SEACAR code review

Marcus W. Beck, [mbeck@tbep.org](mailto:mbeck@tbep.org)

2021-10-15 12:54:29

Table of Contents

View as [HTML document](https://FloridaSEACAR.github.io/SEACAR/review.html)

# 1 Document overview

SEACAR is a collaborative process which involves local, state and federal natural resource managers, data providers, researchers and partners to identify and assess ecological indicators and to develop a decision support tool to better understand the status of aquatic resources throughout the Office of Resilience and Coastal Protection managed areas. SEACAR staff are working with the University of South Florida (USF) Water Institute to create an online platform for data analysis and reporting to support science-based decisions of coastal resources in Florida. Analysis routines have been created to import, format, and summarize status and trends for various metrics of coastal condition reported by habitat, indicators, multiple parameters per indicator, and across 47 management areas. Primary habitats include water column, submerged aquatic vegetation, coral reefs, oyster reefs, and coastal wetlands.

To date, several scripts written using the R statistical programming language have been produced. This document is a review of these analysis scripts, as currently delivered in two separate GitHub repositories (available at <https://github.com/FloridaSEACAR>). This review includes a description of the purpose of each script to ensure the outputs are as expected based on preliminary goals established for the project. In general, status and trends of key indicators is to be quantified using conventional, non-parametric methods following the seasonal Kendall tau test that assesses direction and magnitude of a change within a specified period of time. Minimum requirements for the trend tests for each indicator include at least five years of data with at least two months in common across at least two consecutive years for each managed area. In addition, linear mixed-effects models have been used to assess trends in seagrass coverage.

This review is delivered as a bulleted outline for each script describing the objective, required inputs, processing steps, and output (e.g., graphic, summary table). The tests run or summaries created are described. The extent to which each indicator fulfills the requirements for the applicable trend test is also assessed, e.g., by describing available years of data or if data are insufficient. Each script is also reviewed for coding best practices to identify areas that can be improved or shortened for efficiency and to identify potential mistakes or errors in calculation. Specific lines are referenced as needed to identify areas in need of improvement or further review. Outputs, such as graphics or tabular summaries, are also reviewed to identify potential improvements in future versions of the code.

# 2 Comments by habitat type

## 2.1 Water column

### Scripts:

|  |  |  |
| --- | --- | --- |
| Index | File | Link |
| 1 | 2021.08.21\_FINAL\_SEACAR Water\_Nutrients\_Phosphorus\_FINAL.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.21_FINAL_SEACAR%20Water_Nutrients_Phosphorus_FINAL.R) |
| 2 | 2021.08.21\_FINAL\_SEACAR\_Water Clarity\_CDOM\_Final.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.21_FINAL_SEACAR_Water%20Clarity_CDOM_Final.R) |
| 3 | 2021.08.21\_FINAL\_SEACAR\_Water Clarity\_chla\_Final.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.21_FINAL_SEACAR_Water%20Clarity_chla_Final.R) |
| 4 | 2021.08.21\_FINAL\_SEACAR\_Water Clarity\_Turbidity\_Final.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.21_FINAL_SEACAR_Water%20Clarity_Turbidity_Final.R) |
| 5 | 2021.08.21\_FINAL\_SEACAR\_Water Nutrients\_Nitrogen\_Final.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.21_FINAL_SEACAR_Water%20Nutrients_Nitrogen_Final.R) |
| 6 | 2021.08.22\_FINAL SEACAR\_WaterQuality\_DOsaturation\_Final.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.22_FINAL%20SEACAR_WaterQuality_DOsaturation_Final.R) |
| 7 | 2021.08.22\_FINAL\_SEACAR\_WaterQuality\_DOmgL\_Final.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.22_FINAL_SEACAR_WaterQuality_DOmgL_Final.R) |
| 8 | 2021.08.22\_FINAL\_SEACAR\_WaterQuality\_pH\_Final.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.22_FINAL_SEACAR_WaterQuality_pH_Final.R) |
| 9 | 2021.08.22\_FINAL\_SEACAR\_WaterQuality\_Salinity.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.22_FINAL_SEACAR_WaterQuality_Salinity.R) |
| 10 | 2021.08.22\_FINAL\_SEACAR\_WATERQUALITY\_Temperature\_Final.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.22_FINAL_SEACAR_WATERQUALITY_Temperature_Final.R) |
| 11 | FINAL 2021.09.19\_WaterNekton\_SppRichnessbyMA.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL%202021.09.19_WaterNekton_SppRichnessbyMA.R) |

### Summaries:

1. 2021.08.21\_FINAL\_SEACAR Water\_Nutrients\_Phosphorus\_FINAL.R
   * Objective: Import and format phosphorus data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: Nutrients - Nitrogen and Phosphorus-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 52: load libraries and import phosphorus and regions datasets
     + Lines 53 - 187: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data). NA values are removed for water quality measurements and dates.
     + Lines 188 - 206: boxplot summaries of surface phosphorus by year, year/month for complete dataset
     + Lines 208 - 222: boxplot summaries of surface phosphorus by year/month, separate for managed areas
     + Lines 223 - 239: boxplot summaries of surface phosphorus by year, separate for managed areas
     + Lines 240 - 264: average phosphorus by year, month, managed area to prep for trend test
     + Lines 265 - 322: identify number of unique years in each dataset, remove those with < 10
     + Lines 323 - 349: plots of average year, month phosphorus, separate for managed areas
     + Lines 350 - 368: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 369 - 440: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Nutrients\_P\_MA\_surf\_MonthlyBoxplots.pdf, Nutrients\_P\_MA\_surf\_AnnualBoxplots.pdf, 2Nutrients\_P\_MA\_surf\_plots.pdf, Nutrients\_P\_MA\_surf\_KToutput.csv, Nutrients\_P\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + Line 41: object regions called before import
     + Line 202 (and others): boxplot summaries should be on log-scale
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Values at MDL are not removed, these need to be treated separately before running trend tests, i.e., estimate averages using appropriate statistics that account for left-censored data.
2. 2021.08.21\_FINAL\_SEACAR\_Water Clarity\_CDOM\_Final.R
   * Objective: Import and format CDOM data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: CDOM and FDOM-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 49: load libraries and import CDOM and regions datasets
     + Lines 50 - 186: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data). NA values are removed for water quality measurements and dates.
     + Lines 187 - 205: boxplot summaries of surface CDOM by year, year/month for complete dataset
     + Lines 206 - 216: boxplot summaries of surface CDOM by year/month, separate for managed areas
     + Lines 217 - 238: boxplot summaries of surface CDOM by year, separate for managed areas
     + Lines 239 - 262: average CDOM by year, month, managed area to prep for trend test
     + Lines 263 - 320: identify number of unique years in each dataset, remove those with < 10
     + Lines 321 - 347: plots of average year, month CDOM, separate for managed areas
     + Lines 350 - 366: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 367 - 438: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Clarity\_CDOM\_MA\_surf\_MonthlyBoxplots.pdf, Clarity\_CDOM\_MA\_surf\_AnnualBoxplots.pdf, 2Clarity\_CDOM\_MA\_surf\_plots.pdf, Clarity\_CDOM\_MA\_surf\_KToutput.csv, Clarity\_CDOM\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + Line 201 (and others): boxplot summaries should be on log-scale
     + Line 269: object name has a reserved character (-) and cannot be created
     + Line 359: object name has a reserved character (-) and cannot be created
     + Line 371: object name has a reserved character (-) and cannot be created
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Values at MDL are not removed, these need to be treated separately before running trend tests, i.e., estimate averages using appropriate statistics that account for left-censored data.
3. 2021.08.21\_FINAL\_SEACAR\_Water Clarity\_chla\_Final.R
   * Objective: Import and format chlorophyll data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: Chlorophyll-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 49: load libraries and import chlorophyll and regions datasets
     + Lines 50 - 185: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data). NA values are removed for water quality measurements and dates.
     + Lines 186 - 205: boxplot summaries of surface chlorophyll by year, year/month for complete dataset
     + Lines 206 - 220: boxplot summaries of surface chlorophyll by year/month, separate for managed areas
     + Lines 221 - 237: boxplot summaries of surface chlorophyll by year, separate for managed areas
     + Lines 238 - 261: average chlorophyll by year, month, managed area to prep for trend test
     + Lines 262 - 319: identify number of unique years in each dataset, remove those with < 10
     + Lines 320 - 346: plots of average year, month chlorophyll, separate for managed areas
     + Lines 347 - 365: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 366 - 437: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Clarity\_chla\_MA\_surf\_MonthlyBoxplots.pdf, Clarity\_chla\_MA\_surf\_AnnualBoxplots.pdf, 2Clarity\_chla\_MA\_surf\_plots.pdf, Clarity\_chla\_MA\_surf\_KToutput.csv, Clarity\_chla\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + Line 205 (and others): boxplot summaries should be on log-scale
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Values at MDL are not removed, these need to be treated separately before running trend tests, i.e., estimate averages using appropriate statistics that account for left-censored data.
4. 2021.08.21\_FINAL\_SEACAR\_Water Clarity\_Turbidity\_Final.R
   * Objective: Import and format turbidity data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: Turbidity and TSS-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 50: load libraries and import turbidity and regions datasets
     + Lines 51 - 186: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data). NA values are removed for water quality measurements and dates.
     + Lines 187 - 205: boxplot summaries of surface turbidity by year, year/month for complete dataset
     + Lines 206 - 221: boxplot summaries of surface turbidity by year/month, separate for managed areas
     + Lines 222 - 238: boxplot summaries of surface turbidity by year, separate for managed areas
     + Lines 239 - 262: average turbidity by year, month, managed area to prep for trend test
     + Lines 263 - 320: identify number of unique years in each dataset, remove those with < 10
     + Lines 321 - 347: plots of average year, month turbidity, separate for managed areas
     + Lines 348 - 366: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 367 - 438: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Clarity\_Turbidity\_MA\_surf\_MonthlyBoxplots.pdf, Clarity\_Turbidity\_MA\_surf\_AnnualBoxplots.pdf, 2Clarity\_Turbidity\_MA\_surf\_plots.pdf, Clarity\_Turbidity\_MA\_surf\_KToutput.csv, Clarity\_Turbidity\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + Line 201 (and others): boxplot summaries should be on log-scale
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Values at MDL are not removed, these need to be treated separately before running trend tests, i.e., estimate averages using appropriate statistics that account for left-censored data.
5. 2021.08.21\_FINAL\_SEACAR\_Water Nutrients\_Nitrogen\_Final.R
   * Objective: Import and format nitrogen data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: Nitrogen-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 48: load libraries and import nitrogen and regions datasets
     + Lines 49 - 219: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data). NA values are removed for water quality measurements and dates. Creates TN as sum of TKN and NOx. Values less than zero are flagged as 999999999, which is different from the other water quality variables.
     + Lines 220 - 239: boxplot summaries of surface nitrogen by year, year/month for complete dataset
     + Lines 240 - 255: boxplot summaries of surface nitrogen by year/month, separate for managed areas
     + Lines 256 - 272: boxplot summaries of surface nitrogen by year, separate for managed areas
     + Lines 273 - 296: average nitrogen by year, month, managed area to prep for trend test
     + Lines 297 - 354: identify number of unique years in each dataset, remove those with < 10
     + Lines 355 - 381: plots of average year, month nitrogen, separate for managed areas
     + Lines 382 - 400: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 401 - 472: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Nutrients\_N\_MA\_surf\_MonthlyBoxplots.pdf, Nutrients\_N\_MA\_surf\_AnnualBoxplots.pdf, 2Nutrients\_N\_MA\_surf\_plots.pdf, Nutrients\_N\_MA\_surf\_KToutput.csv, Nutrients\_N\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + Line 115: Date conversion is incorrect, creates all NA preventing downstream analysis
     + Line 235 (and others): boxplot summaries should be on log-scale
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Values at MDL are not removed, these need to be treated separately before running trend tests, i.e., estimate averages using appropriate statistics that account for left-censored data.
     + Many of the plots of the observed or averaged data are suspect, with large jumps in values between years, possibly related to changing lab methods or incorrect sum to get TN.
6. 2021.08.22\_FINAL SEACAR\_WaterQuality\_DOsaturation\_Final.R
   * Objective: Import and format dissolved oxygen saturation data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: Dissolved Oxygen-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 49: load libraries and import dissolved oxygen saturation and regions datasets
     + Lines 50 - 193: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data)
     + Lines 194 - 212: boxplot summaries of surface dissolved oxygen saturation by year, year/month for complete dataset
     + Lines 213 - 228: boxplot summaries of surface dissolved oxygen saturation by year/month, separate for managed areas
     + Lines 229 - 245: boxplot summaries of surface dissolved oxygen saturation by year, separate for managed areas
     + Lines 246 - 269: average dissolved oxygen saturation by year, month, managed area to prep for trend test
     + Lines 270 - 327: identify number of unique years in each dataset, remove those with < 10
     + Lines 328 - 354: plots of average year, month dissolved oxygen saturation, separate for managed areas
     + Lines 355 - 373: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 374 - 445: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Quality\_DOsat\_MA\_surf\_MonthlyBoxplots.pdf, Quality\_DOsat\_MA\_surf\_AnnualBoxplots.pdf, 2Quality\_DOsat\_MA\_surf\_plots.pdf, Quality\_DOsat\_MA\_surf\_KToutput.csv, Quality\_DOsat\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Many of the plots of the observed or averaged data are suspect, with large jumps in values between years.
7. 2021.08.22\_FINAL\_SEACAR\_WaterQuality\_DOmgL\_Final.R
   * Objective: Import and format dissolved oxygen concentration data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: Dissolved Oxygen-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 49: load libraries and import dissolved oxygen concentration and regions datasets
     + Lines 50 - 193: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data)
     + Lines 194 - 212: boxplot summaries of surface dissolved oxygen concentration by year, year/month for complete dataset
     + Lines 213 - 228: boxplot summaries of surface dissolved oxygen concentration by year/month, separate for managed areas
     + Lines 229 - 245: boxplot summaries of surface dissolved oxygen concentration by year, separate for managed areas
     + Lines 246 - 269: average dissolved oxygen concentration by year, month, managed area to prep for trend test
     + Lines 270 - 327: identify number of unique years in each dataset, remove those with < 10
     + Lines 328 - 354: plots of average year, month dissolved oxygen concentration, separate for managed areas
     + Lines 355 - 373: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 374 - 445: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Quality\_DOmgL\_MA\_surf\_MonthlyBoxplots.pdf, Quality\_DOmgL\_MA\_surf\_AnnualBoxplots.pdf, 2Quality\_DOmgL\_MA\_surf\_plots.pdf, Quality\_DOmgL\_MA\_surf\_KToutput.csv, Quality\_DOmgL\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Many of the plots of the observed or averaged data are suspect, with large jumps in values between years.
8. 2021.08.22\_FINAL\_SEACAR\_WaterQuality\_pH\_Final.R
   * Objective: Import and format pH data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: Water Temperature and pH-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 49: load libraries and import pH and regions datasets
     + Lines 50 - 193: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data)
     + Lines 194 - 212: boxplot summaries of surface pH by year, year/month for complete dataset
     + Lines 213 - 228: boxplot summaries of surface pH by year/month, separate for managed areas
     + Lines 229 - 245: boxplot summaries of surface pH by year, separate for managed areas
     + Lines 246 - 269: average pH by year, month, managed area to prep for trend test
     + Lines 270 - 327: identify number of unique years in each dataset, remove those with < 10
     + Lines 328 - 354: plots of average year, month pH, separate for managed areas
     + Lines 355 - 373: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 374 - 445: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Quality\_pH\_MA\_surf\_MonthlyBoxplots.pdf, Quality\_pH\_MA\_surf\_AnnualBoxplots.pdf, 2Quality\_pH\_MA\_surf\_plots.pdf, Quality\_pH\_MA\_surf\_KToutput.csv, Quality\_pH\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + The pH file is 1Gb, not sure this is a huge issue but import time is longer than most files.
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Many of the plots of the observed or averaged data are suspect, with large jumps in values between years.
9. 2021.08.22\_FINAL\_SEACAR\_WaterQuality\_Salinity.R
   * Objective: Import and format salinity data, create summary boxplots, evaluate trends with seasonal Kendall tests.
   * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
   * File inputs: Salinity and Specific Conductivity-2021-Jul-26.csv, SEACARMARegionList.csv
   * Steps by line number:
     + Lines 1 - 49: load libraries and import salinity and regions datasets
     + Lines 50 - 193: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data)
     + Lines 194 - 212: boxplot summaries of surface salinity by year, year/month for complete dataset
     + Lines 213 - 228: boxplot summaries of surface salinity by year/month, separate for managed areas
     + Lines 229 - 245: boxplot summaries of surface salinity by year, separate for managed areas
     + Lines 246 - 269: average salinity by year, month, managed area to prep for trend test
     + Lines 270 - 327: identify number of unique years in each dataset, remove those with < 10
     + Lines 328 - 354: plots of average year, month salinity, separate for managed areas
     + Lines 355 - 373: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
     + Lines 374 - 445: format all trend tests in a simple table, combine with region names, save output as csv
   * File outputs: Quality\_Salinity\_MA\_surf\_MonthlyBoxplots.pdf, Quality\_Salinity\_MA\_surf\_AnnualBoxplots.pdf, 2Quality\_Salinity\_MA\_surf\_plots.pdf, Quality\_Salinity\_MA\_surf\_KToutput.csv, Quality\_Salinity\_MA\_surf\_resultssummarystats2.csv
   * Potential issues:
     + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
     + Many of the plots of the observed or averaged data are suspect, with large jumps in values between years.
10. 2021.08.22\_FINAL\_SEACAR\_WATERQUALITY\_Temperature\_Final.R
    * Objective: Import and format temperature data, create summary boxplots, evaluate trends with seasonal Kendall tests.
    * Packages: dplyr, EnvStats, gt, tidymodels, tidyr, tidyverse
    * File inputs: Water Temperature and pH-2021-Jul-26.csv, SEACARMARegionList.csv
    * Steps by line number:
      + Lines 1 - 49: load libraries and import temperature and regions datasets
      + Lines 50 - 193: Format data for analysis, including removal of NA values, conversion of date column to date objects, and removing entries with flagged QA codes (from WIN and NERRS data)
      + Lines 194 - 212: boxplot summaries of surface temperature by year, year/month for complete dataset
      + Lines 213 - 228: boxplot summaries of surface temperature by year/month, separate for managed areas
      + Lines 229 - 245: boxplot summaries of surface temperature by year, separate for managed areas
      + Lines 246 - 269: average temperature by year, month, managed area to prep for trend test
      + Lines 270 - 327: identify number of unique years in each dataset, remove those with < 10
      + Lines 328 - 354: plots of average year, month temperature, separate for managed areas
      + Lines 355 - 373: run trend tests, first on a “test” site then for all, results saved as raw output from the trend test
      + Lines 374 - 445: format all trend tests in a simple table, combine with region names, save output as csv
    * File outputs: Quality\_Temperature\_MA\_surf\_MonthlyBoxplots.pdf, Quality\_Temperature\_MA\_surf\_AnnualBoxplots.pdf, 2Quality\_Temperature\_MA\_surf\_plots.pdf, Quality\_Temperature\_MA\_surf\_KToutput.csv, Quality\_Temperature\_MA\_surf\_resultssummarystats2.csv
    * Potential issues:
      + The temperature file is 1Gb, not sure this is a huge issue but import time is longer than most files.
      + NA values were also removed for dates that were incomplete. This may possibly be corrected through additional formatting but is not done here.
      + Many of the plots of the observed or averaged data are suspect, with large jumps in values between years.
11. FINAL 2021.09.19\_WaterNekton\_SppRichnessbyMA.R
    * Objective: Import and format nekton data, calculate richness, assess trends with linear mixed effects models.
    * Packages: dplyr, gt, nlme, RColorBrewer, tidyverse
    * File inputs: Nekton\_NE-2021-Jul-26.csv
    * Steps by line number:
      + Lines 1 - 19: load libraries and import nekton data
      + Lines 20 - 30: combine genus, species columns as one, remove any rows with missing genus or management area names
      + Lines 31 - 52: calculate richness metrics management area, region, year
      + Lines 53 - 69: identify number of unique years in each dataset, remove those with < 5
      + Lines 70 - 86: create linear mixed models evaluating richness vs year, using a random region effect
      + Lines 87 - 113: create plot of richness vs year by management area using model output
      + Lines 114 - 141: tabular summary of richness vs year by management area
      + Lines 142 - 166: tabular summary of richness vs year by management area, similar info as previous table
      + Lines 167 - 200: join results from the previous two tables, save output as csv
    * File outputs: Nekton\_SppRichnessbyMAmodel.csv, Nekton\_SppRichnessbyMAmodelPlot.pdf, Nekton\_SppRichnessbyMAmodelTable.pdf.csv
    * Potential issues:
      + Line 38: NA values are incorrectly identified, should do this with a regular expression search.
      + If the goal is to evaluate diversity, this was not done. Only richness was evaluated.
      + There are redundant models created for the plots and tables. These should only be created once and recycled as needed.
      + There are some outlier values for one of the managed areas in the species richness vs year plot that should be verified.

## 2.2 Coral reef

### Scripts:

|  |  |  |
| --- | --- | --- |
| Index | File | Link |
| 1 | 2021.09.19\_CoralNektonRichness\_FINAL.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.09.19_CoralNektonRichness_FINAL.R) |
| 2 | FINAL\_2021.09.18\_DryTort Region\_Any Corals.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_DryTort%20Region_Any%20Corals.R) |
| 3 | FINAL\_2021.09.18\_DryTortugas Region\_Genus.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_DryTortugas%20Region_Genus.R) |
| 4 | FINAL\_2021.09.18\_DryTortugas Region\_Group1.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_DryTortugas%20Region_Group1.R) |
| 5 | FINAL\_2021.09.18\_FLAKEYS\_AnyCorals.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_FLAKEYS_AnyCorals.R) |
| 6 | FINAL\_2021.09.18\_FLKeysRegion\_Group1.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_FLKeysRegion_Group1.R) |
| 7 | FINAL\_2021.09.18\_Keys Region\_Genus.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_Keys%20Region_Genus.R) |
| 8 | FINAL\_2021.09.18\_SE Region\_AnyCorals.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_SE%20Region_AnyCorals.R) |
| 9 | FINAL\_2021.09.18\_SE\_Genus.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_SE_Genus.R) |
| 10 | FINAL\_2021.09.18\_SERegion\_Group1.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL_2021.09.18_SERegion_Group1.R) |

### Summaries:

1. 2021.09.19\_CoralNektonRichness\_FINAL.R
   * Objective: .
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: CorSE.csv, CorFLKeys.csv, CorDryTortugas.csv, Nekton2.csv
   * Steps by line number:
     + Lines 1 - :
   * File outputs: CoralNektonlmeSERegion.csv, CoralNektonlmebyMA.csv, NektonCoral\_SppRichnessbyMAmodelPlot.pdf, NektonCoral\_SppRichnessbyMAmodelTable.csv
   * Potential issues:
     + Requires data inputs created in files FINAL 2021.09.18\_SE Region\_AnyCorals.R, FINAL\_2021.09.18\_FLAKEYS\_AnyCorals.R, FINAL\_2021.09.18\_DryTort Region\_Any Corals.R. Could not locate the file Nekton2.csv.
2. FINAL\_2021.09.18\_DryTort Region\_Any Corals.R
   * Objective: Import and format percent cover data, create summary boxplots, evaluate percent cover trends for any corals across years with linear models.
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: Percent Cover - DRY TORT-2021-Jul-26.csv
   * Steps by line number:
     + Lines 1 - 20: load libraries and import percent cover data for Dry Tortugas
     + Lines 21 - 79: format data for analysis including removal of non-coral taxa, renaming columns, removing NA percent cover data, combining genus/species columns, and removing missing rows for genus/species
     + Lines 80 - 101: create a summary table for percent cover statistics by year, region across all species, save output
     + Lines 102 - 112: create linear model year vs percent cover, save model summary plots to pdf
     + Lines 113 - 128: create boxplots of percent cover for any coral by year, save as pdf
     + Lines 129 - 145: create linear mixed effects model of percent cover for any coral vs year, random effect for program location ID
     + Lines 146 - 179: create linear mixed effects model of percent cover for any coral vs year, random effect for program location ID, separate for coral region, save output as csv
     + Lines 180 - 202: plot model output as estimated linear trend across years by coral region, save as pdf
     + Lines 203 - 301: format and summarize model output to create summary table of model fit, saved as csv
   * File outputs: CoralSummaryStatisticsTable\_AnyCoral\_DRYTORTUGASRegion.csv, CoralSummaryStatisticsPlots\_AnyCoral\_DRYTORTUGASRegion.pdf, CoralBoxplots\_AnyCoral\_DRYTORTUGASRegion.pdf, CorDryTortugas.csv, Corallme\_AnyCoral\_DRYTORTUGASRegion.csv, CorallmePlot\_AnyCoral\_DRYTORTUGASRegion.pdf, Coral\_lmeResults\_AnyCoral\_DRYTORTUGASRegion.csv
   * Potential issues:
     + Lines 106, 109: Year is dependent variable in the linear model, it should be the independent variable.
     + Model results are separate for each coral region, yet the file includes only one region (Dry Tortugas). This does not create incorrect output, but the intention is unclear.
3. FINAL\_2021.09.18\_DryTortugas Region\_Genus.R
   * Objective: Import and format percent cover data, create summary boxplots, evaluate percent cover trends by genus across years with linear models.
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: Percent Cover - DRY TORT-2021-Jul-26.csv
   * Steps by line number:
     + Lines 1 - 18: load libraries and import percent cover data for Dry Tortugas
     + Lines 19 - 83: format data for analysis including removal of non-coral taxa, renaming columns, removing NA percent cover data, combining genus/species columns, and removing missing rows for genus/species
     + Lines 84 - 103: create a summary table for percent cover statistics by year, genus, region, save output
     + Lines 104 - 110: create linear model year vs percent cover, not separated by genus
     + Lines 111 - 125: create boxplots of percent cover by genus, year, save as pdf
     + Lines 126 - 161: remove genera with less than ten years of data, create linear mixed effects model of percent cover for all genera vs year, random effect for program location ID. No output created because no genus meets the minimum year requirement.
     + Lines 162 - 184: create linear mixed effects model of percent cover for each genus vs year, random effect for program location ID, save output as csv. No output created because no genus meets the minimum year requirement.
     + Lines 185 - 210: plot model output as estimated linear trend across years by genus, save as pdf. No output created because no genus meets the minimum year requirement.
     + Lines 211 - 309: format and summarize model output to create summary table of model fit, saved as csv. No output created because no genus meets the minimum year requirement.
   * File outputs: CoralSummaryStatisticsTable\_Genus\_DRYRegion.csv, CoralBoxplots\_ByGenus\_DRYRegion.pdf, Corallme\_ByGenus\_DRYRegion.csv, CorallmePlot\_ByGenus\_DRYRegion.pdf, Coral\_lmeResults\_ByGenus\_DRYRegion.csv
   * Potential issues:
     + Lines 108: Year is dependent variable in the linear model, it should be the independent variable.
     + No output created because no genus meets the minimum year requirement.
4. FINAL\_2021.09.18\_DryTortugas Region\_Group1.R
   * Objective: Import and format percent cover data, create summary boxplots, evaluate percent cover trends by species group across years with linear models.
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: Percent Cover - DRY TORT-2021-Jul-26.csv
   * Steps by line number:
     + Lines 1 - 17: load libraries and import percent cover data for Dry Tortugas
     + Lines 18 - 81: format data for analysis including removal of non-coral taxa, renaming columns, removing NA percent cover data, combining genus/species columns, and removing missing rows for genus/species
     + Lines 82 - 101: create a summary table for percent cover statistics by year, species group, region, save output
     + Lines 102 - 107: create linear model year vs percent cover, not separated by species group
     + Lines 108 - 122: create boxplots of percent cover by species group, year, save as pdf
     + Lines 123 - 158: create linear mixed effects model of percent cover for each species groups vs year, random effect for program location ID
     + Lines 159 - 181: plot model output as estimated linear trend across years by species group, save as pdf
     + Lines 182 - 280: format and summarize model output to create summary table of model fit, saved as csv
   * File outputs: CoralSummaryStatisticsTable\_Group1\_DryTortugasRegion.csv, CoralBoxplots\_ByGroup1\_DRYRegion.pdf, Corallme\_ByGroup1\_DRYRegion.csv, CorallmePlot\_ByGroup1\_DRYRegion.pdf, Coral\_lmeResults\_ByGroup1\_DryTortugasRegion.csv
   * Potential issues:
     + Lines 106: Year is dependent variable in the linear model, it should be the independent variable.
5. FINAL\_2021.09.18\_FLAKEYS\_AnyCorals.R
   * Objective: Import and format percent cover data, create summary boxplots, evaluate percent cover trends for any corals across years with linear models.
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: Percent Cover - FLA KEYS-2021-Jul-26.csv
   * Steps by line number:
     + Lines 1 - 20: load libraries and import percent cover data for Florida Keys
     + Lines 21 - 79: format data for analysis including removal of non-coral taxa, renaming columns, removing NA percent cover data, combining genus/species columns, and removing missing rows for genus/species
     + Lines 80 - 101: create a summary table for percent cover statistics by year, region across all species, save output
     + Lines 102 - 112: create linear model year vs percent cover, save model summary plots to pdf
     + Lines 113 - 127: create boxplots of percent cover for any coral by year, save as pdf
     + Lines 128 - 144: create linear mixed effects model of percent cover for any coral vs year, random effect for program location ID
     + Lines 145 - 178: create linear mixed effects model of percent cover for any coral vs year, random effect for program location ID, separate for coral region, save output as csv
     + Lines 179 - 201: plot model output as estimated linear trend across years by coral region, save as pdf
     + Lines 202 - 300: format and summarize model output to create summary table of model fit, saved as csv
   * File outputs: CoralSummaryStatisticsTable\_AnyCoral\_FLKEYSRegion.csv, CoralSummaryStatisticsPlots\_AnyCoral\_FLKEYSRegion.pdf, CorFLKeys.csv, CoralBoxplots\_AnyCoral\_FLKEYSRegion.pdf, Corallme\_AnyCoral\_FLKEYSRegion.csv, CorallmePlot\_AnyCoral\_FLKEYSRegion.pdf, Coral\_lmeResults\_AnyCoral\_FLKEYSRegion.csv
   * Potential issues:
     + Lines 106, 109: Year is dependent variable in the linear model, it should be the independent variable.
     + Model results are separate for each coral region, yet the file includes only one region (Florida Keys). This does not create incorrect output, but the intention is unclear.
6. FINAL\_2021.09.18\_Keys Region\_Genus.R
   * Objective: Import and format percent cover data, create summary boxplots, evaluate percent cover trends by genus across years with linear models.
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: Percent Cover - FLA KEYS-2021-Jul-26.csv
   * Steps by line number:
     + Lines 1 - 17: load libraries and import percent cover data for Florida Keys
     + Lines 18 - 82: format data for analysis including removal of non-coral taxa, renaming columns, removing NA percent cover data, combining genus/species columns, and removing missing rows for genus/species
     + Lines 83 - 102: create a summary table for percent cover statistics by year, genus, region, save output
     + Lines 103 - 109: create linear model year vs percent cover, not separated by genus
     + Lines 110 - 124: create boxplots of percent cover by genus, year, save as pdf
     + Lines 125 - 160: remove genera with less than ten years of data, create linear mixed effects model of percent cover for all genera vs year, random effect for program location ID
     + Lines 161 - 183: create linear mixed effects model of percent cover for each genus vs year, random effect for program location ID, save output as csv
     + Lines 184 - 209: plot model output as estimated linear trend across years by genus, save as pdf
     + Lines 210 - 308: format and summarize model output to create summary table of model fit, saved as csv
   * File outputs: CoralSummaryStatisticsTable\_Genus\_KEYSRegion.csv, CoralBoxplots\_ByGenus\_KEYSRegion.pdf, Corallme\_ByGenus\_KEYSRegion.csv, CorallmePlot\_ByGenus\_KEYSRegion.pdf, Coral\_lmeResults\_ByGenus\_KEYSRegion.csv
   * Potential issues:
     + Lines 107: Year is dependent variable in the linear model, it should be the independent variable.
7. FINAL\_2021.09.18\_SE Region\_AnyCorals.R
   * Objective: .
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: Percent Cover - SE FL-2021-Jul-26.csv
   * Steps by line number:
     + Lines 1 - :
   * File outputs: CoralSummaryStatisticsTable\_AnyCoral\_SERegion.csv, CoralSummaryStatisticsPlots\_AnyCoral\_SERegion.pdf, CorSE.csv, CoralBoxplots\_AnyCoral\_SERegion.pdf, Corallme\_AnyCoral\_SERegion.csv, CorallmePlot\_AnyCoral\_SERegion.pdf, Coral\_lmeResults\_AnyCoral\_SERegion.csv
   * Potential issues:
8. FINAL\_2021.09.18\_SE\_Genus.R
   * Objective: .
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: Percent Cover - FLA SE-2021-Jul-26.csv
   * Steps by line number:
     + Lines 1 - :
   * File outputs: CoralSummaryStatisticsTable\_Genus\_SERegion.csv, CoralBoxplots\_ByGenus\_SERegion.pdf, Corallme\_ByGenus\_SERegion.csv, CorallmePlot\_ByGenus\_SERegion.pdf, Coral\_lmeResults\_ByGenus\_SERegion.csv
   * Potential issues:
     + Could not locate the input file Percent Cover - FLA SE-2021-Jul-26.csv
9. FINAL\_2021.09.18\_SERegion\_Group1.R
   * Objective: .
   * Packages: dplyr, gt, nlme, tidyverse
   * File inputs: Percent Cover - SE FL-2021-Jul-26.csv
   * Steps by line number:
     + Lines 1 - :
   * File outputs: CoralSummaryStatisticsTable\_Group1\_SERegion.csv, CoralBoxplots\_ByGroup1\_SERegion.pdf, Corallme\_ByGroup1\_SERegion.csv, CorallmePlot\_ByGroup1\_SERegion.pdf, Coral\_lmeResults\_ByGroup1\_SERegion.csv
   * Potential issues:

## 2.3 Coastal wetlands

### Scripts:

|  |  |  |
| --- | --- | --- |
| Index | File | Link |
| 1 | Final 2021.09.19\_SEACAR\_CoastalWetlands\_GroupRichness.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/Final%202021.09.19_SEACAR_CoastalWetlands_GroupRichness.R) |
| 2 | FINAL 2021.09.19\_SEACAR\_CoastalWetlands\_SppRich\_Region.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL%202021.09.19_SEACAR_CoastalWetlands_SppRich_Region.R) |

### Summaries:

1. Final 2021.09.19\_SEACAR\_CoastalWetlands\_GroupRichness.R
2. FINAL 2021.09.19\_SEACAR\_CoastalWetlands\_SppRich\_Region.R

## 2.4 Submered aquatic vegetation

### Scripts:

|  |  |  |
| --- | --- | --- |
| Index | File | Link |
| 1 | FINAL Percent Cover by Secchi.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL%20Percent%20Cover%20by%20Secchi.R) |
| 2 | FINAL Percent Cover by Turbidity.R | [link](https://github.com/FloridaSEACAR/SEACAR/blob/master/FINAL%20Percent%20Cover%20by%20Turbidity.R) |

### Summaries:

1. FINAL Percent Cover by Secchi.R
2. FINAL Percent Cover by Turbidity.R

# 3 General comments

* Leverage more tools from the tidyverse for more concise, streamline coding. For example, use case\_when instead of nested ifelse statements.
* Modularize the parts of copied code in functions. For example, the scripts for water quality variables are all very similar. Functions could be created to remove NAs, filter by QA codes, create summary plots, run trend tests, and format results.
* General issue with portability, environment for analysis is not isolated, i.e., scripts were created on a local computer and specific setup was unknown. Could be solved by package devevelopment, possibly combined with Docker container. Use of rm(list = ls() at the top of many scripts demonstrates the problem.
* Many scripts use install.packages() which is not necessary for running the scripts in production.
* Some libraries are loaded more than once (e.g., tidyverse, then dplyr which is included in tidyverse, as in lines 25-29 [here](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.21_FINAL_SEACAR%20Water_Nutrients_Phosphorus_FINAL.R)). This does not impact performance but creates challenges for tracking dependencies. All relevant libraries should be entered at the top of each script.
* Use of logical values entered as text strings (e.g., line 151 [here](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.21_FINAL_SEACAR%20Water_Nutrients_Phosphorus_FINAL.R)) is incorrect and should use the actual atomic values, i.e., TRUE, FALSE.
* Water quality trend tests are based on average year, month values. Is it possible to run these tests without averaging first? This reduces sample size and removes real variability that could influence the outcome.
* Month, year summary boxplots for water column variables are difficult to interpret as all values are chronological, maybe better to show just month on x-axis, then facet by year.
* Output files created with sink create a csv but this would be more appropriate as a text file since these are not tabular data (e.g., line 361 [here](https://github.com/FloridaSEACAR/SEACAR/blob/master/2021.08.21_FINAL_SEACAR_Water%20Clarity_chla_Final.R)).
* Formatted water column tables with trend statistics for water column data should have number of years for which the trend applies, in addition to min/max years. The time series for each managed area are all different and it’s useful to provide this information.
* Data I/O needs to be evaluated. All scripts use files from a local path and outputs are saved in the working directory. This is an issue to be solved in production.
* For corals, analyses were requested separately by sampling type: fixed or random. Note that this refers to sampling method, not a statistical test. However, the scripts indicate that sample information is unavailable in the input data (e.g., line 22 [here](https://github.com/FloridaSEACAR/SEACAR/blob/main/FINAL_2021.09.18_DryTort%20Region_Any%20Corals.R)). The analyses are not done at this time.
* For corals, trends in percent cover are assessed using linear mixed-effects models. It may be worth confirming trends using non-parametric analogs. There also is not a minimum year requirement for tests in percent cover evaluating any coral and for species groups, unlike trends tests for other habitats (i.e., water column). However, there is a minimum year requirement for tests evaluating genera.