Project Deadline 2: Preliminary Results
Group 4: Sangeet Dhandwar, Jason Duong, Eric Hsieh, Youngsub Lee

1. **Revised Title** (0.5 pts): revised title
   - ➢ Pearson's Correlation between the ~~increased~~ sea surface water temperature and water pH level in Hawaii and Bermuda.

2. **Revised research questions** (1 pt.)
   - Make sure your questions are succinctly worded for statistical analysis (use statistical terminology) and are answerable based on the analysis you plan to perform.
   - In general, your project should have a minimum of 3 specific questions that are connected to specific analyses.
   - You can pose your questions as research questions that can be definitely answered through your analysis (essentially yes or no questions) or you can word them as hypothesis questions using H0 and H1.

   - ➢ Does the relationship between the sea surface water temperature and the water pH level have a significant correlation over the years for Hawaii and Bermuda?

   - ➢ Is the annual difference in water temperature and the rate of change in water pH level increasing, decreasing, or remaining constant? (Hawaii vs Bermuda)

   - ➢ Is the rate of change in water temperature and rate of change in water pH level the same as observed in two different locations? (Hawaii vs Bermuda)

   - ➢ Is the rate of change in difference between the water temperature and water pH faster, slower, or the same at two different locations? (Hawaii vs Bermuda)

3. **Summary** of analysis done to date. This should include (1.25 pts):
   - Any SPECIAL steps needed to import and/or manipulate the data into Matlab (**no need to include any coding, just summarize what your group had to do to achieve your goal**).
     - You do not need to include simple steps you have taken (downloading data, convervios from an xls file etc.).

- ➢ Read and save a copy of the files first before manipulating it. We have three sets of data: Bermuda surface water temperature, Hawaii surface water temperature, and pH levels for Bermuda and Hawaii. Before importing these sets of data into Matlab, we want to remove the header row and any descriptions at the top of the file.
- ➢ Ocean-acidity_fig-1.csv:
  Remove columns 4 to 8 because we are not using the data for Canary Islands.
- ➢ Hawaii_sst.csv:
  Remove columns 2 and 3 because we don't need the latitude and longitude information.
- ➢ Bermuda_SST_1950_To_2021.xls
  Remove columns 2 to 6 because those are unimportant informations for the analysis.
- ➢ How you dealt with missing or outlier data.
  For missing data points, fill in the blank of NULL space with NaN in Matlab. If there are any outlier data, replace the data point with NaN.
- ➢ What your group has done to understand the distribution of the data and plotting of the data.
  As stated above, we graphically analyzed the data to try to observe a trend. We need to prove the trend is statistically significant as correlation does not always mean causation. Additionally we used plots (below) to see various trends among the data sets and still need to statistically analyze them.
- ➢ A brief summary of what your group plans to do to finish the project.
  To complete this project we must analyze the data with regard to the chlorophyll data of Hawaii and Bermuda. This will help us understand if the phytoplankton activity affects carbon dioxide and ultimately pH. Additionally we must calculate the pearson correlation coefficient to see if we can have any significant findings.

4. **Data analysis** (1.25 pts):
   - Give a justification of each step your group has taken in their analysis of the data.

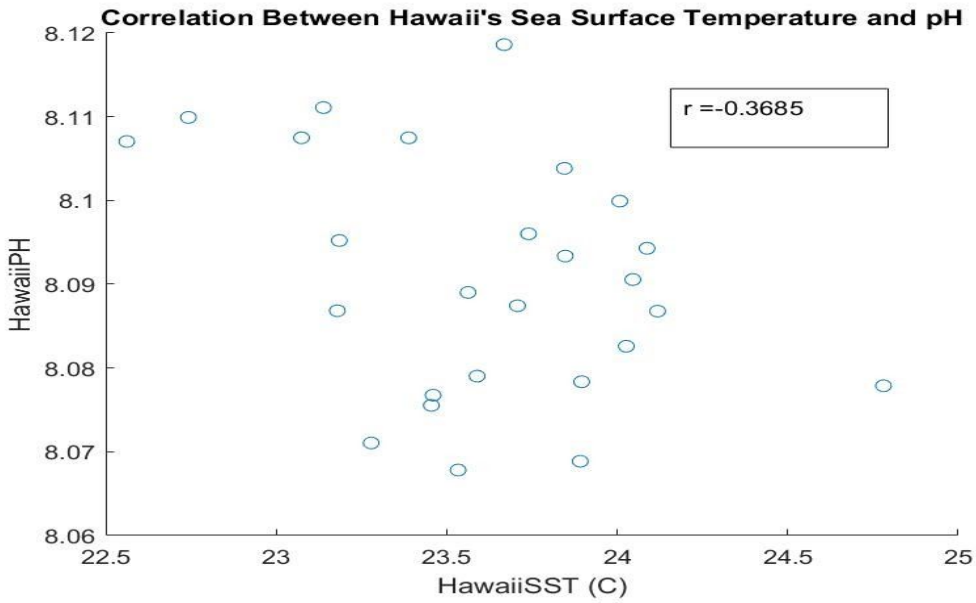
- ➢ Each step taken thus far in our analysis is to narrow down our raw data in order to be able to perform a statistical analysis between Surface Water Temperature, pH Levels, and eventually CO_2 concentration and chlorophyll concentration.
- ➢ To focus our analysis, we removed a lot of data that is not needed from the RAW data that we've found. This allows us to succinctly look at our data to perform any analysis.

- We additionally may restrict the range of data in order to find a statistical correlation between the datasets. For instance, some data has been tracked much longer whilst some data has not been tracked that long. We restrict the range to the data that has not been tracked as long.
- We additionally need to ensure the sample size between any two datasets are the same before performing any analysis.
- For Hawaii plot: We plotted Hawaii's annual average SST vs pH.
  The r-value (-0.3685) indicates that they have a negative correlation
- Warmer Hawaii SST is correlated with lower pH values.

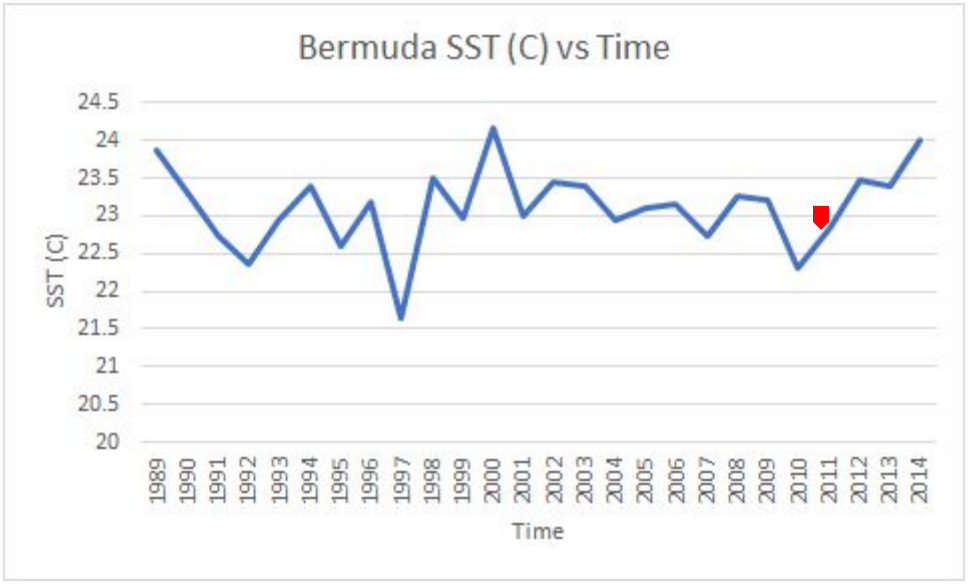
Hawaii pH	Hawaii SST	Bermuda pH	Bermuda SST	Year		HawaiipH	HawaiiSST	BermudapH	BermudaSST
8.10748	23.38833	8.122	23.86137	1989	HawaiipH	1	-0.3685	0.8198	-0.2109
8.11861	23.66833	8.106583	23.29699	1990	HawaiiSST	-0.3685	1	-0.4278	-0.2906
8.11109	23.13833	8.108	22.72658	1991	BermudapH	0.8198	-0.4278	1	-0.0168
8.10748	23.07417	8.102583	22.35128	1992	BermudaSST	-0.2109	-0.2906	-0.0168	1
8.109917	22.7425	8.107333	22.93825	1993					
8.107044	22.56167	8.102846	23.39581	1994					
8.103833	23.845	8.104571	22.59518	1995					
8.09602	23.74	8.101133	23.18952	1996					
8.09992	24.0075	8.103533	21.63261	1997					
8.09429	24.0875	8.096143	23.5095	1998					
8.09522	23.185	8.092385	22.96683	1999					
8.086836	23.17917	8.098308	24.15918	2000					
8.087433	23.7075	8.097357	22.99375	2001					
8.08902	23.5625	8.094583	23.45604	2002					
8.093364	23.8475	8.083154	23.40637	2003					
8.090558	24.04583	8.0896	22.93983	2004					
8.0826	24.02667	8.088625	23.10824	2005					
8.08679	24.11833	8.071467	23.15534	2006					
8.076756	23.46	8.080533	22.73507	2007					
8.079025	23.58917	8.077357	23.26393	2008					
8.07554	23.455	8.069857	23.22055	2009					
8.077878	24.78083	8.073643	22.29671	2010					
8.07836	23.89583	8.084455	22.83068	2011					
8.06782	23.53333	8.078643	23.48115	2012					
8.06887	23.89167	8.086067	23.39233	2013					
8.07105	23.27833	8.0841	24.00877	2014					

- The figure above is the aggregation of our RAW data that we are going to be using. Any statistical analysis and creation of graphs will likely be done from this spreadsheet.

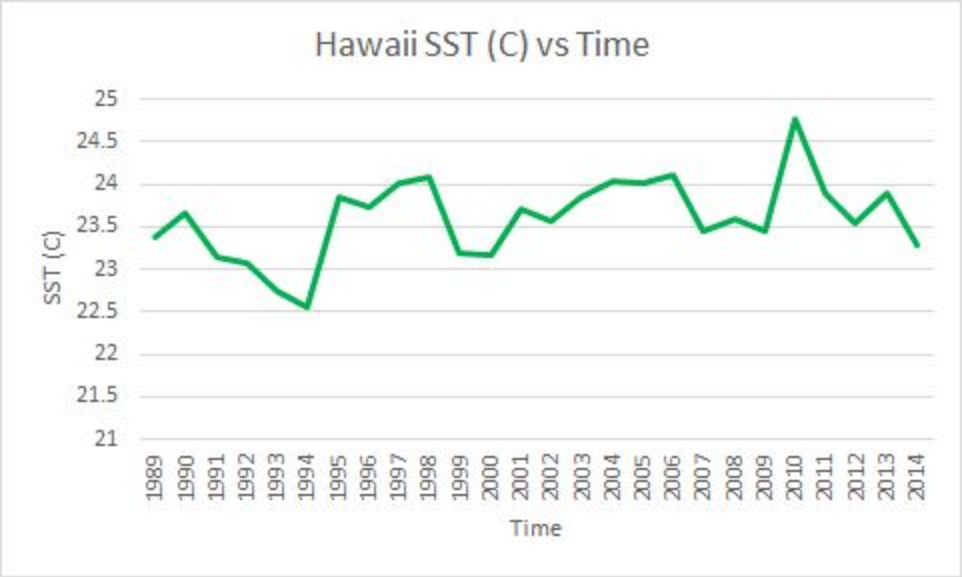
5. **Graphs** (1 pt.)



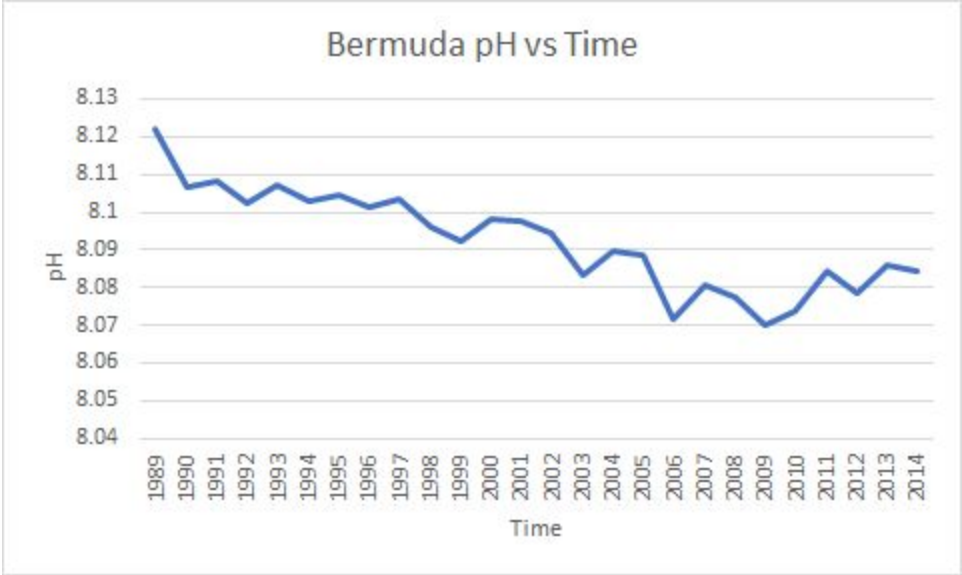
Pearson's correlation between Hawaiian pH and sea surface temperature



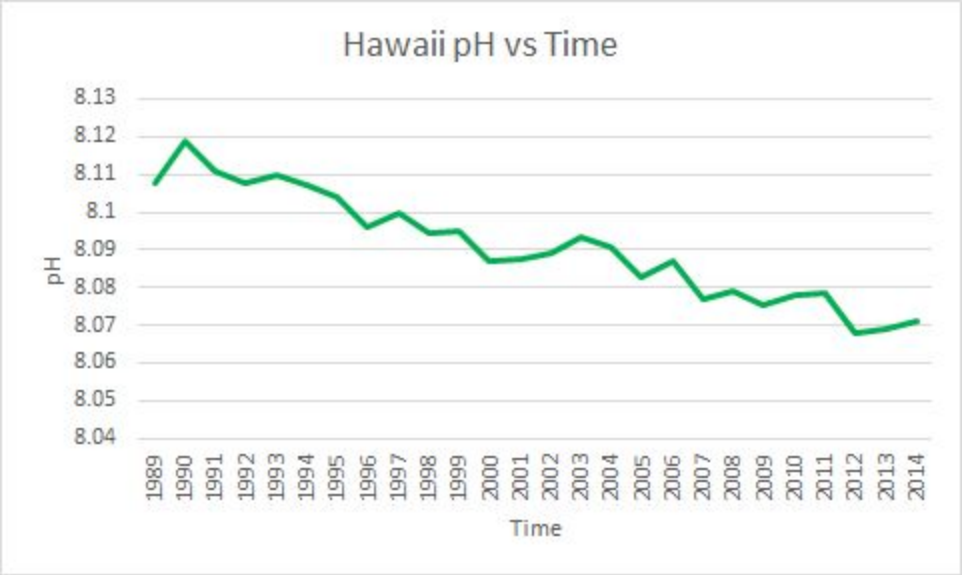
Bermudian sea surface temperature versus time from 1989 to 2014



Hawaiian sea surface temperature versus time from 1989 to 2014

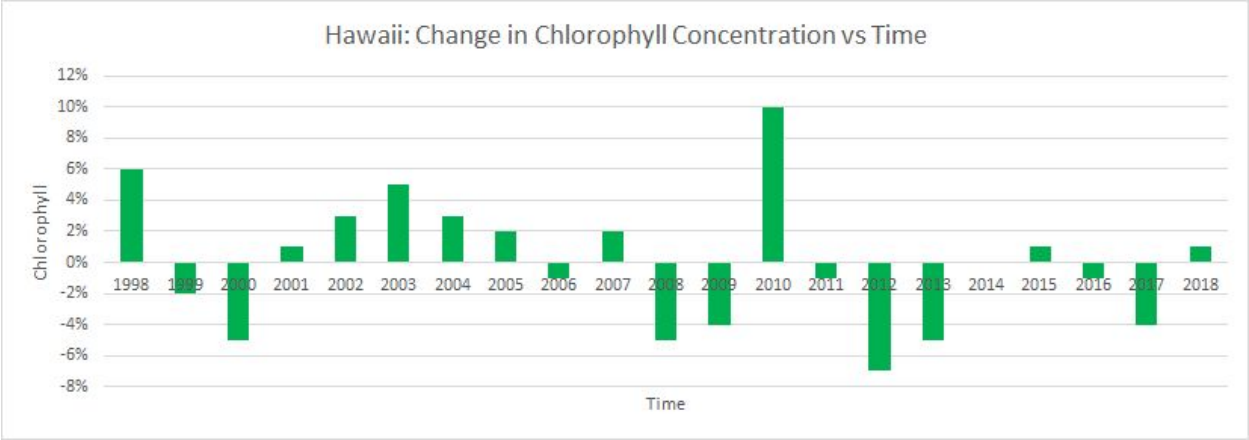


Bermudian pH levels versus time from 1989 to 2014



Hawiiwan pH levels versus time from 1989 to 2014

We can look at the peak or valley of (SST & pH) and try to match with chlorophyll and CO2 data (example: see if chlorophyll boom period overlaps with the peak in SST)



Hawaiiwan change chlorophyll concentration versus time from 1998 to 2018.

Remarks (Please comment if possible Dr. Lipsen)

- Our monthly surface water pH data is low resolution, so we'll be analyzing the annual average instead
- We've found a couple of Chlorophyll data, but they are mostly graphical data so we probably can't do numerical analysis with it. Would any of the following be suitable for analysis?

1. https://data.globalchange.gov/file/5ee52aad-c5d3-43aa-a18f-9820759a2896

2. https://data.globalchange.gov/report/indicator-ocean-chlorophyll-concentrations-2019/table/indicator-ocean-chlorophyll-concentrations-2019-6
3. https://hahana.soest.hawaii.edu/hot/methods/fluor.html
4. https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MY1DMM_CHLORA&year=2020

- With some of this graphical data, we can observe it and do a rough estimate analysis and compare the changes by the decade. We are planning to make a graph that shows a correlation between SST and pH. Additionally, we will be further analyzing/combining this data with chlorophyll and CO_2 data.
- We're wondering if you can help us understand how to read the dates in this csv graph? We can tell the year, but not the month of which the data point was collected. Not every year has 12 data points:

https://www.epa.gov/climate-indicators/climate-change-indicators-ocean-acidity

Eric's comment:

We've found a higher resolution data from BATS. However, I'm not sure if the "Temperature" is for ocean temperature. Under "Core Measurements at BATS" it only says depth range is 0-4200m. I also know they're using SBE-3F sensors, and doing some Google search I think it's for measuring ocean temperatures (but I'm not 100% sure). In the BATS data there is a column for depth (col 7), and one for temperature (col 8). So I'm 99% sure it's the ocean depth, please correct me if I'm wrong...

When accessing the BATS data and selecting a format type, there are multiple data files with extremely high resolutions. For example with the Excel format, for 1989 alone they have it stored in b10004_cts to b10016_cts (13 separate files) with probably thousands of rows each; data are grouped by dates but I can't make out what each specific date is.

I don't know what is the best method to extract and process all these data since there are so many separate files… will I have to download all the files individually and do analysis on them separately?? If so that is A LOT of work