# Provisional Data Access Addendum

## Introduction

Research projects are often conducted in areas that are of high interest to a variety of stakeholders and they are often interested in seeing preliminary data and results as they come available (i.e. provisional) instead of at the end of a project when final publications are prepared and submitted. Making this provisional data available benefits the research in that it increases the likelihood that other users may uncover potential quality issues with the data. It also benefits the project as we are providing useful information to our partners and stakeholders in a time frame that is more beneficial for their end uses. This addendum provides the details of how we will make data from our high-resolution cyanobacteria research available and how we will conduct preliminary quality control checks as those data are collected.

## Data

There are several data sources associated with this project and we plan to make all of them available as provisional data when they come available. All data are collected in two ponds on Cape Cod, Shubael and Hamblin Ponds. Each data source is listed below with a brief description, list of parameters, and the approximate frequency at which they are collected. More details on each are included in the QAPP.

* Water quality buoys: The buoys are NexSens CB-150s and are equipped with a YSI EXO2 Multi- parameter sonde, a Trios NICOs optical nitrate sensor, and an Airmar 200WX weather station. Data is logged and transmitted via the NexSens X2-CB data logger via cellular network. All data are collected every 15 minutes. The buoys are located at the approximate center of each waterbody.
  + Parameters: nitrate, water temperature, pH, specific conductivity, dissolved oxygen, turbidity, chlorophyll, phycocyanin, fdom, barometric pressure, air temperature, wind direction, wind speed
* Fast Limnological Automated Measurements (FLAMe): The FLAMe is an on-board flowthrough system that provides detailed spatial data of the same water quality parameters that are measured by the water quality buoys. We follow a pre-determined grid within the waterbody and take a measurement of each parameter every 1 second, except for nitrate which is measured every 10 seconds. Location data are corrected based on the overall FLAMe residence time and sensor response time.
  + Parameters: nitrate, water temperature, pH, specific conductivity, dissolved oxygen, turbidity, chlorophyll, phycocyanin, fdom
* EXO2 sonde data: At five locations within each waterbody, the buoy location and at four littoral locations, we also collect data from a hand-held EXO2 multi-parameter sonde. These data provide information on possible drift in the buoy measurements and serve as a check for the FLAMe measurements. The same parameters are measured; however, nitrate is measured with a different sensor than on the FLAMe and buoy, the YSI nitra LED.
  + Parameters: nitrate, water temperature, pH, specific conductivity, dissolved oxygen, turbidity, chlorophyll, phycocyanin, fdom
* Water quality grab samples: Grab samples area also collected at the five locations where hand-held sonde data is collected. Grab samples are used to validate and transform some of the sensor based measurements as well as for measuring other parameters that we are not able to measure with sensors.
  + Parameters: extracted chlorophyll, extracted phycocyanin, nutrients (dissolved and total), microcystin
* Zooplankton and Phytoplankton: At the buoy site only, 150 and 50 µm plankton tows are performed to examing zooplankton, and whole water samples are used for phytoplankton
  + Parameters: Genus level abundance for both zoo- and phytoplankton.

## Quality control checks

Quality assurance activities are detailed in the QAPP and all of these will be conducted (e.g. sensor calibration, duplicate samples, etc.) prior to provisional release of data. Other quality control checks, such as outlier detection, are best used once a dataset is closer to completion and thus are not appropriate for provisional datasets. That being said, we do have past data, collected in similar systems, that provides reasonable ranges for each variable to be included in the provisional dataset. Additionally some variables (e.g., wind direction and speed) have other values that can be used for an initial quality control screen and flag. For most variables we flag values that are 25% greater than the prior measured maximum values or 25% less than the prior measured minimum values. The values are listed in Table 1.

**Table 1. Range check values for each collected parameter**

| Parameter | Quality Control Check |
| --- | --- |
| nitrate (NICO/ecoN) | 0 to past + 25% |
| water temperature | 0 to past + 25% |
| pH | 0 to 14 |
| specific conductivity | 0 to past + 25% |
| dissolved oxygen concentration | 0 to past + 25% |
| dissolved oxygen saturation | 0 to past + 25% |
| turbidity | 0 to past + 25% |
| chlorophyll | 0 to past + 25% |
| phycocyanin | 0 to past + 25% |
| fdom | 0 to past + 25% |
| barometric pressure | 0 to past + 25% |
| air temperature | 0 to past + 25% |
| wind direction | 0 to 360 |
| wind speed | 0 to 150 (reasonable expectation) |
| extracted chlorophyll | 0 to past + 25% |
| extracted phycocyanin | 0 to past + 25% |
| nutrients | 0 to past + 25% |
| microcystin | 0 to past + 25% |
| zooplankton abundance | NA |
| phytoplankton abundance | NA |

## Data processing procedures

Data for this project are collected three ways: via telemetry from the buoys, direct download from the FLAMe, or recorded after lab processing and analysis. Details for processing each provisional data stream are below:

* Telemetry from buoys: The data buoys will measure every 15 minutes and send data to WQ Data Live (<https://wqdatalive.com>) via cellular telemetry. A scheduled report will send data every 4 hours from WQ Data Live to a private folder on EPA’s sftp site on GoAnywhere (<https:://newftp.epa.gov>). The data on EPA’s sftp will be retrieved, cleaned, and quality control checks (outlined above) applied every 4 hours via GitHub Actions and a series of R functions. The GitHub Actions will be set to run on a private repository (USEPA/high\_res\_cyano) inside of EPA’s Enterprise GitHub organization. Additionally, all data are stored on the buoys themselves and can be retrieved manually should these automated methods fail.
* Direct download from FLAMe: Data from the FLAMe is transferred from the sensors to a Sutron SatLink data logger. A single FLAMe run collects all measurements every 10 second (or faster) and the location of the collection is also recorded. After a FLAMe run, all data are downloaded to a USEPA field laptop. These data will be checked for quality (see above) after collection with R functions. The data will be added to the private repository (USEPA/high\_res\_cyano) inside of EPA’s Enterprise GitHub organization.
* Recorded after analysis: Extracted pigments, nutrients, microcystin, and plankton counts are all based off of water samples collected in the field and then subsequently analyze in the lab. Details on QA/QC procedures for these variables are outline in QAPP-XX-XXXX and these procedures will be applied prior to any provisional release of these data. In short, though, the procedures rely on direct download via a serial port from the devices for extracted pigments, nutrients, and microcystin. The plankton samples are manually counted and recorded and added to a csv file on an EPA workstation. These data are cleaned and combined with other project data and addedd to the private repository (USEPA/high\_res\_cyano) inside of EPA’s Enterprise GitHub organization.

## Data management

All project data will be managed as a an Apache Arrow formatted columnar data file (<https://arrow.apache.org/>). This format is cross-platform and provides efficient storage and read/write. Changes to this file are managed via the git version control system and ccess to the data for project members are via a private repository (USEPA/high\_res\_cyano) inside of EPA’s Enterprise GitHub organization.

The data file will be generated on a schedule from files committed to the repository. Each of these files is provided in a unique format. We will use a scheduled run, via GitHub Actions, to clean these files and reorganize them into a standard format. All datasets will be combined into the single project data file. The structure of this file is outlined in Table 2.

**Table 2. Data file structure**

| Field Name | Data Type | Details |
| --- | --- | --- |
| date | Date (ISO 8601) | The date data was collected |
| time | Hours, Minutes, Seconds (hms) | The time data was collected |
| waterbody | Character | Name of the waterbody |
| site | Character | Site identifier within the waterbody |
| depth | Numeric | The depth at which the data was collected |
| field\_dups | Numeric | Field duplicate identifier |
| lab\_reps | Numeric | Lab repition idenfitifer |
| variable | Character | The name of the variable |
| units | Character | Units for the variable |
| value | Numeric | The measured value |
| notes | Character | Notes about the measurement, including QA Flags |

This is a long-data format and is incredibly flexible as it allows for adding new data streams without having to adjust the structure of the file. Additionally, supplementary data (e.g. geographic coordinates, minimum detection limit, etc.) are easily linked with this data as each row represents a single sampling event.

## Open data access

As data are collected, processed, and initial checks are run they will be added to the Apache Arrow file described above. This file will then be converted to a .csv file and moved to a public repository within the USEPA Github Enterprise organization on a scheduled basis (frequency to-be-determined) via GitHub Actions. This public repository will also contain a README file that provides a brief description of the data, a data dictionary, details on how to access the data, standard disclaimers, and a designation that the data are provisional and subject to change. This repository will be made available with a Creative Commons Zero public domain desingation.