PHIDIAS

Prototype of HPC/Data Infrastructure for On-demand Services

Analyzing ocean observations in a HPC infrastructure with DIVAnd

Alexander Barth, Charles Troupin
University of Liège

LIÈGE université

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

The ocean is complex...

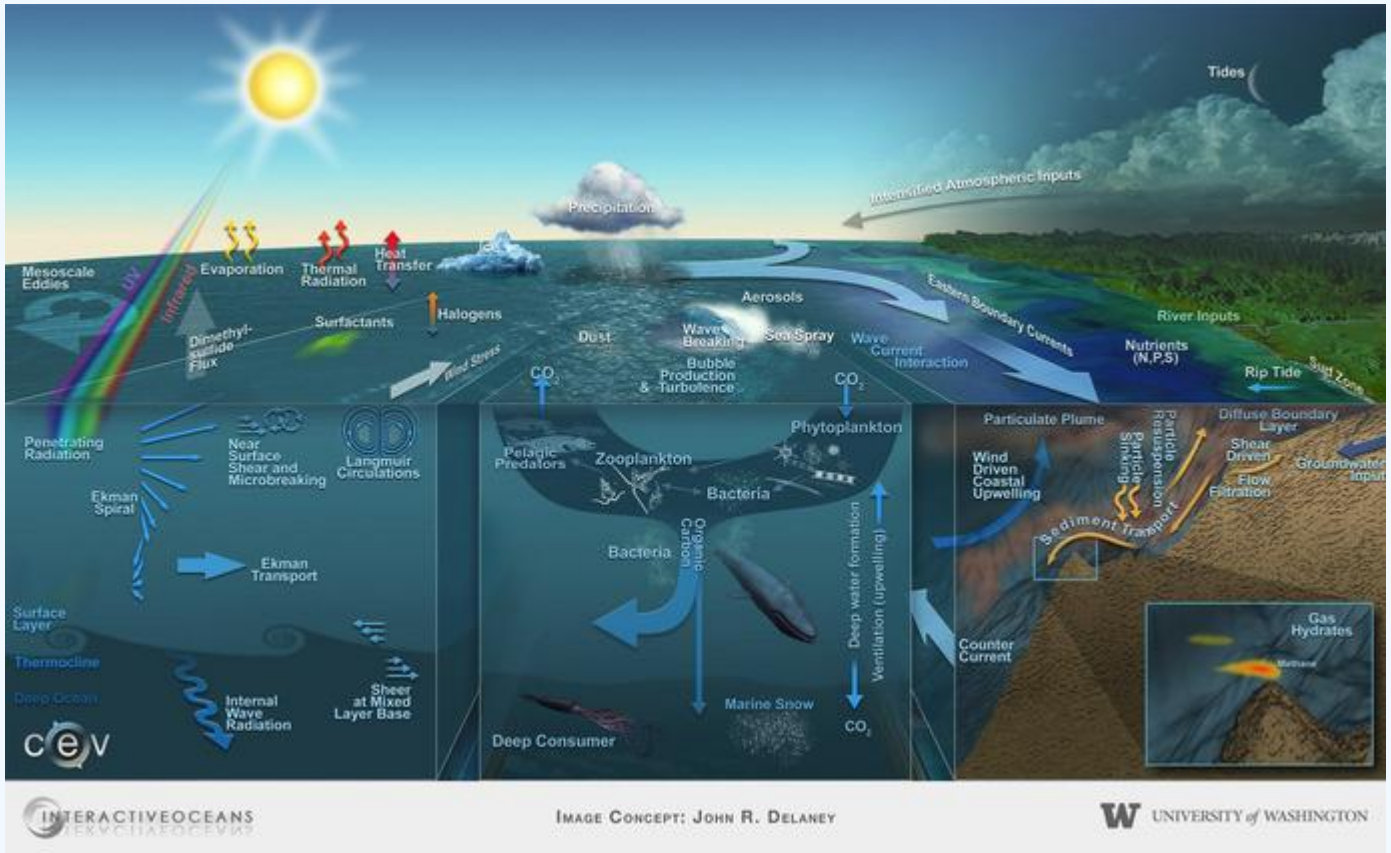


Image creation: Center for Environmental Visualization, University of Washington

- Many ocean processes are present simultaneously
- Non-linear Interaction between them
- Wide time/space spectrum of scales
- → High diversity of ocean observations

… and is complex to observe

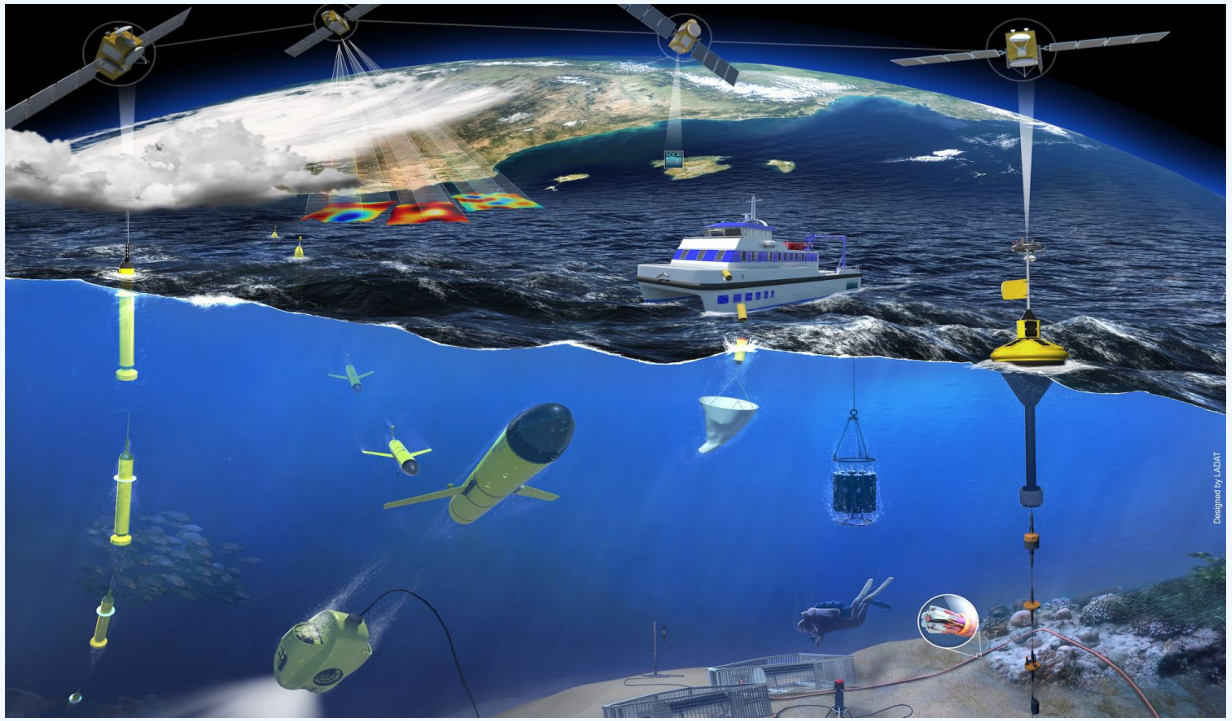


Image credits: ICTS SOCIB

The types of observations observations is quite diverse

Ocean observations are sparse (because expensive)

Yet scientifically very valuable (a measurement not taken it lost forever, the state of the climate and ocean in particular changes)

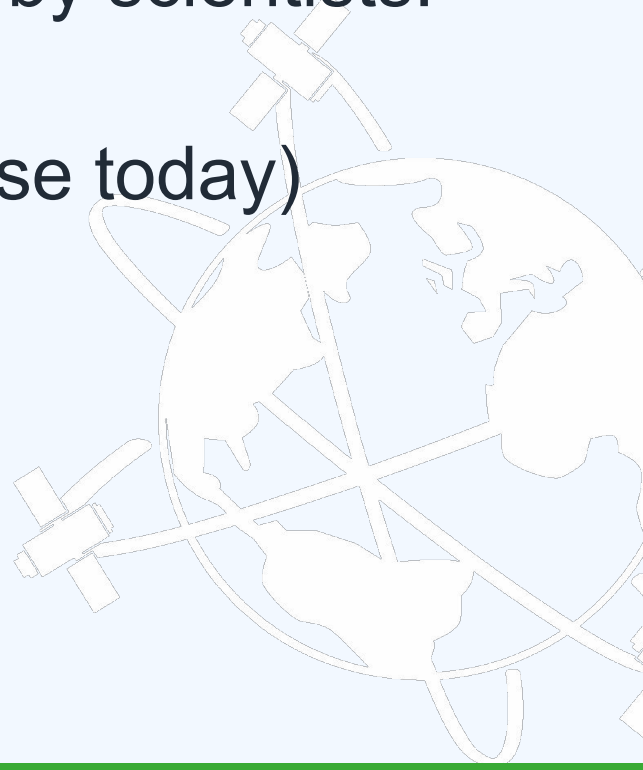
Challenges in ocean data analysis

Fast access to data, multitude of formats, general trend towards netCDF

Different programming environments/languages used by scientists:

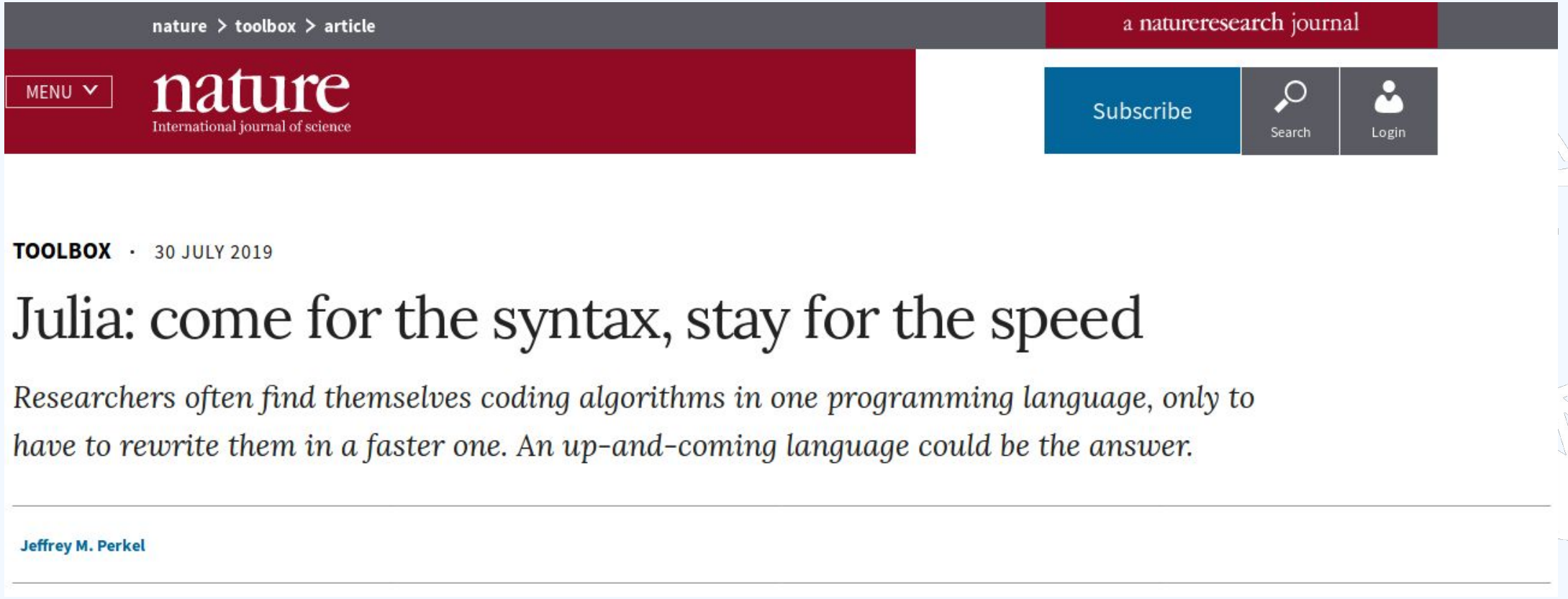
- Fortran (still used in numerical models)
- Matlab (very widespread ~10 years ago, but less use today)
- Python
- R

But also Julia, C, C++, shell scripts,...



Switching to Julia language

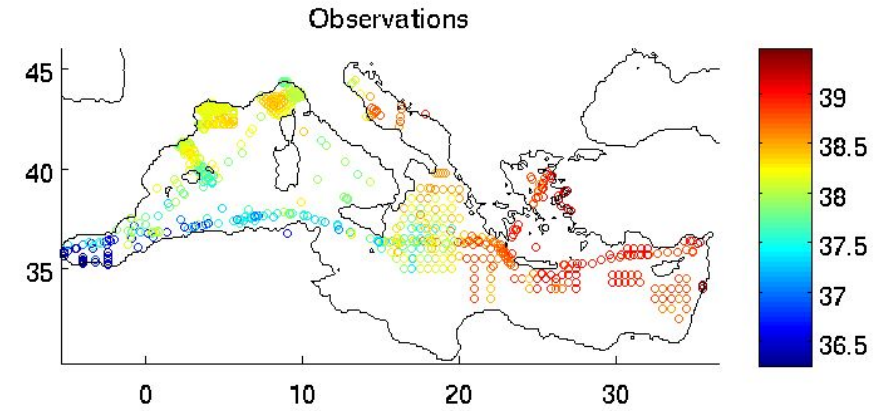
- At GHER, ULiège: started to use Julia to use in 2017
- Julia version 1.0 was released on 8 August 2018

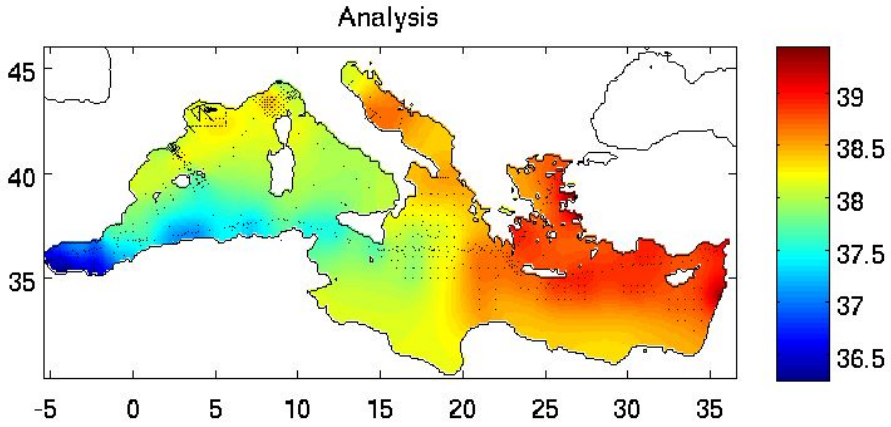


PHIDIAS DIVAnd

Prototype of HPC/Data Infrastructure for On-demand Services

- DIVA: Data Interpolating Variational Analysis
- Objective: derive a **gridded climatology from in situ observations**
- The variational inverse methods aim to derive a continuous field which is:
  - **close to the observations** (it should not necessarily pass through all observations because observations have errors)
  - "**smooth**"
- Spline interpolation



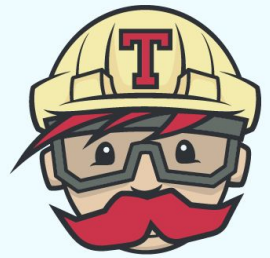


PHIDIAS DIVAnd

Prototype of HPC/Data Infrastructure for On-demand Services

SeaDataNet

- Workshops
- Virtual Research Environment (VRE) in SeaDataCloud
- Jupyter Notebooks
- CI (Continuous Integration) testing (Linux, Mac OS, Windows)
- Docker and Singularity images with preconfigured software

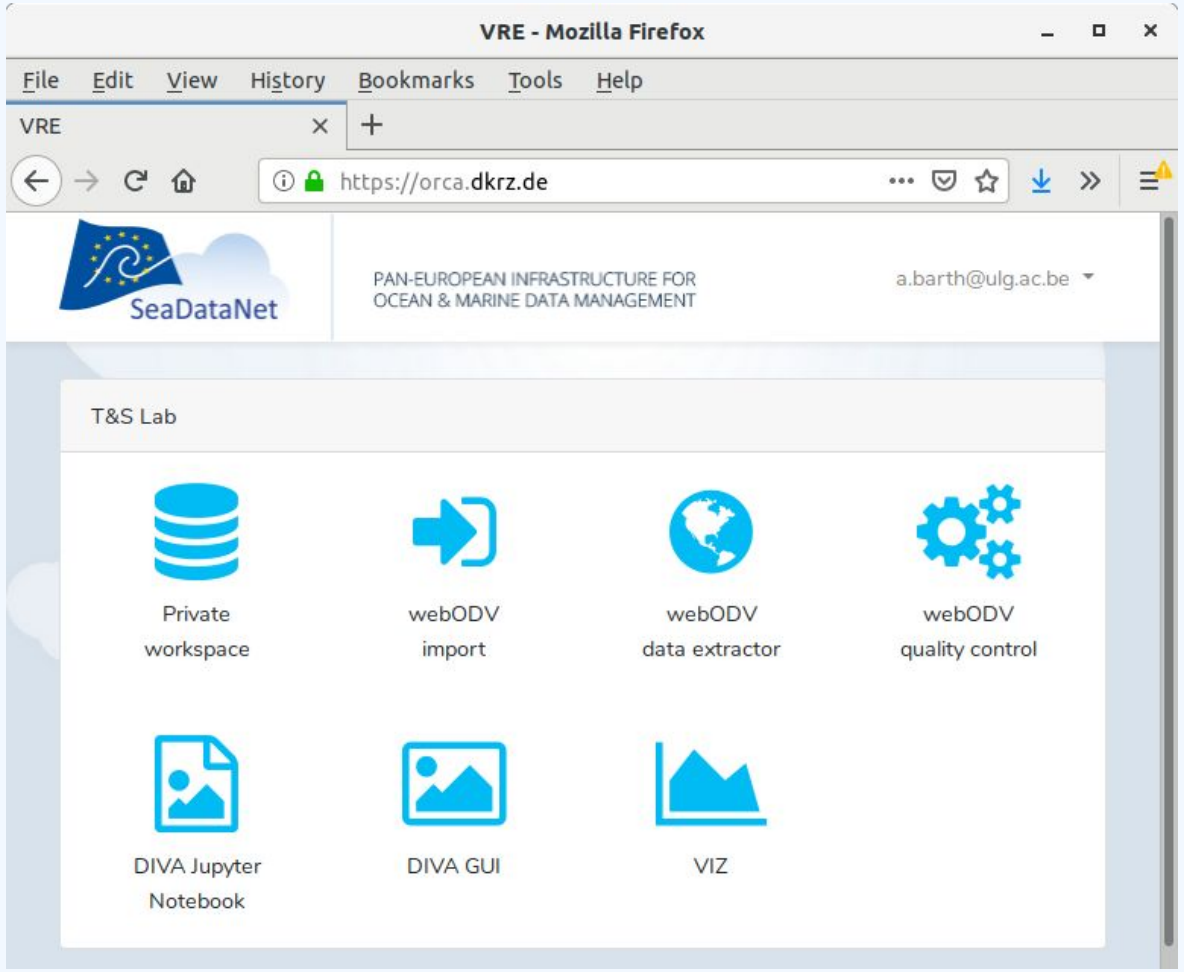


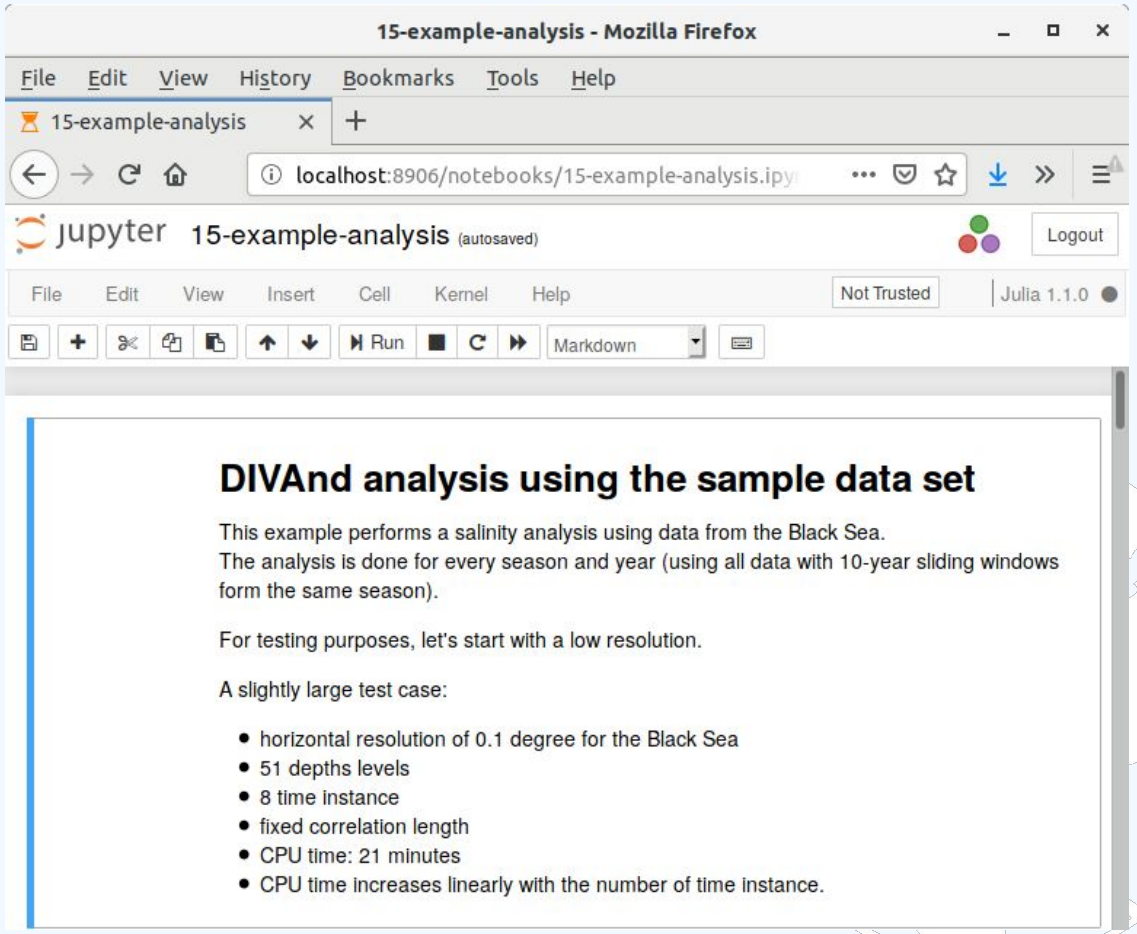
jupyter

docker

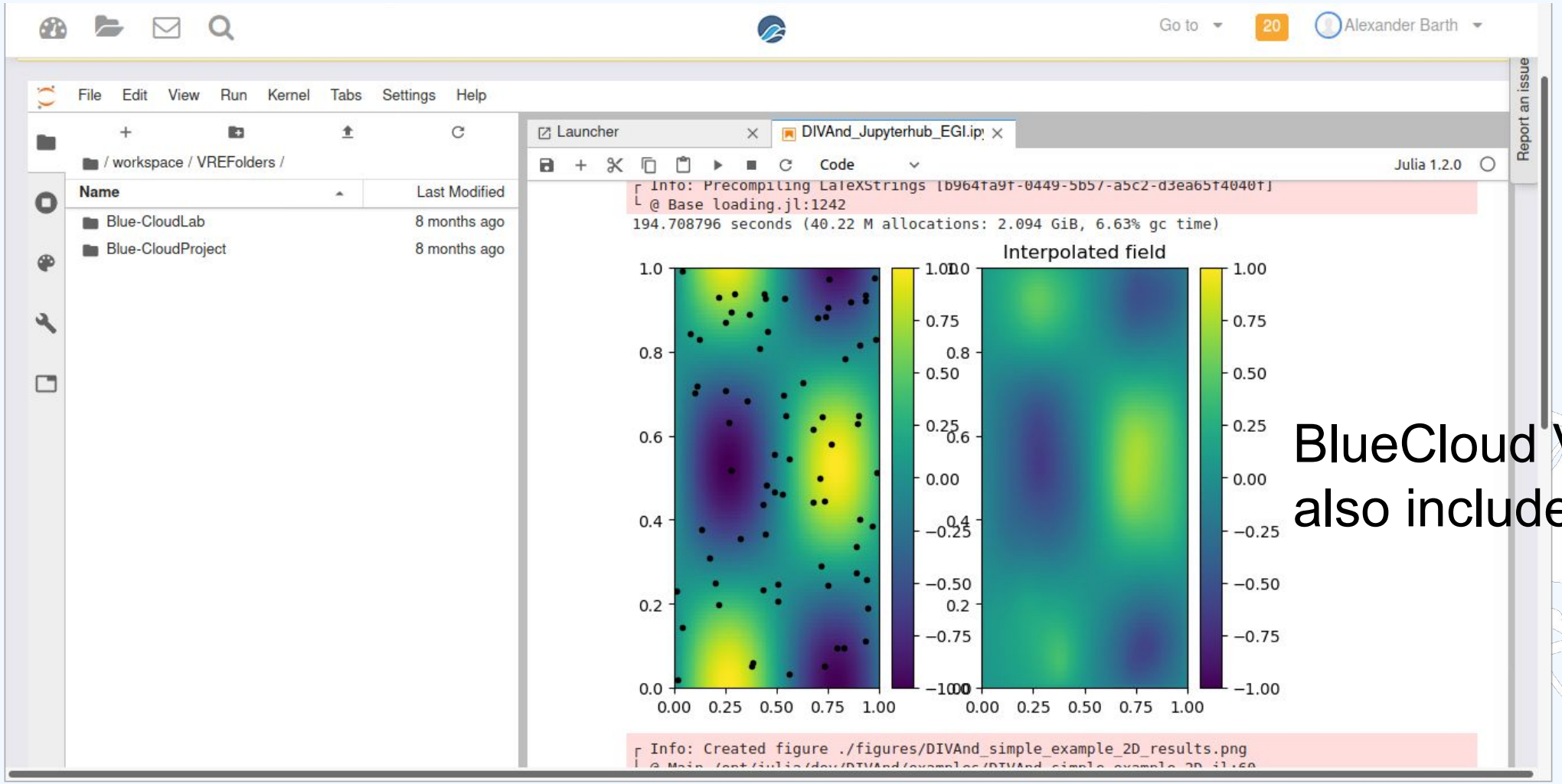
DIVAnd in a virtual research environment

https://vre.seadatanet.org/





BlueCloud VRE



BlueCloud VRE will also include DIVAnd

Computing resource

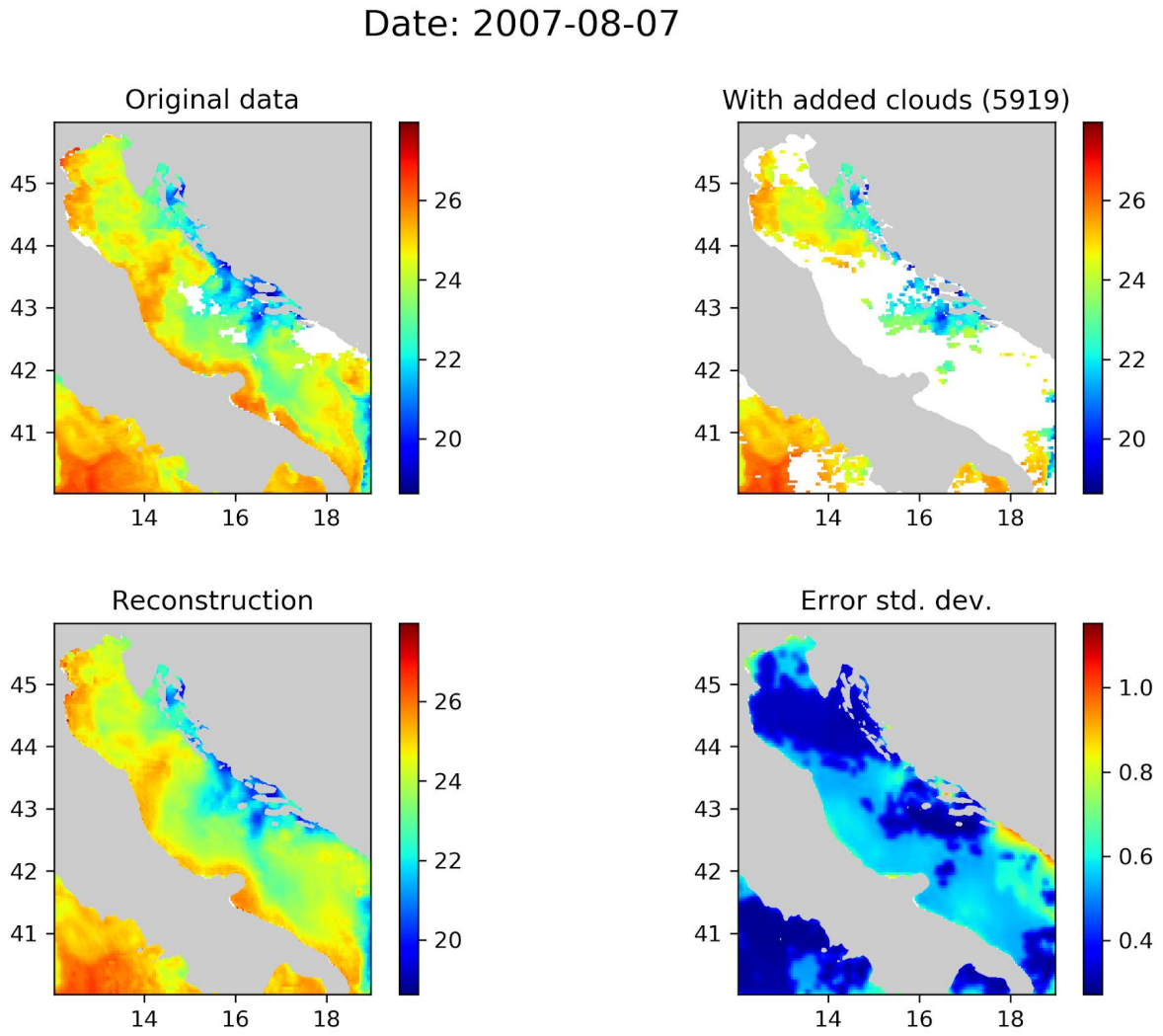
- DIVAnd needs to solve a large matrix system
- The solvers:
  - **direct solver** (`SuiteSparse`, `Cholmod`) requiring a significant amount of memory but a very fast
  - **iterative solvers** (preconditioned conjugate gradient) are more memory efficient but slower
- In practice: the direct solver is preferred as long as the problems fits into the available memory
- But having access to **computing resources with sufficient memory** has been a problem for our users (SeaDataCloud, EMODnet Chemistry)
- Code portability via **Singularity container**

PHIDIAS DINCAE

- Paper: Data INterpolating Convolutional Auto-Encoder
- **Neural network** to reconstruct missing data in satellite images (in particular clouds in remotely sensed Sea Surface Temperature)
- Originally written in Python using `TensorFlow 1`
- Many changes in `TensorFlow 2` -> better alternatives?
- Use Julia and with the `Knet` library
- Training time of the network was **reduced from 3.5 hours to 1.9 hours** (on a NVidia 1080 GPU)
- We use "**data augmentation**" (in particular perturbing input data, add additional clouds,...) using vectorized `numpy` code, but it could be made significantly faster by using Julia instead.

Some results with DINCAE

- Sea Surface Temperature (SST) reconstruction with DINCAE
- **Some data is withheld during the reconstruction** (i.e. additional clouds)
- SST is reconstructed and a **reliable the expected error standard deviation is computed**



DINCAE reconstruction using MODIS sea surface temperature in the Adriatic

PHIDIAS Conclusions

- The types of available **ocean data is quite diverse**
- Fortran is still widely used in the oceanographic HPC community
  - But there are significant challenges to support users outside of a typical HPC environment
  - Julia has been a good fit for us for data analysis
- The original Fortran tool DIVA has been rewritten in Julia (DIVAnd)
- **Jupyter notebooks** provide the users a convenient interface that can also be used in a Virtual Research Environment (especially for data exploration)
- In future: adapt existing tools or adopt new algorithm able to **leverage GPUs** (or other accelerators)

PHIDIAS
Prototype of HPC/Data Infrastructure for On-demand Services

Questions?