Thomas Grothues

Jacques Cousteau Reserve & Rutgers University

Date: September 19, 2018
Time: 3.00 – 4.00 p.m. (EST)

National Estuarine Research Reserve System Science Collaborative

Trend Analysis of SWMP Temperature with Missing Data

Summary Points:

Dr. Tom Grothues has a Research Faculty appointment as a fish ecologist at Rutgers University and joined the Jacques Cousteau National Estuarine Research Reserve as Research Coordinator in Fall 2018.

He researches the abundance and distribution of fishes in response to physical factors. These responses include involuntary responses such as:

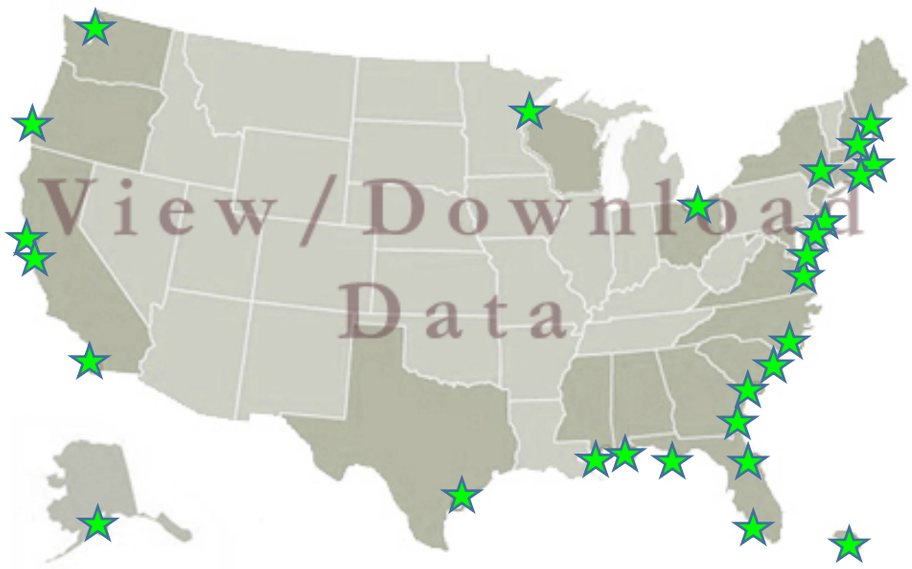
1. Distribution of larvae by ocean currents;
2. Mortality or loss of reproductive capacity in unsuitable environments; and
3. Voluntary behavioral responses, such as migration, ranging, and sheltering.

Tom has been using National Estuarine Research Reserve System (NERRS) System Wide Monitoring Program data in peer-reviewed publications about fish habitat use, migration, and recruitment since 2007.

What is SWMP?

- **S**ystem **W**ide **M**onitoring **P**rotocol
- 4 water quality loggers x 28 Reserves
- 1 weather station x 28 Reserves
- Vegetation/GIS Maps
- Same parameters
- Same schedule
- Same QA/QC
- Same Data Portal
- Same tools

He'eia, HI



Summary Points:

The National Estuarine Research Reserve System is comprised of 29 reserves nationwide.

As of September 2018, 28 of the reserves follow a System Wide Monitoring Program (SWMP) protocol. The newest reserve, He'eia Reserve in Hawaii, will be participating in 2019.

Each of these 28 reserves has four water quality loggers and one weather station, as well as vegetation maps, GIS maps, and other resources.

Each reserve collects data in the same way and at the same timestamps, using the same water quality loggers and following the same parameters for maintaining and calibrating them. All data are shared through a common data portal, with common tools for data management available to the reserves.

Terminology:

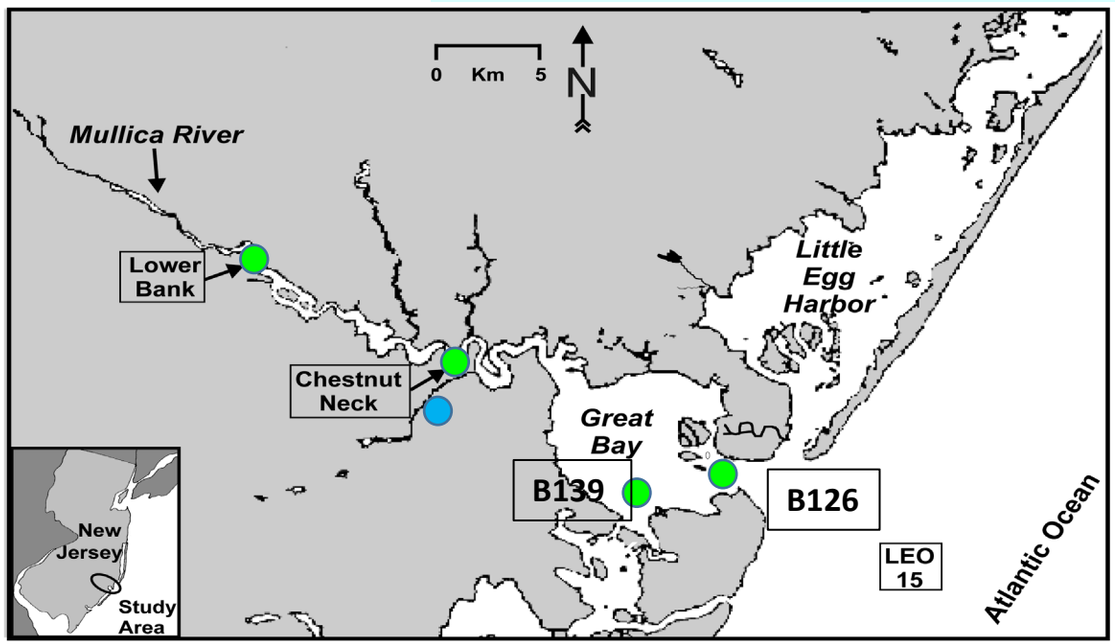
- **Water quality loggers**: Tools that can be used to monitor a range of environmental water quality data, such as pH, temperature, conductivity, and turbidity.

Summary Points:

SWMP data are managed by the NERRS Centralized Data Management Office (CDMO), which is housed at the University of South Carolina's Baruch Institute for Marine and Coastal Sciences.

You can access SWMP data at cdmo.baruch.sc.edu.

SWMP in the Jacques Cousteau National Estuarine Research Reserve



- Weather – Wind, Barometric Pressure, Humidity, Precipitation, PAR
- Water Quality – Salinity, Temperature, pH, DO, Tide, (Nutrients)

Summary Points:

This map shows the distribution of the four water quality data loggers and a weather station at the Jacques Cousteau NERR in New Jersey.

The data discussed in this webinar come from Buoy 126, seen on lower right side of the slide.

The physical structure of Buoy 126 is shown in the photo. The logger is attached to a Coast Guard-maintained Aid to Navigation within Great Bay.

SWMP Water Quality is…

- HI RES – 15 minute interval!
- Salinity, temperature, pH, DO, depth
- Freely available
- Centrally managed
- With Metadata

  https://cdmo.baruch.sc.edu/

DateTimeStamp	Historical	Provisiona	F_Record	Temp	F_Temp	SpCond	F_SpCond	Sal	F_Sal	DO_Pct	F_DO_P
1/1/2008 0:00	0	1		6.1	<1> [GNF]	44.15	<1> [GNF]	28.1	<1> [GNF]	100.9	<1> [GN
1/1/2008 0:15	0	1		6.1	<1> [GNF]	44.07	<1> [GNF]	28	<1> [GNF]	100.8	<1> [GN
1/1/2008 0:30	0	1		6.1	<1> [GNF]	44.3	<1> [GNF]	28.2	<1> [GNF]	100.6	<1> [GN
1/1/2008 0:45	0	1		6.1	<1> [GNF]	44.51	<1> [GNF]	28.3	<1> [GNF]	100.5	<1> [GN
1/1/2008 1:00	0	1		6.2	<1> [GNF]	44.72	<1> [GNF]	28.5	<1> [GNF]	100.4	<1> [GN
1/1/2008 1:15	0	1		6.2	<1> [GNF]	44.89	<1> [GNF]	28.6	<1> [GNF]	100.3	<1> [GN
1/1/2008 1:30	0	1		6.2	<1> [GNF]	45.15	<1> [GNF]	28.8	<1> [GNF]	100.2	<1> [GN
1/1/2008 1:45	0	1		6.3	<1> [GNF]	45.32	<1> [GNF]	28.9	<1> [GNF]	100.1	<1> [GN
1/1/2008 2:00	0	1		6.3	<1> [GNF]	45.7	<1> [GNF]	29.2	<1> [GNF]	99.8	<1> [GN
1/1/2008 2:15	0	1		6.3	<1> [GNF]	45.83	<1> [GNF]	29.3	<1> [GNF]	99.6	<1> [GN
1/1/2008 2:30	0	1		6.3	<1> [GNF]	45.96	<1> [GNF]	29.4	<1> [GNF]	99.4	<1> [GN
1/1/2008 2:45	0	1		6.3	<1> [GNF]	46.09	<1> [GNF]	29.4	<1> [GNF]	99.2	<1> [GN
1/1/2008 3:00	0	1		6.3	<1> [GNF]	46.14	<1> [GNF]	29.5	<1> [GNF]	98.9	<1> [GN
1/1/2008 3:15	0	1		6.3	<1> [GNF]	45.39	<1> [GNF]	29	<1> [GNF]	98.8	<1> [GN
1/1/2008 3:30	0	1		6.4	<1> [GNF]	46.19	<1> [GNF]	29.5	<1> [GNF]	99	<1> [GN
1/1/2008 3:45	0	1		6.3	<1> [GNF]	45.35	<1> [GNF]	28.9	<1> [GNF]	99	<1> [GN
1/1/2008 4:00	0	1		6.3	<1> [GNF]	45.92	<1> [GNF]	29.3	<1> [GNF]	99.2	<1> [GN
1/1/2008 4:15	0	1		6.3	<1> [GNF]	46.01	<1> [GNF]	29.4	<1> [GNF]	99.2	<1> [GN
1/1/2008 4:30	0	1		6.2	<1> [GNF]	45.78	<1> [GNF]	29.2	<1> [GNF]	99.2	<1> [GN
1/1/2008 4:45	0	1		6.2	<1> [GNF]	45.57	<1> [GNF]	29.1	<1> [GNF]	99.3	<1> [GN
1/1/2008 5:00	0	1		6.1	<1> [GNF]	45.23	<1> [GNF]	28.8	<1> [GNF]	99.4	<1> [GN
1/1/2008 5:15	0	1		6.1	<1> [GNF]	45.1	<1> [GNF]	28.7	<1> [GNF]	99.5	<1> [GN
1/1/2008 5:30	0	1		6	<1> [GNF]	45.02	<1> [GNF]	28.7	<1> [GNF]	99.4	<1> [GN
1/1/2008 5:45	0	1		6	<1> [GNF]	44.75	<1> [GNF]	28.5	<1> [GNF]	99.3	<1> [GN
1/1/2008 6:00	0	1		5.9	<1> [GNF]	44.53	<1> [GNF]	28.3	<1> [GNF]	99.3	<1> [GN
1/1/2008 6:15	0	1		5.9	<1> [GNF]	44.26	<1> [GNF]	28.1	<1> [GNF]	99.1	<1> [GN
1/1/2008 6:30	0	1		5.9	<1> [GNF]	44.09	<1> [GNF]	28	<1> [GNF]	99.1	<1> [GN
1/1/2008 6:45	0	1		5.8	<1> [GNF]	44.15	<1> [GNF]	28	<1> [GNF]	99	<1> [GN
1/1/2008 7:00	0	1		5.8	<1> [GNF]	43.92	<1> [GNF]	27.9	<1> [GNF]	99	<1> [GN
1/1/2008 7:15	0	1		5.8	<1> [GNF]	43.96	<1> [GNF]	27.9	<1> [GNF]	99	<1> [GN
1/1/2008 7:30	0	1		5.8	<1> [GNF]	44.03	<1> [GNF]	27.9	<1> [GNF]	98.9	<1> [GN
1/1/2008 7:45	0	1		5.8	<1> [GNF]	43.02	<1> [GNF]	27.2	<1> [GNF]	98.8	<1> [GN
1/1/2008 8:00	0	1		5.8	<1> [GNF]	43.36	<1> [GNF]	27.5	<1> [GNF]	98.8	<1> [GN

Summary Points:

SWMP data collection at the Jacques Cousteau NERR began in 1996.

Each water quality logger collects data on salinity, temperature, pH, dissolved oxygen (DO), and depth in 15-minute intervals.

The data and metadata can be accessed at cdmo.baruch.sc.edu. When downloading the data, users receive a table similar to the one shown here.

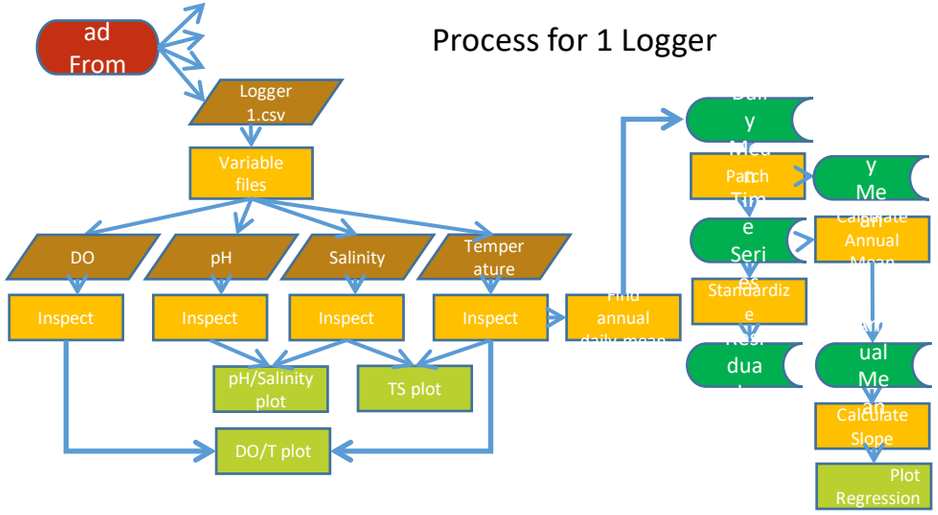
In the header row of the table, there are columns titled “F” (e.g. “F_Temp” and “F_Sal”). These columns are quality assurance and quality control (QAQC) flags, which allow filtering by determinants of data quality that have been marked by technicians’ statistical analyses. In the example on the slide, the QC tag <1> indicates “suspect data,” and the description [GNF] indicates a water quality-specific error in which there was a clog in the deployment tube preventing flow. More information on QAQC codes is available at cdmo.baruch.sc.edu/data/qaqc.cfm.

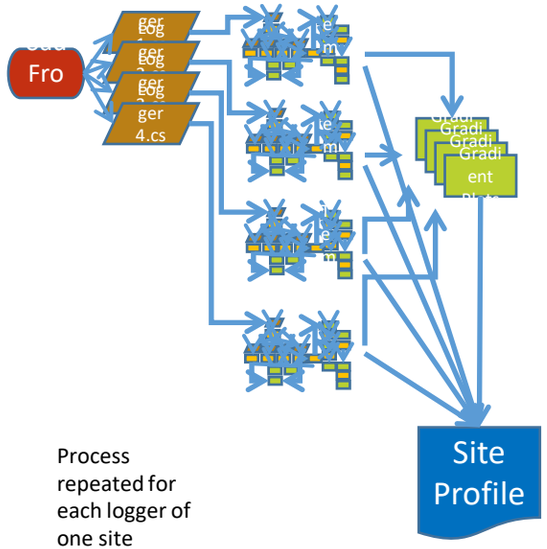
When using the data, users can choose to include or exclude provisional data, which are data that:

- May be good, but were flagged for being potentially suspicious; or
- Have not been checked.

Transfer Project-
Transfer of SWMP analytical products and capability

- Funded through a National Estuarine Research Reserve Science Transfer Grant
- Provided MATLAB and R code for working with SWMP data
- Provided workshop for transfer





Summary Points:

Collecting and managing SWMP data according to a universal protocol across reserves facilitates standardized studies at multiple sites. However, reserves have different capacities for working with and analyzing SWMP data.

All reserves are funded by federal grants and managed by local custodians, including universities, states, or federal agencies like U.S. Fish and Wildlife Service. Since software like MATLAB is expensive and typically only available at universities, reserves that are not linked to universities typically do not have the capacity to work with software and develop code to analyze their SWMP data.

A 2015 Science Transfer grant from the NERRS Science Collaborative allowed Tom and a colleague to write code that could be disseminated to reserves that do not have partnerships with universities.

The project team provided MATLAB and R code for working with SWMP data and transferred this to reserves in a two-day workshop. R code is free, widely available, and very popular.

Results of Transfer Project and Developments

- Code developed, taught, customized, translated to R
- Transferred to private consulting company for development of automated reporting
- Incorporated in SWMPr
- Elucidated challenges…

Summary Points:

Code developed through the project included scripts for managing and examining temperature, dissolved oxygen, and pH data. To date, these data have been widely shared and subsumed into SWMPr. A number of the tools the project team developed were also subsumbed into SWMPr and turned into larger R code packages.

All the tools the project team developed were also given to a private consultant hired to generate auto-reporting and auto-documentation for each reserve's Annual Report.

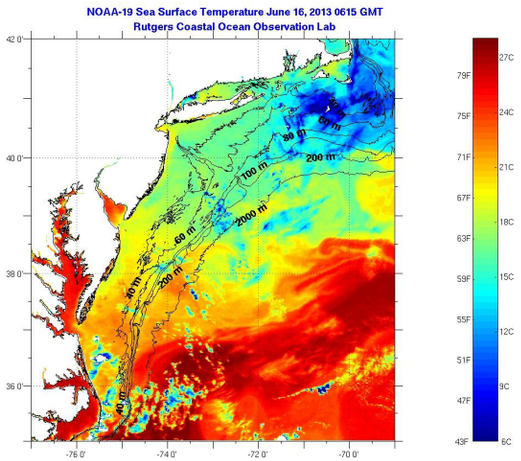
This webinar focuses on one important topic that came out of interacting with reserve staff at the two-day workshop: treatment of missing temperature data.

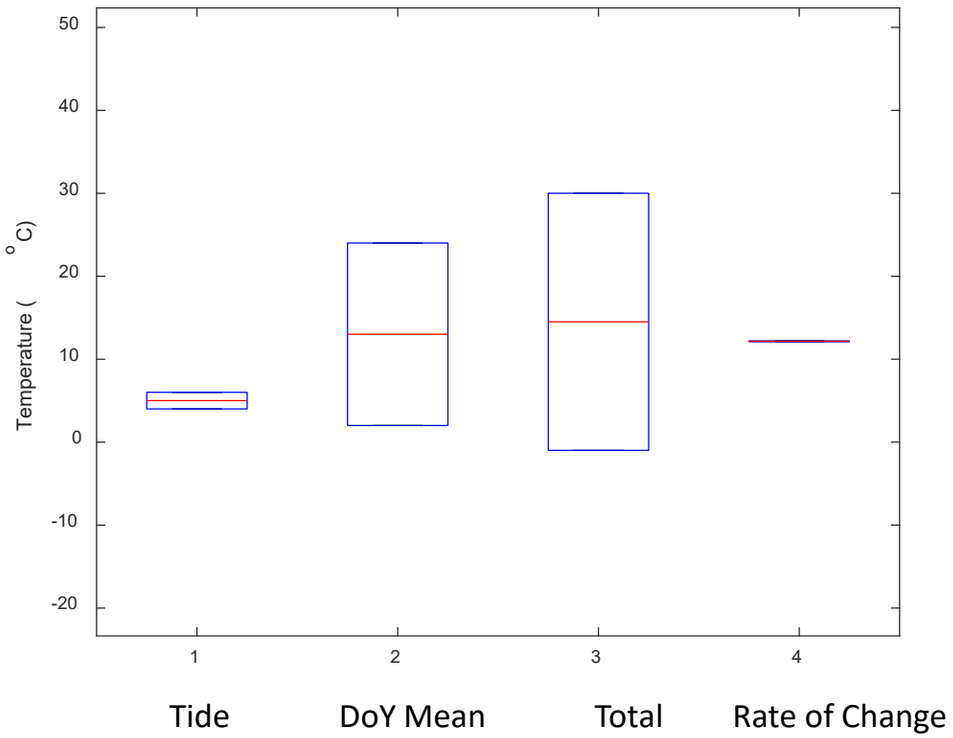
Terminology"

- **SWMPr**: an R package for retrieving, organizing, and analyzing estuary monitoring data from SWMP.

Temperature Variation Scales in Mid-Atlantic

- Range due to tide ~ 6 ºC
- Range of daily mean ~ 22 ºC
- Total range ~ 31 ºC
- Range of annual mean ~ 4 ºC
- Interannual rate of change ~ 0.1 º C





Summary Points:

These graphs help explain how dynamic temperature is relative to the sources of variation, and why it matters how missing data are treated.

The Mid-Atlantic Bight is the region between Cape Hatteras, North Carolina and Cape Cod, Massachusetts. It experiences the highest annual temperature cycle of any oceanographic province worldwide. The total range is 31° C in any given year, and the daily mean range is 22° C.

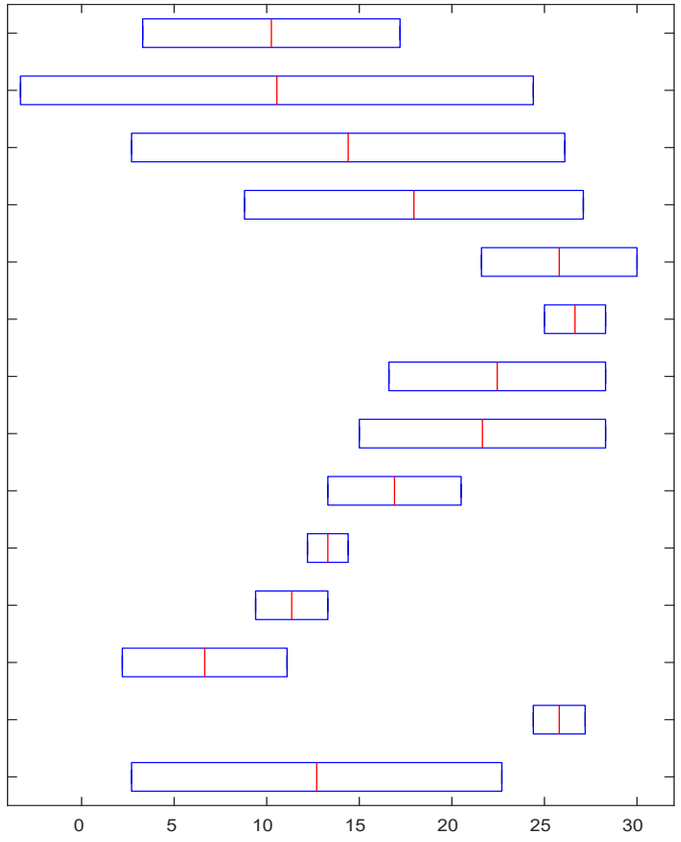
As the interannual rate of change is so small in terms of degrees Celsius - compared to the ranges from other sources - it is difficult to detect as a signal within the “noise” of variation.

Terminology:

- **Day-of-Year (DoY) Mean**: An average mean temperature for a specific day of the year, calculated by averaging the mean temperature on a given day in each year from 1997 to 2017.

Temperature Variation Elsewhere –NODC.NOAA.gov

Location	Min Average	Max Average
Bar Harbor ME	3.3	17.2
Bergen NY	-3.3	24.4
Baltimore MD	2.7	26.1
Virginia Beach, VA	8.8	27.1
Myrtle Beach, SC	8.8	27.1
Miami Beach FL	21.6	30
San Juan, PR	25	28.3
St. Petersburg FL	16.6	28.3
Port Aransas TX	15	28.3
Oceanside CA	13.3	20.5
Morro Bay CA	12.2	14.4
Newport, OR	9.4	13.3
Juneau, SEAK	2.2	11.1
Honolulu, HI	24.4	27.2
Atlantic City, NJ	2.7	22.7



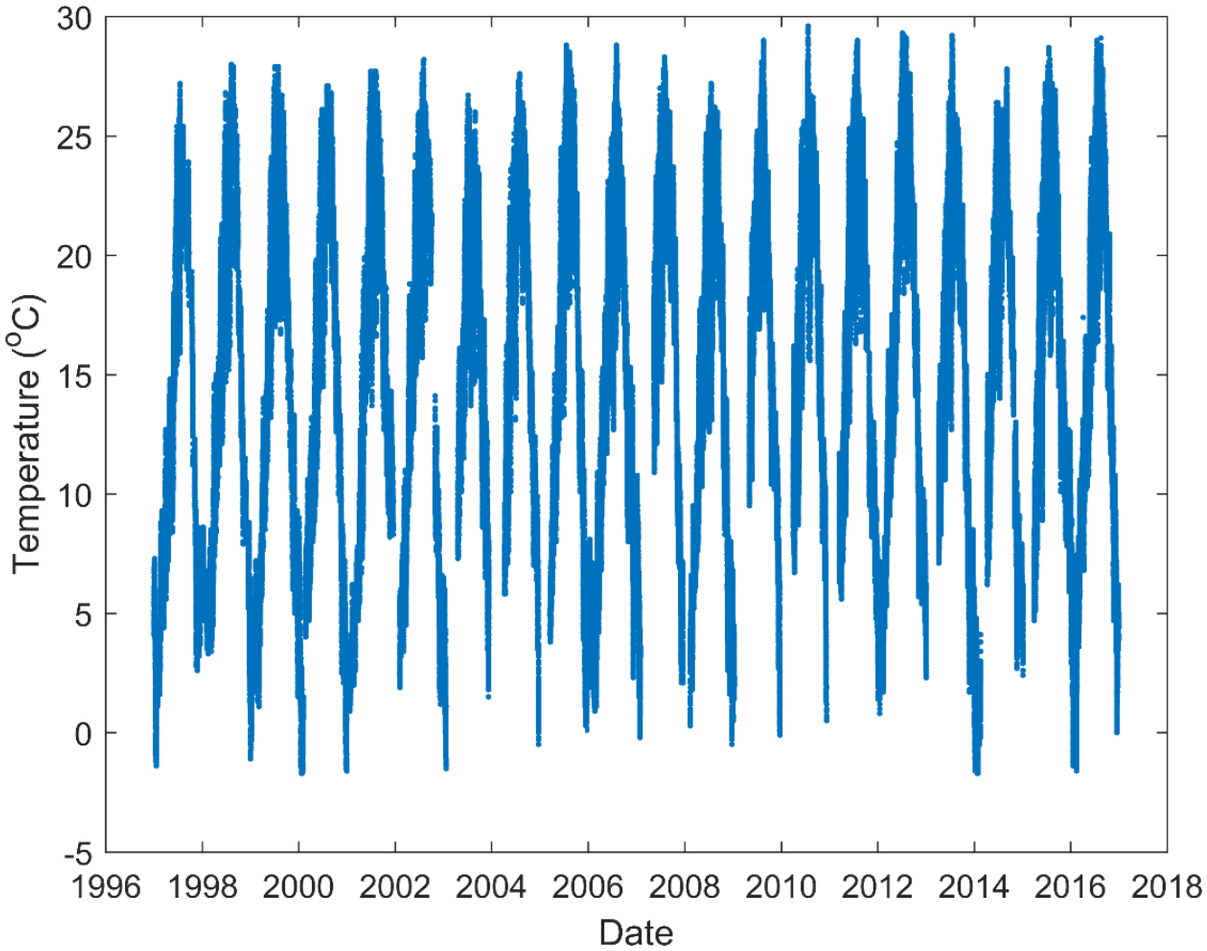
Summary Points:

The chart and table show how temperature varies elsewhere in the United States.

Atlantic City, New Jersey is close to Jacques Cousteau Reserve. The annual mean daily temperature ranges here, and at other locations in the Mid-Atlantic, are very high; because of this inherent range, any bias in the collection of temperature data can make it difficult to accurately calculate means and trends.

JCNERR
Logger “B6” at
Channel Marker 126

1997-2017
Filtered using QA/QC
flag notations

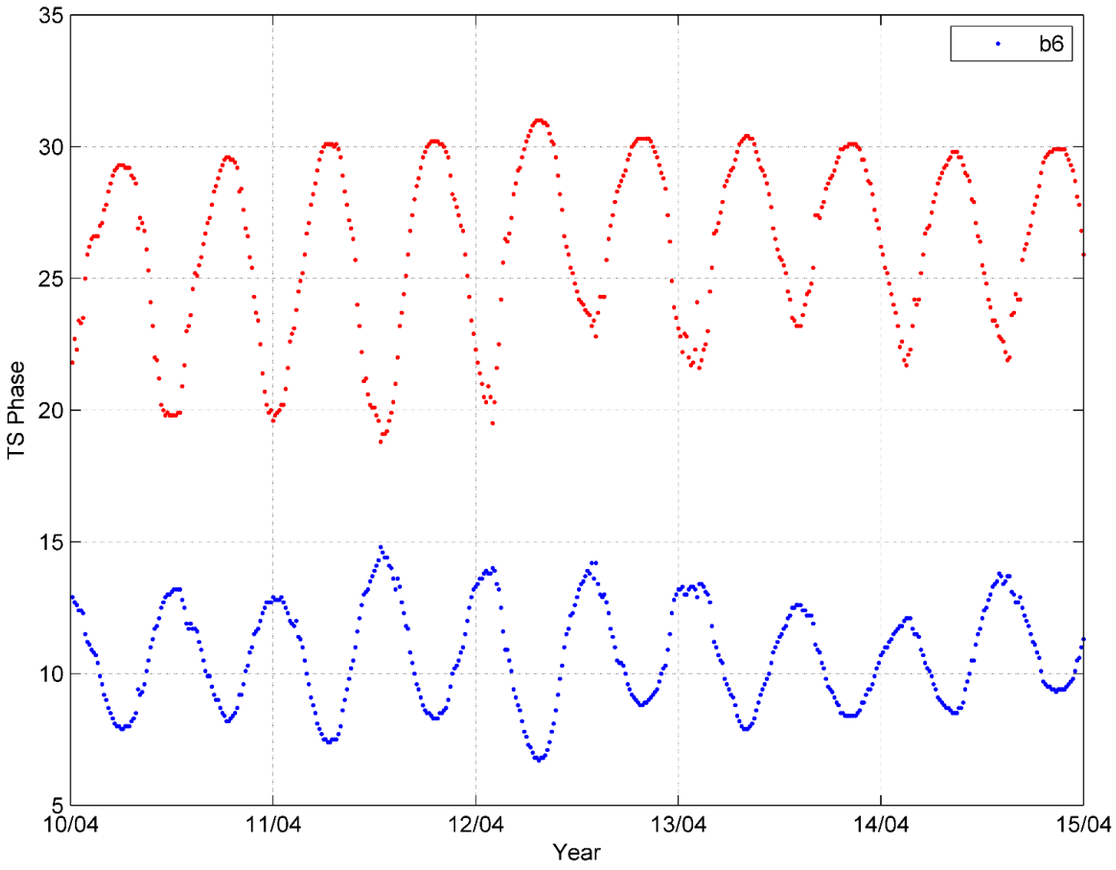


Summary Points:

This plot depicts temperature data at Jacques Cousteau Reserve from 1996-2017. This includes every data point that passed the quality-assurance/quality-control (QAQC) check for high quality data.

The peaks and absolute lows exceed daily averages because, in some cases, there is a range in temperature of as much as 6° C as the tide ebbs and flows.

The data show that there are clear, and sometimes large, gaps in temperature-data collection over the given time.



Salinity

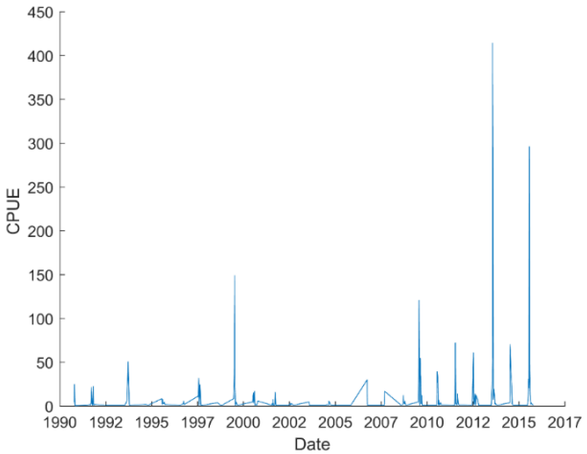
Temperature

Summary Points:

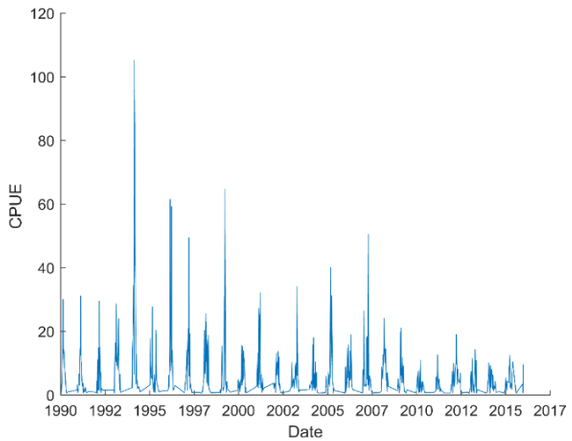
These graphs show some of the temperature and salinity fluctuations at Buoy 126.

Salinity rises and falls as the tide ebbs and flows, and is perfectly out of phase with temperature. As the tide comes in and the estuary receives an influx of cold ocean water, salinity increases and temperature decreases.

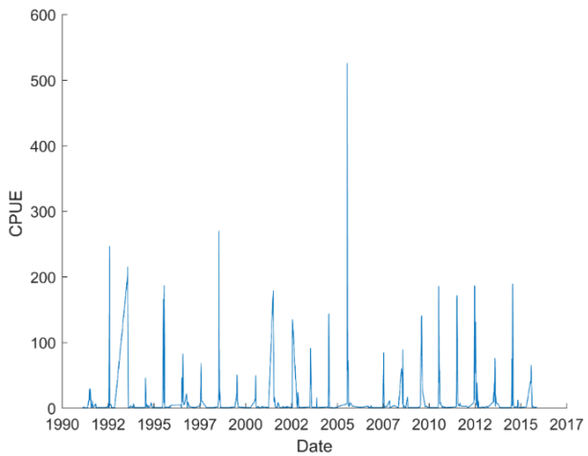
Challenge Example- Phenology



Anchoa hepsetus



Anguilla rostrata



Gobiosoma bosc

Need - Fit temperature to variance in abundance or timing of ingress for long term data sets

Summary Points:

SWMP temperature data are used for a variety of projects, many of which are tied to climate change. The project team is interested in even the smallest differences in temperature change, particularly relating to phenology.

At Jacques Cousteau NERR, the team has been conducting ongoing fish sampling since 1989 and have noted more than 135 fish species.

The graphs shown depict three examples from bridge netting at the reserve, where staff collect larvae entering the estuary by fishing off the bridge with a plankton net. The team takes three samples one night per week, providing a dense, long-term record of data.

The team is particularly interested in assessing how the arrival time of larvae entering the estuary is affected by temperature and other variables.

SWMP data are typically used in multivariate analyses; however, if any data are missing, the associated dependent data must also be dismissed, even if the other data are intact. For example, a gap in temperature data, but not salinity or pH data, would still mean the salinity and pH data points would need to be discarded unless treating the temperature data with some sort of replacement.

Assumptions of rigor for analytical processes

- Adequate sample size
- Fair sampling
  - Independence, free of spatio-temporal bias
- Homoscedasticity

OLS Regressions are parametric!

Summary Points:

These are some important data set assumptions for providing confidence in the results of a statistical analysis.

Particularly worth noting: The data need to be free of bias in regard to when and where they were collected; this especially applies to parametric statistical tools used for testing hypotheses.

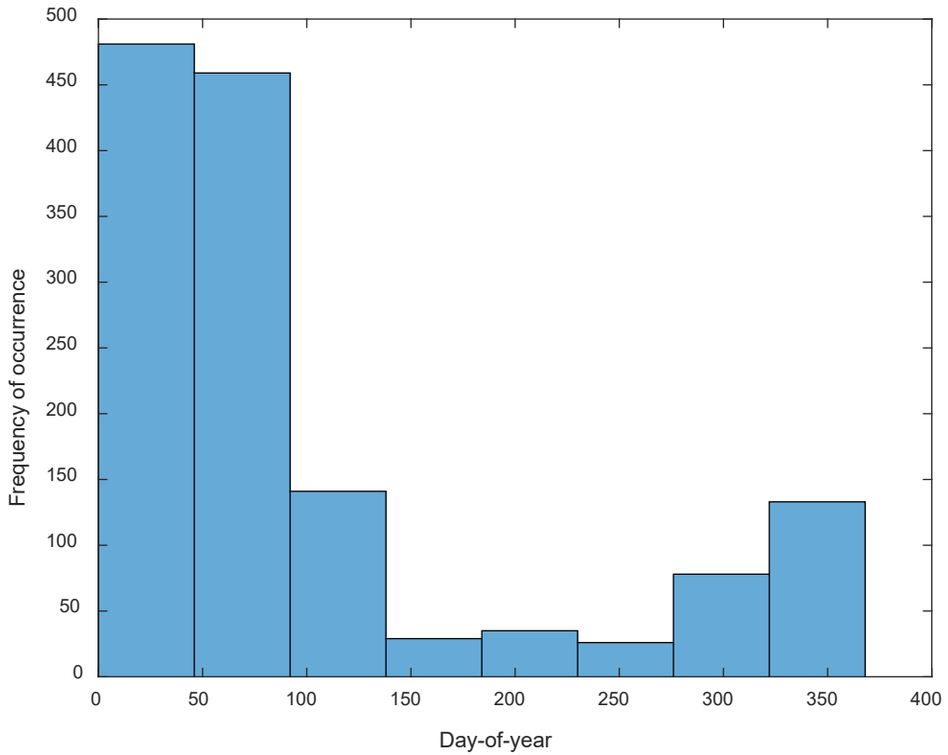
Terminology:

- **Homoscedasticity**: Same variance; an assumption meaning that the variance around the regression line is the same for all values of the predictor variable (X).
- **Ordinary Least Squares (OLS) Regression**: A statistical method for estimating unknown parameters in a linear regression model using existing data.

Missing data

- Occurs from logger loss or damage
- Occurs from planned removal
- Both are seasonally biased
- For JC B126 - 1382 days of 6939 had partial missing or full missing days

Kolmogorov-Smirnov GoF test K = 0.9871, p< 0.001



Summary Points:

Timeseries data can be incomplete for a variety of reasons, including loss or planned removal of - or damage to - a logger, and quality issues that don't destroy the logger but render data suspect.

Missing data in the Jacques Cousteau NERR are seasonally biased. For the logger at Buoy 126, 1382 of 6939 days had partially- or fully-missing data. The histogram shows that there is a clear bias in the frequency of missing data points each year, with the majority of missing data occurring between day 0 and day 120 - roughly January through April, but particularly January through March - of a given year.

Summary Points:

At the Jacques Cousteau NERR, there are frequently missing data in the first few months of each year because the estuary is filled with ice. Loggers go missing because they are forcefully dislodged by ice, or they are intentionally removed due to an issued ice warning.

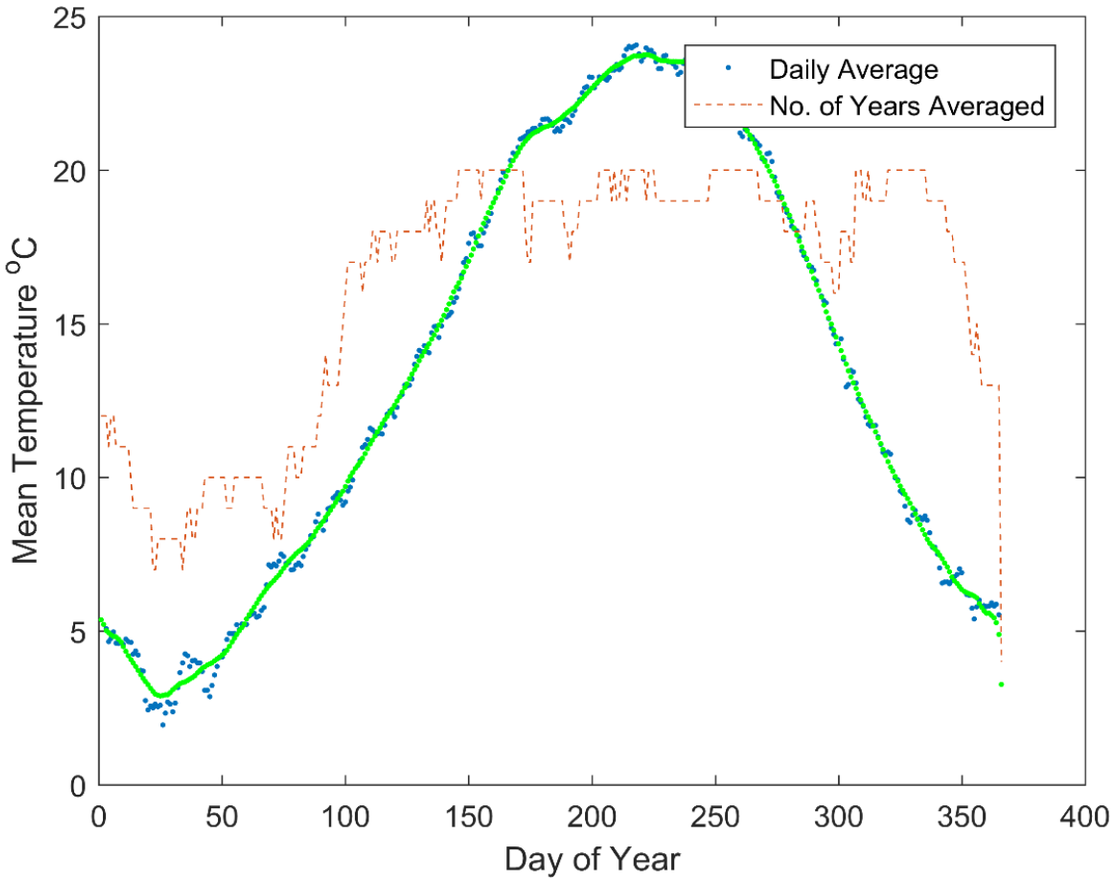
Treatment for missing data in Analysis: Options

- Ignore it
- Use Annual Julian Intersect
- Use benchmark Annual Julian dates
- Replace it with annual DoY or decimal Julian averages
- Substitute proximal data

Summary Points:

There are a number of different ways to address missing data. Some of these methods give trends in, but not a particular number of, instances of missing data for the year in need of patching.

At Jacques Cousteau NERR, the team has some proximal data from daily handheld temperature measurements taken nearby, which allows them to check for any fit between their data and those proximal data. However, proximal data are not always available.



1997 -2017 Day-of-Year average

LOESS smoothed

DoY smoothed average will change depending on years included

Summary Points:

Tom averaged the daily mean temperature data for each day from 1997 to 2017, which produced average day-of-year mean temperature data for 366 days-of-year; this includes a leap year every four years.

The day-of-year mean temperatures are plotted as blue dots. The red hatched line shows how many years provided available data points for the average day-of-year mean temperature calculation. The green line is a smooth version of the day-of-year averages, which was necessary due to the clear variation in available data for the daily day-of-year averages.

The days-of-year that show particularly large portions of missing data, especially day 0 to day 50 - or January 1 to late February - are cold days with higher likelihood of ice. Removing those data stands to introduce bias because, by failing to treat data from the coldest years, annual mean temperature calculations will appear erroneously warmer, which in turn will change the slope of the calculated trend.

Terminology:

- **Locally Estimated Scatterplot Smoothing (LOESS)**: A non-parametric method for fitting a smooth curve between two variables, or fitting a smooth surface between an outcome and up to four predictor variables.

Methodology – benchmark dates

- Choose and use a series of DoY rather than all days
- Limit date set to seasons without bias in missing data
  - Useful as covariates to seasonally restricted phenomena

Given a data set of Year, Day-of-Year, and Temperature

```
% find the set of nth year days
benchdates = [60,86];
index = find(yd > benchdates(1) & yd < benchdates(2));
benchset = [yr(ind) yd(ind) temperature(ind)];

for n = 1 : length(y)
   y(n)
   mean_set = benchset(benchset(:,1) == y(n),3);
   %Column 1 finds year and 3 is the temperature
   annual_mean(n) = nanmean(mean_set);
   annual_std(n) = nanstd(mean_set);
end
annual_mean
```

Summary Points:

Using the benchmark dates method, users can choose and use a series of the day-of-year means rather than all days.

The example shown on the right side of the slide uses 26 days around the middle of the year to calculate a trend.

One problem is that, with the selected benchmark dates, there is potential for data gaps due to missing individual dates.

This method is useful if the user is interested in data from one season. For instance, if a user is studying the phenology of a species that recruits in the spring and only needs the spring temperatures, this method works well. However, if a user is looking at the phenology of a species that has some annual persistence, they would need whole-year data and this method is less useful.

Methodology – Intersect

- Find the set of annual Julian datetimes common to every year
- Unbiased estimator of seasonally biased trend
- Intercept will be offset, so annual mean is not a good estimate of actual annual mean
- Computationally expensive

```
[~, mo, dy, hr, mn, ~] = datevec(jday);
% deconstruct jday to make all the same year
jday = datenum(1900, mo, dy, hr, mn, 0);

F = fieldnames(Z)  % These are the names of each year's DoY list
H = Z.(F{1});   % Initialize X as the first years list of DOY

% Find the intersecting timestamps of each year
% to the previous intersect set
for ii = 2:length(y)
    H = intersect(H,Z.(F{ii}));
end

For each year
    Find the year set
            [~, indyrset, indH]  = intersect(yrset(:,1), H);
            annual_mean = mean(year_set
     end
end
```

Summary Points:

An alternative method is the intersect method, which allows users to find data points that are shared between two data sets.

Process:

1. Find the intersect of the first year and the second year and keep that as a new intersect data array.
2. Repeat step 1, using this new intersect data array with the third year to create a new intersect array.
3. Iterate this process through the loop until all years are finished.

Each time the user iterates through the loop, they lose more days from their data set, so this method very quickly shortens the total intersect.

This method can give a clear, unbiased view of the temperature trend, but the intercept will be offset because winter days missing from one annual set are excluded for all years as “unmatched.” As such, the calculated annual mean is not a good estimate of the actual annual mean.

This method is also computationally expensive, and often needs to run overnight with large data sets.

Methodology – Replace missing data with DoY means

- Find missing DoY, index them and keep track of them as a set for each year

```
for n = min(yr):max(yr);
   %disp(n)
   ind = yr == n;
   for j = 1:366
      indyd = find(yd(ind,1) == j, 1);
      if isempty(indyd)
         %disp(j);
         miss = [n, j];
         missingyd = [missingyd;miss];
      end
   end
end
```

- Calculate daily averages

```
aa=[];
av_temp = zeros(366,1); % preallocate for efficiency

for n = 1:366
  tempset= x(x(:,3)== n,:);  % find all rows for each yd
  %calculate the mean across all those rows
  av_temp(n) = nanmean(tempset(:,1));
  %this will return the years of the dates in the selection
  yrs = tempset(:,2);
  a = unique(yrs);
  aa =[aa;a];
  [r(n),c(n)] = size(a);
  clear c
end
```

Summary Points:

The most conservative method is to replace missing data with the day-of-year means.

Step 1: Find the missing days-of-year for each year, index them, and then keep track of them as a set for each year.

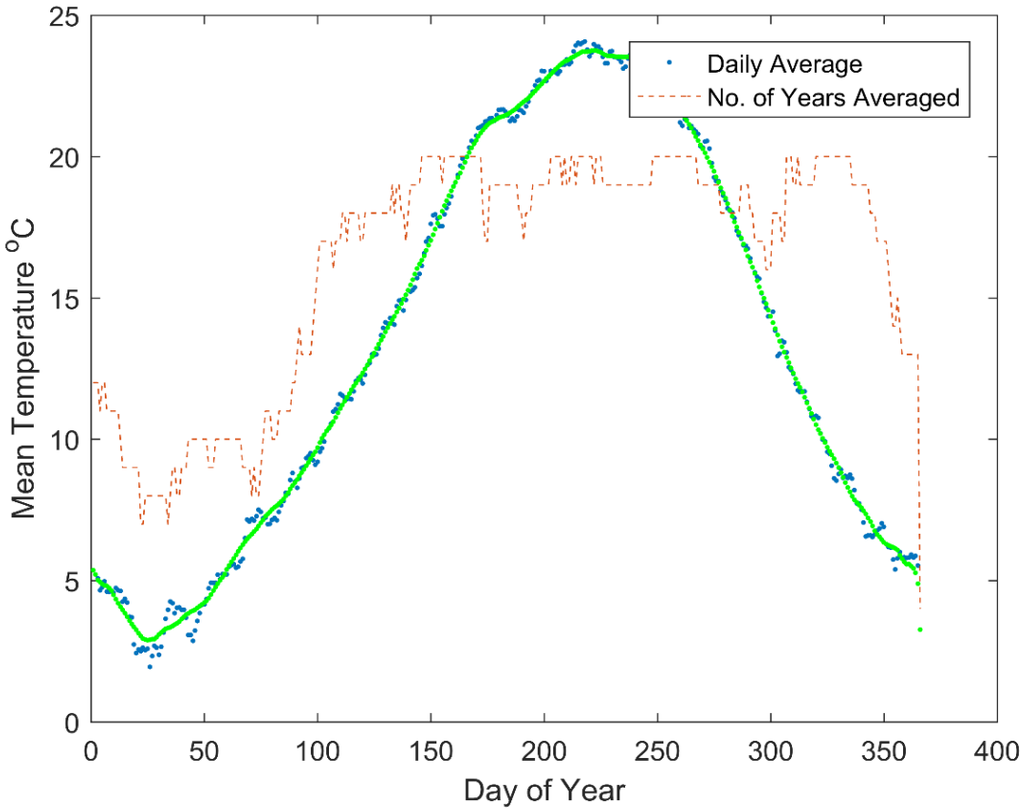
Step 2: Calculate the daily averages from available data. The more years for which mean temperatures are available for a given day-of-year, the higher the confidence that the calculated average reflects the mean.

Note: Even the temperature from a day-of-year sampled only once in all available years is an estimate.

Methodology – Replace missing data with DoY means

- Plot the averages and LOESS smooth them
- Smoothing window is locally chosen

```
figure; plot(av_temp,'.'); hold on
plot(r,'--') %shows the number of years in the average
hold on
smoothed_av_temp = smooth(av_temp,20);
%20 points is 4 points x 5 hrs
plot(smoothed_av_temp,'g.')
```



Summary Points:

Step 3: Plot the averages and smooth them using LOESS.

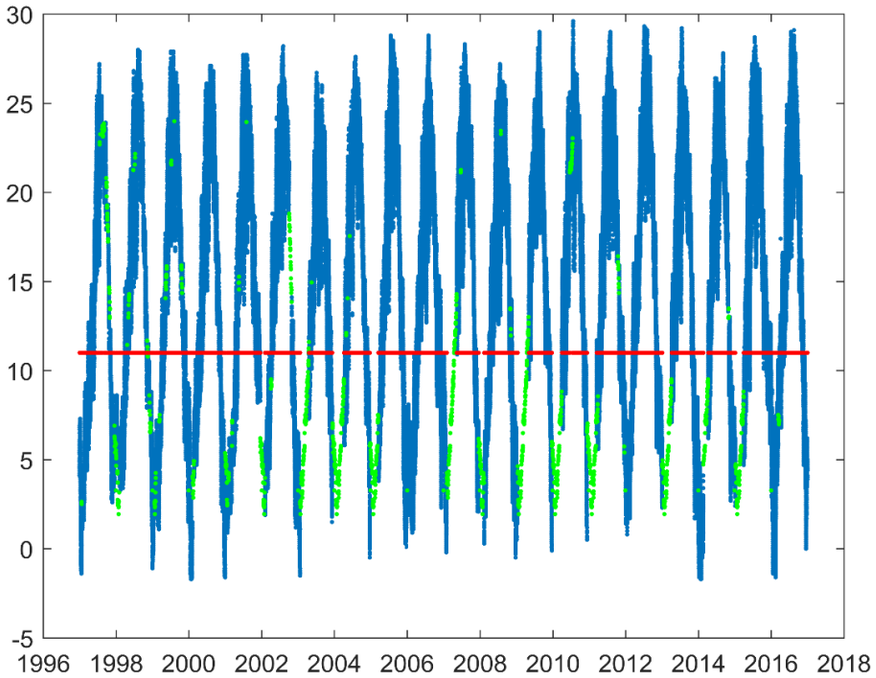
LOESS gives a moving window average. For example, starting at position 10 in a window of 20, the calculation would use the 10 days before and the 10 days after the starting position, weight the average of the temperature at the center day more than the averages at the fringe of the range, then move the window over one position and start again.

This process flattens out waves in the data, as shown by the green curve.

Methodology – Replace missing data with DoY means

- Join the actual data set with the replacement set and check it

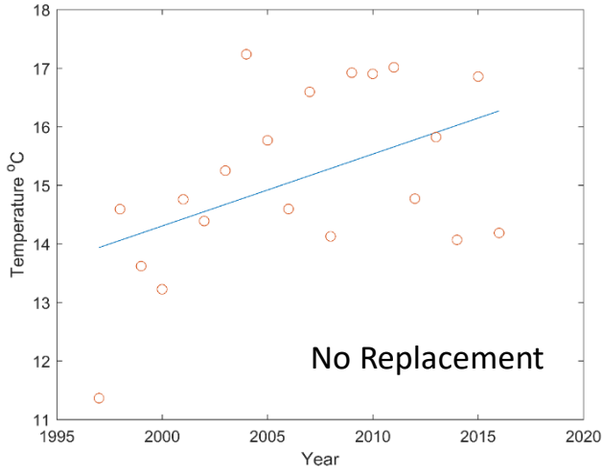
```
for n = 1:366
   ind = find(missingyd(:,2) == n); %index the rows
of yd that equal n, starting with n
   valin = av_temp(n);
   missingyd(ind,4) = valin;
end

plot(jday,wq,'.')
hold on
plot(missingyd(:,3), missingyd(:,4),'g.')
hold on
plot(jday,11,'r.')
```

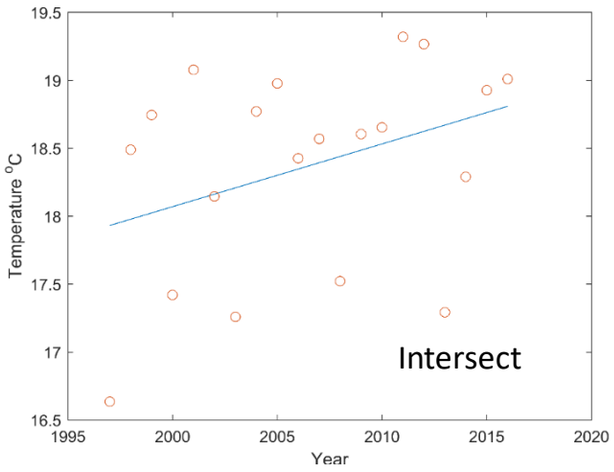


Summary Points:

Step 4: Finally, plot the missing data with the original data and inspect it.

Note: Again, the original data have a greater range because they include raw data that make up daily averages, but the peak highs and lows collapse toward the mean of the given day when calculating temperature trends.





For all

Fit the trend to year by OLS

Summary Points:

To examine each method's sensitivity to deviation, Tom fitted a trend line across years to the data sets - patched using each method - using ordinary least squares.

Notably, the benchmark method had no data points for some years, but the trend can still be useful in some analyses.

Also worth noting: If a goal is to join these data to another data set, or use them as part of a multivariate set, the user still would not have a usable value and thus would still have to either discard the other dependent data, such as those for salinity or pH, or use the expected value from the fitted line.

Terminology:

- **Ordinary least squares**: A statistical method for estimating unknown parameters in a linear regression model using existing data.

Method	Slope	R^2	Calculated Change over 20 years (°C)	P value	Comment
As available	0.1229	0.2238	2.46	0.0352*	Test for bias first
Warmest	0.0872	0.3512	1.74	0.0059*	Seasonal mean, no annual mean
Set intersect	0.0462	0.1305	0.92	0.1176	Could produce seasonal mean, no annual mean
Benchmark dates only	0.1128	0.2941	2.25	0.0685*	Some years skipped, could produce seasonal mean and annual mean
Replace with DoY mean	0.1225	0.2300	2.45	0.0324*	

Summary Points:

This table shows the results from the regression fitted to each of the different "patched" data sets.

Intepreting statistical significance:

- **P value**: A measure of the certainty that the results of a given study were not due random chance. A statistically significant P value is less than 0.05, meaning that there is less than a 5 percent chance that the results obtained from a study occurred by random chance.
- **R^2 value**: Also known as the coefficient of determination, this value is an indication of how close the data are to a fitted regression line. A value of 1 indicates that the regression model perfectly fits the data, while a value of 0 indicates no fit.

Considerations

- Do you need annual or only seasonal data?
- Is there distributional bias in the data set you will use?
- Do you need the trend, or actual covariates?
- What are your computational resources

Summary Points:

If only in need of seasonal data: Only use that portion of the data set. By using a smaller data set, there is decreased chance for bias in the data, but users should always check for the distribution of missing data within the set.

If in need of covariates: Use a method such as replacement of missing data in order to generate the means that go with the data set for which the user needs a covariate.

Note: It is important to consider computational resources. Some computers cannot handle replacement of missing data. MATLAB loops can slow computers down considerably, even though it is a powerful program.

Next

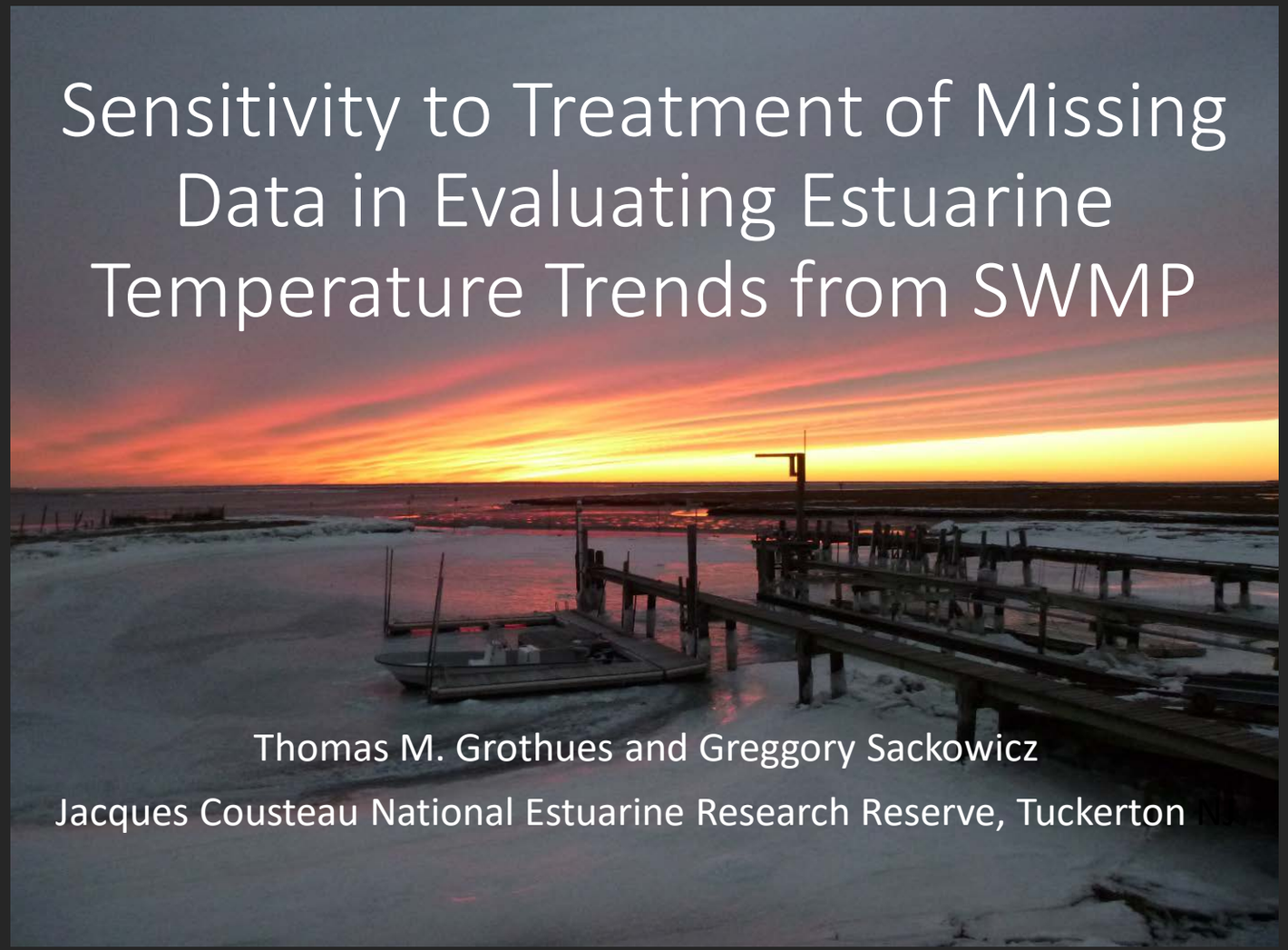
- Synthesize a data set as sine function, vary ranges
- Add noise
- Add trend
- Remove data using biased and psuedo-RNG
- Recover trends with different treatments, examine for patterns relative to parameters of range, trend, smoothing etc.

Summary Points:

The team's next step is to figure out what they missed and what should have been done before replacing the data.

The team has a graduate student working on a project synthesizing a data set as a sine function, adding noise and trend, and then recovering that trend. They will then remove data using biased and pseudo-random number generators to create a sprinkling of missing data or whole patches that can be randomly or non-randomly distributed. The team will then recover trends and examine them for patterns relative to parameters of range, trend, and smoothing.

Note: When one replaces values with the mean of the data set, the overall data set is pulled back toward the mean; means cannot increase a trend in a data set.



Questions:

Can you share links for your MATLAB and R code?
The data and code referenced are available from Tom at grothues@marine.rutgers.edu.