

```
In [1]: import pandas as pd
from IPython.display import HTML
import base64, io, IPython
from PIL import Image as PILImage
from IPython.display import Image
from IPython import display
```

Name:

TJ Sipin

Contributors:

Preeti Kulkarni and Gian Taparan

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.

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 [2]: # show a few rows of clean data
pd.read_csv('out').head()
```

Out[2]:

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

The dataset above contains amounts of various nutrients (like ammonia, nitrogen, and phosphate) and levels of productivity (via chlorophyll A levels) in several bodies of water across the US during the summer months of 2010. Each observation includes date collected as well as longitude/latitude.

Part 2: exploratory analysis

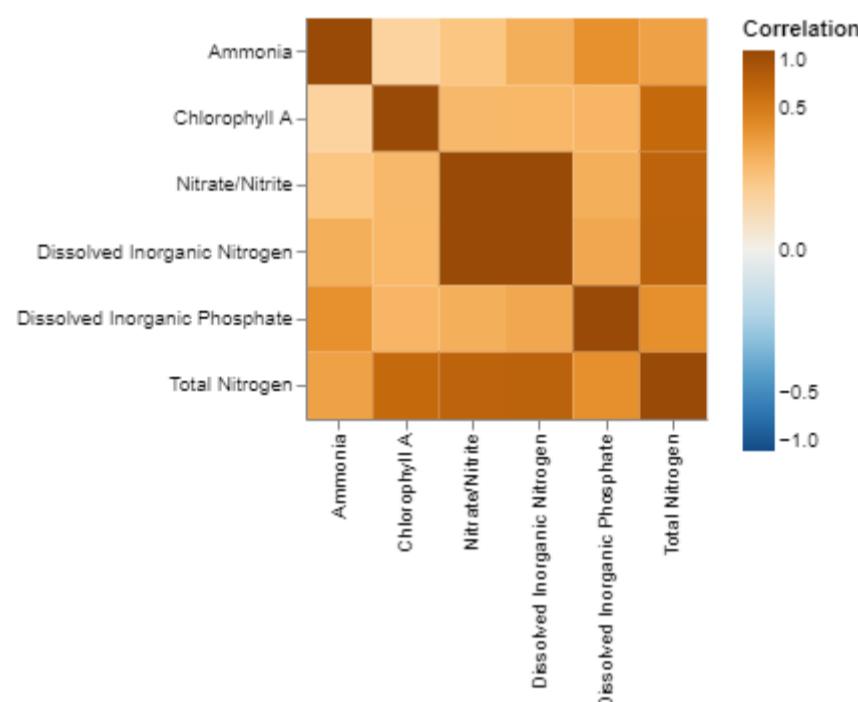
Answer each question below and provide a visualization supporting your answer. A description and interpretation of the visualization should be offered.

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.

Total nitrogen and total phosphorous are moderately correlated with each other. Ammonia has low-moderate correlation with nutrient availability (i.e. nitrogen and phosphorus levels), while chlorophyll A has moderate-high correlation with nutrient availability.



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

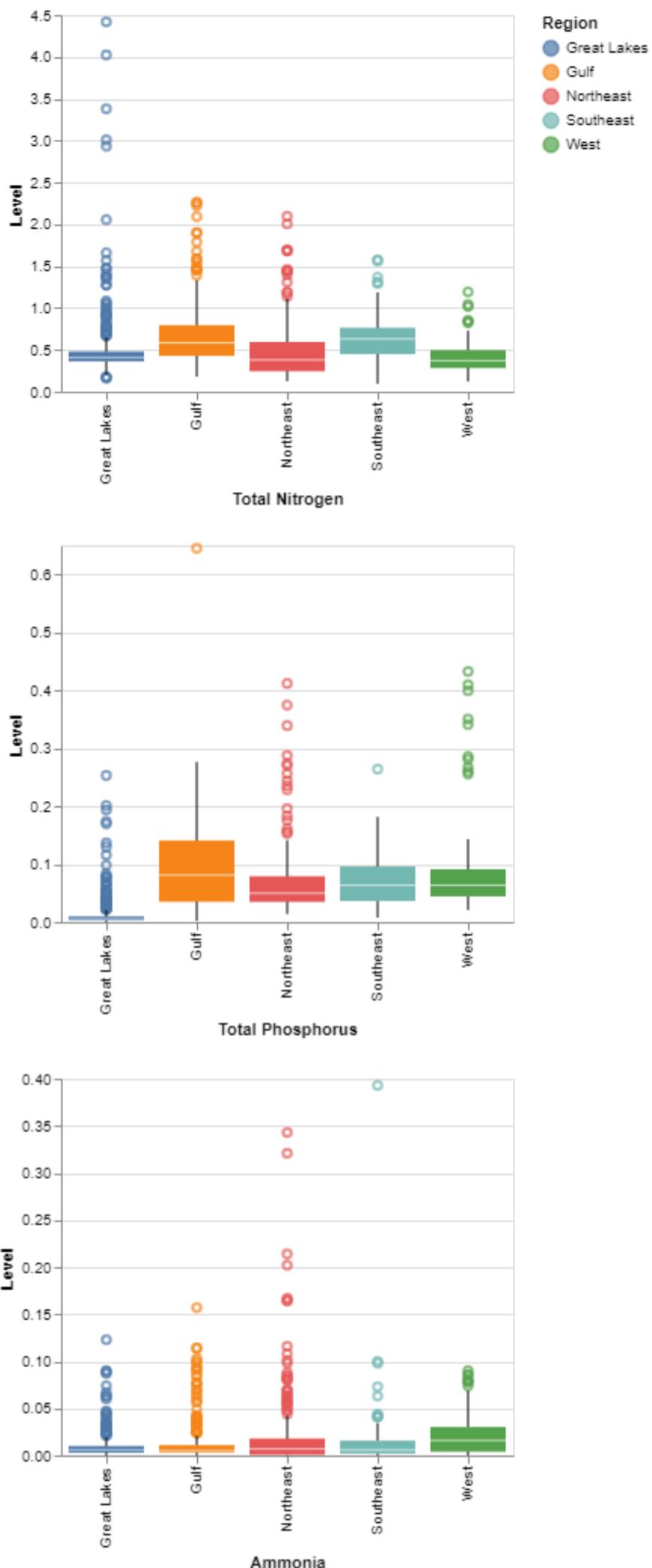
Generally, the west has slightly higher phosphorus and ammonia levels, while the east has higher nitrogen levels. Upon further investigation, the west may have higher phosphorus levels due to agricultural malpractice, such as high runoff rates.

The west also has lower variability in ammonia levels, whereas the east has many outliers, specifically the northeast.

One can see that for nitrogen and phosphorus, but nitrogen especially, the Great Lakes tend to have many more outliers than the other regions. This may be due to a lower flow of fresh water to carry nutrients compared to the currents of the oceans, which allows higher opportunity for buildup of nutrients.

For the Gulf, the distribution is incredibly gaussian, but there is one outlier that stands apart from the rest. This would be interesting to investigate.

It must be noted that all but one outlier are in the upper-tail.

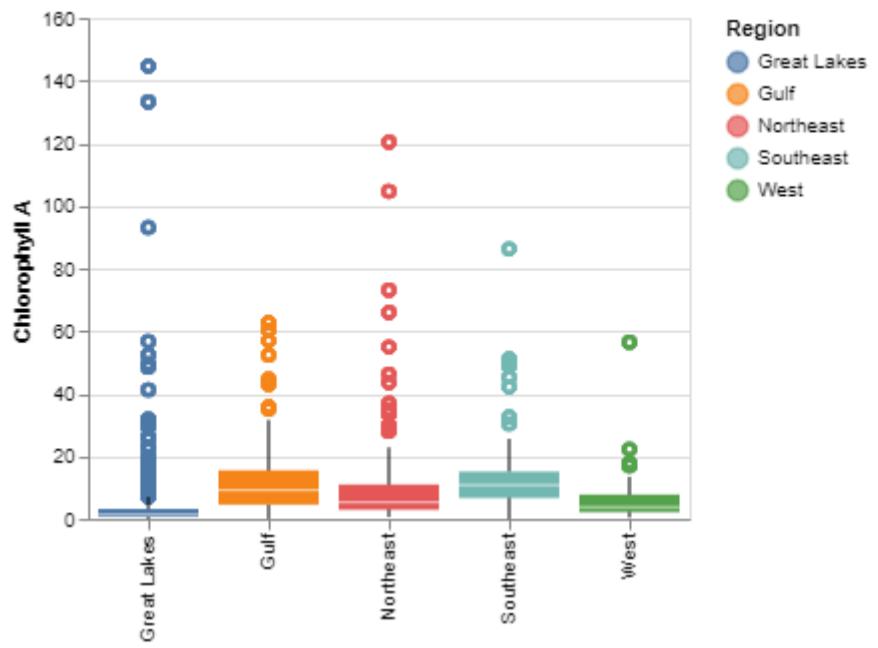


Based on the 2010 data, does productivity seem to vary geographically in some way?

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

Just like for the nutrients, the Great Lakes region features the most outliers with the lowest median chlorophyll A levels. However, each region features many outliers. The west does not seem to have high chlorophyll concentrations compared to the east and the gulf.

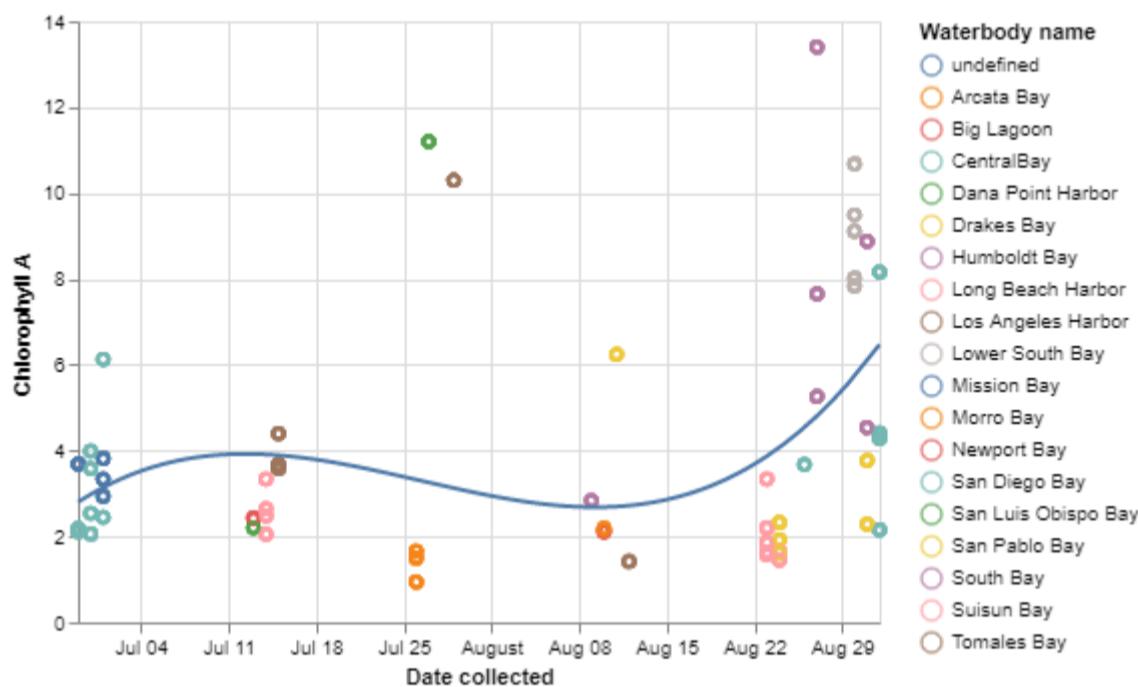
This is most likely due to the corresponding nutrient levels.

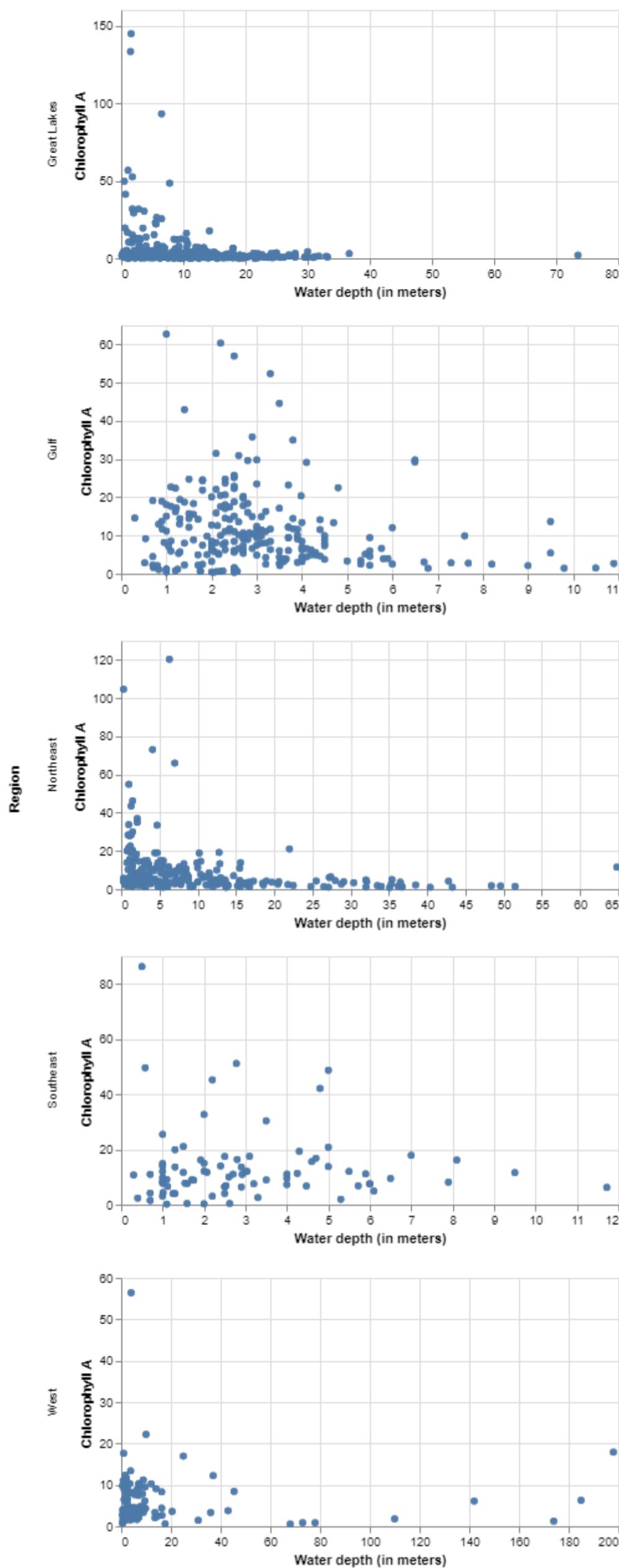


How does primary productivity in California coastal waters change seasonally in 2010, if at all?

Does your result make intuitive sense?

There is a slight upward trend, though the variance is high. This might make sense as harvesting in California is typically done in the autumn, so there is a higher risk for fertilizer runoff as mentioned before. This runoff of fertilizer, which is high in phosphorus, may be responsible for the high productivity levels.





Codes

```
In [3]: import pandas as pd
import numpy as np
import altair as alt

ncca_raw = pd.read_csv('assessed_ncca2010_waterchem.csv')
Loading [MathJax]/extensions/Safe.js .read_csv('assessed_ncca2010_siteinfo.csv')
```

```
In [4]: ncca_raw[ncca_raw.PARAMETER_NAME == 'Ammonia'].UNITS.describe()
```

```
Out[4]: count      1091
unique       1
top        mg N/L
freq      1091
Name: UNITS, dtype: object
```

1

```
In [5]: ncca_raw
```

```
Out[5]:   UID SITE_ID STATE DATE_COL      BATCH_ID PARAMETER PARAMETER_NAME    RESULT UNITS      MDL      MRL      PQL DATE_ANALYZED
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
...   ...
7871 16731 NCCA10-1108  CA  6/29/2010  100707.1      NTL Total Nitrogen  0.228750 mg N/L  0.0150  0.0300  NaN  7/7/2010
7872 16731 NCCA10-1108  CA  6/29/2010  100707.1      PTL Total Phosphorus  0.041821 mg P/L  0.0012  0.0024  NaN  7/7/2010
7873 16731 NCCA10-1108  CA  6/29/2010  100702.1      SRP Dissolved Inorganic Phosphate  0.033000 mg P/L  0.0010  0.0054  NaN  7/2/2010
7874 16731 NCCA10-1108  CA  6/29/2010  100701.1      NH3 Ammonia  0.016000 mg N/L  0.0030  0.0060  NaN  7/1/2010
7875 16731 NCCA10-1108  CA  6/29/2010  100702.1  NO3NO2 Nitrate/Nitrite  0.012000 mg N/L  0.0020  0.0040  NaN  7/2/2010
```

7876 rows × 18 columns

```
In [6]: 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 [7]: data_mod1 = pd.merge(ncca_raw, ncca_sites,
                           how='right',
                           on = ['UID', 'SITE_ID', 'STATE',
                                 'DATE_COL'])
data_mod1
```

Out[7]:

	UID	SITE_ID	STATE	DATE_COL	BATCH_ID	PARAMETER	PARAMETER_NAME	RESULT	UNITS	MDL	...	NPSPIKE	PANEL	STATUS!
0	59	NCCA10-1111	CA	1-Jul-10	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Base	Target_Samp
1	60	NCCA10-1119	CA	1-Jul-10	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Base	Target_Samp
2	61	NCCA10-1123	CA	1-Jul-10	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Base	Target_Samp
3	62	NCCA10-1127	CA	1-Jul-10	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Base	Target_Samp
4	63	NCCA10-1133	NC	9-Jun-10	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Revisit	Target_Samp
...
1099	2010099	NCCAGL10-GLBA10-174	MI	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Over	Target_Presun
1100	2010110	NCCAGL10-GLBA10-183	MI	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Over	Target_Presun
1101	2010113	NCCA10-2326	LA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Over	Target_Presun
1102	2010135	NCCA10-2328	LA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Over	Target_Presun
1103	2010141	NCCAGL10-GLBA10-179	MI	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Over	Target_Presun

1104 rows × 45 columns

```
In [8]: data_mod1a = pd.merge(ncca_raw, ncca_sites,
                           how='right',
                           on = 'UID'
                           )
data_mod1a
```

Out[8]:	UID	SITE_ID_x	STATE_x	DATE_COL_x	BATCH_ID	PARAMETER	PARAMETER_NAME	RESULT	UNITS	MDL	...	NPSPARK	PANEL
0	59	NCCA10-1111	CA	7/1/2010	100714.1	NTL	Total Nitrogen	0.407500	mg N/L	0.0150	...	NaN	Base T
1	59	NCCA10-1111	CA	7/1/2010	100708.1	NO3NO2	Nitrate/Nitrite	0.014000	mg N/L	0.0020	...	NaN	Base T
2	59	NCCA10-1111	CA	7/1/2010	100708.1	SRP	Dissolved Inorganic Phosphate	0.028000	mg P/L	0.0027	...	NaN	Base T
3	59	NCCA10-1111	CA	7/1/2010	IM_CALCULATED	DIN	Dissolved Inorganic Nitrogen	0.014000	mg N/L	NaN	...	NaN	Base T
4	59	NCCA10-1111	CA	7/1/2010	100714.1	PTL	Total Phosphorus	0.061254	mg P/L	0.0012	...	NaN	Base T
...

7883 2010099 NaN ... NaN Over Tai

7884 2010110 NaN ... NaN Over Tai

7885 2010113 NaN ... NaN Over Tai

7886 2010135 NaN ... NaN Over Tai

7887 2010141 NaN ... NaN Over Tai

7888 rows × 48 columns

```
In [9]: vars_to_keep_1a = ['UID', 'STATE_x', 'DATE_COL_x',
'PARAMETER_NAME', 'RESULT', 'WTBDY_NM',
'NCCR_REG', 'STATION_DEPTH', 'ALAT_DD',
'ALON_DD', 'PROVINCE']
```

```
In [10]: data_mod2 = data_mod1a.loc[:,vars_to_keep_1a]
```

```
In [11]: data_mod2
```

Out[11]:

	UID	STATE_x	DATE_COL_x	PARAMETER_NAME	RESULT	WTBDY_NM	NCCR_REG	STATION_DEPTH	ALAT_DD	ALON_DD	PROVINCE
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 × 11 columns

In [12]:

```
data_mod3 = data_mod2[data_mod2.STATE_x.notna()]
data_mod3
```

Out[12]:

	UID	STATE_x	DATE_COL_x	PARAMETER_NAME	RESULT	WTBDY_NM	NCCR_REG	STATION_DEPTH	ALAT_DD	ALON_DD	PROVINCE
0	59	CA	7/1/2010	Total Nitrogen	0.407500	Mission Bay	West		2.5	32.77361	-117.21471
1	59	CA	7/1/2010	Nitrate/Nitrite	0.014000	Mission Bay	West		2.5	32.77361	-117.21471
2	59	CA	7/1/2010	Dissolved Inorganic Phosphate	0.028000	Mission Bay	West		2.5	32.77361	-117.21471
3	59	CA	7/1/2010	Dissolved Inorganic Nitrogen	0.014000	Mission Bay	West		2.5	32.77361	-117.21471
4	59	CA	7/1/2010	Total Phosphorus	0.061254	Mission Bay	West		2.5	32.77361	-117.21471
...
7873	16731	CA	6/29/2010	Total Nitrogen	0.228750	San Diego Bay	West		4.8	32.66243	-117.12712
7874	16731	CA	6/29/2010	Total Phosphorus	0.041821	San Diego Bay	West		4.8	32.66243	-117.12712
7875	16731	CA	6/29/2010	Dissolved Inorganic Phosphate	0.033000	San Diego Bay	West		4.8	32.66243	-117.12712
7876	16731	CA	6/29/2010	Ammonia	0.016000	San Diego Bay	West		4.8	32.66243	-117.12712
7877	16731	CA	6/29/2010	Nitrate/Nitrite	0.012000	San Diego Bay	West		4.8	32.66243	-117.12712

7876 rows × 11 columns

In [13]:

```
data_mod4 = data_mod3.pivot(
    index = data_mod3.drop(['PARAMETER_NAME', 'RESULT'], axis = 1).columns,
    columns = 'PARAMETER_NAME',
    values = 'RESULT'
).reset_index(
).rename_axis(
    columns = {'PARAMETER_NAME': ''}
)
data_mod4
```

Out[13]:

	UID	STATE_x	DATE_COL_x	WTBDY_NM	NCCR_REG	STATION_DEPTH	ALAT_DD	ALON_DD	PROVINCE	Ammonia	...	Nitrate	Nitrate/Nitrite
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	Californian Province	0.000	...	NaN	0.014
1	60	CA	7/1/2010	San Diego Bay	West	3.5	32.71424	-117.23527	Californian Province	0.010	...	NaN	0.010
2	61	CA	7/1/2010	Mission Bay	West	2.2	32.78372	-117.22132	Californian Province	0.000	...	NaN	0.009
3	62	CA	7/1/2010	San Diego Bay	West	9.5	32.72245	-117.20443	Californian Province	0.000	...	NaN	0.010
4	63	NC	6/9/2010	White Oak River	Southeast	1.0	34.75098	-77.12117	Carolinian Province	0.002	...	NaN	0.028
...
1087	16727	MI	6/18/2010	Lake Michigan	Great Lakes	0.6	44.98607	-85.64046	Great Lakes Province	0.003	...	NaN	0.257
1088	16728	MI	6/25/2010	Lake Michigan	Great Lakes	2.3	44.94789	-85.94790	Great Lakes Province	0.005	...	NaN	0.230
1089	16729	MI	6/16/2010	Lake Michigan	Great Lakes	31.2	44.83721	-85.52862	Great Lakes Province	0.010	...	NaN	0.240
1090	16730	CA	6/29/2010	San Diego Bay	West	4.1	32.66443	-117.13879	Californian Province	0.017	...	NaN	0.011
1091	16731	CA	6/29/2010	San Diego Bay	West	4.8	32.66243	-117.12712	Californian Province	0.016	...	NaN	0.012

1092 rows × 24 columns

In [14]: `data_mod4[data_mod4['Total Dissolved Nitrogen'].notna()]`

Out[14]:

	UID	STATE_x	DATE_COL_x	WTBDY_NM	NCCR_REG	STATION_DEPTH	ALAT_DD	ALON_DD	PROVINCE	Ammonia	...	Nitrate	Nitrate/Nitrite
221	587	VA	7/13/2010	Warwick River	Northeast	3.0	36.899760	-76.458730	Virginian Province	0.004	...	0.0010	0.001
222	588	VA	7/13/2010	Lower James River	Northeast	1.5	36.954960	-76.273370	Virginian Province	0.025	...	0.0050	0.005
258	639	VA	7/8/2010	Back Bay	Northeast	1.5	36.610230	-75.981980	Carolinian Province	0.009	...	0.0020	0.002
397	819	VA	8/5/2010	Broad/Linkhorn Bay	Northeast	0.9	36.890760	-76.070580	Virginian Province	0.010	...	0.0080	0.008
398	820	VA	7/27/2010	Elizabeth River	Northeast	10.4	36.769190	-76.296590	Virginian Province	0.008	...	0.0240	0.146
399	822	VA	7/27/2010	Lower James River	Northeast	15.6	36.880450	-76.335060	Virginian Province	0.008	...	0.0070	0.013
659	1235	VA	7/21/2010	Pocomoke Sound	Northeast	5.2	37.381930	-76.010570	Virginian Province	0.012	...	0.0020	0.002
660	1236	VA	7/22/2010	Milford Haven	Northeast	10.2	37.625233	-76.206816	Virginian Province	0.014	...	0.0020	0.002
661	1237	VA	7/22/2010	Potomac River	Northeast	10.0	37.678733	-76.262283	Virginian Province	0.011	...	0.0020	0.002
662	1238	VA	7/22/2010	Pocomoke Sound	Northeast	11.5	37.620250	-76.079017	Virginian Province	0.030	...	0.0060	0.006
663	1239	VA	7/20/2010	Pocomoke Sound	Northeast	5.8	37.115433	-76.019100	Virginian Province	0.022	...	0.0030	0.003
664	1240	VA	7/21/2010	Milford Haven	Northeast	2.7	37.443450	-76.239730	Virginian Province	0.012	...	0.0040	0.004
665	1241	VA	7/20/2010	Pocomoke Sound	Northeast	3.4	37.238600	-76.043500	Virginian Province	0.012	...	0.0030	0.003
671	1251	VA	9/16/2010	Hog Island Bay	Northeast	2.5	37.365020	-75.724020	Virginian Province	0.008	...	0.0080	0.009
923	1794	VA	6/30/2010	Chickahominy River	Northeast	3.5	37.293100	-76.893183	Virginian Province	0.007	...	0.0000	0.000
924	1796	VA	8/25/2010	Pocomoke River	Northeast	1.3	37.942600	-75.642600	Virginian Province	0.012	...	0.0030	0.005
925	1797	VA	7/21/2010	Milford Haven	Northeast	9.9	37.397550	-76.165530	Virginian Province	0.017	...	0.0030	0.003
926	1798	VA	9/9/2010	Rappahannock River	Northeast	3.7	37.579900	-76.385900	Virginian Province	0.026	...	0.0051	0.067
927	1799	VA	7/22/2010	Pocomoke Sound	Northeast	3.1	37.595380	-75.937333	Virginian Province	0.013	...	0.0030	0.003
928	1800	VA	7/21/2010	York River	Northeast	7.2	37.294017	-76.327500	Virginian Province	0.068	...	0.0050	0.005
929	1801	VA	8/25/2010	Upper James River	Northeast	6.3	37.262570	-76.981650	Virginian Province	0.025	...	0.1450	0.229
942	1836	VA	8/5/2010	Upper James River	Northeast	1.0	37.363930	-77.268250	Virginian Province	0.085	...	0.2370	0.261
946	1856	VA	7/15/2010	York River	Northeast	2.2	37.160067	-76.302633	Virginian Province	0.006	...	0.0030	0.003

23 rows × 24 columns

In [15]: `(data_mod4.notna().sum()/len(data_mod4)) > 0.9`

```
Out[15]: UID           True
STATE_X        True
DATE_COL_X     True
WTBDY_NM       True
NCCR_REG       True
STATION_DEPTH  True
ALAT_DD        True
ALON_DD        True
PROVINCE       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 [16]: data_mod5 = data_mod4[data_mod4.columns[(data_mod4.notna().sum()/len(data_mod4)) > 0.9]]
data = data_mod5.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',
        'PROVINCE': 'Province'
    }
)
data
```

Out[16]:

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

1092 rows × 16 columns

```
In [17]: data_csv = data.to_csv('out', index=False)
```

2

What is the apparent relationship between nutrient availability and productivity?

```
Out[18]: DataTransformerRegistry.enable('default')
```

```
In [19]: data.head(1)
```

```
Out[19]:
```

UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Province	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite	
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	Californian Province	0.0	3.34	0.014	0.028	0.014

```
In [20]: # df for scatter
```

```
scatter_df = data.melt(  
    id_vars = ['UID', 'State', 'Date collected',  
              'Waterbody name', 'Region', 'Water depth (in meters)',  
              'Latitude', 'Longitude', 'Chlorophyll A', 'Province'],  
    var_name = 'Nutrient',  
    value_name = 'Level'  
)
```

```
In [21]: scatter_df
```

```
Out[21]:
```

	UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Chlorophyll A	Province	Nutrient	Level
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	3.34	Californian Province	Ammonia	0.000000
1	60	CA	7/1/2010	San Diego Bay	West	3.5	32.71424	-117.23527	2.45	Californian Province	Ammonia	0.010000
2	61	CA	7/1/2010	Mission Bay	West	2.2	32.78372	-117.22132	3.82	Californian Province	Ammonia	0.000000
3	62	CA	7/1/2010	San Diego Bay	West	9.5	32.72245	-117.20443	6.13	Californian Province	Ammonia	0.000000
4	63	NC	6/9/2010	White Oak River	Southeast	1.0	34.75098	-77.12117	9.79	Carolinian Province	Ammonia	0.002000
...
6547	16727	MI	6/18/2010	Lake Michigan	Great Lakes	0.6	44.98607	-85.64046	0.75	Great Lakes Province	Total Phosphorus	0.000000
6548	16728	MI	6/25/2010	Lake Michigan	Great Lakes	2.3	44.94789	-85.94790	2.27	Great Lakes Province	Total Phosphorus	0.006249
6549	16729	MI	6/16/2010	Lake Michigan	Great Lakes	31.2	44.83721	-85.52862	1.11	Great Lakes Province	Total Phosphorus	0.000000
6550	16730	CA	6/29/2010	San Diego Bay	West	4.1	32.66443	-117.13879	2.11	Californian Province	Total Phosphorus	0.044127
6551	16731	CA	6/29/2010	San Diego Bay	West	4.8	32.66243	-117.12712	2.19	Californian Province	Total Phosphorus	0.041821

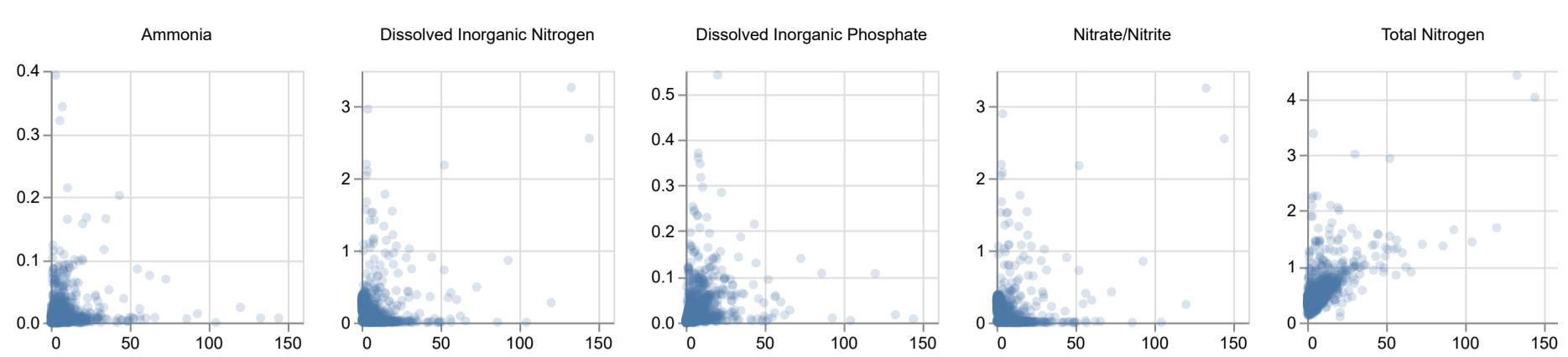
6552 rows × 12 columns

```
In [22]: # panel
```

```
scatter_panel_ammonia = alt.Chart(scatter_df).mark_circle(opacity = 0.2).encode(  
    x = alt.X('Ammonia:Q', scale = alt.Scale(zero = True), title = ''),  
    y = alt.Y('Level', scale = alt.Scale(zero = True), title = '')  
)  
.properties(  
    width = 150,  
    height = 150  
)  
.facet(  
    column = alt.Column('Nutrient', title = 'Ammonia Levels mg N/L')  
)  
.resolve_scale(x = 'independent', y = 'independent')  
  
# panel  
scatter_panel_Chl = alt.Chart(scatter_df).mark_circle(opacity = 0.2).encode(  
    x = alt.X('Chlorophyll A', scale = alt.Scale(zero = True), title = ''),  
    y = alt.Y('Level', scale = alt.Scale(zero = True), title = '')  
)  
.properties(  
    width = 150,  
    height = 150  
)  
.facet(  
    column = alt.Column('Nutrient', title = 'Chlorophyll A Levels ug/L')  
)  
.resolve_scale(x = 'independent', y = 'independent')
```

```
In [23]: scatter_panel_Chl
```

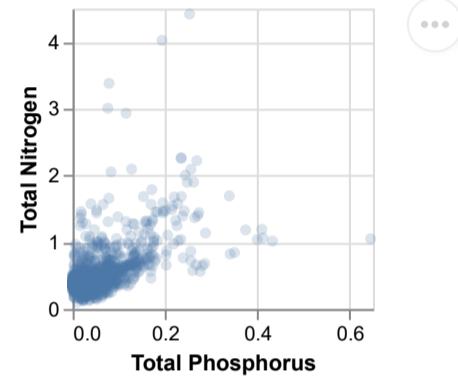
Out[23]:



In [24]:

```
scatter_panel_P_N = alt.Chart(data).mark_circle(opacity = 0.2).encode(
    x = alt.X('Total Phosphorus', scale = alt.Scale(zero = True)),
    y = alt.Y('Total Nitrogen', scale = alt.Scale(zero = True)))
).properties(
    width = 150,
    height = 150
)
```

In [25]:



In [26]:

```
x_mx = data.iloc[:, 8:15].drop(columns = 'Province')

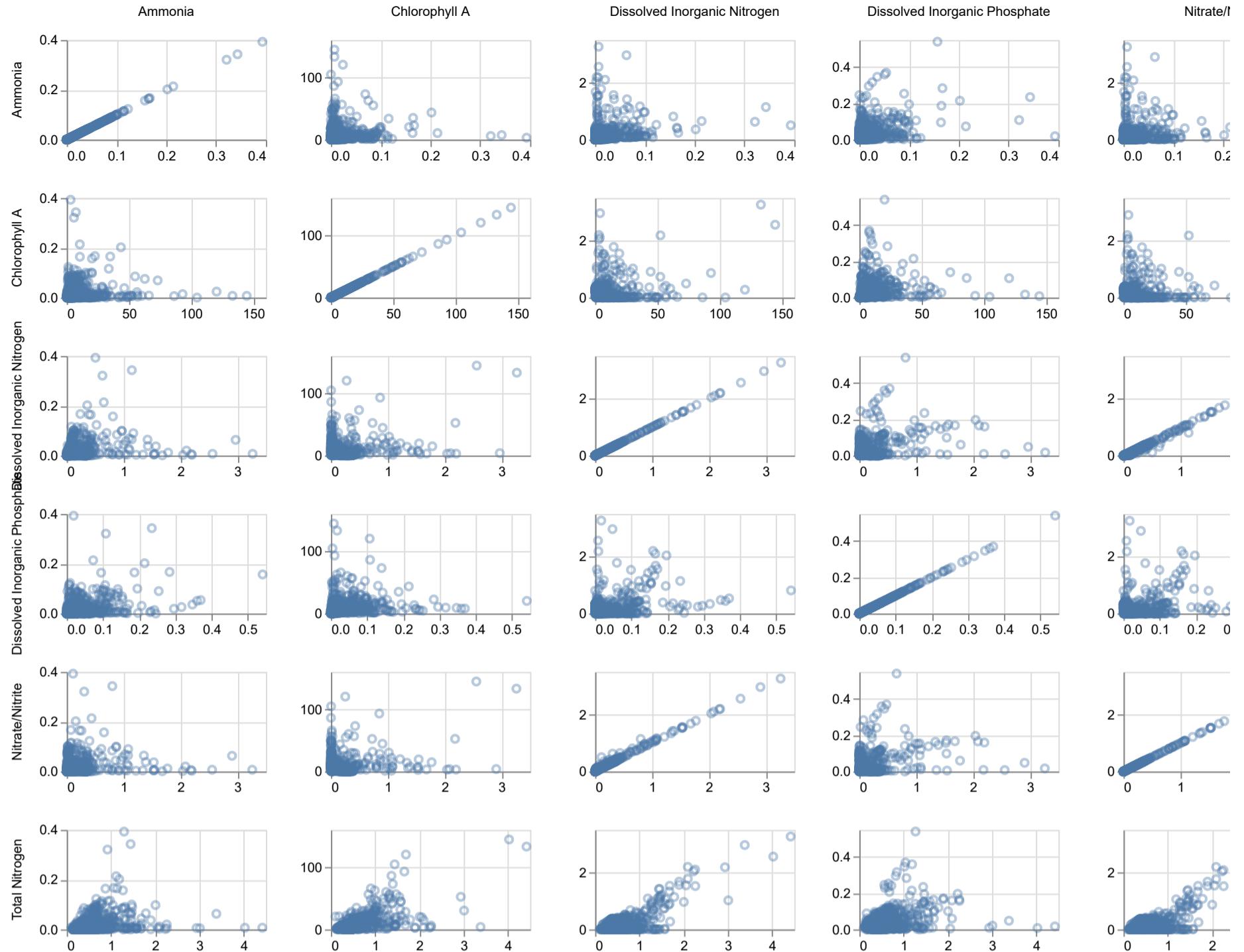
# Long form dataframe for plotting panel
scatter_df_long = x_mx.melt(
    var_name = 'row',
    value_name = 'row_index'
).join(
    pd.concat([x_mx, x_mx, x_mx, x_mx, x_mx,
              x_mx, x_mx], axis = 0).reset_index(),
).drop(
    columns = 'index'
).melt(
    id_vars = ['row', 'row_index'],
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

# panel
scatter_panel = alt.Chart(scatter_df_long).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')
```

In [27]:

```
# Pairwise relationship for plotting panel
scatter_panel
```

Out[27]:



In [28]: `scatter_panel.save('variance_scatter.html')`

In [29]: `# Correlation matrix for just nutrients and productivity
x_mx.corr()`

Out[29]:

	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite	Total Nitrogen
Ammonia	1.000000	0.076214	0.223906	0.373070	0.128686	0.288228
Chlorophyll A	0.076214	1.000000	0.188035	0.196624	0.185112	0.641165
Dissolved Inorganic Nitrogen	0.223906	0.188035	1.000000	0.258240	0.995142	0.716507
Dissolved Inorganic Phosphate	0.373070	0.196624	0.258240	1.000000	0.224840	0.378746
Nitrate/Nitrite	0.128686	0.185112	0.995142	0.224840	1.000000	0.700950
Total Nitrogen	0.288228	0.641165	0.716507	0.378746	0.700950	1.000000

In []:

In [30]: `# correlation matrix for all quantitative variables
data.iloc[:,2:15].corr()`

Out[30]:

	Water depth (in meters)	Latitude	Longitude	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite	Total Nitrogen
Water depth (in meters)	1.000000	0.306774	-0.078211	-0.122657	-0.144925	-0.025702	-0.141074	-0.014072	-0.215607
Latitude	0.306774	1.000000	-0.020342	-0.100746	-0.241976	0.102828	-0.325458	0.114923	-0.279980
Longitude	-0.078211	-0.020342	1.000000	-0.042832	0.053520	-0.026439	-0.282131	-0.021377	0.027846
Ammonia	-0.122657	-0.100746	-0.042832	1.000000	0.076214	0.223906	0.373070	0.128686	0.288228
Chlorophyll A	-0.144925	-0.241976	0.053520	0.076214	1.000000	0.188035	0.196624	0.185112	0.641165
Dissolved Inorganic Nitrogen	-0.025702	0.102828	-0.026439	0.223906	0.188035	1.000000	0.258240	0.995142	0.716507
Dissolved Inorganic Phosphate	-0.141074	-0.325458	-0.282131	0.373070	0.196624	0.258240	1.000000	0.224840	0.378746
Nitrate/Nitrite	-0.014072	0.114923	-0.021377	0.128686	0.185112	0.995142	0.224840	1.000000	0.700950
Total Nitrogen	-0.215607	-0.279980	0.027846	0.288228	0.641165	0.716507	0.378746	0.700950	1.000000

In [31]:

```
# store correlation matrix
corr_mx = x_mx.corr()

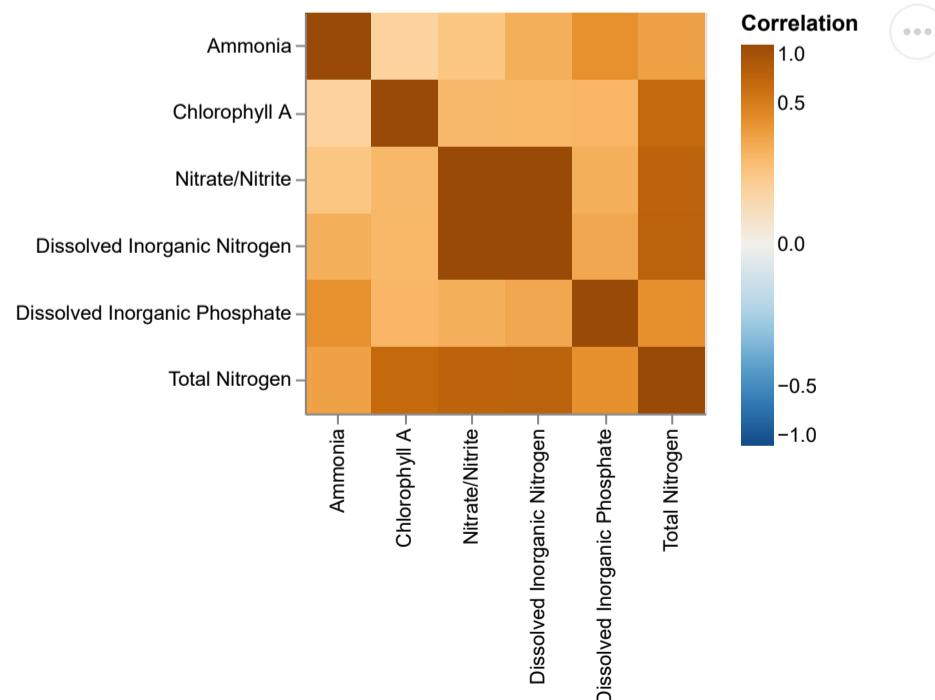
# melt to long form
corr_mx_long = corr_mx.reset_index().rename(
    columns = {'': 'row'}
).melt(
    id_vars = 'row',
    var_name = 'col',
    value_name = 'Correlation'
)

# 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.LegendTickCount = 5)
).properties(width = 200, height = 200)

# 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.LegendTickCount = 5)
).properties(width = 200, height = 200)
```

In [32]: heatmap

Out[32]:



In [33]: corr_mx

Out[33]:

	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite	Total Nitrogen
Ammonia	1.000000	0.076214	0.223906	0.373070	0.128686	0.288228
Chlorophyll A	0.076214	1.000000	0.188035	0.196624	0.185112	0.641165
Dissolved Inorganic Nitrogen	0.223906	0.188035	1.000000	0.258240	0.995142	0.716507
Dissolved Inorganic Phosphate	0.373070	0.196624	0.258240	1.000000	0.224840	0.378746
Nitrate/Nitrite	0.128686	0.185112	0.995142	0.224840	1.000000	0.700950
Total Nitrogen	0.288228	0.641165	0.716507	0.378746	0.700950	1.000000

In [34]: heatmap_html = heatmap.save('heatmap.html')

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

In [35]: data

Out[35]:

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

1092 rows × 16 columns

In [36]:

```

N_box = alt.Chart(scatter_df[scatter_df['Nutrient'] == 'Total Nitrogen']).mark_boxplot(
    size = 50
).encode(
    x = alt.X('Region', title = 'Total Nitrogen'),
    y = alt.Y('Level'),
    color = alt.Color('Region')
).properties(
    width = 300,
    height = 250
)

P_box = alt.Chart(scatter_df[scatter_df['Nutrient'] == 'Total Phosphorus']).mark_boxplot(
    size = 50
).encode(
    x = alt.X('Region', title = 'Total Phosphorus',
              scale = alt.Scale(zero = False)),
    y = alt.Y('Level'),
    color = alt.Color('Region')
).properties(
    width = 300,
    height = 250
)

NH3_box = alt.Chart(scatter_df[scatter_df['Nutrient'] == 'Ammonia']).mark_boxplot(
    size = 50
).encode(
    x = alt.X('Region', title = 'Ammonia',
              scale = alt.Scale(zero = False)),
    y = alt.Y('Level'),
)

```

Loading [MathJax]/extensions/Safe.js

```

        color = alt.Color('Region')
    ).properties(
        width = 300,
        height = 250
    )

```

In [37]: total_region_nutrients = N_box & P_box & NH3_box

In [38]: total_region_nutrients.save('total_region_nutrients.html')

In [39]: scatter_df

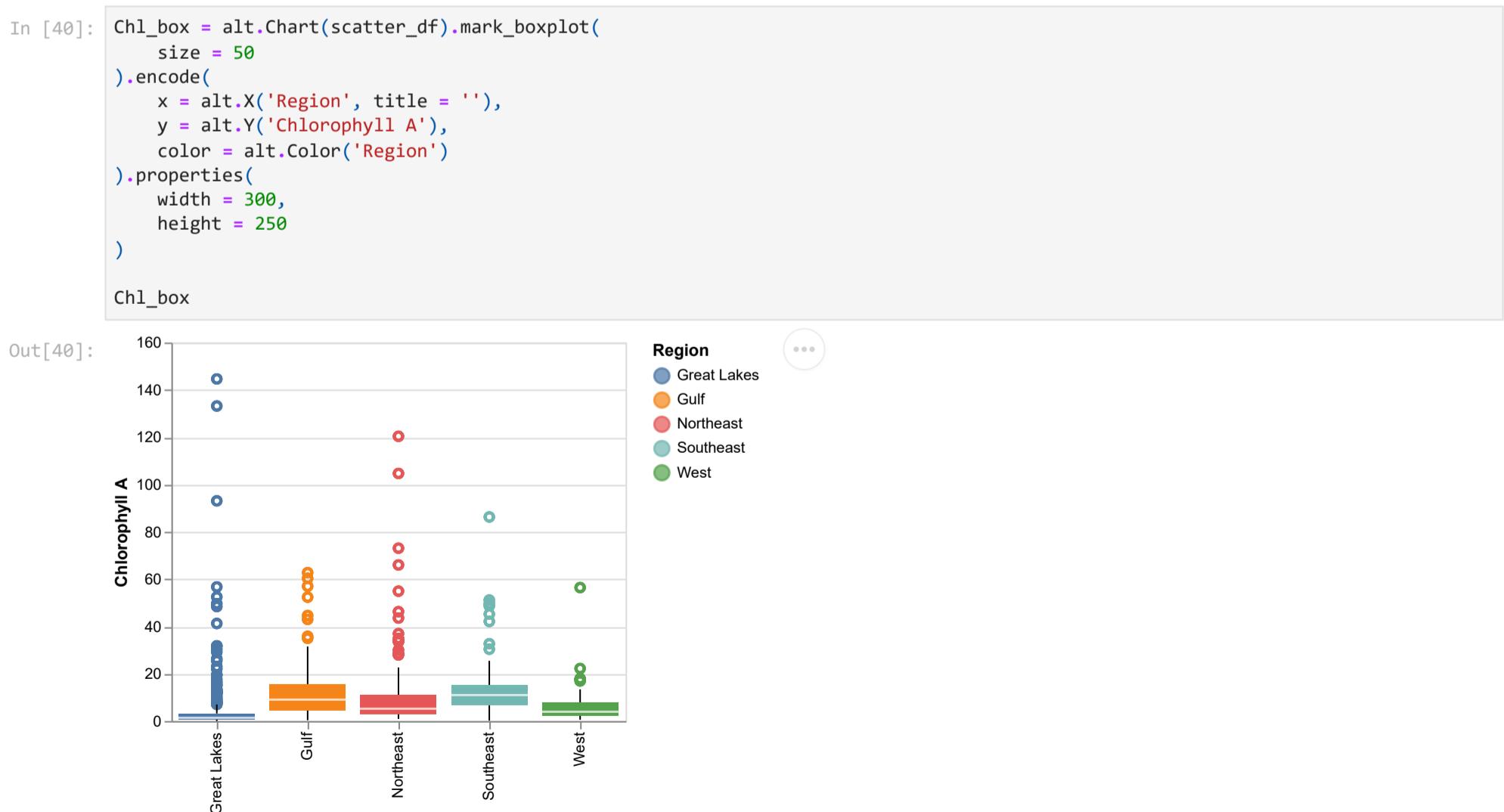
Out[39]:

	UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Chlorophyll A	Province	Nutrient	Level
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	3.34	Californian Province	Ammonia	0.000000
1	60	CA	7/1/2010	San Diego Bay	West	3.5	32.71424	-117.23527	2.45	Californian Province	Ammonia	0.010000
2	61	CA	7/1/2010	Mission Bay	West	2.2	32.78372	-117.22132	3.82	Californian Province	Ammonia	0.000000
3	62	CA	7/1/2010	San Diego Bay	West	9.5	32.72245	-117.20443	6.13	Californian Province	Ammonia	0.000000
4	63	NC	6/9/2010	White Oak River	Southeast	1.0	34.75098	-77.12117	9.79	Carolinian Province	Ammonia	0.002000
...
6547	16727	MI	6/18/2010	Lake Michigan	Great Lakes	0.6	44.98607	-85.64046	0.75	Great Lakes Province	Total Phosphorus	0.000000
6548	16728	MI	6/25/2010	Lake Michigan	Great Lakes	2.3	44.94789	-85.94790	2.27	Great Lakes Province	Total Phosphorus	0.006249
6549	16729	MI	6/16/2010	Lake Michigan	Great Lakes	31.2	44.83721	-85.52862	1.11	Great Lakes Province	Total Phosphorus	0.000000
6550	16730	CA	6/29/2010	San Diego Bay	West	4.1	32.66443	-117.13879	2.11	Californian Province	Total Phosphorus	0.044127
6551	16731	CA	6/29/2010	San Diego Bay	West	4.8	32.66243	-117.12712	2.19	Californian Province	Total Phosphorus	0.041821

6552 rows × 12 columns

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.



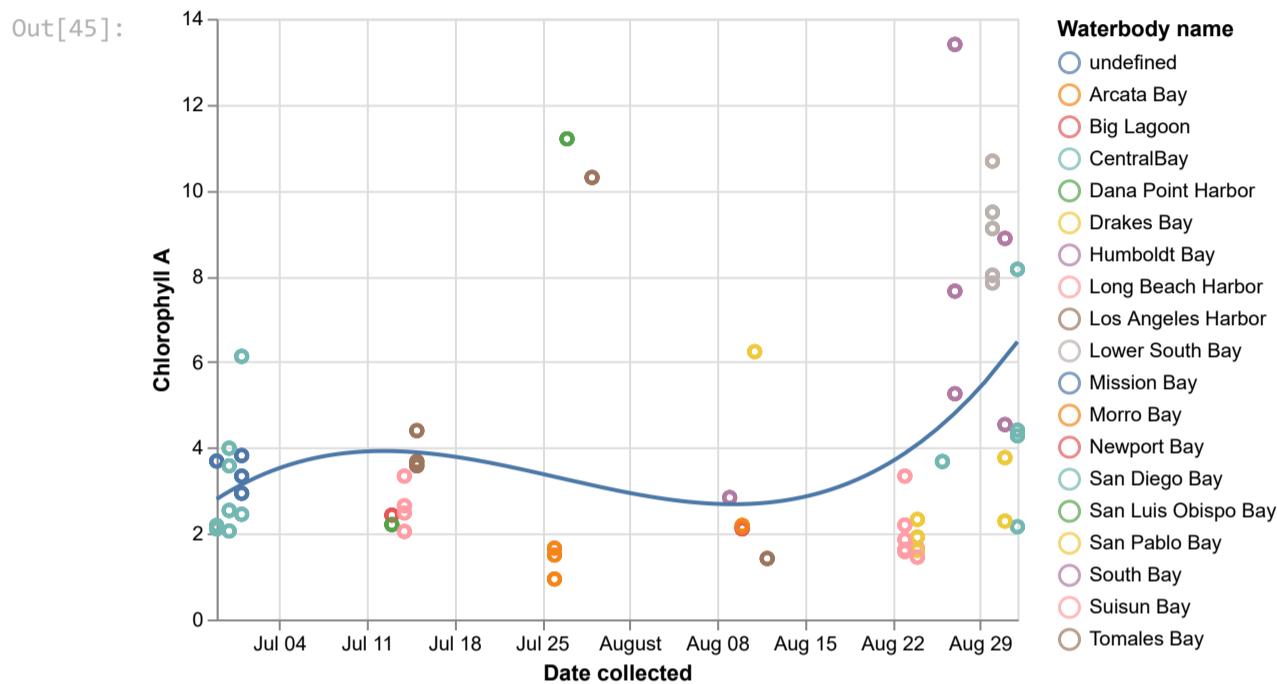
In [41]: Chl_box.save('chl_box.html')

Out[42]:	UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Chlorophyll A	Province	Nutrient	Level
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	3.34	Californian Province	Ammonia	0.0

```
In [43]: ca_scatter = alt.Chart(scatter_df[scatter_df['State'] == 'CA']).mark_point().encode(
    x = 'Date collected:T',
    y = 'Chlorophyll A',
    color = 'Waterbody name'
)
```

```
In [44]: ca_total = ca_scatter + ca_scatter.transform_regression('Date collected',
    'Chlorophyll A',
    method = 'poly',
    order = 3).mark_line()
```

```
In [45]: ca_total
```



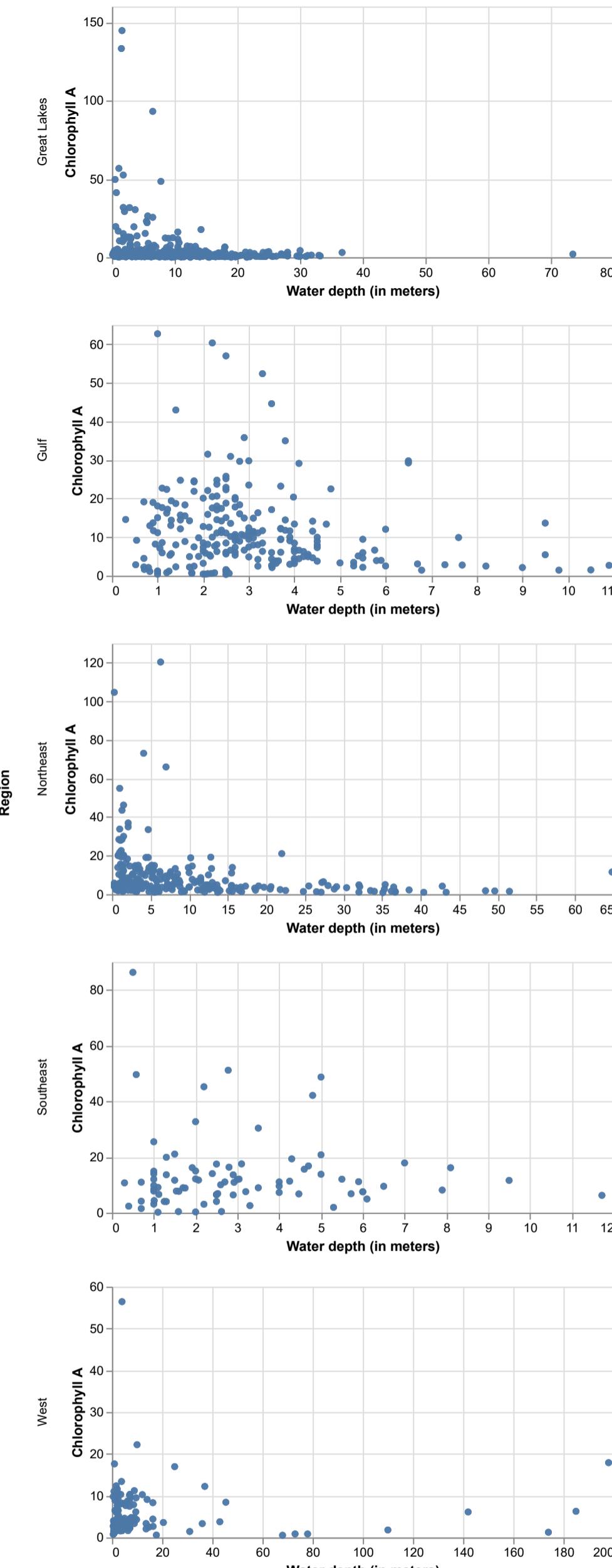
```
In [46]: ca_total.save('ca_total.html')
```

```
In [47]: data.head(1)
```

Out[47]:	UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Province	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitrite
0	59	CA	7/1/2010	Mission Bay	West	2.5	32.77361	-117.21471	Californian Province	0.0	3.34	0.014	0.028	0.014

```
In [48]: depth_scatter_facet = alt.Chart(scatter_df).mark_circle(
    opacity = 0.4
).encode(
    x = alt.X('Water depth (in meters)',
        scale = alt.Scale(zero = True)),
    y = alt.Y('Chlorophyll A',
        scale = alt.Scale(zero = True))
).properties(
    width = 400,
    height = 200
).facet(
    row = 'Region'
).resolve_scale(x = 'independent', y = 'independent')
depth_scatter_facet
```

Out[48]:

In [49]: `data[data['Water depth (in meters)'] > 180]`

Out[49]:

UID	State	Date collected	Waterbody name	Region	Water depth (in meters)	Latitude	Longitude	Province	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Nitrate/Nitr	
118	422	WA	7/10/2010	Puget Sound	West	185.0	47.80405	-122.457533	Columbian Province	0.001	6.26	0.009	0.010	0.0
245	624	WA	7/7/2010	Puget Sound	West	198.0	47.39325	-122.344383	Columbian Province	0.011	17.90	0.023	0.009	0.0

Loading [MathJax]/extensions/Safe.js

```
In [50]: depth_scatter = alt.Chart(scatter_df).mark_circle(
    opacity = 0.4
).encode(
    x = alt.X('Water depth (in meters)',
              scale = alt.Scale(zero = True)),
    y = alt.Y('Chlorophyll A',
              scale = alt.Scale(zero = True))
).properties(
    width = 400,
    height = 200
)
depth_scatter
```

