WP5 – Training session: RGeostats

Ellip Solutions for IMR dataset

INTAROS – General Assembly

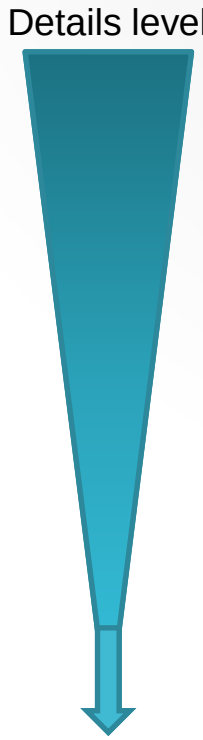
Haus der Wissenschaft,
Sandstrasse 4/5
28195 Bremen, Germany
14.00-17.30
January 10th 2019

ORS Fabien, ARMINES (lead)
RENARD Didier, ARMINES (co-lead)

Outline

Thursday 10th

1. Creating iAOS Processing Services
2. Geostatistics and RGeostats
3. Ellip Notebooks using RGeostats
4. Ellip Worflow using RGeostats
5. IMR Case Study - RGeostats in Action!



Friday 11th

Geostatistics Course & Exercices

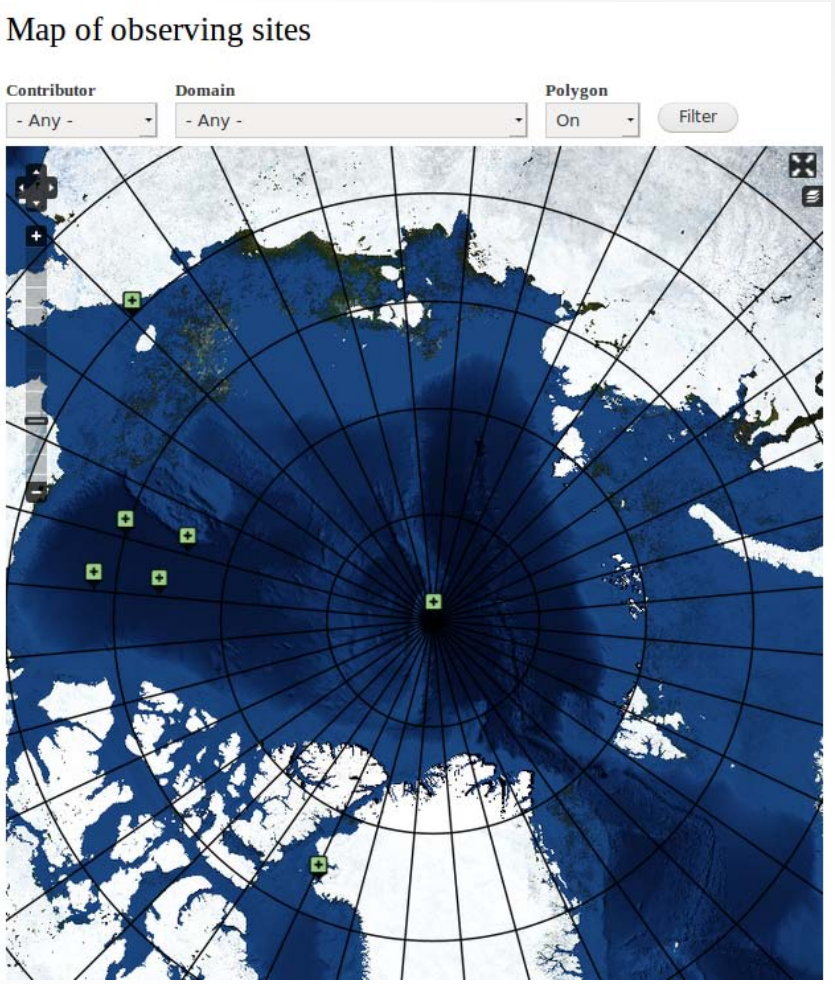
Part #1

Creating iAOS Processing Services

Creating iAOS Processing Services

Why?

- Visualizing and analyzing Pan-Arctic data on the iAOS
- Studying spatial and temporal correlations
- Mapping multiple variables
- Combining different data sources
- Learning more from available data with a deeper analysis



Creating iAOS Processing Services

For who?

- Developers:
  - INTAROS partners (data providers, iAOS developers, WP6 projects…)
  - INTAROS arctic scientific community (researchers, scientists…)
- iAOS users (*visitors*):
  - Scientists (climate, meteorology, biology, oceanography,…)
  - Companies (tourism, earth resources, environments…)
  - Non Governmental Organization (climate change, pollution, life preservation…)
  - European commission (politics)
  - World citizens

Creating iAOS Processing Services

For who?

- **Developers:**
  - **INTAROS partners (data providers, stakeholders…)**
  - **Arctic scientific community (researchers, scientists)**
- iAOS users (*visitors*):
  - Scientists (climatologists, meteorologists, biologists, oceanologists,…)
  - Companies (tourism, earth resources, environments)
  - Non Governmental Organization (climate change, pollution, life preservation)
  - European commission (politics)
  - Curious world citizens

→ The rest of this talk is mainly targeted to **developers**

Creating iAOS Processing Services

New Ellip Solution provided by TERRADUE (sept. 2018)

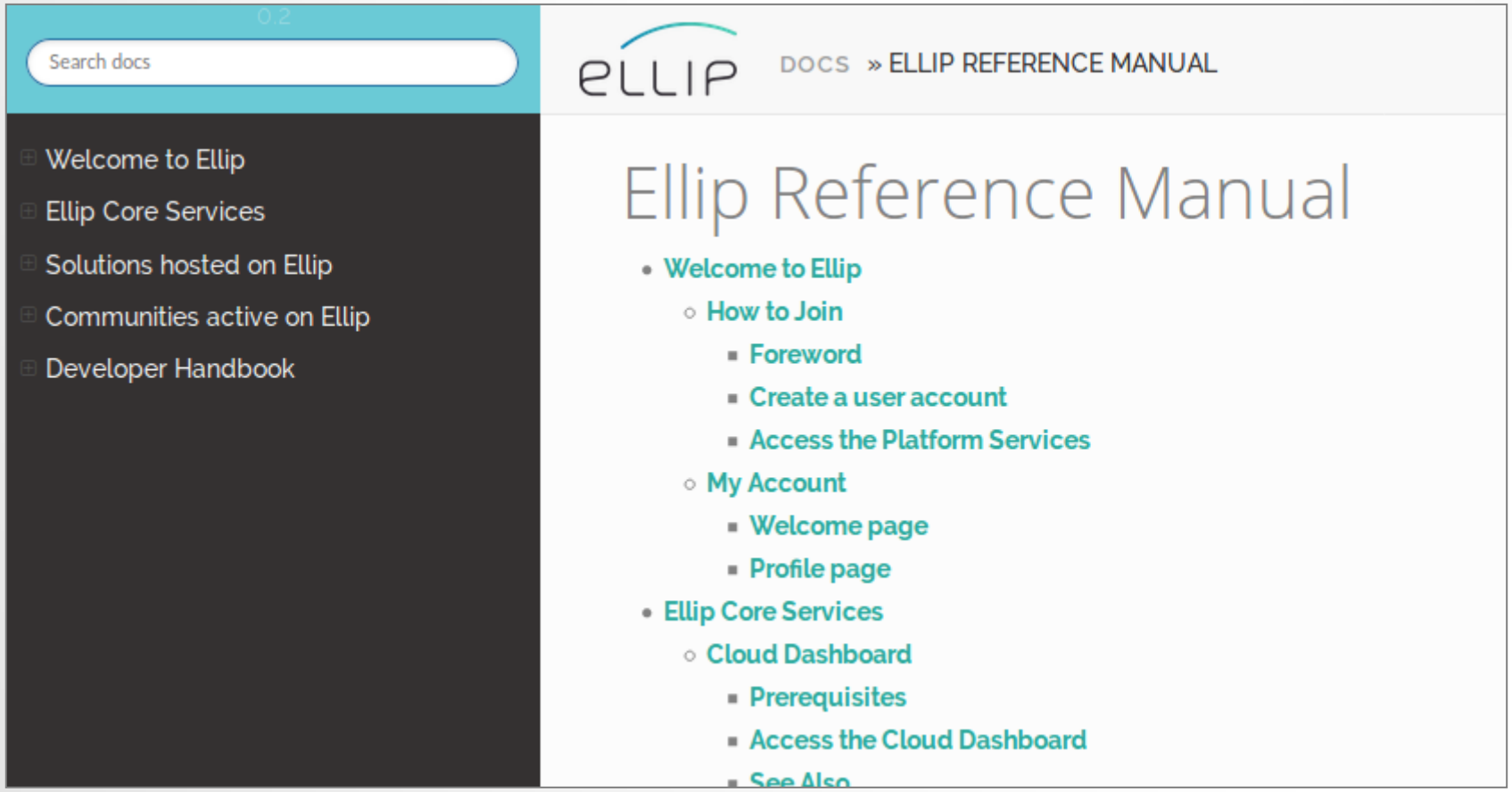
https://notebooks-qa.terradue.com

Objectives:

- Make data processing service **design easier**
- **Build**, **deploy** and **monitor** data processing services

Creating iAOS Processing Services

Full set of Ellip documentation

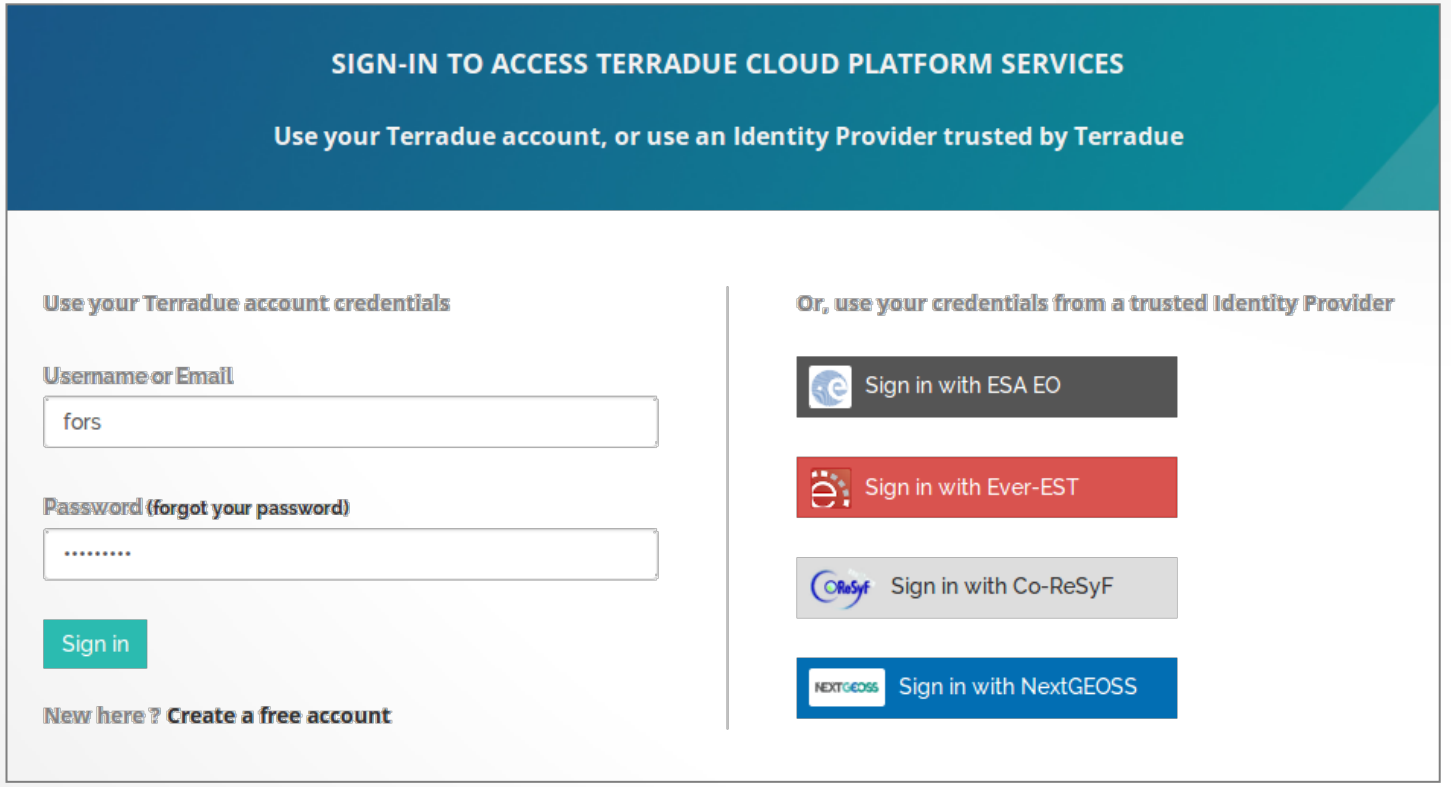


https://docs.terradue.com/ellip

Creating iAOS Processing Services

Become an Ellip user

1. Request iAOS access to Terradue's Ellip Notebooks solution
2. Install Google Chrome browser
3. Sign-in to the Terradue Portal



https://www.terradue.com

Creating iAOS Processing Services

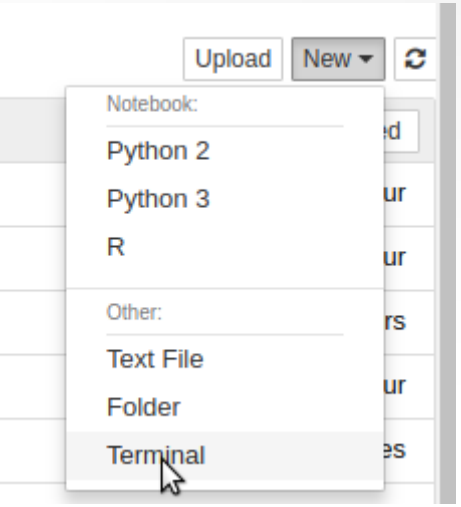
Configure the workshop

1. Open Google Chrome and sign-in to Terradue:

   https://www.terradue.com

2. Browse to Jupyter Notebook and open a Terminal:

   https://notebooks-qa.terradue.com



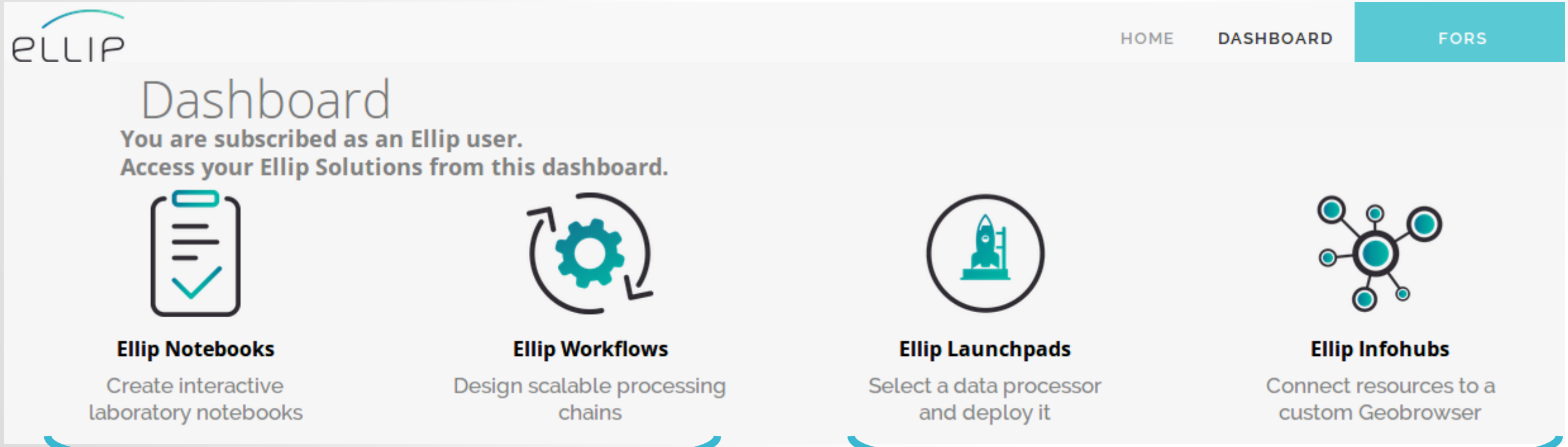
2. Execute the 3 following commands:

```
$ rm –f configure_workshop.sh
$ wget http://rgeostats.free.fr/doc/Files/configure_workshop.sh
$ chmod +x configure_workshop.sh
$ ./configure_workshop.sh
```

This may take several minutes…

Creating iAOS Processing Services

4 steps to create and use data processors

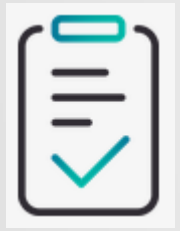


Designing

Deploying on iAOS

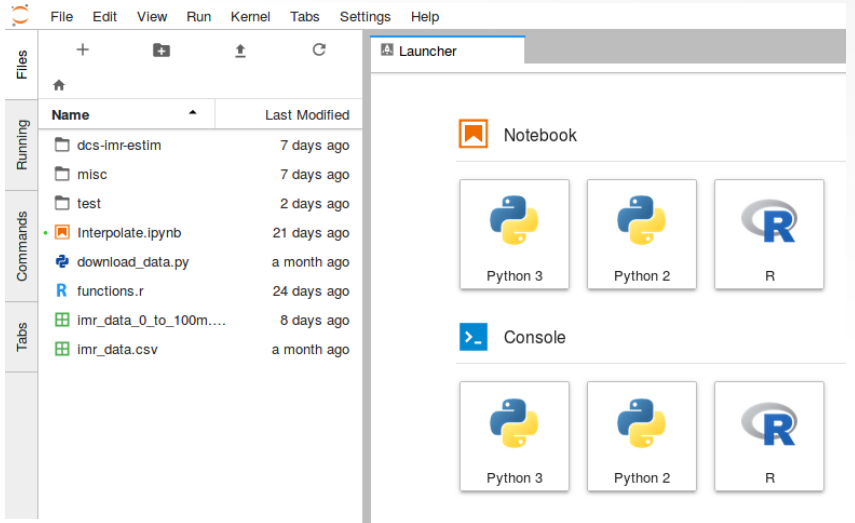
Creating iAOS Processing Services

1. **Designing** data processors

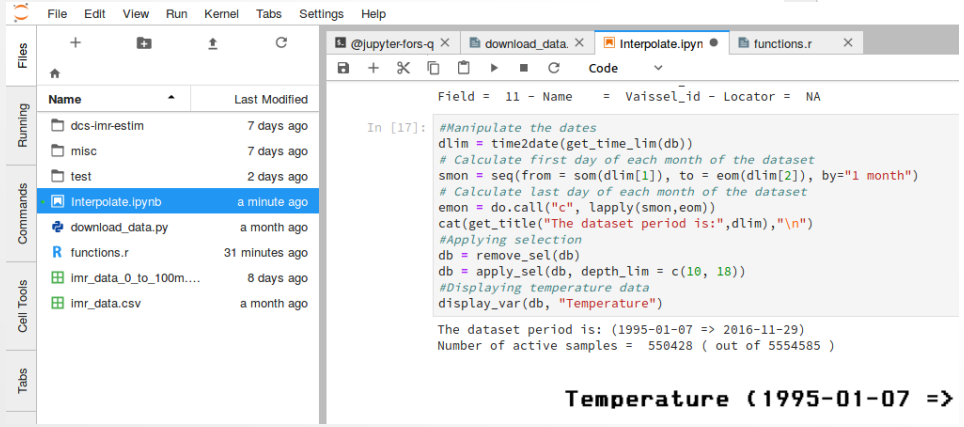


Notebooks

1. Access **Jupyter Notebooks** and your lab: https://notebooks-qa.terradue.com
2. Configure your Jupyter environment
3. Write your Notebooks (Python or R)
4. Test your Notebooks and analyze results



- Python Kernel
- R Kernel
- Terminal



Creating iAOS Processing Services

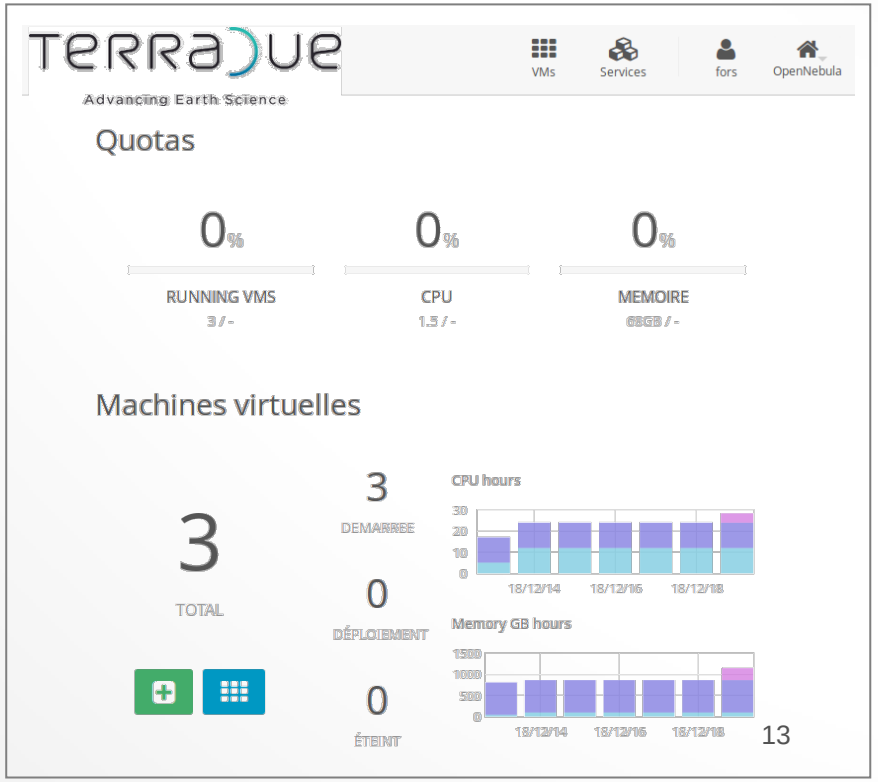
2. Building your workflow (1/2)

Workflows

1. Access to your Cloud Dashboard https://cloud.terradue.com/
2. Create an **Ellip Workflows** Virtual Machine (VM)

While the VM is deploying…

3. Install SSH key pairs
4. Configure the VPN client (**openvpn**)

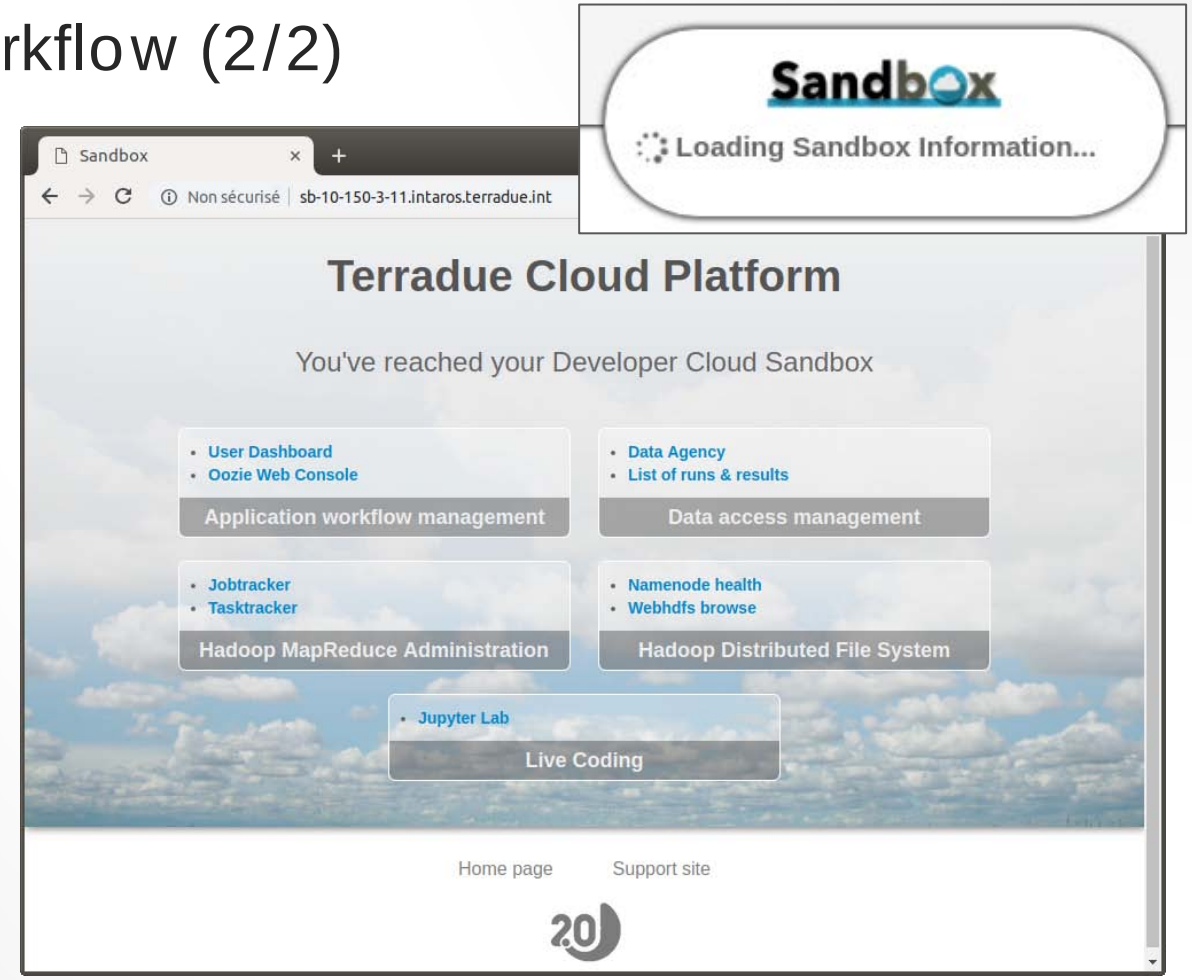


Creating iAOS Processing Services

2. **Building** your workflow (2/2)

Workflows

1. Connect to Terradue's **VPN**
2. Connect via **ssh** to your VM
3. Configure your VM system environment
4. Create an Ellip workflow application using **ciop**, **git** and **mvn**
5. Test your application with the **Sandbox**

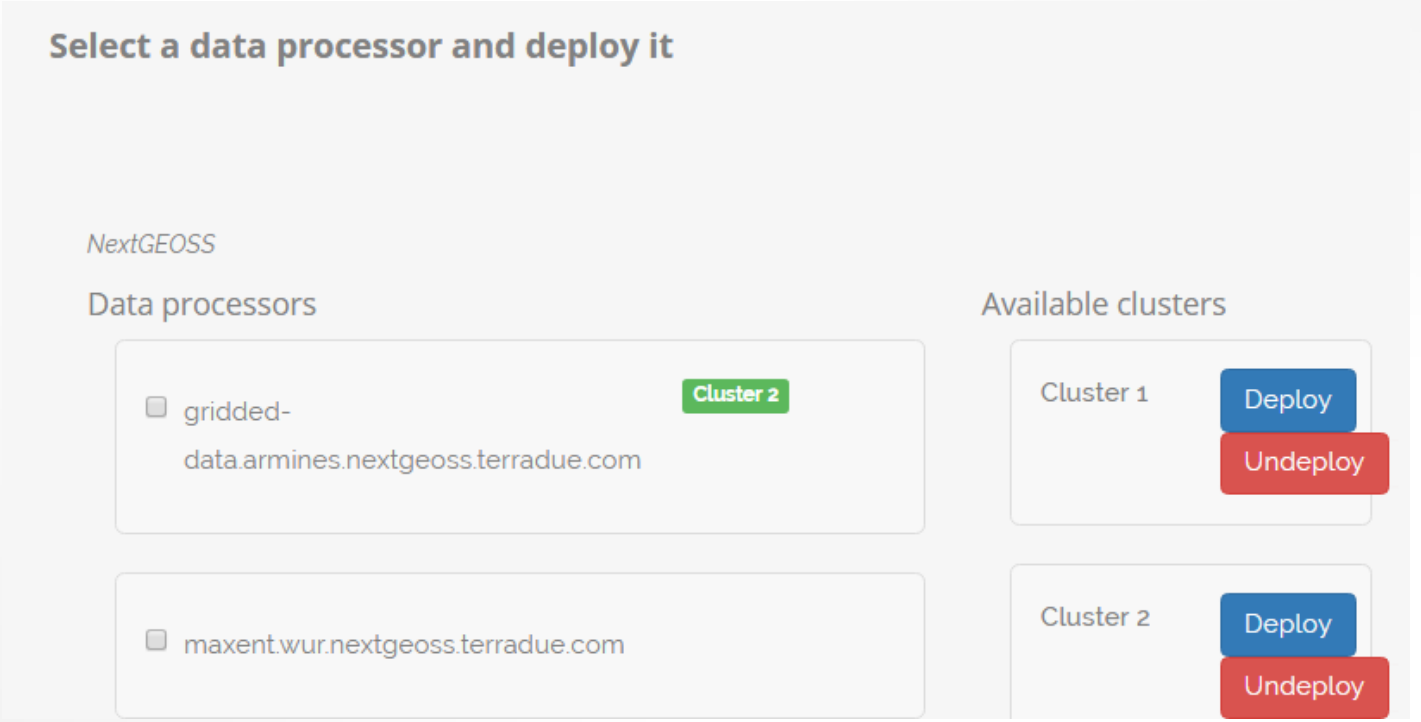


Creating iAOS Processing Services

3. **Deploying** your workflow

Launchpads

1. Manage successive workflow releases (**ciop-release**)
2. Package it in a Web Processing Service (**WPS**)
3. Make it available to geo-portals

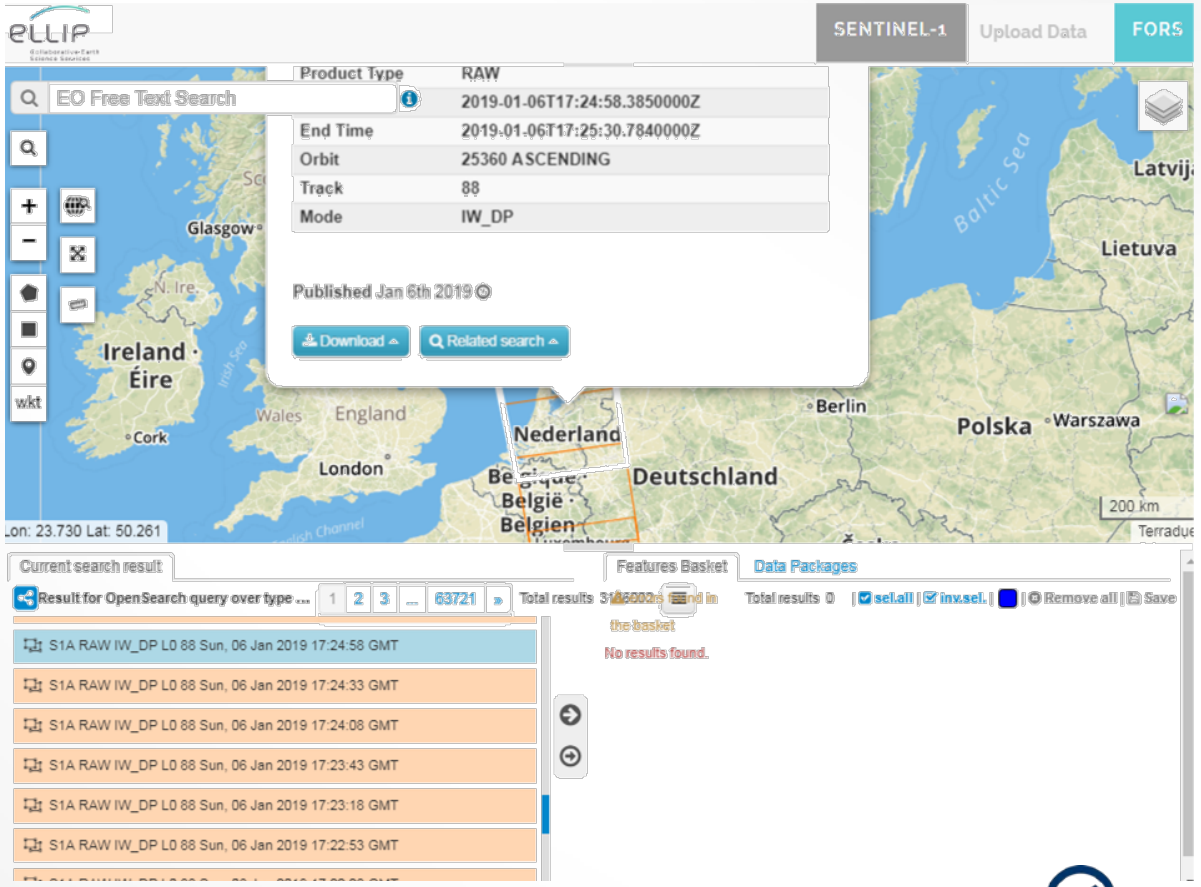


Creating iAOS Processing Services

4. Running your workflow

Infohubs

1. Create a geo-browser application
2. Connect to your data
3. Launch your deployed Web Processing Services



This could be an example of interface for iAOS visitors

Geostatistics and RGeostats

Who are we?

MINES ParisTech | PSL

R Geostats

MINES ParisTech

60 Bd St Michel in Paris

- *Training of engineers*
- *Research centers*
  - **Fontainebleau** — *Earth sciences & Applied mathematics*
    - Centre de Géosciences
  - **Corbeil - Evry**
  - **Sophia Antipolis**

INTAROS

Who are we?

Centre de Géosciences in Fontainebleau: Géostatistique

Names in Geostatistics

- D.G. Krige (mining engineer) & H.S. Sichel (statistician):
  Gold mining (1950)
- L. Gandin:
  Objective Analysis of Meteorological Fields (1965)
- B. Matérn:
  Importance of spatial dependence and variation in forestry(1960)
- G. Matheron:
  - Formalization of the theory of regionalized variables, Ecole des Mines de Paris (1962)
  - Inception of Mathematical Morphology discipline
- Many fellows / students:
  - M. David, A. Journel, J. Davis, D. Merriam, R. Webster, N. Cressie, …

Geostatistics

- Describe the spatial characteristics of the variable (variogram): classification, spatial correlation
- Estimation, interpolation (kriging)
- Simulations: possible alternative scenarios
- Appraisal of uncertainty
- Risk assessment

Applications

- Soil Science, Topography, Geology
- Mining, Petroleum
- Hydrology
- Biology, Epidemiology, Health, Ecology
- Environmental monitoring and assessment
- Forestry, Fishing, Agronomy
- Atmospheric Sciences
- Any discipline with spatial data

RGeostats

- Created in Centre de Géosciences of MINES ParisTech
- Initiated for Fish Industry within a European project: GEFA (2001)
- Named RGeoS and expanded possibilities:
  - Based on a separate commercial library Geoslib (written in C)
  - Mapping its functions in R language (using Rcpp layer)
- Renamed in RGeostats in 2014
- Used for testing new methodologies and/or for teaching purpose
- (Free) Download from: http://cg.ensmp.fr/rgeostats
- Provides demonstration scripts, FAQ, forum for posting questions, asking for help and benefiting from users experience

Trainings

More extensive courses in Geostatistics are available:

- *Les méthodes de la Géostatistique* (in French): October 2019 in Paris & Fontainebleau
- *C.F.S.G.* (in English) : ten months from September to July
- *Mining professional training* (during academic year in English)
- *Introduction to Geostatistics* (in English): 18-22 Fev 2019 in Paris

More information:
http://www.geosciences.mines-paristech.fr

Contact:
nathalie.dietrich@mines-paristech.fr

Bibliography

G. Matheron:

- 1962-1963: *Treatise of applied geostatistics* (in French), Technip and BRGM editions, Paris
- 1965: *Regionalized variables and their estimation* (in French), Masson, Paris
- 1967: *Elements for a theory of porous media* (in French), Masson, Paris
- 1968: *Treatise of applied geostatistics* (in Russian), MIR, Moscow
- 1969: *Theory of random sets* (in French), Ecole des Mines de Paris
- 1969: *Geostatistics course* (in French), Mines Paris
- 1969: *Universal kriging* (in French), Mines Paris
- 1970: *Mathematical morphology* (in French), Mines Paris
- 1970: *The theory of regionalized variables and its applications*, Mines Paris
- 1972-1975: *Random sets and integral geometry*, Wiley, New York
- 1978-1989: *Estimating and choosing*, Springer, Berlin

Bibliography

From others:

- Armstrong M., Galli A., Beucher H., Le Loc'h G., Renard D., Doligez B., Eschard R., Geffroy F., *Plurigaussian Simulations in Geosciences*, 2011. 176p.
- Chilès J.P., Delfiner P. *Geostatistics : modeling spatial uncertainty*. N.Y. : Wiley, 1999. 695p. 2nd edition 2012.
- Journel A., Huijbregts C. *Mining geostatistics*. London : Academic Press, 1978. 600p.
- Lantuéjoul C. Geostatistical simulation : models and algorithms. Berlin : Springer, 2001. 256p.
- Rivoirard J. Introduction to disjunctive kriging and non-linear geostatistics. Oxford : Clarendon Press, 1994. 181p.
- Wackernagel H. Multivariate geostatistics: an introduction with applications. 3rd ed. Berlin : Springer, 2003. 387p.

Part #3

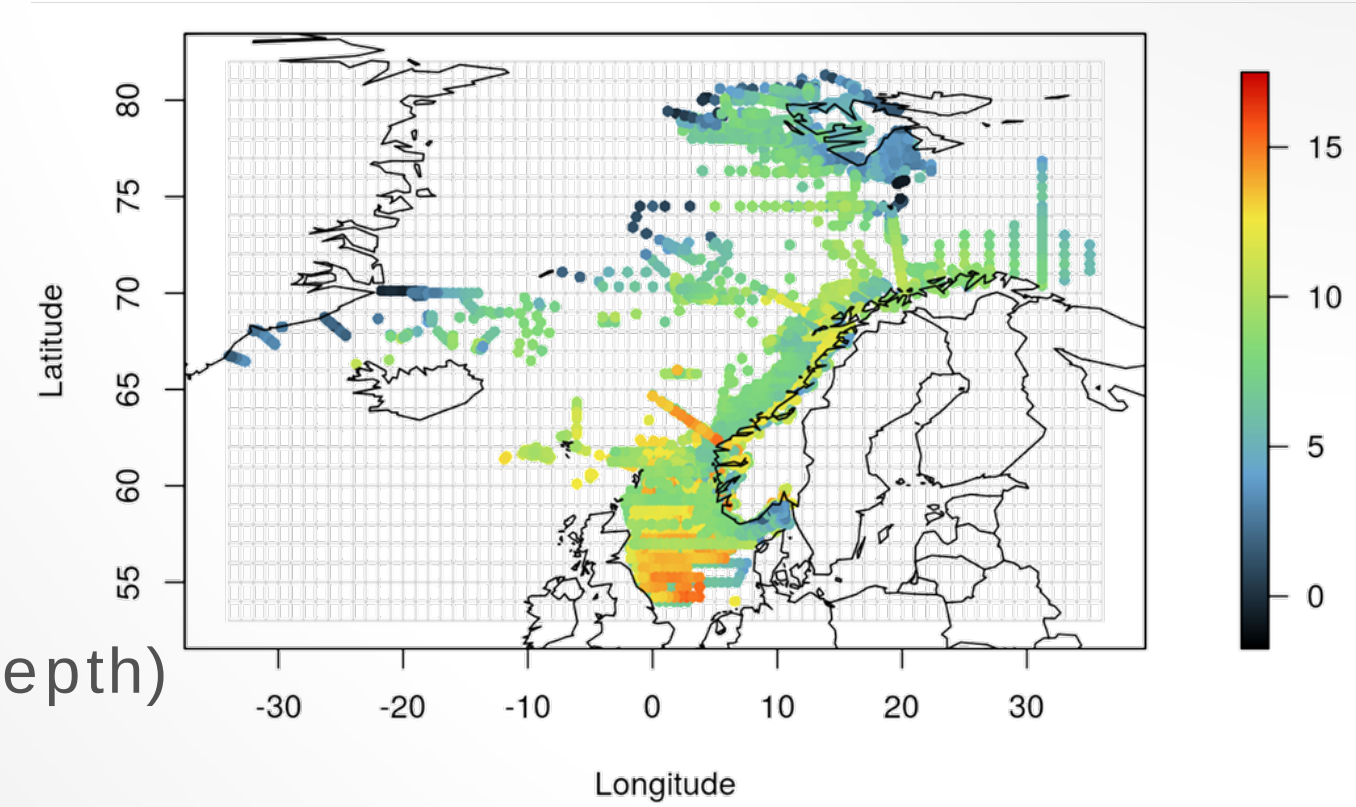
Ellip Notebooks using RGeostats

Ellip Notebooks using RGeostats

IMR Dataset global overview

- 7 vessels
- from 1995 to 2016
- 3 variables measured:
  - Temperature
  - Salinity
  - Conductivity

- 63 500 positions {long, lat}
- 63 500 vertical profiles (in depth)
- A few million samples
- 84 NetCDF files (~60 Mb each)



Ellip Notebooks using RGeostats

Work in the cloud!

IMR OpenDap Server

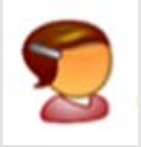
data

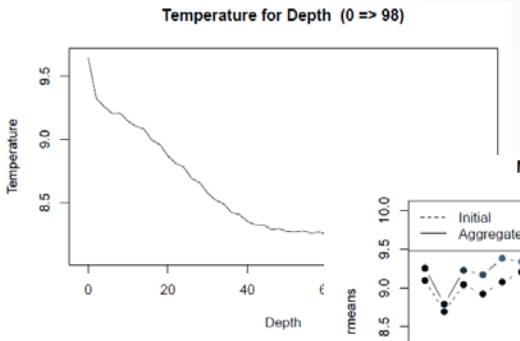
Jupyter Notebooks Server

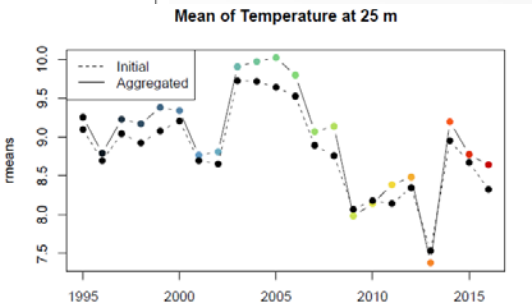
RGeostats

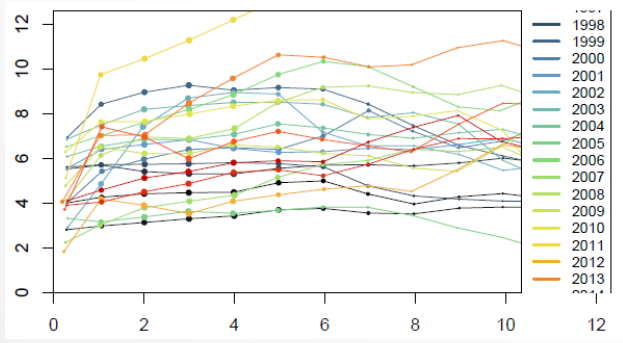
jupyter

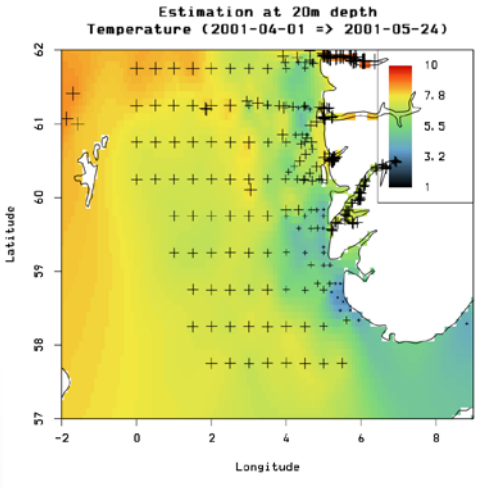
Internet











Ellip Notebooks using RGeostats

Available unitary jobs for IMR dataset

One **Python** job for fetching data:

1. Subsetting by location, depth or time period from IMR OpenDAP server

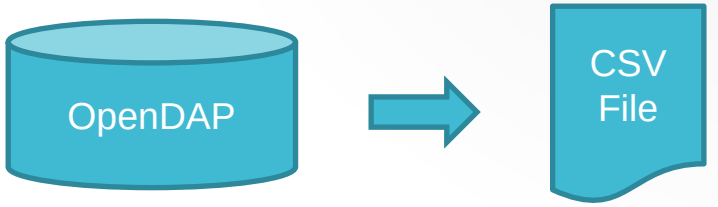
Independent RGeostats **R scripts** jobs:

2. Display basemap of a variable
3. Subsetting / Filtering
4. Statistics by cells
5. Multi-variable correlation
6. 2D aggregation along time or depth
7. Variable ploting through time or depth
8. Interpolation map using kriging
9. …

Ellip Notebooks using RGeostats

Python job for downloading data

Python script fetching data into a CSV file



IMR OpenDAP server:
Thredds server

Catalog URL:
http://opendap1-test.nodc.no/thredds/catalogs/physics/physics_point_yearly.xml

File URLs:
http://opendap1-test.nodc.no/thredds/dodsC/physics/point/yearly/58xx_CTD_yyyy.nc

Variables dumped in CSV:
Longitude, Latitude, Depth, Time, Vessel_id, Vessel_name, Profile_id, **Temperature**, **Salinity**, **Conductivity**

Ellip Notebooks using RGeostats

Download known issues

IMR OpenDAP access directly through R:

- **Rdap** package: Not operational and not maintained
- **Pydap** module through **reticulate** package: no translation of Pydap indexing within reticulate

⇒ Current workaround: **netCDF4** python module

⇒ Fetching / subsetting data with python script

⇒ Creating CSV intermediate file

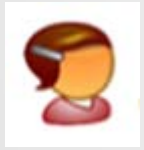
⇒ Loading this CSV file for creating the RGeostats Db

But:

- No numpy advanced indexing support
- No meta dataset covering all boats (yet)
- IMR Remote access to Hyrax OpenDAP server is slow (has to be solved)

Ellip Notebooks using RGeostats

Wrapping RGeostats functions



imr_study.ipynb

Jupyter Notebooks

Unitary jobs for each notebook cell:

Statistics / Aggregation / **Interpolation**

CSV File

functions.r

IMR dedicated functions

read_csv / write_csv / load_data
date2time / time2date / aggregate_depth
apply_sel / remove_sel / **interpolate_2D**
display_var / display_stats / **display_result**

Db

RGeostats

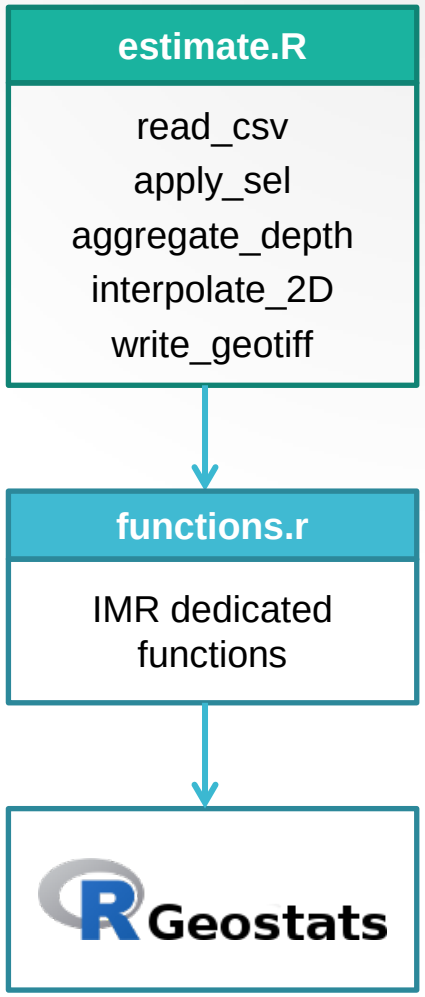
*db.create / db.locate / db.sel / **db.plot***
vario.calc / model.auto / kriging
blockstat / db.regularize

INTAROS

Ellip Notebooks using RGeostats

Interpolation service example : **estimate.R**

```
# Load the CSV file
dbr = read_csv(csv_file)

# Select measures in the required interval (time and aera)
dbr = apply_sel(dbr, …)

# Aggregate all measures around 20m depth
dbr = aggregate_depth(dbr, 20, …)

# Interpolate Temperature values
res = interpolate_2D(dbr, "Temperature", …)

# Display results
display_result(dbr, res,…)
```



Part #4

Ellip Worflow using RGeostats

Ellip Worflow using RGeostats

Application creation

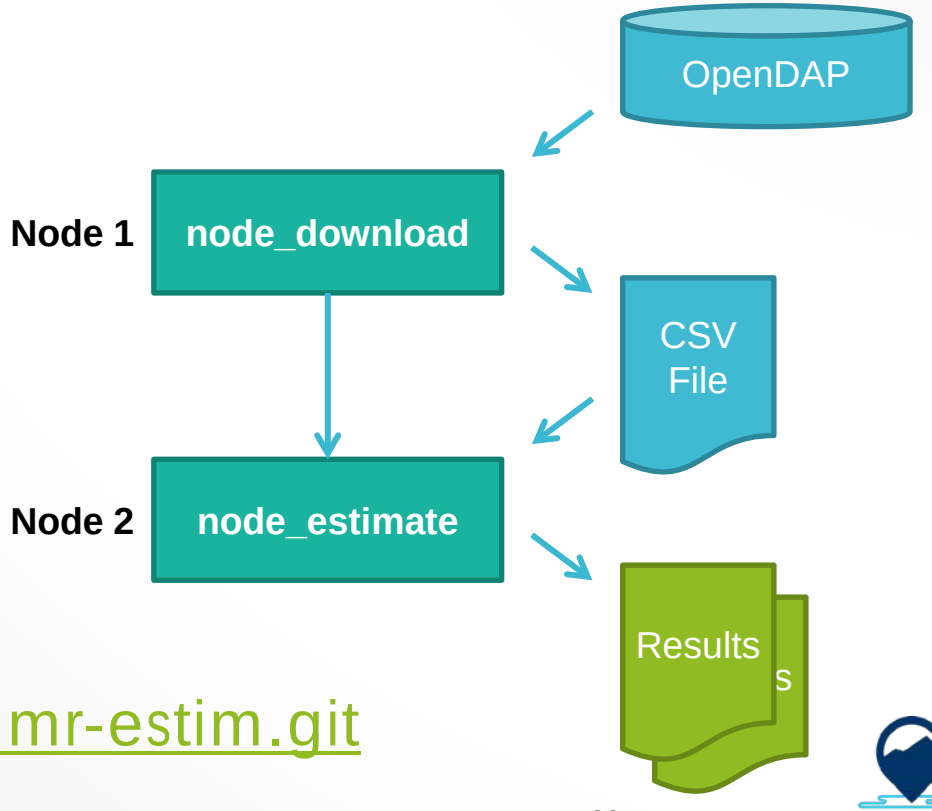
New workflow: `mvn archetype:generate`

Type: **3** (R workflow with 2 nodes)

Group Id: **com.terradue**

Community: **ec-intaros**

Name: **dcs-imr-estim**



Git repository:

https://gitlab.com/ec-intaros/dcs-imr-estim.git

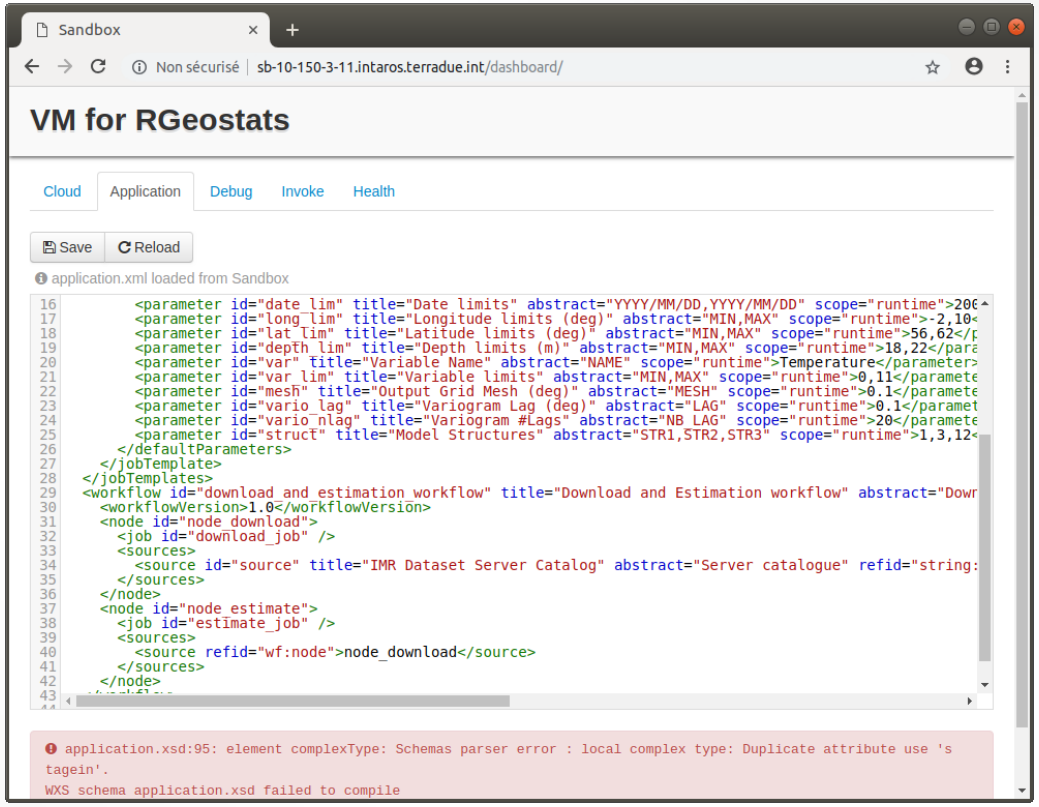
Ellip Worflow using RGeostats

Application.xml file

- Job Templates:
  - Streaming executable = R entry points:

    node_download/run.R
    node_estimate/run.R
  - Job parameters

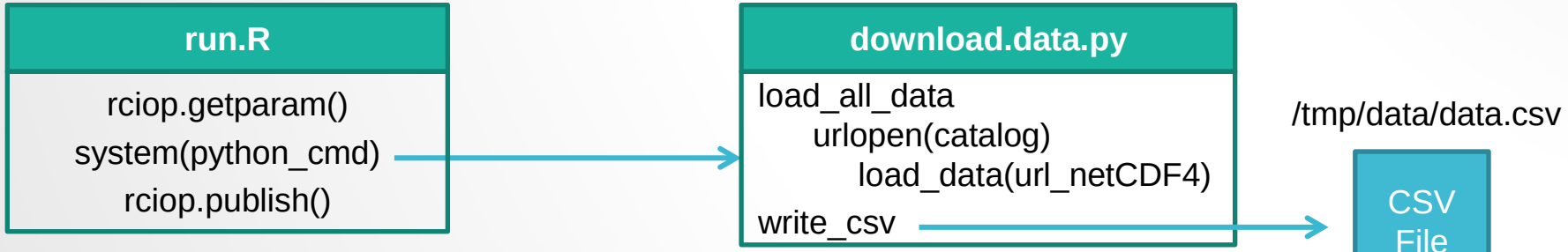
    (**long_lim**: Longitude limits
    vario_nlag: Variogram lags)
- Workflow sequence:
  - Nodes identification
  - Data sources
    (catalog URL or results from previous node)



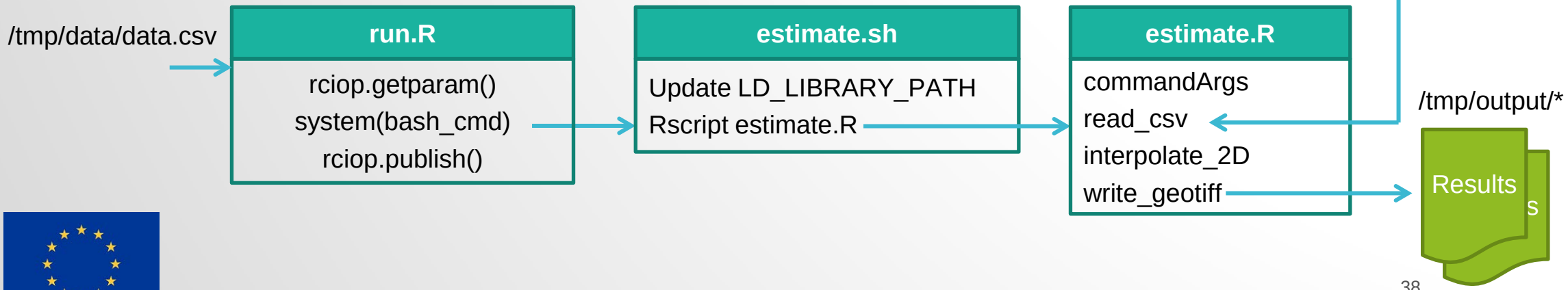
Ellip Worflow using RGeostats

Nodes description

- node_download:



- node_estimate:



Ellip Worflow using RGeostats

Run the workflow

1. Synchronise the application with your git repository (**git pull**)
2. Install the application (**mvn clean install**)
3. Launch the application (**ciop-run**)

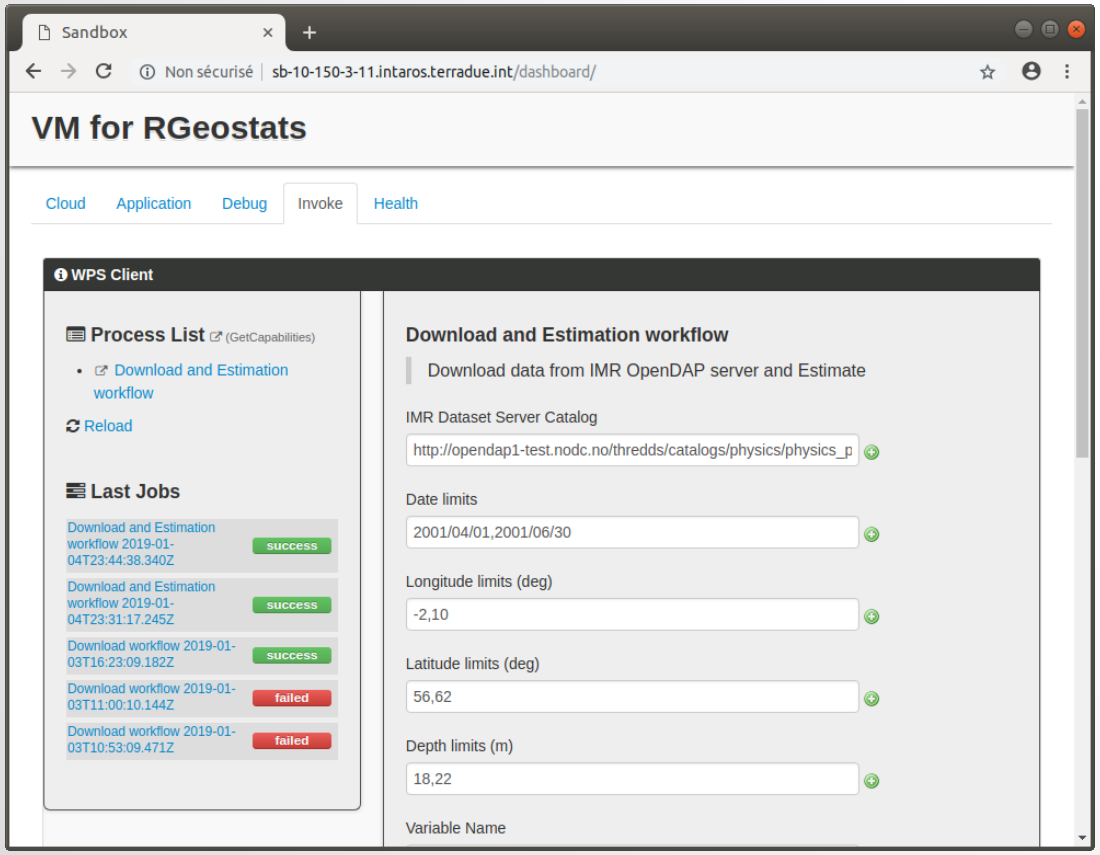
Ellip Worflow using RGeostats

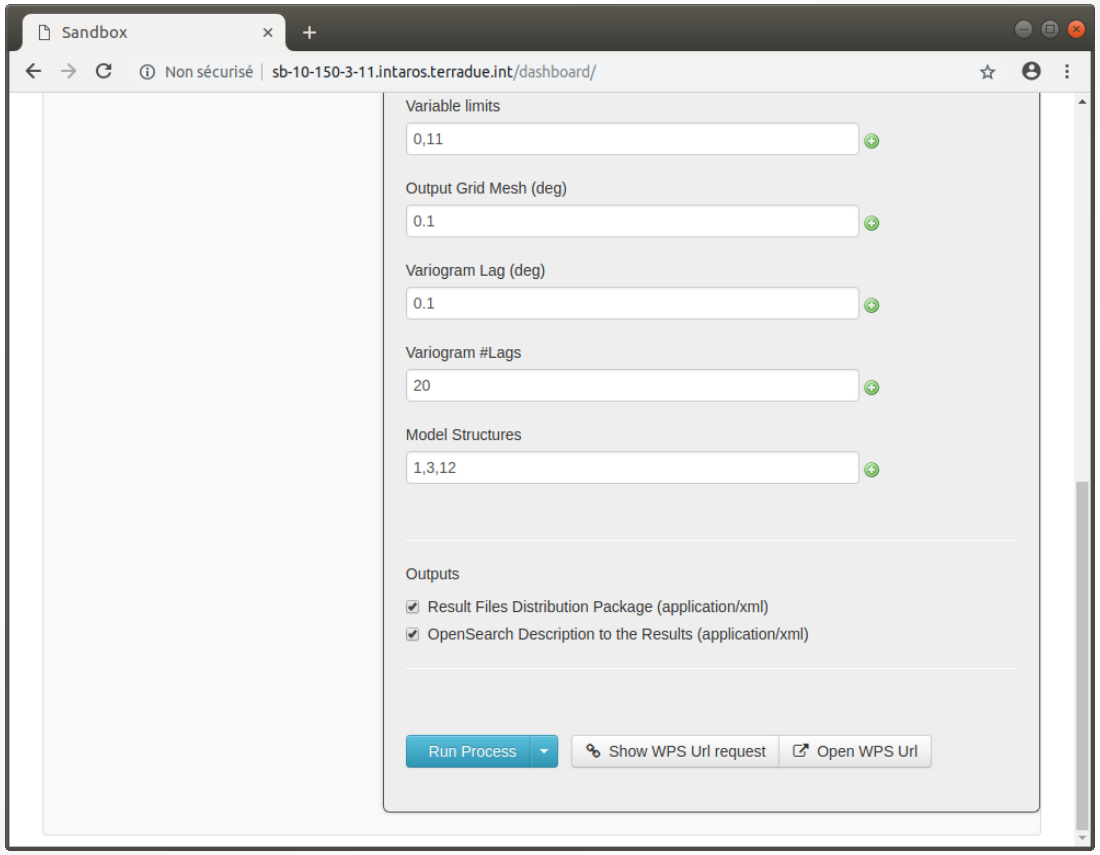
Run the workflow

```
$ ciop-run

2019-01-06 09:13:51 [INFO ] - Workflow submitted
2019-01-06 09:13:51 [INFO ] - Closing this program will not stop the job.
2019-01-06 09:13:51 [INFO ] - To kill this job type:
2019-01-06 09:13:51 [INFO ] - ciop-stop 0000065-190101000008245-oozie-oozi-W
2019-01-06 09:13:51 [INFO ] - Tracking URL:2019-01-06 09:13:51 [INFO ] - http://sb-10-
150-3-11.intaros.terradue.int:11000/oozie/?job=0000065-190101000008245-oozie-oozi-W

Node Name      :  node_download
Status         :  OK

Node Name      :  node_estimate
Status         :  OK

Publishing results...
2019-01-06 09:19:52 [INFO ] - Workflow completed.
2019-01-06 09:19:52 [INFO ] - Output Metalink: http://sb-10-150-3-
11.intaros.terradue.int:50070/webhdfs/v1/ciop/run/download_and_estimation_workflow/000
0065-190101000008245-oozie-oozi-W/results.metalink?op=OPEN
```

Ellip Worflow using RGeostats

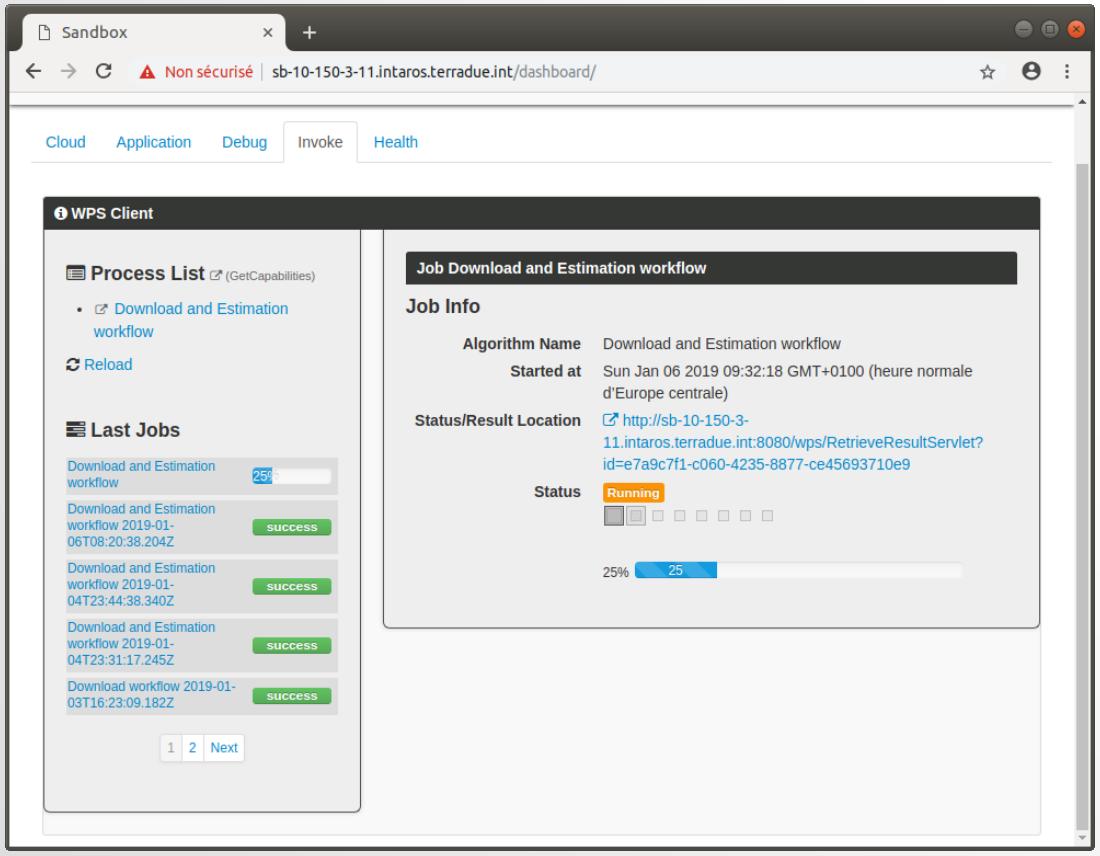
Invoke the workflow (1/2)

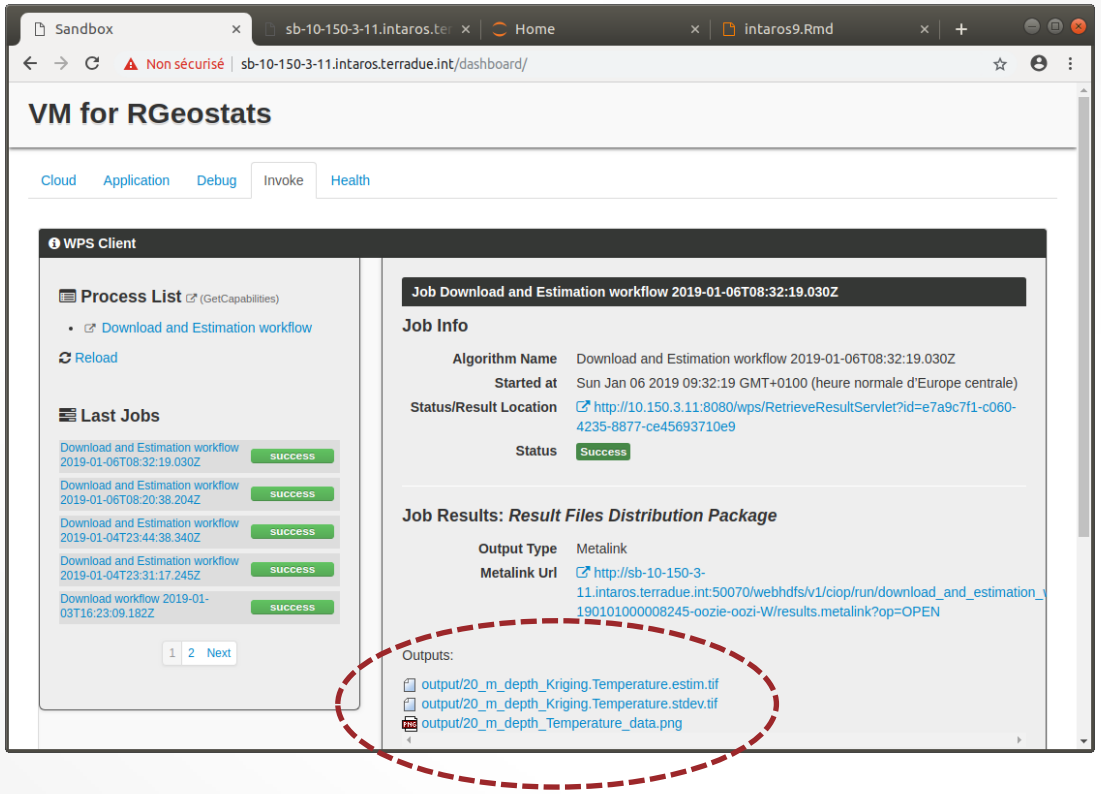




Ellip Worflow using RGeostats

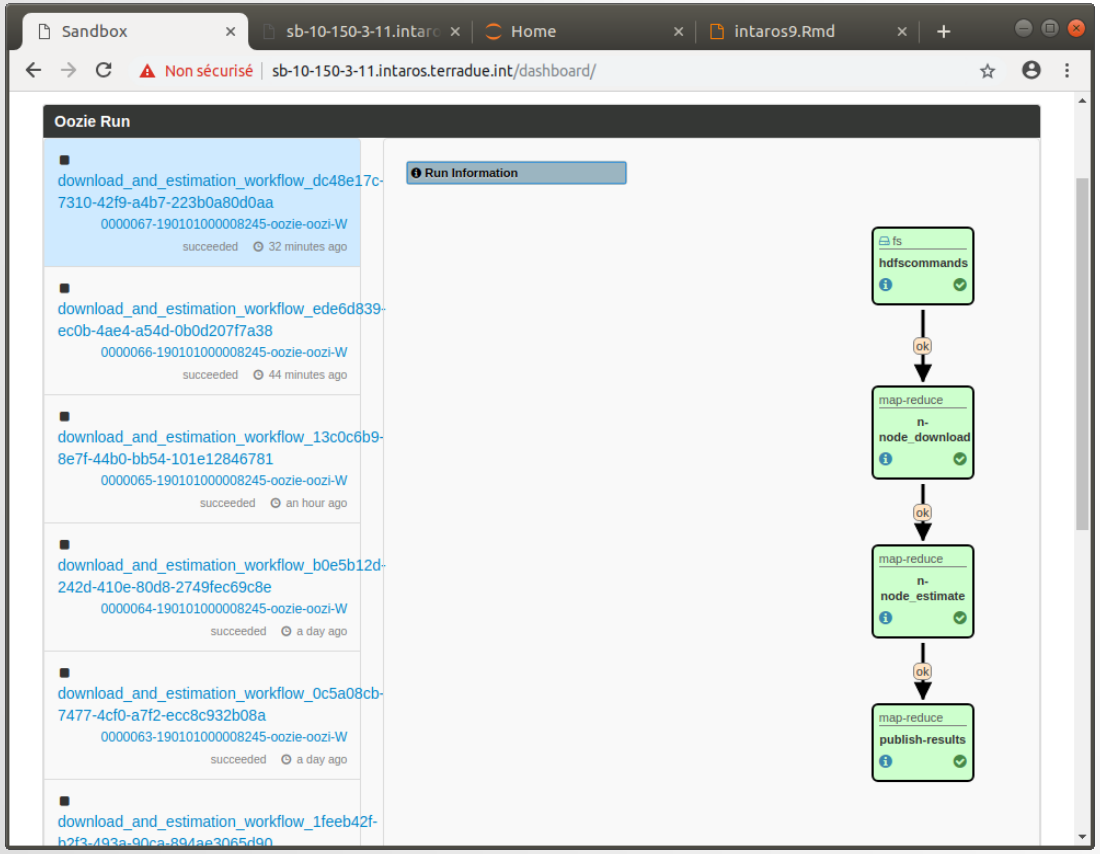
Invoke the workflow (2/2)

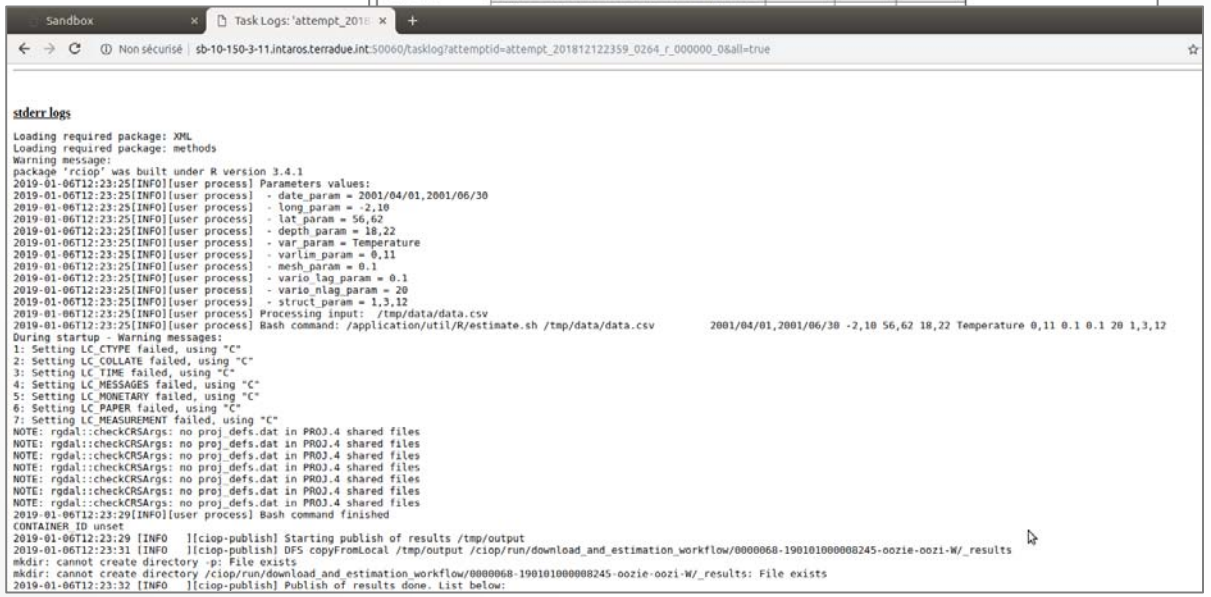




Ellip Worflow using RGeostats

Debug the workflow

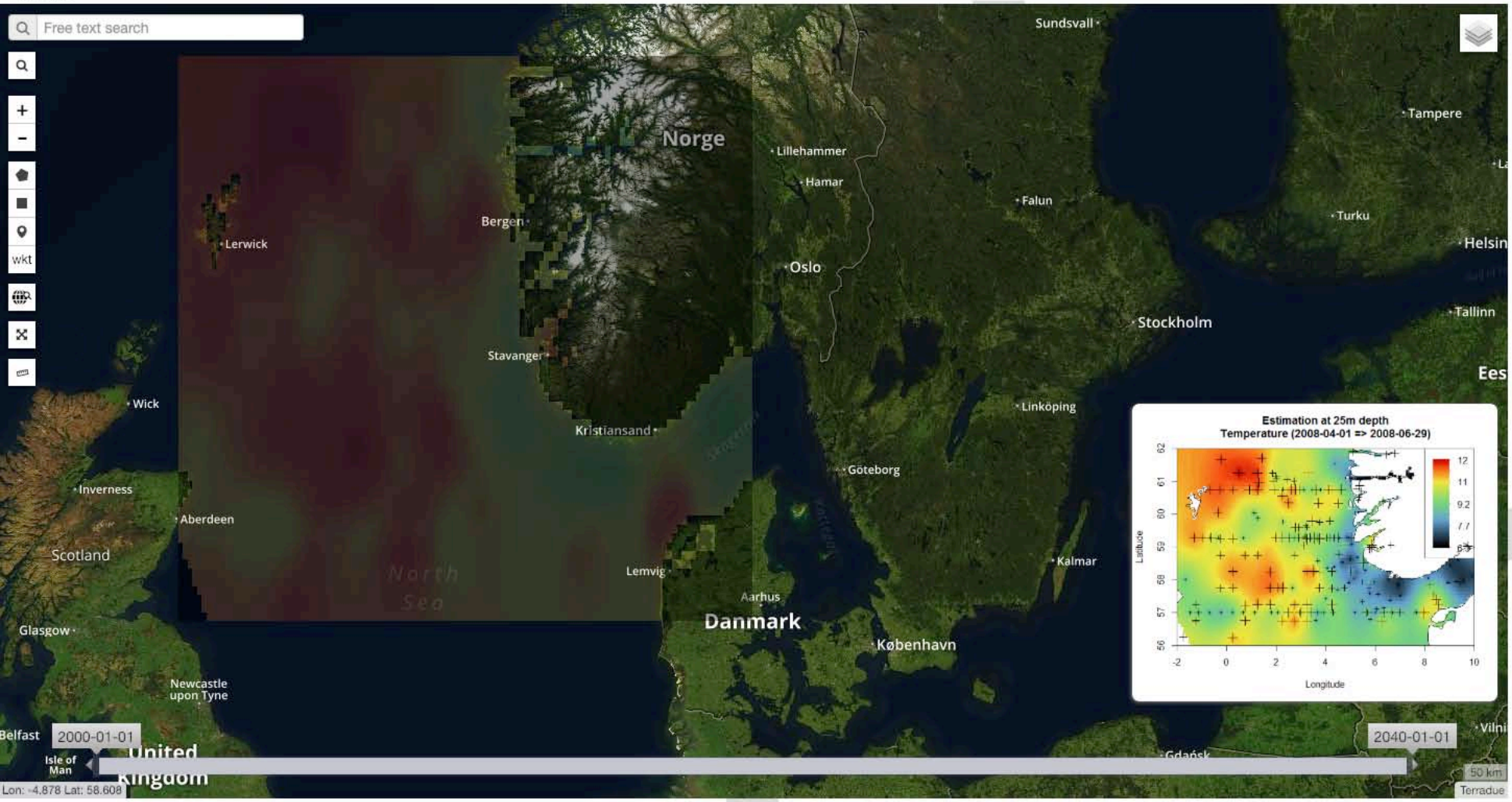




Ellip Worflow using RGeostats

Deploy the workflow

```
$ ciop-release

***
*** Application release for 'dcs-imr-estim'
***


...
...

Writing objects: 100% (1/1), 159 bytes, done.
Total 1 (delta 0), reused 0 (delta 0)
To https://gitlab.com/ec-intaros/dcs-imr-estim.git
 * [new tag]         1.0 -> 1.0

***
-> Stage 1. Checking your repository 'dcs-imr-estim' DONE
-> Stage 2. Setting the release and development versions DONE
-> Stage 3. Releasing 'dcs-imr-estim' DONE
-> Stage 4. Syncronising the remote repository DONE
```



Current search result

Discovery feed for local data — Total results 2

- 20_m_depth_Kriging_Temperature_estim.tif
- 20_m_depth_Kriging_Temperature_stdev.tif

Features Basket | Data Packages

Total results 0 | sel.all | inv.sel. | Remove all | Save

No results found.

Id	2e773f7d-6d91-4dd4-90c4-18a3778afc20
Processing service	Download and Estimation workflow
Started at	Jan 9th 2019 10:17
Finished at	Jan 9th 2019 10:24
Created by	crossi
Status/Result Location	
Status	Success
Visibility	restricted
Share	

Parameters

Name	Value
source	http://opendap1-test.nodc.no/thredds/catalogs/physics/physics_point_yearly.xml
date_lim	2001/04/01,2001/06/30
long_lim	-2,10
lat_lim	56,62
depth_lim	18,22
var	Temperature
var_lim	0,11
mesh	0.1
vario_lag	0.1
vario_nlag	20
struct	1,3,12

Resubmit Job

Success
The job was completed successfully.

Results

Found layers in the result. Show results

XML Result

Support

Ellip Worflow using RGeostats

Useful commands (1/3)

Connect to your VM

```
sudo openvpn --config client.ovpn
ssh 10.150.3.11
```

=> Connect to VPN
=> Connect to your VM with no password (ssh key pairs)

Creating and installing application

```
mvn archetype:generate
mvn clean install
```

=> Create a new workflow
=> Install the application

Ellip Worflow using RGeostats

Useful commands (2/3)

Local git repository

```
git init                                  => Initialize a local repository
git config credential.helper store        => Store credentials (no more password)
git add –A .                              => Add content to a local repository
git commit –m <msg>                       => Commit changes to the local repo
git branch <branch>                       => Create a branch named <branch>
git checkout <branch>                     => Work on the branch named <branch>
```

Remote git repository interaction

```
git remote add origin <url>               => Synchronize with remote repository
git push                                  => Push local changes to server
git pull                                  => Update local repo from the server
git clone                                 => Create a local repo from the server
```

Ellip Worflow using RGeostats

Useful commands (3/3)

Ciop tool

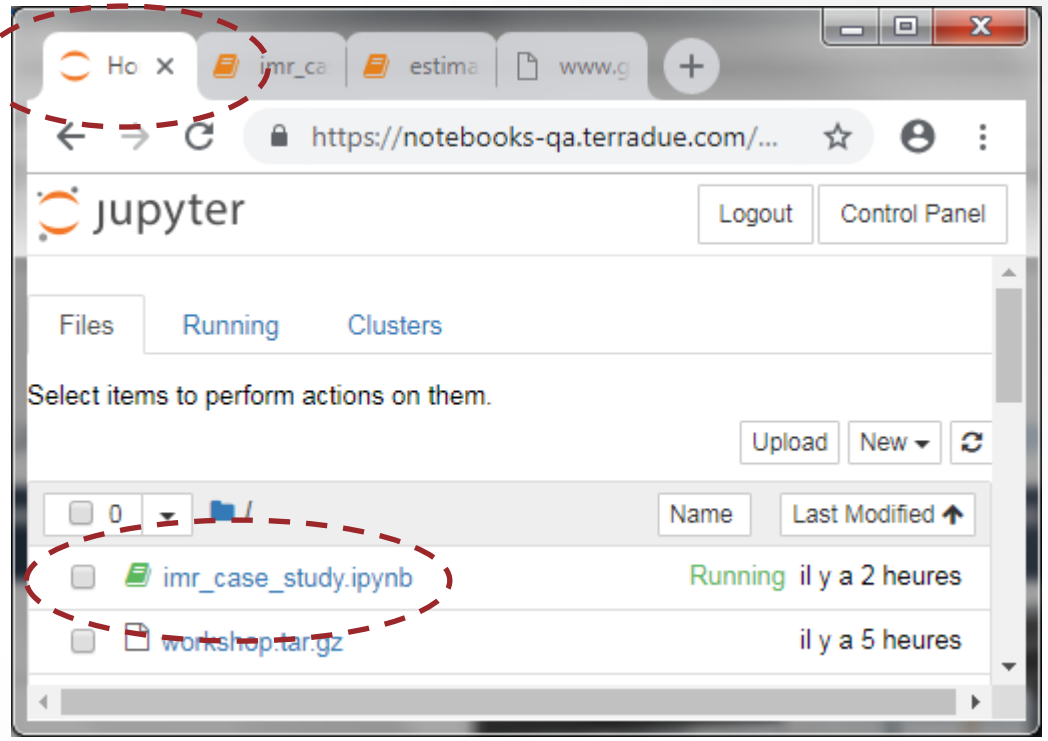
`ciop-getparam`	=> Retrieve job's parameter value (user input)
`ciop-log`	=> Log something to job log files
`ciop-copy`	=> Copy online source file to local file system
`ciop-publish`	=> Publish job productions as output results
`ciop-run`	=> Execute a job or a complete workflow
`ciop-release`	=> Create and deploy a new application release

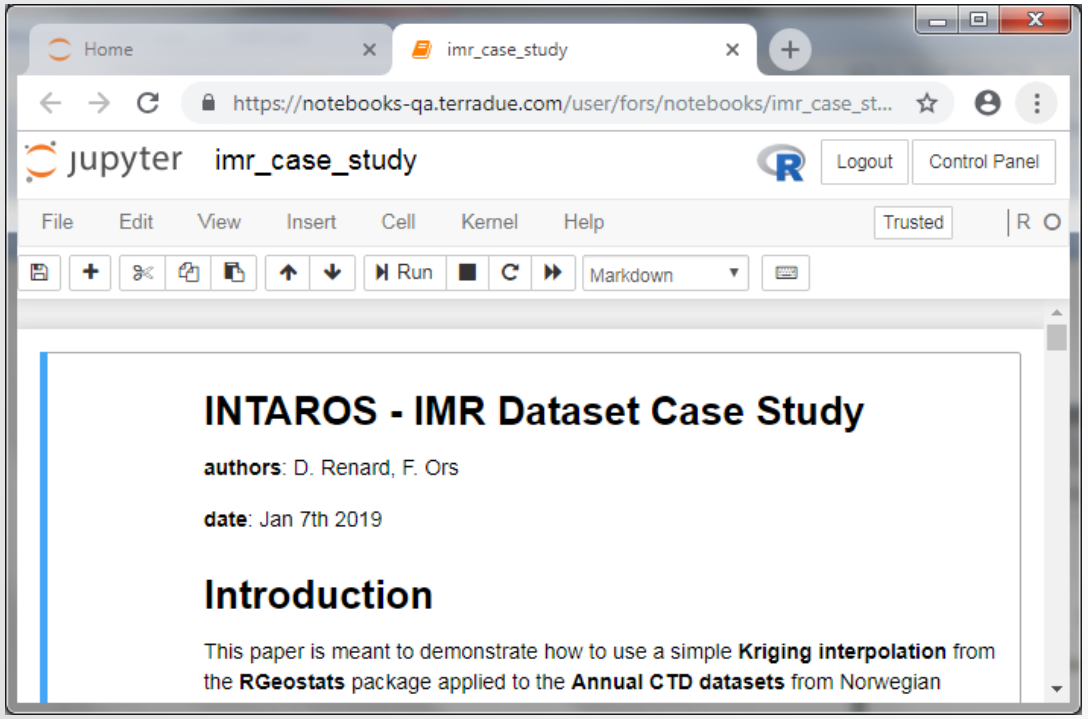
Part #5

IMR Case Study – RGeostats in Action!

Getting Ready for Action!

- Click on the **Home** tab of your Jupyter Notebooks (in Chrome)
- Click on the **imr_case_study.ipynb**





Note: You may prefer the 'lab view': replace **tree** by **lab** in URL

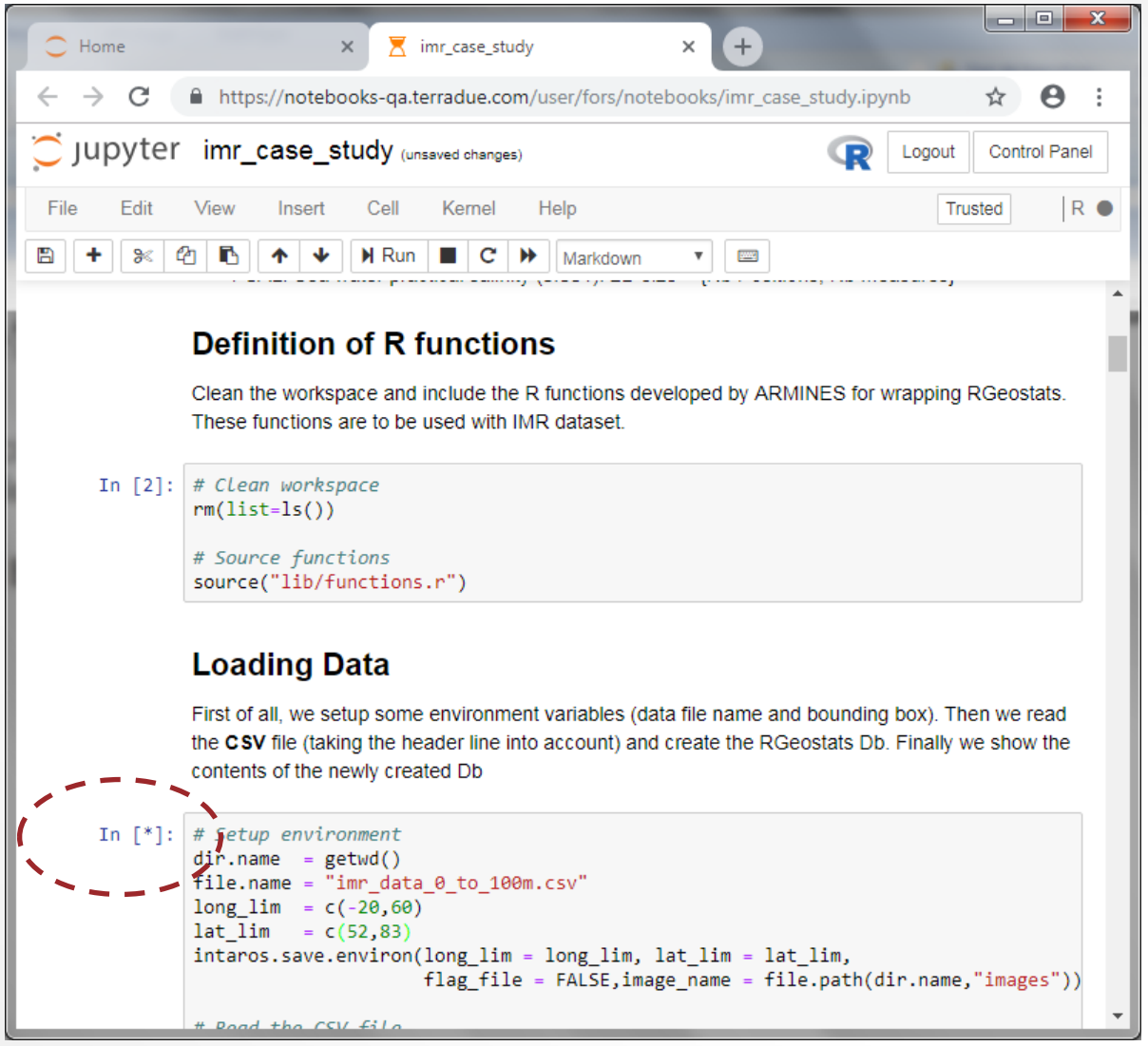
Getting Ready for Action!

- Execute all code cells in the **Introduction** section:

  => Hitting **Shift + Enter**

  - Definition of R functions
  - Loading Data (slow)
  - Dataset global statistics (2 cells)

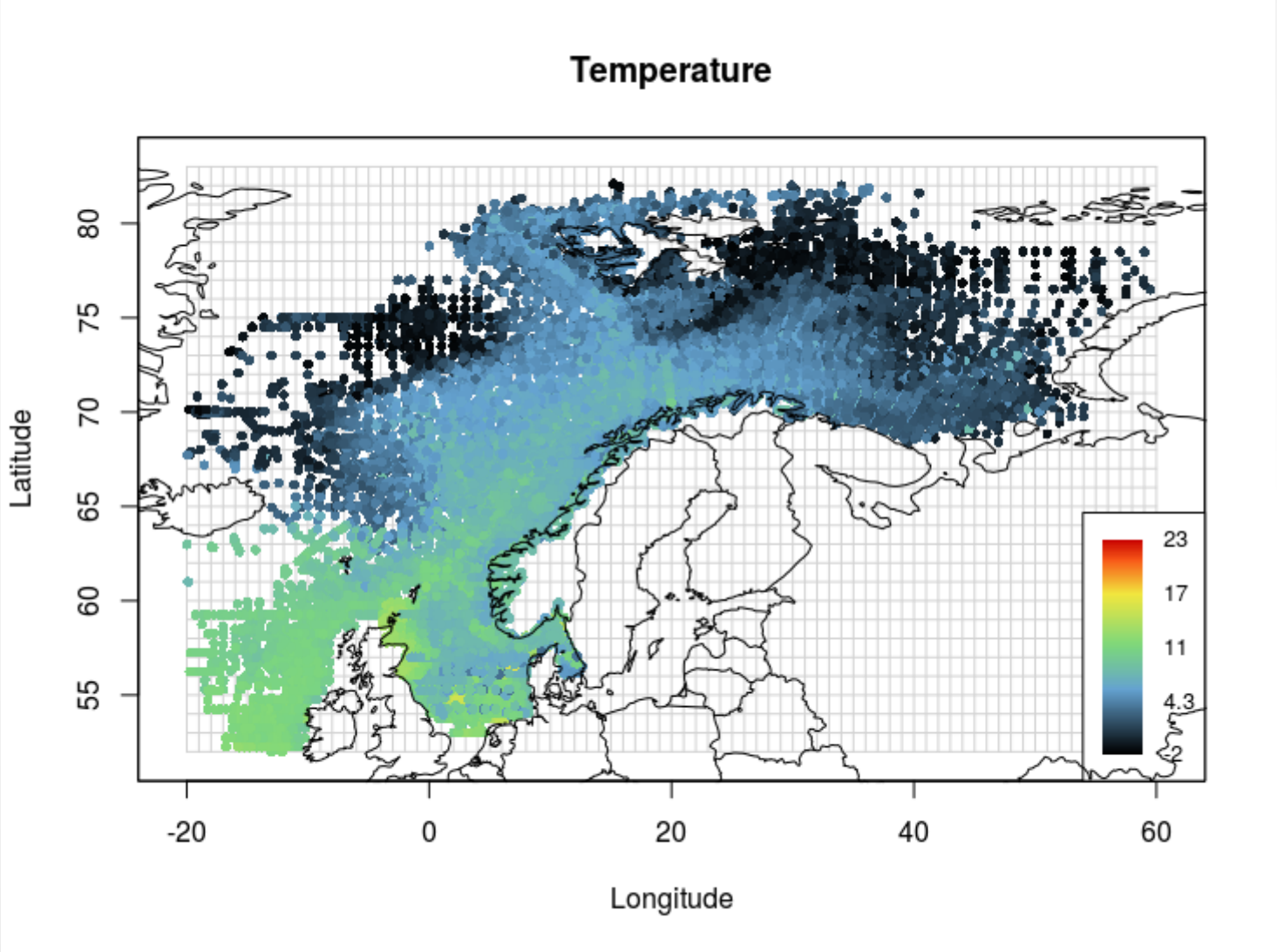
Note: The symbol **In [*]** indicates that a cell is running



Studying Temperature variable

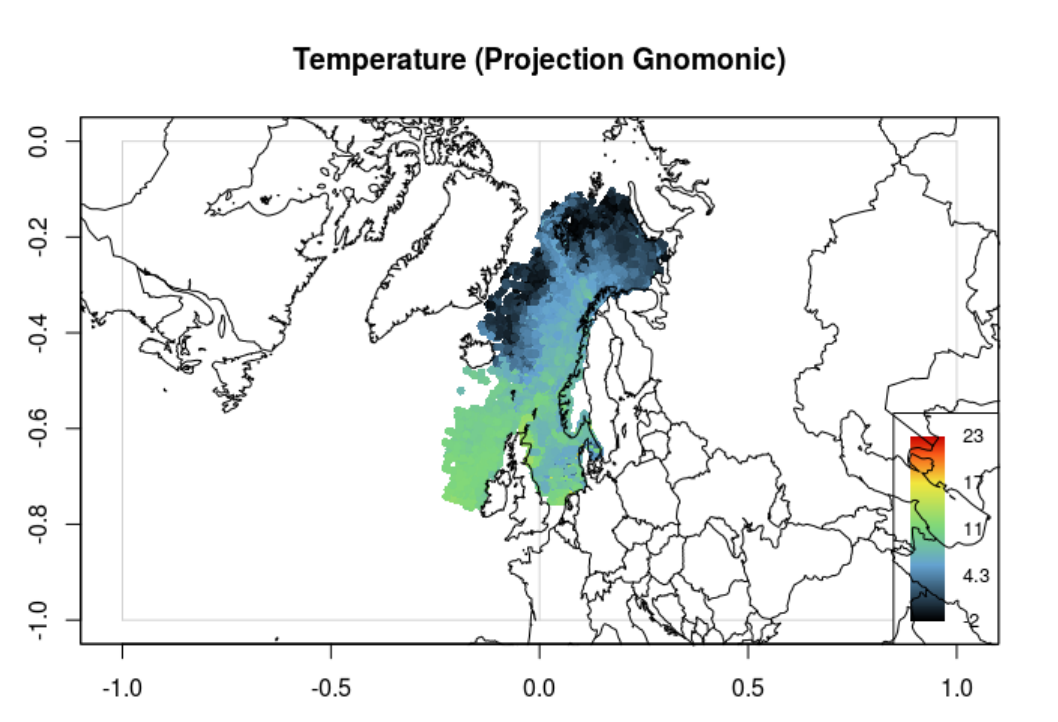
All Database

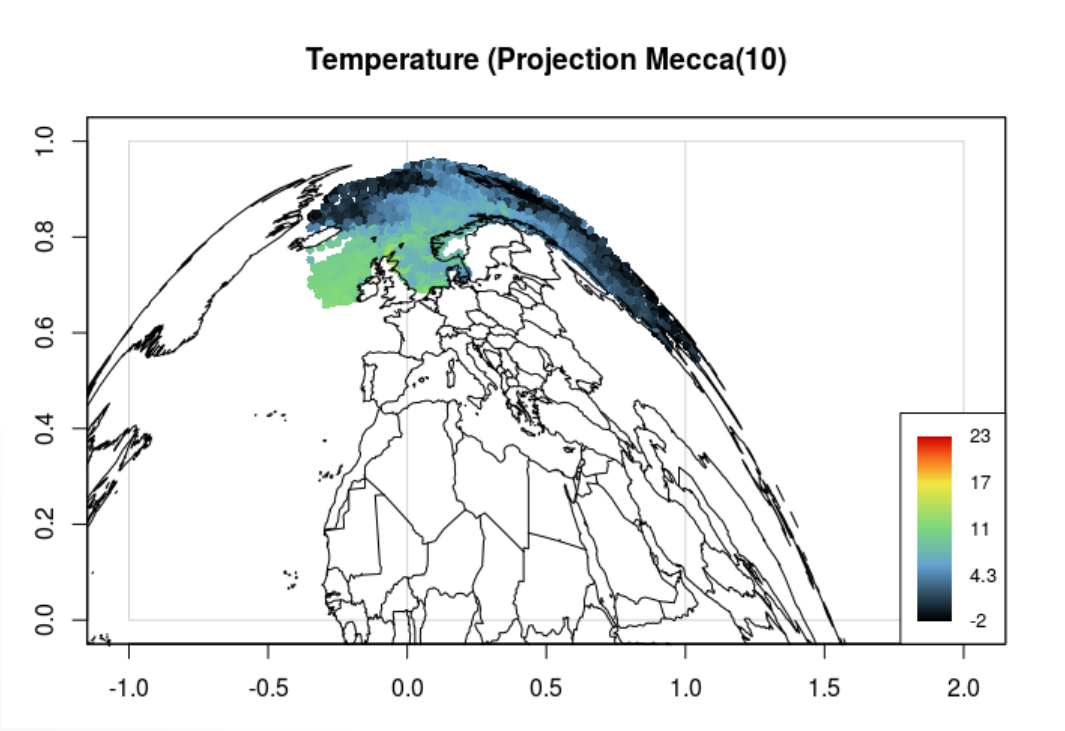
- Display Temperature Variable
- 5 million samples displayed



Different Projections

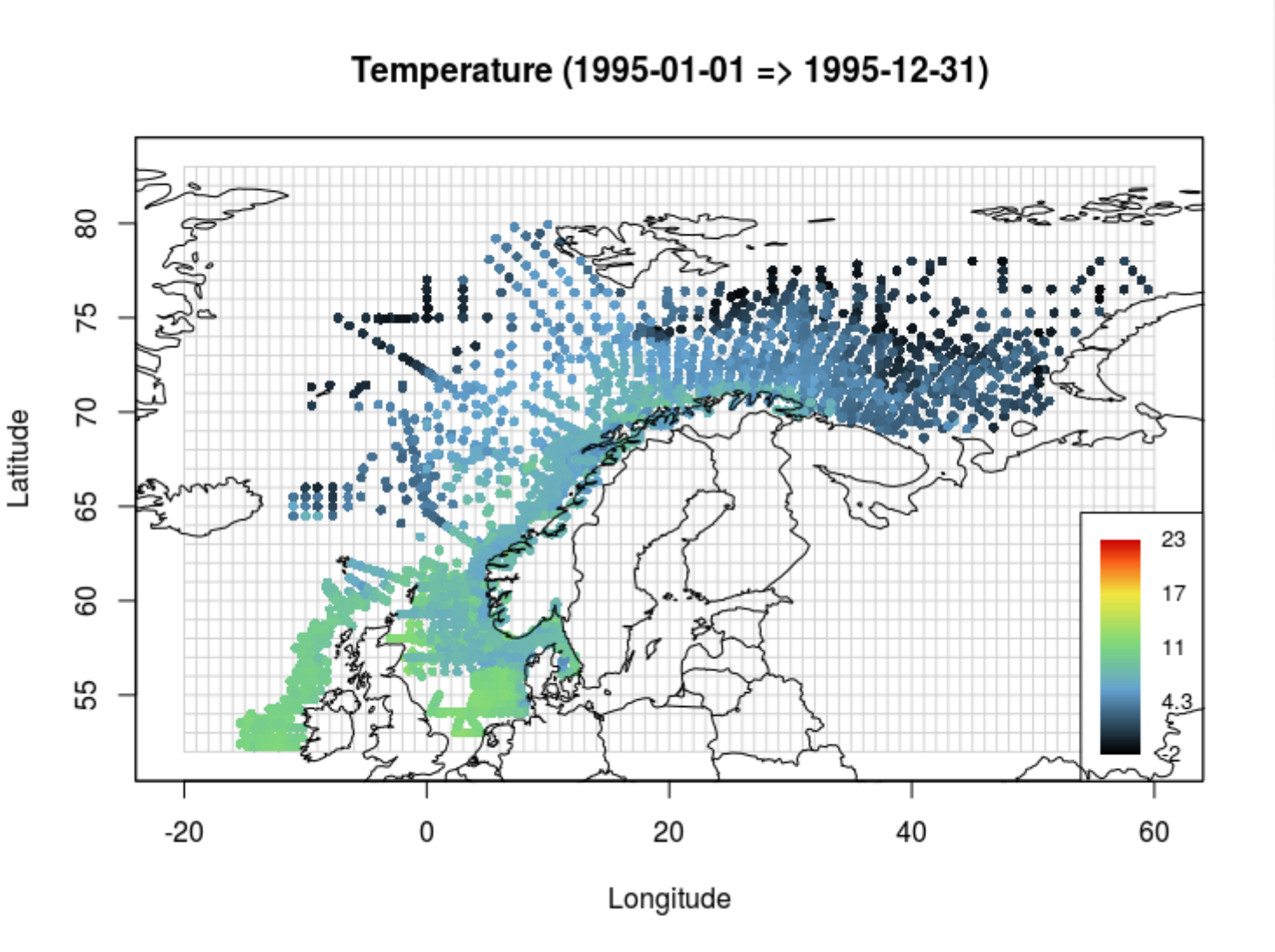
- Define New Projections (Gnomonic and Mecca)





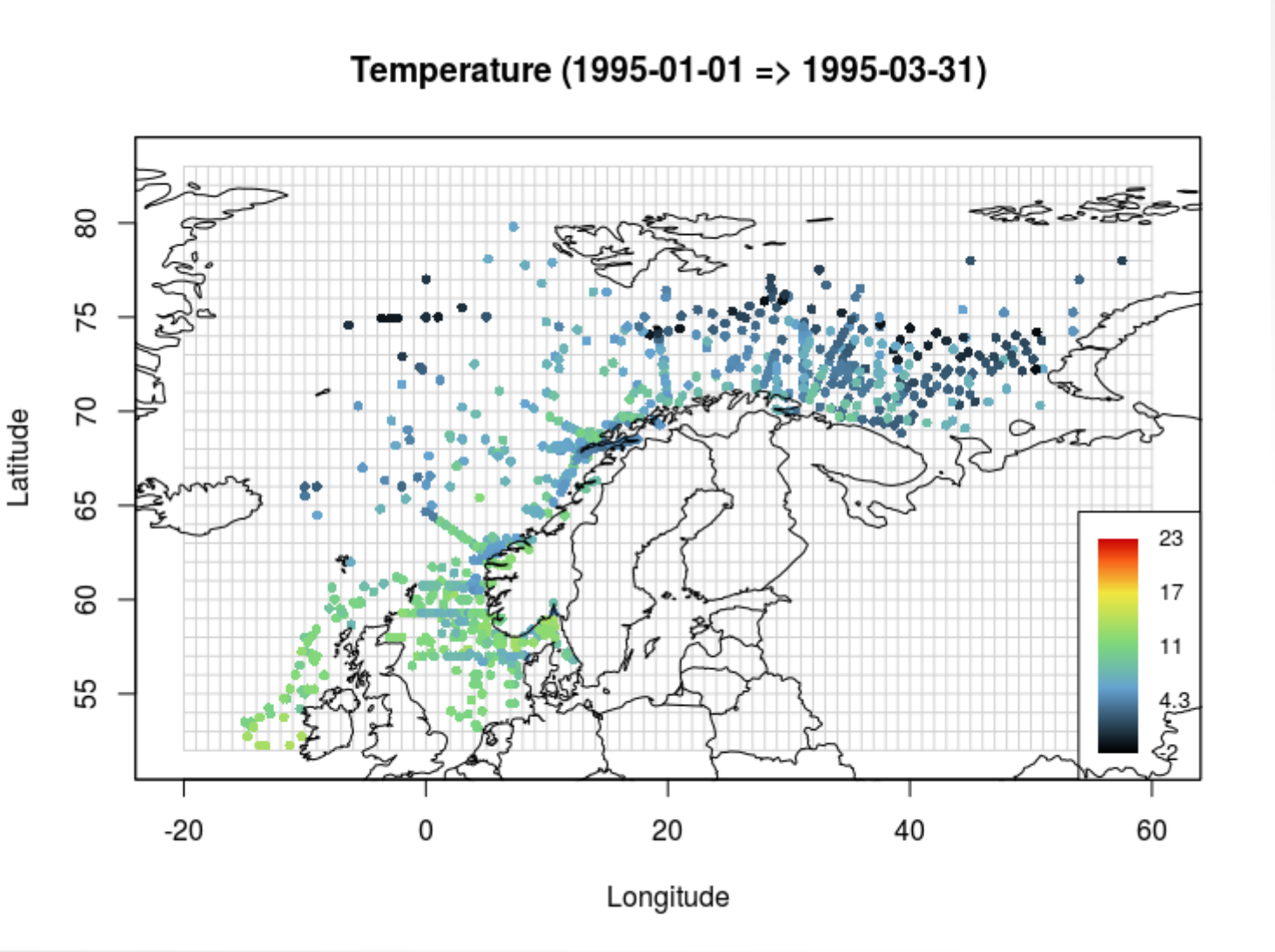
Year Campaign

- Define Target Year (1995)
- Display Variable



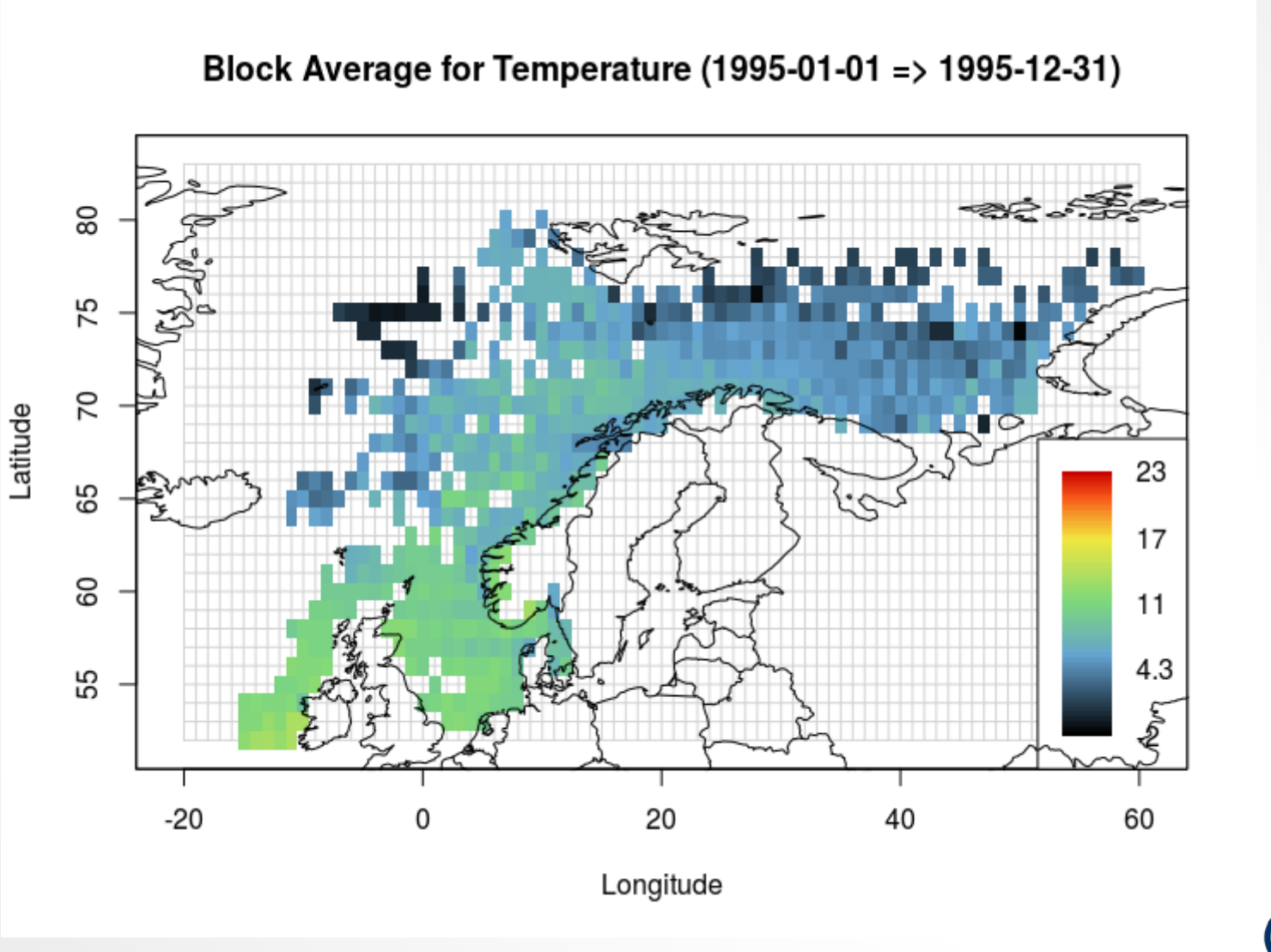
Year/Trimester at 20m Depth

- Select Depth Interval (20m)
- Define Target Year (1995)
- Define Target Trimester (1st)
- Display Variable



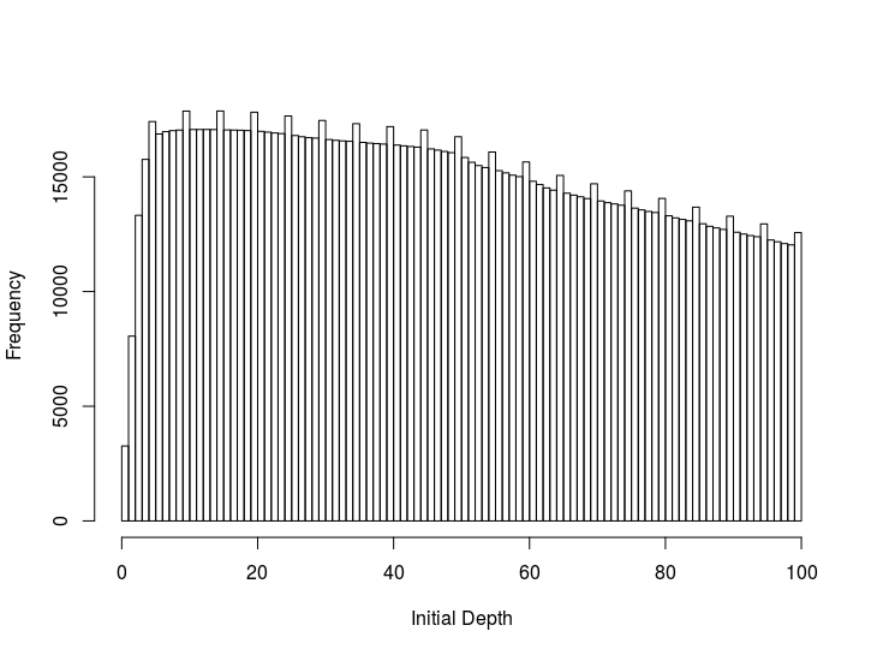
Block Average at 20m Depth

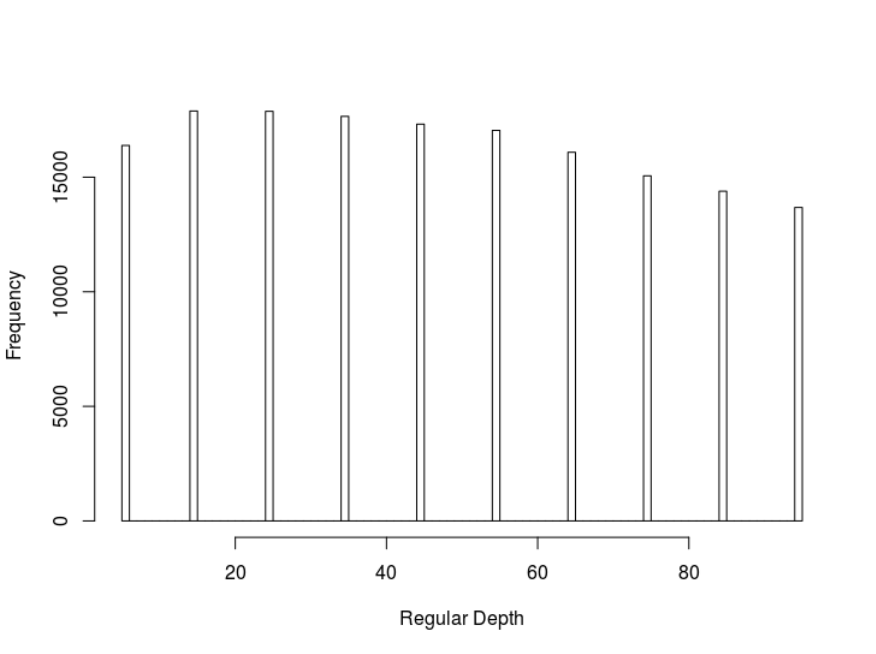
- Select Depth Interval (20m)
- Define Target Year (1995)
- Average per Cell (1 degree)
- Display Variable



Histogram of Measurement Depths

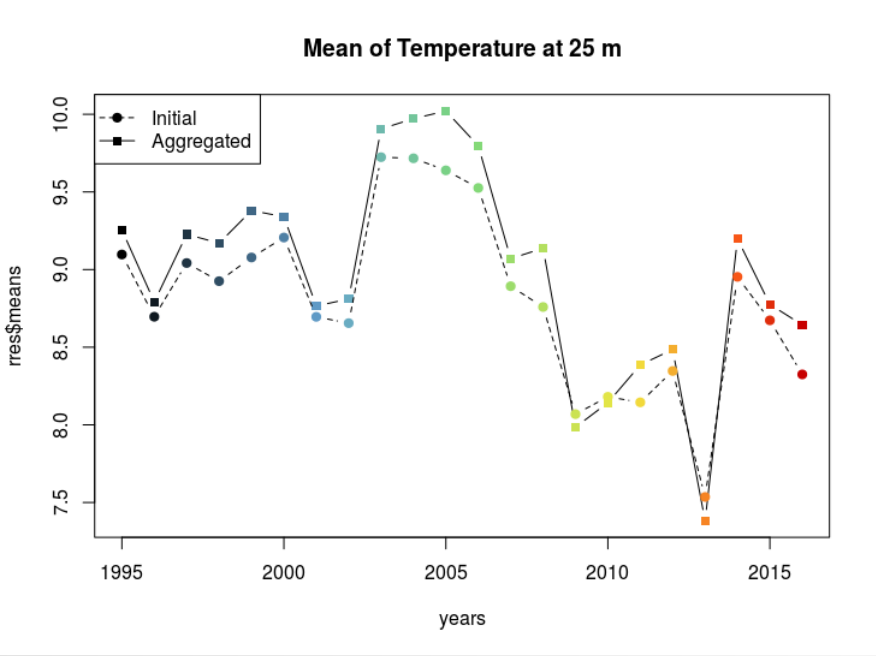
- Focus on a Smaller Area (South West of Norway)
- Aggregate per Depth Intervals (10m)
- Histogram of Depths (before and after Aggregation)

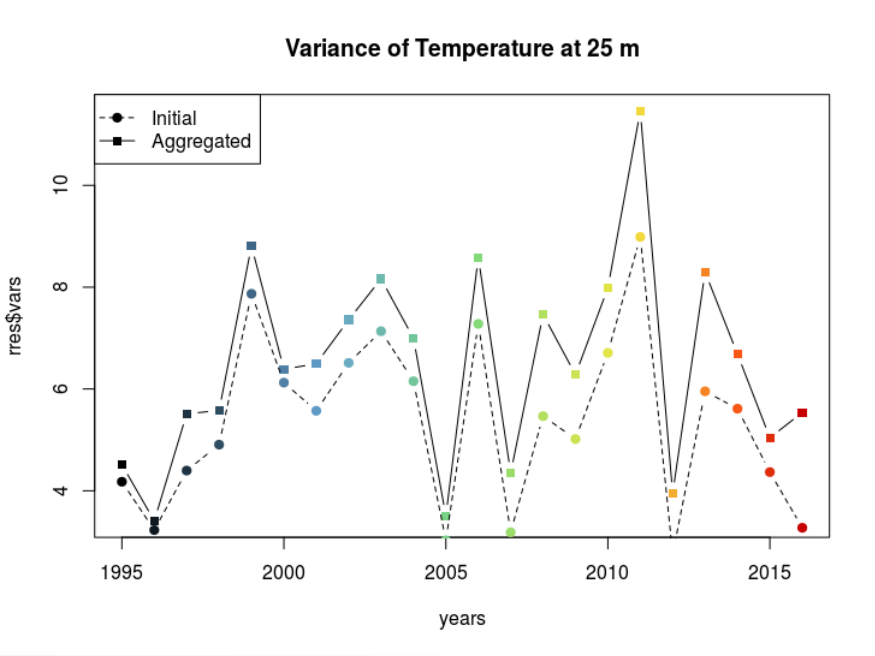




Statistics per Year

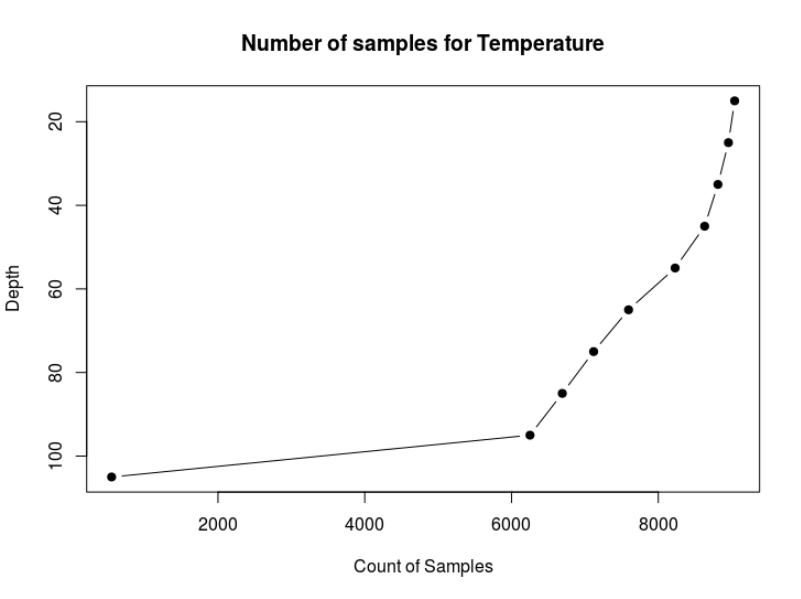
- Select Depth Interval (25m)
- Aggregate per Time Intervals (1 year)
- Mean and Variance of Temperature (before and after Aggregation)

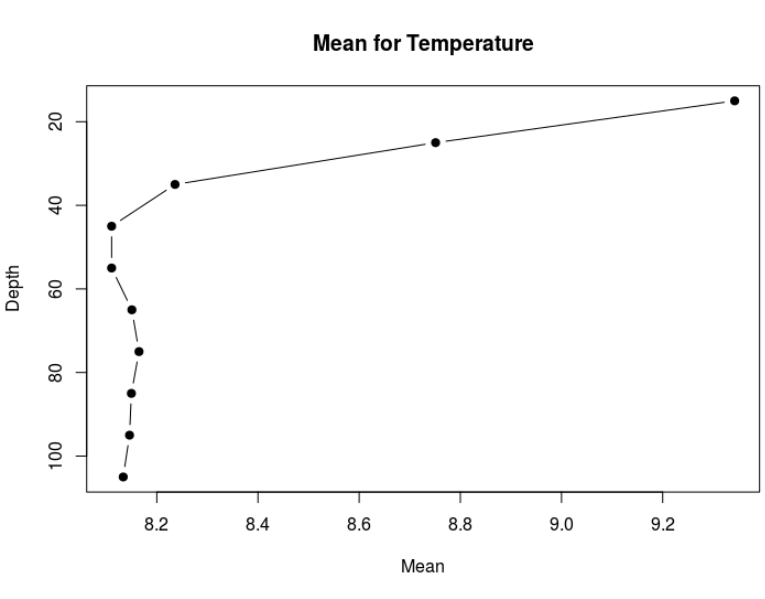


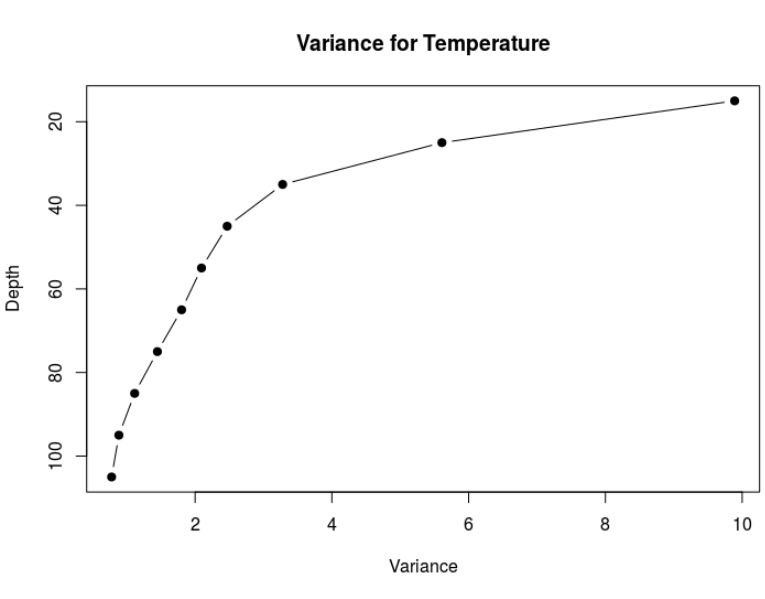


Statistics per Depth

- Select Year (2008)
- Average per Depth Interval (10m)
- Count, Mean and Variance of Temperature per Depth

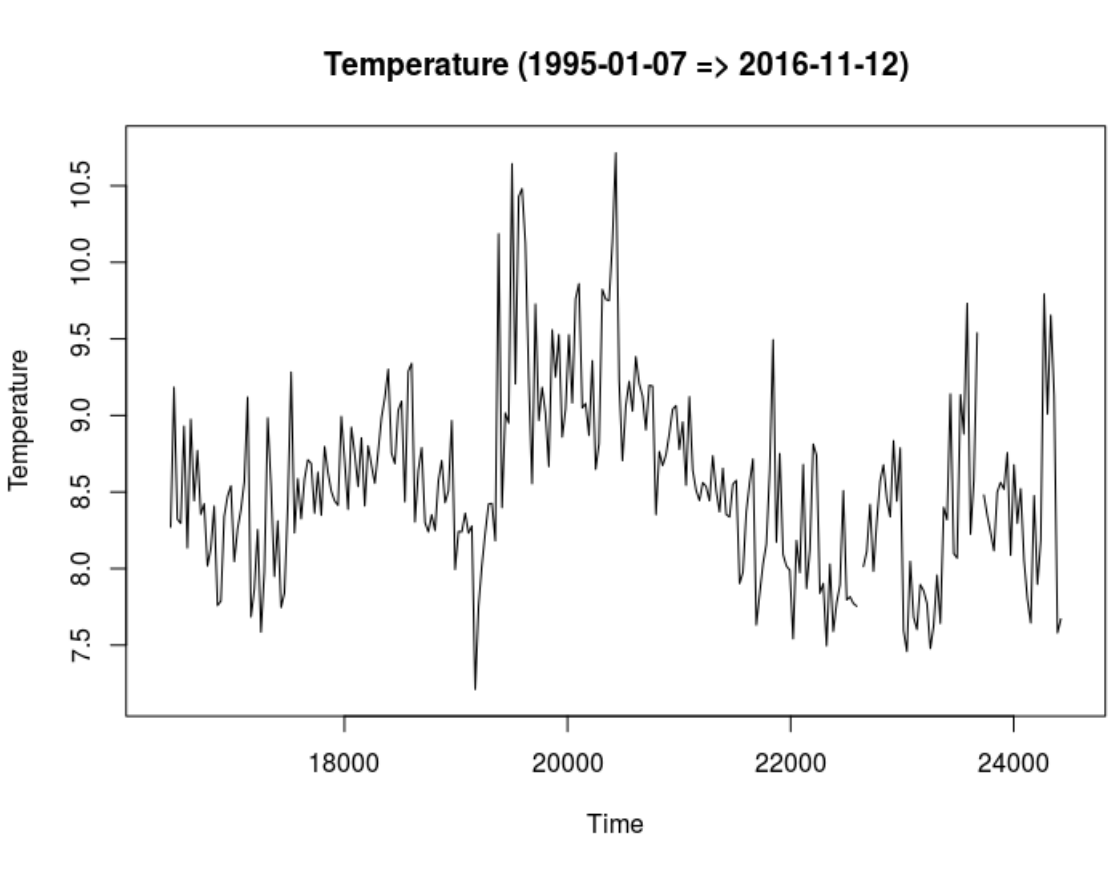






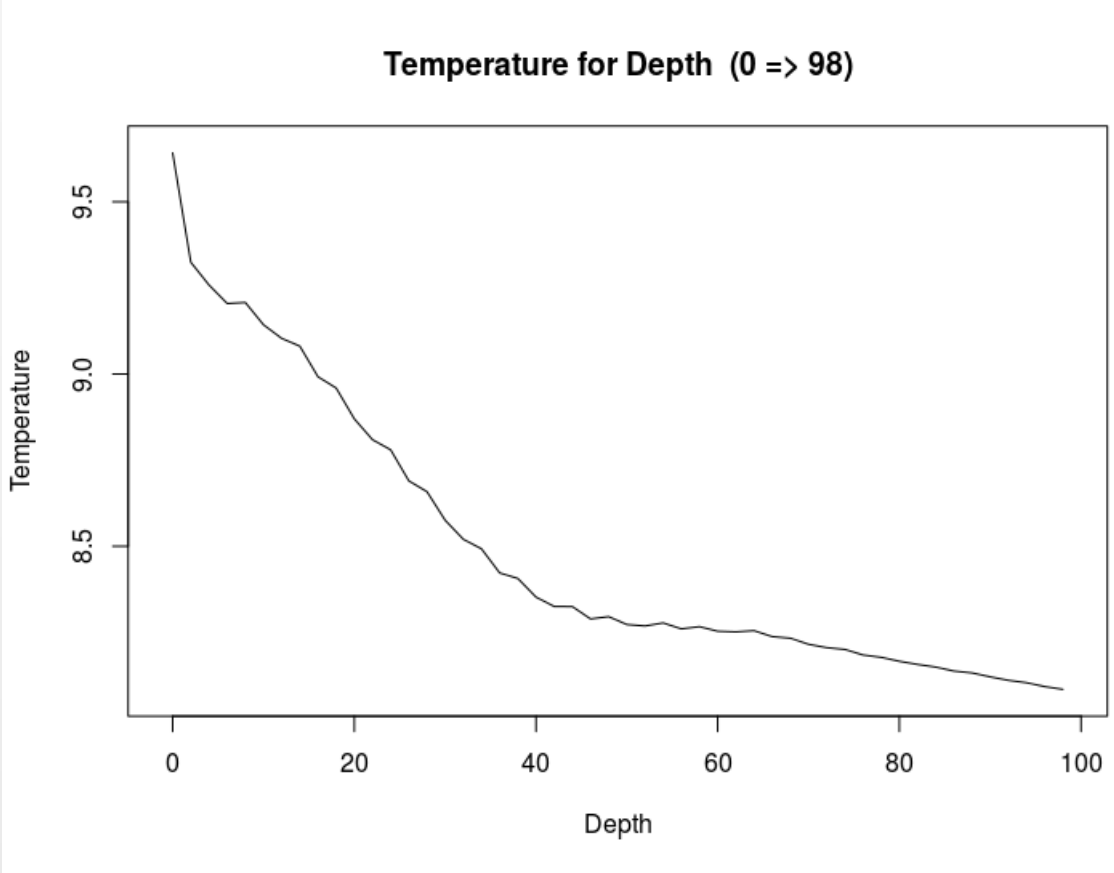
Variation along Time

- Aggregate by Time Intervals (30 days)



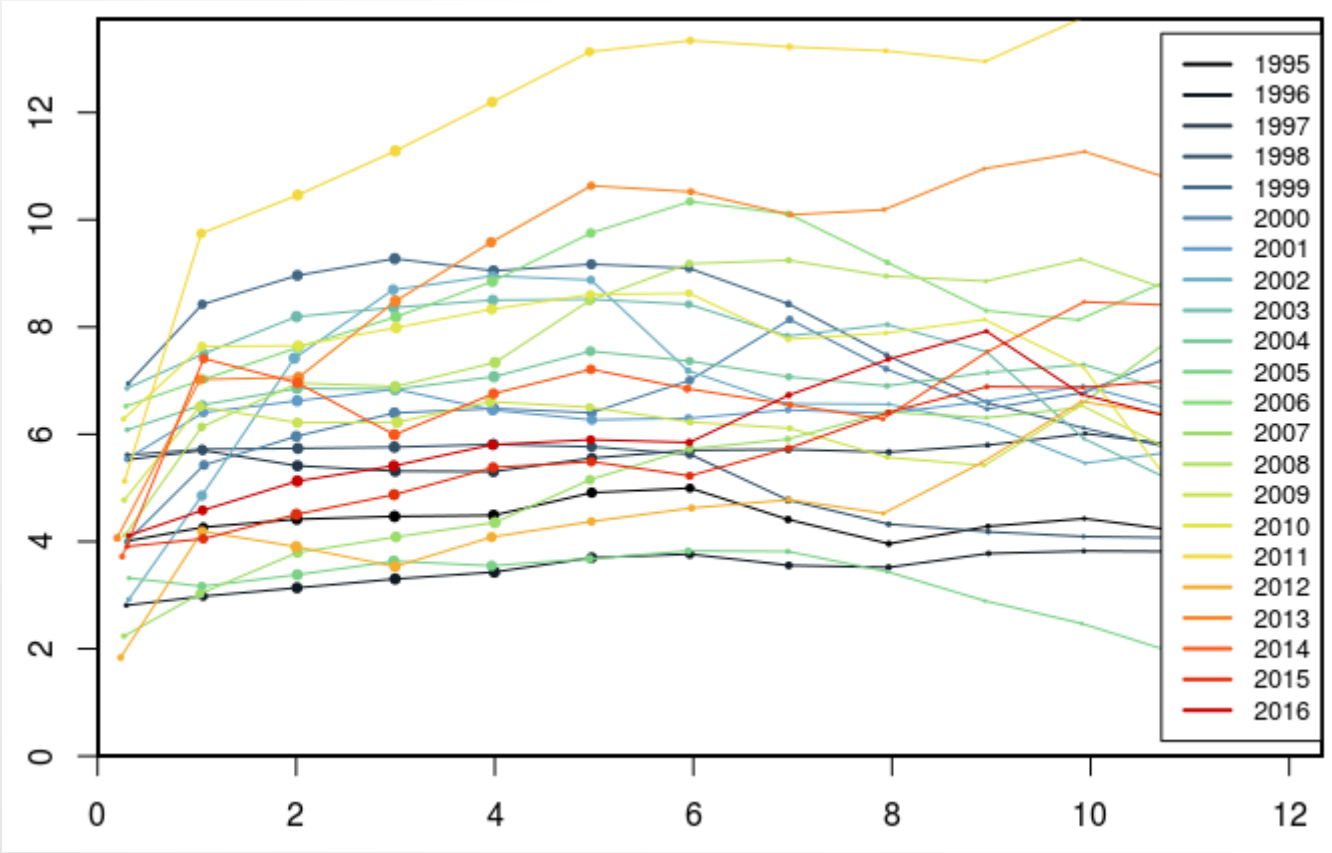
Variation along Depth

- Aggregate by Depth Intervals (2m)



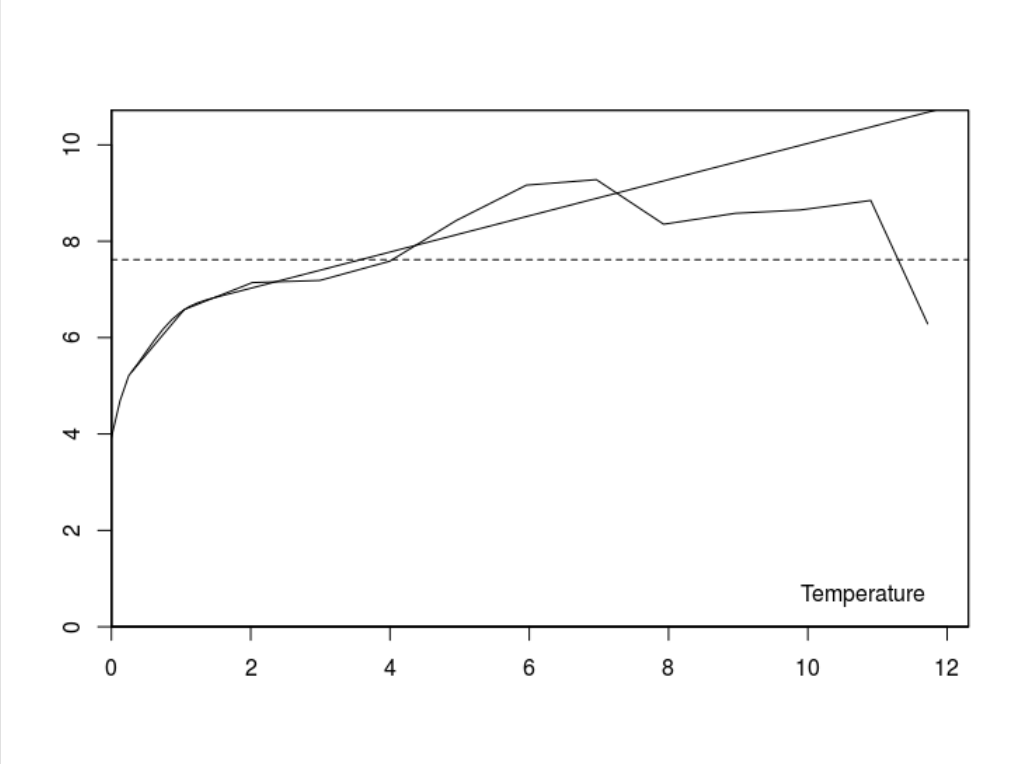
Horizontal Variogram per Year

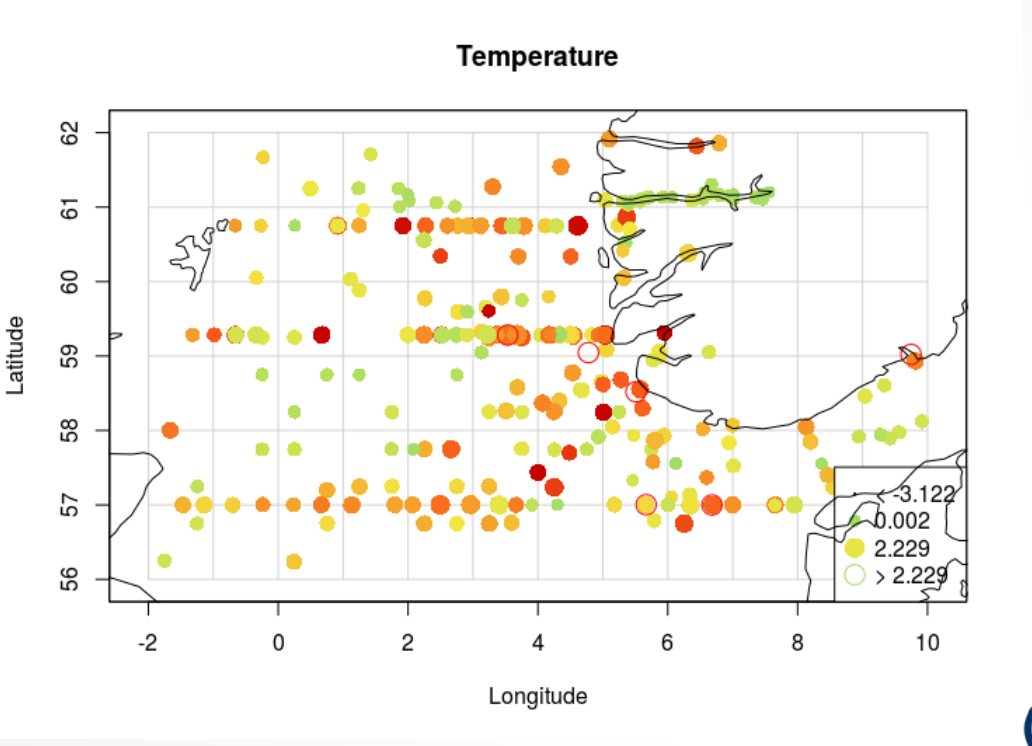
- Select Depth Interval (25m)
- Omni-directional Variogram per Year for Temperature



Cross-Validation

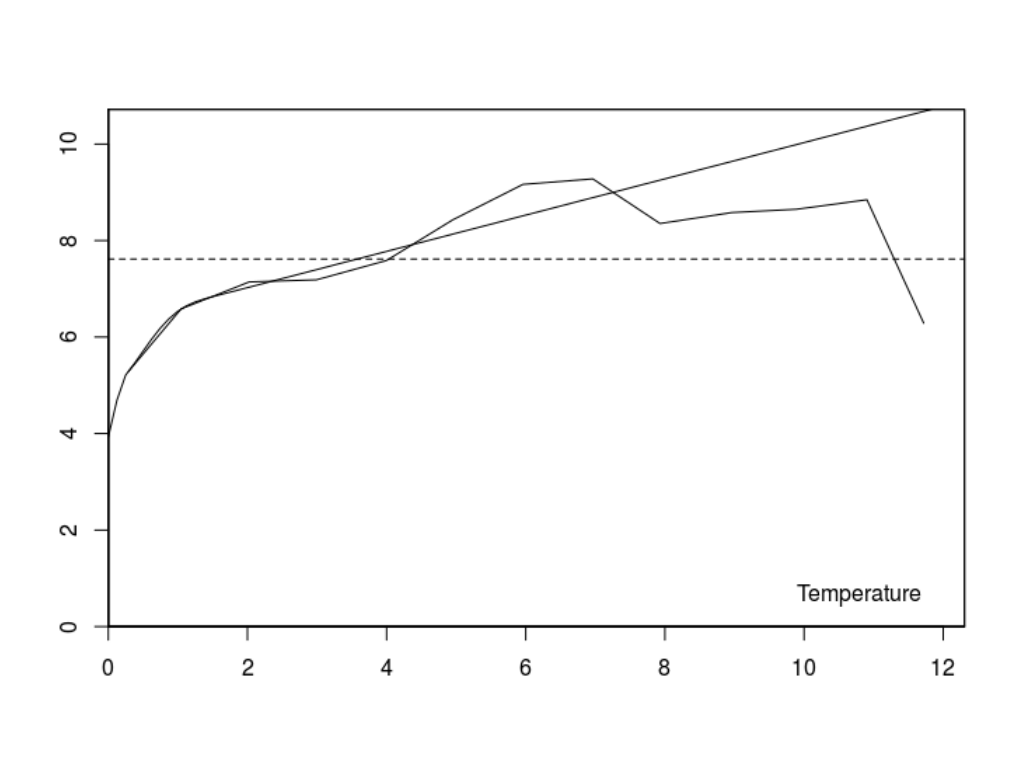
- **From now**: use Aggregated database per Depth Intervals (10m)
- Calculate Average Horizontal Variogram and Fit the Model of Temperature
- Perform the Cross-Validation at Data Samples

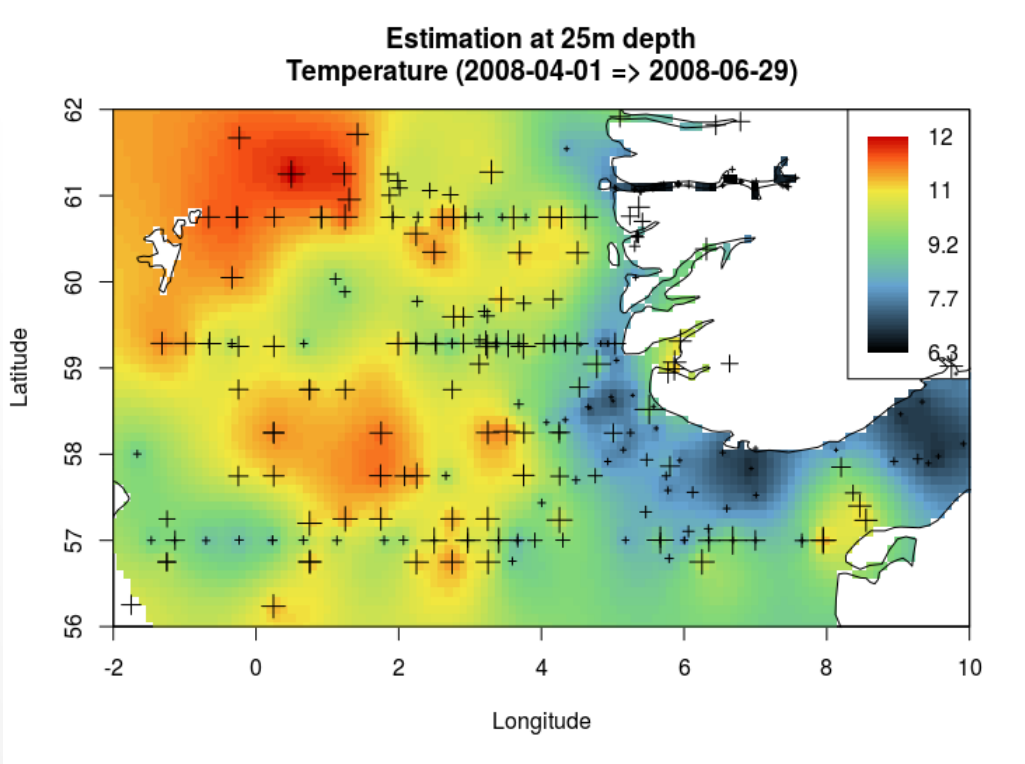




2-D Estimation of Temperature

- Select Depth Interval (25m) and Focus on 2nd Trimester of Year 2008
- Calculate Average Horizontal Variogram and Fit the Model for Temperature
- Estimate the Temperature in 2-D in South-West of Norway

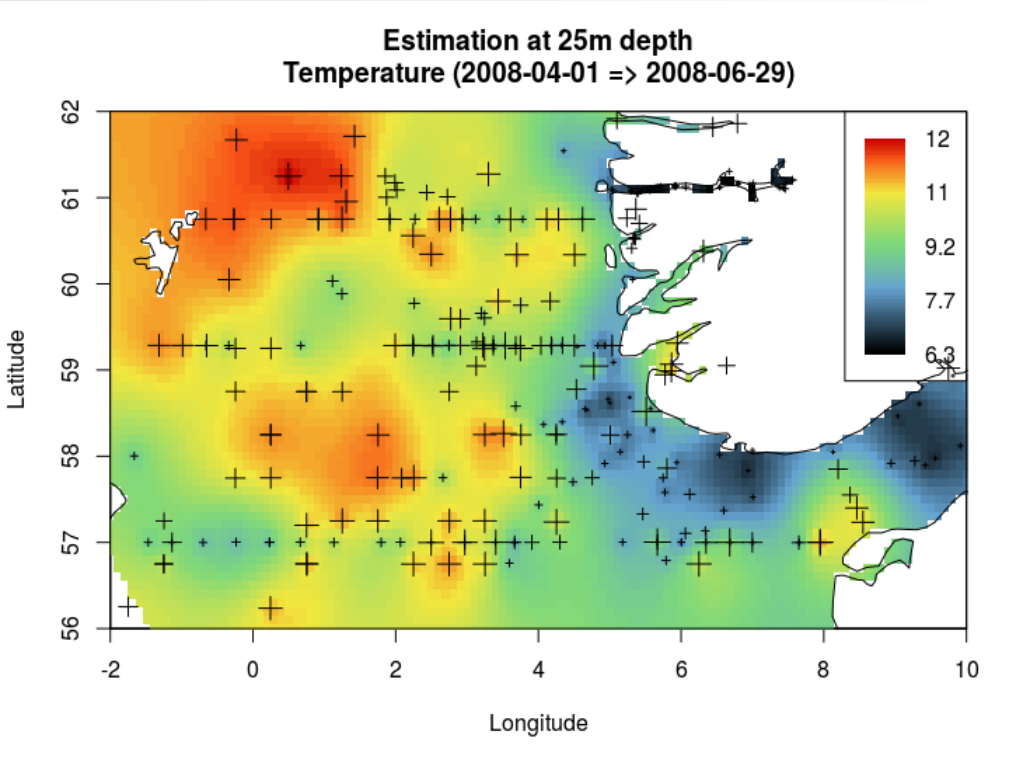


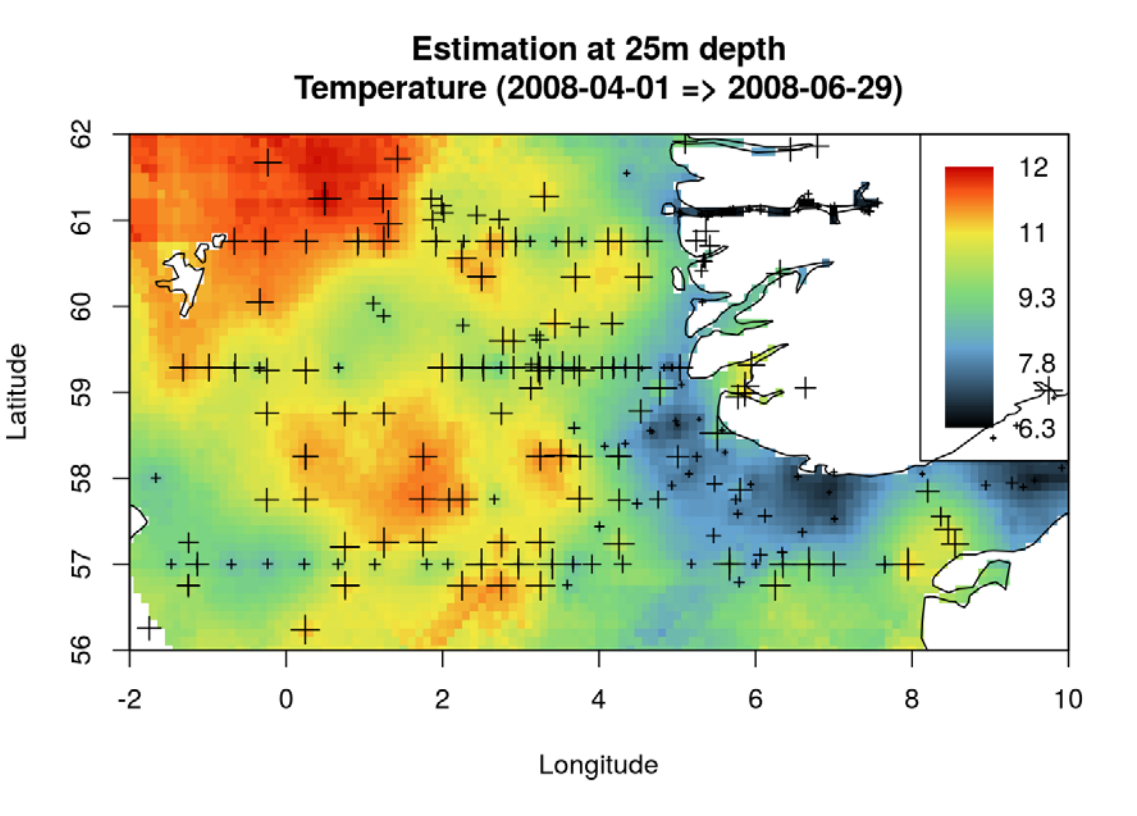


Moving Neighborhood

Comparing Unique (left) and Moving (right) Neighborhoods

- 2-D Moving Neighborhood: Maximum Number of Samples = 40

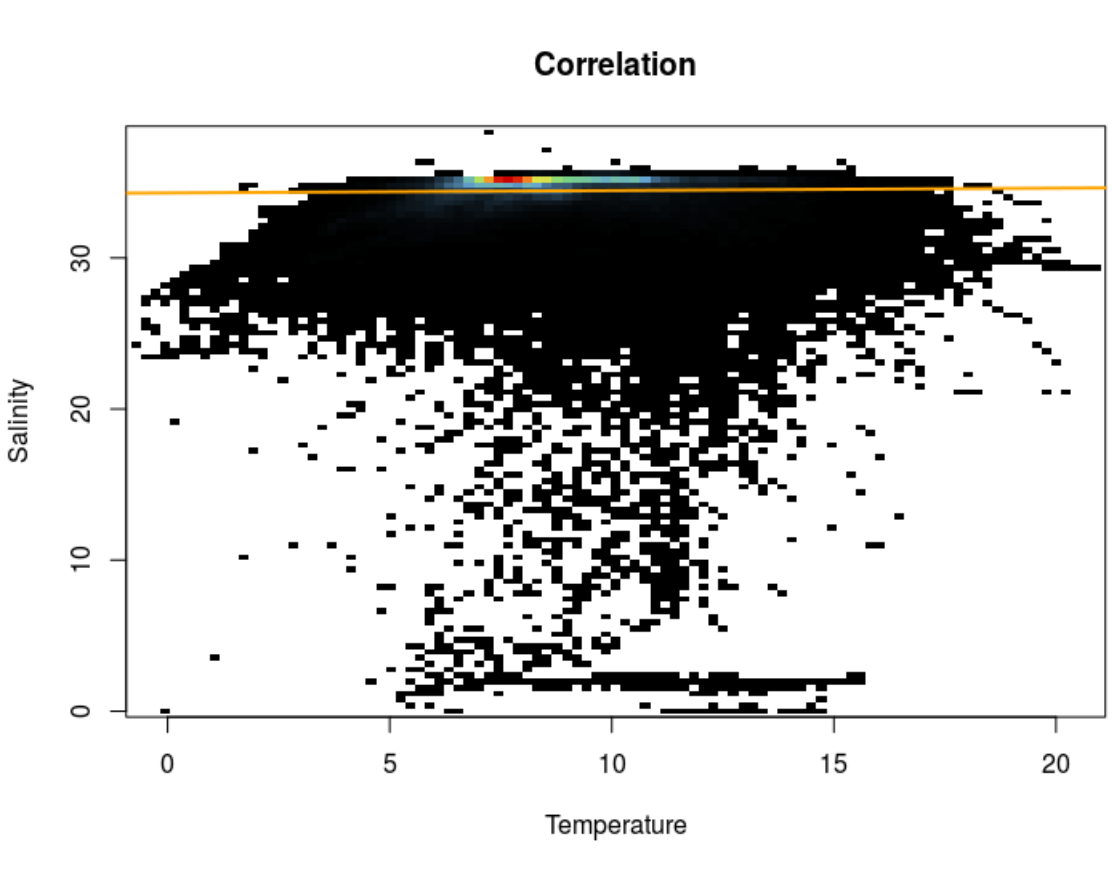




Bivariate Approach: Temperature & Salinity

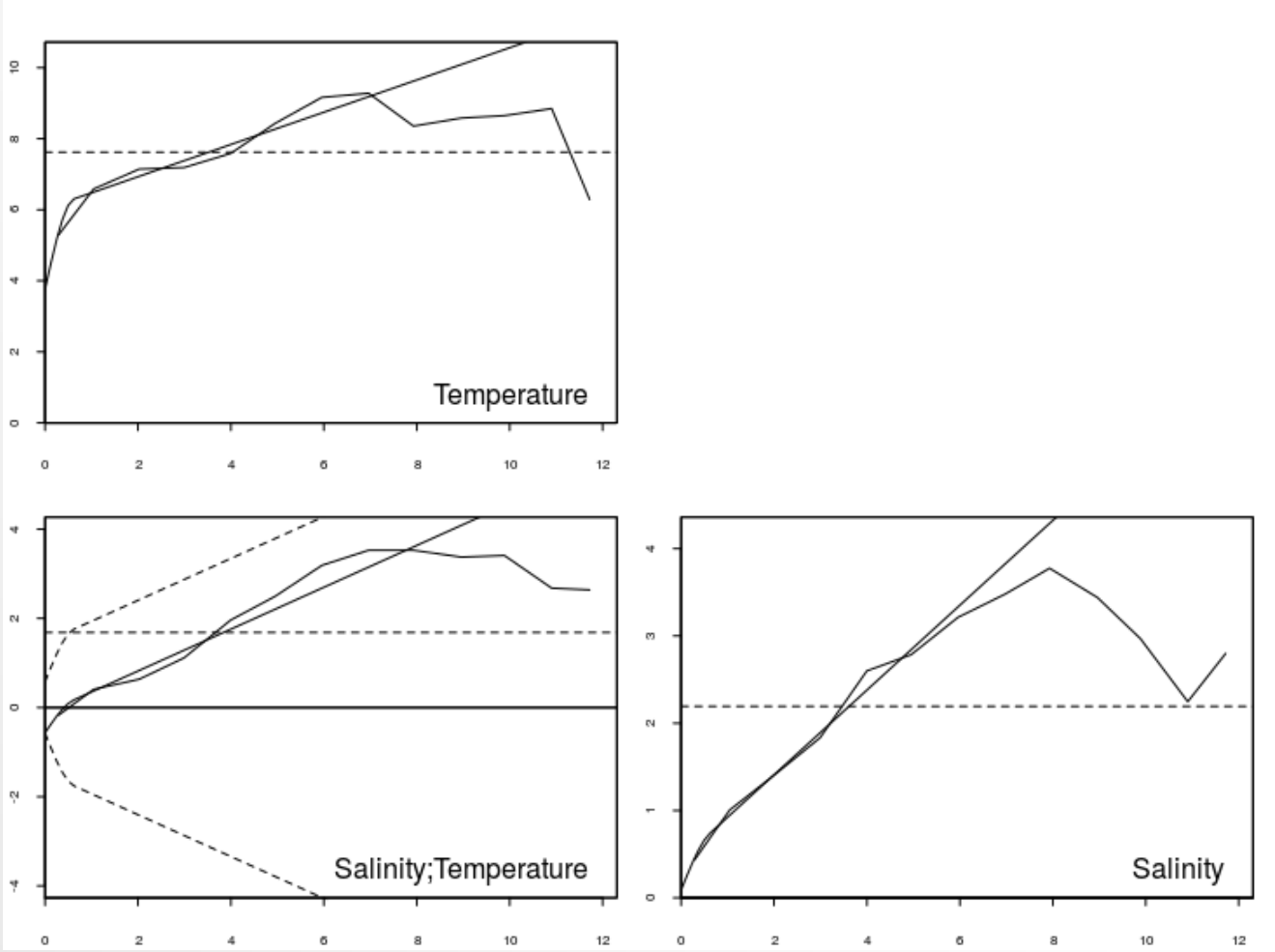
Temperature vs. Salinity

- Define New Variables of Interest
- Correlation and Linear Regression



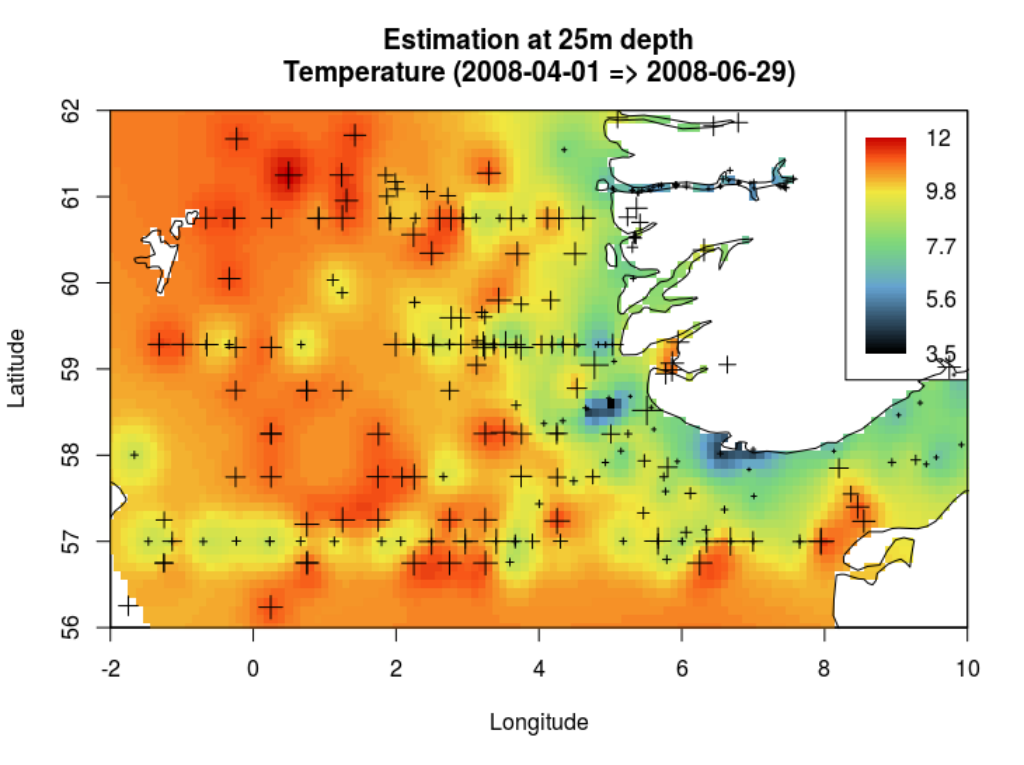
2-D Bivariate Variogram

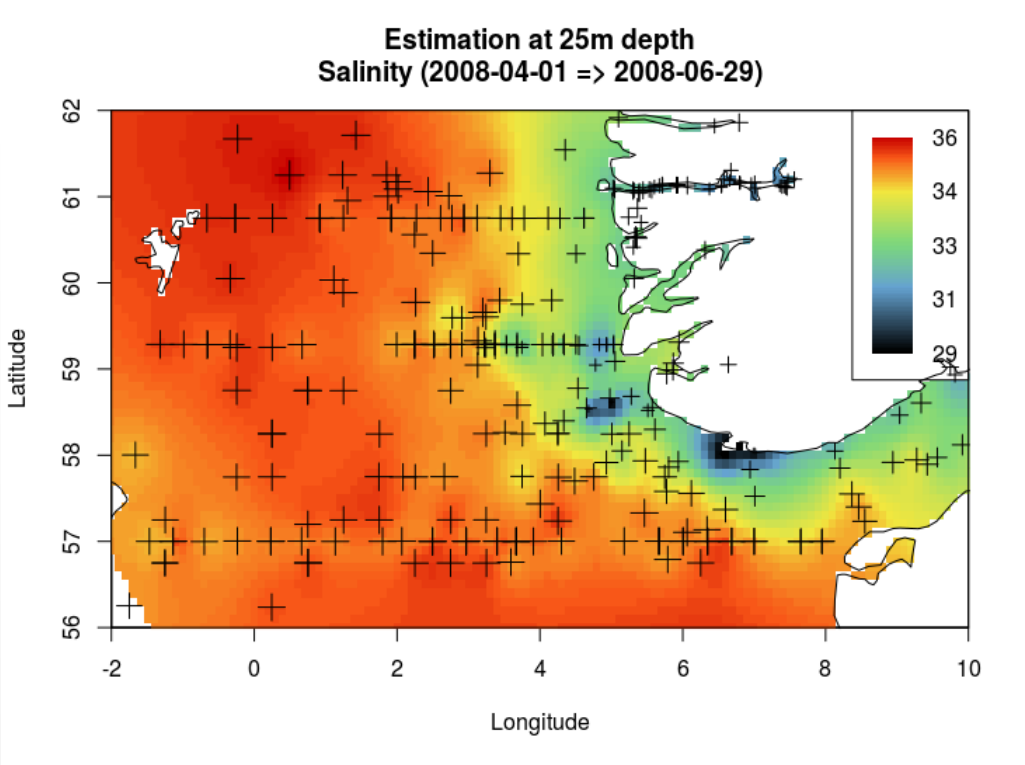
- Select Depth Interval (25m) and Focus on 2nd Trimester of Year 2008
- Bivariate Variogram



2-D Estimation (Bivariate)

- Select Depth Interval (25m) and Focus on 2nd Trimester of Year 2008
- Calculate Average Horizontal Bivariate Variogram and Fit the Model
- Estimate the Temperature and the Salinity at 25m Depth by Cokriging

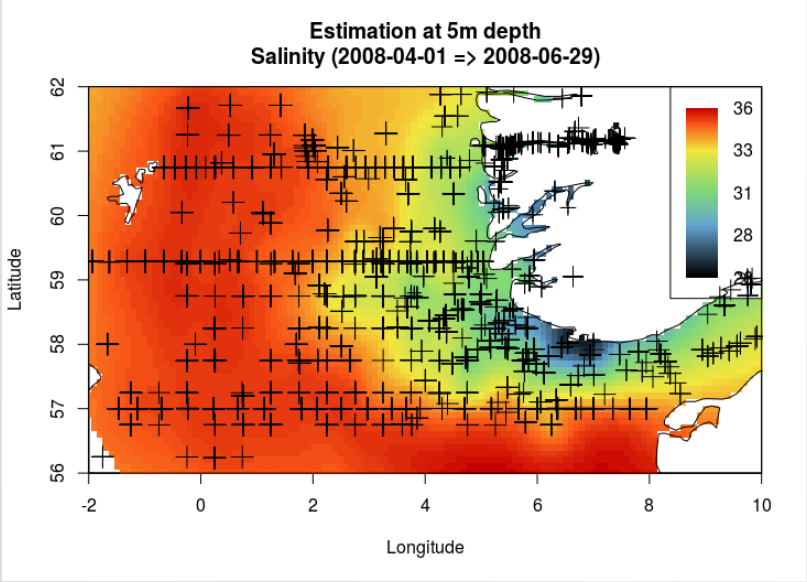


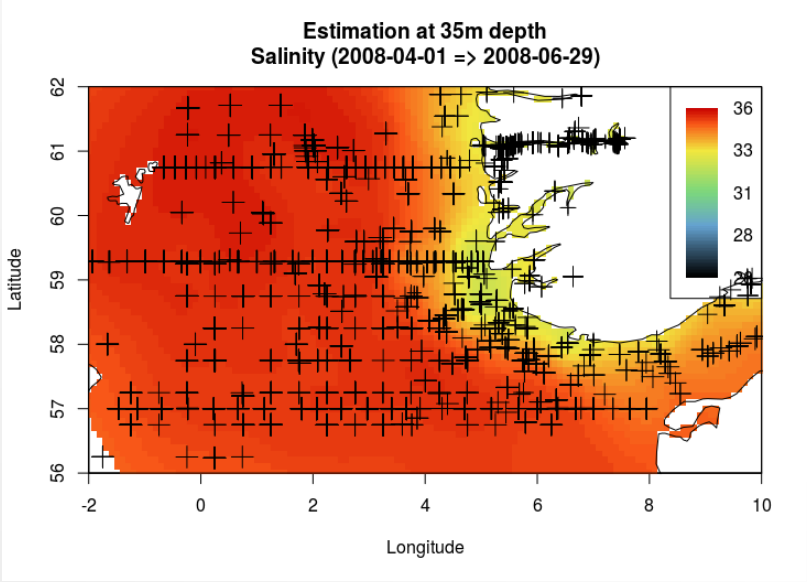


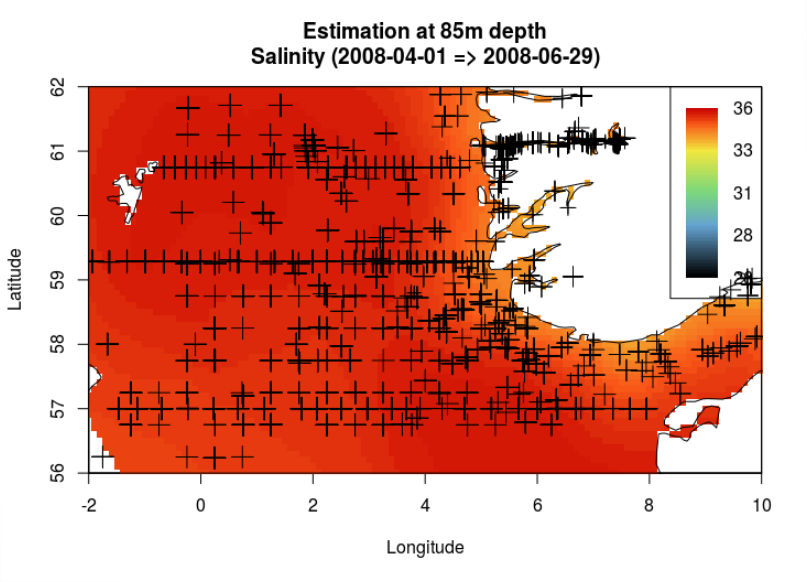
3-D Estimation of Salinity

3-D Estimation of Salinity

- Define New Variable of Interest
- Focus on 2nd Trimester of Year 2008
- Calculate 3-D Variogram (by Bench) and Fit the Model for Salinity
- Estimate Salinity by Benches of 10m Thick from 5m to 95m Depth

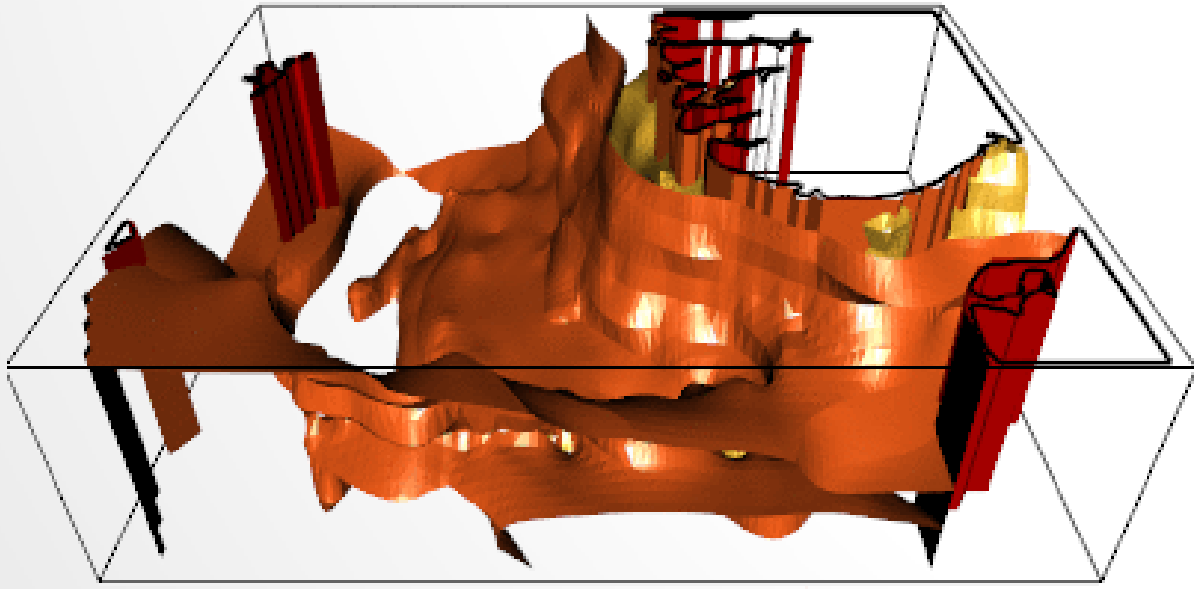






3-D Estimation by Iso-Surfaces

- Show 3-D Estimation of Salinity with Iso-Surface



X

End of presentation

Thank you for attention!