



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

Context for DIVA and 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
- 2 Julia language
- 3 Only 2 (!!!) input files
- 4 Applications as Jupyter notebooks

 Barth et al. 2014
instead of Fortran
data & bathymetry
all in one

What have we improved since then?



Founder Collective

@fcollective

Follow



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 ...](https://github.com/JuliaLang/julia)
[#ProudInvestor](#)



11:59 PM - 8 Aug 2018

What have we improved since then?



Founder Collective

@fcollective

Follow



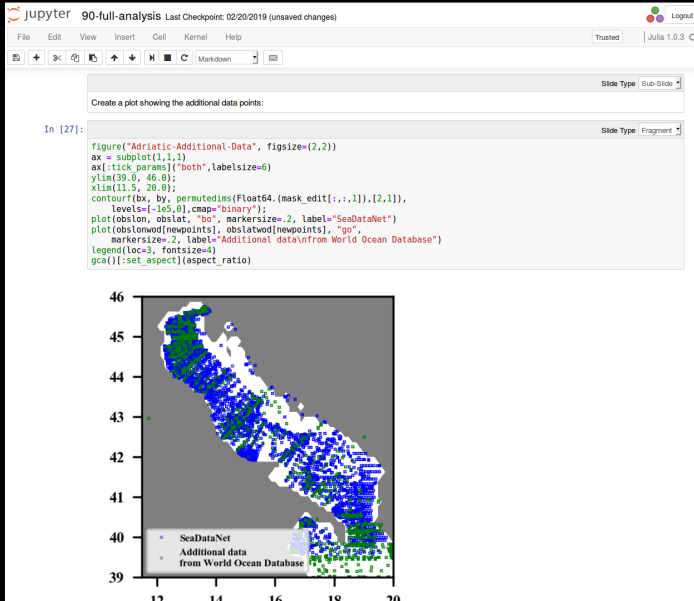
Congrats to the [@JuliaLanguage](#) team on their 1.0 release! We look forward to watching the [@JuliaComputing](#) team use it to smash the competition like so much bœuf à la Bourguignonne! [github.com/JuliaLang/juli ...](https://github.com/JuliaLang/julia)
[#ProudInvestor](#)



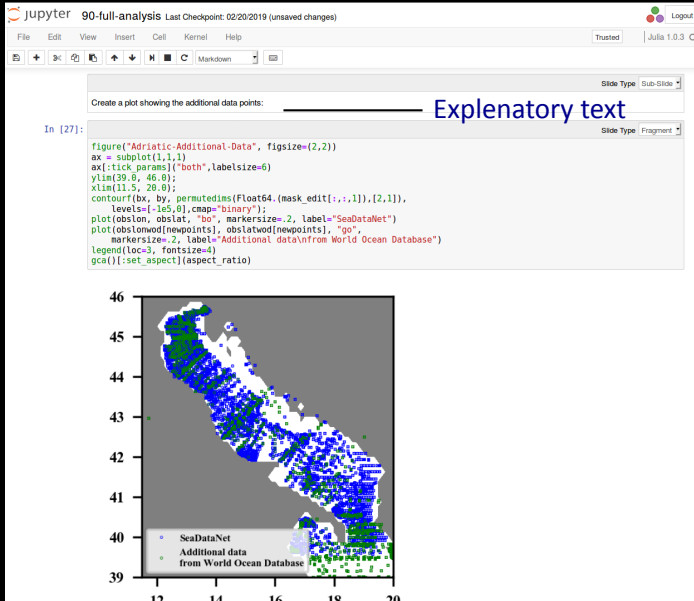
11:59 PM - 8 Aug 2018



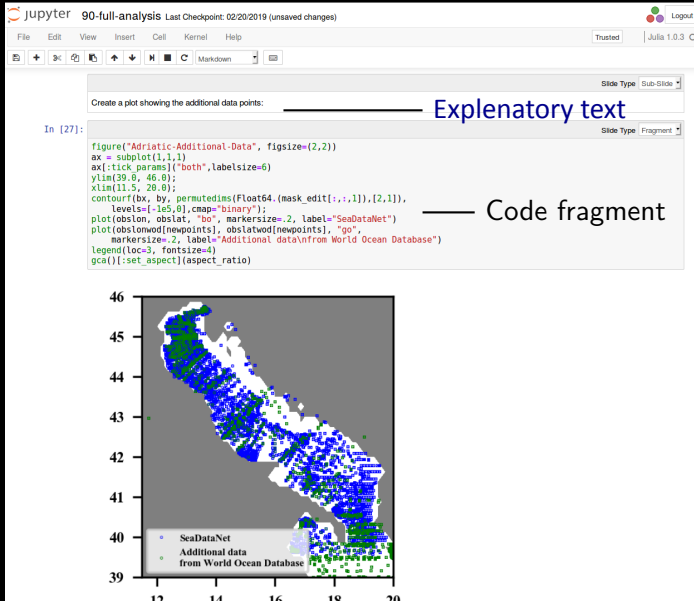
Jupyter notebooks as guidelines



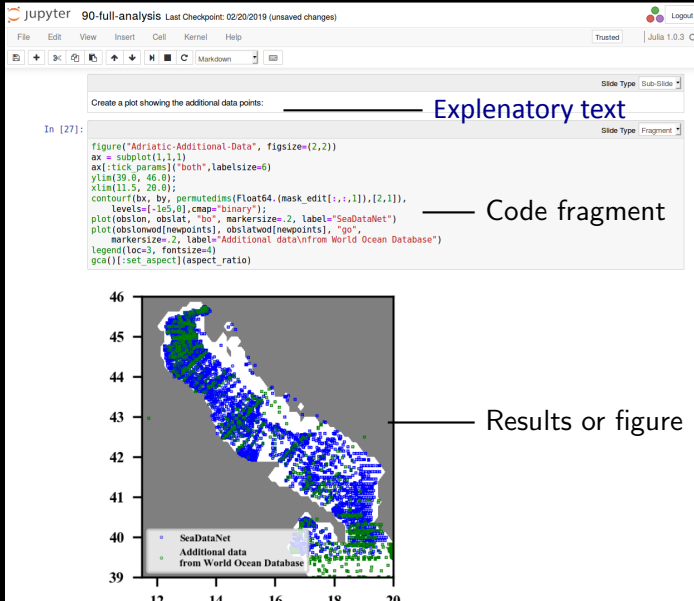
Jupyter notebooks as guidelines



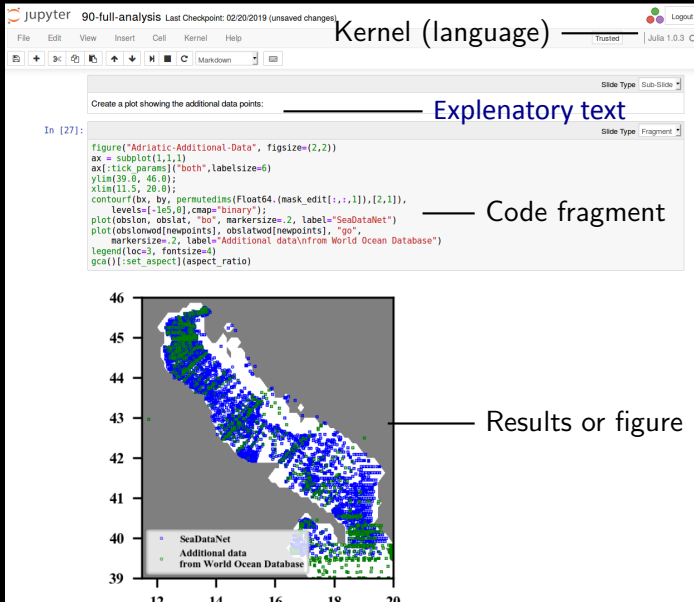
Jupyter notebooks as guidelines



Jupyter notebooks as guidelines



Jupyter notebooks as guidelines



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!




DIVAnd
in a few
clicks

 DIV And in a few clicks


DIVAnd in a few clicks


Dashboard


 PAN-EUROPEAN INFRASTRUCTURE FOR OCEAN & MARINE DATA MANAGEMENT


[Dashboard](#) [Contact](#) [Charles Troupin](#)


Dashboard



Private
workspace



Jupyter
Notebook



Survey



webODV
Import



webODV
Data Extractor



webODV
Data Explorer


DIVA Jupyter
Notebook


VIZ

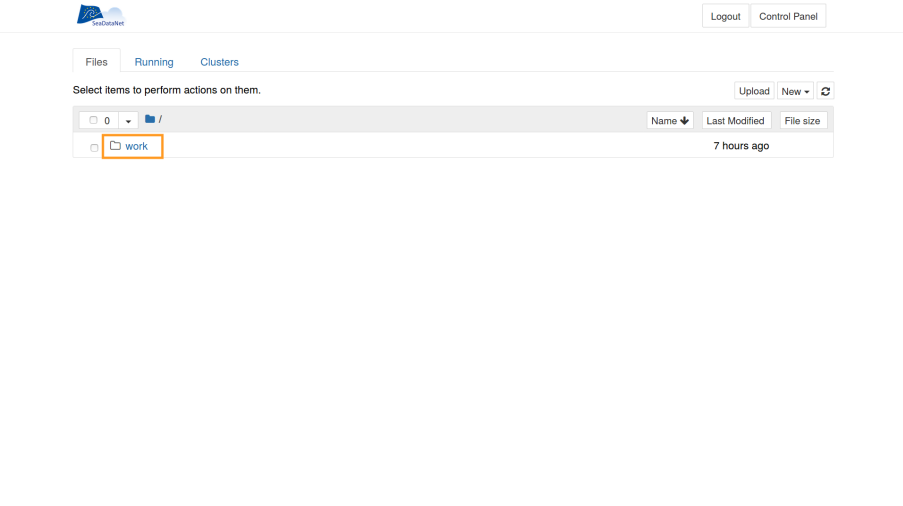

ERDDAP


BioQC


DIVA HP Jupyter
Notebook

DIVAnd in a few clicks

Work directory



The screenshot displays the SeaDataNet web interface. At the top left is the SeaDataNet logo. At the top right are 'Logout' and 'Control Panel' buttons. Below the logo is a navigation bar with 'Files', 'Running', and 'Clusters' tabs. The 'Files' tab is active. Below the navigation bar is the text 'Select items to perform actions on them.' To the right of this text are 'Upload', 'New', and a refresh icon. Below this is a table listing files and folders. The first row shows a folder icon, a checkbox, the number '0', a dropdown arrow, a folder icon, and a slash '/'. The second row shows a folder icon, a checkbox, the name 'work', and the text '7 hours ago'. The 'work' folder name is highlighted with an orange rectangle.

SeaDataNet

Logout Control Panel

Files Running Clusters

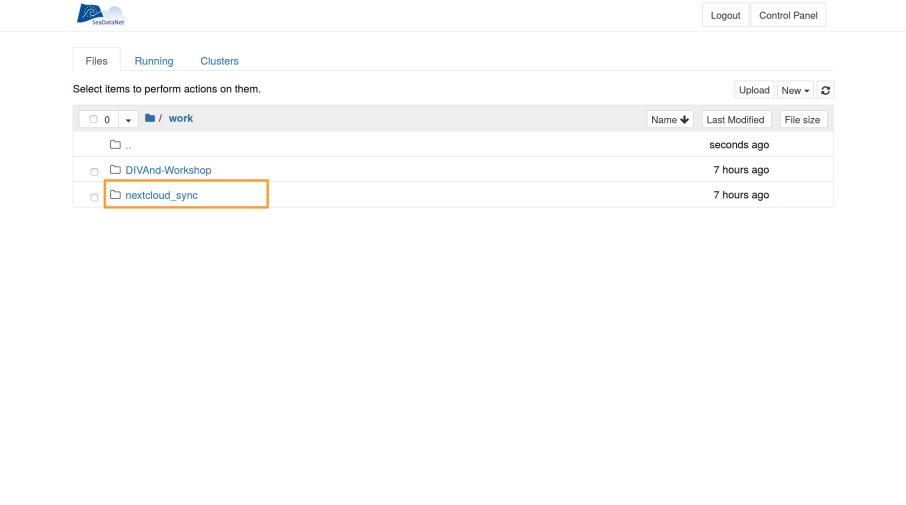
Select items to perform actions on them.

Upload New ↻

				Name ↓	Last Modified	File size
<input type="checkbox"/>	0	▼	📁 /			
<input type="checkbox"/>			📁 work		7 hours ago	

DIVAnd in a few clicks

nextcloud_sync allows users to access their private workspace

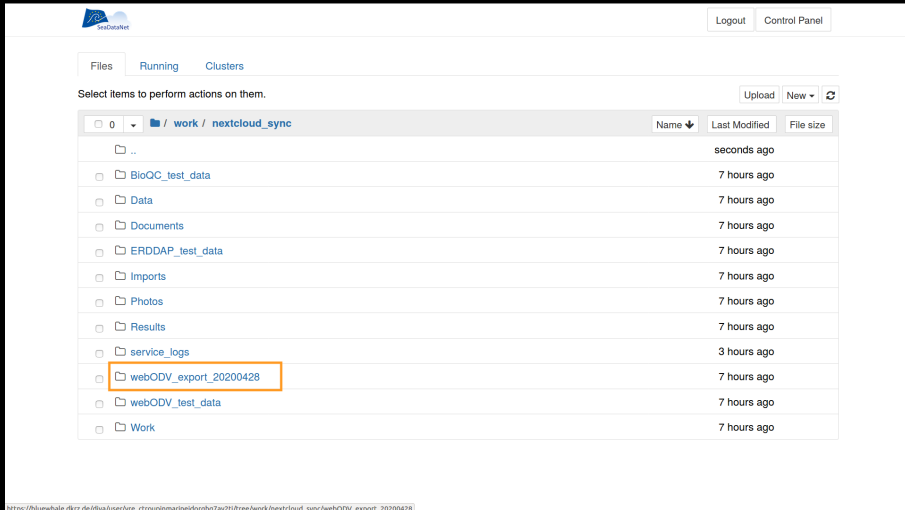


The screenshot displays the SeaDataNet web interface. At the top left is the SeaDataNet logo. At the top right are 'Logout' and 'Control Panel' buttons. Below the header, there are tabs for 'Files', 'Running', and 'Clusters'. A message states 'Select items to perform actions on them.' To the right of this message are 'Upload', 'New', and a refresh icon. The main content area shows a file browser for the '/ work' directory. It contains a table with three rows: '..' (seconds ago), 'DIVAnd-Workshop' (7 hours ago), and 'nextcloud_sync' (7 hours ago). The 'nextcloud_sync' entry is highlighted with an orange border. The table has columns for 'Name', 'Last Modified', and 'File size'.

	Name ↓	Last Modified	File size
<input type="checkbox"/> 0	..	seconds ago	
<input type="checkbox"/>	DIVAnd-Workshop	7 hours ago	
<input type="checkbox"/>	nextcloud_sync	7 hours ago	

DIVAnd in a few clicks

directory where webODV gave me the data files



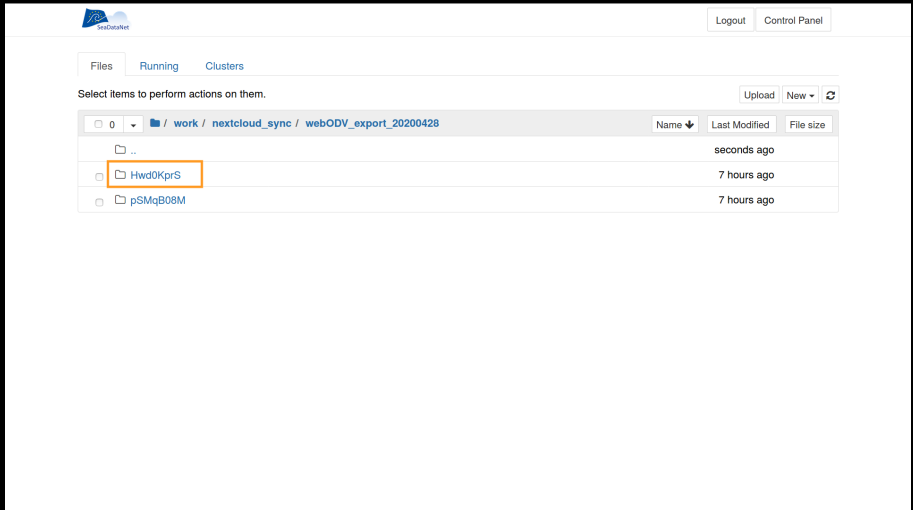
The screenshot shows the SeaDataNet web interface. At the top right, there are links for "Logout" and "Control Panel". Below the header, there are tabs for "Files", "Running", and "Clusters". A message says "Select items to perform actions on them." with buttons for "Upload", "New", and a refresh icon. The main content area shows a file directory structure. The current path is "/ work / nextcloud_sync". The directory listing includes folders like "..", "BioQC_test_data", "Data", "Documents", "ERDDAP_test_data", "Imports", "Photos", "Results", "service_logs", "webODV_export_20200428" (highlighted with an orange box), "webODV_test_data", and "Work". The "Last Modified" column shows times like "seconds ago", "7 hours ago", and "3 hours ago".

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-ctroupinmarineidorghq7ay21/tree/work/nextcloud_sync/webODV_export_20200428

▶ DIVAnd in a few clicks

sub-directory where webODV gave me the data files

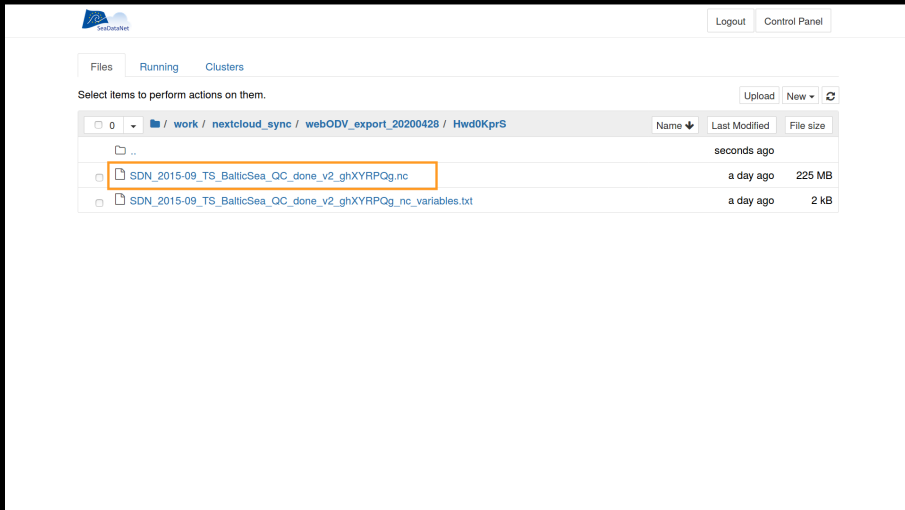


The screenshot shows the SeaDataNet web interface. At the top left is the SeaDataNet logo. At the top right are buttons for "Logout" and "Control Panel". Below the logo are tabs for "Files", "Running", and "Clusters". A message says "Select items to perform actions on them." To the right of this message are buttons for "Upload", "New" (with a dropdown arrow), and a refresh icon. Below this is a file browser showing a directory structure: "0" (with a dropdown arrow) / "work" / "nextcloud_sync" / "webODV_export_20200428". To the right of the path are buttons for "Name" (with a dropdown arrow), "Last Modified", and "File size". The file list contains three items: ".." (seconds ago), "Hwd0KprS" (7 hours ago, highlighted with an orange box), and "pSMqB08M" (7 hours ago).

Name	Last Modified	File size
..	seconds ago	
Hwd0KprS	7 hours ago	
pSMqB08M	7 hours ago	

DIVAnd in a few clicks


data file to use in the notebooks

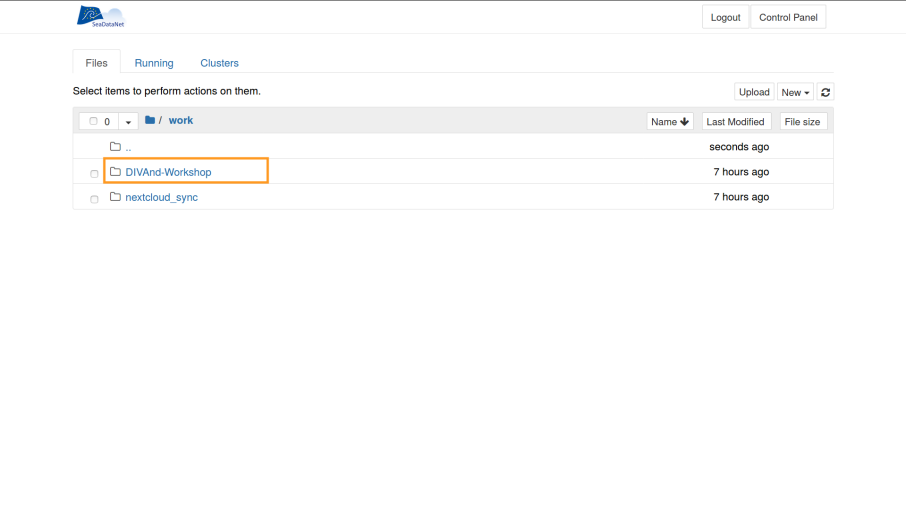


The screenshot shows the SeaDataNet web interface. At the top left is the SeaDataNet logo. At the top right are links for "Logout" and "Control Panel". Below the logo are tabs for "Files", "Running", and "Clusters". A message says "Select items to perform actions on them." To the right of this message are buttons for "Upload", "New" (with a dropdown arrow), and a refresh icon. Below this is a file browser showing the path: `/ work / nextcloud_sync / webODV_export_20200428 / Hwd0KprS`. The file list has columns for "Name", "Last Modified", and "File size". The file `SDN_2015-09_TS_BalticSea_QC_done_v2_ghXYRPQg.nc` is highlighted with an orange box. Below it is the file `SDN_2015-09_TS_BalticSea_QC_done_v2_ghXYRPQg_nc_variables.txt`.

	Name	Last Modified	File size
<input type="checkbox"/>	..	seconds ago	
<input type="checkbox"/>	SDN_2015-09_TS_BalticSea_QC_done_v2_ghXYRPQg.nc	a day ago	225 MB
<input type="checkbox"/>	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




The screenshot shows the SeaDataNet web interface. At the top, there is a logo for SeaDataNet and two buttons: "Logout" and "Control Panel". Below the logo, there are three tabs: "Files", "Running", and "Clusters". The "Files" tab is selected. Below the tabs, there is a text prompt: "Select items to perform actions on them." To the right of this prompt are three buttons: "Upload", "New", and a refresh icon. Below the prompt, there is a table showing the contents of the current directory. The table has three columns: "Name", "Last Modified", and "File size". The table shows three items: a directory named "..", a directory named "DIVAnd-Workshop" (highlighted with an orange box), and a directory named "nextcloud_sync". The "Last Modified" column shows "seconds ago" for "..", "7 hours ago" for "DIVAnd-Workshop", and "7 hours ago" for "nextcloud_sync".

	Name	Last Modified	File size
<input type="checkbox"/>	0 / work		
<input type="checkbox"/>	..	seconds ago	
<input type="checkbox"/>	DIVAnd-Workshop	7 hours ago	
<input type="checkbox"/>	nextcloud_sync	7 hours ago	

DIVAnd in a few clicks

Notebooks divided into subdirectories



LogoutControl Panel

FilesRunningClusters

Select items to perform actions on them.

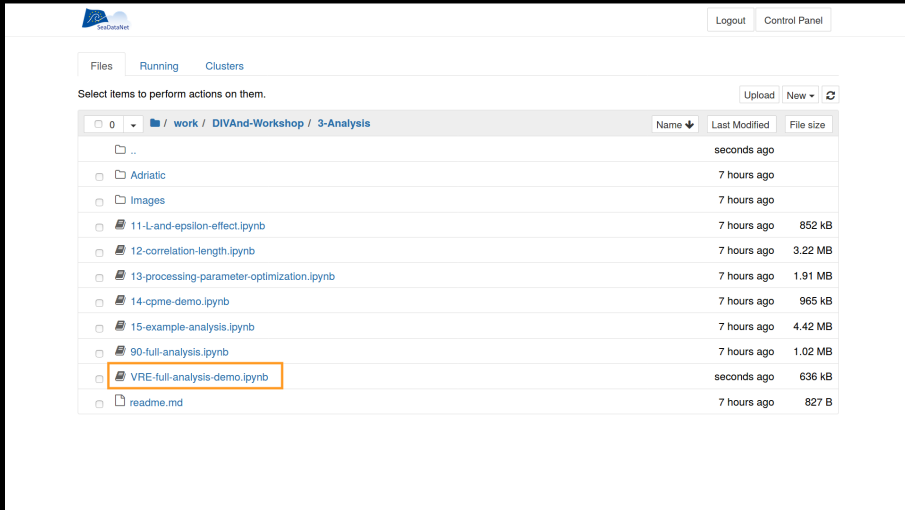
UploadNew↺

<input type="checkbox"/> 0 ▾	/ work / DIVAnd-Workshop	Name ▾	Last Modified	File size
<input type="checkbox"/>	..		seconds ago	
<input type="checkbox"/>	1-Intro		7 hours ago	
<input type="checkbox"/>	2-Preprocessing		7 hours ago	
<input type="checkbox"/>	3-Analysis		7 hours ago	
<input type="checkbox"/>	4-Postprocessing		7 hours ago	
<input type="checkbox"/>	5-AdvancedTopics		7 hours ago	
<input type="checkbox"/>	6-UnderConstruction		7 hours ago	
<input type="checkbox"/>	Adriatic		7 hours ago	
<input type="checkbox"/>	data		7 hours ago	
<input type="checkbox"/>	Exercises		7 hours ago	
<input type="checkbox"/>	Images		7 hours ago	
<input type="checkbox"/>	cressman2.jl		7 hours ago	1.63 kB
<input type="checkbox"/>	OptimInterp.jl		7 hours ago	9.32 kB
<input type="checkbox"/>	README.md		7 hours ago	3.26 kB
<input type="checkbox"/>	Makefile		7 hours ago	791 B

https://bluewhale.dkrz.de/diva/user/vre_ctroupinmarineldorhg7ay2ti/tree/work/DIVAnd-Workshop/3-Analysis

DIVAnd in a few clicks




Full-analysis using the extracted data file















The screenshot displays the SeaDataNet web interface. At the top left is the SeaDataNet logo. On the top right are 'Logout' and 'Control Panel' buttons. Below the logo is a navigation bar with 'Files', 'Running', and 'Clusters' tabs. A message states 'Select items to perform actions on them.' To the right of this message are 'Upload', 'New', and a refresh icon. The main content area shows a file browser for the path '/ work / DIVAnd-Workshop / 3-Analysis'. A table lists the files and folders in this directory, with columns for 'Name', 'Last Modified', and 'File size'. The file 'VRE-full-analysis-demo.ipynb' is highlighted with an orange box.

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

Jupyter notebooks to produce climatologies

 VRE-full-analysis-demo Last Checkpoint: Last Wednesday at 7:31 PM (autosaved)  

File Edit View Insert Cell Kernel Widgets Help Trusted Julia 1.4.1

           Markdown 

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)
```

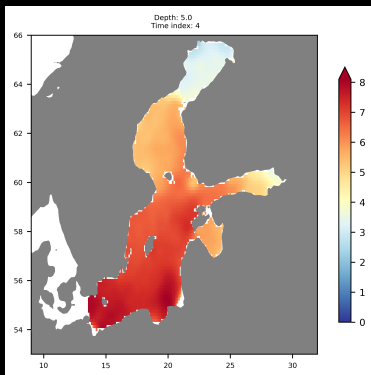
```
Out[6]: true
```

```
In [7]: @time obsval, obslon, obslat, obsdepth, obstime, obsid = NCODV.load(Float64, datafile, "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 - 48.491733891219184 %
```

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)





Improving DIVAnd in the VRE

- 1 Updated mathematical formulation
- 2 Julia language
- 3 Only 2 input files
- 4 User guides = Jupyter notebooks

 Barth et al. 2014

data & bathymetry
all in one

Providing users access to notebooks



- 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?



 <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.

4



DIVAnd.jl package

v2.6.0 (April 27, 2020)

DOI [10.5281/zenodo.3768663](https://doi.org/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 scalefactor is NaN resulting in an empty output variable

Example of heatmaps

Data points are locations acquired by GPS device during running activities

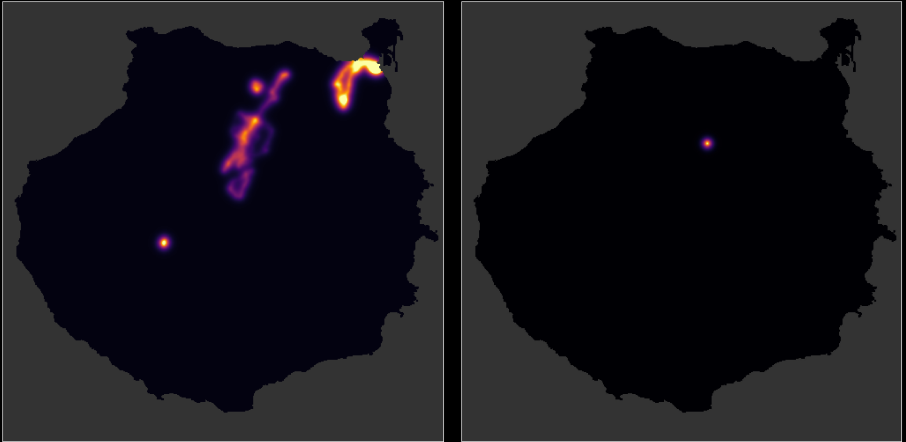


Figure 1: Left: heatmap before the confinement; Right: after the confinement

Training notebooks (Diva-Workshop)

v1.2 (April 27, 2020)

DOI [10.5281/zenodo.3769037](https://doi.org/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

The screenshot displays a JupyterLab environment with a Julia 1.4.1 kernel. The 'Kernel' menu is open, showing options: Interrupt, Restart, Restart & Clear Output, Restart & Run All, Reconnect, Shutdown, and Change kernel. A tooltip for 'Change kernel' lists the following options:

- Julia (DIVAnd precompiled) 1.4.1
- Julia (DIVAnd precompiled, 4 CPUs) 1.4.1
- Julia 1.4.1
- Julia with 4 CPUs 1.4.1

The background shows a notebook with the following code and output:

```
1. Read your
Adapt the :datadir

In [6]: datadir = "../.."
        datafile = joinp
        isfile(datafile)

Out[6]: true

In [7]: @time obsval, obslon, obslat, obsdep
```

The output of the last cell shows a progress bar for 160000 iterations, with the first few lines visible:

```
0 out of 288789 - 0.0 %
10000 out of 288789 - 3.4636952779442276 %
20000 out of 288789 - 6.927390555088455 %
30000 out of 288789 - 10.391085833832683 %
40000 out of 288789 - 13.85478111177691 %
50000 out of 288789 - 17.318476389721138 %
60000 out of 288789 - 20.782171667665366 %
70000 out of 288789 - 24.245866945609592 %
80000 out of 288789 - 27.70956222355382 %
90000 out of 288789 - 31.17325750149885 %
100000 out of 288789 - 34.636952779442275 %
110000 out of 288789 - 38.1006480573865 %
120000 out of 288789 - 41.56434333533873 %
130000 out of 288789 - 45.02803861327496 %
140000 out of 288789 - 48.491733891219184 %
150000 out of 288789 - 51.955429169163416 %
160000 out of 288789 - 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