

The Open Ocean

A PhD in oceanography
and open source software

Callum Rollo



University of East Anglia



National
Oceanography
Centre



COAS
Centre for Ocean and
Atmospheric Sciences



NEXUSS



Before open source



After open source

Contents

- Example of poor data/code sharing (content warning: excel)
- An attempt at doing better
- Why you should care & what you can do

My first data set

- 1000s of files scattered across directories
- No canonical version of processed files
- Processing in excel spreadsheets
- No README

This is not a database →

	A	B	C	D	E	F	G	H	I	J	K	L
1									1.00158934			
2									-3.01383708			
3					.mrk file	O2	Winkler	residual				CTD 9.11
4		<u>Niskin</u>	<u>Botella</u>	<u>Presion</u>	Temperatura	Salinidad	21/09 A	O2 μmol/Kg				Tmp CTD
5	21/09/2010	11 6A		5	16.88	35.63	254.9	252.64	4.86			16.8878
6		8 5A		12.5	16.05	35.67	238.3			234.80	6.19	16.2192
7		7 4A		20	14.3	35.84	205.7	208.79	-0.42			14.2785
8		5 3A		50	13.33	35.85	197.8	200.94	-0.40			13.3336
9		3 2A		80	13.08	35.83	188.2	194.21	-3.35			13.0304
10		1 1A		127	12.98	35.82	170.5	174.11	-0.92			12.9874
11							21/09 B					
12	21/09/2010	11 6B		5	16.9	35.62	254.7	257.37	-0.02			16.9368
13		9 5B		12.5	16.45	35.64	249.5					16.0897
14		7 4B		20	14.68	35.82	212.6	212.01	3.24			14.5952
15		5 3B		50	13.31	35.84	199.3	202.00	-0.01			13.2942
16		4 2B		80	13.06	35.83	188.7	190.09	1.35			13.0512
17		1 1B		124	13.03	35.83	178.2	183.01	-2.13			13.0263
18												
19												
20		<u>Latitud</u>	<u>Longitud</u>	<u>Z (real)</u>	Temperatura	Salinidad	15/09 loc	O2 μmol/Kg				Tmp CTD
21	15/09/2010	42.23	-8.788	5	17.0025	35.5145	253.0474			243.55	12.12	17.0862
22				10	15.6445	35.6085	190.5372			176.40	16.87	15.5691

My first data set

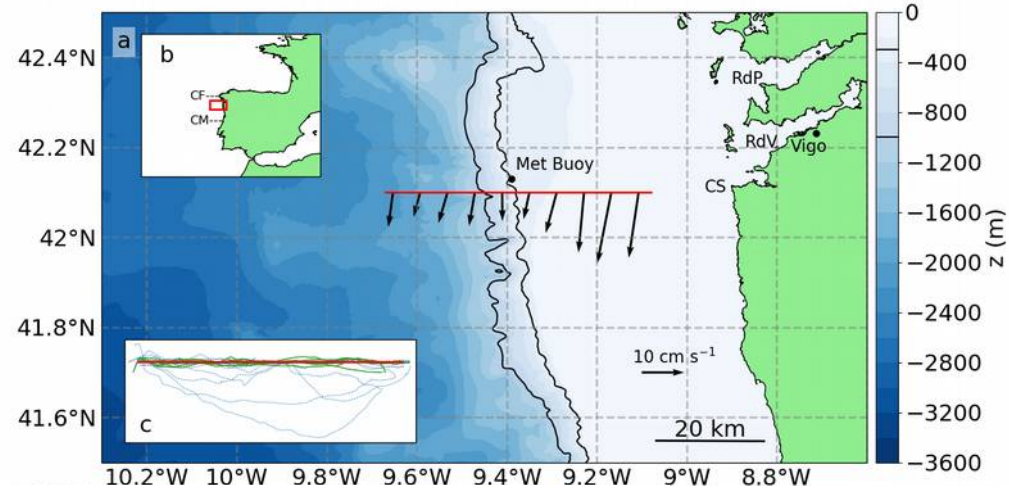
End result >> Reprocess from scratch

This took two years

Substantial duplication of effort

My first data set

A happy ending
at least



JGR Oceans

Research Article | [Open Access](#) |

Glider Observations of the Northwestern Iberian Margin During an Exceptional Summer Upwelling Season

Callum Rollo , Karen J. Heywood, Rob A. Hall, Eric Desmond Barton, Jan Kaiser

First published: 22 June 2020 | <https://doi.org/10.1029/2019JC015804>



Volume 125, Issue 8

August 2020

e2019JC015804



Figures



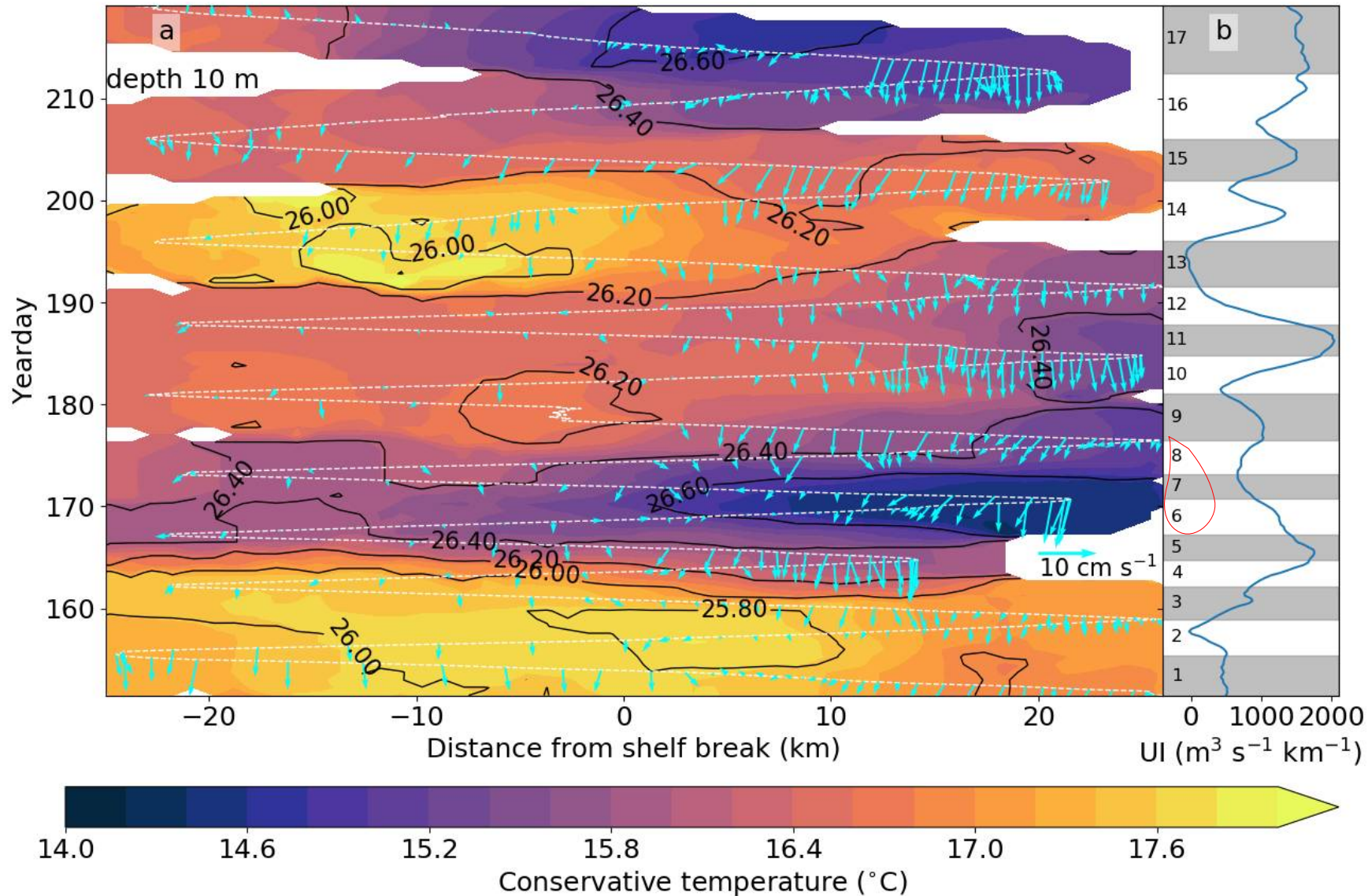
References

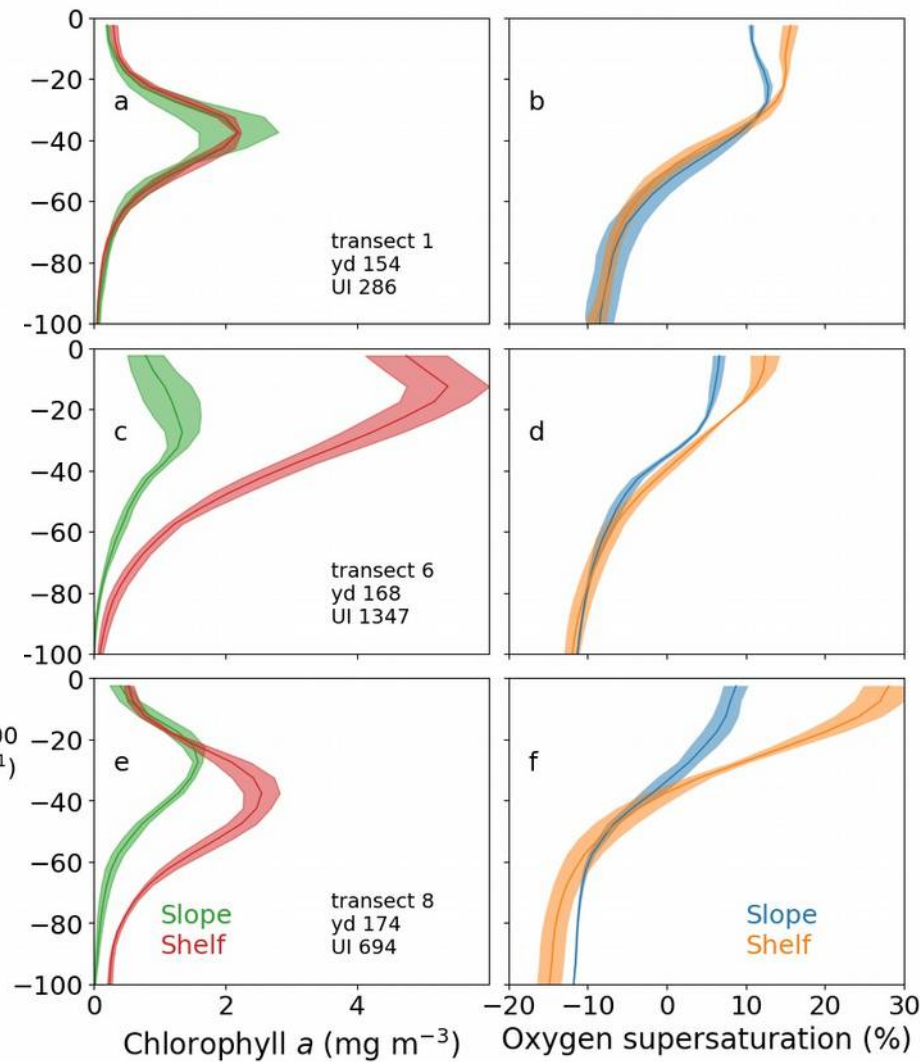
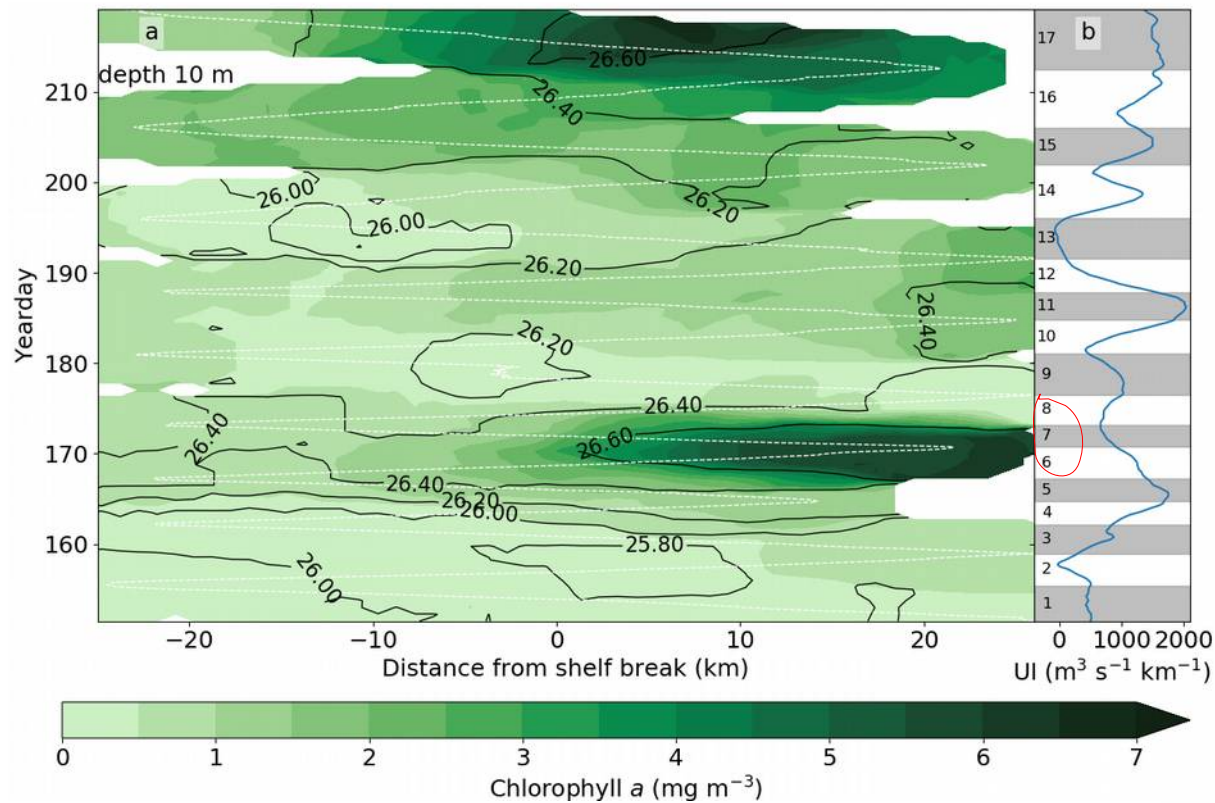


Related



Information





Why is this an problem?

- Effort to collect data >> effort to preserve/share data
- Undocumented datasets easily lost
- Duplication of effort
- “Bus factor”
- Shared code often poorly documented/unusable
- The replication crisis
- When a global pandemic stops field/lab work, how much work can you do?

Principle Issues

Poor data organisation

Multiple conflicting data files

No data archival

No processing notes

Use of proprietary/closed formats

Combined storage and processing

No environment information

Principle Issues

Poor data organisation

Multiple conflicting data files

No data archival

No processing notes

Use of proprietary/closed formats

Combined storage and processing

No environment information

>> Data and scripts
are not **FAIR**



Principle Issues

Poor data organisation

Multiple conflicting data files

No data archival

No processing notes

Use of proprietary/closed formats

Combined storage and processing

No environment information

Solutions I use

Add a README and structure data logically

One canonical dataset, no duplication

Archive at BODC or zenodo

Work in Jupyter Notebooks

Data in csv or NetCDF

Absolute border between data and scripts

Anaconda environment stored with scripts

A better way, “me and my glider”

SG637 *Omura* (rare species of fin whale)

- Custom sensor integration
- No documentation
- No development notes
- No processing code
- No tests

So, let's make them!



How is it organised?

- All tracked with git
- Includes READMEs
- Manual written in plain text
- Converted to pdf, html etc. automatically with makefile
- Hosted on Github, updates pushed live
- Code to generate plots included
- Timestamped copies archived on zenodo
- Open source FOSS license (GPL3)

What does it look like?

adcp-glider-manual

- manual
 - images
 - sensor-diagram.png
 - signal-test.png
 - ...
 - adcp-manual.md
 - adcp-manual.pdf
 - adcp-manual.html
 - Makefile
- **scripts**
 - README.txt
 - script.py
 - helper-script.sh
- README.md
- LICENSE.txt
- .git

What does it look like?

main 1 branch 1 tag Go to file Add file Code

callumrollo	added custom state change schematic	1a9cffe now	14 commits
manual	added custom state change schematic		now
tele_file_parsers	initial commit		2 months ago
.gitignore	added transducer test to manual		2 months ago
LICENSE.txt	initial commit		2 months ago
README.md	Added zenodo doi and license badges		28 days ago

README.md

DOI 10.5281/zenodo.4147102 License GPLv3

Seaglider Nortek 1MHz AD2CP unofficial guide

This an unofficial guide to bench testing, deploying and analysing the data from a Nortek AD2CP integrated onto a Seaglider.

This guide is based solely on the author's experience as a PhD student tasked with working on an ADCP glider without

About

Unofficial guide to the ADCP glider SG637 Omura

Readme

GPL-3.0 License

Releases 1

First public release Latest 28 days ago

Packages

No packages published
Publish your first package

Languages



What does it look like?

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Settings

main







1 branch

1 tag

Go to file

Add file

Code

 callumrollo added custom state change schematic 1a9cffe now 🕒 14 commits
 manual added custom state change schematic now
 tele_file_parsers initial commit 2 months ago
 .gitignore added transducer test to manual 2 months ago
 LICENSE.txt initial commit 2 months ago
 README.md Added zenodo doi and license badges 28 days ago

README.md

DOI [10.5281/zenodo.4147102](https://doi.org/10.5281/zenodo.4147102) License [GPLv3](#)

Seaglider Nortek 1MHz AD2CP unofficial guide

This an unofficial guide to bench testing, deploying and analysing the data from a Nortek AD2CP Integrated onto a Seaglider.

This guide is based solely on the author's experience as a PhD student tasked with working on an ADCP glider without

About

Unofficial guide to the ADCP glider SG637 Omura

Readme

GPL-3.0 License

Releases 1

First public release Latest
28 days ago

Packages

No packages published
[Publish your first package](#)

Languages



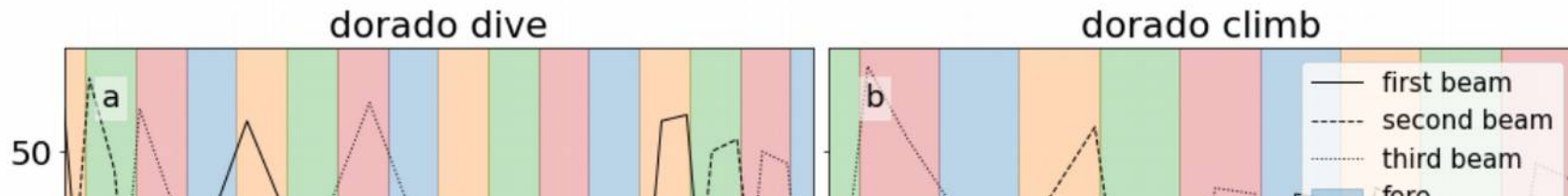
Beam bench test

Following bench test instructions in the [ADCP glider manual](#) transducers can be tested for firing order.

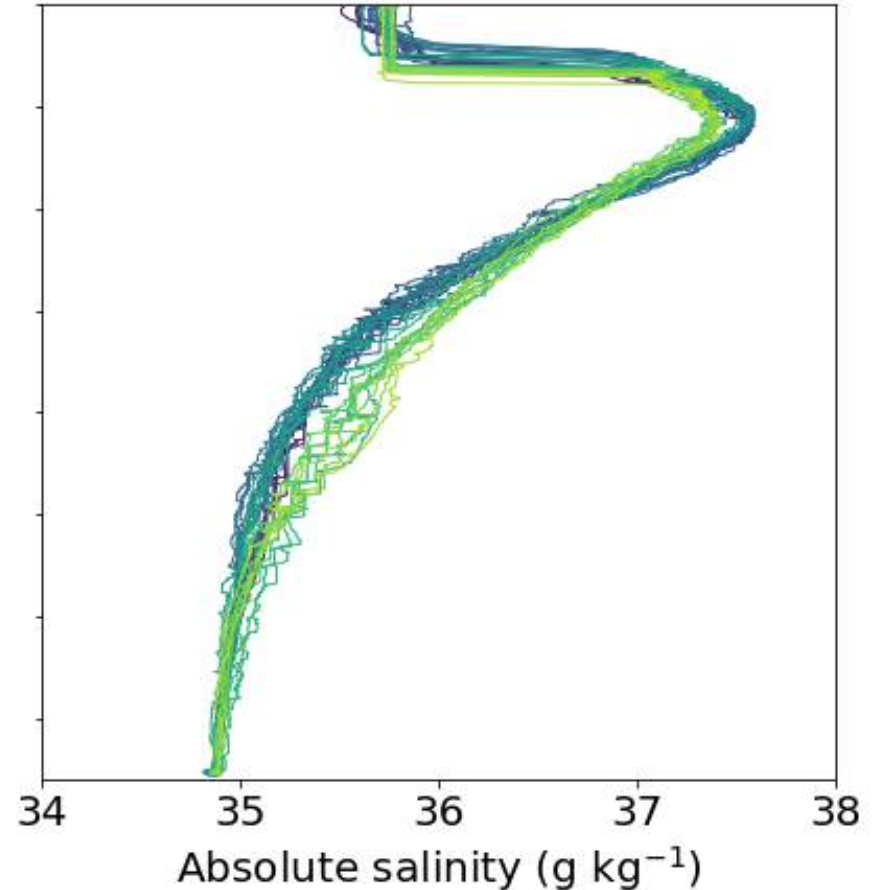
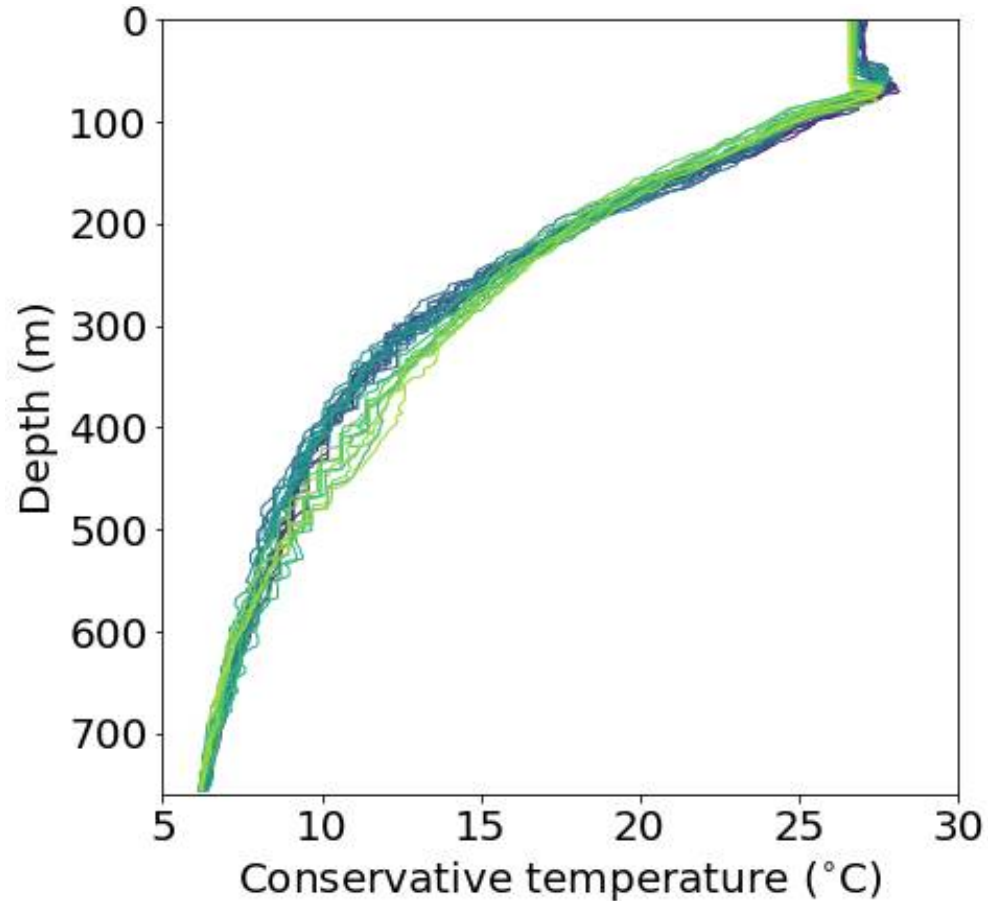
The following plot shows the results when operating under the DORADO and EAGLERAY firmware versions

```
fig, ax = plt.subplots(2,2,figsize=(16,10), sharey='row', sharex=False)
fig.subplots_adjust(hspace=0.1, wspace=0.02)

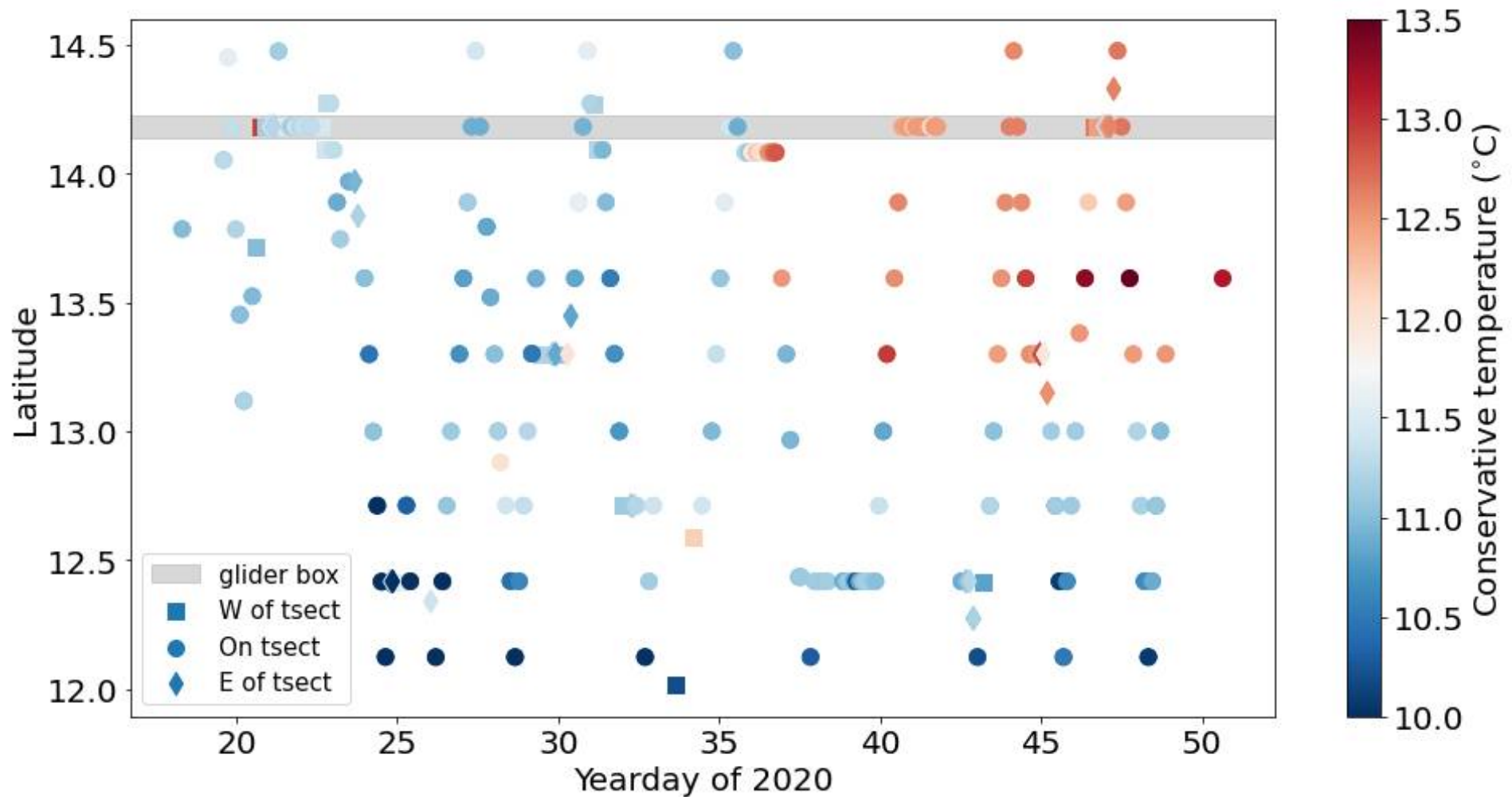
ax = ax.ravel()
datasets = [dorado_dive, dorado_climb, eagle_dive, eagle_climb]
ds_names = ["dorado dive", "dorado climb", "eagle dive", "eagle climb"]
for ax_no, dataset in enumerate(datasets):
    ax[ax_no].plot(dataset.time, dataset.amp_beam[:,1,0], color='k', linestyle='-', label="first beam")
    ax[ax_no].plot(dataset.time, dataset.amp_beam[:,1,1], color='k', linestyle='--', label="second beam")
    ax[ax_no].plot(dataset.time, dataset.amp_beam[:,1,2], color='k', linestyle='dotted', label="third beam")
    ax[ax_no].set(xlim=[dataset.time[0], dataset.time[15]], xticks=[], title=ds_names[ax_no])
    beam_shade(ax[ax_no])
    fig.autofmt_xdate()
ax[2].set(xlabel='Time', ylabel='Return amp (dB)')
ax[1].legend(loc=1);
letterboxes(ax)
fig.savefig('../reports/bench_test_figures/transducer_tests.png')
```



Awkward sidestep into science



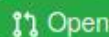
Awkward sidestep into science



What's this got to do with open source?

- Open source library to read CTD data
- Library lacked a feature I needed
- I implemented it
- Pushed to Github
- This would be impossible with proprietary software

add start datetime to as_DataFrame method #61



Open

callumrollo wants to merge 1 commit into `castelao:master` from `callumrollo:df_datetime`

Conversation 1



Commits 1



Checks 0



Files changed 1

**callumrollo** commented on 19 Jul

`as_DataFrame` did not copy any datetime information to the created dataframe. This checks that the attribute `start_time` exists; if so, it creates a column in the dataframe the same fashion as `lon` and `lat`.

Name `datetime_first_scan` chosen as I believe this is the timestamp from the first scan of the cast.

This commit fixes [#60](#)



add start datetime to as_DataFrame method

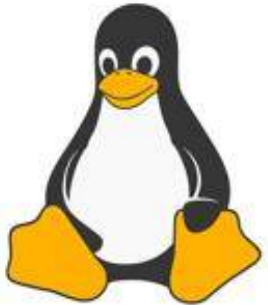
✓ 0ac56aa

**codecov-commenter** commented on 19 Jul • edited

Open source is good (and fun).
You should use it



git



It doesn't have to be perfect

- Every step is worthwhile. Start with a README
- Easy if you do it from the beginning of a project
- There are courses to help e.g. git, Python
- Start with a small/personal project
- This presentation has been shared, typos and all

It doesn't have to be perfect

We all write horrible messy code. It's ok

The CRAPL

“ Academics rarely release code, but I hope a license can encourage them.”

“An open source license for academics... should absolve authors of shame, embarrassment and ridicule for ugly code.”

- Matt Might

Sounds like effort, why should I do this?

- Will remind you in 6 months what your code means
- Reduce duplication of effort
- Lead by example/golden rule
- Encourage collaboration
- Recognition outside of academia
- The wind blows toward open science
- Increasingly a requirement of publication/funding



OK I'm sold, what now?

An hour

- Comment your code and add a README
- Note the environment you ran it in
- Archive to zenodo

A day

- Learn to use git
- Refactor code for readability
- Upload to github*

A new project

- Use open source languages and formats
- Structure your project e.g. cookiecutter
- Make your code into a reusable module and share it

*Other Hosting Services are Available



Better coding
practices

==

Better
science

How did you get into this?

- Leading the [UEA Python group](#)
- Running a server
- Syncing data during glider deployments
- Automating boring stuff
- [Shell scripting](#)
- Making a [personal website](#)
- Breaking things and fixing them
- Attending [Ocean Hack Week \(blog post\)](#)
- Contributing to projects on [GitHub](#)
- Using Linux



Training you can take

- Jennifer Graham and Tiago Silva run an excellent introductory git training course for ENV and Cefas researchers
https://github.com/CefasRepRes/Git_Training_2020/wiki
- Join the UEA [Scientific Python group](#) (contact Callum). We teach half day courses in Python and hold occasional seminars and drop in sessions
- Self teach from the [3 day intro to Python](#) we run with Cefas
- Take a PPD course in R, particularly good for statistical analysis
- Attend [Ocean Hack Week](#)

Links and resources

- [Good enough practices for scientific computing](#) – basically this talk but good
- Getting started with [GitHub](#)
- Top quality data science courses from [software carpentries](#)
- ANDS' [FAIR data webinars](#)
- JPO [letter from editors](#) on supporting FAIR principles November 2020
- Zenodo data archival [policies](#) (any project up to 50 GB, indefinite storage and free DOI)
- [Pangeo](#) project for open source atmospheric science and oceanography
- [Software sustainability institute](#), Southampton
- The journal of open source software ([JOSS](#))
- [Cookiecutter](#) template for data science
- Matt Might's [CRAPL academic licence](#)
- Great [minimalist example](#) of open source from geosciences. Commented code, readme and licence Try it out on [Binder](#) here (launches interactive notebook in your browser).
- [Ten Years Reproducibility Challenge](#)
- [Stackoverflow](#) all of your programming questions answered

The wall of logos on slide 22 are:

git github zenodo
Python netCDF R
Linux Project Jupyter British Oceanographic Data Centre

Footnotes

This talk was presented as a seminar at the Centre for Atmospheric and Ocean Sciences, University of East Anglia 2020-11-27

All of the (poorly articulated) views herein are solely those of Callum Rollo and do not necessarily represent the views of his employer, supervisors, co-workers, acquaintances, or AIs.

All contents licensed under CC-BY-NC 4.0 (attribution, non commercial) except:

- FAIR graphic on slide 11 from [ANDS](#)
- Image of Bartelbey the Scrivener on slide 25 from [Broken Brilliant](#)