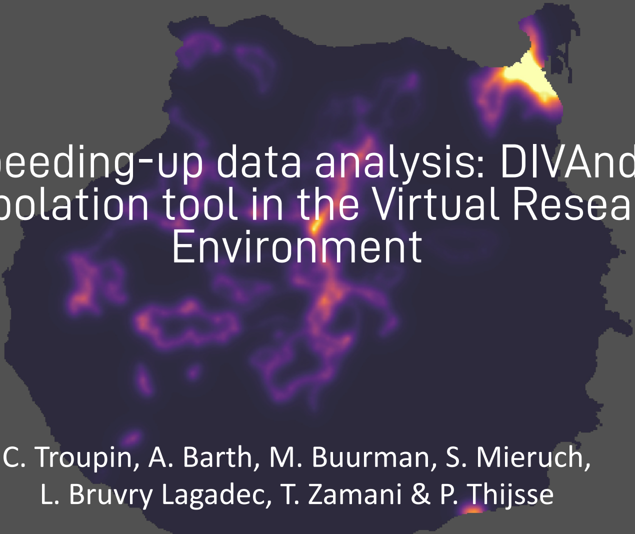


Speeding-up data analysis: DIVAnd interpolation tool in the Virtual Research Environment

C. Troupin, A. Barth, M. Buurman, S. Mieruch,
L. Bruvry Lagadec, T. Zamani & P. Thijsse

A few definitions before we start

DIVA: software tool designed for spatial interpolation
https://github.com/gher-ulg/DIVA

DIVAnd: n-dimensional version of DIVA
https://github.com/gher-ulg/divand.jl

Julia: high-level, high-performance dynamic programming language for numerical computing
https://julialang.org/

Notebook: documents that contain live code, equations, visualizations and narrative text

Jupyter: open-source web application to create and share notebook
http://jupyter.org/

Jupyterhub: multi-user server for Jupyter notebooks
https://jupyterhub.readthedocs.io/en/latest/

Virtual Research Environment (VRE)

Definition: a VRE is made up of software tools, datasets and other resources that several users can access and use in a cooperative way.

Advantages:

- Availability of tools – no need to install anything
- Availability of data – no download needed
- Availability of computing power

1

Context for DIVAnd in the VRE

Installing DIVA was sometimes...painful

Back in 2007

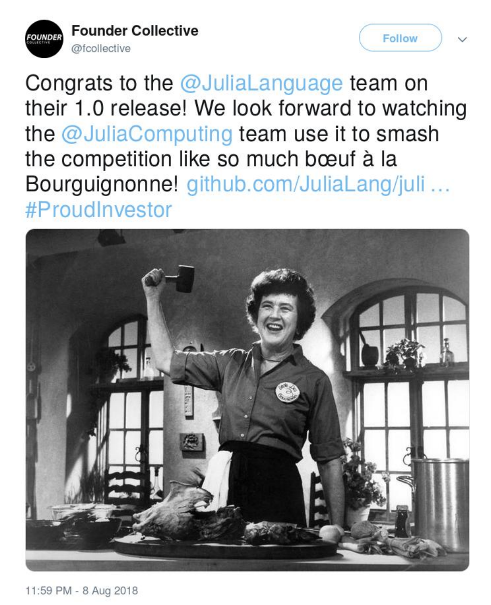
Installing DIVA was sometimes...painful

or in 2013

What have we improved since then?

1. New mathematical formulation	Barth et al. 2014
2. Julia language	instead of Fortran
3. Only 2 (!!!) input files	data & bathymetry
4. Applications as Jupyter notebooks	all in one

What have we improved since then?

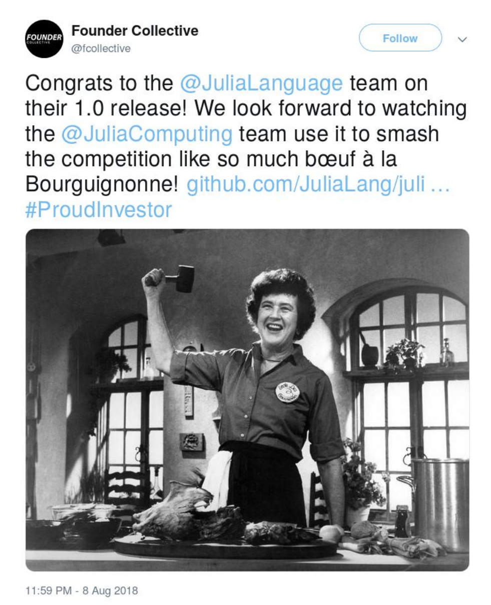


Founder Collective
@fcollective

Congrats to the @JuliaLanguage team on their 1.0 release! We look forward to watching the @JuliaComputing team use it to smash the competition like so much bœuf à la Bourguignonne! github.com/JuliaLang/juli … #ProudInvestor

11:59 PM - 8 Aug 2018

What have we improved since then?

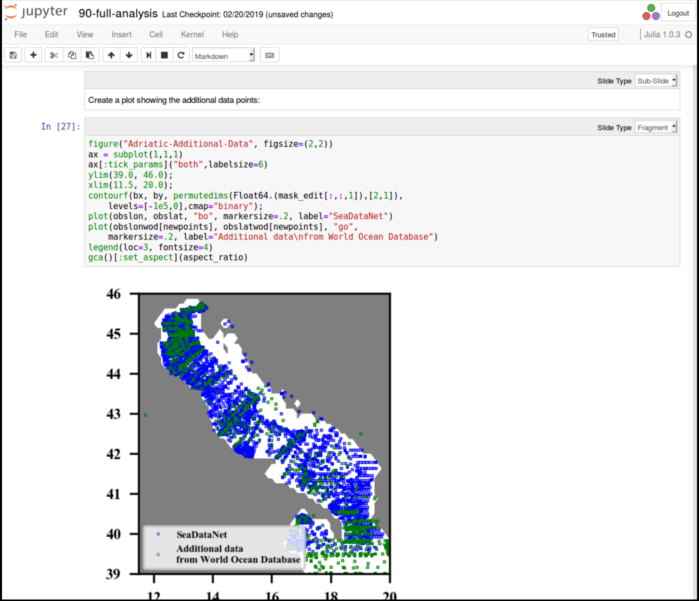


Founder Collective
@fcollective

Congrats to the @JuliaLanguage team on their 1.0 release! We look forward to watching the @JuliaComputing team use it to smash the competition like so much bœuf à la Bourguignonne! github.com/JuliaLang/juli … #ProudInvestor

11:59 PM - 8 Aug 2018

Jupyter notebooks as guidelines



90-full-analysis Last Checkpoint: 02/20/2019 (unsaved changes)

Slide Type: Sub-Slide

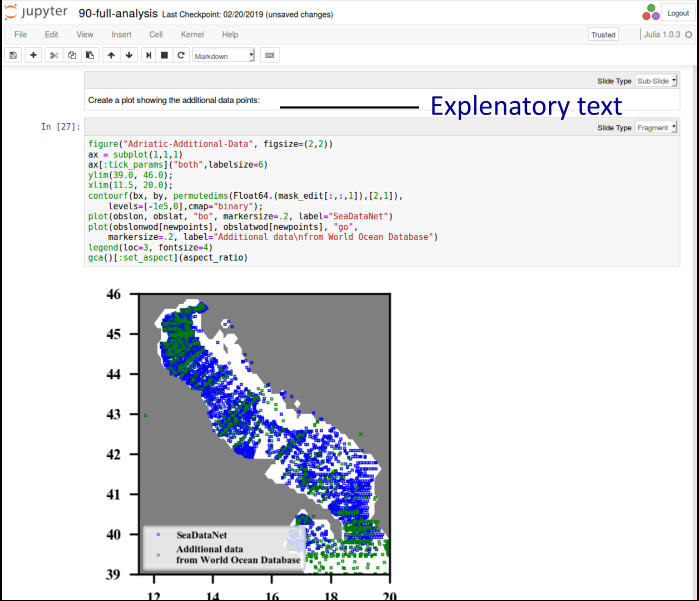
Create a plot showing the additional data points:

In [27]:

Slide Type: Fragment

```
figure("Adriatic-Additional-Data", figsize=(2,2))
ax = subplot(1,1,1)
ax[:tick_params]("both",labelsize=6)
ylim(39.0, 46.0);
xlim(11.5, 20.0);
contourf(bx, by, permutedims(Float64.(mask_edit[:,:,1]),[2,1]),
    levels=[-1e5,0],cmap="binary");
plot(obslon, obslat, "bo", markersize=.2, label="SeaDataNet")
plot(obslonwod[newpoints], obslatwod[newpoints], "go",
    markersize=.2, label="Additional data\nfrom World Ocean Database")
legend(loc=3, fontsize=4)
gca()[:set_aspect](aspect_ratio)
```

Jupyter notebooks as guidelines



90-full-analysis Last Checkpoint: 02/20/2019 (unsaved changes)

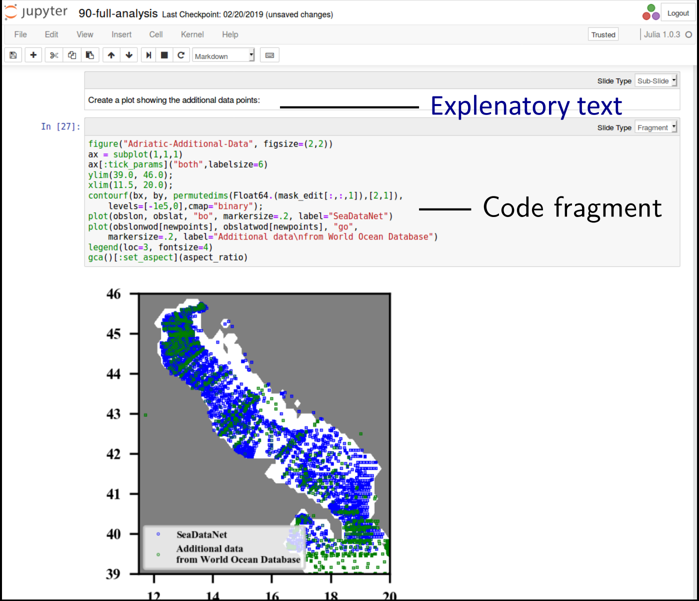
Create a plot showing the additional data points:

Explanatory text

In [27]:

```julia
figure("Adriatic-Additional-Data", figsize=(2,2))
ax = subplot(1,1,1)
ax[:tick_params]("both",labelsize=6)
ylim(39.0, 46.0);
xlim(11.5, 20.0);
contourf(bx, by, permutedims(Float64.(mask_edit[:,:,1]),[2,1]),
    levels=[-1e5,0],cmap="binary");
plot(obslon, obslat, "bo", markersize=.2, label="SeaDataNet")
plot(obslonwod[newpoints], obslatwod[newpoints], "go",
    markersize=.2, label="Additional data\nfrom World Ocean Database")
legend(loc=3, fontsize=4)
gca()[:set_aspect](aspect_ratio)
```

Jupyter notebooks as guidelines



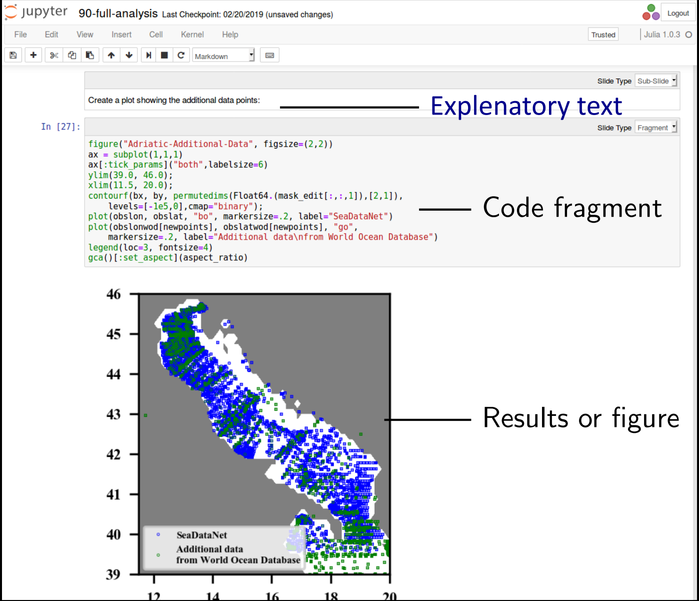
Create a plot showing the additional data points:

Explanatory text

```
figure("Adriatic-Additional-Data", figsize=(2,2))
ax = subplot(1,1,1)
ax[:tick_params]("both",labelsize=6)
ylim(39.0, 46.0);
xlim(11.5, 20.0);
contourf(bx, by, permutedims(Float64.(mask_edit[:,:,1]),[2,1]),
    levels=[-1e5,0],cmap="binary");
plot(obslon, obslat, "bo", markersize=.2, label="SeaDataNet")
plot(obslonwod[newpoints], obslatwod[newpoints], "go",
    markersize=.2, label="Additional data\nfrom World Ocean Database")
legend(loc=3, fontsize=4)
gca()[:set_aspect](aspect_ratio)
```

Code fragment

Jupyter notebooks as guidelines



Create a plot showing the additional data points:

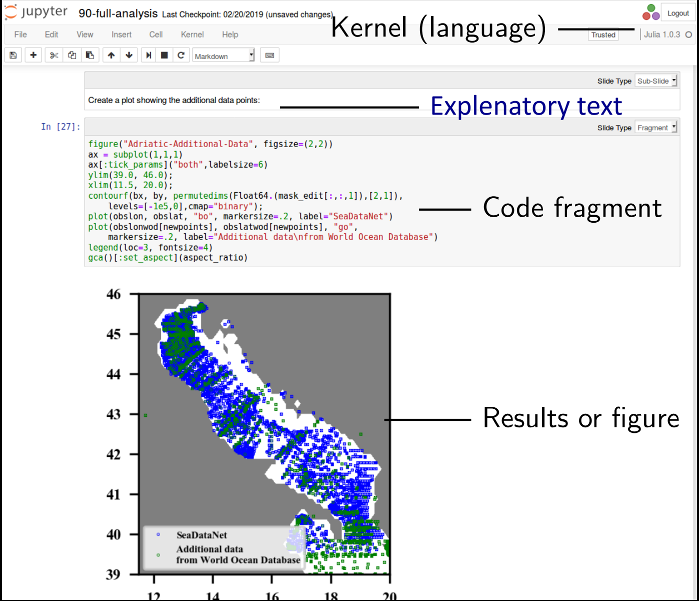
```
figure("Adriatic-Additional-Data", figsize=(2,2))
ax = subplot(1,1,1)
ax[:tick_params]("both",labelsize=6)
ylim(39.0, 46.0);
xlim(11.5, 20.0);
contourf(bx, by, permutedims(Float64.(mask_edit[:,:,1]),[2,1]),
    levels=[-1e5,0],cmap="binary");
plot(obslon, obslat, "bo", markersize=.2, label="SeaDataNet")
plot(obslonwod[newpoints], obslatwod[newpoints], "go",
    markersize=.2, label="Additional data\nfrom World Ocean Database")
legend(loc=3, fontsize=4)
gca()[:set_aspect](aspect_ratio)
```

Explanatory text

Code fragment

Results or figure

Jupyter notebooks as guidelines



90-full-analysis Last Checkpoint: 02/20/2019 (unsaved changes)

Kernel (language)

Create a plot showing the additional data points:

Explanatory text

```
figure("Adriatic-Additional-Data", figsize=(2,2))
ax = subplot(1,1,1)
ax[:tick_params]("both",labelsize=6)
ylim(39.0, 46.0);
xlim(11.5, 20.0);
contourf(bx, by, permutedims(Float64.(mask_edit[:,:,1]),[2,1]),
    levels=[-1e5,0],cmap="binary");
plot(obslon, obslat, "bo", markersize=.2, label="SeaDataNet")
plot(obslonwod[newpoints], obslatwod[newpoints], "go",
    markersize=.2, label="Additional data\nfrom World Ocean Database")
legend(loc=3, fontsize=4)
gca()[:set_aspect](aspect_ratio)
```

Code fragment

Results or figure

What should we improve?

What should we improve?

+ Access to computing power

+ Data availability

+ Documentation

What should we improve?

+ Access to computing power
+ Data availability
+ Documentation

= Virtual Research Environment!

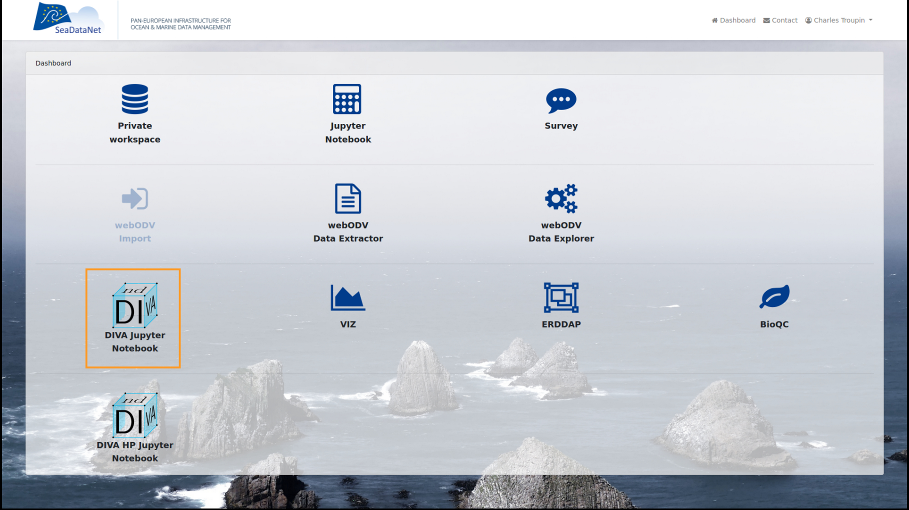
2

DIVAnd in a few clicks

DIVAnd in a few clicks

DIVAnd in a few clicks

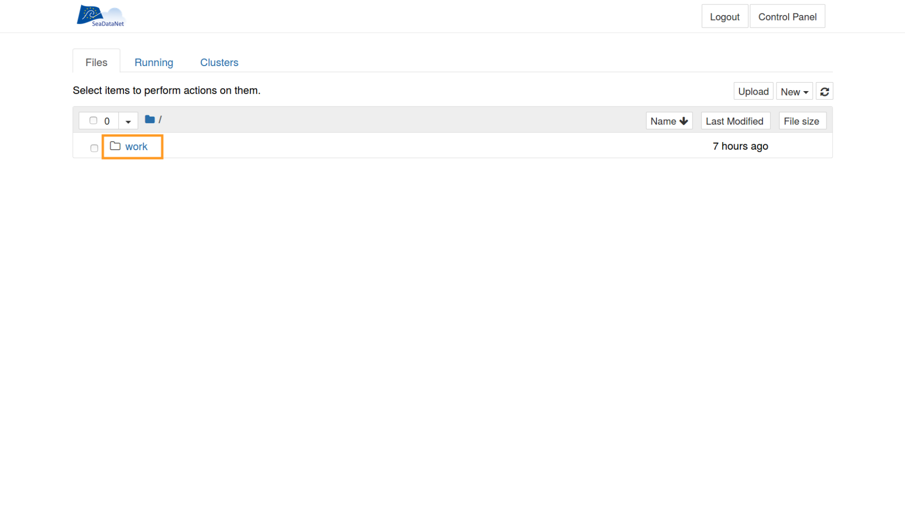
Dashboard



PAN-EUROPEAN INFRASTRUCTURE FOR
OCEAN & MARINE DATA MANAGEMENT

DIVAnd in a few clicks

Work directory



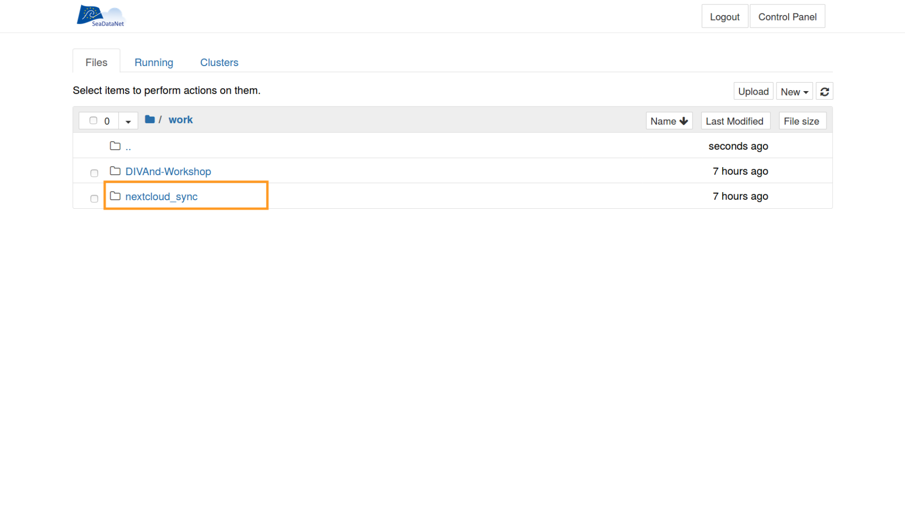
Select items to perform actions on them.

□ 0 /

□ work 7 hours ago

DIVAnd in a few clicks

nextcloud_sync allows users to access their private workspace



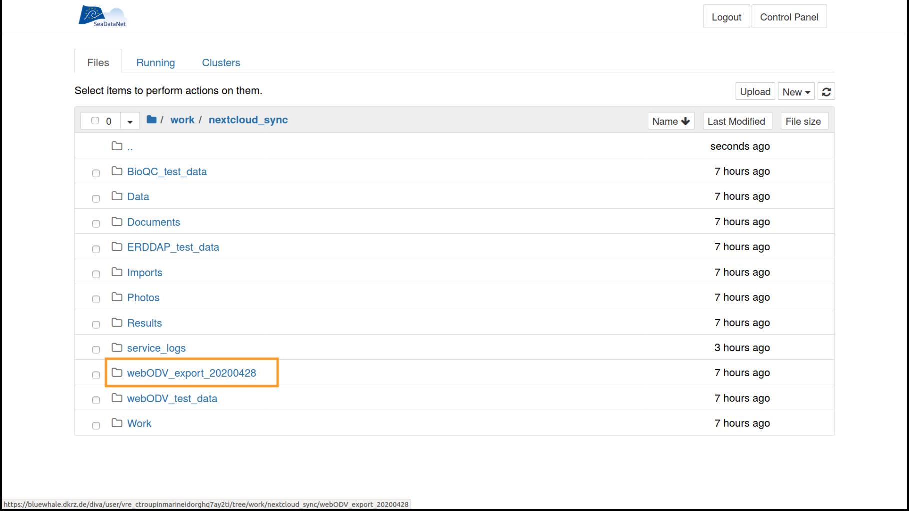
Select items to perform actions on them.

□ 0 / work

Name	Last Modified
..	seconds ago
□ DIVAnd-Workshop	7 hours ago
□ nextcloud_sync	7 hours ago

DIVAnd in a few clicks

directory where webODV gave me the data files



Select items to perform actions on them.

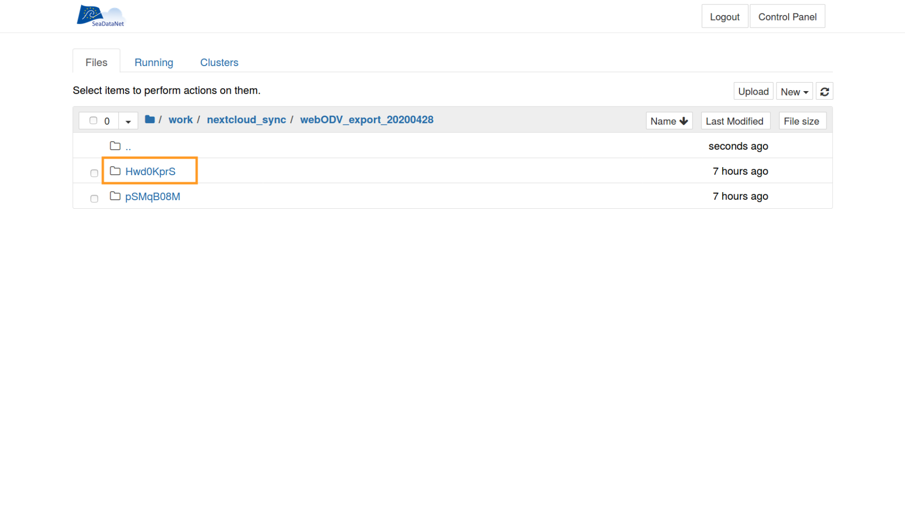
/ work / nextcloud_sync

	Name	Last Modified	File size
	..	seconds ago	
□	BioQC_test_data	7 hours ago	
□	Data	7 hours ago	
□	Documents	7 hours ago	
□	ERDDAP_test_data	7 hours ago	
□	Imports	7 hours ago	
□	Photos	7 hours ago	
□	Results	7 hours ago	
□	service_logs	3 hours ago	
□	webODV_export_20200428	7 hours ago	
□	webODV_test_data	7 hours ago	
□	Work	7 hours ago	

https://bluewhale.dkrz.de/diva/user/vre_ctroupinmarineidorghq7ay2ti/tree/work/nextcloud_sync/webODV_export_20200428

DIVAnd in a few clicks

sub-directory where webODV gave me the data files



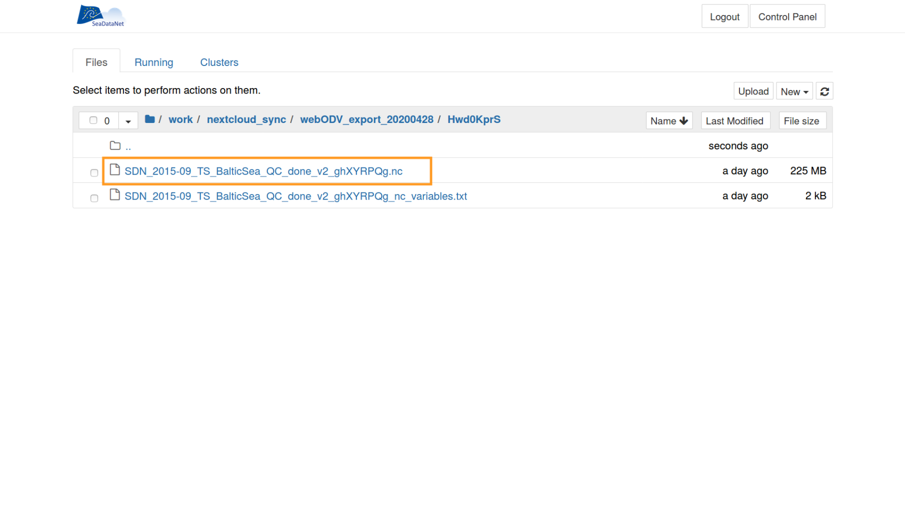
Select items to perform actions on them.

□ 0 / work / nextcloud_sync / webODV_export_20200428

Name	Last Modified
..	seconds ago
□ Hwd0KprS	7 hours ago
□ pSMqB08M	7 hours ago

DIVAnd in a few clicks

data file to use in the notebooks



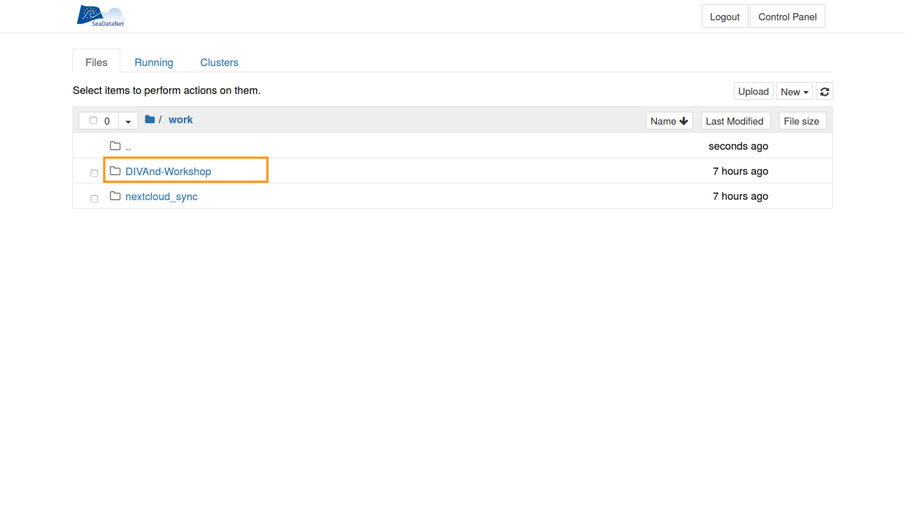
Select items to perform actions on them.

/ work / nextcloud_sync / webODV_export_20200428 / Hwd0KprS

```
..                                                          seconds ago
□ SDN_2015-09_TS_BalticSea_QC_done_v2_ghXYRPQg.nc              a day ago   225 MB
□ SDN_2015-09_TS_BalticSea_QC_done_v2_ghXYRPQg_nc_variables.txt a day ago     2 kB
```

DIVAnd in a few clicks

DIVAnd-Workshop = repository cloned into the VRE



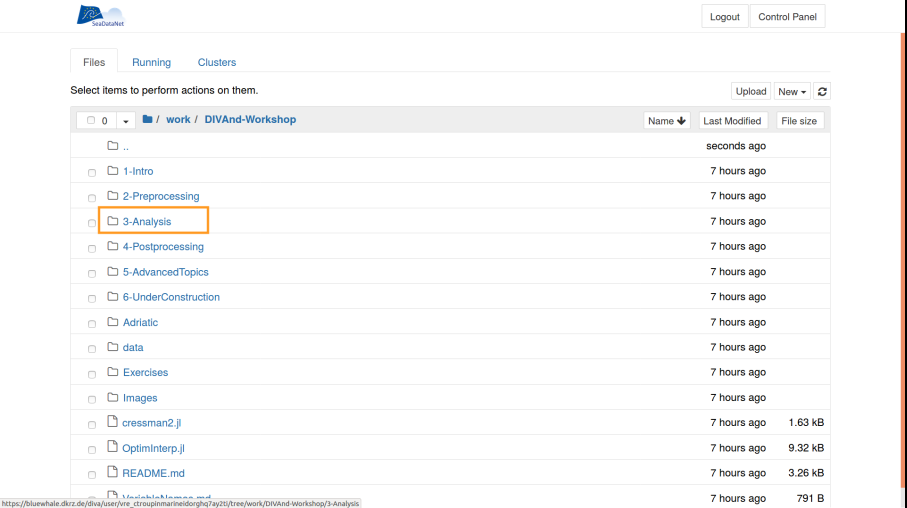
Select items to perform actions on them.

□ 0 / work

Name	Last Modified
..	seconds ago
□ DIVAnd-Workshop	7 hours ago
□ nextcloud_sync	7 hours ago

DIVAnd in a few clicks

Notebooks divided into subdirectories



Select items to perform actions on them.

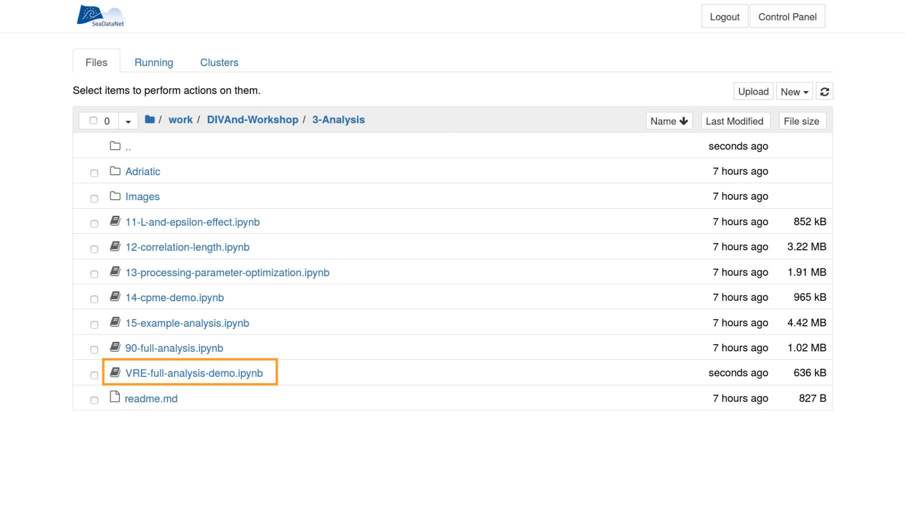
/ work / DIVAnd-Workshop

Name	Last Modified	File size
..	seconds ago	
1-Intro	7 hours ago	
2-Preprocessing	7 hours ago	
3-Analysis	7 hours ago	
4-Postprocessing	7 hours ago	
5-AdvancedTopics	7 hours ago	
6-UnderConstruction	7 hours ago	
Adriatic	7 hours ago	
data	7 hours ago	
Exercises	7 hours ago	
Images	7 hours ago	
cressman2.jl	7 hours ago	1.63 kB
OptimInterp.jl	7 hours ago	9.32 kB
README.md	7 hours ago	3.26 kB
VariableNames.md	7 hours ago	791 B

https://bluewhale.dkrz.de/diva/user/vre_ctroupinmarineidorghq7ay2ti/tree/work/DIVAnd-Workshop/3-Analysis

DIVAnd in a few clicks

Full-analysis using the extracted data file

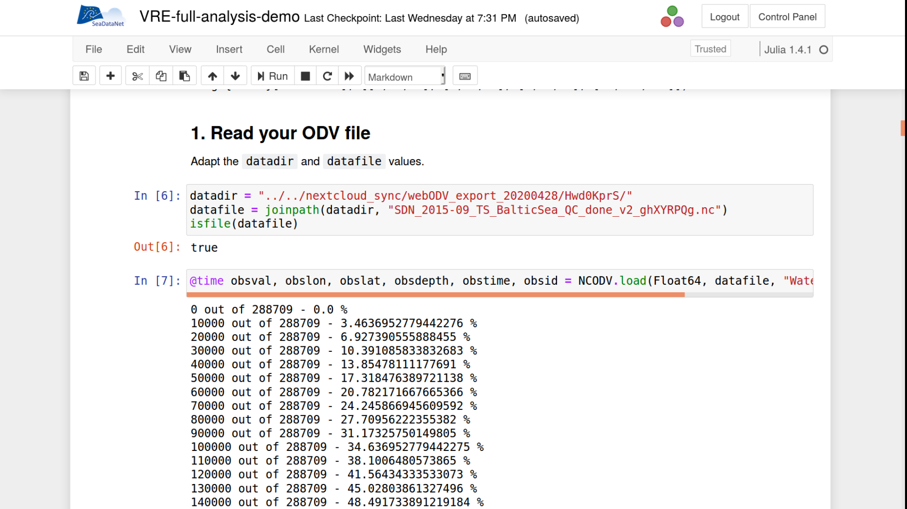


Select items to perform actions on them.

/ work / DIVAnd-Workshop / 3-Analysis

Name	Last Modified	File size
..	seconds ago	
Adriatic	7 hours ago	
Images	7 hours ago	
11-L-and-epsilon-effect.ipynb	7 hours ago	852 kB
12-correlation-length.ipynb	7 hours ago	3.22 MB
13-processing-parameter-optimization.ipynb	7 hours ago	1.91 MB
14-cpme-demo.ipynb	7 hours ago	965 kB
15-example-analysis.ipynb	7 hours ago	4.42 MB
90-full-analysis.ipynb	7 hours ago	1.02 MB
VRE-full-analysis-demo.ipynb	seconds ago	636 kB
readme.md	7 hours ago	827 B

Jupyter notebooks to produce climatologies



1. Read your ODV file

Adapt the `datadir` and `datafile` values.

In [6]:

```
datadir = "../../nextcloud_sync/webODV_export_20200428/Hwd0KprS/"
datafile = joinpath(datadir, "SDN_2015-09_TS_BalticSea_QC_done_v2_ghXYRPQg.nc")
isfile(datafile)
```

Out[6]:

```
true
```

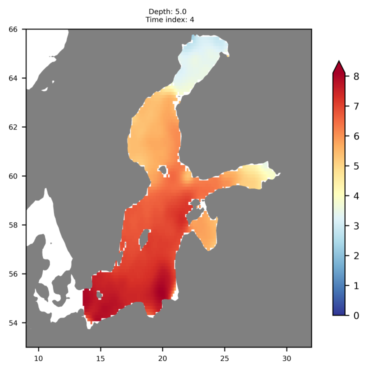
In [7]:

```
@time obsval, obslon, obslat, obsdepth, obstime, obsid = NCODV.load(Float64, datafile, "Wate
```

```
0 out of 288709 - 0.0 %
10000 out of 288709 - 3.4636952779442276 %
20000 out of 288709 - 6.927390555888455 %
30000 out of 288709 - 10.391085833832683 %
40000 out of 288709 - 13.85478111177691 %
50000 out of 288709 - 17.318476389721138 %
60000 out of 288709 - 20.782171667665366 %
70000 out of 288709 - 24.245866945609592 %
80000 out of 288709 - 27.70956222355382 %
90000 out of 288709 - 31.17325750149805 %
100000 out of 288709 - 34.636952779442275 %
110000 out of 288709 - 38.1006480573865 %
120000 out of 288709 - 41.56434333533073 %
130000 out of 288709 - 45.02803861327496 %
140000 out of 288709 - 48.491733891219184 %
```
]

DIVAnd outputs

1. A netCDF file storing the analysis fields and the observation ID's
2. A XML file that can be used for the Sextant catalog
3. Figures (here Autumn salinity at 5 m depth)



3

Improving DIVAnd in the VRE

DIVAnd in 2020

1. Updated mathematical formulation — Barth et al. 2014
2. Julia language
3. Only 2 input files — data & bathymetry
4. User guides = Jupyter notebooks — all in one

Providing users access to notebooks

jupyterhub ≡ jupyter × 👥

2017	1st deployment at **CINECA**
2019	Deployment transferred to **DKRZ** for the first training workshop
2020	Multiple machines at **DKRZ** **GRNET** **STFC**

Multiple copies of a Docker container are run (https://hub.docker.com/r/abarth/divand-jupyterhub)

What’s shipped inside the container?

docker
julia
netCDF
nd
DIVA

🔗 https://github.com/gher-ulg/DIVAnd-jupyterhub

☑ Libraries: netCDF, unzip, git, ...

What’s shipped inside the container?

https://github.com/gher-ulg/DIVAnd-jupyterhub

- ☑ Libraries: netCDF, unzip, git, ...
- ☑ Julia language (V1.4.1, released on April 14, 2020))

What’s shipped inside the container?

🔗 https://github.com/gher-ulg/DIVAnd-jupyterhub

- ☑ Libraries: netCDF, unzip, git, ...
- ☑ Julia language (V1.4.1, released on April 14, 2020))
- ☑ Julia packages: PyPlot, NCDatasets, DataStructures, ...

What's shipped inside the container?

🔗 https://github.com/gher-ulg/DIVAnd-jupyterhub

- ☑ Libraries: netCDF, unzip, git, ...
- ☑ Julia language (V1.4.1, released on April 14, 2020))
- ☑ Julia packages: PyPlot, NCDatasets, DataStructures, ...
- ☑ DIVAnd.jl (#master)

What's shipped inside the container?

https://github.com/gher-ulg/DIVAnd-jupyterhub

- ☑ Libraries: netCDF, unzip, git, ...
- ☑ Julia language (V1.4.1, released on April 14, 2020))
- ☑ Julia packages: PyPlot, NCDatasets, DataStructures, ...
- ☑ DIVAnd.jl (#master)
- ☑ DIVAnd *workshop* notebooks (#master)

4 components for improvement

1. Julia language

4 components for improvement

1. Julia language
2. DIVAnd package

4 components for improvement

1. Julia language
2. DIVAnd package
3. Training notebooks (Diva-Workshop)

4 components for improvement

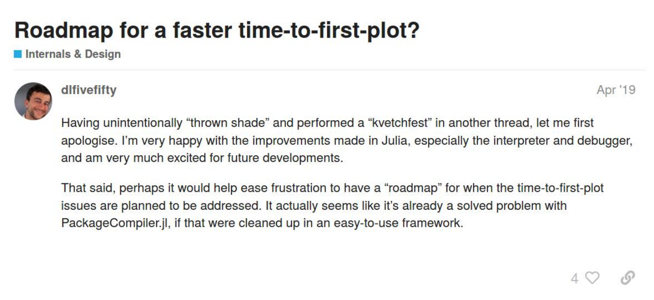
1. Julia language
2. DIVAnd package
3. Training notebooks (Diva-Workshop)
4. Docker container in the VRE

Julia language

v1.4.1 (April 14, 2020)

New features: changes in language features, multi-threading, build systems, library functions, ...

PackageCompiler.jl : allows to pre-compile the modules
Reduce the famous "*time-to-first-plot*":
when you first load and run a package in a session,
Julia needs to compile it first



Roadmap for a faster time-to-first-plot?

Internals & Design

dlfivefifty — Apr '19

Having unintentionally "thrown shade" and performed a "kvetchfest" in another thread, let me first apologise. I'm very happy with the improvements made in Julia, especially the interpreter and debugger, and am very much excited for future developments.

That said, perhaps it would help ease frustration to have a "roadmap" for when the time-to-first-plot issues are planned to be addressed. It actually seems like it's already a solved problem with PackageCompiler.jl, if that were cleaned up in an easy-to-use framework.

DIVAnd.jl package

v2.6.0 (April 27, 2020)

DOI 10.5281/zenodo.3768663

New feature: heatmap generation

Closed issues:

- DIVAnd_qc doesn't work with advection constrain
- Missing surfextend documentation
- Wrong long name for lat
- netCDF scalefactore is NaN resulting in an empty output variable

Example of heatmaps

Data points are locations acquired by GPS device during running activities

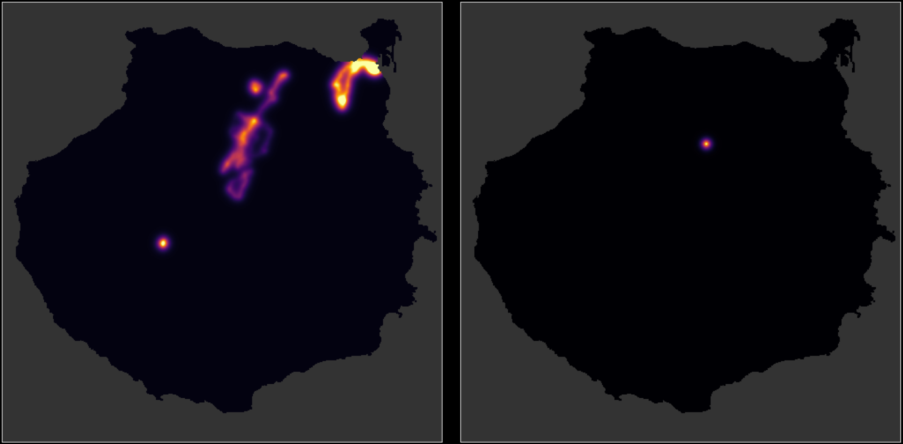


Figure 1: Left: heatmap before the confinement; Right: after the confinement

Training notebooks (Diva-Workshop)

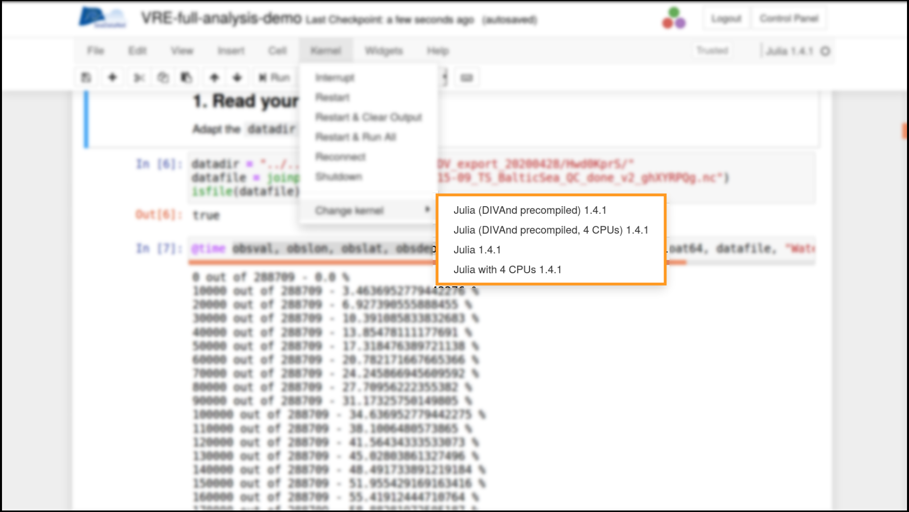
v1.2 (April 27, 2020)

DOI 10.5281/zenodo.3769037

- Compatibility with Julia 1.4
- CI (Travis)
- Cleaned notebooks
- Sub-directories with notebooks arranged by topics.
- Minor bugs

Docker container (DIVAnd-jupyterhub)

Precompiled DIVAnd with PackageCompiler



1. Read your

Adapt the `datadir`

```
In [6]: datadir = "../..                    DV_export_20200428/HwdBKpr5/"
        datafile = joinp                    15-09_TS_BalticSea_QC_done_v2_ghXYRPQg.nc")
        isfile(datafile)

Out[6]: true

In [7]: @time obsval, obslon, obslat, obsdep                    .oat64, datafile, "Wat

0 out of 288709 - 0.0 %
10000 out of 288709 - 3.4636952779442276 %
20000 out of 288709 - 6.927390555888455 %
30000 out of 288709 - 10.391085833832683 %
40000 out of 288709 - 13.85478111177691 %
50000 out of 288709 - 17.318476389721138 %
60000 out of 288709 - 20.782171667665366 %
70000 out of 288709 - 24.245866945609592 %
80000 out of 288709 - 27.70956222355382 %
90000 out of 288709 - 31.17325750149805 %
100000 out of 288709 - 34.636952779442275 %
110000 out of 288709 - 38.1006480573865 %
120000 out of 288709 - 41.56434333533073 %
130000 out of 288709 - 45.02803861327496 %
140000 out of 288709 - 48.49173389121918 %
150000 out of 288709 - 51.955429169163416 %
160000 out of 288709 - 55.41912444710764 %
```

Conclusions

1. DIVAnd is deployed in the SeaDataCloud VRE
2. Continuous improvements in:
   Julia, DIVAnd, Training notebooks and in the Docker
3. Making available data, tools and computing power is essential to speed up research

Thanks
for your attention

🐦 CharlesTroupin

#shareEGU20