https://doi.org/10.15027/0002041270

Using particle size distribution (PSD) to automate imaging flow cytobot (IFCB) data quality in coastal California, USA

Kendra Hayashi[1*], Jamie Enslein[1,2], Alle Lie[3], Jayme Smith[3], and Raphael M. Kudela[1].

[1]Ocean Sciences Department, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, United States
[2]University of California, Berkeley, 306 Stanley Hall MC #1762, Berkeley, CA 94720, United States
[3]Southern California Coastal Water Research Project, 3535 Harbor Blvd., Suite 110, Costa Mesa, CA 92626, United States

[*]corresponding author's email: khayashi@ucsc.edu

Abstract

The use of imaging flow cytobots (IFCBs) for plankton research is increasing worldwide and coordinated IFCB networks are being developed to monitor harmful algal blooms (HABs) in several coastal regions. Monitoring programs with IFCBs designed to run continuously can generate up to 70 samples per day creating a wealth of image data. Ideally, data streams are monitored daily (real-time) for data quality assurance and quality control (QA/QC). However, front end data QA/QC can be cumbersome for personnel and is often left for a later date once thousands of data files have accumulated. Particle size distribution (PSD) is used to inform food web dynamics, calculate total community biomass, and calculate radiative transfer in ocean remote sensing. PSD can be generated from equivalent spherical diameter (ESD), a measure derived from IFCB image processing, and in previous work, anomalous IFCB generated PSDs identified bloom events in San Francisco Bay, CA. We propose that variations in PSDs also reveal "bad" data to allow for some automation in backend QA/QC procedures. As more and larger IFCB networks come online worldwide, the use of automated data QA/QC is prudent to increase the efficiency of working with these datasets. While full automation of IFCB data QA/QC is unlikely, using PSD to automatically flag data allows users to focus their efforts on a reduced number of data to determine whether they are questionable or reflect shifts in community structure.

Keywords: imaging flow cytobot (IFCB), particle size distribution (PSD), data QA/QC

Introduction

Imaging flow cytobots (IFCBs; McLane Research Laboratories, Inc., USA) are plankton imagers designed to autonomously collect samples in the environment. The high-resolution images collected at high temporal frequency result in a powerful tool being employed by monitoring networks worldwide. When sampling continuously, a single IFCB has the potential to collect >10,000 sample files per year creating a wealth of data needing quality assurance and quality control (QA/QC). Ideally each file would be manually checked real-time to ensure the IFCB was operating correctly, however this is time consuming, inefficient, and not a standard practice, so users are left with thousands of files to QA/QC after the fact, resulting in a need to automate the QA/QC process.

Particle size distribution (PSD) is a critical component to understanding the optical properties of the water column. It is used to inform food web dynamics, calculate total community biomass, and calculate radiative transfer in ocean remote sensing (Reynolds et al. 2010). PSD can be generated by plotting estimated spherical diameter (ESD), a measurement calculated during the post-processing of IFCB images, against particle concentration (Fig. 1). Previous work demonstrated that variations from theoretical PSD, specifically an increased number of particles of a certain size, revealed phytoplankton blooms in San Francisco Bay, California (Hayashi and Kudela 2022). Here we developed code using PSD to QA/QC IFCB data and evaluated its performance at 3 sites in the California (CA) IFCB Network.

Materials and Methods

Site Information

Three datasets from the CA IFCB Network were used to evaluate the automated data QA/QC code: Newport Beach Pier (NBP), Santa Cruz Wharf

(SCW), and the Hog Island Oyster Company (HIOC). The SCW and HIOC deployments are above water using a peristaltic pump to bring water to the IFCB. Data from the SCW span a 4-month stretch (3969 files, 9Aug to 4Dec2016) and were used to build and set thresholds for the QA/QC code. The HIOC dataset spanned about a 1-month deployment period (1835 files, 20Apr to 25May2023). The NBP has an underwater IFCB deployment configuration and had the largest data set evaluated, collected over about 2 years (14,969 files, 9Jul2021 to 13Aug2023).

All datasets had a subset of data manually checked. Data files from NBP and HIOC were manually checked by someone not familiar with the IFCB, but followed a specific set of guidelines. SCW data were manually checked by someone familiar with the IFCB and phytoplankton and had an additional 'Bad XY' flag that was not included in the automated code.

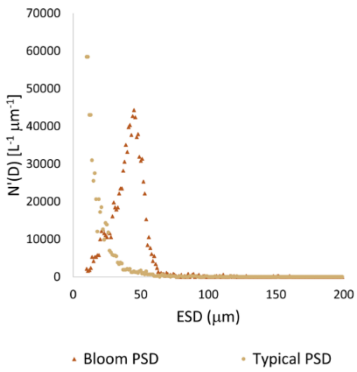

Fig. 1. PSDs generated from IFCB data. A typical PSD is represented by the tan dots and the dark orange triangles are the PSD of a sample with a 'Bloom' auto flag.

QA/QC code Information and Requirements
The QA/QC code was written in Python and the version used to evaluate the datasets in this study is available to download in GitHub (https://github.com/kudelalab/PSD). The code requires access to the IFCB's raw data files (.hdr and .adc) and corresponding v2 feature files generated from MATLAB code (https://github.com/hsosik/ifcb-analysis/; Sosik and Olson 2007). PSD was generated from 1 µm binned equivalent spherical diameter (ESD) and data were fit to a power law or Junge distribution. A total of 7 flags are available using the QA/QC code, their descriptions and thresholds are as follows:

- Beads – File with 'sampleType: Beads' or a power fit constant greater than a user defined threshold (generally greater than 10^7)
- Bloom – ESD difference between the maximum particle concentration and smallest ESD used in the curve fit is greater than 5 µm (Fig. 1)
- Bubbles – ESD with the maximum particle concentration is greater than 150 µm
- Incomplete Run – Runtime of a sample is less than half the expected runtime
- Low biomass – Maximum particle concentration is less than 1000 L^{-1} µm^{-1}
- Low R^2 – $r^2 < 0.5$ and the poor fit is not associated with low biomass or bloom
- Missing Cells – Image:Trigger count ratio is less than 0.8

It is important to note that the 'Bloom' and 'Low biomass' are not meant to denote bad data files. These flags are included because they explain why the power fit has a low r^2 value and to inform the users of unusual events.

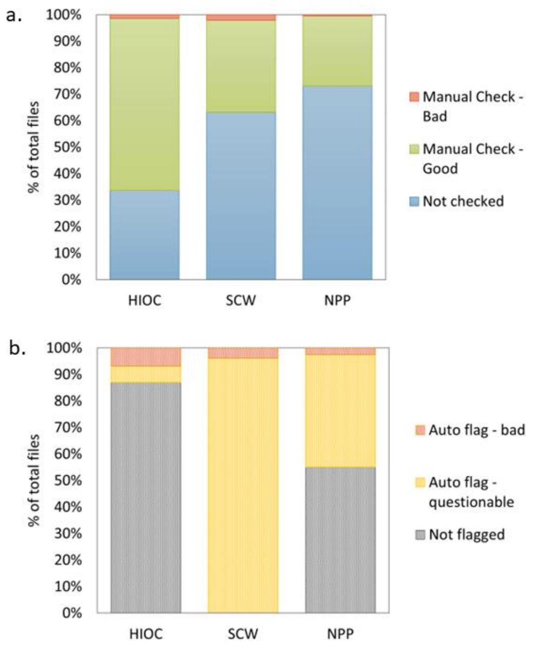

Fig. 2. Stacked bar graphs representing the distribution of bad data flags in manual checked (a) and auto flagged (b) datasets from the HIOC, SCW, and NBP. 'Bead,' 'Bubble,' 'Incomplete Run,' and 'Missing Cells' flags were combined into the 'Auto flag-bad' category. 'Auto flag-questionable' includes 'Low biomass,' 'Low R^2,' and 'Bloom' flags.

Results and Discussion

Overall Data Quality

A minimum of 27% of the data from each site was manually checked (Fig. 2a) using the online IFCB dashboard interface (https://ifcb.caloos.org/dashboard). Collectively this effort took over 300 hours, however, that estimate included the time for files to load. When webpage load times were removed, manual checking speeds ranged from 25-200 files per hour (depending on level of experience). Less than 2% of the manually checked files at all 3 sites were flagged as bad data and a majority of those flagged files (>70%) were bead samples.

The automated flags showed a similar result with 3-7% of the files in the dataset being flagged as bad data (Fig. 2b; bad data include the following flags: 'Bead', 'Bubble', 'Incomplete Run', and 'Missing Cells'). It should be noted that the proportion of data flagged by the QA/QC code varied widely between sites, ranging from 13% (at HIOC) to 100% (at SCW). However, when we look at the break down of flag type (Fig. 3), >78% of the automated flags in the SCW and NBP datasets were 'Bloom' and do not indicate bad data. At both sites, the 'Bloom' flagged files were collected during periods when *Akashiwo sanguinea* or *Lingulodinium polyedra*, dominated the phytoplankton assemblage, indicating the flag's potential to identify HABs. Occasionally, the 'Bloom' flags were applied to files with peak size concentrations closer to the low end of the curve fit range (10 μm) and did not represent a bloom. While this is not ideal, it is not detrimental, as the intention of the flag was to indicate files with interesting community structure and not bad data.

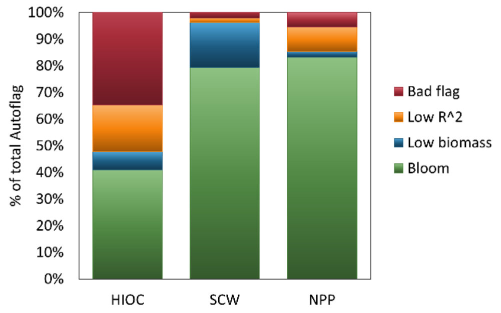

Fig. 3. Stacked bar graph representing the distribution of flag types (orange and yellow bars in Fig. 2b) in the QA/QC code generated flags for the HIOC, SCW, and NBP datasets. 'Bead,' 'Bubble,' 'Incomplete Run,' and 'Missing Cells' flags were combined into the 'Bad Flag' category.

Collectively, both the manual and automated flag results indicate that when IFCBs are deployed and operating, they are generating quality data over 95% of the time.

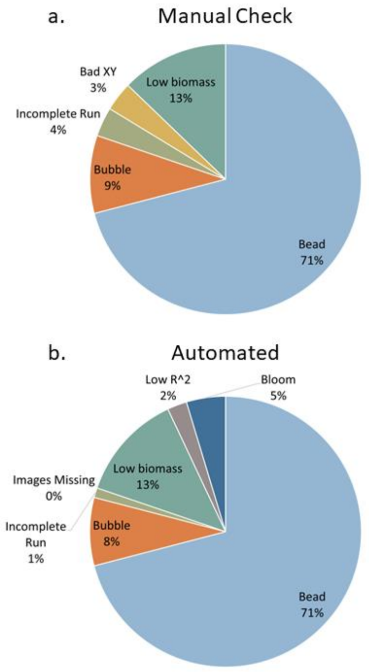

Fig. 4. Pie charts representing the distribution of flag types for the 86 matched manual check files (a) and the automated code flags (b) from the SCW dataset.

Automated vs Manual flags

To evaluate the QA/QC code performance, the manual and automated flags from the SCW dataset were directly compared by looking at the distribution of flag types (Fig. 4). Eighty of the 86 match ups had identical manual and automated flags. Of the 6 mismatches, 4 files were flagged as 'Bloom' and 2 files were flagged as 'Low R^2' by the code. The manual flag for 2 of the files (1 auto flagged 'Bloom', 1 auto flagged 'Low R^2') was 'Incomplete Run' and was likely miscategorized in the code due to an order of operations and can be fixed to elevate the 'Incomplete Run' flag, as it is more critical for data quality. The remaining auto flagged 'Bloom' samples correctly fell under the bloom criteria but were manually flagged for triggering too far to the left in XY space. The other incorrect 'Low R^2' file was manually flagged as 'Bubbles' (Fig. 5). The 'Bubbles' criteria were set based on the assumption that a sample would be composed primarily of large bubble images. This specific mismatched sample had some large bubble images, but most of the images had ESDs <100 µm (real cells and partial bubbles) and we intend to change the criteria for the 'Bubble' flag to look for any particles with an ESD over 150 µm.

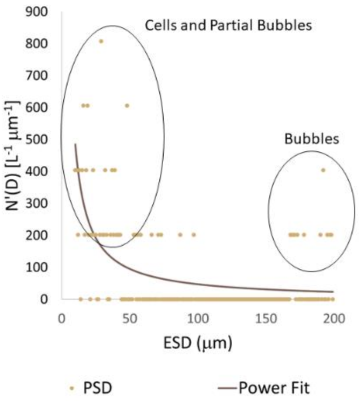

Fig. 5. An example PSD where full bubble images (circled on the right) were not the majority of the images collected in the sample. This file was auto flagged with a 'Low R^2' but should have a 'Bubble' flag.

Automated flags were manually spot checked to ensure the flags were working as expected in all 3 datasets and were found to consistently meet the flag criteria as the direct comparison revealed. More attention was directed towards the 'Low R^2' flag because goodness of fit is not something that can be manually assessed from raw images. The 'Low R^2' flag was generally assigned to files that did not have many images. A low biomass sample is not necessarily a bad sample, so care should be taken to not remove files during periods with low chlorophyll. 'Low R^2' and 'Low biomass' flags collectively made up 11-21% of the total automated flags (Fig. 3) and generally clumped together in time. We recommend files assigned these flags be manually checked before removing them from the dataset because most of the 'Low biomass' flagged files were considered good data.

A benefit to the automated code was that it was able to catch several legitimate bad data files that were missed by a human checker unfamiliar with the IFCB. For the NBP dataset, in addition to bead files caught by manual checking, the automated code caught 1 more bead file and 2 'Missing Cells' files than manual checking. The human checker and automated flags were more disparate for the HIOC dataset with the human checker missing most of the 'Missing Cells' files caught by the automated code and the automated code missing all the 'Bubble' flags caught by the human checker. Human error increases with inexperience and increasing file number whereas the code cannot provide a nuanced evaluation of images. The discrepancies in the HIOC and NBP datasets represent the risks associated with relying on only 1 QA/QC method.

As IFCBs become more widely used and monitoring networks grow larger, the need for more automated data management increases. Combining information derived from the raw IFCB files (.hdr and .adc) and information from the PSD of the sample was the most successful approach to automating the QA/QC process. However, we believe manually checking files needs to be a part of the QA/QC process and the code generated in this project is best used as a screening tool, significantly reducing the number of files to be manually checked thereby focusing effort and reducing human error.

Acknowledgements

The authors thank L. Cruz Pascual and C. Park for performing manual checks for the NBP and HIOC datasets as a part of their California College Corps Fellowship. We also thank P. Daniel and G. Zirkel for their help with accessing data from HIOC. This work was supported by Cal-HABMAP which represents multiple organizations and partners who contribute to the operation and maintenance of the IFCB network. Support was also provided by SCCWRP JPA Research Funding. Core funding has been provided by NOAA, including the IOOS, ECOHAB, and MERHAB programs, NASA, California Sea Grant, California Ocean Protection Council, and numerous in-kind contributions from our partners and stakeholders. Initial purchase and development of the IFCBs was supported by NA16NMF4270263 (Saltonstall Kennedy) and NA14NOS0120148 (NOAA OTT).

References

Hayashi, K. and Kudela, R.M. (2022) U.S. Symposium on Harmful Algae. Albany, NY, United States. https://doi.org/10.5281/zenodo.11212206

Reynolds, R.A., Stramski, D., Wright, V.M., Woziak, S.B. (2010) J. Geophys. Res. 115, C08024.

Sosik, H. M. and Olson, R. J. (2007) *Limnol. Oceanogr:Methods 5*, 204–216.