Coordinated monitoring of the Piney Point wastewater discharge into Tampa Bay: Data synthesis and reporting

Marcus W. Beck, Tampa Bay Estuary Program, 263 13th Ave S, Saint Petersburg, FL, [mbeck@tbep.org](mailto:mbeck@tbep.org)

Maya C. Burke, Tampa Bay Estuary Program, 263 13th Ave S, Saint Petersburg, FL, [mburke@tbep.org](mailto:mburke@tbep.org)

Gary E. Raulerson, Tampa Bay Estuary Program, 263 13th Ave S, Saint Petersburg, FL, [graulerson@tbep.org](mailto:graulerson@tbep.org)

Sheila Scolaro, Tampa Bay Estuary Program, 263 13th Ave S, Saint Petersburg, FL, [sscolaro@tbep.org](mailto:sscolaro@tbep.org)

Edward T. Sherwood, Tampa Bay Estuary Program, 263 13th Ave S, Saint Petersburg, FL, [esherwood@tbep.org](mailto:esherwood@tbep.org)

Joe Whalen, Tampa Bay Estuary Program, 263 13th Ave S, Saint Petersburg, FL, [jwhalen@tbep.org](mailto:jwhalen@tbep.org)

From March 30th to April 9th, 2021, an estimated 215 million gallons of wastewater from the Piney Point phosphogypsum stacks were released into Tampa Bay to avoid catastrophic failure of the holding ponds. Ammonium concentrations in the wastewater were measured in excess of 200 mg/L and it was estimated that ~205 tons of total nitrogen were exported to Lower Tampa Bay, exceeding typical annual nitrogen load estimates in a matter of days. In response to these events, a coordinated environmental monitoring effort consisting of multiple government, university, and private sector partners began to assess conditions of surface waters around Piney Point to understand conditions prior to, during, and after the wastewater release. These efforts included sampling of surface water chemistry, surveys of algal community response, assessments of wastewater contaminants, and biological surveys of seagrass, macroalgae, benthic, and nekton communities. This resulted in many disparate data sets coming from multiple sources, which required the use of robust synthesis methods for assessment of current conditions, comparisons with the decades of baseline data available in Tampa Bay, and development of forward-facing reporting products to convey results of field sampling in near real time. This talk will discuss the open-source tools that were used to synthesize the environmental monitoring data, the online dashboard that was developed to report the data, and how these products were used to inform management response to rapidly changing environmental conditions. The specific challenges in combining data, both from a technical and philosophical perspective, will also be discussed as lessons learned from the Piney Point experience to inform future event-based monitoring responses, both in Tampa Bay and elsewhere.

Last manuscript build 2022-05-18 17:28:40

# Introduction

* Piney Point background
* Monitoring response
* Challenges with data synthesis and need for rapid assessment tools, TBEP commitment to OS

This paper describes the data synthesis tools that were used to inform partner response and public awareness of the 2021 Piney Point release into Tampa Bay, Florida. We focus specifically on the open science methods and products that were created to organize, contextualize, and communicate these data, including the Piney Point environmental monitoring dashboard developed by the Tampa Bay Estuary Program to rapidly deliver results for data collected in near real-time. A discussion of the technical and philosophical challenges of synthesizing data from multiple disparate sources, as well as the lessons learned, is also provided to inform future response-based monitoring efforts.

# Data collection and synthesis

Monitoring of bay resources in response to release from Piney Point began in late March of 2021 and continued through the summer and early fall. Several partners from public, private, and academic institutions conducted monitoring of water quality, phytoplankton (including the red tide organism *Karenia brevis*), seagrasses, macroalgae, and benthic habitats (Figure 1). Monitoring partners included the Florida Department of Environmental Protection (FDEP), Environmental Protection Commission (EPC) of Hillsborough County, Parks and Natural Resources Department of Manatee County, Pinellas County Division of Environmental Management, Fish and Wildlife Research Institute (FWRI) of the Florida Fish and Wildlife Conservation Commission (FWC), City of St. Petersburg, Tampa Bay Estuary Program (TBEP), Sarasota Bay Estuary Program, Environmental Science Associates, University of South Florida, University of Florida, and New College of Florida (Figure 1a). Coordination among monitoring agencies was facilitated by the TBEP and under guidance of a plume simulation model run by the Ocean Circulation Lab at the University of South Florida (USF), College of Marine Science ([Chen et al. 2018](#ref-Chen18), [2019](#ref-Chen19)). Initial monitoring was focused near the release site at Port Manatee (Figure 1), but later expanded throughout middle and lower Tampa Bay, Boca Ciega Bay, Manatee River, and Sarasota Bay as the simulation model predicted dispersal of the plume away from Piney Point ([Liu et al. 2021](#ref-Liu21)). Sampling effort was most intense in April, with less frequent sampling from May through September, although seagrass and macroalgae sampling was consistent throughout the study period (Table 1).

Water quality monitoring focused on parameters that were expected to indicate a response to elevated nutrient loads from the Piney Point release. Because Tampa Bay is a nitrogen-limited estuary ([Wang et al. 1999](#ref-Wang99); [Greening & Janicki 2006](#ref-Greening06); [Greening et al. 2014](#ref-Greening14)), water quality samples were taken for total nitrogen and inorganic forms (i.e., total ammonia nitrogen, nitrate/nitrite), chlorophyll-a as an indication of phytoplankton growth, and additional parameters expected to respond to nutrient loading (e.g., secchi depth, dissolved oxygen). Phytoplankton monitoring included both qualitative (presence/absence) and quantitative (cells/L) summaries of taxa enumerated by microscopy from water samples taken at similar frequency and distribution as the water quality sites. Phytoplankton data also included event-based sampling for *Karenia brevis*. Finally, seagrass and macroalgae data were collected by transect sampling at locations covering the shoreline north of Piney Point and extending to the south into northern Sarasota Bay (Figure 1b). Transect sampling included identification of seagrass species and macroalgae, typically to genus, and Braun-Blanquet abundance estimates at 10 m distances along each 50 m transect. Details on all sampling methods are provided in Beck et al. ([2022](#ref-Beck22)).

The synthesis of data from multiple partners included several challenges that TBEP staff addressed by employing open science principles within the time and resource limitations of the event. First, a centralized location for sharing data on Google Drive was required for the synthesis effort (Figure 2, file drops). Partner agencies were requested to voluntarily upload their data to a publicly accessible folders, or share via email with TBEP staff for upload if access issues to the main drive were encountered. Datasets were also downloaded by TBEP staff from source (e.g., FDEP, <https://floridadep.gov/dear/dear/content/tampa-bay-sampling-response-and-results>) and uploaded to the Google Drive. Using more formal data hosting services that include full metadata documentation was not possible while monitoring was ongoing given the volume and variety of data collected by partners. As such, Google Drive was a useful interim solution that provided 1) online sharing of data in different formats in a familiar file-based organization, and 2) an entry point for downstream data synthesis products created by TBEP. Changing user permissions also allowed read/write access to only those necessary (i.e., partner agencies for data upload), while allowing a read only option for all files so that source data were publicly available to view as needed.

A second synthesis challenge was unifying the multiple data types hosted on Google Drive into a usable format for the dashboard and additional analyses (Figure 2, data synthesis). Several folders on Google Drive were created that included relevant files for each type of monitoring data (e.g., water quality, phytoplankton, etc.). Each folder typically contained dozens of files saved as Google Sheets, often uploaded as a single sampling event by an individual monitoring agency. For example, the folder for the water quality sampling included 91 separate files. Extensive data “wrangling” was required to combine the files into a tidy format ([Wickham 2014](#ref-Wickham14c)) and a custom processing script was created by TBEP using the open source R Statistical Programming Language ([R Core Team 2021](#ref-RCT20)). An RStudio project was created with full version control that included a single script, dat\_proc.R, that imported all files from Google Drive for synthesis into individual binary .RData files for use with the dashboard (i.e., individual file for water quality, phytoplankton, etc.). Several R packages were indispensable in the data synthesis effort and these included 1) the dplyr, tidyr, lubridate, and readxl packages included with the tidyverse suite of packages ([Wickham et al. 2019](#ref-Wickham19)), and 2) the googlesheets4 ([Bryan 2020](#ref-Bryan20)) and googledrive ([D’Agostino McGowan & Bryan 2020](#ref-DAgostino20)) R packages for connecting and downloading files from Google Drive. Importantly, all source files were kept on Google Drive and were unaltered from their original format to isolate the original data from any downstream analysis.

Combining the source files on Google Drive into a synthesized format in R also required a consistent set of variables and naming conventions that were applied to each file. For example, parameters included in each water quality file were converted to a common set of names and all units were standardized to commonly used values (e.g., total nitrogen as tn and mg/L, chlorophyll-a as chla and g/L, etc.). This required developing custom processing code for data files from each partner to address each unique use case. As a result, the final processing file included almost 3000 lines of code to create the synthesized data products. Each partner agency was also assigned a unique identifier in the synthesized files so effort could be tracked and reported in the dashboard (e.g., Pinellas County as pinco, Florida Department of Environmental Protection as fldep, etc.). All synthesized files were saved in a long and tidy format, where each observation was given its own row, each variable was in its own column, and only one value was assigned to a cell ([Wickham 2014](#ref-Wickham14c)). This allowed for easily filtering observations of interest (as required with the dashboard) using the standardized naming conventions applied to each source file. More importantly, the developed workflow was used iteratively, often several times each week as new data files were uploaded to the Drive folders. This required executing the synthesis code locally and often making edits to the code to address new data formats or issues encountered when new files were uploaded.

The local workflow created by TBEP was hosted online in a GitHub repository (<https://github.com/tbep-tech/piney-point>, [Beck 2021](#ref-Beck21d)) that allowed use of additional resources that were critical to the synthesis effort (Figure 2, testing and sharing). Hosting the RStudio project on GitHub served multiple purposes of 1) using version control to document and track changes using Git, 2) making the code publicly accessible for transparency with the larger research and management community, 3) allowing automated testing of repository content, and 4) serving as an access point for deploying the dashboard on Shiny Server (see below). The use of automated testing was an important service to identify potential errors or mistakes in the synthesized datasets. The testthat R package ([Wickham 2011](#ref-Wickham11)) was used to write several checks for each dataset that were considered essential for maintaining integrity of the information. Each time changes were made to the repository and pushed to GitHub, tests were run automatically with GitHub Actions to verify each dataset included appropriate and accurate results. For example, a test was used to verify that all water quality parameters were appropriately named and did not include any missing values. A notification was sent via email to TBEP staff if an error was encountered. This automated service reduced the amount of time required to manually check the data and provided assurance that the reported results in the dashboard were correct.

# The Piney Point Dashboard

## Setup

The Piney Point environmental monitoring dashboard was created to communicate the results from the synthesized datasets (Figure 3, [Beck et al. 2021](#ref-Beck21c)). The dashboard was developed “on the fly” as data became available and the synthesis workflow described in the previous section was used to update input datasets during the course of the 2021 monitoring effort. All code used to synthesize the datasets from Google Drive and the code used to create the dashboard were contained in the same repository, allowing the workflow in Figure 2 to be used iteratively to update the content. The dashboard was created using the shiny ([Chang et al. 2021](#ref-Chang21)) and flexdashboard ([Iannone et al. 2020](#ref-Iannone20)) R packages to create the server and user interface components of the app. The flexdashboard package was particularly useful in creating an intuitive user interface to organize related components of the monitoring effort. The dashboard was deployed using Shiny Server by pulling the GitHub repository to a remote server maintained by TBEP each time an update was made (Figure 2, app deployment). This allowed anyone to access the dashboard using the url <https://shiny.tbep.org/piney-point>.

Both the server and user interface (UI) components of the dashboard were included in a single R Markdown ([Xie et al. 2020](#ref-Xie20)) file totaling 2346 lines of code, with 71% and 29% of the code dedicated to the server and UI, respectively. The server components of the dashboard included a setup component (346 lines, 15% of the total) and reactive component (1319, 56% of the total). The setup imported all R package dependencies (15 total), the synthesized input data from Google Drive (20 .RData files), an R script with custom functions to summarize or plot the results, five log files, and various “static” objects used in other parts of the dashboard that did not depend on user inputs. The rest of the server included thirty-one reactive objects (1319 lines, 56% of the total) that received different user inputs and returned a shiny data object sent back to the user interface. The imported log files were created during the data synthesis steps external to the dashboard and were specific to each data type (e.g., water quality, phytoplankton, etc.). Each log file was used to display when a data type was last changed, to inform users when updates were made to the input data.

The remainder of the R Markdown file was used to create the UI or the dashboard (681 lines, 29% of total). The flexdashboard package rendered with R Markdown was used to place shiny content in specific CSS elements, such as boxes or tabs that are scaled automatically to a user’s browser, to create a UI with easily navigable content. Specifically, the UI code organized the dashboard content into 1) a landing page when a user first accesses the website (Figure 3), 2) tabs for navigating to current results for data from a specific monitoring type, and 3) tabs for viewing historical baseline data for comparison with the response-based monitoring data. The landing page included a text overview of the dashboard, a map showing the monitoring locations, and several summary boxes at the bottom showing current effort to date (e.g., number of sites sampled, total days sampling, etc., Figure ). Each tab for the current data was organized specifically for each monitoring data type, although common elements included maps and tabular summaries where the displayed data could be chosen by date ranges or selected parameters. Finally, the baseline data tab displayed long-term monitoring data for water quality and seagrass to provide a comparison with the response-based monitoring data.

An important service provided by the dashboard was creating context for the response-based monitoring results relative to baseline conditions in Tampa Bay. This allowed dashboard users to quickly assess if the 2021 conditions were abnormal relative to typical seasonal patterns observed in the bay. In addition to the baseline data tabs, all water quality data from the response-based monitoring were compared to the long-term monthly averages for data at the nearest ambient monitoring station ([Beck et al. 2022](#ref-Beck22)). Elements of the UI and summary plots allowed a user to quickly determine if the results were outside of the expected range. Specifically, map points for a sampling event were displayed with a solid outline if results were out of range and without an outline if within the normal range. A user could also click on an individual map point to view a time series of the 2021 data for the selected location that included upper and lower limits for the long-term monthly averages. This functionality required additional modifications to the UI, in addition to developing custom code in the synthesis workflow that summarized the normal ranges prior to uploading data to the dashboard. Ultimately, these changes improved the quality of information provided by the response-based monitoring by leveraging the decades of long-term monitoring data available for the bay.

## Analytics

In addition to providing a centralized location for viewing results of the 2021 monitoring, the TBEP also had a need to understand how the broader community was engaging with the dashboard content. The Google Analytics service was used to track the number of users over time, their approximate location, how long a user engaged with the dashboard, and how the dashboard was accessed. From May 13th through September, 2,856 users visited the dashboard, representing 654 cities, 70 states, and 19 countries (Table 2). Understandably, most users were local, with the state of Florida accounting for a majority of visits. An average of one minute and 21 seconds was spent on the dashboard by each user, although this time varied throughout the summer, with about an even split of users accessing through a mobile device or a web browser on a PC. The number of users also varied over time, with the highest number of visits in July (Figure 4a), which corresponded to severe red tide conditions and observable fish kills in densely populated areas of the bay ([Beck et al. 2022](#ref-Beck22)). Most users accessed the website through referrals from other websites (e.g., Figure 4b), although a substantial amount of visitors also accessed the dashboard directly (e.g., bookmarks or entering the URL in a web browser) or through email links. Accessing the dashboard through social media or organically through a search engine was less common. Trends in how users accessed the dashboard did not appear to change over time.

# Data Archive

Finally, the synthesized datasets used by the dashboard were documented with complete metadata and uploaded to a federated data repository in early 2022. The TBEP has a commitment to open science defined in the Strategic Management Plan and an essential part of his effort is ensuring that data are available and documented appropriately for reuse by others. Notably, past spill events from Piney Point have occurred and the documentation and data describing these events is lacking. Making the data available from the 2021 event, in addition to publishing an open access journal article describing the results (i.e., [Beck et al. 2022](#ref-Beck22)), ensures that the broader community can learn from this event, both to inform research of estuarine impacts from inorganic nutrient inputs and efforts in response-based monitoring conducted by the management community.

The synthesized datasets were uploaded to the Knowledge Network for Biocomplexity (KNB) data repository with metadata created using the Ecological Metadata Language (EML).

* Rationale, maybe make relative to past PP events (no data) and highlight importance of metadata
* Workflow and GitHub repository
* EML, KNB location, download stats

# Discussion

* Added value of the app
  + analytics speak for themselves, but not the whole picture. How did users get to our dashboard (some of this info is in Analytics, but which of our comms products drove the bus)? How was it used (e.g., screenshots for pres, K. Hubbard pers. comm.)?
  + how can we better assess the value of these tools more holistically? What is the true measure of success in terms of behavior change or driving policy, if any?
* Relevant considerations
  + What worked - use of OS tools, but still very time-consuming (e.g., R/dat\_proc.R is thousands of lines of code)
  + Need for unified datasets, use of tidy principles, etc. Also mention that this relied on partners sending us data/putting data on GDrive, not everyone did that (note issues with the latter, i.e., access from elsewhere can be unreliable, but we also didn’t have all the data). Highly value of partnerships - data collection and sharing.
  + Value of historical data and contextualized summaries
  + Exposure of additional data needs (e.g., macroalgae)
* Future of Piney Point, next steps, etc.

# Figures

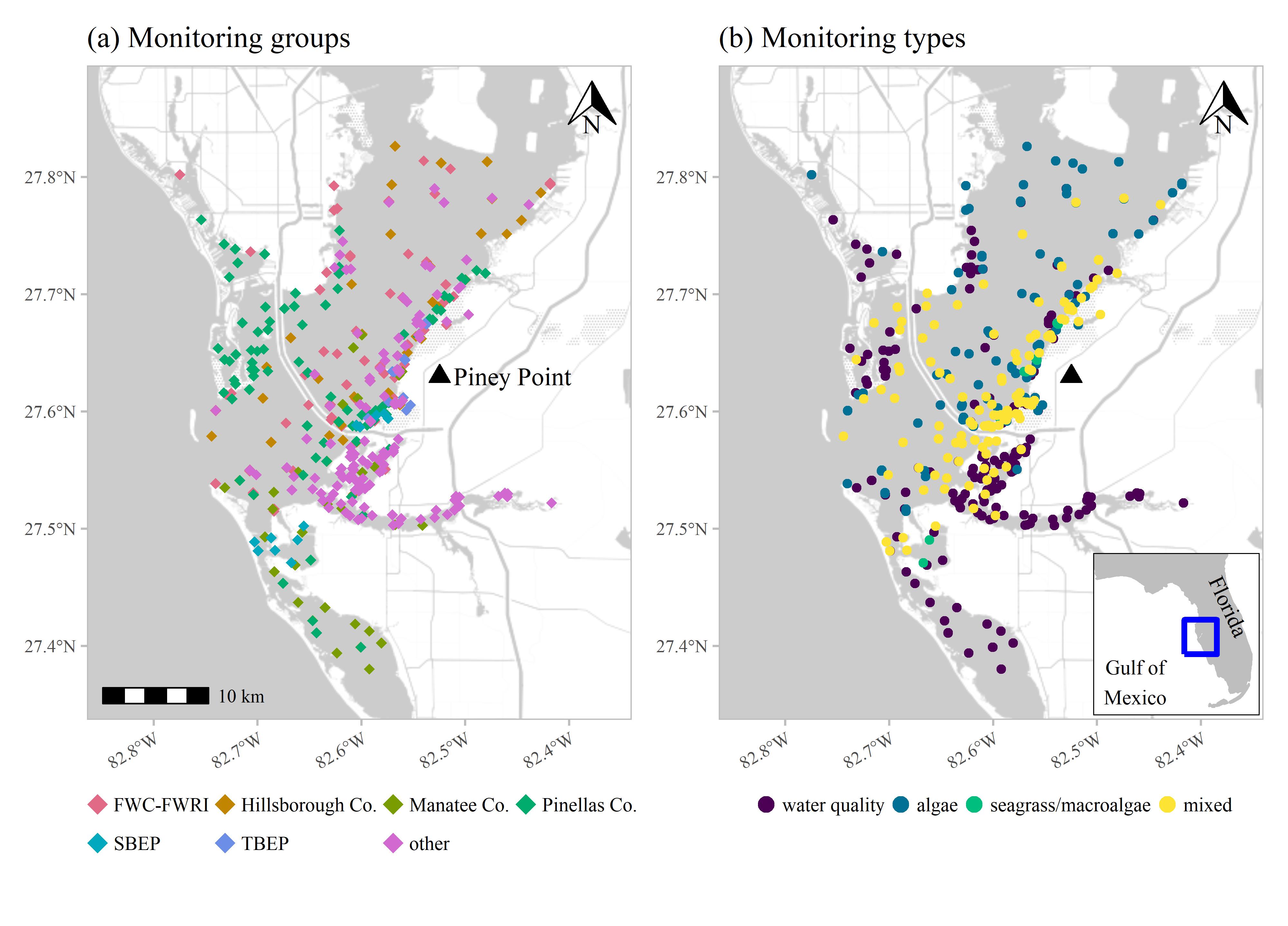


Figure 1: Response-based monitoring effort near Piney Point from March through September 2021 with (a) effort by monitoring group and (b) monitoring data type. Monitoring groups included City of St. Petersburg, Environmental Science Associates, Florida Department of Environmental Protection, Fish and Wildlife Commission, Fisheries and Wildlife Research Institute (FWC-FWRI), Hillsborough Co., Manatee Co., New College of Florida, Pinellas Co., Sarasota Bay Estuary Program (SBEP), Tampa Bay Estuary Program (TBEP), University of Florida, and University of South Florida. Monitoring data included algae sampling, seagrass and macroalgae, water quality (field-based and laboratory samples), and mixed monitoring (algae, seagrass and macroalgae, water quality). Inset shows location of Tampa Bay on the Gulf coast of Florida, USA.

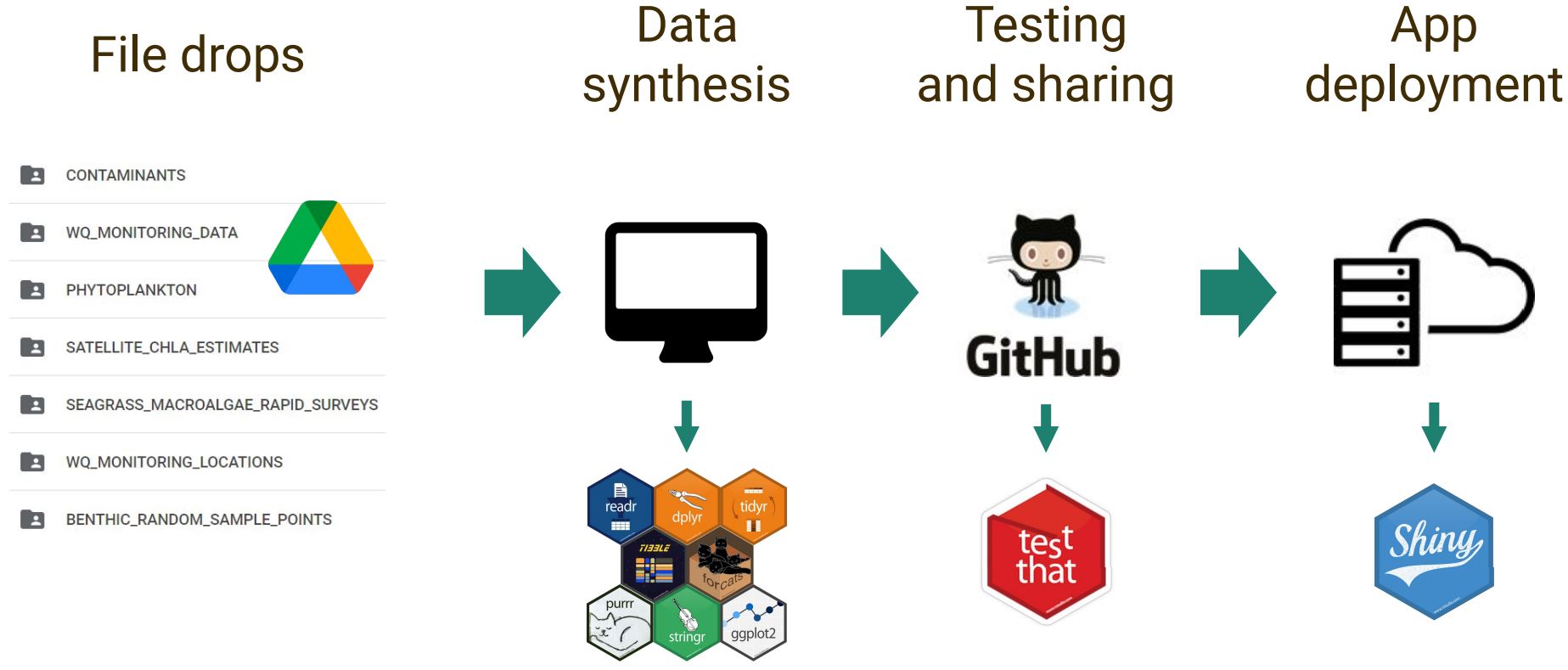


Figure 2: Data synthesis workflow for Piney Point data. Raw data were uploaded by partners to Google Drive, data were synthesized using R-based tools, data and dashboard code were uploaded in a unified format on GitHub that included automated testing, and then the app was deployed using Shiny Server.

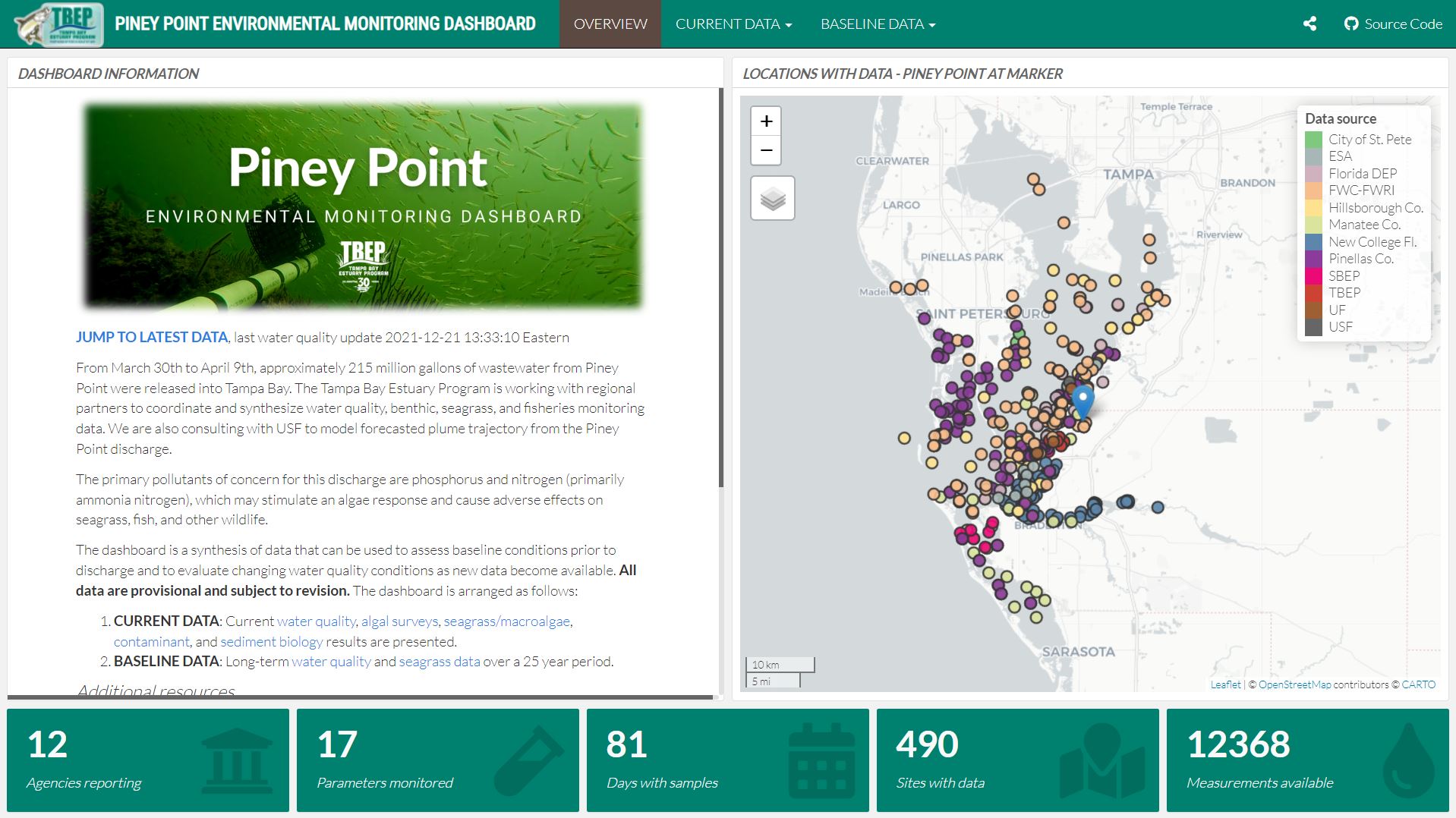


Figure 3: Landing page for the Piney Point dashboard. Users can view a summary of effort to date and access additional tabs to view specific datasets.

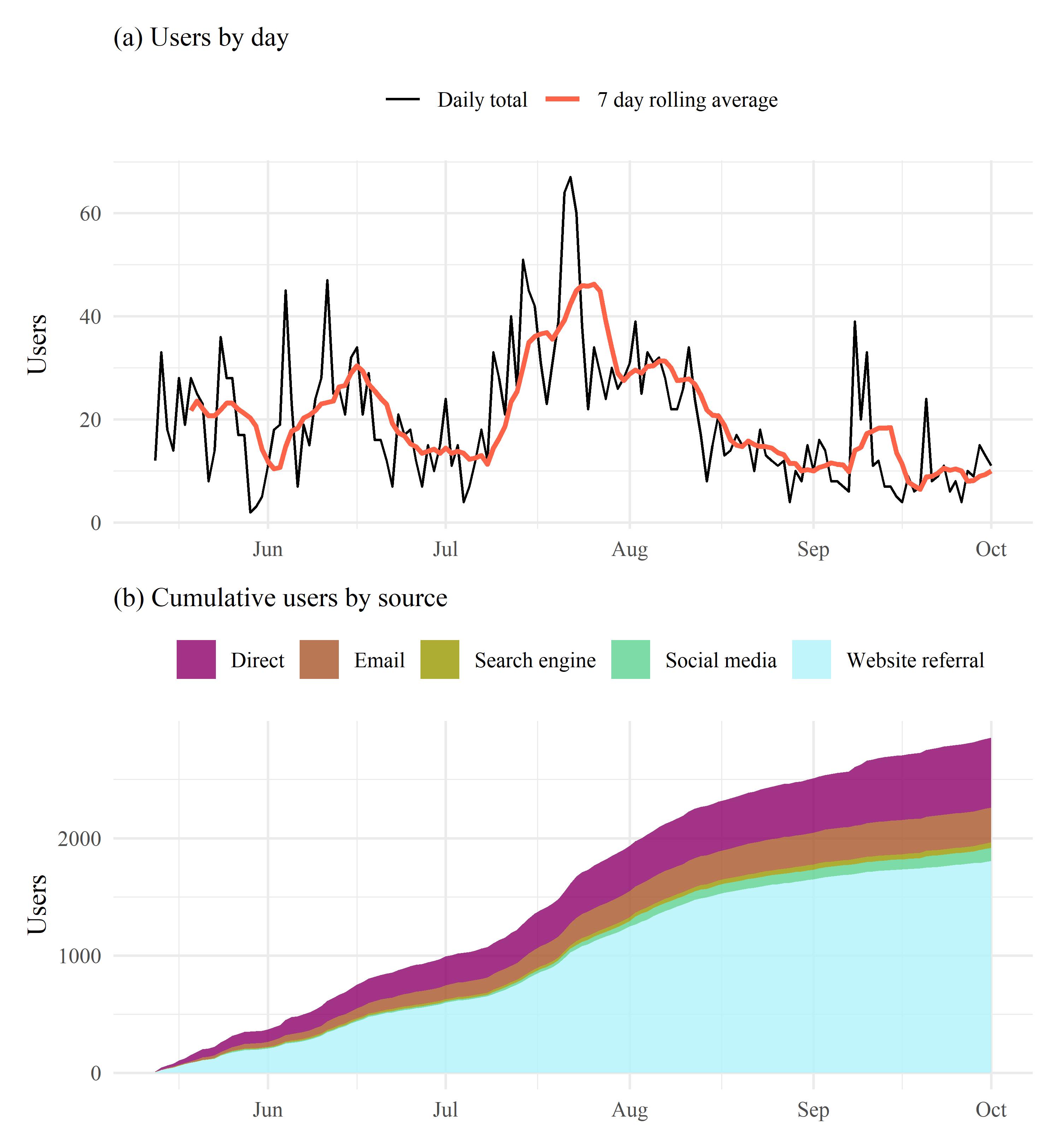


Figure 4: Piney Point dashboard (a) users by day and (b) cumulative users by source. Source refers to how users accessed the dashboard (e.g., direct URL to the dashboard, email links, etc.). Analytics were included in the dashboard beginning on May 13th, 2021.

# Tables

Table 1: Sampling effort for response-based monitoring by partners from April 1st 2021 through September. Values show the total number of unique sites samples for water quality, algae (phytoplankton), and seagrass/macroalgae in Tampa Bay. Values in parentheses show the percentage of the total for each monitoring type.

| Month | Water Quality | Algae | Seagrass/macroalgae |
| --- | --- | --- | --- |
| Apr | 223 (45.1) | 181 (41.8) | 30 (17.4) |
| May | 82 (16.6) | 62 (14.3) | 36 (20.9) |
| Jun | 55 (11.1) | 40 (9.2) | 35 (20.3) |
| Jul | 60 (12.1) | 54 (12.5) | 18 (10.5) |
| Aug | 43 (8.7) | 49 (11.3) | 28 (16.3) |
| Sep | 31 (6.3) | 47 (10.9) | 25 (14.5) |

Table 2: Top ten locations of users by city, state, and country for the Piney Point dashboard. Numbers show the total users and percentages from the total for all locations. Total cities, states, and countries that accessed the dashboard are shown in the column headers in parentheses. Summaries are from May 13th, 2021 when analytics were included in the dashboard to October 1st, 2021.

| City (654) | Users | State (70) | Users | Country (19) | Users |
| --- | --- | --- | --- | --- | --- |
| St. Petersburg | 325 (11.4) | Florida | 2005 (69.9) | United States | 2812 (98.3) |
| Tampa | 282 (9.9) | California | 107 (3.7) | Canada | 11 (0.4) |
| Bradenton | 81 (2.8) | New York | 65 (2.3) | United Kingdom | 9 (0.3) |
| Largo | 68 (2.4) | Texas | 57 (2) | Germany | 8 (0.3) |
| Orlando | 66 (2.3) | Georgia | 56 (2) | Finland | 3 (0.1) |
| Clearwater | 52 (1.8) | Ohio | 44 (1.5) | India | 3 (0.1) |
| Roseville | 49 (1.7) | Illinois | 36 (1.3) | Ireland | 2 (0.1) |
| Pinellas Park | 46 (1.6) | Pennsylvania | 35 (1.2) | Netherlands | 2 (0.1) |
| Sarasota | 43 (1.5) | Indiana | 31 (1.1) | Spain | 2 (0.1) |
| Miami | 42 (1.5) | Virginia | 27 (0.9) | Australia | 1 (0) |

# References

Beck MW. 2021. [2021 Piney Point Sampling in Tampa Bay. Knowledge Network for Biocomplexity. urn:node:KNB.](https://doi.org/doi:10.5063/F1959G05)

Beck MW, Altieri A, Angelini C, Burke MC, Chen J, Chin DW, Gardiner J, Hu C, Hubbard KA, Liu Y, Lopez C, Medina M, Morrison E, Phlips EJ, Raulerson GE, Scolaro S, Sherwood ET, Tomasko D, Weisberg RH, Whalen J. 2022. [Initial estuarine response to inorganic nutrient inputs from a legacy mining facility adjacent to Tampa Bay, Florida](https://doi.org/10.1016/j.marpolbul.2022.113598). Marine Pollution Bulletin 178: 113598.

Beck MW, Burke MC, Raulerson GE, Scolaro SC, Sherwood ET, Whalen JD. 2021. [Piney Point environmental monitoring dashboard. tbep-tech/piney-point: v1.0.0](https://doi.org/10.5281/zenodo.4666494). Zenodo.

Bryan J. 2020. [googlesheets4: Access Google Sheets using the Sheets API V4. R package version 0.2.0. https://CRAN.R-project.org/package=googlesheets4](https://CRAN.R-project.org/package=googlesheets4).

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B. 2021. [shiny: Web application framework for R](https://CRAN.R-project.org/package=shiny).

Chen J, Weisberg RH, Liu Y, Zheng L. 2018. [The Tampa Bay Coastal Ocean Model performance for Hurricane Irma](https://doi.org/10.4031/MTSJ.52.3.6). Marine Technology Society Journal 52: 33–42.

Chen J, Weisberg RH, Liu Y, Zheng L, Zhu J. 2019. [On the momentum balance of tampa bay](https://doi.org/10.1029/2018JC014890). Journal of Geophysical Research: Oceans 124: 4492–4510.

D’Agostino McGowan L, Bryan J. 2020. [googledrive: An interface to Google Drive. R package version 1.0.1. https://CRAN.R-project.org/package=googledrive](https://CRAN.R-project.org/package=googledrive).

Greening H, Janicki A. 2006. [Toward reversal of eutrophic conditions in a subtropical estuary: Water quality and seagrass response to nitrogen loading reductions in Tampa Bay, Florida, USA](https://doi.org/10.1007/s00267-005-0079-4). Environmental Management 38: 163–178.

Greening H, Janicki A, Sherwood E, Pribble R, Johansson JOR. 2014. [Ecosystem responses to long-term nutrient management in an urban estuary: Tampa Bay, Florida, USA](https://doi.org/10.1016/j.ecss.2014.10.003). Estuarine, Coastal and Shelf Science 151: A1–A16.

Iannone R, Allaire J, Borges B. 2020. [flexdashboard: R Markdown format for flexible dashboards](https://CRAN.R-project.org/package=flexdashboard).

Liu Y, Weisberg RH, Zheng L, Sun Y, Chen J. 2021. Nowcast/forecast of the Tampa Bay, Piney Point effluent plume: A rapid response, *In* Abstract (OS35B-1036) presented at AGU fall meeting, december, 2021. New Orleans, Lousiana.

R Core Team. 2021. [R: A language and environment for statistical computing, v4.1.2](https://www.R-project.org/). R Foundation for Statistical Computing, Vienna, Austria.

Wang P, Martin J, Morrison G. 1999. [Water quality and eutrophication in Tampa Bay, Florida](https://doi.org/10.1006/ecss.1999.0490). Estuarine, Coastal and Shelf Science 49: 1–20.

Wickham H. 2011. [testthat: Get started with testing](https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf). The R Journal 3: 5–10.

Wickham H. 2014. [Tidy data](https://doi.org/10.18637/jss.v059.i10). Journal of Statistical Software 59: 1–23.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. [Welcome to the tidyverse](https://doi.org/10.21105/joss.01686). Journal of Open Source Software 4: 1686.

Xie Y, Dervieux C, Riederer E. 2020. [R markdown cookbook](https://bookdown.org/yihui/rmarkdown-cookbook). Chapman; Hall/CRC, Boca Raton, Florida.