



CheckEM user guide

Brooke Gibbons

Updated: October 2023

brooke.gibbons@uwa.edu.au

CheckEM

Upload data

✓ Check metadata & periods

✓ Create & check MaxN

✓ Check length & 3D points

≡ Compare MaxN & length

✓ Create & check mass

Download data and QC score

Schema downloads

User guide

Edit maximum lengths

Feedback

Change log

Acknowledgements

100

Sample metadata score

62

Samples in the Sample Metadata

0

Sample(s) without points data

0

Sample(s) without lengths

0

Sample(s) in points file missing metadata

0

Sample(s) in lengths or 3D points file missing metadata

0

Period(s) without an end

0

Sample(s) without a period

0

Point(s) outside periods

0

Length(s) or 3D point(s) outside periods

Enter your correct period time (mins):

60

2

Periods not 60 mins long

+

-

OpCode: 61

Status: No-take

Depth: 79.8 m

Site: NA

Location: NA

Date/Time: 2022-05-26T08:23:34+08:00

Marine parks

Table of Contents

Definitions.....	3
General.....	5
Introduction.....	5
Using this manual.....	5
Version Information.....	5
Format required for upload.....	5
Sample Metadata.....	6
Defining a sample.....	6
Column requirements.....	7
Metadata tips.....	7
Annotation data.....	12
EventMeasure Database Output.....	12
Generating database output from EventMeasure.....	12
“Generic” annotation data.....	15
Uploading data.....	17
How to upload.....	17
Setting the format of uploads.....	17
Setting the life history information and spatial zoning to use.....	18
Reading the QC assessments.....	19
Check Metadata & Periods.....	21
Sample Metadata Score.....	22
Create & Check MaxN (point methods only).....	24
Check Length & 3D points.....	26
Compare MaxN & length (point methods only).....	28
Downloading Data and QC score.....	32
Downloading report of all potential ‘errors’.....	32
Downloading final data.....	33
Quality Control Score.....	34
Editing Maximum Lengths of Fish Species.....	36
Feedback.....	38

Table of Figures

Figure 1. Example of errors in sample metadata.....	7
Figure 2. Generating database output from EventMeasure.....	13
Figure 3. Example settings to export database output from EventMeasure.....	13
Figure 4. Example summary whilst exporting database output from EventMeasure.....	14
Figure 5. Example database output from EventMeasure.....	14
Figure 6. Setting the format of data to be uploaded.....	18
Figure 7. Setting the life history information to use.....	19
Figure 8. Sample metadata preview in CheckEM.....	19
Figure 9. Example of more detail that is shown when clicking a QC assessment box.....	20
Figure 10. Screenshot from the Check Metadata & periods tab showing the QC assessments...	21
Figure 11. Screenshot from the Check Metadata & periods tab of a map.....	22
Figure 12. Screenshot from the Create & Check MaxN tab showing the QC assessments.....	24
Figure 13. Screenshot from the Create & Check MaxN tab showing the spatial and status plot..	25
Figure 14. Screenshot from the Check Length & 3D points tab showing the QC assessments...	26
Figure 15. Screenshot from the Check Length & 3D points tab showing an example length histogram.....	27
Figure 16. Screenshot from the Compare MaxN and length tab showing the QC assessments..	30
Figure 17. Screenshot from the Compare MaxN and length tab showing the length measurements versus 3D points plot.....	30
Figure 18. Screenshot from the Compare MaxN and length tab showing the length measurements versus 3D points as a proportion plot.....	31
Figure 19. Adding a project name.....	32
Figure 20. Setting the limits to download all potential errors.....	33
Figure 21. Setting the errors to remove from the final downloaded data.....	34
Figure 22. Example Quality Control Score plots.....	36
Figure 23. The form to edit species maximum lengths.....	37
Figure 24. The form to give feedback on CheckEM.....	38

Table of Tables

Table 1. Column requirements for the \$_Metadata.csv file.....	9
Table 2. An example of the first five rows of a \$_Metadata.csv file.....	11
Table 3. Column requirements for the \$_Count.csv file.....	15
Table 4. Column requirements for the \$_Length.csv file.....	16
Table 5. First 5 rows of a \$_Count.csv file for an example stereo-BRUVs campaign.....	16
Table 6. First 5 rows of a \$_Length.csv file for an example stereo-BRUVs campaign.....	17
Table 7. QC assessments on the Metadata tab.....	22
Table 8. QC assessments on the MaxN tab.....	25
Table 9. QC assessments on the Length tab.....	27
Table 10. QC assessments on the Compare count and length tab.....	31

Definitions

Campaign

- A discrete set (temporal and spatial) of samples.
- All samples within a campaign use the same sampling and image analysis methods.

CampaignID

- A *CampaignID* is a unique identifier for a *Campaign* made up of YYYY-MM_Campaign_Method (\$ is used to denote a *CampaignID* throughout this guide) e.g. 2023-12_Rottnest_stereo-BRUVs.
- All files that are to be uploaded in CheckEM need to start with the *CampaignID* (YYYY-MM_Project.name-Method_Metadata.csv). The *CampaignID* is case sensitive and needs to match between files (for example the \$_Count and \$_Metadata files from the same *Campaign* need to have the exact same *CampaignID*).

Count data

- A data set containing the number of each species present in a sample.

EMObs file

- An EventMeasure measurement file. Has the file extension .EMObs.

EventMeasure

- A commonly used software for logging and reporting events occurring in digital underwater imagery. Available from www.seagis.com.au

Floating point number

- A positive or negative whole number with a decimal point. For example, 5.5, 0.25, and -103.342. Decimal point can "float" to any position necessary.

Generic annotation data

- Historical data that predates EventMeasure or data that has been changed or corrected outside of the annotation files.

GlobalArchive

- An online repository of ecological data and science communications. A platform for users to store and share ecological datasets and synthesis products. Accessed via globalarchive.org

Integer

- A whole number. Not a fraction. For example 1, 16 or 72.

Length data

- A data set containing the length of each individual present in a sample.

MaxN

- The maximum number of a particular family, genus, and species measured in any single image.

OpCode

- Short for Operation Code. A code associated with an EventMeasure measurement file. The user enters the OpCode in the information fields for each EventMeasure measurement file.

Period

- A portion of time set by the user in an EventMeasure file. The user sets the period start time, end time and a name.

Sample

- A single observational unit (e.g. a single stereo-BRUV deployment or stereo-DOV transect).

Sampling method

- The technique or equipment used to collect the ecological data e.g. stereo-BRUVs (baited remote underwater stereo-video) and stereo-DOVs (diver operated stereo-video).

Sample metadata

- A spreadsheet of properties of all *samples* in the *Campaign* including where the sample is in space and time.

Stage

- Development stage. Either AD, F, M or J (Adult, Female, Male or Juvenile respectively).

General

Introduction

CheckEM is an open-source web based application which provides quality control assessments on metadata and image annotations of fish stereo-imagery. It is available at marine-ecology.shinyapps.io/CheckEM. The application can assess a range of sampling methods and annotation data formats for common inaccuracies made whilst annotating stereo imagery. CheckEM creates interactive plots and tables in a graphical interface, and provides summarised data and a report of potential errors to download.

CheckEM runs up to 30 quality control assessments on uploaded data, for example it warns users if their metadata is missing information, flags species that are outside of their documented distribution or are larger than their maximum size listed on FishBase. It provides a list of all the potential errors in the annotation files for analysts to easily integrate into their quality control workflow.

CheckEM is not fully automated; it only flags suspicious annotations. It is then up to the user to go back to the raw annotations and review the imagery and annotations. This ensures that the link between the raw annotations and the data exported is maintained.

A video summarising this user guide is available on [vimeo](#) and walks you through all steps of preparing data, uploading data and downloading data.

Using this manual

This manual details the data format requirements, how to set up an upload, how to read the QC assessments and how to download data from CheckEM. Because the manual uses links within the document and to external websites it is best used in electronic form. This manual can be accessed through the *User guide* tab on CheckEM or through the PDF download.

Version Information

See the *Changelog* tab for a summary of the recent changes between versions of the application.

Format required for upload

All files to upload need to be saved in one folder (can be on your computer, harddrive or a network folder). CheckEM reads files based on the suffix (e.g. `_Metadata.csv` or `_Points.txt`), therefore all files uploaded to CheckEM need to be named consistently (CheckEM is case sensitive).

CheckEM can check multiple campaigns at a time but can only check one type of [annotation data format](#).

Sample Metadata

Both [annotation data types](#) require a sample metadata spreadsheet saved as a CSV file. The sample metadata is just as important as the annotations. It is a record of where and when each sample was collected, if the sample was successfully annotated for count data and/or length data, who annotated the sample and if the sample was collected in an area closed to fishing. The sample metadata is partially collected in the field (e.g. `date_time`, `latitude_dd` and `longitude_dd`) and partially filled out during annotating (e.g. `observer_count` and `successful_count`). We suggest that the sample metadata collected in the field is maintained in a shareable, online spreadsheet (e.g. GoogleSheets or OneDrive) so multiple annotators can fill out the extra columns whilst they are annotating, an example google sheet version of a template sample metadata spreadsheet is available [here](#).

Defining a sample

There should be one row in the sample metadata file for every sample collected (e.g. one stereo-BRUV deployment or one stereo-DOV transect). CheckEM will match the annotation data to the sample metadata using the columns that are present in the sample metadata. There are three ways to define a unique sample in the sample metadata:

1. *opcode* only (see example 1)
2. *period* only (see example 2)
3. *opcode* and *period* (see example 3)

Example 1. A campaign using stereo-BRUVs.

The annotator creates one EMObs file per stereo-BRUV deployment. They have filled out the “OpCode” field in the information fields with a unique identifier, therefore they need to include the *opcode* column in their sample metadata. They have used periods to define the sampling duration (60 minutes) but they have named this column with information that is not needed to match the samples to the sample metadata (e.g. “Time seabed” or “seabed”), they do not include the *period* column in their sample metadata.

Example 2. A campaign using drop cameras.

The cameras in the drop camera were not turned off throughout a sampling day. To save time, the annotator has created one EMObs file per field day rather than creating one EMObs file for each deployment. They have entered a value into the information field “OpCode” that is not used to define the sample, therefore they do not need to include *opcode* in their sample metadata. They have set a period start and end, and entered a unique identifier into the period name for each deployment, therefore they need to include a *period* column in their sample metadata.

Example 3. A campaign using stereo-DOVs.

The annotator has created one EMObs per site (six transects). They have filled out the “OpCode” field in the information fields with the site code. They have set a period start and end, and entered the transect number in the period name (e.g. T1, T2, T3). As both *opcode* and *period* name are needed to create a unique identifier for each sample, they need to include both *opcode* and *period* column in their sample metadata.

It is very important that the *opcode* and *period* in the sample metadata match EventMeasure exactly (case sensitive). Care should be taken when entering the information fields and period names into EventMeasure, or use the 'skeleton' EventMeasure measurement file generation feature in EventMeasure (see the EventMeasure manual for instructions). Mismatches will be flagged in CheckEM as missing sample metadata or missing annotations.

Column requirements

The required and optional columns to be included in the sample metadata are listed in [Table 1](#). These are the same columns required by [GlobalArchive](#). If you are missing any required columns CheckEM will add the columns but they will be blank and you will receive an error message (as shown in [Figure 1](#)). **Any errors in the metadata should be fixed before proceeding.** Additional columns with campaign specific information can be added to the end of the sample metadata as required. An example of a stereo-BRUV sample metadata file is given in [Table 2](#) and a google sheet version with examples for a stereo-BRUV, stereo drop camera and stereo-DOV are available [here](#).

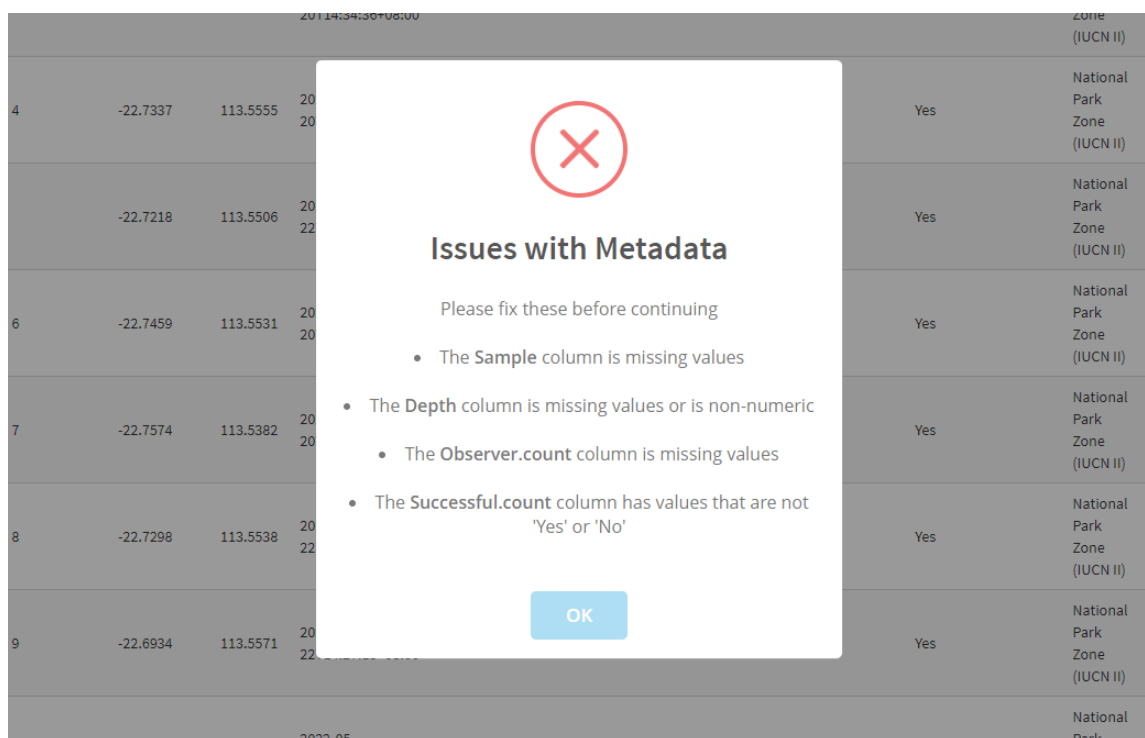


Figure 1. Example of errors in sample metadata.

These should be fixed before continuing.

Metadata tips

- We recommended using the sample metadata template for new Campaigns to reduce data manipulation and re-formatting before uploading to CheckEM and [GlobalArchive](#).

- If your longitude and latitude are not in the required format use a batch converter tool such as: www.earthpoint.us. A free account can be requested if it is for educational purposes.
 - Errors in latitude and longitude will prevent CheckEM plotting the sample metadata and annotations spatially or determining the marine region of the samples.
- Beware - Excel will parse any recognisable date and time data into the computer's system default format - which may not match the required format for the *\$_Metadata.csv*.
 - You may think you have the date format correct, but if you open the file in Excel it will change the format!

Table 1. Column requirements for the \$ _Metadata.csv file.

Transposed (rows for columns) for formatting convenience.

Column name	Format	Column required
opcode	String	✓ if opcodes were used to define a sample. DON'T include this column if it is not required to define a sample.
period	String	✓ if periods were used to define a sample. DON'T include this column if it is not required to define a sample.
latitude_dd	Decimal degrees. Must be between -90 to 90.	✓
longitude_dd	Decimal degrees. Must be between -180 to 180.	✓
date_time	YYYY-MM-DDThh:mm:ssTZD YYYY = four-digit year MM = two-digit month (01=January, etc.) DD = two-digit day of month (01 through 31) T being a required literal character. hh = two digits of hour (00 through 23) mm = two digits of minute (00 through 59) ss = two digits of second (00 through 59) TZD = time zone designator (Z or +hh:mm or -hh:mm)	✓
site	String. The scale of sites are up to the user to define.	✗
location	String. The scale of locations are up to the user to define.	✗
status	MPA status (must be Fished, No-take, I, II, III, IV, V, VI)	✓
depth_m	Floating point number (metres)	✓
successful_count	Was the sample annotated for count and will that data be included in any analysis? String ("Yes", "No" or blank).	✓
successful_length	Was the sample annotated for length and will that data be included in any analysis? String ("Yes", "No" or blank).	✓
observer_count	String (Full name of analyst). Only required if successful_count = "Yes"	✓
observer_length	String (Full name of analyst). Only required if successful_length = "Yes"	✓
visibility_m	Floating point number (metres)	✗
inclusion_probability	Floating point number. The probability of including that sample in a spatially balanced sampling design.	✗
observer_habitat_forward	String (Full name of analyst)	✗
observer_habitat_backward	String (Full name of analyst)	✗
observer_habitat_downward	String (Full name of analyst)	✗
successful_habitat_forward	String ("Yes" or "No")	✗
successful_habitat_backward	String ("Yes" or "No")	✗
successful_habitat_downward	String ("Yes" or "No")	✗

Table 2. An example of the first five rows of a \$_Metadata.csv file.

This is an example for a stereo-BRUVs campaign with additional backwards facing cameras for habitat annotation where the sample is defined using the opcode column only.

opcode	latitude_dd	longitude_dd	date_time	site	location	status	depth_m	successful_count	successful_length	observer_count	observer_length	successful_habitat_forward	successful_habitat_backward	observer_habitat_forward	observer_habitat_backward
35	-34.1315	114.9236	2023-03-15T07:36:19+08:00	Site 1	South-west Corner	No-take	39.6	Yes	Yes	Hannah Williams	Gidget Mirrabelle	Yes	Yes	Hannah Williams	Hannah Williams
5	-34.1295	114.9292	2023-03-15T07:49:41+08:00	Site 1	South-west Corner	No-take	42.7	Yes	Yes	Hannah Williams	Gidget Mirrabelle	Yes	Yes	Hannah Williams	Hannah Williams
26	-34.1272	114.9284	2023-03-15T07:54:35+08:00	Site 1	South-west Corner	No-take	36	Yes	Yes	Gidget Mirrabelle	Hannah Williams	Yes	Yes	Hannah Williams	Hannah Williams
23	-34.1283	114.9189	2023-03-15T08:01:12+08:00	Site 2	South-west Corner	Fished	41	Yes	Yes	Gidget Mirrabelle	Hannah Williams	Yes	Yes	Hannah Williams	Hannah Williams
29	-34.1229	114.9105	2023-03-15T08:07:51+08:00	Site 2	South-west Corner	Fished	42.6	Yes	Yes	Levi Peters	Gidget Mirrabelle	Yes	Yes	Hannah Williams	Hannah Williams

Annotation data

There are two types of annotation formats that you can upload to CheckEM: EventMeasure database outputs or “Generic” files. An EventMeasure database upload requires the user to have access to EventMeasure and the EMObs created during annotation. A “Generic” upload is a much simpler format and allows users who haven’t used EventMeasure to QC their annotation data.

We recommend using the EventMeasure upload if the EMObs files are available and up to date. There are more assessments possible with an EventMeasure upload than with a “Generic” upload. If you have used EventMeasure software to annotate your samples but have made corrections on the exported data (e.g. in Excel), this corrected data is now the true copy of the data and you should import your data as “Generic” annotation files (e.g. count and length).

The *opcode* and *period* names in the annotation data must **exactly match** the names in the sample metadata if they are [required to define a sample](#).

CheckEM can assess multiple campaigns at a time but can only handle one type of [annotation data format](#).

EventMeasure Database Output

Before exporting an EventMeasure database output you should make sure that all your EMObs are in one folder, and you know where that folder is saved on your computer/harddrive/network, note down how many EMObs files are within this folder. We suggest that you export one database output per Campaign (e.g. keep a separate folder of EMObs for each Campaign).

You should also create a new folder where you would like to save the outputs to.

Generating database output from EventMeasure

- Open EventMeasure
- Go to *Program*
- *Generate database output* ([Figure 2](#)).

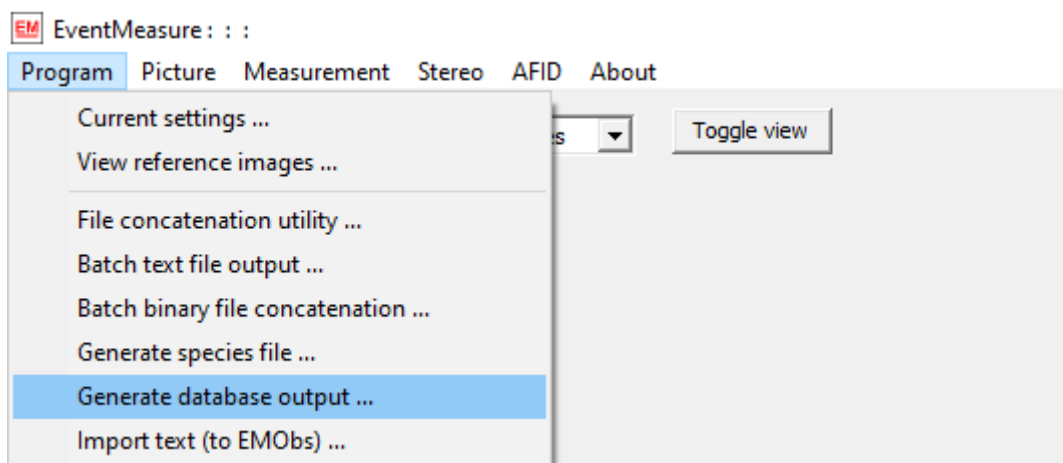


Figure 2. Generating database output from EventMeasure.

- Add the directory where your EMObs from one campaign are saved in the *Input file directory* ([Figure 3](#)).
- Add the directory where you would like to save the database outputs to in the *Output file directory* ([Figure 3](#)).
- Add the CampaignID as the *Base name* (ensure this is spelt exactly the same as the CampaignID prefix of the metadata file ([Figure 3](#)).

Name	Data	Extra info	
Input file directory	✓ Z:\Project Folders\2022-05_PtCloates_BOSS_BRUVs\BRUVs\Working\Video Analysis\EventMeasure	Directory Selection	Directory where EMObs files are located
Output file directory	✓ Z:\Project Folders\2022-05_PtCloates_BOSS_BRUVs\BRUVs\Working\Video Analysis\EM Output	Directory Selection	Directory where database files are generated
Base name	✓ 2022-05_PtCloates_stereo-BRUVs		

Figure 3. Example settings to export database output from EventMeasure.

- Click *Process*
- Check the summary table that the correct number of EMObs have been used (it should equal the number of EMObs in the input file directory) and that there are no errors shown. The summary table should look like the following screenshot ([Figure 4](#)).



Figure 4. Example summary whilst exporting database output from EventMeasure.

- Navigate to the *Output file directory* in your file explorer. You will see the eight files exported from EventMeasure for the first campaign ([Figure 5](#)). Repeat the above process for each campaign you would like to assess in CheckEM.

	2022-05_PtCloates_stereo-BRUVS_3DPoints	2023-08-01 09:48	TXT File	117 KB
	2022-05_PtCloates_stereo-BRUVS_ImagePtPair	2023-08-01 09:48	TXT File	180 KB
	2022-05_PtCloates_stereo-BRUVS_Info	2023-08-01 09:48	TXT File	2 KB
	2022-05_PtCloates_stereo-BRUVS_Lengths	2023-08-01 09:48	TXT File	350 KB
	2022-05_PtCloates_stereo-BRUVS_MovieSeq	2023-08-01 09:48	TXT File	39 KB
	2022-05_PtCloates_stereo-BRUVS_Period	2023-08-01 09:48	TXT File	5 KB
	2022-05_PtCloates_stereo-BRUVS_Points	2023-08-01 09:48	TXT File	801 KB
	2022-05_PtCloates_stereo-BRUVS_Source	2023-08-01 09:48	TXT File	4 KB

Figure 5. Example database output from EventMeasure.

The files needed for an EventMeasure upload to CheckEM are:

- \$_3DPoints.txt
- \$_Lengths.txt
- \$_Period.txt
- \$_Points.txt

“Generic” annotation data

If you do not use EventMeasure or have historical data that predates EventMeasure or you have made changes to data or corrected errors outside of the annotation files and this changed data is now the true copy of the data (e.g. in excel) you will have to import “Generic” files.

The files needed for a “Generic” upload are:

- \$_Count.csv
- \$_Length.csv

The column requirements for the \$_Count.csv and the \$_Length.csv file are in [Tables 3 and 4](#) respectively. Examples of the \$_Count.csv and the \$_Length.csv file are given in [Tables 5 and 6](#) respectively.

Table 3. Column requirements for the \$_Count.csv file.

Transposed (rows for columns) for formatting convenience.

Name	Format	Required
opcode	String	✓ <i>if opcodes were used to define a sample. DON'T include this column if it is not required to define a sample.</i>
period	String	✓ <i>if periods were used to define a sample. DON'T include this column if it is not required to define a sample.</i>
family	String	✓
genus	String	✓
species	String	✓
count	Integer	✓
stage	String (either AD, F, M, J, or blank)	x
code	Integer (CAAB code or AphialD)	x

Table 4. Column requirements for the \$_Length.csv file.

Transposed (rows for columns) for formatting convenience.

Name	Format	Required
opcode	String	✓ if opcodes were used to define a sample. DON'T include this column if it is not required to define a sample.
period	String	✓ if periods were used to define a sample. DON'T include this column if it is not required to define a sample.
family	String	✓
genus	String	✓
species	String	✓
count	Integer	✓
length_mm	Floating point number (in mm)	✓
stage	String (either AD, F, M, J, or blank)	x
code	Integer (CAAB code or AphialD). Can be blank	x

Table 5. First 5 rows of a \$_Count.csv file for an example stereo-BRUVs campaign.

opcode	family	genus	species	count	stage	code
NCB606	Nemipteridae	Pentapodus	porosus	2	AD	37347007
NCB606	Labridae	Coris	caudimacula	3	AD	37384092
NCB606	Nemipteridae	Pentapodus	porosus	2	AD	37347007
NCB607	Nemipteridae	Pentapodus	porosus	3	AD	37347007
NCB607	Labridae	Coris	caudimacula	10	AD	37384092

Table 6. First 5 rows of a `$_Length.csv` file for an example stereo-BRUVs campaign.

opcode	family	genus	species	count	length _mm	stage	code
NCB606	Nemipteridae	Pentapodus	porosus	1	150	AD	37347007
NCB606	Nemipteridae	Pentapodus	porosus	1	140	AD	37347007
NCB606	Labridae	Coris	caudimacula	1	190	AD	37384092
NCB606	Labridae	Coris	caudimacula	1	180	AD	37384092
NCB606	Labridae	Coris	caudimacula	1	165	AD	37384092

Uploading data

CheckEM does not save any data that is uploaded. Once the application is closed or refreshed all uploaded data will be erased from the applications memory. All uploaded data remains the property of the user.

How to upload

- Navigate to [CheckEM](#). We recommend using Google Chrome.

Setting the format of uploads

- On the *Upload data* tab, fill out the *Format of data* box ([Figure 6](#)):
 - **Type of upload:** If you are uploading unedited data from EventMeasure choose *EventMeasure*, otherwise choose *Generic* (Generic is not available for transect campaigns at the time of writing). For more information on types of upload see [Annotation data](#).
 - **Type of method:** If you are uploading data from a stationary sampling platform choose a *single point* (this will calculate MaxNs) sampling method. If you are uploading data from a transect based sampling method (e.g. stereo-DOVs or ROVs) choose *transect* (will not calculate MaxNs).
 - **How did you record the sample name:** The options for this question will depend on the type of method chosen (single point: either *opcode* OR *period*, transect: either *period* OR *opcode and period*). Choose the column names used in EventMeasure or the “Generic” files to identify a sample. See [defining a sample](#) for more information.
 - **Did you use periods?** You will only be asked this question if you are uploading *single point* data that uses *opcodes* to define the sample. If you used periods to standardise the sampling duration (e.g. Made sure each stereo-BRUV deployment was 60 minutes long) then choose *Yes* otherwise choose *No*.

- Once you have answered the above questions **click on the ‘Select directory...’ box.**
- **Navigate to the folder/directory** where you have saved your sample metadata and annotation files. CheckEM requires the sample metadata and chosen annotation files (EventMeasure or “Generic”) to be located in one folder.

Format of data

Choose the type of upload:

☒ EventMeasure ☐ Generic (only single point methods)

Choose the type of sampling method used:

☒ Single point e.g. BRUV & BOSS ☐ Transect e.g. DOV & ROV

How did you record the sample name in EventMeasure:

☒ OpCode ☐ Period

Did you use periods to standardise the sampling duration?

☒ Yes ☐ No

Select directory with sample metadata and EM exports

Figure 6. Setting the format of data to be uploaded.

Setting the life history information and spatial zoning to use

- On the *Upload data* tab, fill out the *Life history information...* box ([Figure 7](#)):
 - **Keep or generate ‘status’ column:** If you have filled out the status column correctly you can keep the status column, or you can choose to generate the column based on the shapefiles hosted in CheckEM (**caution:** there are some known issues in CAPAD and the WDPa, we suggest using the uploaded status column).
 - **Life history and regions:** Choose the Australian life history sheet based on the Codes for Australian Aquatic Biota (CAAB) or the Global list based on fishbase and the World Register of Marine Species (WoRMS). The Australian list uses the CAAB distributions but at [Australia’s marine bioregion](#) scale, the global list uses the fishbase distributions at the [Food and Agriculture Organisation of the United Nations Major Fishing Areas](#) scale.
 - **Marine region for your data:** Once you have chosen if you want to use the Australian or Global marine regions, you need to choose if you require one marine region per campaign or a marine region per sample. If your data crosses multiple marine regions (e.g. entire West Coast of Australia or Multiple Countries) you should choose a marine region per sample.

Life history Information, Marine Spatial Planning & Marine Regions

Would you like to keep the uploaded 'status' column or generate one from shapefiles?

☒ Uploaded

☐ Generate from shapefiles

Which life history list & regions would you like to check your annotations against?

☒ Australian List (Based on the Codes for Australian Aquatic Biota) & Australian Marine Regions

☐ Global List (Based on FishBase and the World Register of Marine Species) & FAO Major Fishing Areas

How would you like to retrieve the marine region for your data?

☒ One marine region per campaignID (uses the average lat/lon of a campaign)

☐ One marine region per sample (uses the lat/lon for each sample, takes much longer to run but recommended if your data crosses multiple marine regions)

Figure 7. Setting the life history information to use.

- You can check that the upload and marine regions have worked by inspecting the preview of the sample metadata. A new *marine_region* column will be in your metadata. This column will be used to identify species that haven't been observed in that marine region before ([Figure 8](#)).

date_time	site	location	status	depth_m	successful_count	successful_length	zone	marine_region
2022-05-22T10:03:24+08:00			No-take	93.9	Yes	Yes	National Park Zone (IUCN II)	North-west
2022-05-22T14:16:25+08:00			No-take	77.3	Yes	Yes	National Park Zone (IUCN II)	North-west
2022-05-20T14:34:36+08:00			No-take	78.3	Yes	Yes	National Park Zone (IUCN II)	North-west

Figure 8. Sample metadata preview in CheckEM.

- If you are missing any required columns in the sample metadata CheckEM will add the columns but they will be blank and you will receive an error message (as shown in [Figure 1](#)). **Any errors in the metadata should be fixed before proceeding.**
- If you haven't received any error messages you are now ready to start checking your data.

Reading the QC assessments

Each QC assessment in CheckEM is displayed in a coloured box. CheckEM uses a 'traffic light' system to colour the boxes which allows users to easily identify issues that need their attention.

The colours correspond to:

- **Blue**: General information (e.g. the number of samples or fish observed).
- **Green**: Passed the QC assessment (e.g. zero samples are missing metadata).
- **Orange**: Should be reviewed, possibly not a problem (e.g. a sample in the sample metadata is missing fish).
- **Red**: Did not pass the assessment and needs to be reviewed (e.g. fish outside of the predefined range).

We suggest clicking through the tabs on the left hand side of the application and checking each of the QC assessment boxes:

- *Check Metadata & periods* →
- *Create & check MaxN* (only displayed if uploading [single point](#) data) →
- *Check length & 3D points* →
- *Compare MaxN & length* (only displayed if uploading [single point](#) data) →
- *Create & check mass*

Most QC boxes can be clicked, which will display a box detailing more information, see [Figure 9](#) for an example. The further information can be downloaded by clicking the *Download as csv* button (or [download all errors](#) at once on the *Download data and QC score* tab).

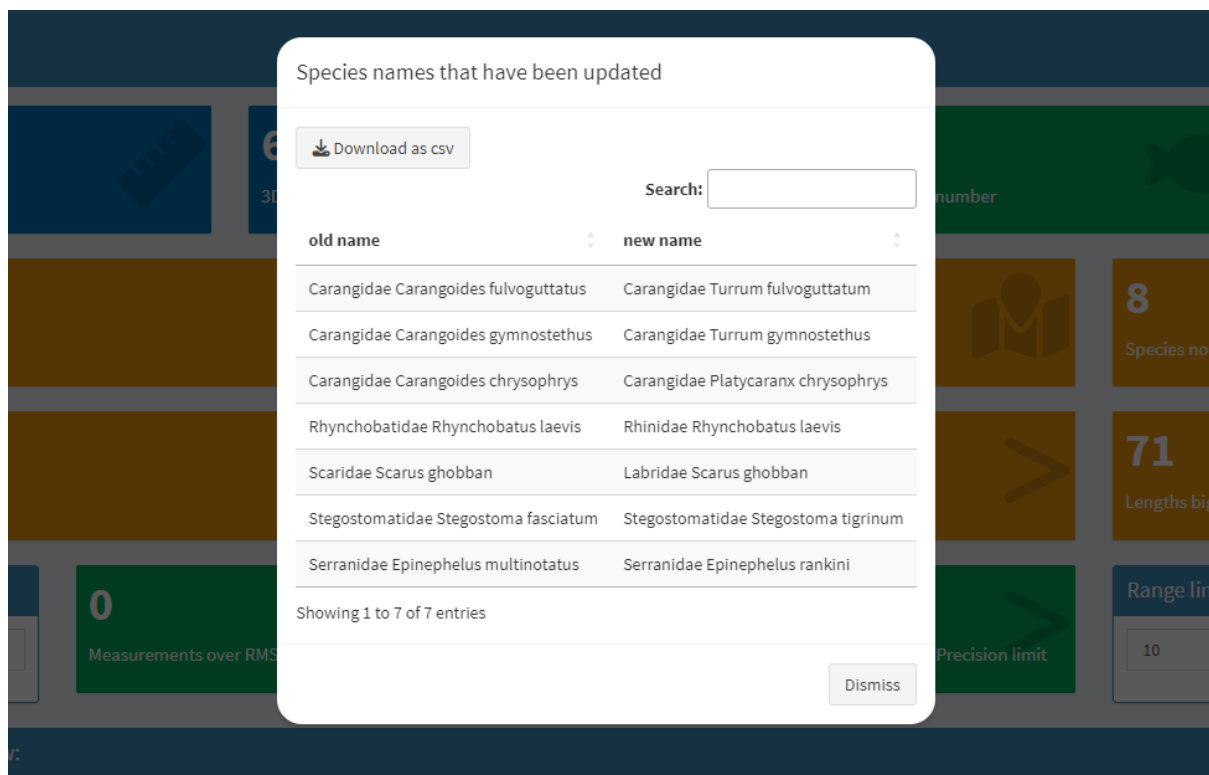


Figure 9. Example of more detail that is shown when clicking a QC assessment box. In this example the “Species names updated” box was clicked.

In the following sections we describe each of the QC assessments on each tab. Not every QC assessment is appropriate for each sampling method (point or transect) or for each type of upload (EventMeasure or “Generic”), [Table 7](#), [Table 8](#), [Table 9](#) and [Table 10](#) indicate if the QC assessment is performed for each sampling method and type of upload.

Check Metadata & Periods

On the *Check Metadata & periods* tab up to 11 QC assessment boxes will be displayed (the number is dependent on the selections made on the *upload* tab, [Figure 10](#)). Each assessment is described in [Table 7](#). A map of the sample metadata is displayed ([Figure 11](#)) to allow users to conduct a visual check of the distribution of samples (make sure none are on land or in a different hemisphere). Clicking on the marker icon will display a pop-up of the metadata for that sample). For Australian datasets the spatial zoning will also be displayed, clicking on the polygons will display the type and IUCN category of the zone (e.g. National Park (IUCN II), Multiple Use (IUCN VI) etc.).

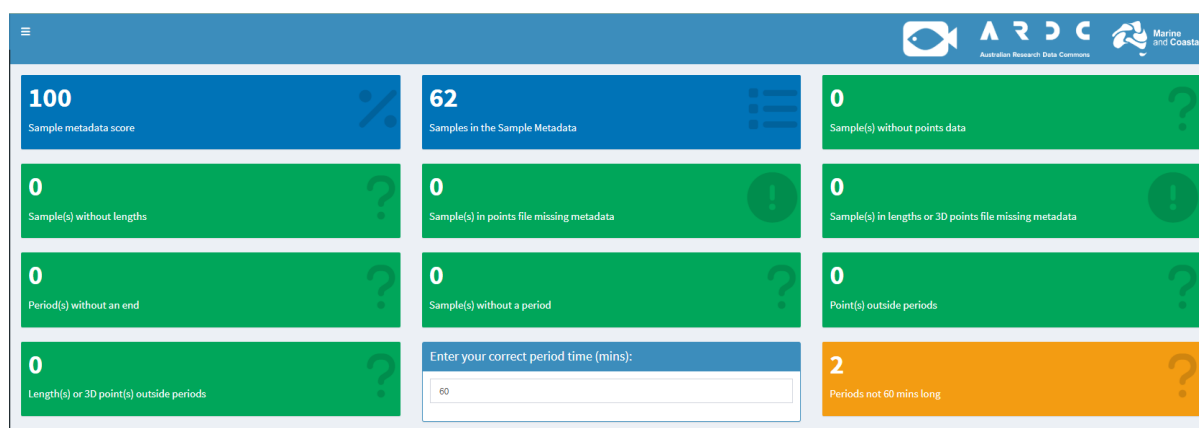


Figure 10. Screenshot from the *Check Metadata & periods* tab showing the QC assessments.

In this example a stereo-BRUVs campaign that used periods to define the sampling duration was uploaded using EventMeasure database outputs.



Figure 11. Screenshot from the Check Metadata & periods tab of a map.

Sample Metadata Score

The *Sample metadata* score indicates if your sample metadata meets the criteria outlined in the formatting requirements. For example if you have a dataset with 50 samples, and you are missing the depth information for 10 samples (the cells are blank) you will receive a score of 80% (as only 80% of your sample metadata includes all the necessary metadata).

$$\frac{\text{number of rows in the sample metadata **without** missing values or formatting errors}}{\text{total number of rows in the sample metadata}} \times 100$$

Table 7. QC assessments on the Metadata tab.

Assessment	Point		Transect	
	Gen	EM	Gen	EM
The Sample metadata score indicates if your sample metadata meets the criteria outlined in the formatting requirements. For example if you have a dataset with 50 samples, and you are missing the depth information for 10 samples (the cells are blank) you will receive a score of 80% (as only 80% of your sample metadata includes all the necessary metadata).	✓	✓	☐	✓
Number of samples in the sample metadata. This is the total number of unique samples in the sample metadata (it should also be the number of rows).	✓	✓	☐	✓

Number of sample(s) without points or count data. This could be due to samples not observing fish (not an error) or a sample that should be marked as Successful.count = No. It could also be due to a sample name spelt incorrectly in the count/points or sample metadata file.	✓	✓		
Number of sample(s) without length or 3D measurements. Could be due to samples not observing fish (not an error) or a sample that should be marked as Successful.length = No. It could also be due to a sample name spelt incorrectly in the count/points or sample metadata file.	✓	✓	<input type="checkbox"/>	✓
Number of sample(s) in points or count data missing metadata. This could be due to a spelling mistake in the sample name in the count/points or sample metadata file or a row missing from the sample metadata.	✓	✓		
Number of sample(s) in length or 3D point (EM only) data missing metadata. This could be due to a spelling mistake in the sample name in the count/points or sample metadata file or a row missing from the sample metadata.	✓	✓	<input type="checkbox"/>	✓
Number of period(s) without an end (<i>this will only be displayed if you indicated that you used periods to define the sampling duration</i>). The period end time has not been set. This will also cause errors in calculation the period length.		✓		✓
Number of sample(s) without periods (<i>this will only be displayed if you indicated that you used periods to define the sampling duration</i>). Samples in the sample metadata that do not have a period.		✓		✓
Number of point(s) outside a period (<i>this will only be displayed if you indicated that you used periods to define the sampling duration</i>). The number of point annotations that are not within a period.		✓		✓
Number of length measurement(s) or 3D point(s) outside periods (<i>this will only be displayed if you indicated that you used periods to define the sampling duration</i>). The number of length or 3D point annotations that are not within a period.		✓		✓
Number of period(s) not XX minutes long. (<i>this will only be displayed if you indicated that you used periods to define the sampling duration</i>). After the user sets the correct period length (e.g. 60 minutes) this error will display all periods that are not that length.		✓		

Create & Check MaxN (point methods only)

On the *Create & Check MaxN* tab up to five QC assessment boxes will be displayed (the number is dependent on the selections made on the *upload* tab, [Figure 12](#)). Each assessment is described in [Table 8](#).

A plot of the 15 most abundant species is created ([Figure 12](#)), the number can be changed using the *Species to plot* number input. CheckEM also creates a spatial 'bubble-plot' for each species, with a circle for each sample, the bigger the sample the higher the abundance at that sample ([Figure 13](#)). Hovering over the circle will display the MaxN for that sample. To update the species plotted choose a new one in the dropdown (the dropdown is ordered by total abundance with the most abundant species appearing first in the dropdown) above the map.

The species dropdown will also update the four bar plots showing the average abundance per sample for each status, location and site (taken from the sample metadata) and zone (from the spatial zoning chosen on the *upload* tab).

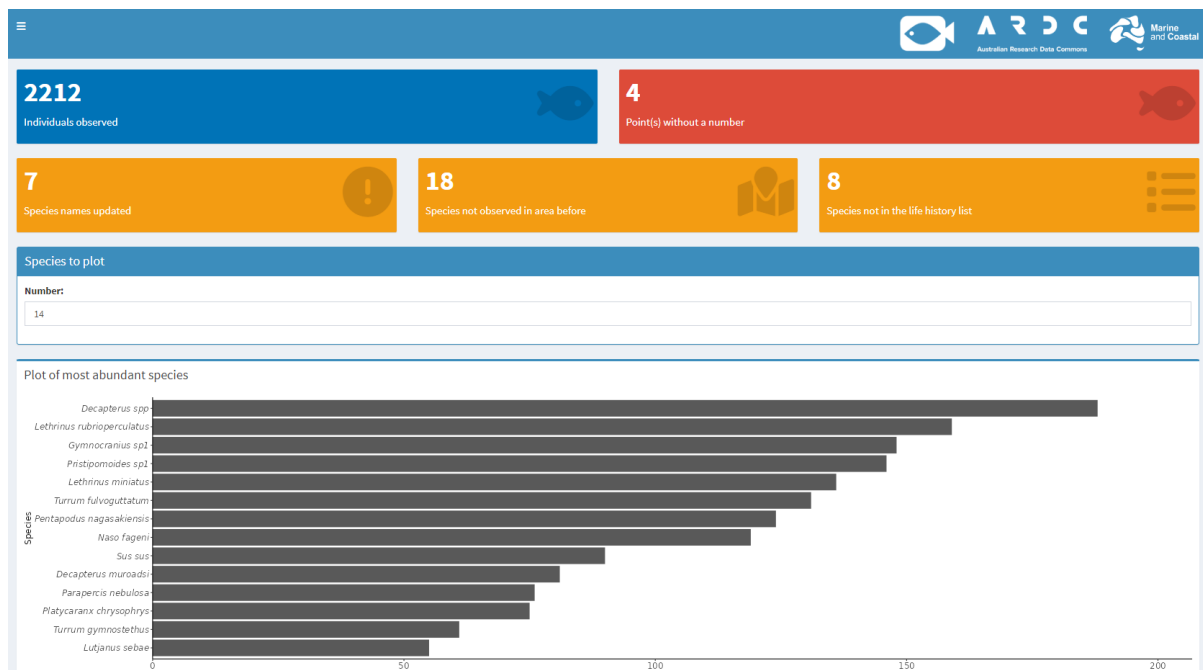


Figure 12. Screenshot from the *Create & Check MaxN* tab showing the QC assessments.

In this example a stereo-BRUvs campaign that used periods to define the sampling duration was uploaded using EventMeasure database outputs.

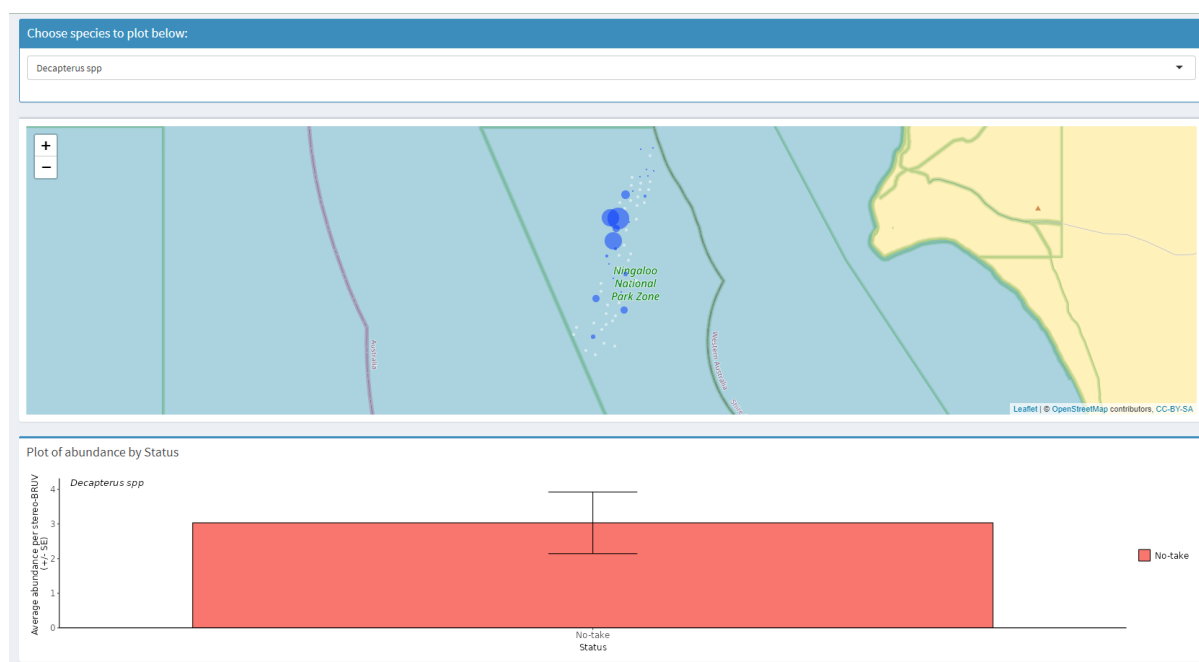


Figure 13. Screenshot from the Create & Check MaxN tab showing the spatial and status plot.

Table 8. QC assessments on the MaxN tab.

Assessment	Point		Transect	
	Gen	EM	Gen	EM
Number of individuals observed. This is a sum of the MaxNs of every species for every sample uploaded.	✓	✓		
Number of point(s) without a number. Most of the time this is annotators adding a point for an animal that is not a fish that may be of interest (e.g. a squid or turtle), but this should be reviewed in case it is a fish that was mistakenly put down without a number.		✓		
Number of species names updated using the list of synonyms. This could be due to a name change (e.g. In Australia <i>Pagrus auratus</i> changed to <i>Chrysophrys auratus</i>) or it could be a common spelling mistake.	✓	✓		
Number of species not observed in the area before from the chosen life history list. This could be due to a misidentified species (for example an Eastern Fiddler Ray on the west coast of Australia), a range shift, or the species distribution in CAAB or WoRMs not being accurate. The literature should be consulted or species identifications double checked.	✓	✓		

Number of species not in the chosen life history list. This could be due to the species not being listed on CAAB or WoRMS or a spelling mistake. Please check that you have spelled the name correctly and check either CAAB or WoRMS if the family, genus, and species is listed.	✓	✓		
--	---	---	--	--

Check Length & 3D points

On the *Check Length & 3D points* tab up to 12 QC assessment boxes will be displayed (the number is dependent on the selections made on the *upload* tab, [Figure 13](#)). Each assessment is described in [Table 9](#).

CheckEM creates a length histogram for each species, with a vertical line for the maximum length listed on FishBase (if available) and 15% and 85% of this value for analysts to visualise the length distribution of each species and identify any abnormalities in the length data ([Figure 15](#)). To update the species plotted choose a new one in the dropdown (the dropdown is ordered by total abundance with the most abundant species appearing first in the dropdown) above the histogram. Another histogram faceted by status is created to see if this pattern is consistent inside and outside of no-take areas.

The species dropdown will also update the two box plots showing the length for each status (taken from the sample metadata) and zone (from the spatial zoning chosen on the *upload* tab).

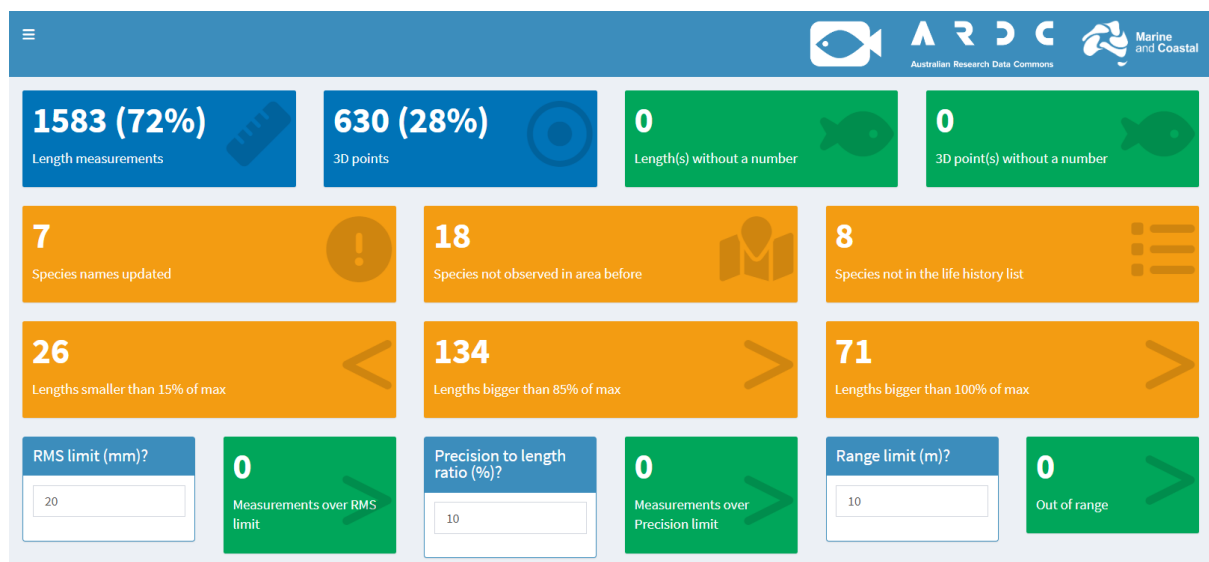


Figure 14. Screenshot from the *Check Length & 3D points* tab showing the QC assessments.

In this example a stereo-BRUvs campaign was uploaded using EventMeasure database outputs.

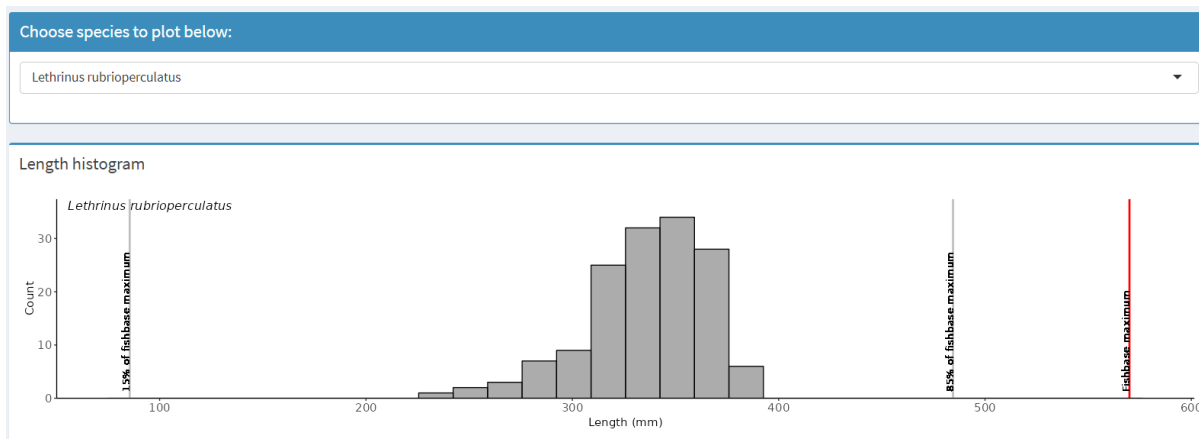


Figure 15. Screenshot from the Check Length & 3D points tab showing an example length histogram.

Table 9. QC assessments on the Length tab.

Assessment	Point		Transect	
	Gen	EM	Gen	EM
Number of length measurement(s) without a number. All length measurements should have a number.		✓		✓
Number of 3D-point(s) without a number. These may be 3D points that were used to synchronise the cameras or 3D points where the number is not meant to be missing. All 3D points of fish should have a number.		✓		✓
Number of species names updated using the list of synonyms. This could be due to a name change (e.g. In Australia <i>Pagrus auratus</i> changed to <i>Chrysophrys auratus</i>) or it could be a common spelling mistake.	✓	✓	<input type="checkbox"/>	✓
Number of species not observed in the area before from the chosen life history list. This could be due to a misidentified species (for example an Eastern Fiddler Ray on the west coast of Australia), a spelling mistake, a range shift, or the species distribution in CAAB or WoRMs not being accurate. The literature should be consulted or species identifications double checked.	✓	✓	<input type="checkbox"/>	✓
Number of species not in the chosen life history list. This could be due to the species not being listed on CAAB or WoRMS or a spelling mistake. Please check that you have spelled the name correctly and check either CAAB or WoRMS if the family, genus, and species is listed.	✓	✓	<input type="checkbox"/>	✓

Number of length measurement(s) less than 15% of the maximum length for that species on FishBase. This is to double check any measurements that are too small.	✓	✓	<input type="checkbox"/>	✓
Number of length measurement(s) greater than 85% of the maximum length for that species on FishBase. This is to check any measurements that are at the upper limits of that species maximum length. You can provide evidence of species with a greater known length by entering the new maximum length and the source into the google form hosted in CheckEM.	✓	✓	<input type="checkbox"/>	✓
Number of length measurement(s) greater than the maximum length for that species on FishBase. This is to check any measurements that are greater than the known maximum length. You can provide evidence of species with a greater known length by entering the new maximum length and the source into the google form hosted in CheckEM.	✓	✓	<input type="checkbox"/>	✓
Number of 3D points and length measurements over the user-defined RMS limit (default is set to 20 mm). These will need to be fixed in the EMObs. Setting the recommended length rules in EventMeasure will prevent these.		✓		✓
Number of length measurements over the user-defined precision limit (default is set to 10%). These will need to be fixed in the EMObs. Setting the recommended length rules in EventMeasure will prevent these.		✓		✓
Number of measurements outside of the user-defined range (default is set to 10 m). These will need to be fixed in the EMObs. Setting the recommended length rules in EventMeasure will prevent these.		✓		✓
Number of measurements outside the user-defined width of the transect (default is set to 5 m). These will need to be fixed in the EMObs. Setting the recommended length rules in EventMeasure will prevent these.				✓

Compare MaxN & length (point methods only)

We recommend the [length measurement procedures outlined in the field and video annotation guide for stereo-BRUVs](#) for all stereo-video methods where MaxN is used. The guide outlines that if fish cannot be measured, a 3D point measurement may be used for annotation, which records the 3D location of the fish to ensure it is within the sampling area. To create a relative abundance metric standardised to a consistent sample area, abundance should be summed from the lengths and 3D points at the MaxN for each species. Therefore

it is important to check if the MaxN for each species in each sample is equal to the sum of the length measurements and the 3D points.

On the *Compare MaxN & length* tab up to three QC assessment boxes will be displayed (the number is dependent on the selections made on the *upload* tab, [Figure 16](#)). Each assessment is described in [Table 10](#).

The first plot compares the MaxN to the number of measurements (length measurements + 3D point measurements) by plotting the MaxN of each species from each sample against the number of length measurements ([Figure 16](#)). Each dot represents a species in a sample, if the MaxN = the number of length measurements the dot will be on the red line. If a dot is below the red line the MaxN is greater than the number of length measurements and if the dot is above the line there are more length measurements than the MaxN for that species in that sample.

Arguably the most important data from stereo-video data is the length measurements, it allows you to estimate age, biomass or fishing pressure. The second plot visualises the ratio of length measurements to 3D point measurements for each sample ([Figure 17](#)). The green segment of the bar represents the number of length measurements, the blue represents the number of 3D points, the pink represents individuals in the MaxN that were not measured. A * denotes a sample where the sum of the length measurements and 3D points is greater than the MaxN. If you are following the [length measurement procedures outlined in the field and video annotation guide for stereo-BRUVs](#) these any MaxNs that haven't been measured should be, and any excess measurements should be removed. A similar plot displaying the same data but as a proportion is displayed to easily compare if there are any differences in the ratio of length measurements to 3D points across the samples ([Figure 18](#)). An option to facet the plots by *observer_length* is available to investigate if the proportion of length measurements to 3D points differs between length analysts.

The above three plots are available for each species in the campaign. Use the dropdown to select a species.

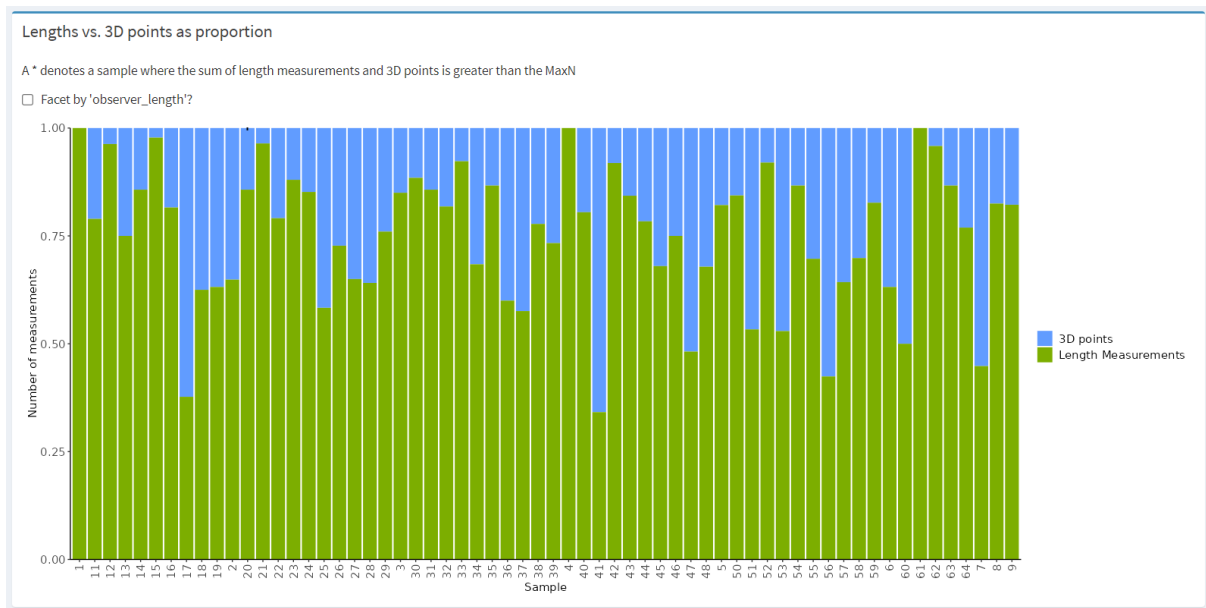


Figure 18. Screenshot from the Compare MaxN and length tab showing the length measurements versus 3D points as a proportion plot.

Table 10. QC assessments on the Compare count and length tab.

Assessment	Point		Transect	
	Gen	EM	Gen	EM
Number of rows in the count dataframe where the MaxN doesn't equal the number of lengths + number of 3D points. This is to ensure that every individual at the time of MaxN (for that species) in every sample has been measured.	✓	✓		
Number in count VS. Number in length (+3D Point if it is an EventMeasure upload) score. This is a percentage score of the above assessment. A score of 100% indicates that every individual at the time of MaxN has been length measured or 3D pointed.	✓	✓		
Percent of measurements (length measurements and 3D point measurements) that have length. If you have not used 3D points to ensure that every individual at the time of MaxN has been accounted for then ignore this check. If you have, we suggest aiming for a score above 70%, a score below this could indicate low visibility, habitat obstruction or other issues preventing length measurements and should be investigated.		✓		

Downloading Data and QC score

Once you are ready to download a report of all the 'errors' (or the final 'cleaned' data), add a project name ([Figure 19](#)).

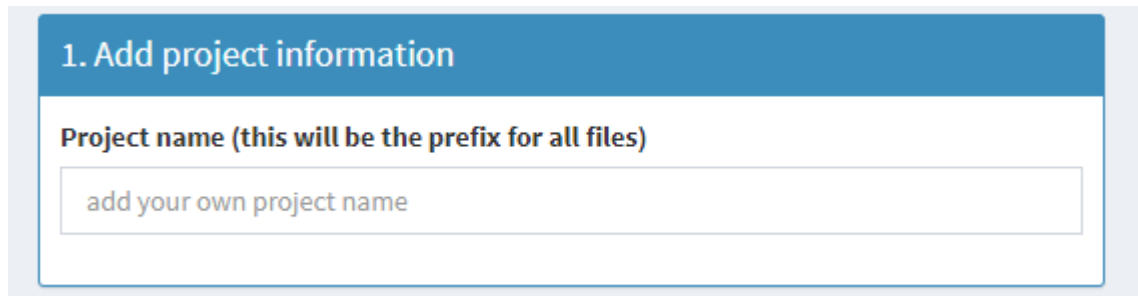


Figure 19. Adding a project name.

Downloading report of all potential 'errors'

A report of all potential errors is available to download. Before downloading the 'error' report, you need to set some limits, the limits will depend on what type of sampling method you have used and if you have used periods to define the sampling duration ([Figure 20](#)). For example, for a *single point* campaign, using EventMeasure exports, that has used periods to define the sampling duration you will need to set the correct period time, the acceptable range limit, the RMS limit and the precision:length ratio limit. Once you have entered the limits, click the download all errors in the EMObs button to download a CSV file of all the potential errors. The list is ordered by opcode and period so analysts can fix the issues in one EMObs at a time. Once you have edited the EMObs or generic data to fix up the errors, we recommend running the data through CheckEM once more to ensure you have removed all the errors.

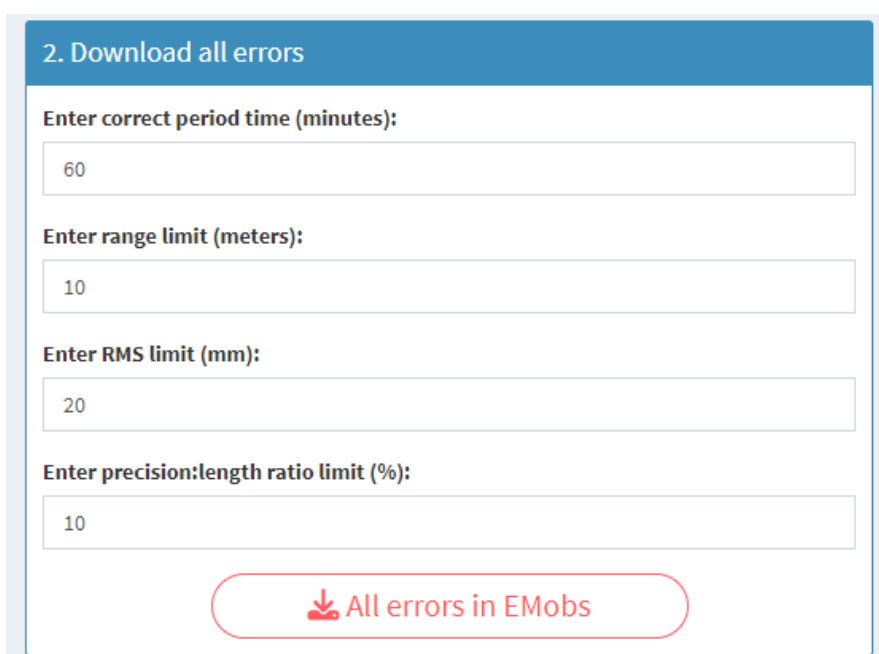


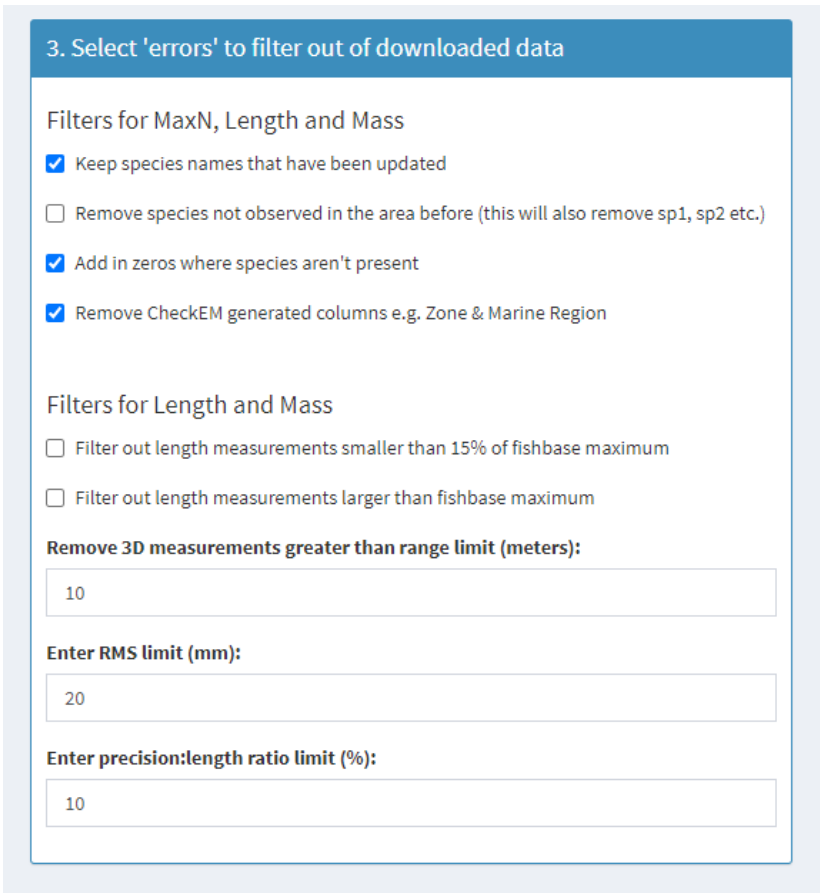
Figure 20. Setting the limits to download all potential errors.

Downloading final data

Once you are happy that you have tidied the data and double checked all the errors have been fixed as much as you can outside of CheckEM, you can download the formatted data. Although we recommend fixing the data outside of CheckEM, you can use CheckEM to remove some “errors”, for example species that haven’t been observed in the area before or filter out length measurements that are too small or too big (based on the FishBase maximum sizes). Filtering out these errors is turned off by default, to filter them out you need to check the checkbox next to the error you would like to remove ([Figure 21](#)).

The default download in CheckEM adds zeros for every species for every sample that it was not present in. If the default is selected, the files downloaded will have the sample metadata included in them e.g. the count.csv file downloaded will have the count columns as well as the sample metadata columns. If adding the zeros is not selected, the sample metadata file will be downloaded as a separate csv file.

Clicking the Download all files button will download a zip file containing files for each campaign that was uploaded.



3. Select 'errors' to filter out of downloaded data

Filters for MaxN, Length and Mass

- ☒ Keep species names that have been updated
- ☐ Remove species not observed in the area before (this will also remove sp1, sp2 etc.)
- ☒ Add in zeros where species aren't present
- ☒ Remove CheckEM generated columns e.g. Zone & Marine Region

Filters for Length and Mass

- ☐ Filter out length measurements smaller than 15% of fishbase maximum
- ☐ Filter out length measurements larger than fishbase maximum

Remove 3D measurements greater than range limit (meters):

Enter RMS limit (mm):

Enter precision:length ratio limit (%):

Figure 21. Setting the errors to remove from the final downloaded data.

Quality Control Score

The quality control score infographic is a quick and easy way to visualise your data against five key QC assessments. This is only available for *single point* uploads, with *transect* based uploads coming soon. We have chosen the five assessments that we think are the most important. The five scores are defined below:

- The **Sample metadata** score indicates if your sample metadata meets the criteria outlined in the formatting requirements. For example if you have a dataset with 50 samples, and you are missing the depth information for 10 samples (the cells are blank) you will receive a score of 80% (as only 80% of your sample metadata includes all the necessary metadata).

$$\frac{\text{number of rows in the sample metadata *without* missing values or formatting errors}}{\text{total number of rows in the sample metadata}} \times 100$$

- The **Count** score indicates if all the samples in your count data have a match in the sample metadata. For example if you have a count dataset with 10 samples, and three samples do not have a match in the metadata because (they are misspelt) you will receive a score of 70% (as only 70% of your count data matches the sample metadata). Matching the sample metadata is important because you need to be able to match both datasets together to analyse patterns in depth, no-take status or locations.

$$\frac{\text{number of Counts} - \text{Number of Counts where the sample does not match the sample metadata}}{\text{number of Counts}} \times 100$$

- The **Count vs. length** score indicates if the MaxNs are equal to the number of length measurements + number of 3D point measurements. A score of 100% indicates that every individual at the time of MaxN has been length measured or 3D pointed once. A score of 0% means that there are either too many or not enough length measurements compared to the MaxN.

$$\frac{\text{number of MaxNs} - \text{number of times MaxN does not equal the number of measurements}}{\text{number of MaxNs}} \times 100$$

- The **% Length score** indicates the proportion of measurements that have a value in length. If you have not used 3D points to ensure that every individual at the time of MaxN has been accounted for then ignore this check. If you have, we suggest aiming for a score above 70%, a score below this could indicate low visibility, habitat obstruction or other issues preventing length measurements and should be investigated. If you upload “generic” data you will not have this segment present in the QC plot.

$$\frac{\text{total number of length measurements}}{\text{total number of length measurements} + \text{total number of 3D point measurements}} \times 100$$

- The **Length score** if all the samples in your length data have a match in the sample metadata. For example if you have a length dataset with 10 samples, and three samples do not have a match in the metadata because (they are misspelt) you will receive a score of 70% (as only 70% of your length data matches the sample metadata). Matching the sample metadata is important because you need to be able to match both datasets together to analyse patterns in depth, no-take status or locations.

$$\frac{\text{number of Lengths} - \text{Number of Lengths where the sample does not match the sample metadata}}{\text{number of Lengths}} \times 100$$

Three example quality control score plots are shown below ([Figure 22](#)). In example A and C the user has uploaded sample metadata where every row is formatted correctly and is not missing data. In example B 96.77% of the samples in the sample metadata have all the required information. The solid green line represents the score users should aim for, in example A and C this goal is met, in example B the goal is not met.

In example A and B, every sample in the count data has a match in the sample metadata. In example B, only 99.14% of samples in the count data have a match in the sample metadata. The solid yellow line represents the score users should aim for.

In all examples, the user has uploaded data where a MaxN does not equal the number of measurements, therefore they have not received a score of 100% for Count vs. Length and the segment is below the solid orange line.

In example A, 71.53% of the measurements have length, this is above the goal of 70%. In example B, only 59.57% of measurements had length, this is below the target of 70%. The dark blue segment is not included in example C because it is a “generic” upload (does not have 3D points).

In example A and C, every sample in the length data has a match in the sample metadata (light blue segment). In example B, 99.66% of samples in the length data have a match in the sample metadata. The solid light blue line represents the score users should aim for.

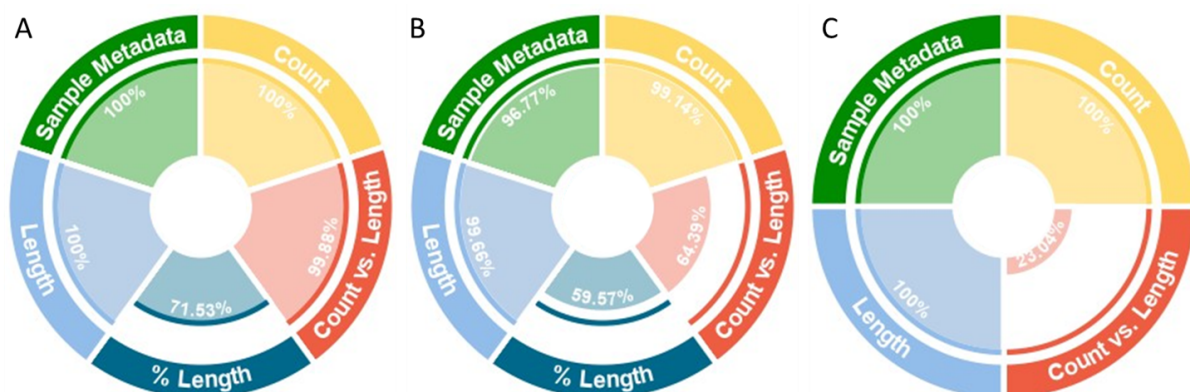


Figure 22. Example Quality Control Score plots.

The five quality control assessments are shown as different segments of the circle. The solid lines in each segment represent the target for that quality control assessment.

Editing Maximum Lengths of Fish Species

CheckEM uses maximum lengths for fish species from FishBase, sometimes this length is smaller than that listed on [Fishes of Australia](#) or other reputable sources. A google form is hosted in CheckEM to collect information on maximum lengths that are greater than those on FishBase. The form is available through the CheckEM tab *Edit maximum lengths* or through [this link](#). To fill out the form you will need the Family, Genus and Species of the fish species to update, the new maximum length (in centimetres), the type of length measure (e.g. Total Length or Fork Length) and a link to the source material ([Figure 23](#)). Any maximum lengths submitted will not be immediately reflected in CheckEM and will go through a verification process.

Update a species maximum length

Please use this form if you believe the fishbase maximum length needs to be updated.
Please supply a reference for the update.

NOTE: these updates will not be reflected in CheckEM immediately and will need to be verified.

brooke.gibbons@marineecology.io [Switch accounts](#)

Not shared

* Indicates required question

Full name *

Your answer

Organisation *

Your answer

Family *

Your answer

Genus *

Your answer

Species *

Your answer

New maximum length (in cm) *

Your answer

Type of length measure *

☐ FL

☐ TL

☐ SL

☐ WD

☐ OT

☐ NG

☐ Other:

Please provide a link to reference information e.g. website or paper *

Your answer

Submit

Clear form

Never submit passwords through Google Forms.

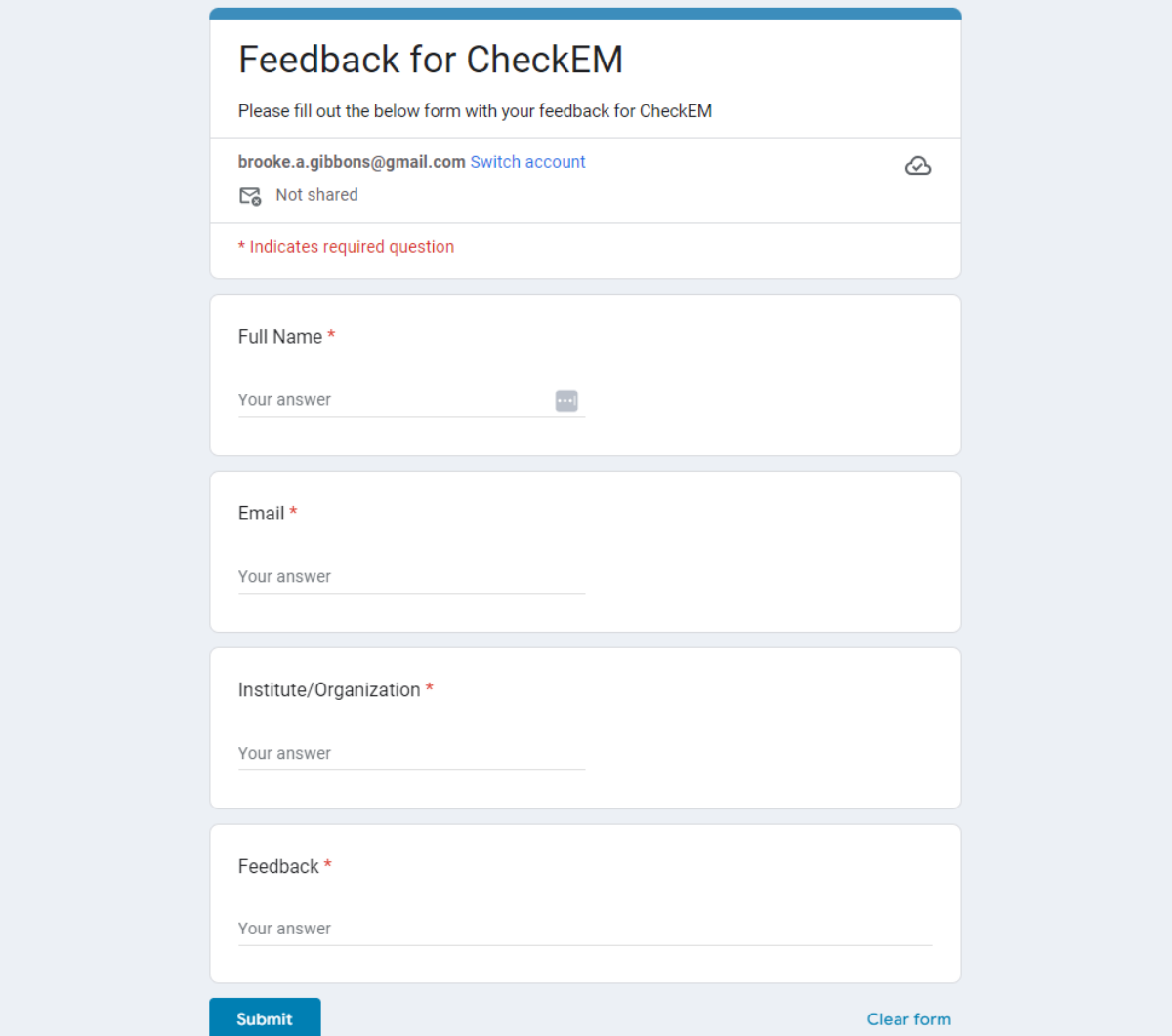
This form was created inside MarineEcology.io. [Report Abuse](#)

Google Forms

Figure 23. The form to edit species maximum lengths.

Feedback

Our development process has been heavily influenced by user feedback. We strongly encourage and incorporate feedback on CheckEM, a feedback form is available on the *Feedback* tab in CheckEM ([Figure 24](#)) or through this [link](#).



Feedback for CheckEM

Please fill out the below form with your feedback for CheckEM

brooke.a.gibbons@gmail.com [Switch account](#)

Not shared

* Indicates required question

Full Name *

Your answer

Email *

Your answer

Institute/Organization *

Your answer

Feedback *

Your answer

Submit [Clear form](#)

Figure 24. The form to give feedback on CheckEM.