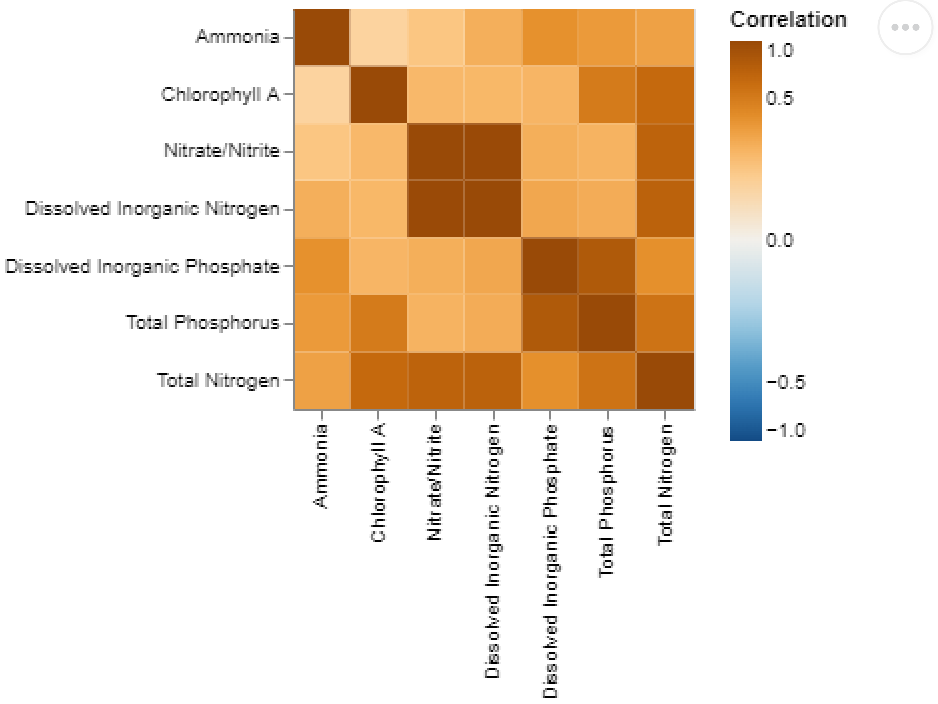
In [28]:
```
# melt to long form
corr_mx_long = corr_mx.reset_index().rename(
    columns = {'': 'row'}
).melt(
    id_vars = 'row',
    var_name = 'col',
    value_name = 'Correlation'
)

# visualize
heatmap = alt.Chart(corr_mx_long).mark_rect().encode(
    x = alt.X('col', title = '', sort = {'field': 'Correlation', 'order': 'ascending'}),
    y = alt.Y('row', title = '', sort = {'field': 'Correlation', 'order': 'ascending'}),
    color = alt.Color('Correlation',
                      scale = alt.Scale(scheme = 'blueorange',
                                        domain = (-1, 1),
```

```
                    type = 'sqrt'),
          legend = alt.Legend(tickCount = 5))
).properties(width = 200, height = 200)

heatmap
```

Out[28]:



2b.Are there any notable differences in available nutrients among U.S. coastal regions?

In [33]:

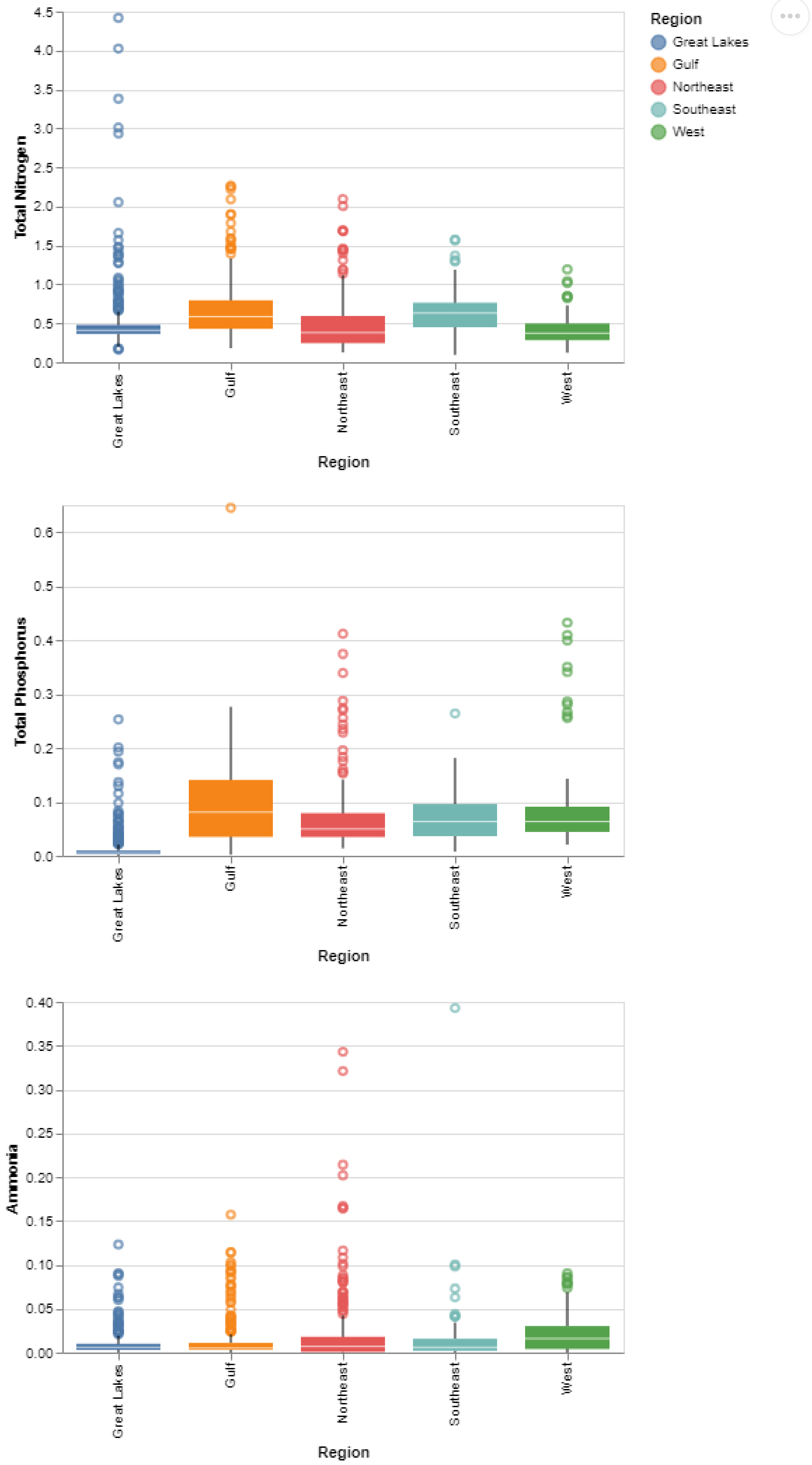
```
tot_nitrogen=alt.Chart(data).mark_boxplot(size=60).encode(
    x='Region',
    y='Total Nitrogen',
    color='Region'
).properties(width=400, height=250)

tot_phosphorus=alt.Chart(data).mark_boxplot(size=60).encode(
    x='Region',
    y='Total Phosphorus',
    color='Region'
).properties(width=400, height=250)

tot_ammonia=alt.Chart(data).mark_boxplot(size=60).encode(
    x='Region',
    y='Ammonia',
    color='Region'
).properties(width=400, height=250)

totals=tot_nitrogen & tot_phosphorus & tot_ammonia
totals
```

Out[33]:



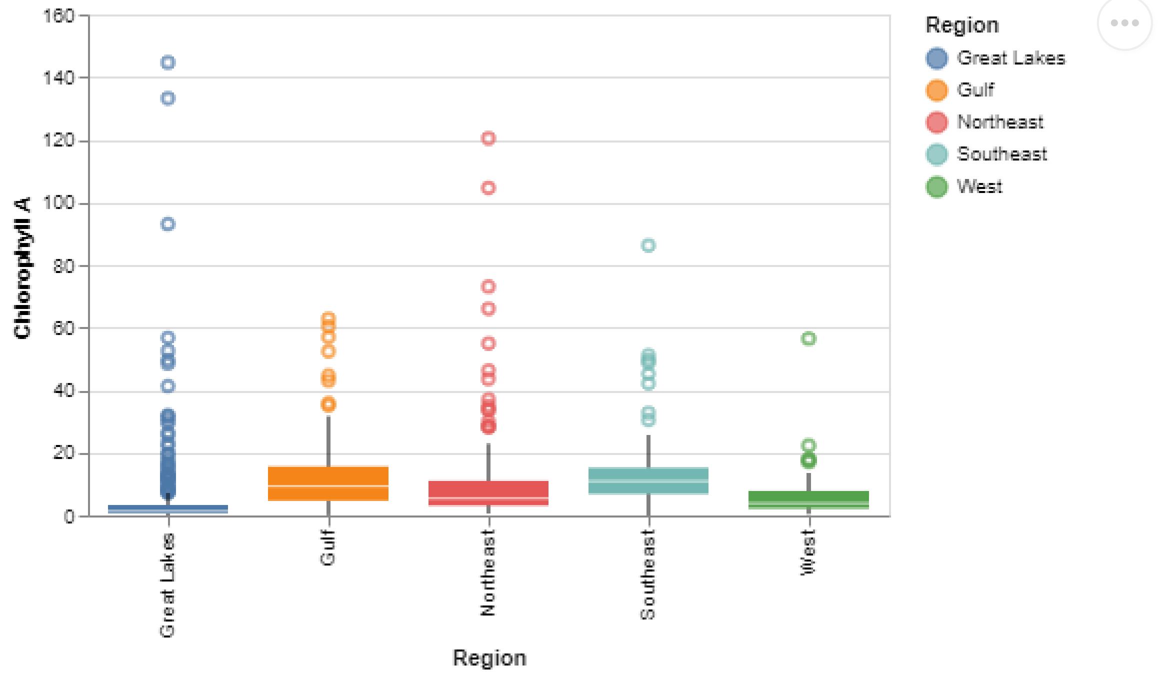
2c. Based on the 2010 data, does productivity seem to vary geographically in some way?

In [30]:

```
tot_chlorophyll=alt.Chart(data).mark_boxplot(size=60).encode(
    x='Region',
    y='Chlorophyll A',
    color='Region'
).properties(width=400, height=250)

tot_chlorophyll
```

Out[30]:



2d. How does primary productivity in California coastal waters change seasonally in 2010, if at all?

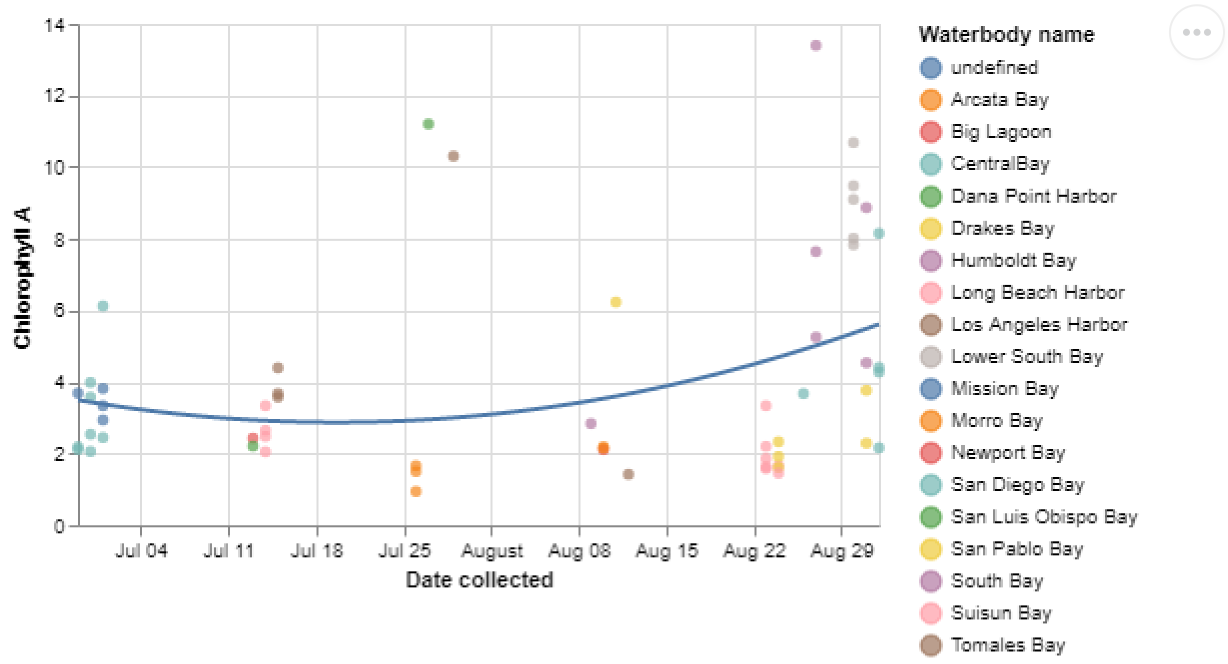
Does your result make intuitive sense?

In [35]:
```
scatter = alt.Chart(data[data['State']== 'CA']).mark_circle(color="black").encode(
    x='Date collected:T',
    y='Chlorophyll A',
    color = 'Waterbody name'
).properties(width=400, height=250)

smooth = scatter.transform_regression(
    'Date collected', 'Chlorophyll A', method= 'quad'
    ).mark_line(color = 'blue')

scatter + smooth
```

Out[35]:



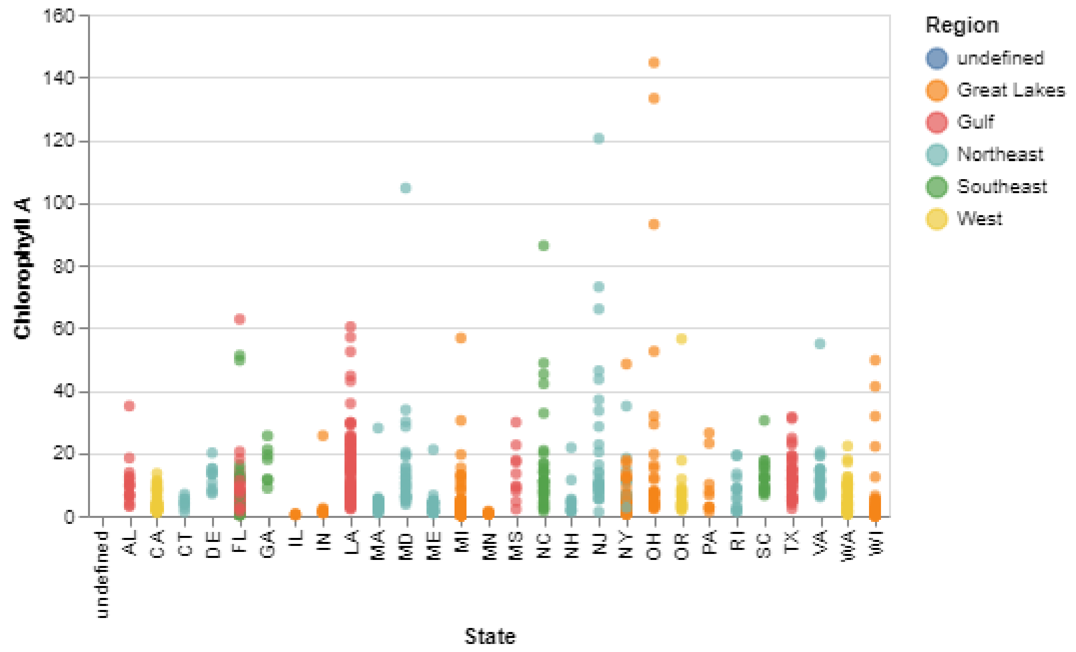
2e. Pose and answer one additional question. Which state is generally most and least productive? Does this intuitively make sense?

In [36]:
```
scatter = alt.Chart(data).mark_circle(color="black").encode(
    x='State',
    y='Chlorophyll A',
    color = 'Region'
).properties(width=400, height=250)

smooth = scatter.transform_regression(
    'Date collected', 'Chlorophyll A', method= 'quad'
    ).mark_line(color = 'blue')

scatter + smooth
```

Out[36]:


Chlorophyll A
0
20
40
60
80
100
120
140
160
State
undefined
AL
CA
CT
DE
FL
GA
IL
IN
LA
MA
MD
ME
MI
MN
MS
NC
NH
NJ
NY
OH
OR
PA
RI
SC
TX
VA
WA
WI
Region
undefined
Great Lakes
Gulf
Northeast
Southeast
West