eReefs R data handling [SEC=Official]-20230725_141729-Meeting Recording

Adam Shand, Eric Lawrey, Brandon Ford, Marc Hammerton

Meeting starts 2:00pm

(Introductions, pleasantries, agreement to record). One key point missed: Adam asks "When searching for variables, what resources are available for reference variables names. print.nc() returns all aspects of the NetCDF not just variable names. Response is when recording begins.

Recording begins 2:17pm

Question 1: Best resource for understanding all the variables that are available?

Answer 1: There is a tutorial online that has most variables (about 110 of the 250) for BGC model (note that the Hydro model only has ~10 so not such an issue for that model as you can manually search the print.nc() output)). Tutorial with list of variables.

Some of the variables are not really of use as they act as intermediates to other variables, or are used to derive other variables, or are essentially repeats of variables. So, the 110ish that are listed in the link are the important ones. There is also the AIMS GitHub Repo.

Further, the model data are provided from multiple locations. Raw curvilinear data is available for download from the NCI THREDDS service (this has all 250+ variables), regridded and aggregated data is available from the AIMS eReefs THREDDS service (this has only what AIMS considers the important variables ~160), or a csv time series extraction can be created from the AIMS Data Extraction tool.

Note: Later, in Question 3, we discuss the pros and cons of each of these data services.

Question 2: Currently Murray Logan's script (we have the pdf report output from this script) produces water quality (inshore, midshelf, and offshore) for the Reef Report Card focusing on the Secchi Depth and Chlorophyll-a variables (see here for example). Is this script something we can tack a few more variables onto, to expand what it reports on?

(Context, the RRC network reports on more than just secchi and chla in the inshore zone).

Answer 2: Not exactly. The model data used for the Reef Report Card is not the same model data that is publicly available for download. The model run (and data produced) is optimised specifically for the secchi and chla variables that are being used in the report card. CSIRO integrates additional information such as satellite imagery, catchment runoff, and weather to improve the reliability of the output. Therefore, when doing this they might not be checking the outputs of other variables are as accurate (as they aren't going to be used anyway). You also must assess if you are making the metric better or worse by adding additional variables in, particularly those that might not be as carefully constrained.

Note: Although we (the RRC) don't currently use the outputs of the model we are discussing above, it does seem like a viable option to simply take the results that are produced for the Reef Report Card and recycle them straight into our reports. However, by creating this specific model run, it means that data is no longer provided in real time and reporting cycles are then limited to the Reef Report Card cycle of 2 years. This is in contrast to the publicly available models which are provided in near real time (hydro is up to date, the BGC model is being reworked so currently only goes to 2019).

Question 2.1: As an alternative, what would the workflow look like to assess the publicly available, near real time model? For example, comparing water samples for Cleveland Bay against model outputs for Cleveland Bay?

Answer 2.1: Yes, you would definitely want to compare "real" samples against the modelled eReefs data before deciding how to use the data. Particularly when getting down to smaller areas as this can also get a bit complicated because you might be looking at one sample point against a 4km-by-4km grid cell. There are a lot of dynamics that can happened across the grid cell, particularly inshore these gradients can be quick dramatic. Given the resolution of the model compared to a single point, even if the model was perfect, it would never align with the data.

We would recommend looking at using the data extraction tool (link in Q1) to pull time series of relevant variables and compare them to the actual water quality data. You might find that eReefs does an ok job on average but doesn't pick up certain small scale/small resolution dynamics. In which case, if you only report on more reliable parameters that it does ok on at a larger scale then it is ok to use.

The issue is for the model to get the water quality parameters just right, there are many steps in a row that have to be just right. The hydrodynamic model has to move the water in the just the right way, the water coming off the land has to come off the right way, the wind has to be going in exactly the right direction, and when going to smaller and smaller scales that is harder and harder to pick up. The wind data that is worked in has a 12km resolution. As an example, when looking at flood plumes that tend to hug the coast there might be 6km of variable in where the model estimates the edge of the plume and where the plume actually is. Which can result in quite significant errors at a small scale. However, if you are looking at the average of a large area, or the average over a longer period of time those dynamics tend to average out and the model can give a good estimate. And you might even create your index using eReefs.

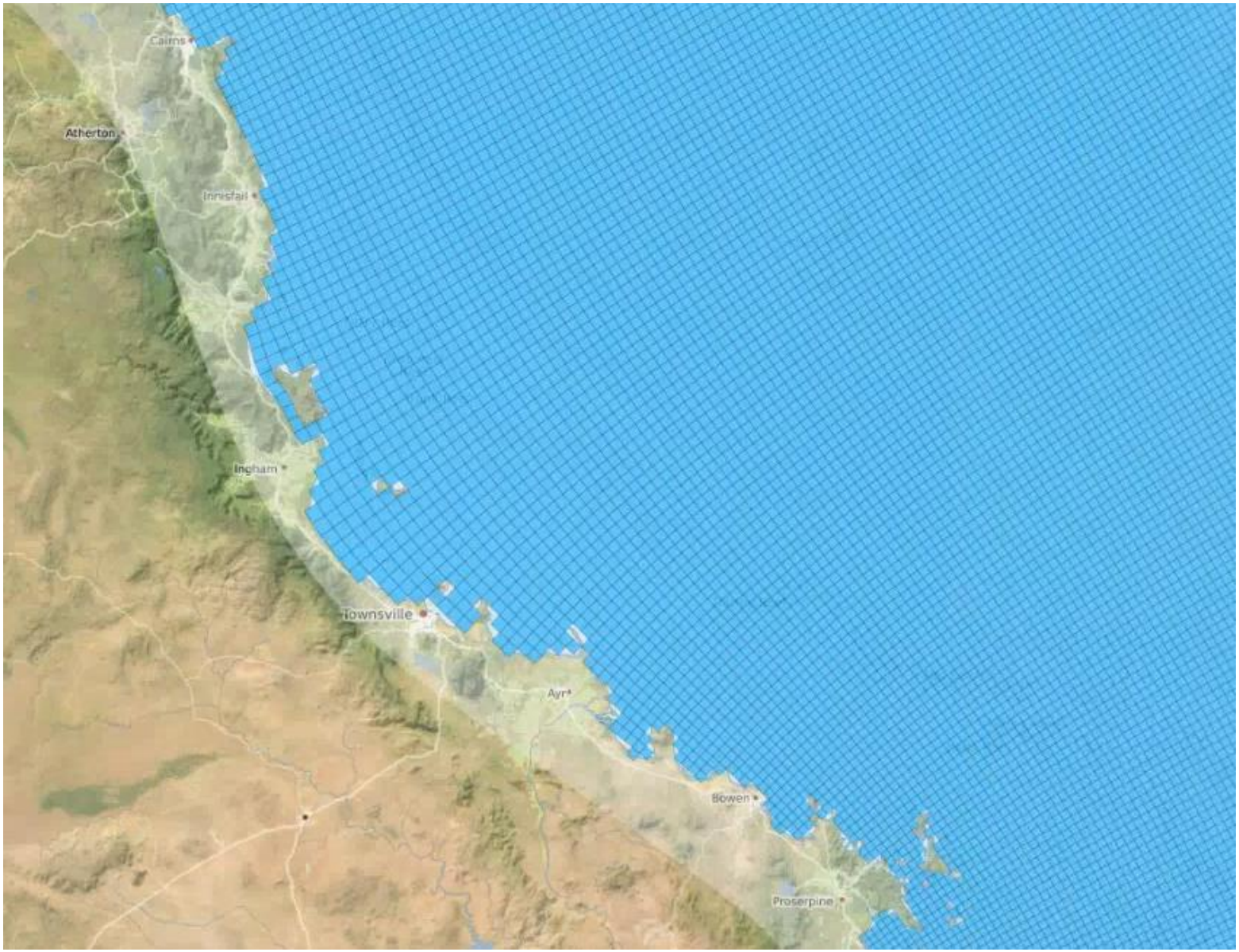
Note: The points I draw from this are that inshore and small areas will be hard to confidently model, but averaging larger offshore areas, and/or averaging over time would work better. But it is very important that wherever we have data, we compare the data and the model.

Question 3: What are the key differences between the modelled data provided from the NCI service compared to the AIMS service?

Answer 3: The data provided by NCI is the "raw" curvilinear data which does not adhere to a regular grid of data. (Eric shared his screen to demonstrate – picture below). Basically, this data is optimised to best capture the gradient moving away from the coastline, curving the cells to fit nicely. This also means that they don't need as many grid cells to achieve the same level of performance as a regularly grid, making it slightly more computationally efficient. However, it also means that the files themselves are very complicated. For example, in a regular grid (used in "normal" NetCDF files) you can pick a value (e.g., the 2nd row, 10th column), and know spatially where that is, because each cell corresponds to an exact latitudinal and longitudinal step. However, for curvilinear data the values are keep in a table or matrix of numbers, but to know where each element of the matrix is spatially, there are two more matrices of numbers, one for lat, and one for long that need to be references to determine cell position.

The data provided by AIMS is a regraded product of the curvilinear output onto a regular grid using an inverse distance weighted interpolation between the closest four neighbouring pixels. The resolution is also slightly increase to account for any weird artifacts. However, this increases the size of the final product. So, to adjust for this AIMS doesn't regraded models 1:1, for example, the hydrodynamic model data is produced hourly, and AIMS will aggregate the data into a daily product, remove the deep depth layers, and then for the BGC model, remove some of the variables.

Note: the curvilinear BGC model is produced daily, and the AIMS regraded version is also daily.


Cairns
Atherton
Innisfail
Ingham
Townsville
Ayr
Bowen
Proserpine

Question 3.1: What are the impacts of the AIMS regrading, how would this effect the work we are trying to do?

Answer 3.1: The biggest differences would be noted along extremely sharp discontinuous boundaries – as when the interpolation is done it can sort of smooth that out. You could think of it as an inverse distance error (the smaller the distance the larger the impact). But it is a bit hard to say what the right answer is anyway because the model is not a perfect representation anyway.

Just in terms of practical usage, if your tools have been developed to process a regular grid then just go with the regular grid. The error introduced by the interpolation is relatively small compared to the inherent errors in the model. Because we have the AIMS daily regular grid that should be more than enough. However, one thing to note is that the daily data is effectively data at noon – i.e., it is essentially a snapshot.

Note: Currently AIMS have only regraded the BGC model to 2019 – this will be updated eventually but as of right now if we want to use BGC outputs we will have to regrade the curvilinear data ourselves.

Note: Marc comments that the actual conversion between curvilinear to regular might not be as daunting as it sounds. There are R libraries that have been written to help with this (by AIMS), but they might not be working at the moment.

Question 4: Can I have a bit of background on how the model is currently being applied and developed further. E.g., this modelling has been going on for a while now, but we (the RRC) haven't really been connected to the process despite the outputs seeming perfect for our usage. What's happening there?

Answer 4: There was the satellite-based stuff that was being done by the Bureau and CSIRO which was run for a few years but there was sort of drift in the satellite, and it needed recalibrating, but BOM couldn't do that, and CSIRO had moved on. So, the Bureau shut down the service. But they never made an equivalent service to what was before. Which I'm not quite sure why they haven't done that as they could if they wanted to. But yes, eReefs certainly has been calculates those variables all the way out to the coral sea. but because it is only done for the report card run it is only updated every two years now.

Note: Adam comments so I guess they key point to make here is that we (the RRC) would like able to piggyback off those model runs when they are done to complement our current water quality measures. Note that no formal arrangements have been made here and this remains an action.

Question 4.1: And what about each of the different models available, I've seen there are pre-industrial, baseline, and target models? How do each of these different again to the Reef Report Card specific model?

Answer 4.1: So, to start off with there is the reef report card model, this one is not real time and is only available every two years with the release of the report. It integrates satellite data etc. and is tailor to give the best outputs for secchi and chla. If you want data in between that you have to rely on the near real time models – and currently AIMS is not processing (producing the regular grid) the near real time BGC model at moment. There are also multiple BGC models focused on specific scenarios such as an improved land practice model (i.e., how does water quality look with better land practices, and what's the relative effect it would have), then one with just the current conditions, and other to match what we think pre-industrial conditions would have been. They run the model three times, with exactly the same weather, exactly the same model, the only difference is the runoff from the land. And then comparisons between these models can be made. They (CSIRO) are also working on a new version of the model. They are updating the hydrodynamic model, and

once there is an improved version of that, the BGC model can then be rerun and will be slightly better. The data is also split into a hindcast dataset and the near real time dataset for more information go here.

Note: As mentioned earlier AIMS is not currently regrading the near real time BGC data – this is currently an issue for us. This is an important snippet of the conversation that should be read as is:

[Start Transcript]

40:04
And then at some point in the future, they'll go, OK, let's make this hindcast longer because we've now got better data. So it's kind of split that way. We're currently waiting for CSIRO to finish and publish the new round of the BGC model from the new hydro model, and then we'll push all those into our (AIMS) systems.

40:28
We haven't done any of that yet because they haven't published them yet.

40:36
so that's kind of in the works.

40:38
And for the current BGC near real time is. Do you know if there is one currently sitting on NCI Mark?

40:56
Yeah. So it is there. But of course, it's got the curve linear grid, Yeah. Yeah. I mean, I guess for you, if you need the water quality, uh, the variables, then probably you have to work with those files. Yeah, because we only have files BCC and until 2019.

[End Transcript]

Question 5: Given that it looks like I will have to work with the curvilinear data, what resources already exist that I could utilise?

Answer 5: Something like this has been done in R before and I feel like we should almost cut out part of Barbara's (AIMS researcher) script that just deals with the curvilinear stuff. I once had a look at the library, and she is doing a pretty simple algorithm to work out the location of each pixel on a regular grid. It does all the mapping once with a very exhaustive search that takes a long time, but then saves the result to use as a lookup.

But we need to make a tutorial about that, we need to work out how to do that ourselves first!

Final Discussion: Agreeing to establish specific problem that we (those in the meeting) can then meet again to address and develop solutions to.

Actions for this: Continue working with the eReefs data to:

a) Compare model outputs with real world data.

b) Establish clear problems/questions to be workshopped (e.g., curvilinear to gridded in R).

c) Maintain contact with Eric, Marc and Brandon with issues faced as user of eReefs.