Spatio-temporal Relationships of Water Quality Measurements in the Bay Delta: A Time and Frequency Domain Approach

Sara Stoudt

1 Introduction

The questions that I will be considering in this study are:

1. Within a measurement station in the Bay Delta, are different nutrients significantly correlated? This knowledge can help scientists generate hypotheses of water quality mechanisms at play locally.
2. Across measurement stations in the Bay Delta, are the same nutrients significantly correlated? This knowledge can help scientists assess connectivity and predict where water quality damage could spread.
3. Across measurement stations in the Bay Delta, are different nutrients significantly correlated? This knowledge can help scientists generate hypotheses of more big picture water quality mechanisms at play in the whole system.

   Long-term time series of water quality measurements spanning multiple locations in the San Francisco Estuary are available. The challenge is to leverage this data to learn how nutrient concentrations vary across space and time in the Delta, what the major drivers of this variability are, and if any mechanistic insights can be gained by exploring the relationships between various variables collected.

2 Data

I have data from collaboration with the San Francisco Estuary Institute for a different project that I did last summer. The data is monthly water nutrient measurements from multiple stations spread across the San Francisco Bay Delta. The left panel of Figure 1 shows all stations marked with black dots. I am narrowing my focus to the stations marked in red as they have more continual records and are in the part of the Bay Delta system where we would expect the largest correlations based on previous work.

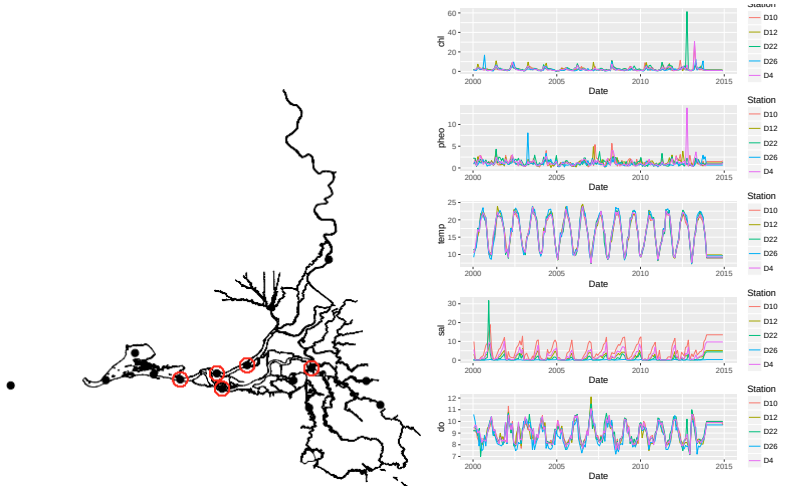


Figure 1:

I will narrow my focus to simplify the combinatorics of pairwise comparisons to five nutrient and water quality measurements, namely chlorophyl, dissolved oxygen, pheophytin a, temperature, and salinity. These variables have more continual records and capture both water quality and nutrient aspects of interest. The original series for these per station are shown in the right panel of Figure 1.

2.1 Context and Motivation

Ecologists working at the San Francisco Estuary Institute are interested in having their hypothesized mechanisms of how nutrients interact in the Bay Delta supported (or refuted) by the data, so that their arguments are stronger when they are advocating for policy decisions in the Bay Delta as there has been contention over these decisions. Knowing how different nutrients and water quality measures interact with one another across different areas in the Delta can help us make data driven policy decisions. Therefore, the goal of this project is to pick out significant correlations and cross correlations between nutrients across different stations, noting the lags and frequencies in context. We are less concerned with being "wrong" (due to less than ideal assumptions) and more interested in systematically sifting through all possible station-nutrient pairs and synthesizing some preliminary results that can reasoned about in context by ecologists.

2.2 Preprocessing

For correlation and coherence analysis, we assume that our data is stationary and that there are complete records. Before starting my analysis, I dealt with the missingness and adjusted the original series to obtain approximately stationary series to work with moving forward.

2.2.1 Missingness

I examined each station's series for missingness. There are often stretches of time where a few variables at each station are missing together as the sensor broke. To avoid making imputation the focus of my project I decided to work with the period of time between January 2010 to June 2015 as these records were the most complete across the stations of interest. For the few missing values in these records, I used the average of the values of that variable at the same station in the month before and after the missing value.

2.2.2 Detrending and Deseasonalizing

To make the stationarity assumption more tenable, I first detrend and deseasonalize each station-variable series by fitting a Generalized Additive Model (GAM) to each series with date as a decimal value (date.dec) and day of year (doy) as the predictors and taking the residuals as my series to work with moving forward.

$$nutrient \sim Station + s(date.dec, bs = "cs", by = Station, k = 25) + s(doy, bs = "cs", by = Station, k = 25)$$

This GAM allows for a station level effect, a different smooth of date.dec and doy by station. Standard diagnostics were checked to ensure that the basis dimension (k) was not too small. The left panel of Figure 2 shows some examples of the original series and the components that were removed via the GAM.

2.2.3 Variance Stabilization

I plotted the residual series and tried variance stabilization transformations, looking at the transformed series to see if this helped. The log transform was used after shifting the series up to be all positive for pheophytin a and chlorophyll for all stations and for the salinity of station D22. The log transform did not seem to help the other station-nutrient combinations. I also tried the Box Cox transform for the rest of the station-nutrient combinations, but the confidence interval for the tuning parameter included one, which

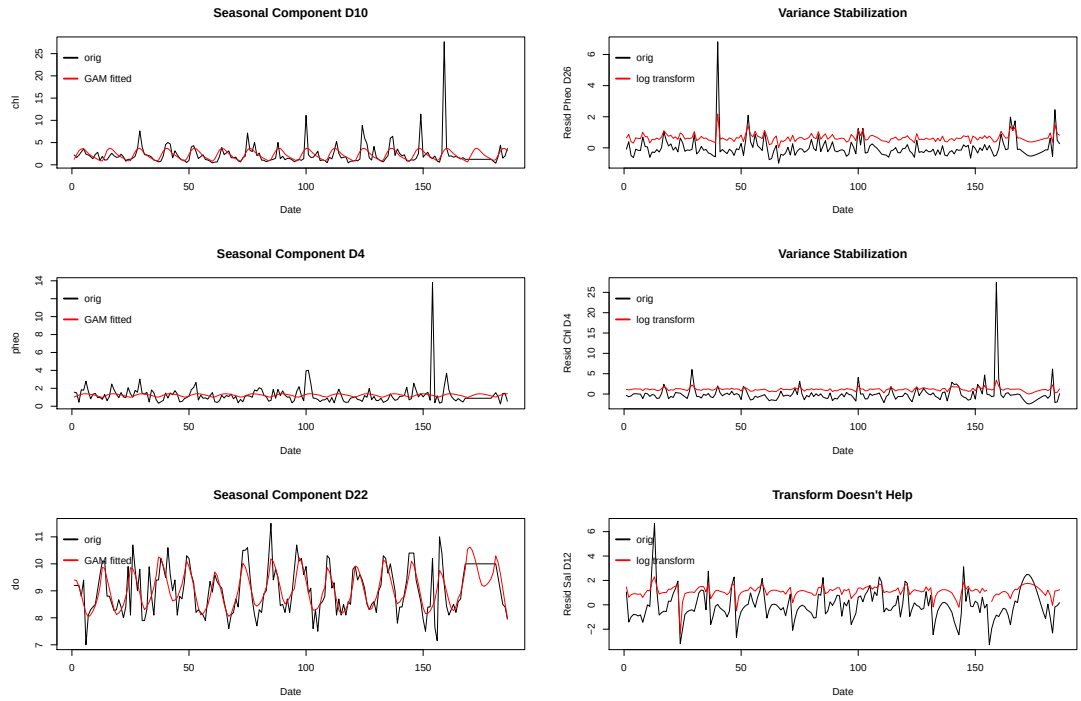


Figure 2:

is consistent with no transformation. The right panel of Figure 2 shows some examples of the residual series and their transformations.

3 Analyses

3.1 Assessing Significance and Dealing with Multiple Testing

When comparing two series in the time domain, I find the lag that has the maximum absolute value of cross correlation. I limit my search to positive lags up to 12 because we are most interested in effects within a year. I then use a rotation test (Dorum et al. [2009],Dorum et al. [2014]) to get the p-value for that maximum value. The rotation test preserves the time-series structure by taking the first value of the series and rotating it to the end of the series, taking the first two values of the series and rotate them to the end of the series, etc. This gives us $N - 1$ "permutations" for a non-parametric significance test. This is many more samples than we could have obtained using a block bootstrap, a typical approach for time series applications, allowing for a larger power by yielding a smaller minimum p-value.

I test every possible station-nutrient combination, breaking them into three groups:

within station different nutrient, across stations same nutrient, and across stations different nutrient. Because of this serious multiple testing, I then apply the Benjamini-Yekutieli (B-Y) procedure to adjust the p-values for multiple testing (Benjamini and Yekutieli [2001]) within groups. I chose to adjust for multiple testing by group because each group answers a different scientific question. I am not worried about being "wrong" in terms of identifying too many promising pairs, as my goal is to merely narrow down the possible combinations so that ecologists can focus their attention on a subset and apply their domain knowledge to further eliminate based on mechanistic hypotheses.

In the frequency domain, the rotation test leads to over sensitive results because the rotation causes a large discontinuity in the Fourier transform. Therefore, I use the theoretical null distribution based on the F distribution to calculate the p-values. As we see below, the frequency domain approach is more sensitive and picks out more statistically significant pairs. Part of this is most likely due to this cruder measure of statistical significance. Because I am using a cruder test statistic for assessing significance that is less tenable if the stationary assumption is met, I check that pre-whitening the series doesn't change how many significant pairs are found. I use the prewhite function in the psd package to fit a low order ARMA model automatically and prewhiten before assessing cross coherence. When I do this, the number of significant pairs remains almost completely unchanged. A more sophisticated approach would be to try to threshold the discontinuity away in the rotation testing, but I leave that for future work.

3.2 Q1: Different Nutrients Within a Station

3.2.1 Time Domain: Cross Correlation

Out of 100 possible combinations, 23 within station different variable combinations have p-values less than my cut-off of 0.05. After the multiple testing adjustment, only one was significant: station D12 dissolved oxygen and temperature. The lag with the largest autocorrelation was 8 months.

As water temperature rises, oxygen solubility decreases, allowing for dissolved oxygen to increase (Fondriest.Environmental.Inc.), so this positive autocorrelation makes sense

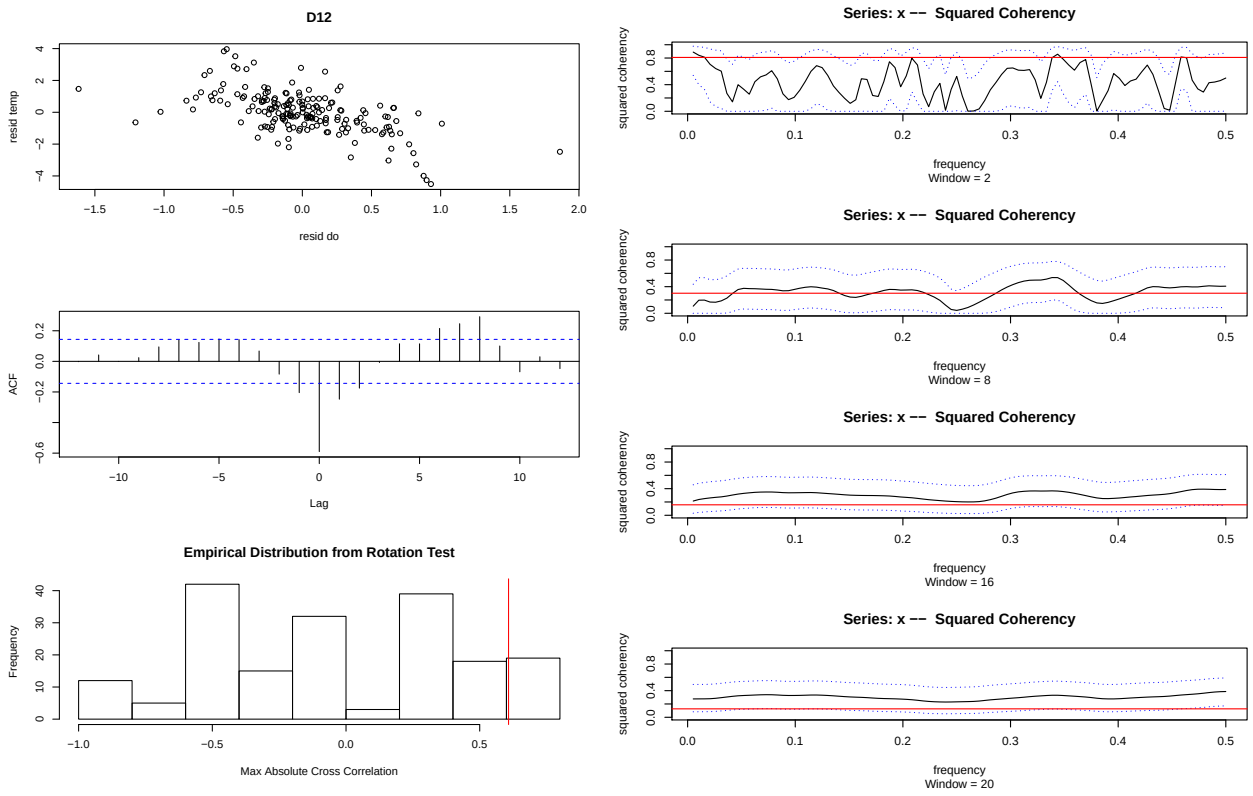


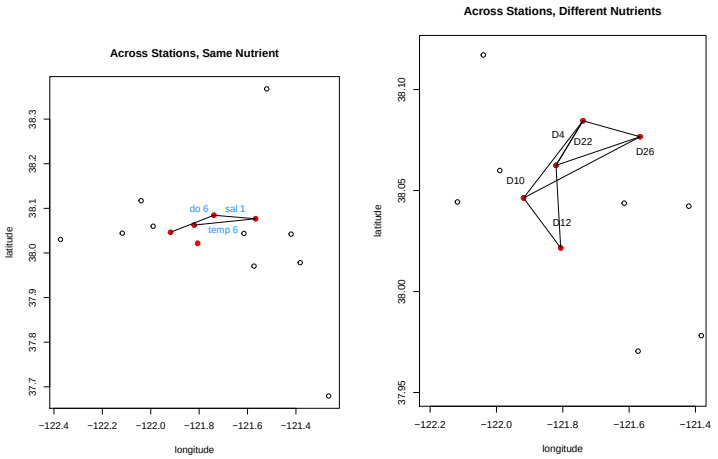
Figure 3:

in context. The lag of 8 months corresponds to two-thirds of a year. The left panel of Figure 3 shows the process: a simple plot of each variable's residual values, the cross correlation function plot, and the empirical distribution of maximum cross correlations under the rotation test with our observed value marked by the red line.

3.2.2 Frequency Domain: Cross Coherence

Out of 100 possible combinations, 65 within station different variable combinations have p-values less than my cut-off of 0.05. After the multiple testing adjustment, 26 were significant, including D12 dissolved oxygen and temperature as identified in the time domain. Dissolved oxygen-temperature and chlorophyl-pheophytin a are the most common pairs. See the Appendix for a table of these.

For all of the frequency domain analysis, I taper the series and use a double-pass kernel smoother before estimating the spectrum. I experimented with different tapering and windows for the kernel smoothing until I found parameters that gave me smooth estimates of the spectrum. The right panel of Figure 3 shows the cross coherence plots for different window choices. Since I wanted to isolate a single maximum, I use a double-



	Station 1	Station 2	Nutrient 1	Nutrient 2	Lag Max
1	D10	D22	do	chl	2
2	D26	D22	sal	chl	2
3	D10	D12	sal	do	4
4	D4	D22	temp	do	4
5	D10	D26	temp	do	7
6	D4	D22	pheo	sal	3
7	D26	D4	pheo	sal	3
8	D12	D4	do	temp	8

Figure 4:

pass window of size 16. There is not a large difference between this and a window of 20 while windows of 2 and 8 allow for too many details that may not be robust.

3.3 Q2: Same Nutrient Across Stations

3.3.1 Time Domain: Cross Correlation

Out of 100 possible combinations, 18 across station same variable combinations have p-values less than my cut-off of 0.05. After the multiple testing adjustment, only three were significant. In the first panel of Figure 4 we can see where in the Delta network, these significant relationships occur and what lag yielded the maximum absolute cross correlation.

We can see that dissolved oxygen, temperature, and salinity can be linked across stations. The difference in the lags that produced the maximum absolute cross correlation can be interpreted as a mix of flow distance and the strength of the flow between stations. For example, there is only a lag of one month for salinity while the other two variables have a six month lag. However, the link for salinity has a shorter flow path. We note that we do not have any triplets where the same variable is linked across three stations in a row. This suggests that the flow of nutrients is not particularly strong across far distances.

3.3.2 Frequency Domain: Coherence

Out of 100 possible combinations, all 100 across station same variable combinations have p-values less than my cut-off of 0.05. After the multiple testing adjustment, 90 were

	Station 1	Station 2	Nutrient 1	Nutrient 2
1	D26	D10	pheo	pheo
2	D26	D12	pheo	pheo
3	D10	D26	pheo	pheo
4	D12	D26	pheo	pheo
5	D26	D10	sal	sal
6	D26	D22	sal	sal
7	D10	D26	sal	sal
8	D22	D26	sal	sal
9	D4	D26	sal	sal
10	D26	D4	sal	sal

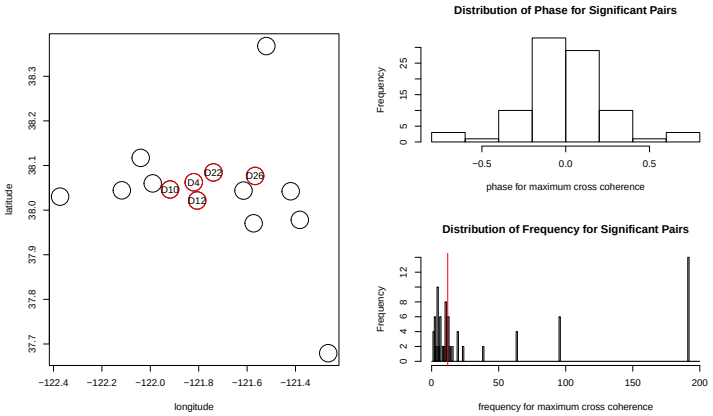


Figure 5:

significant. For simplicity, the left panel of Figure 5 reports the ten combinations that **were not** statistically significant and the right panel displays summaries of the phase and frequency of the maximum cross coherence for the 90 significant pairs. The middle panel reminds us where each station is. We can see that pheophytin a and salinity across stations do not have strong cross coherences across larger distances as D26 is the furthest from D10 and D12. However, salinity does not have strong cross coherence at even smaller distances than pheophytin a since even station D22 and D4 appear in the non-significant list when paired with D26. When we look at the distribution of the phase where the maximum cross coherence occurs in the significant pairs, we see that the shifts between the two series are negligible (less than the finest level of data granularity, one month). Similarly some of the frequencies where the maximum cross coherence occurs in the significant pairs are very large, in which case we suspect these are spurious or at least not informative in context. There are quite a few pairs that fall to the left of the vertical red line (12 months). These are the ones we want to compare to the time domain results to see if they overlap. See the appendix for a table of these 46 pairs. Only one of the three found in the time domain (dissolved oxygen pair) is also found in the frequency domain.

3.4 Q3: Different Nutrient Across Stations

3.4.1 Time Domain: Cross Correlation

Out of 400 possible combinations, 83 across station different variable combinations have p-values less than my cut-off of 0.05. After the multiple testing adjustment, only eight were significant. The second and third panels of Figure 4 show the location connections, the nutrients, and lags with maximum absolute cross correlation. All nutrients have some connection across two stations.

Dissolved oxygen and temperature are the most common pair, which again makes sense in context as we saw above in the within station analysis. Pheophytin a and salinity are connected in a triangle like path; D4's pheophytin a is matched with D22's salinity while D4's salinity is matched with D26's pheophytin a. Yet D26 and D24 are not matched based on pheophytin a. We expect the maximum cross correlations to be negative for these as higher salt values are known to decrease pheophytin a levels in other scientific contexts (Saida et al. [2014]), and we see this. Chlorophyll is matched to dissolved oxygen and salinity. We do not expect a strong link between dissolved oxygen and chlorophyll (Morgan et al. [2006], Rocha et al. [2009], Kunlasak et al. [2013]), *so this may be something to study further*. Overall, it is reassuring to see that the larger lags occur between stations that are further apart on the flow path. This makes sense as it would take longer for water to flow between far away stations, leading to slower changes in the water quality and chemistry.

3.4.2 Frequency Domain: Coherence

Out of 400 possible combinations, all 266 across station same variable combinations have p-values less than my cut-off of 0.05. After the multiple testing adjustment, 60 were significant. Again, this is too many pairs to synthesize easily, so I see how many of each station combination and nutrient pair there are in the first and second panel of Figure 6. The third panel shows the distribution of the phase and frequency where the maximum cross coherence occurs in the significant pairs.

Noting that station combinations should be symmetric (confirmed by the first table in Figure 6) D4-12 and D4-D22 have the most significant pairs across nutrient combinations. We can refer back to the second panel of Figure 5 to remember how far away different

	Station Combination	Number of Significant Pairs
1	D12_D4	5
2	D22_D4	5
3	D4_D12	5
4	D4_D22	5
5	D10_D22	4
6	D12_D22	4
7	D22_D10	4
8	D22_D12	4
9	D10_D12	3
10	D12_D10	3
11	D26_D4	3
12	D4_D26	3
13	D10_D26	2
14	D10_D4	2
15	D26_D10	2
16	D4_D10	2
17	D12_D26	1
18	D22_D26	1
19	D26_D12	1
20	D26_D22	1

	Nutrient Combination	Number of Significant Pairs
1	do_temp	17
2	temp_do	17
3	chl_sal	4
4	sal_chl	4
5	pheo_sal	3
6	sal_pheo	3
7	chl_pheo	2
8	pheo_chl	2
9	sal_temp	2
10	temp_sal	2
11	chl_do	1
12	do_chl	1
13	do_sal	1
14	sal_do	1

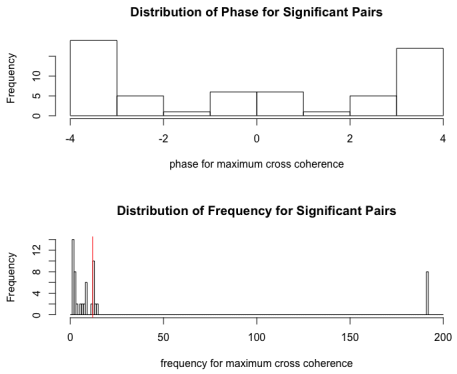


Figure 6:

stations are from one another. In terms of distance, these results make sense as D12, D22, and D4 are close to one another. The difference between the tallies of station pairs are fairly close to one another, but D26-D22 and D26-D12 come in last place with only one significant pair apiece, and this also makes sense as these are pairs that are further away on the station network.

We see slightly more of a spread in variable combinations. Temperature and dissolved oxygen have overwhelmingly the most significant pairs, and this matches what we saw in the time domain and the scientific context. Dissolved oxygen paired both with chlorophyll and salinity are tied for the least number of significant pairs. We saw above that this is not surprising for chlorophyll, but we do expect increased salinity to lead to decreased dissolved oxygen (APEC.Water), *so this may be interesting to look into further.*

Are the eight pairs found to be statistically significant in the time domain still significant in the frequency domain? Only two are (D4-D22 temp-do and D10-D22 temp-do). But now when we look at the phase for the maximum cross coherence per pair, the distribution shows practically significant shifts (up to +/- four months). Similarly, looking at the frequency for the maximum cross coherence per pair, most fall to the left of the red line representing frequencies of interest (<12 months), so the signals are stronger here, from a practical in context perspective, than they were for across station same nutrients pairs in the frequency domain.

4 Conclusions and Future Work

In this work we have tested all possible combinations between a subset of station and nutrient records to look for significant cross correlations and cross coherences that would help us infer mechanisms of how nutrients and other water quality measurements interact with one another across the Bay Delta. Ecologists can now focus their attention on a smaller subset of possible combinations to further reason about mechanisms that make sense in a chemical context.

Future work would be to try to strengthen our claims about correlation and coherence to causation by testing the significant pairs found here further using mutual information techniques (Brillinger [2002]). Getting a causal interpretation would help strengthen scientists' cases when recommending water quality management decisions.

My answer to my proposed questions are:

1. A dissolved oxygen-temperature connection was found by both the time and frequency domain analysis while the frequency domain also identified chlorophyll-pheophytin a pairs as having statistically significant cross coherence at many stations.
2. The dissolved oxygen, temperature, and salinity at different stations that are closer to one another in the network are statistically significantly cross-correlated in the time domain while pheophytin a and salinity are notably not statistically significantly coherent in the frequency domain.
3. Statistically significant connections between temperature and dissolved oxygen across stations are the most prominent in both the time and frequency domain.

References

APEC.Water. Does salt concentration or salinity of water affect solubility of oxygen? URL http://www.freedrinkingwater.com/water_quality/water-science/j-9-08-does-salt-concentration-salinity-water-affect-solubility-oxygen.htm.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

D. R. Brillinger. Second-order moments and mutual information in the analysis of time series. *Recent Advances in Statistical Methods*, pages 64–76, 2002.

G. Dorum, L. Snipen, M. Solheim, and S. Saebo. Rotation testing in gene set enrichment analysis for small direct comparison experiments. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

G. Dorum, L. Snipen, M. Solheim, and S. Saebo. Rotation gene set testing for longitudinal expression data. *Biometrical Journal*, 56.6:1055–1075, 2014.

Fondriest.Environmental.Inc. Fundamentals of environmental measurements: Dissolved oxygen. URL http://www.fondriest.com/environmental-measurements/parameters/water-quality/dissolved-oxygen/.

K. Kunlasak, C. Chitmanat, N. Whangchai, J. Promya, and L. Lebel. Relationships of dissolved oxygen with chlorophyll-a and phytoplankton composition in tilapia ponds. *International Journal of Geosciences*, 4:46–53, 2013.

A. M. Morgan, T. V. Royer, M. B. David, and L.E. Gentry. Relationships among nutrients, chlorophyll-, and dissolved oxygen in agricultural streams in illinois. *Journal of Environmental Quality*, 35(4):1110–1117, 2006.

R.R.A. Rocha, S.M. Thomaz, P. Carvalho, and L.C. Gomes. Modeling chlorophyll-a and dissolved oxygen concentration in tropical floodplain lakes (paraná river, brazil). *Brazilian Journal of Biology*, 69.2:491–500, 2009.

C. Saida, B. Houria, and B. Mebarek. Interactive effects of salinity and potassium on physio-morphological traits of tomato. *Agriculture and Biology Journal of North America*, 2014.

NOTE: I discussed some advanced techniques for dealing with multiple testing with Aaditya Ramdas, a post-doc here. He suggested that the B-Y adjusted p-value is the most appropriate for my pairwise testing scenario. He also pointed me to the rotation test for significance testing of time-series given my relatively small sample size ($N = 186$).