Appendix 2: An empirical investigation of spatiotemporal patterns in dissolved inorganic macronutrients in the Sacramento–San Joaquin River Delta

Prepared by:

Philip Bresnahan, Thomas Jabusch, David Senn, Philip Trowbridge, Micha Salomon, Emily Novick

San Francisco Estuary Institute – Aquatic Science Center
4911 Central Ave
Richmond, CA 94804

An empirical investigation of spatiotemporal patterns in dissolved inorganic macronutrients in the Sacramento–San Joaquin River Delta

Abstract

Here we present an empirical analysis of trends, both seasonal and inter-annual, in water quality parameters in the Sacramento–San Joaquin River Delta, California. Water quality parameters, including dissolved inorganic macronutrient (*e.g.*, nitrate, ammonium, phosphate, silicate) concentrations, and ancillary physical/biogeochemical quantities (temperature, conductivity, dissolved oxygen, and chlorophyll-*a* concentrations) have been recorded for almost four decades at eleven stations. While it is well known that these water quality parameters have a large annual cycle and have undergone various longer-term changes in the past forty years, the magnitudes of these changes are more poorly quantified. We implement a type of factor analysis/dimension reduction known as non-negative matrix factorization in order to tease apart various contributions to water quality variability. Non-negative matrix factorization proved particularly useful in the physical interpretation of various drivers and proved capable of elucidating important trends that aren't otherwise readily identifiable.

Introduction

The Delta has been divided into subregions for different purposes by different programs. Many of these subregions are similar or overlap suggesting that a small number of subregions are sufficient to characterize variability in the Delta. It is expected that the dominant factors affecting nutrient concentrations are similar within each subregion and are distinct from those in other subregions.

Here we characterize the different trends in both space and time across the Delta and illustrate how these trends exhibit themselves within different regions. Task 1 describes the rationale for subdividing the Delta as we show here. Using those subregions, we demonstrate the heterogeneity of this unique ecosystem. We take several approaches to examine this heterogeneity, from calculating monthly means of different time-series

Within a complex dataset consisting of time-series at many stations across many variables, there are often "latent" or hidden drivers that account for significant amounts of variability in the dataset but cannot be easily spotted via visual inspection of the many time-series. Water quality variability can be thought of as the sum or superimposition of various processes. Whereas principal component analysis (PCA) allows for the sum and subtraction of processes which can lead to non-physical interpretation, non-negative matrix factorization (NMF) constrains the factors to be positive contributions only. For example, an eigenvector (*i.e.*, a "mode") in PCA can have both negative and positive values, suggesting that a given mode can contribute negatively to a given site but positively to another; contrarily, NMF allows for additive processes only. For instance, non-negative matrix factorization can extract seasonal and interannual processes and superimpose them according to the strengths of their modes across stations and parameters. Importantly, the constraint of non-negativity does not imply that removal processes (*e.g.*, nutrient removal via assimilation or transformation) cannot occur.

Methods

The purpose of this portion of the analysis is to determine the "status and trends" of the Sacramento–San Joaquin River Delta (hereafter the Delta). Stations have been maintained at the following locations from 1975–present. Many other stations have been sampled as well but are no longer active. See Novick

et al. (2015) for background on additional stations as well as mass balance and 1-D hydrodynamic and water quality modeling approaches

Figure 1 – Map of stations in Delta DWR-IEP water quality monitoring program.

Description of parameters

In this analysis, we focus on inorganic macronutrients—nitrate + nitrite (NO_{2+3}^-), ammonium (NH_4^+), phosphate (PO_4^{3-}), and silicate—as they are the primary drivers of phytoplankton productivity. We also examine chlorophyll *a* (chl-*a*) and dissolved oxygen (O_2) concentration time-series in order to more directly probe eutrophication and its effects on the Delta ecosystem. Temperature and conductivity are examined in order to illustrate physical drivers of variability.

Per-site deseasonalization

In order to inspect longer trends more closely and perform normalized site-to-site comparisons, we remove the seasonal (*i.e.*, monthly) cycle on a per-site basis. The seasonal cycle for a given parameter *P* at a given station, P_s, is calculated as the mean of all measurements of *P* in a given month, $\underline{P_{s,m}}$. Similarly, the standard deviation of all measurements of *P* in a given month is reported as $sd_{p_{s,m}}$. We remove the seasonal mean and standard deviation by subtracting the mean then dividing by the standard deviation, such that a normalized parameter, $P_{s,n}$, is defined as:

$P_{s,n} = \frac{P_s - \underline{P_{s,m}}}{sd_{P_{s,m}}}$. The normalized parameters can be thought of as local, monthly indices with a mean of zero and standard deviation of one across the entire time-series.

Factor analysis

In order to retrieve the dominant, latent factors driving each time-series, we next performed a matrix decomposition on the raw time-series using NMF. PCA and NMF are similar in concept in that they both decompose a matrix whose first dimension (e.g., rows of the matrix) represents time and second (columns) represents variables (one or more parameters at one or more locations) into components (both temporal and spatial) that can be linearly recombined to retrieve the original matrix. They differ in NMF's major constraint of non-negativity on both input data and resulting components. This constraint implies that in the linear recombination, PCA's components can be added and subtracted whereas NMF's can be added (*i.e.*, superimposed) only. Data were input to the NMF calculator after subtracting the minimum and then dividing by the standard deviation of each time-series. While factor analysis does lend itself to simultaneous analysis of multiple parameters, we chose to treat parameters separately in order to find common and distinct drivers across sites on a per-parameter basis. Station D19 is dropped from the factor analysis due to large data gaps which artificially skew the results.

The number of NMF modes can range from one to the number of input parameters. We attempt to reconstruct the original time-series using the NMF model by superimposing modes: that is, we multiply the weight vector (which represents the strength of a given mode across stations) by the time-series expression of that mode in order to get the weighted time-series at each station. These reconstructed time-series (one time-series per mode per station per parameter) are still expressed in their minimum/standard deviation-removed transformations. We superimpose all modes for a given parameter at a given station by multiplying the modes by that original time-series' standard deviation and adding a baseline of that time-series' minimum value. The time-series reconstructed through this linear recombination is referred to as the NMF model. The relative contribution of each mode to the full reconstruction is calculated on a per-site basis as the spatial weight of that mode at that site times the average of the time-series of that mode divided by the sum of those contributions from each mode. We assess the NMF model's ability to match observed data by examining the NMF model vs. observed data plot's R^2 and coefficient of variation of the root mean square error: $CV(RMSE) = \frac{RMSE}{\underline{P}}$ where $\underline{P}$ is the average of a given time-series.

All analysis was performed using Python 2.7 using scikit-learn, pandas, and numpy toolboxes.

Results

We first examine the "raw" time-series of dissolved inorganic nutrients, dissolved oxygen, and chlorophyll-*a* concentrations at our eleven stations (Figure 2). There are several salient features in this observational record. First, the time-series of each parameter tends to be dominated by the behavior at one or two sites. Stations C10, P8 and MD10, the southeastern-most stations, have striking features that obscure the other stations.

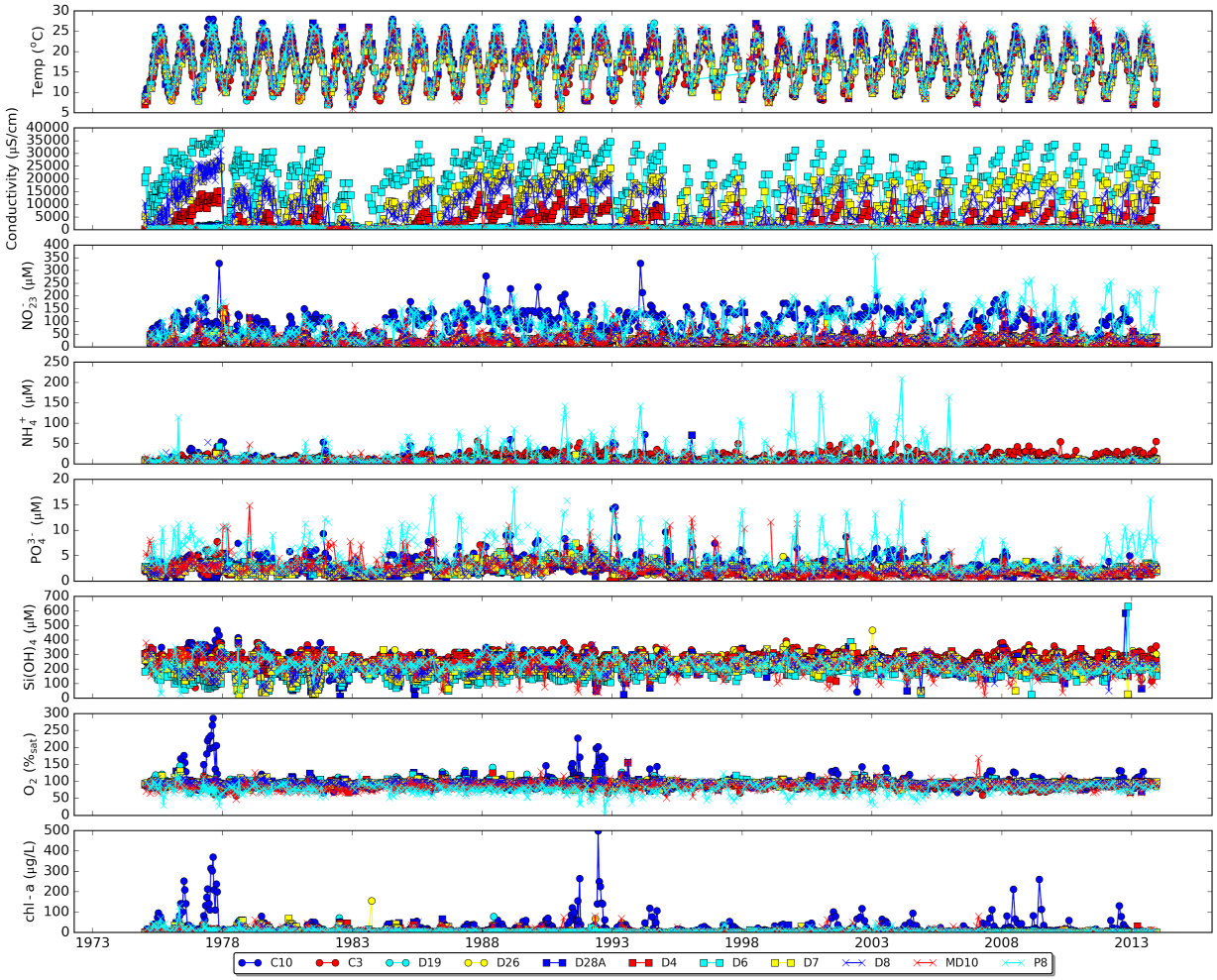

Figure 2 – Time-series plots of temperature, conductivity, NO_{2+3}^-, NH_4^+, PO_4^{3-}, $Si(OH)_4$, O_2, and chl-a at 11 active Delta stations.

We next plot the same parameters at eight stations, removing the dominant C10, P8, and MD10 time-series and zooming in on the most recent decade, in order to examine the other patterns more closely.

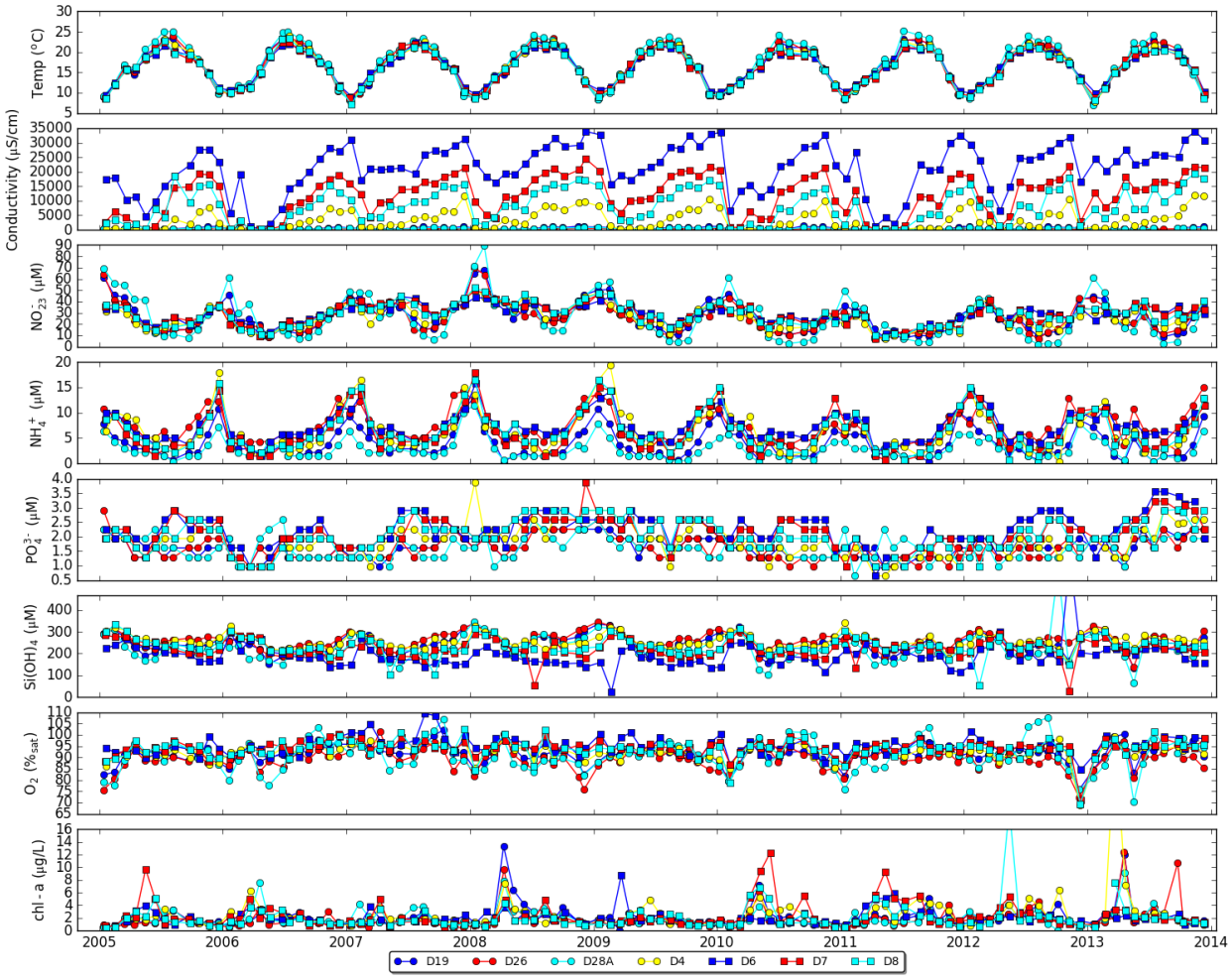

Figure 3 – Same as Figure 2, zoomed in on both time (2003–2013) and y-axes and removing C10, P8, MD10, and C3 for closer visual inspection.

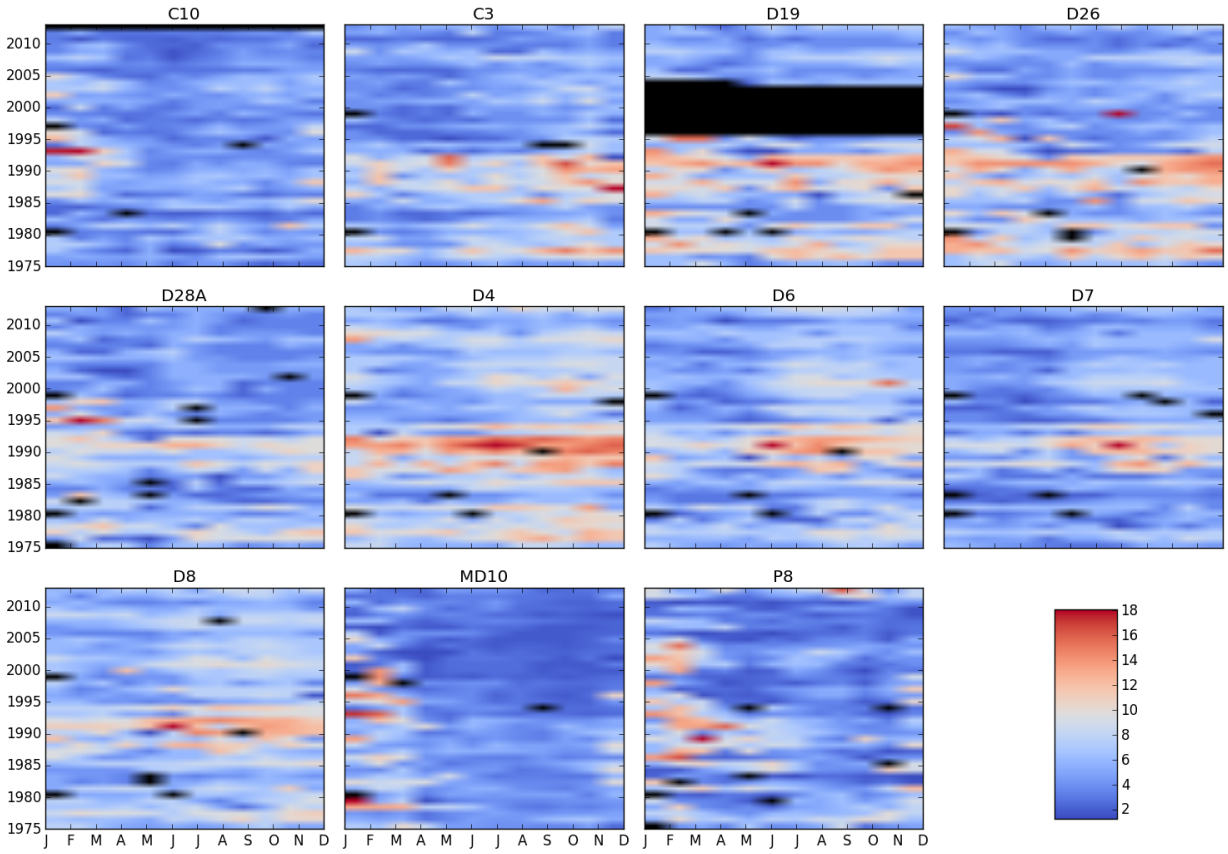

Figure 4 – Untransformed PO_4^{3-} (µM) with respect to year and month. Black denotes gaps in data record.

Note the banding in both horizontal and vertical dimensions in Figure 4 corresponding to seasonal trends (apparent as vertical bands) and longer-term (multi-year) trends (apparent as horizontal bands). A dominant feature in the data for all parameters in Figure 2 and Figure 4 is the seasonal cycle.

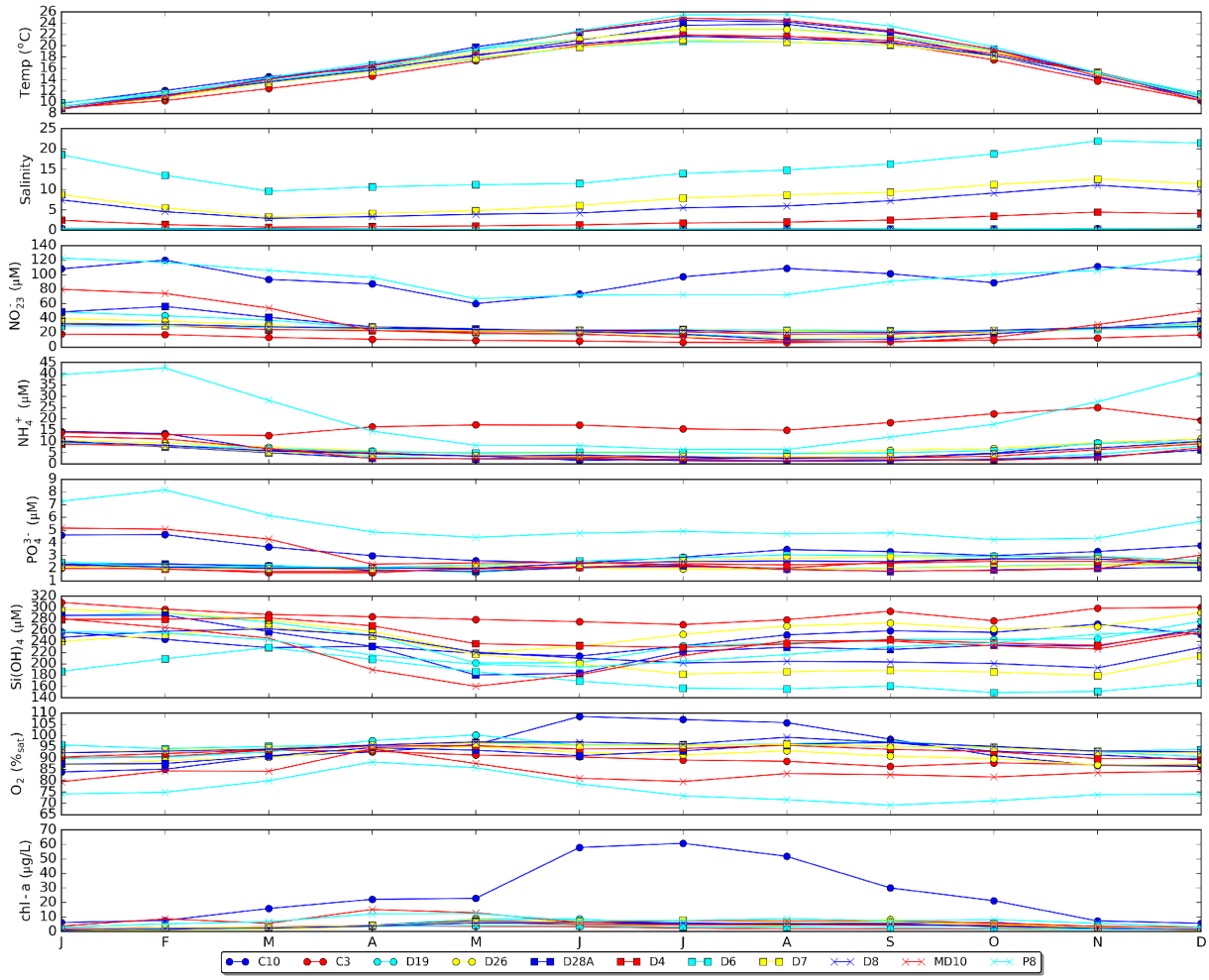

Figure 5 – Seasonal mean climatologies of temperature, conductivity, NO_{2+3}^-, NH_4^+, PO_4^{3-}, $Si(OH)_4$, O_2, and chl-a at 11 active stations with stations divided into two groups to show the radically different behavior at stations C10, C3, MD10, and P8 compared to the others which exhibit visually similar patterns in seasonal trends.

DIN, DIP, O2, and chl-*a* (Figure 5) show moderate summer–winter seasonality, again, largely obscured by the dominant C10, P8, MD10, and C3 patterns.

We next normalize the time-series data, $P_{s,m}$, by subtracting the mean and dividing by the standard deviation such that all normalized monthly parameters have a mean of zero and standard deviation of one. The purpose of performing this normalization is twofold: first, it allows us to inspect the seasonal mean and standard deviation across all parameters at all sites and, second, it allows subsequent analysis to weigh all variables equally.

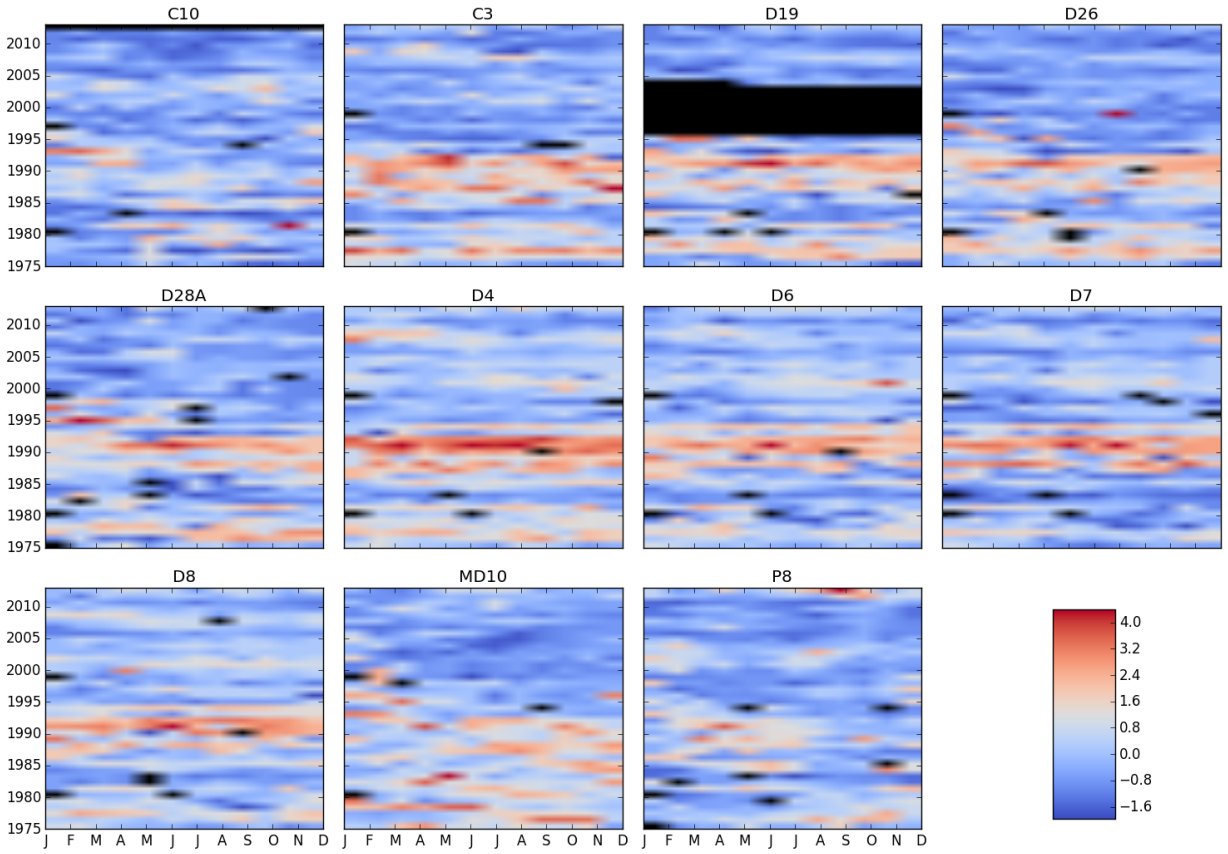

Figure 6 – Deseasonalized/normalized PO_4^{3-} with respect to year and month. Colorbar represents the unitless PO_4^{3-} patterns. Black represents absence of data.

Following this normalization, we see that all banding (Figure 7) appears in the horizontal direction, corresponding to annual trends. The strongest (most positive) anomalies tend to occur prior to 1995, with the five years from 1986–1991 showing the most distinct increase across all stations.

We next zoom in on the most recent twenty years in order to observe greater detail.

Once we normalize all variables by their monthly mean and standard deviation (Figure 7), we are able to examine the parameters more closely for multi-year trends. Taking the annual means of each of the deseasonalized variables (or, equivalently, taking the average of horizontal bands in Figure 7–Figure 14) and applying a five-year rolling mean in order to smooth the data, we begin to see longer-term trends, indicative of anthropogenic forcing.

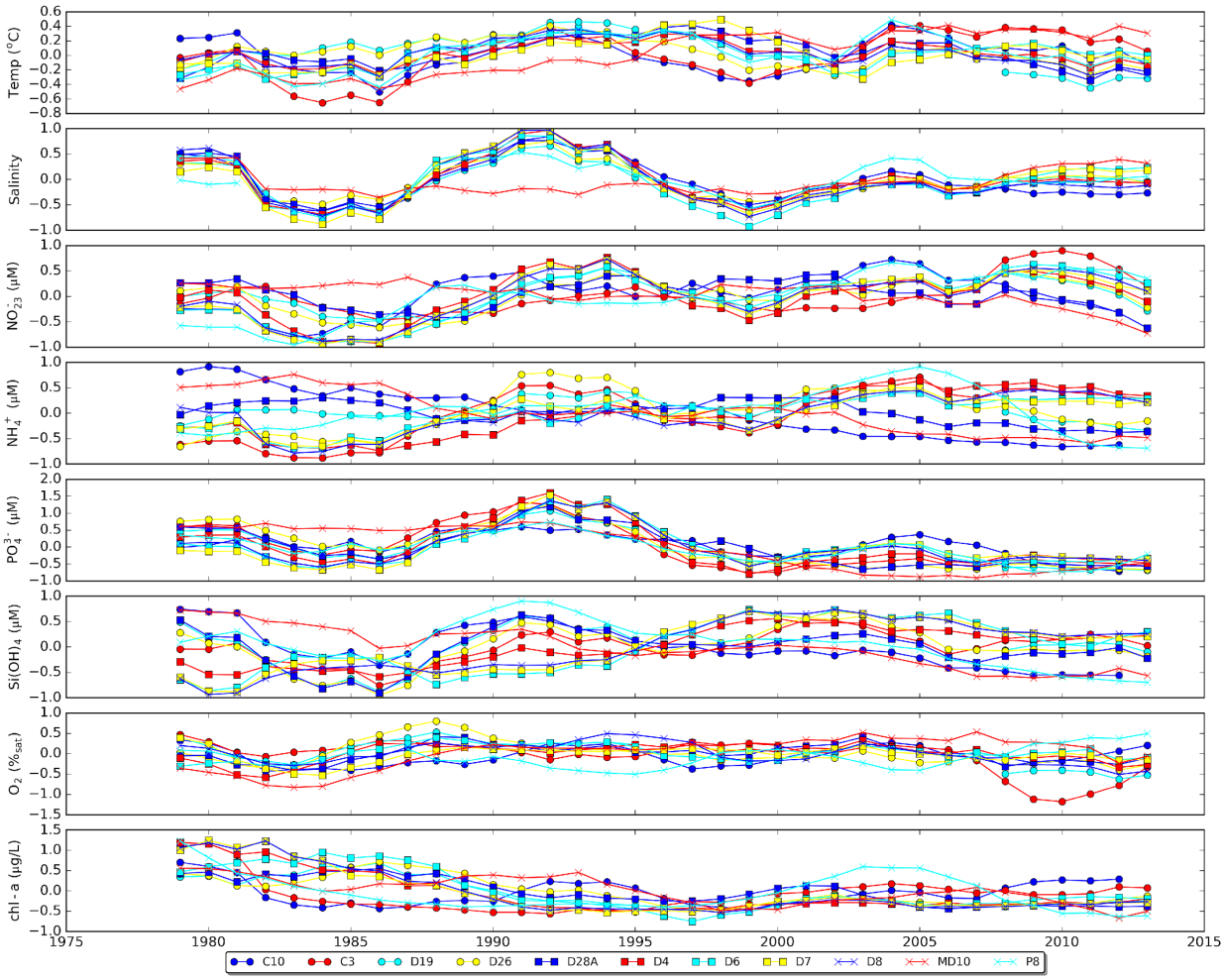

Figure 7 – Yearly means of all deseasonalized/normalized parameters at all stations. Again, signals are largely dominated by several stations, obscuring some commonalities among interannual trends; however, it is readily seen that many parameters behave similarly across the region.

Figure 8 shows water quality trends over the full dataset. Note again that this represents the mean and standard deviation-removed data; in order to retrieve the site and month-specific magnitude and variability, one must multiply these results by the seasonal mean and standard deviation matrices represented by Figure 5 and Figure 6.

Non-negative matrix factorization

We next turn to the NMF analysis in order to attempt to extract additional information not as readily apparent in the raw or climatological time-series. As described in the Methods, the number of modes can range from one to the number of input variables/stations. Visual inspection using a range of modes tested here suggested that four modes successfully captured sufficient variability across all variables; at the end of this section, we will describe additional quantitative assessments of the NMF model using the full range of modes.

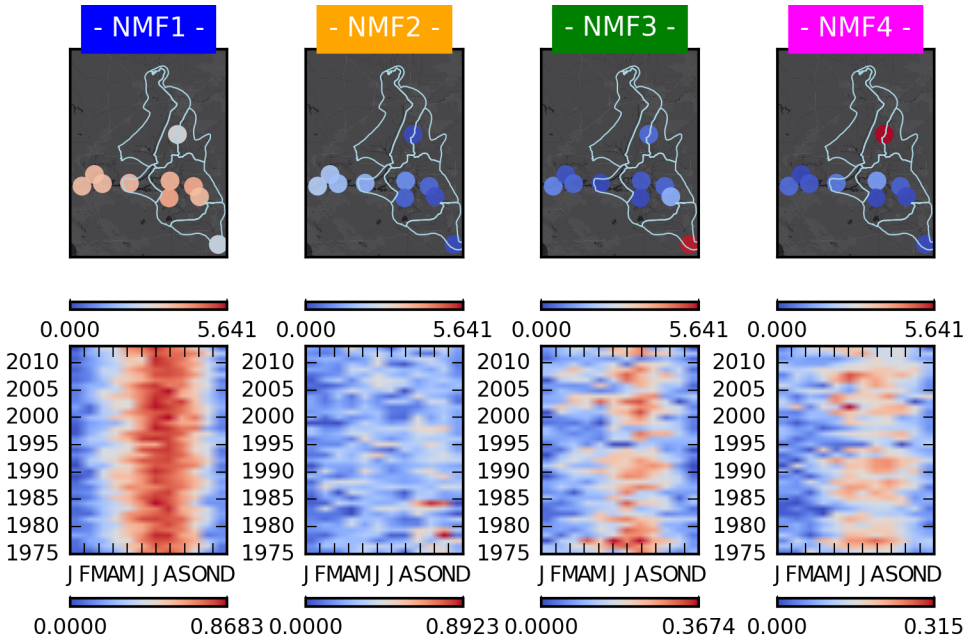

Figure 8 – An illustration of the capabilities of NMF using temperature as a test case. Columns represent the different modes; four modes were used in this analysis. The top row shows the spatial weight vector for each mode where the color of the dot corresponds to the local weight of that column's mode. The bottom row illustrates the time-series vector of each mode plotted with magnitude as a function of year and month; magnitude ranges from 0 (blue) to that mode's maximum value (red).

First examining temperature, we note that the NMF analysis behaves as would be expected from the time-series plots (Figure 2). As can be seen in Figure 8, mode 1 represents the strong seasonality which is both intuitive and evident in all temperature records across the region. Less obvious in the full time-series, however, but extracted in modes 2–4 here are additional features which represent both event-scale (locally high patterns evident as red blotches in mode 2) and interannual cycles (evident as larger red blotches with some horizontal banding in modes 3–4). It is notable that NMF 3 is expressed solely at Station C10 (the only red dot in the NMF 3 regional plot) while NMF 4 is expressed solely at Station C3. In other words, while the time-series may line up visually, they are mathematically distinct in the NMF analysis and these additional features (deeper reds in June–July of NMF 3 and slightly more spread out peaks in NMF 4) suggest important subregional differences, even in temperature.

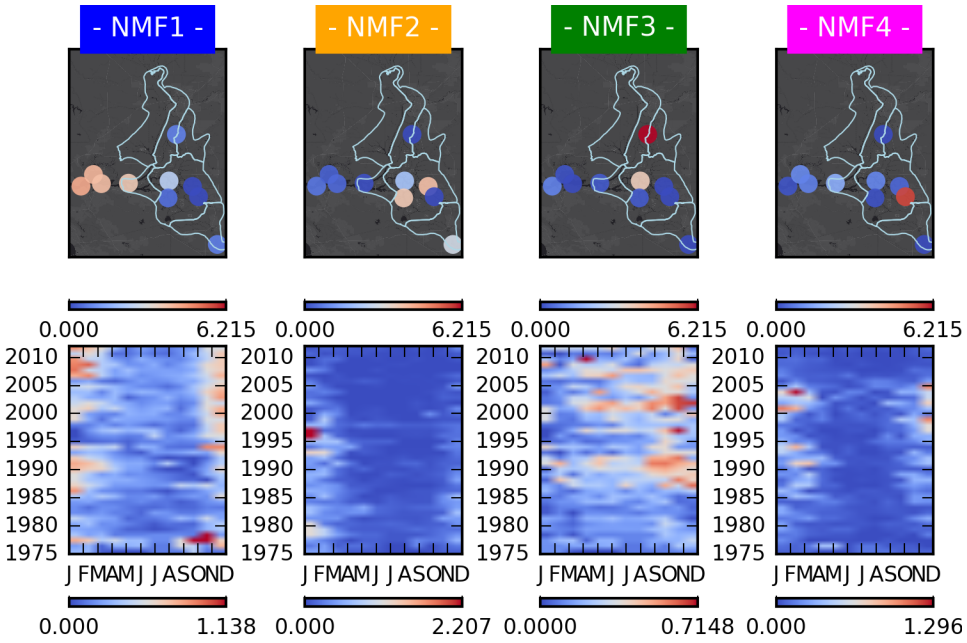

Figure 9 – NMF modes representing the dominant factors in NH_4^+ variability. The colored title labels correspond to the reconstructed time-series plots below.

The NMF analysis of ammonium (Figure 9) shows some additional features beyond the seasonality evident in the time-series plots. First, the grouping of the weights across stations (top row) suggests that, while the Suisun Bay subregion stations behave similarly, stations in the Central Delta subregion tend to be more heterogeneous (evident in the heterogeneous colors of the dots in the top row of NMFs 2–4). Also noteworthy is that, while seasonality appears to be a primary driver of variability (Figure 2, Figure 3), that seasonality is extracted as mathematically unique across different subregions, and even among different stations within a given subregion, in the NMF analysis (Figure 9). Modes 1, 2, and 4 all have a notable seasonal component (as evident in the vertical banding in the bottom row of Figure 9); however, there is significant interannual variability in this seasonality that also varies across stations.

NMF plots such as those shown in Figure 8 and Figure 9 are shown in the Appendix for the remaining parameters (conductivity, nitrite+nitrate, phosphate, silicate, dissolved oxygen, and chlorophyll).

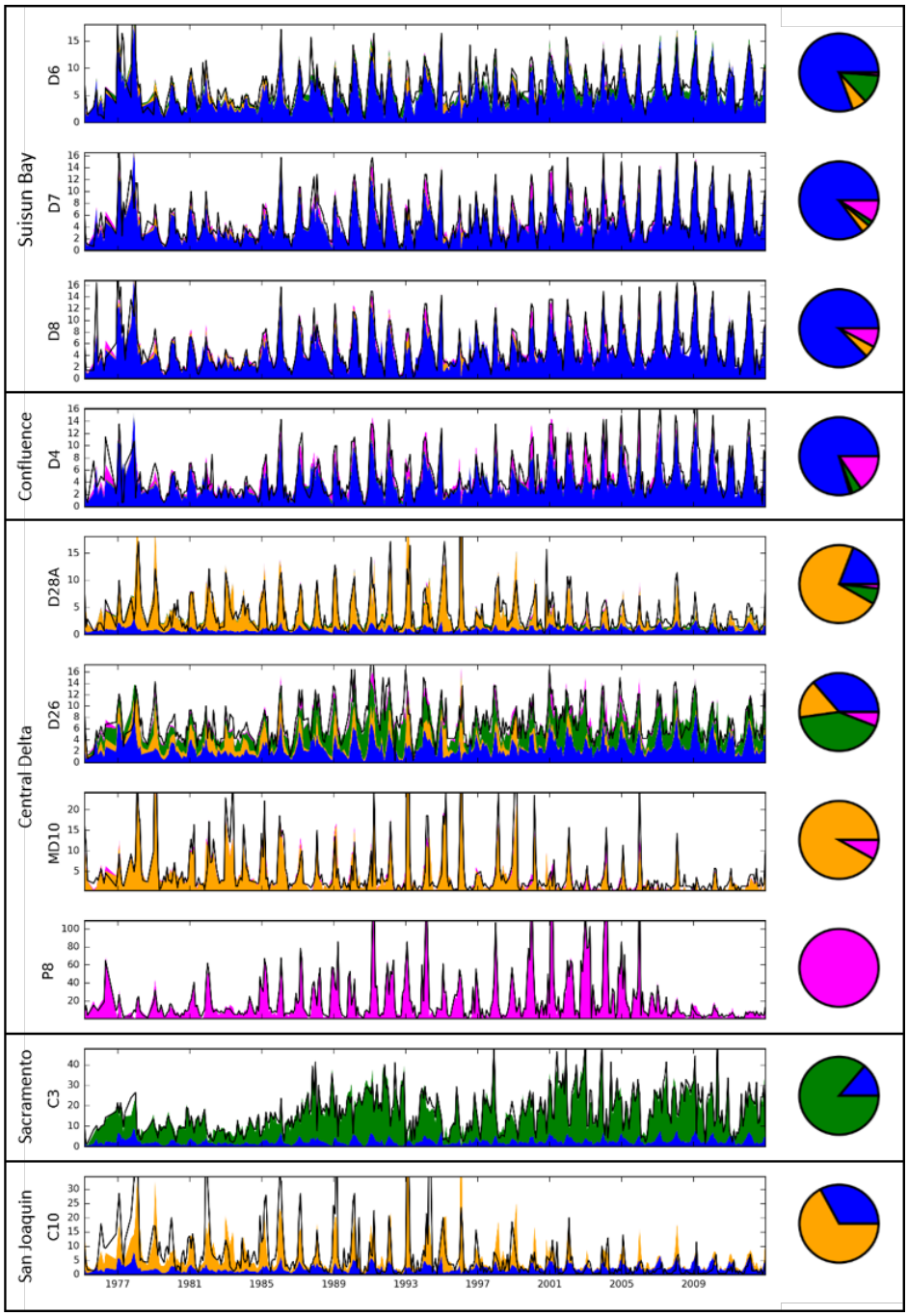

Figure 10. Time-series (left) and percent contribution of modes (right) for ammonium (in units of micromolar). The time-series are reconstructed by superimposing the four factors. Colors represent contributions from the NMF modes—1-blue, 2-orange, 3-green, 4-magenta—and the black line represents observations. Please see the Appendix to Task 2 for additional details on superimposition of NMF modes and additional figures.

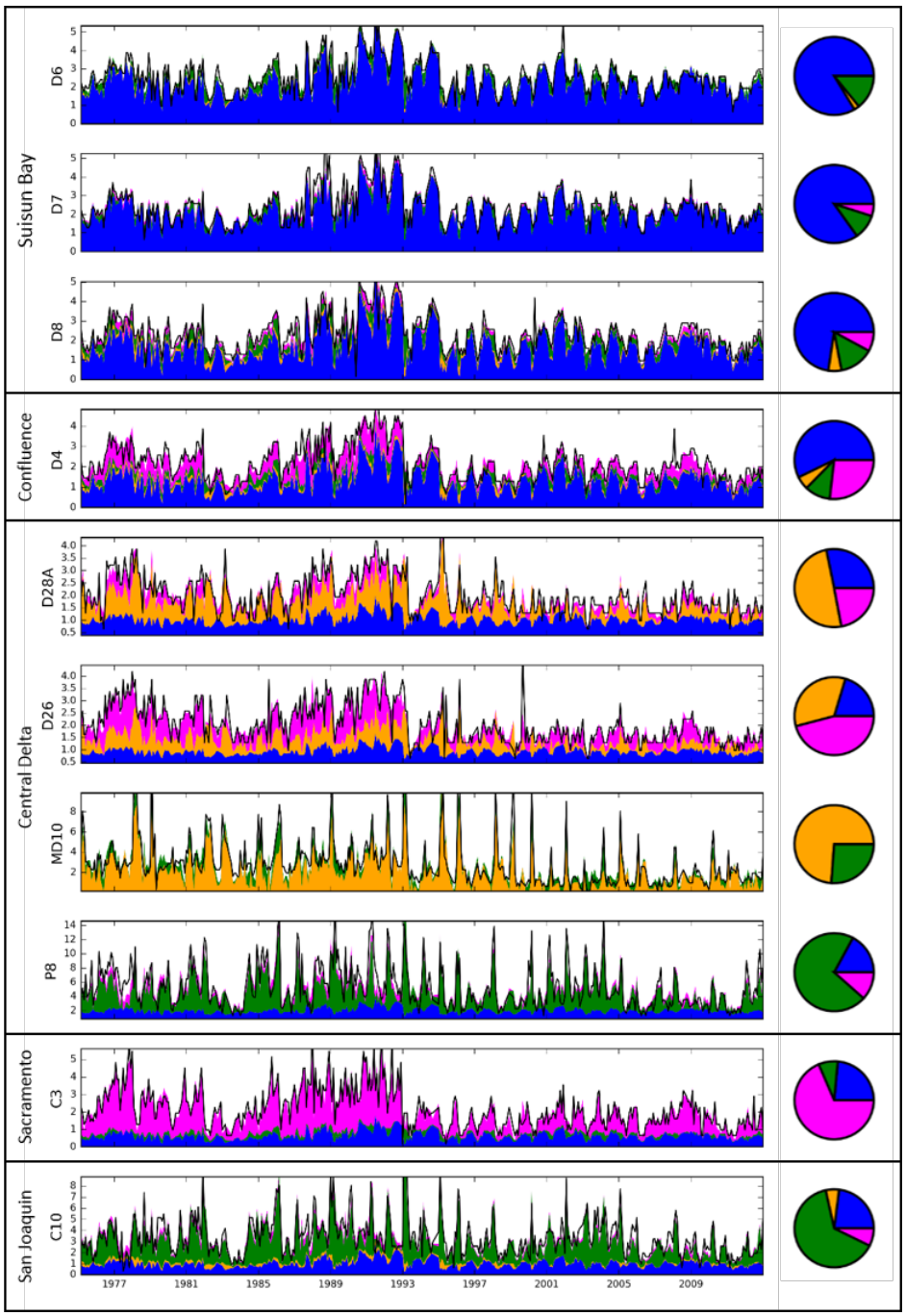

Figure 11. Time-series (left) and percent contribution of modes (right) for phosphate (in units of micromolar). The time-series are reconstructed by superimposing the four factors. Colors represent contributions from the NMF modes—1-blue, 2-orange, 3-green, 4-magenta—and the black line represents observations. Please see the Appendix to Task 2 for additional details on superimposition of NMF modes and additional figures.

An important trait of this factor analysis approach is that, since factors are all non-negative, they can be reconstructed through a straightforward superimposition. Continuing with the ammonium example, we attempt to reconstruct the original time-series for each station according to the procedure described in the Methods. This time-series reconstruction helps to emphasize several features from the NMF mode

plot (Figure 9). For example, it is immediately clear that, while the absolute magnitudes of ammonium concentrations vary across stations, similar drivers (*i.e.*, modes) are found in many of the stations.

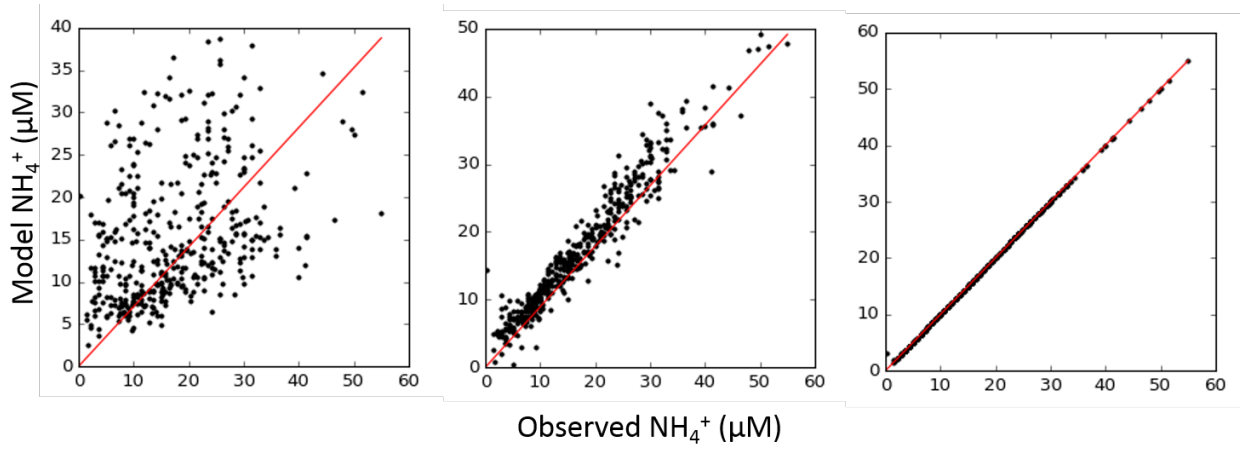

Figure 12 – Fit between NMF reconstructed (model) and observed ammonium at Station C3 using 1, 4, and 10 modes (left, center, right, respectively).

Table 1 – Statistics representing the goodness of fit for one to ten modes in the NMF analysis of each parameter.

1-CV(RMSE)/ R^2	Parameter							
Number of modes	Temperature	Conductivity	Nitrite+nitrate	Ammonium	Phosphate	Silicate	Oxygen	Chlorophyll a
1	0.947/ 0.967	0.639/ 0.590	0.611/ 0.540	0.299/ 0.476	0.651/ 0.491	0.830/ 0.387	0.920/ 0.389	-0.147/ 0.352
2	0.960/ 0.982	0.741/ 0.724	0.703/ 0.708	0.449/ 0.655	0.738/ 0.697	0.885/ 0.697	0.930/ 0.510	-0.003/ 0.499
3	0.967/ 0.987	0.789/ 0.804	0.758/ 0.812	0.521/ 0.739	0.781/ 0.793	0.898/ 0.767	0.940/ 0.620	0.157/ 0.616
4	0.974/ 0.991	0.821/ 0.869	0.800/ 0.865	0.619/ 0.810	0.816/ 0.851	0.912/ 0.831	0.957/ 0.707	0.265/ 0.701
5	0.977/ 0.993	0.854/ 0.930	0.830/ 0.903	0.719/ 0.873	0.858/ 0.891	0.925/ 0.874	0.968/ 0.784	0.374/ 0.773
6	0.981/ 0.995	0.890/ 0.962	0.877/ 0.938	0.761/ 0.912	0.902/ 0.925	0.938/ 0.910	0.971/ 0.848	0.500/ 0.839
7	0.985/ 0.997	0.926/ 0.979	0.913/ 0.961	0.860/ 0.947	0.920/ 0.946	0.958/ 0.945	0.978/ 0.902	0.648/ 0.895
8	0.989/ 0.998	0.949/ 0.989	0.938/ 0.977	0.922/ 0.974	0.943/ 0.967	0.968/ 0.966	0.988/ 0.950	0.795/ 0.955
9	0.993/ 0.999	0.964/ 0.995	0.966/ 0.991	0.961/ 0.990	0.966/ 0.985	0.979/ 0.985	0.992/ 0.979	0.915/ 0.989
10	0.998/ 1.000	0.987/ 0.999	0.987/ 0.999	0.990/ 1.000	0.993/ 1.000	0.986/ 0.995	0.995/ 0.997	0.997/ 1.000

It is clear from the results presented in Table 1 that while some parameters are effectively modeled with only a few modes representing variability across the entire region (*e.g.*, temperature), most parameters an only be well characterized using more modes. We chose four modes for the presentation of the results shown here as that allows for $R^2 > 0.8$ across the nutrient parameters which we were most concerned with characterizing in the analysis. Chlorophyll variability requires six modes in order to fit the reconstruction at $R^2 > 0.8$; this finding reinforces our understanding that the regional biology is particularly heterogeneous in comparison to physical and chemical parameters which can be described with fewer modes.

Attribution of NMF modes

The remaining NMF figures are shown in the Appendix; here we synthesize the major findings.

Table 2 – Attribution of parameter–mode combinations to known and hypothesized physical/chemical/biological drivers. Green represents best understood drivers, not necessarily most important.

		Nitrite+nitrate	Ammonium	Phosphate	Chlorophyll
NMF Mode #	1	Mild interannual cycling, largely open bay effect	Seasonal freshwater flow/dilution effects	Freshwater flow/dilution effects	Clams move into Suisun
	2	Seasonal freshwater flow/dilution effects	Spike in 1996-97	Seasonal cycle changes due to freshwater/ loading	(sub)regional similarity among eastern Central Delta and Sacramento River
	3	Nitrification + something happening in upper watershed, flow effects (inversely related to freshwater input)	Highly localized effect at C3/D26, Sacramento River/watershed effect, this should in theory look very similar to NMF 4 because the known drivers are changing in a similar way. The fact that it doesn't suggests that there are other processes at play here.	Seasonal cycle changes due to freshwater/ loading	(sub)regional similarity among western Central Delta stations - could this be clam effect as well?
	4	Spike in Feb 2006-07	Local wastewater treatment, nitrification	Phase-out of phosphates in detergents, flow	Independent trend (summertime blooms vary in magnitude on interannual basis) in San Joaquin

Conclusions

We examined the time-series of dissolved inorganic macronutrients, dissolved oxygen, and chlorophyll-a concentration at eleven stations across the Sacramento–San Joaquin River Delta over the past four decades with a focus on elucidating important patterns in both space and time.

Among the many well-known drivers of biogeochemical variability in the Delta are several hidden or latent drivers. Their variability in time and space is obscured by other, perhaps larger or more readily apparent, drivers. Non-negative matrix factorization is shown to be capable of extracting these latent drivers and determining their relative importance in biogeochemical variability—a necessary step toward fully characterizing this heterogeneous ecosystem.

Several of the most striking features of this dataset included the following. 1) We observed anomalous behavior at multiple sites in comparison with the others, especially at stations C10, P8, and MD10. 2) There is strong seasonality in most parameters, a pattern extracted in the climatological analysis and also seen in many individual NMF modes. 3) Interannual variability can be observed in both deseasonalized/normalized annual trend plots (Figure 7) and frequently in individual NMF modes. These interannual trends can be attributed to natural cycles (*e.g.*, El Niño/La Niña) and management actions (*e.g.*, phase-out of phosphates in detergent, changes to nitrification of wastewater). And 4) additional latent drivers that are hard to detect through other means can be found in the dataset utilizing the NMF approach described here. NMF allows a careful assessment on a common footing of variability among parameters with very different statistics.

We also wish to emphasize that while NMF analysis was applied to this already well characterized ecosystem and was used to illustrate several known changes in biogeochemistry, it is clearly suited for a similar type of analysis in a less studied ecosystem. The NMF methodology does not inherently include spatial proximity (*i.e.*, it doesn't "know" subregions *a priori*) but proved capable of demonstrating similarity and heterogeneity within and across subregions. While this analysis was performed on data with a long time dimension and a shorter spatial dimension, it could be similarly applied to data with different (or greater) dimensionality (including any of: depth, greater resolution in latitude, and/or longitude).

Recommendations

Furthermore, while the NMF analysis suggests that the Suisun Bay stations behave fairly similarly, there is significant heterogeneity across the remainder of the region, both in comparison to Suisun Bay and in comparison to each other station. This feature is evident in the scatterplot maps of the NMF plots (*e.g.*, Figure 9) where, even for a given mode within a given subregion, the dots (*i.e.*, weights of the NMF modes) are different colors. We demonstrate that the most variability is observed in the Central Delta subregion, where NMF modes are frequently weighted differently, sometimes covering the full range of variability in a given mode just within that subregion (*e.g.*, Figure 9, NMF 4). We are unable to characterize variability in the Northwest (Cache Slough and Deep Water Shipping Channel) and NorthEast (Mokelumne and Cosumnes Rivers) subregions as there are no stations within those. We recommend that each subregion have at least two stations in order to characterize heterogeneity both within and among subregions. The Central Delta has proven to be particularly heterogeneous through the NMF analysis and we therefore recommend that at least four time-series stations be maintained there. Stations D7 and D8 behave most similarly across all parameters throughout the NMF analysis; if

any station must be moved to accommodate the recommendations above, we suggest that one of two be relocated as the biogeochemical information collected there appears to be largely redundant.

Works Cited

Novick, E. et al., 2015. Characterizing and quantifying nutrient sources, sinks and transformations in the Delta: synthesis, modeling, and recommendations for monitoring, San Francisco Estuary Institute, Richmond, CA.

Supplementary Data

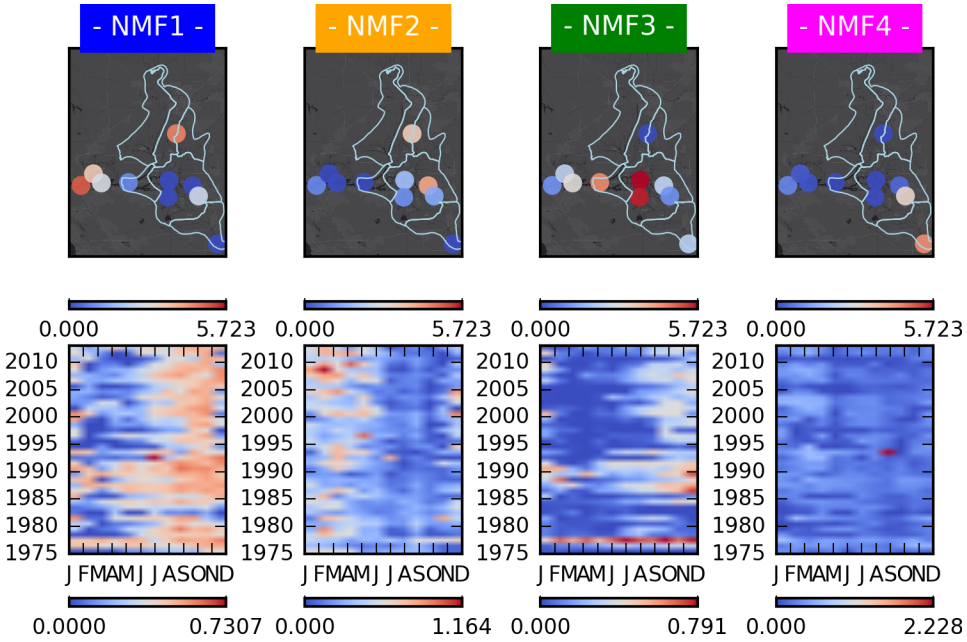

Figure 13 – NMF analysis of conductivity.

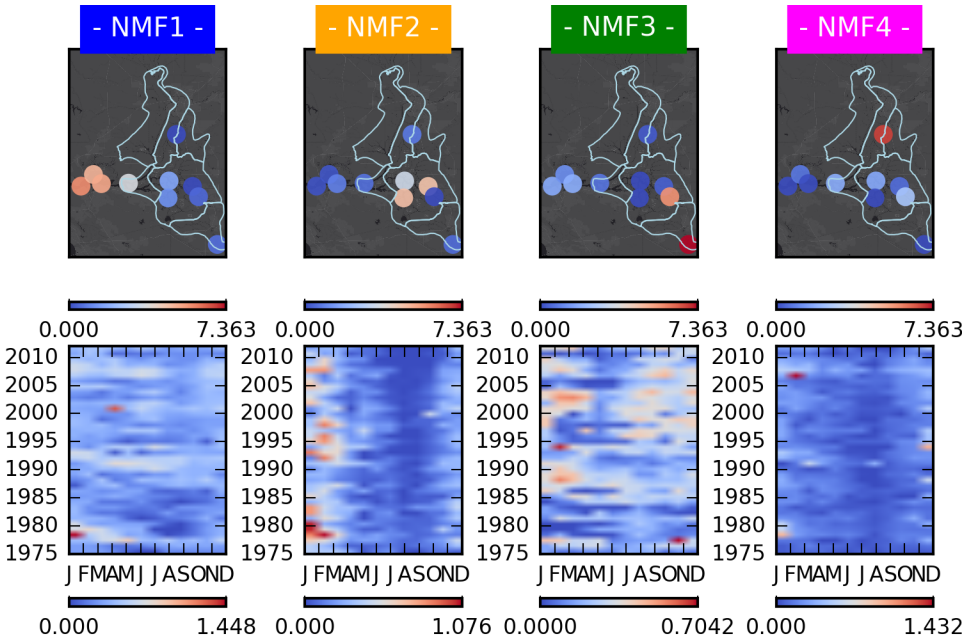

Figure 14 – NMF analysis of nitrite+nitrate.

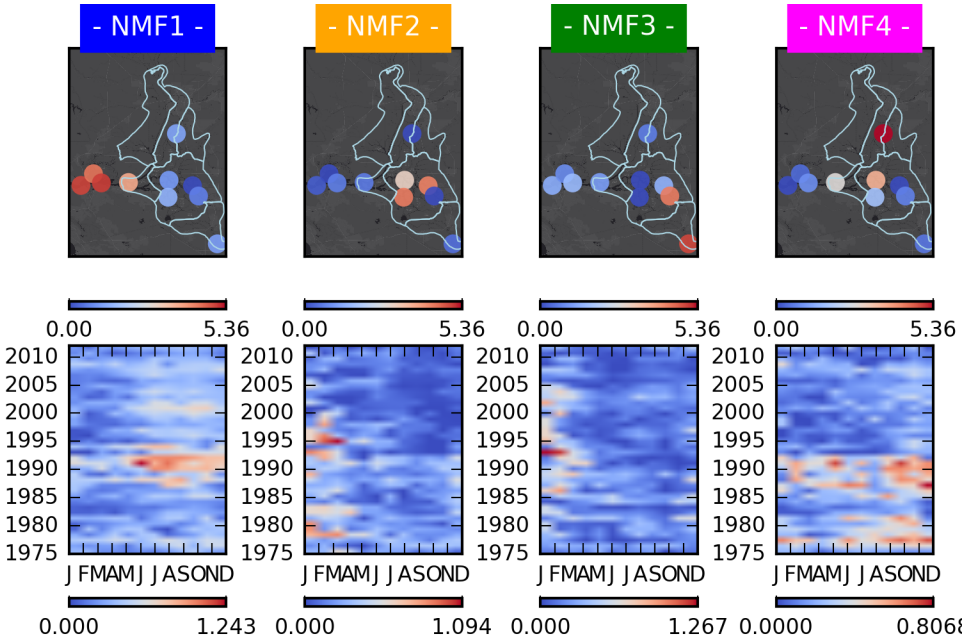

Figure 15 – NMF analysis of phosphate.

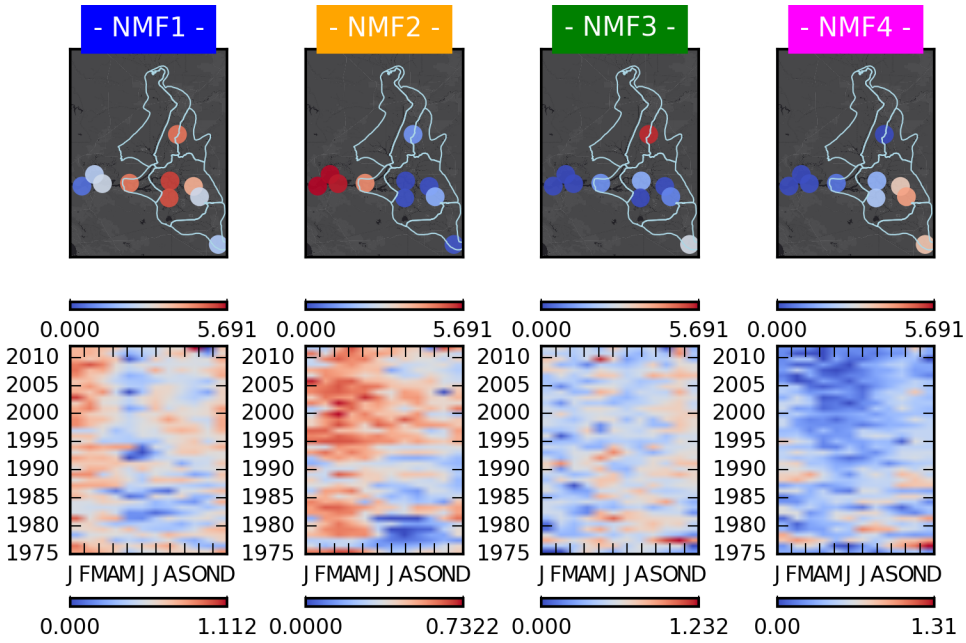

Figure 16 – NMF analysis of silicate.

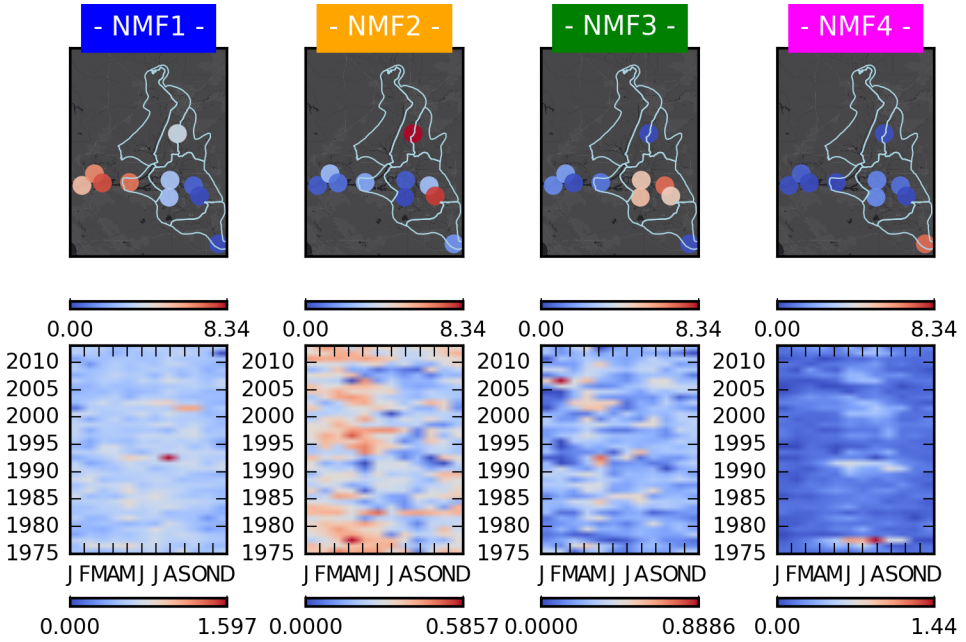

Figure 17 – NMF analysis of dissolved oxygen (% saturation).

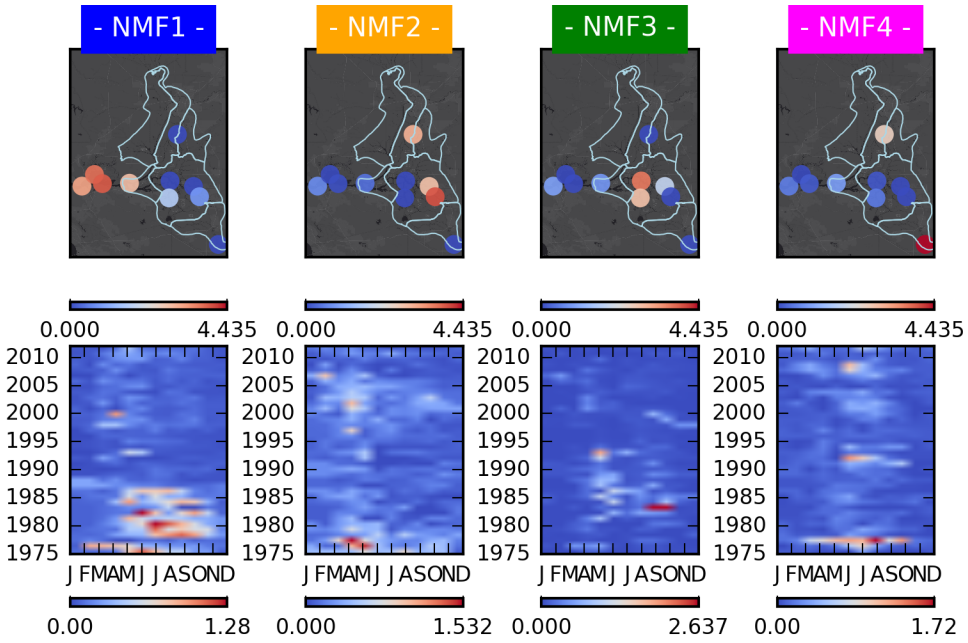

Figure 18 – NMF analysis of chlorophyll a concentration.