Integrated Arctic Observation System

Research and Innovation Action under EC Horizon2020
Grant Agreement no. 727890

Project coordinator:
Nansen Environmental and Remote Sensing Center, Norway

Deliverable 5.10

Geostatistical Library V2

Start date of project:	01 December 2016	Duration:	63 months
Due date of deliverable:	31 May 2021	Actual submission date: Resubmission after review:	12 July 2021 16 Aug 2022
Lead beneficiary for preparing the deliverable:		ARMINES	
Person-months used to produce deliverable		> 2 pm	

Authors: Didier Renard (ARMINES), Fabien Ors (ARMINES), Hervé Caumont (TERRADUE)

Reviewers: Hanne Sagen (NERSC), Torill Hamre (NERSC)

Version	DATE	CHANGE RECORDS	LEAD AUTHOR
0.1	2021/06/09	First version	D. Renard, F. Ors, H. Caumont
0.2	2021/06/15	Second version	D. Renard, F. Ors, H. Caumont
0.3	2021/07/05	Review by coordination team	H. Sagen, T. Hamre
1.0	2021/07/09	Version for delivery to EC	F. Ors, H. Caumont
1.1	2022/08/04	Version after review from EC	F. Ors, H. Caumont, D. Renard

Approval	Date: 11/08/2022	Sign. Stein Sandven	Stein Sandven

USED PERSON-MONTHS FOR WRITING THIS DELIVERABLE					
No	***Beneficiary***	***PM***	***No***	***Beneficiary***	***PM***
1	NERSC	x	24	TDUE	x
2	UiB		25	GINR	
3	IMR		26	UNEXE	
4	MISU		27	NIVA	
5	AWI		28	CNRS	
6	IOPAN		29	U Helsinki	
7	DTU		30	GFZ	
8	AU		31	ARMINES	1
9	GEUS		32	IGPAN	
10	FMI		33	U SLASKI	
11	UNIS		34	BSC	
12	NORDECO		35	DNV GL	
13	SMHI		36	RIHMI-WDC	
14	USFD		37	NIERSC	
15	NUIM		38	WHOI	
16	IFREMER		39	SIO	
17	MPG		40	UAF	
18	EUROGOOS		41	U Laval	
19	EUROCEAN		42	ONC	
20	UPM		43	NMEFC	
21	UB		44	RADI	
22	UHAM		45	KOPRI	
23	NORUT		46	NIPR	
			47	PRIC	

X: contribution without specifying pm

DISSEMINATION LEVEL		
PU	Public, fully open	**X**
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

EXECUTIVE SUMMARY

The INTAROS project develops an integrated Arctic Observation System (iAOS) by extending, improving and unifying existing systems in the different regions of the Arctic. Within INTAROS, WP5 (Data integration and management) is tasked with designing and implementing evolutions of the cloud platform with geo statistical tools for services as part of the iAOS. The work effort in Task 5.4 has been to design, develop and distribute a Geostatistical Library to be used as a scientific toolbox in some different data analysis workflows (e.g. for the iAOS showcase applications).

The first version of the Geostatistical Library (v1) was considering CTD data provided by IMR for estimating ocean temperature maps in the Barents Sea (D5.6).

This document describes the new version of the Geostatistical Library (v2) where some additional capabilities have been developed.

A set of new functions deals with weather stations and AROME Arctic model in order to estimate snow depth maps in Isfjorden (Svalbard). Another set supports the combination of ICES CTD dataset, GINR trawls dataset and GEBCO bathymetry for estimating bottom ocean temperature maps along the West Greenland coast.

Table of Contents

Table of Figures

Acronyms and abbreviations

AROME	Weather forecasting system for the European Arctic
CSV	Comma Separated Value
CTD	Conductivity Temperature Depth
FMI	Finnish Meteorological Institute
GINR	Greenland Institute of Natural Resources
HTTP	HyperText Transfer Protocol
GEBCO	General Bathymetric Chart of the Ocean
iAOS	integrated Arctic Observation System
ICES	International Council for the Exploration of the Sea
IMR	Institute of Marine Research
NERSC	Nansen Environmental and Remote Sensing Center
NetCDF	Network Common Data Form
OpenDAP	Open Data Access Protocol
SSH	Secure Shell protocol
SYNOP	Surface Synoptic Observation (Weather station data format)
THREDDS	Thematic Real-time Environmental Distributed Data Services
VM	Virtual Machine
VPN	Virtual Private Network
WPS	Web Processing Service

1. Introduction

1.1. Purpose and scope of the document

This document covers the activities of the INTAROS Task 5.4. It has been written as an extension of the **D5.6 - Geostatistical Library v1 report**.

All the background, generalities and important information still relevant for this report have been kept from that previous D5.6 document. This is the best way to keep all related information in a unique consistent document. Each time a section is coming from the D5.6 report, an introduction sentence explains the motivation. The new sections of the present report are dedicated to the work performed in support of two additional iAOS Showcase applications using cloud technology, conducted in collaboration with the INTAROS Task 6.4 (FMI) for mapping Snow Depth for Svalbard Avalanche Forecast Modelling, and Task 6.8 (Aarhus University) for mapping Bottom Temperature in the Baffin Bay.

The aim of the Task 5.4 is to build, deploy and make a Geostatistical Library available to the INTAROS community. The Geostatistical Library is a tool that offers a lot of computing procedures useful for analyzing, estimating and simulating spatiotemporal phenomena. This tool is provided as a software component that can be plugged in any Cloud processing service integrated as part of the development of iAOS applications, or directly used on a laptop computer. An important objective of the Task 5.4 is to disseminate this tool to the INTAROS community and even further to make the geostatistical methodologies available to anyone for their own data analysis needs.

The purpose of this document is to describe version 2 of the Geostatistical Library and how it has been used in several iAOS showcase applications, in collaboration with the INTAROS WP6 partners.

1.2. Intended audience for this document

The main target audience of this document are the INTAROS partners integrating processing services and developing showcase applications using the iAOS cloud platform. However, The Geostatistical Library is intented for a broad audience within the research community integrating heterogenous data in their analysis as well as service developers exploiting multisource data in their products tailored for different stakeholders.

1.3. Document Structure

This section describes the document structure:

Section 1 is this introduction section.

Section 2 briefly describes the motivation of developing a Geostatistical Library in the INTAROS project framework. Then, this section gives an overview of the main features of the new version of the Geostatistical Library.

Section 3 presents the capabilities offered by Geostatistical methodologies and explains how deeper analysis of spatio-temporal phenomenon can be done.

Section 4 presents three iAOS showcase applications realized in collaboration with WP6 partners in order to demonstrate how the Geostatistical Library can be exploited in some concrete use cases.

Section 5 describes how the Geostatistical Library has been deployed as a service in the Ellip Cloud platform environment part of the iAOS.

Section 6 is the conclusion of this report.

2. Activities for iAOS

2.1. Benefits of Geostatistics for INTAROS

The two following sub-sections have been kept as-is from the previous INTAROS D5.6 report because the benefits of using Geostatistics for INTAROS remain unchanged. A third sub-section has been added in order to list some general situations where using Geostatistics can really help.

2.1.1. What is Geostatistics?

Geostatistics is a branch of statistics focusing on spatial or spatiotemporal datasets. Developed originally for the mining industry, it is currently applied in diverse disciplines including geography, hydrology, meteorology, oceanography and any discipline regarding environmental control.

Its main goal is to analyze the spatial characteristics of the variable of interest (its continuity, regularity, ...) and capitalize on this spatial model when interpolating this variable, from a few measured samples to a much larger zone of interest. A complementary approach is to use the same spatial model to predict all the possible variability of the variable in order to establish probability maps or risk curves.

The main geostatistical algorithms are now familiar to any earth scientist and have been incorporated in many places, including geographic information systems (GIS) and on many platforms such as INTAROS.

2.1.2. Some examples use cases

Here are some examples among thousands of Geostatistics applications than can benefits to INTAROS partners and user's community:

- Interpolation of ocean temperature or salinity fields for a given depth and time interval (from scattered data) in order to validate climate model projection,
- Filtering of the measurement error of buoys or other in-situ sensors,
- Calculation of the probability of exceeding a given sea ice thickness threshold (useful for example for planning icebreaker cruises),
- Conditional simulations of snow precipitation and resulting snow depth in order to assess the risk of avalanche,
- Filtering of the masking effect of clouds on satellite image,
- Evolution of fish stocks in time and space,
- Seasonal plankton concentration analysis,
- Estimation of the repartition of white bears from in-situ observation…

Some previous real environmental case studies performed by the Geostatistical group of MINES ParisTech are described in the Section 3.6.

2.1.3. General situations where Geostatistics can help

Whatever the application field, Geostatistics is the key discipline for taking advantage of your spatiotemporal data. Here are general situations where Geostatistics can help you a lot:

- Mix data coming from various sources and sampling resolution
- Provide better estimation of indicators using multiple variables at different scales
- 'Calibrate' physical models to field observations
- Improve quality of data analysis by taking into account non-stationarities
- Assess uncertainties and risk management for resource estimations
- Update and validate models using data assimilation

2.2. User driven definition of the Geostatistics Library

All the well-known Geostatistics methodologies are available in an R package named **RGeostats** and built by MINES ParisTech - ARMINES. It can be freely downloaded from its official website http://cg.ensmp.fr/rgeostats. The Geostatistical Library, specifically developed within the INTAROS project, designates an **additional R package named RIntaros** which relies on RGeostats. This package has been designed to make accessible all required RGeostats features that have been used in the several showcases described in the section 4.

RGeostats and **RIntaros** packages are freely available from anaconda websites: https://anaconda.org/Terradue/r-rgeostats and https://anaconda.org/Terradue/r-rintaros, as Linux packages to be used on cloud computing environments such as the iAOS Cloud Platform.

2.3. The Geostatistical Library v2

This section describes the new developments used for INTAROS project and implemented in the RIntaros and RGeostats packages since Geostatistical Library v1 - look at section §1.3 of the INTAROS D5.6 report. (The based version is Bremen Workshop (2019-01): RGeostats 11.2.8 and RIntaros 1.1)

RIntaros and RGeostats main improvements:

- Global environment parameters have been added for:
  - Bounding box limits (Lon/Lat) and
  - File / Image generation parameters
- Coordinates projections support
- Time zones support for date/time conversion
- Geotiff exportation procedure

Dedicated to Svalbard Snow Depth Showcase support (see section §4.2):

- NetCDF AROME files download and load procedures
- Load of SYNOP files from weather stations
- Dedicated functions for change of supports, modeling and estimation

Dedicate to Off Greenland Bottom Temperature showcase support (see section §4.3):

- CTDs and Trawl loading and cleanup procedures

- Time referential management
- GEBCO bathymetry loading procedure
- Dedicated functions for change of supports, modeling and estimation

2.4. Using the Geostatistical Library in a Cloud Processing Service

The Geostatistical Library v1 has been deployed as a service in the Ellip cloud computing environment used for the iAOS cloud platform. It has been done in the framework of the Task 6.2 IMR Showcase and presented in the INTAROS D5.6 report. For keeping this report consistent, the whole section 5 here has been kept from the D5.6 report.

2.4.1. RIntaros and RGeostats

In order to simplify the use of the RGeostats package and improve the accessibility to geostatistical methods for the INTAROS community, a Geostatistical Library has been created. It is a new R package named RIntaros, which relies on RGeostats. The first version or RIntaros is dedicated to IMR dataset.

The general description of the features of RGeostats is done in §3.2.2. Those for RIntaros are described below for comparison.

The RIntaros package has been explicitly developed to complement RGeostats when manipulating spatial data coming from IMR Data Base or for packaging sets of specific workflow to reduce their complexity for non-expert users. We can distinguish:

- Utilities for processing IMR data (reading according to specific formats, date conversion, selection based on time or depth intervals)
- Specific workflows for:
    - Basic statistics benefiting for spatial coarse gridding or analysis of time series
    - Modeling the spatial structure of target variable(s)
    - Estimation in 2-D (by vertical layers for example) or in 3-D
    - Blind test or Cross-validation

RGeostats and RIntaros software packages have been made available on standard online repositories (on anaconda.org) for use on **Cloud Computing environments such as defined for iAOS** (cf. D5.5 - iAOS requirements and architecture consolidation V2).

Evolutions of the RGeostats and RIntaros software are being maintained by ARMINES and regularly updated on the following online repositories for the convenience of INTAROS partners developing new iAOS Processing Services:

- Repository of latest Conda package build for RGeostats:
  https://anaconda.org/Terradue/r-rgeostats/files
  e.g. **linux-64/r-rgeostats-14.0.5-r35_5.tar.bz2**, uploaded on 2022/07/28
- The related build recipe is documented here:
  https://github.com/ec-intaros/r-rgeostats

and

- Repository of latest Conda package build for RIntaros:
  https://anaconda.org/Terradue/r-rintaros/files
  e.g. **linux-64/r-rintaros-2.4-r35h39e3cac_1.tar.bz2**, uploaded on 2022/07/28
- The related build recipe is documented here:
  https://github.com/ec-intaros/r-rintaros

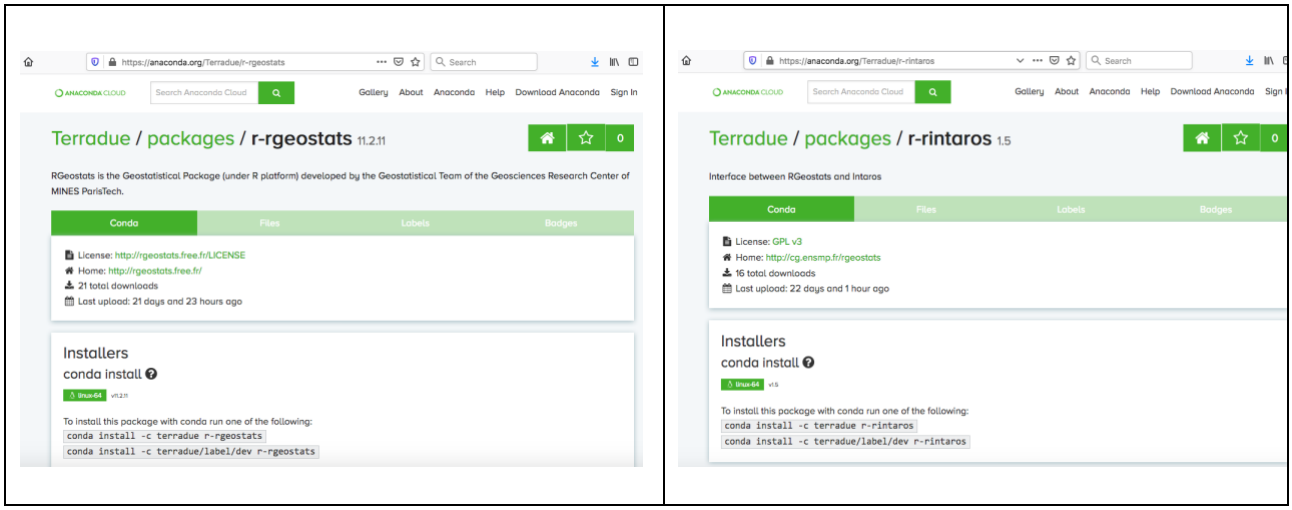


Figure 1. RGeostats and RIntaros Conda packages repositories

r-rgeostats and r-rintaros anaconda packages are available for the INTAROS community to be freely downloaded from their repositories since September 2019.

Their adoption by the INTAROS community is being promoted through the joint work of WP5 and WP6 on the development of showcase applications for the iAOS.

2.4.2. Using the Ellip Solutions to build new iAOS Processing Services

The Ellip Solutions, provided by the INTAROS Partner Terradue as part of the iAOS environment, enable subscribers to work with a Platform-as-a-Service environment, providing them with an integrated user experience for the design and test of their data processing unitary functions (including for sharing towards partners as a reproducible experiments), for the design, integration and test of scalable processing chains, for the packaging and deployment of validated processing services on Production servers from a selected Cloud Provider, as well as the monitoring of the execution and result generation of the deployed processing services.

The Ellip Dashboard provides an integrated online access to the different services, and from there, a typical scenario is for the user to follow the journey as depicted hereafter:

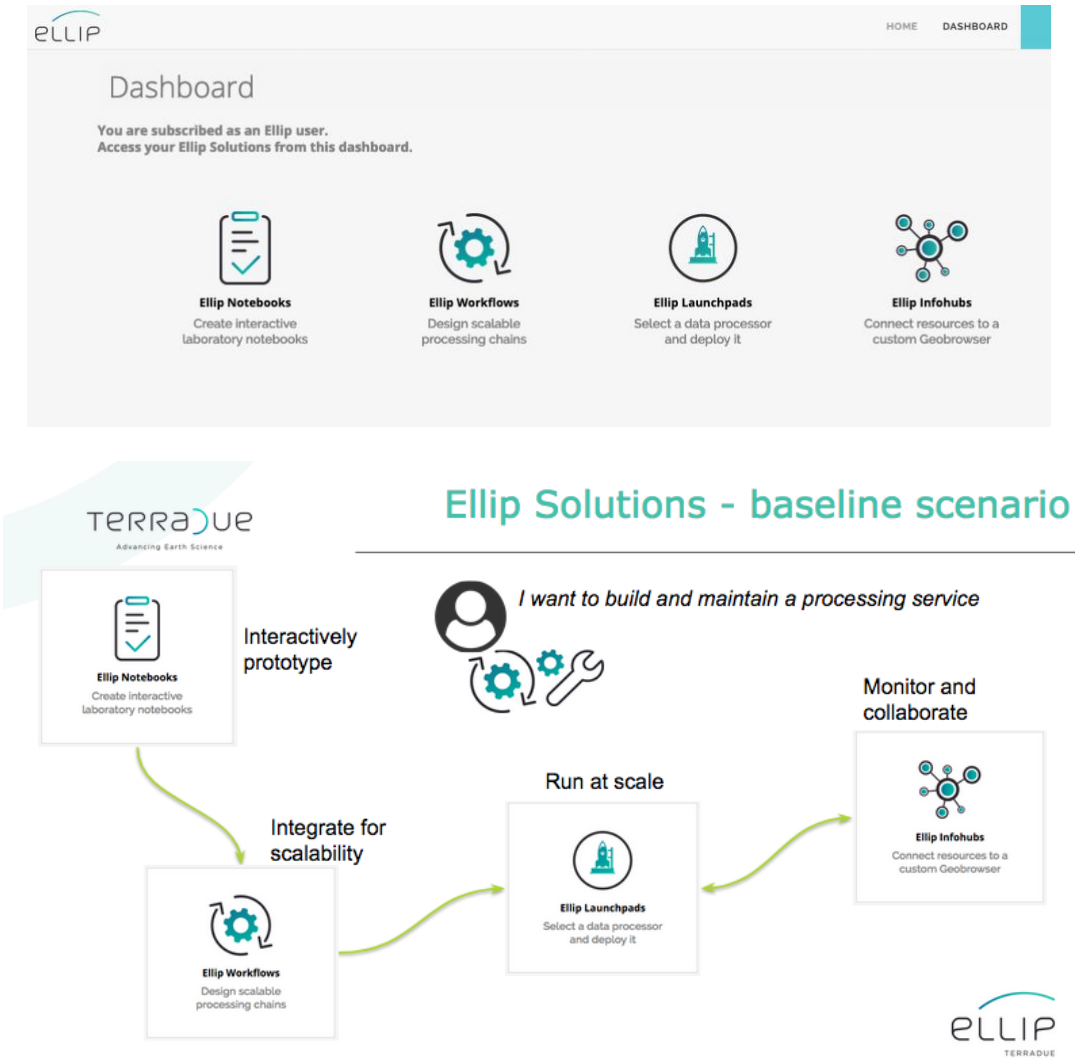


Figure 2. Ellip Solutions user dashboard & usage baseline scenario

An overview of the Ellip solutions is provided online here:

- https://www.terradue.com/portal/ellip (core services and solutions portfolio)
- https://ellip.terradue.com (Ellip Dashboard, subscription-based access only)
- https://docs.terradue.com/ellip (online documentation)

Note 1: the Ellip Solutions have been introduced in more details within the INTAROS Deliverable D5.5 - iAOS requirements and architecture consolidation V2

2.4.3. Application Design

As part of the INTAROS project, a collaboration with the project partner IMR provided the iAOS with a first online version of a data server (based on the OpenDAP standard protocol) and delivering the IMR datasets introduced in the previous section (temperature, salinity and conductivity in the North Sea).

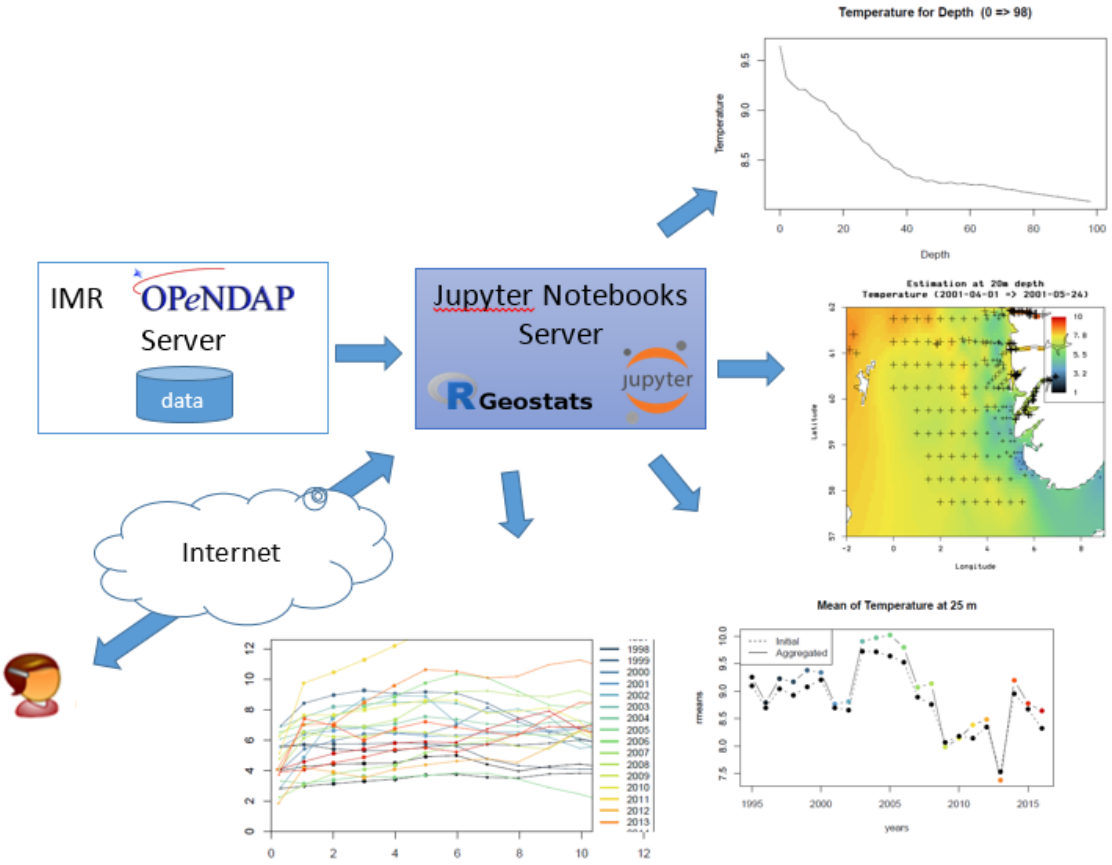


Figure 3. Application design based on remote data access and Cloud-based Jupyter Notebooks

By accessing the Ellip Notebooks solution, a series of data processing functions have been designed and tested on a JupyterLab environment.

On this work environment, the RIntaros and RGeostats libraries have been pre-installed by ARMINES with the support of Terradue, in order to exploit the selected IMR dataset.

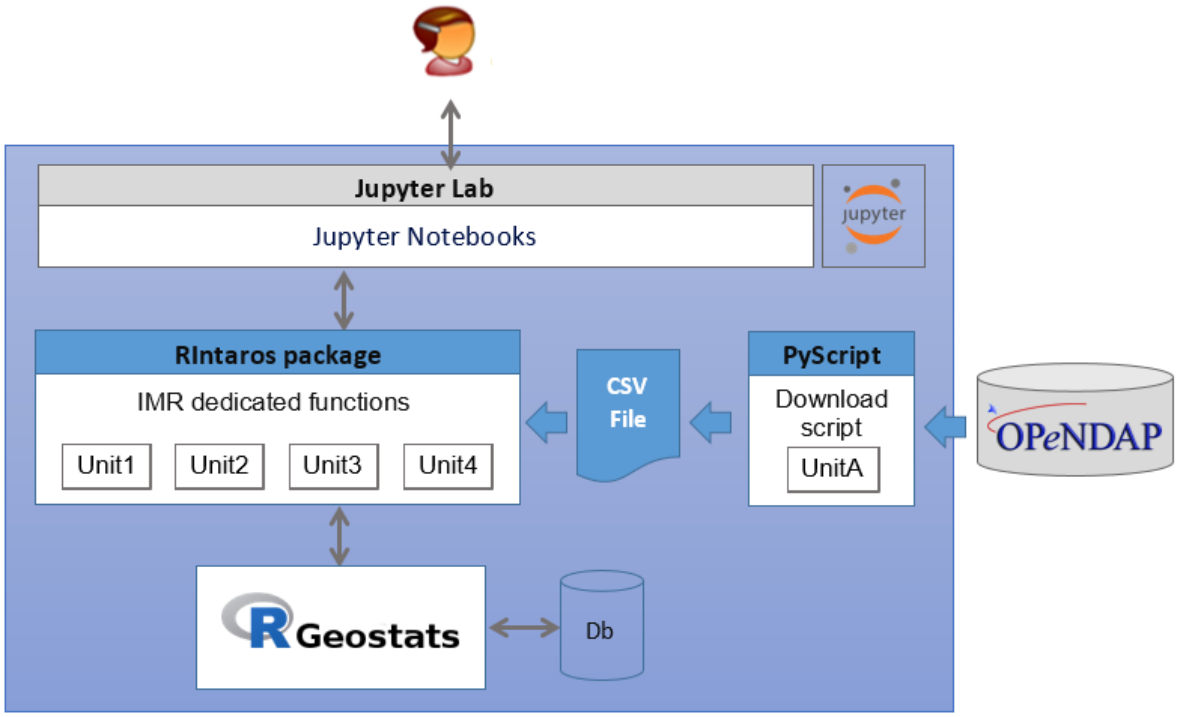


Figure 4. Elaboration of unitary data processing jobs

The scripts developed by ARMINES address the different operations of data filtering and data processing useful for exploring and analyzing dataset contents.

We provide hereafter an overview of the 'downloadData.py' and 'estimate.R' functions and their software dependencies, especially for the later onto the RIntaros and RGeostats libraries.

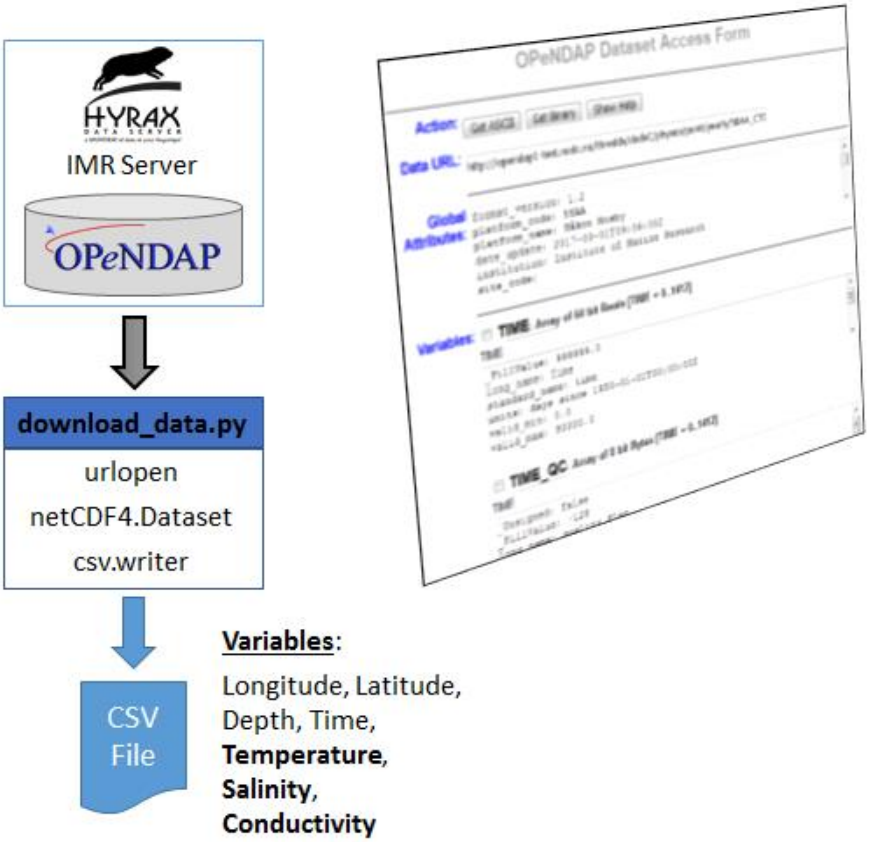


Figure 5. Data access from IMR OPeNDAP server

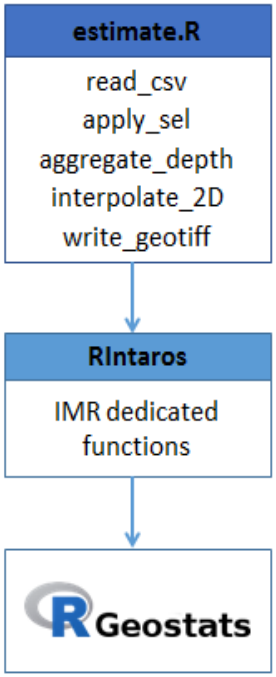


Figure 6. estimate.R script using RIntaros functions

A set of data management scripts resulting from this activity are available as Jupyter Notebook files (.ipynb) and are synchronized online on a public Git repository:

https://github.com/ec-intaros/RGeostats-workshop

They can be easily installed in a JupyterLab user workspace on the Ellip Notebooks solution.

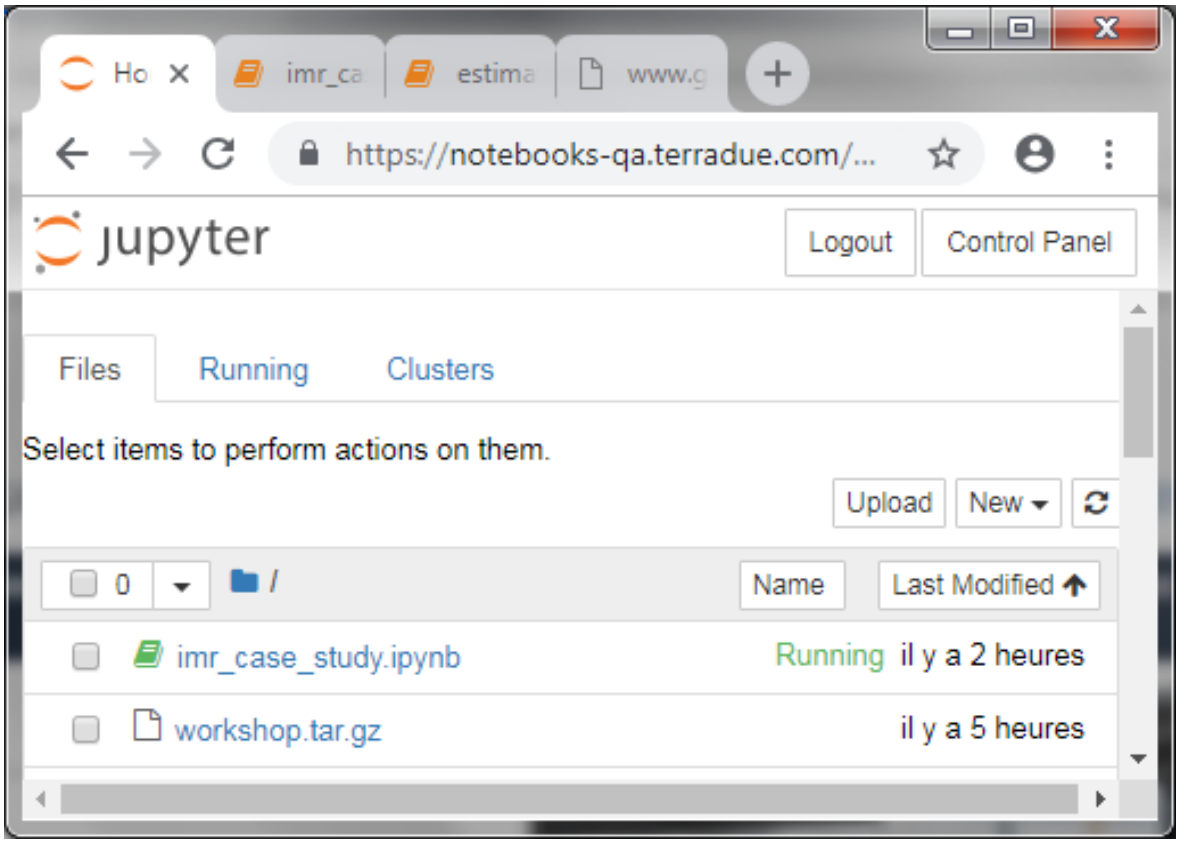
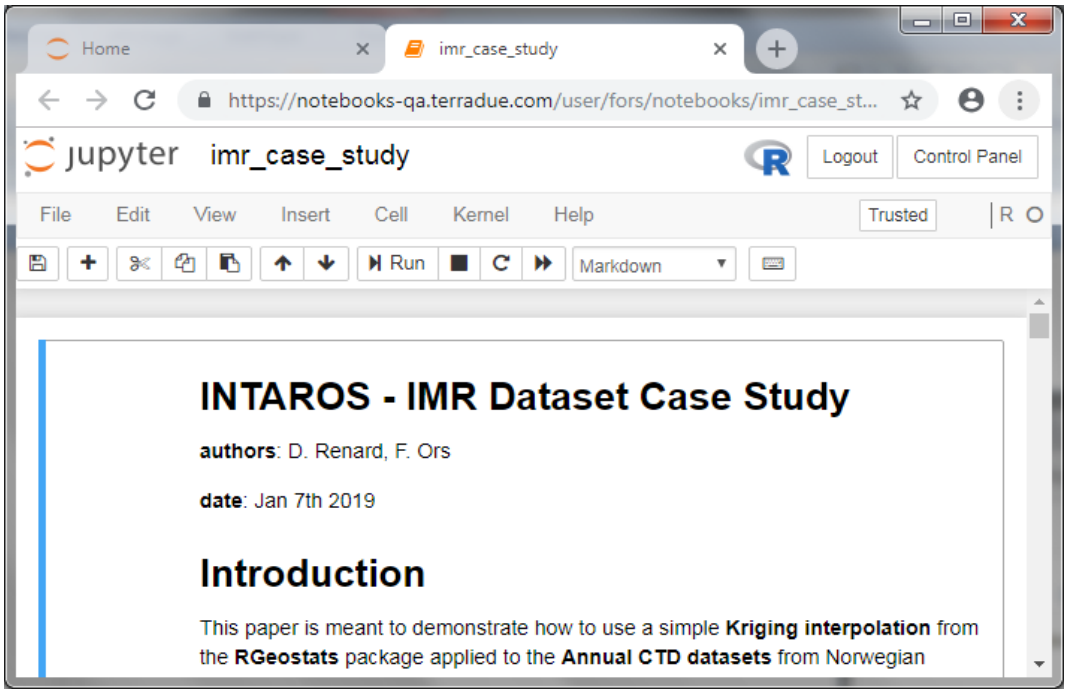


Figure 7. JupyterLab workspace provided by the Ellip Notebooks solution

Once loaded, the INTAROS IMR Case study comes with Workshop material such as a Geostats Course and a description of the IMR Dataset in scope of the RGeostats Course, as well as the executable Notebook files that are embedding the data management scripts.



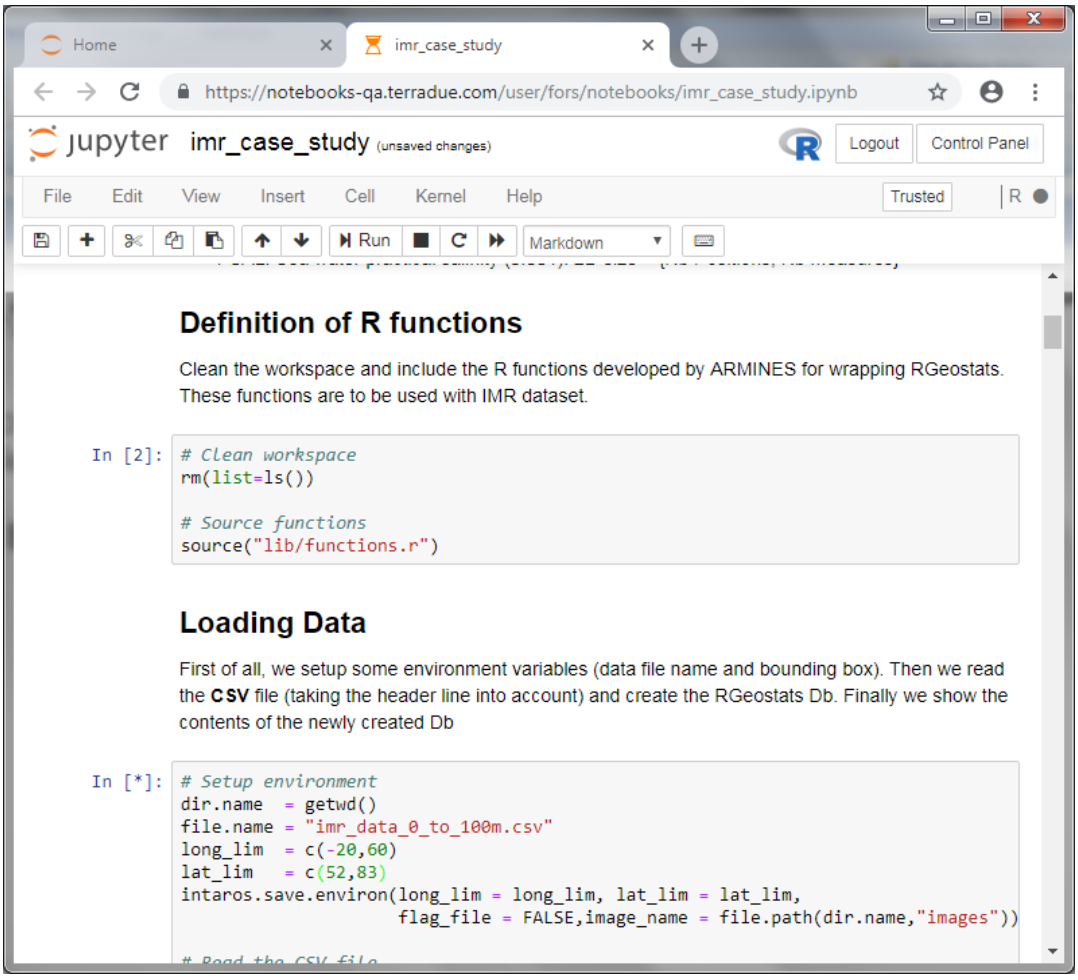


Figure 8. Use of Jupyter Notebooks (IMR Case Study)

2.4.4. Application Workflow integration and tests

From the initial design and test activities presented in the previous chapter, a next step is to integrate these functions as part of a scalable data processing service.

Here the aim is to have the resulting data processing Workflow deployed and operated as-a-Service, for example accessed online by a range of users from a dedicated iAOS Web portal.

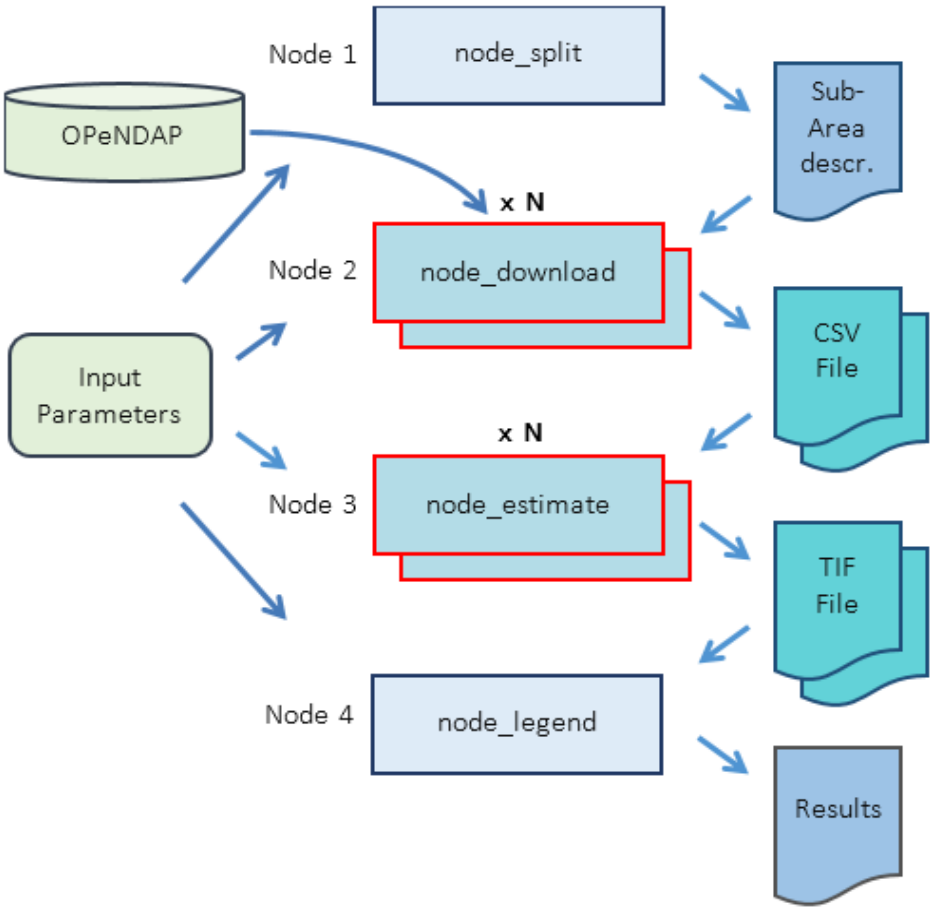


Figure 9. Workflow integration and design of parallelization nodes

Once the data processing chain is integrated using the tools and APIs provided by the Ellip Workflows solution, it can be tested from within the 'Sandbox' virtual machine where it is integrated. The number 'N' of sub-areas is automatically calculated according to the dimension of the output estimation map and the number of available cores of the cloud computing cluster.

Application runs can be tested from the Browser (accessing the Virtual Machine through secured VPN connection).

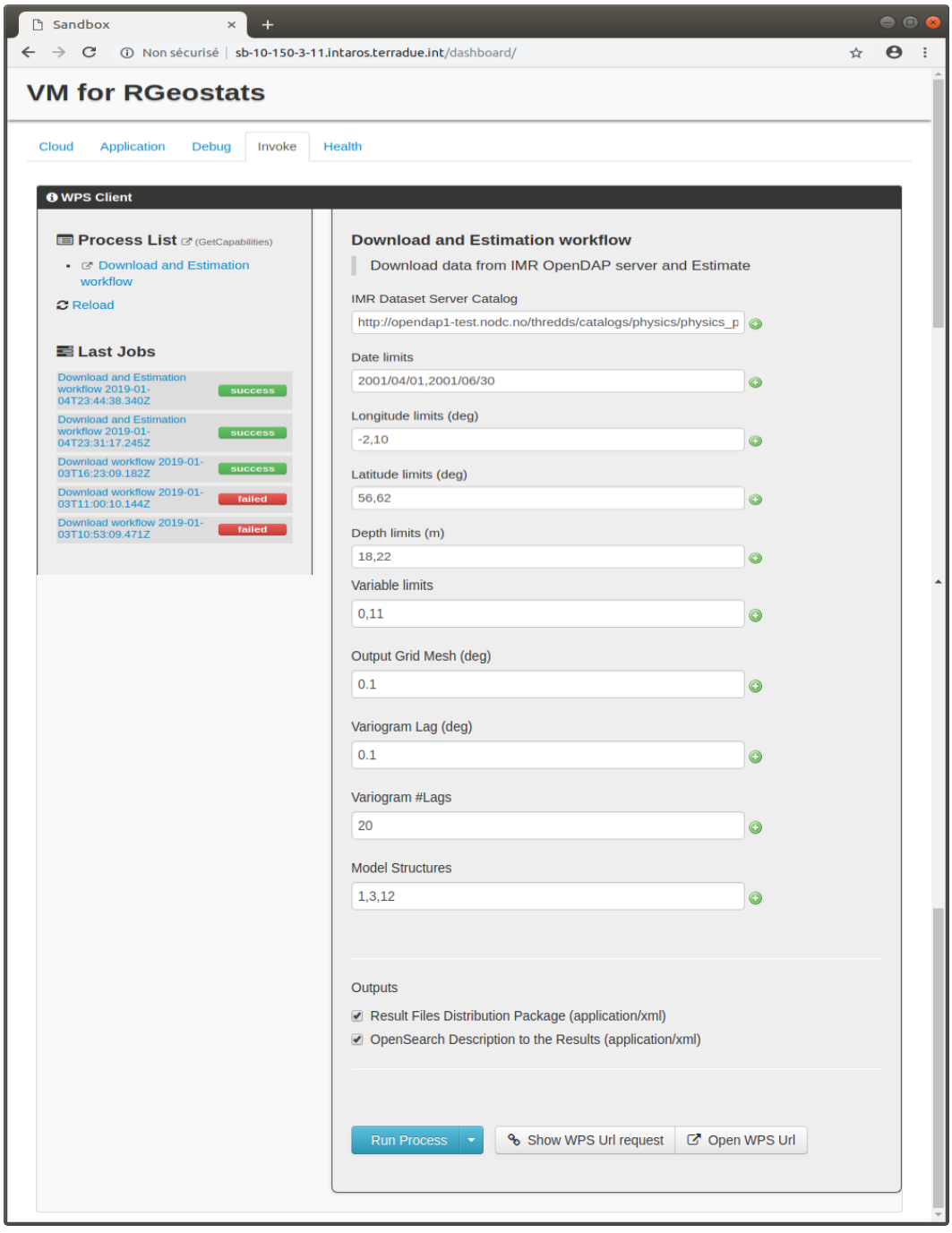


Figure 10. Workflow input parameters and application run (Ellip VM - test client view)

Application runs can also be tested from the Console (accessing the Virtual Machine through secured SSH connection).

```
$ ciop-run

2019-10-16 16:29:06 [INFO ] - Upload results? null jar:file:/usr/lib/ciop-run/ciop-joozie-
1.2.jar!/schemas/oozie-workflow-0.1.xsd

2019-10-16 16:29:06 [INFO ] - Workflow submitted

2019-10-16 16:29:06 [INFO ] - Closing this program will not stop the job.

2019-10-16 16:29:06 [INFO ] - To kill this job type:

2019-10-16 16:29:06 [INFO ] - ciop-stop 0000020-191005043900365-oozie-oozi-W

2019-10-16 16:29:06 [INFO ] - Tracking URL:

2019-10-16 16:29:06 [INFO ] - http://sb-10-150-3-19.intaros.terradue.int:11000/oozie/?job=0000020-
191005043900365-oozie-oozi-W


Node Name      :  node_split

Status         :  OK


Node Name      :  node_download

Status         :  OK


Node Name      :  node_estimate

Status         :  OK


Node Name      :  node_legend

Status         :  OK


Publishing results...


2019-10-16 16:38:29 [INFO ] - Workflow completed.

2019-10-16 16:38:29 [INFO ] - Output Metalink: http://sb-10-150-3-
19.intaros.terradue.int:50070/webhdfs/v1/ciop/run/download_and_estimation_workflow/0000020-
191005043900365-oozie-oozi-W/results.metalink?op=OPEN
```

Figure 11. Application run for generation of test results (Ellip VM – console view)

2.4.5. Application Workflow deployment for user access

As part of the Ellip Solutions, the PaaS environment offers a capacity to:

- Deploy the processing service on production servers, where it can be run at scale on a large Cloud Computing cluster.

- Simply generate a parameterized Geobrowser application, that can be used to reference the deployment of the Processing Service, run it, monitor the data processing jobs that are interactively launched, and visualize/analyse the generated results.

The definition of the processing service output product files, metadata files and legend has to comply to a simple set of design conventions (part of the online Ellip documentation) in order to be automatically discovered and handled by the Geobrowser application.

The Geobrowser, part of the Ellip Solutions, can then use the metadata of a job processing results to handle data visualization on the map. It actually exploits the Terradue Cloud Platform API for this task. For instance, the processing service output product files, metadata files and legend have to comply with the following guidance.

2.4.5.1. Files aggregation

As a first process, the dataset files are listed to find all similar files and regroup them as a single result entry. The filename without the extension is used for this aggregation.

For the IMR Case Study, the application outputs are grouped as shown hereafter, following the convention provided by Terradue for the Ellip Workflows applications:

Figure 12. Aggregation of filenames per output type and tile index (here tile 7 shown)

2.4.5.2. Data properties

A dataset shall also contain a Java properties file (key=value), named as the data file it described by replacing the extension or suffixing with ".properties", to give additional information about a result file. This additional information will be added into the metadata entry of the output file as metadata information. All keywords / value from the .properties file are added as a table to the summary element (used for metadata display on the Geobrowser).

For the IMR Case Study, the product metadata properties are defined as shown hereafter, following the convention provided by Terradue for the Ellip Workflows applications:

```
title=Estimation of Temperature
variable=Temperature.estim
date=2001/04/01,2001/06/30
depth=[18,22]m
bbox=-2,56,10,62
processors=RGeostats & RIntaros packages
url=http://cg.ensmp.fr/rgeostats
```

Figure 13. Metadata properties defined for a job processing output

2.4.5.3. Legend

In order to attach a legend to describe the dataset, a file suffixed with .legend.png enables a map functionality that displays the legend when the data is selected.

For the IMR Case Study, the product legends are defined based on the normalized range of values from the output products, to with a color scale is applied, as shown hereafter, following the convention provided by Terradue for the Ellip Workflows applications:

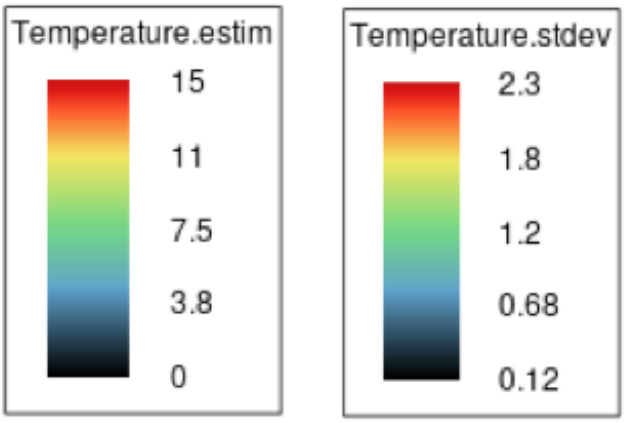


Figure 14. Legend scales and rendering for a job processing output

After the application release and deployment based on such implementation step in the application code, the processing service output products can then be automatically discovered and handled by the Geobrowser application, for rendering and visual exploitation by users of the service.

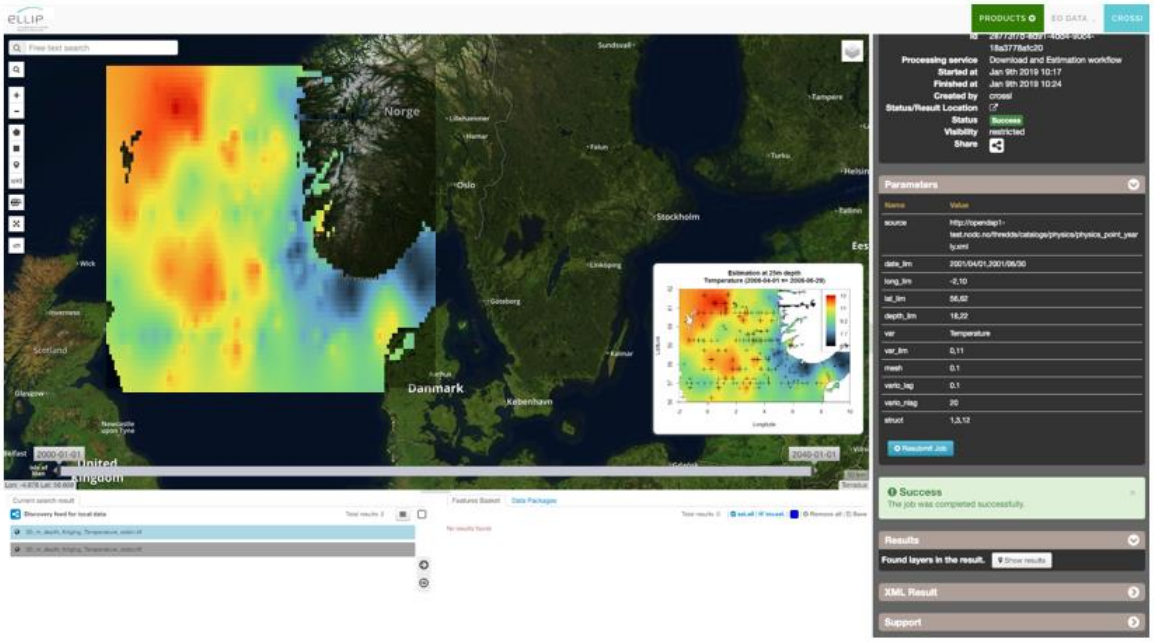


Figure 15. Geobrowser with access to the application as-a-service, and visualization of job results

The Geobrowser application used as an example here is a private application provided to ARMINES on the Ellip Infohubs solution (operated by Terradue for iAOS, see D5.5 and D5.8).

ARMINES can control the user sharing configuration for such a Geobrowser application.

The technical solution planned for accessing the EWF-IMR-ESTIM processing service (deployed on the iAOS Cloud Platform) uses the interoperability protocol "OGC WPS" that is natively supported by Ellip-powered processing services. After the service has been published with a WPS endpoint the service can be be started from Web Applications (typically a Web portal) supporting this standard protocol. The produced files from the service must comply with standard geo reference formats used in web-mapping. This will allow the iAOS Portal to display the analysis results on a map.

3. Geostatistics - an overview

This section provides an overview of Geostatistics methodologies. It has been copy-pasted from the INTAROS D5.6 report, in order to keep the theoretical description and showcase applications presentation as part of the present document.

3.1. Generalities about Geostatistics

Geostatistics is a rather recent scientific discipline, which stands as a branch of statistics applied to spatial or spatiotemporal data. Originally developed for tackling problems in the Mining Industry, it rapidly expanded towards many other domains of application, such as Petroleum Industry, Hydrogeology, Meteorology, Climatology and Environmental Control.

The geostatistical techniques are applied on (almost) any type of data, usually called *variable*, as long as they are defined in a space of any dimension (possibly including time): i.e. a *regionalized variable*. The information is provided as a set of data measurements: this data set can contain only few samples or large quantities of samples. Each sample provides the value of the variables of interest, measured at a given location or time, and on a given volume (measurement support).

The geostatistical procedures usually apply to a target variable. Most of them can be extended to the case of several variables, benefiting from samples measured on all these variables (the count of samples per variable can be different, i.e. *heterotopic* case) and from the joint spatial characteristics between these variables. This multivariate concept can even be extended further when one of the variables is measured exhaustively and can serve as a *shape factor* for the others (external drift method).

Geostatistics covers a whole set of produces such as:

- **Modeling**: capability of describing a physical phenomenon and its spatial characteristics.
- **Interpolation**: extending the knowledge to a whole field starting from a small set of measurements. This corresponds to the traditional mapping concept.
- **Estimation**: predicting the variable contents (grade for example) in an extraction volume (mining selective unit) from samples collected on much smaller support (core samples).
- **Risk analysis**: predicting the probability that the field of interest exceeds a given threshold.

Geostatistics has gained momentum and can now be found in many commercial software offers. The most well-known is certainly ISATIS© commercialized by Geovariances. Some functions can be found on Internet here and there, which perform some of the tasks mentioned above.

Another software named RGeostats is freely available on the web. It has been developed since 2001 by the Geosciences center – a common research center of ARMINES and MINES-ParisTech.

3.2. RGeostats Package

3.2.1. What is RGeostats?

RGeostats stands as the largest collection of geostatistical functions available within a single R package. It can be downloaded for free from http://cg.ensmp.fr/rgeostats. Once installed in the local R environment, the user can then establish some R scripts which call RGeostats functions for a specific activity (Data analysis, Modeling, Estimation, etc…).

RGeostats is an R package which relies on a C/C++ library called Geoslib. The interface between RGeostats R functions and the C/C++ code is made possible by the Rcpp package.

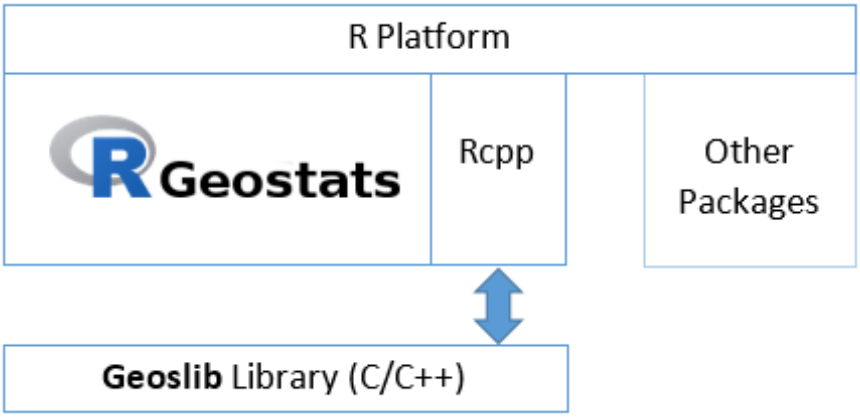


Figure 16. RGeostats R package architecture

The RGeostats website proposes several services that can guide the non-expert users: demonstration scripts in a particular context (vignettes), online help for more than 500 functions available, discussion forum for troubleshooting and News. The RGeostats user community (academic and researchers) has more than 700 members in the world.

The RGeostats package and guidance for the downloading or the loading from RStudio of the R Package archive file is provided on the RGeostats website.

Figure 17. RGeostats website

As part of the INTAROS WP5 activities, the RGeostats software package has been made available on standard online repositories (on anaconda.org) for use on Cloud Computing environments such as defined for iAOS (cf. D5.5 - iAOS requirements and architecture consolidation V2). Please look at the §4.1 for more details.

3.2.2. RGeostats Features

RGeostats (and the R platform in general) proposes a large number of statistical tools that allows the user to perform a sound analysis of its data. When the data are structured spatially (in space and/or in time), Geostatistics are then useful.

This package is the basis of the contribution of ARMINES to INTAROS project. It contains all the general purpose geostatistical functionalities (developed for any space dimension), among which those requested by the project.

They can be grossly subdivided into several themes:

- The functions which are used to handle the **Geostatistical objects**, such as the:
  - The normal score transformation of the distribution of the target variable (or Gaussian Anamorphosis)
  - The (numerical) Database where all the input and results are stored. This Database can be organized as a grid or simply a set of isolated points (possibly organized along lines).
  - The particular Database for storing a network of fractures (connected or not)
  - The meshing of a 2-D or 3-D space (i.e. triangulation or tetrahedrization)

  - The (geostatistical) Model which describes the spatial characteristics of the target (set of) variable(s)
  - The Neighborhood parameters which describe the set of samples to be considered for conditioning estimations and/or simulations: it can vary from a dozens of close samples (moving) to the whole data set (unique).
  - The Polygon which delineates the area of interest where the estimation and/or simulation must be performed
  - The lithotype rule which defines the vicinity relationship between classes of a categorical variable
  - The experimental covariance or variogram derived from the values at samples and which give an experimental point of view on the spatial characteristics of the (set of) variable(s) of interest.
- The functions used to follow the main steps of a **Geostatistical workflow**:
  - Data exploration:
    Calculation of basic statistics (reporting, histogram, correlations…)
    Detection of possible trends
    Possible data transformation (Declustering, Normal score …)
  - Modeling the spatial characteristics:
    Computing the experimental quantities (variogram, covariance…) for a single or a set of target variables
    Fitting an authorized model
  - Estimation of the (set of) variable(s) on a set of targets:
    Several traditional interpolation algorithm (used for comparison): inverse distance, voronoï tessellation, trend surface…)
    Estimation by minimizing the estimation variance of any linear transformation of the target variable: Kriging (punctual, block average, over a territory)
    Enhanced:
    - in the case of multiple target variables: CoKriging
    - in the case of trends (Universal Kriging) or similarities with explanatory variables (External Drift)
  - Simulations:
    This corresponds to a large spectrum of features depending on the type of the target variable(s). For example, let us mention, among others:
    - Turning Bands or Spectral methods for continuous Gaussian variables
    - Boolean technique for object based simulation
    - SPDE: for simulating random fields which obey to partial derivative equations
    - PluriGaussian for simulating a categorical variable

3.3. Data Spatial Structure

All geostatistical techniques rely on the analysis of the spatial behavior of the target variable(s). These characteristics are measured experimentally from the information carried by the samples. Several statistical tools can be considered: the most basic one is the *covariance* (centered or not); the most well-known in the geostatistical community is the *variogram*; but we can also name the

variogram of increments (in the case of a non-stationary variable) or other tools such as the *madogram*, the *rodogram*, ...

This experimental quantity is then summarized in the spatial *Model*, which is a parametric formulation, which depends on a limited number of parameters.

3.3.1. Experimental Quantity

In this paragraph, only the experimental variogram is described. First the variogram cloud is established where all pairs of data points are compared (up to a given maximum distance equal to one third of the total field extension). Each pair is represented by the variability as a function of the distance between the two samples. The variability is obtained as follows: $\frac{1}{2}(Z_1 - Z_2)^2$ where Z_1 stands for the value of the target variable at the first point while Z_2 is the value at the second point. Finally, the variability of these pairs is averaged by classes of distance, in order to obtain the experimental variogram. They can also be averaged by classes of direction, in order to depict some possible *anisotropy*.

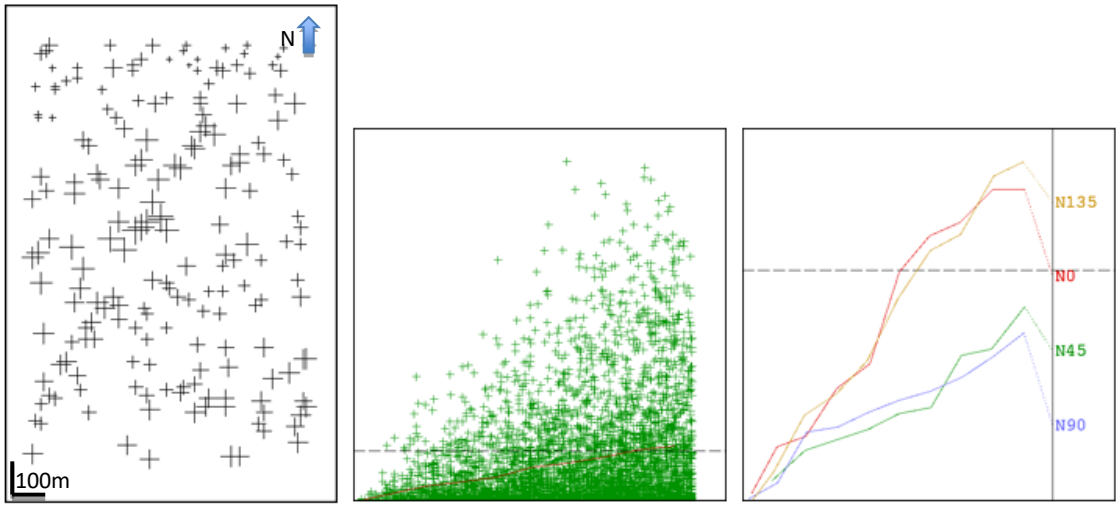


Figure 18. Various views of the Data set and Spatial structure
Left: Data base map (Field extension: 1km along X and 1.5km along Y)
Middle: Variogram Cloud giving variability as a function of distance for pairs of samples
Right: Experimental Directional Variograms

This being considered as a simple example, the exact data point coordinates or the exact values of the variances are not necessary for the understanding.

The left part is a base map (data point location) where the horizontal axis covers 1km and the vertical axis 1.5km.

The middle picture is a variogram cloud describing the 2-D point statistics. As recommended in literature, the horizontal axis (distance) covers one third of the field extension. The vertical axis gives (half of) the maximum variability of a pair of points. The experimental variance of the dataset is represented as a dashed line sitting close to the bottom of the graph.

The right picture is a variogram calculated in four main directions (i.e. N0, N45, N90 and N135). The horizontal axis is the same as for the variogram cloud. The vertical axis covers the maximum

variogram value (reached in directions N135 and N0). The dashed line corresponds again to the experimental variance of the data.

3.3.2. Fitting a Model

The experimental variogram calculated for several classes of distance and orientation must then be fitted by a parametric function called the Model. This function must have some nice mathematical properties (definite positiveness), which ensure that it can be used for variance calculation (always producing positive results).

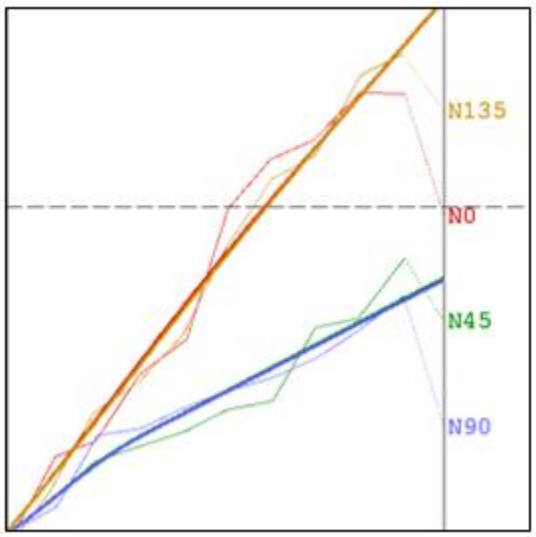


Figure 19. Experimental Variograms and Fitted Anisotropic Model (calculated over half of the field extension)

3.4. Estimation

When the spatial characteristics of the variable of interest have been captured in the Model, we can envisage addressing the estimation / interpolation task performed in the Geostatistical framework. The method named *Kriging* can be used in order to predict the variable in any target location throughout the field, starting from a few measurement points. The targets can be a set of locations of interest or, more generally, the nodes of a regular grid used for mapping.

Kriging presents several virtues:

- It is unbiased: it does not present any tendency to under-estimate or over-estimate.
- It is optimal: on average, it minimizes the error that inevitably occurs in the prediction (according to the spatial characteristics provided by the Model).
- It is an exact interpolator: the result coincides with the measured value when target and data measurement coincide.

As a by-product, Kriging also provides an evaluation of the variance of the error committed in this prediction at each target site. We usually prefer using its square root (the standard deviation), which is expressed in the same unit as the variable itself.

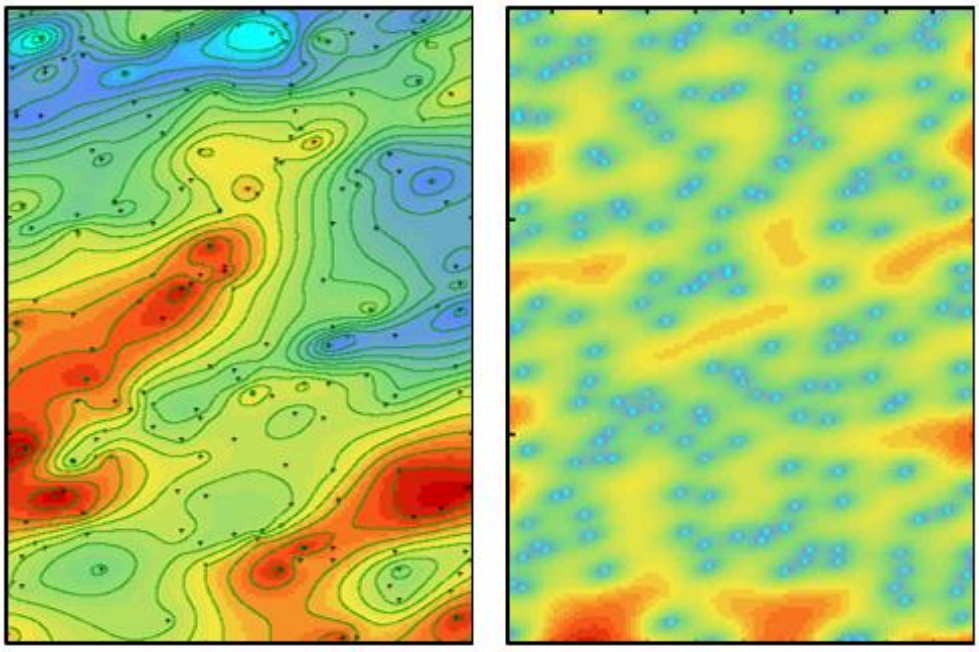


Figure 20. Estimation (left) – Standard Deviation of Estimation Error (right)

3.4.1. Definition of the Neighborhood

The estimation procedure is obtained by solving a linear algebraic system. Its dimension is equal to the number of samples. Therefore in the case of a large data set, this may become intractable. Instead the estimation for each target site can be performed using only a subset of the data located close to this target: this refers to the *Moving Neighborhood* concept.

When the data set is not too large, we may prefer using the *Unique Neighborhood* (where all samples are used for each target site), which offers some nice algebraic properties enabling fast processing.

3.4.2. Different Types of Estimations

We must also mention that the Kriging estimation procedure can be used in many different circumstances. The one described above corresponds to the prediction of the variable in a set of target sites (e.g. the nodes of a regular grid): it refers to the *punctual* estimation. This technique can be enhanced with the *block average* estimation used, for example, to predict the average grade over each selective mining unit (the volume unit during the exploitation) and decide if it must be sent to the mill or to the waste. Kriging can also be used to calculate gradients: this makes sense when predicting the gradient of pressure (wind) starting from atmospheric pressure measured at gauges.

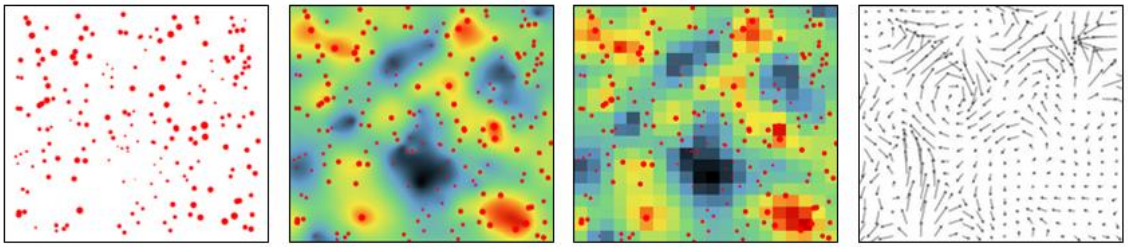


Figure 21. From left to right: Data Set – Estimation Map where target sites are the nodes of a regular underlying grid – Estimation of average value over the cells of a regular grid – Gradient estimation

More generally, Kriging procedure can estimate any quantity linearly related to the measured variable. Finally Kriging can produce the *global estimation* for instance for estimating the total abundance of a fish species in a given portion of ocean for example.

3.4.3. Estimation Properties

As mentioned earlier, Kriging (like any estimation based upon some optimality criterion) tends to produce estimates that are smoother than reality. This is what is demonstrated next by first producing an exhaustive data set in a field (considered as the reality), then sampling it and finally estimating it back on the field again.

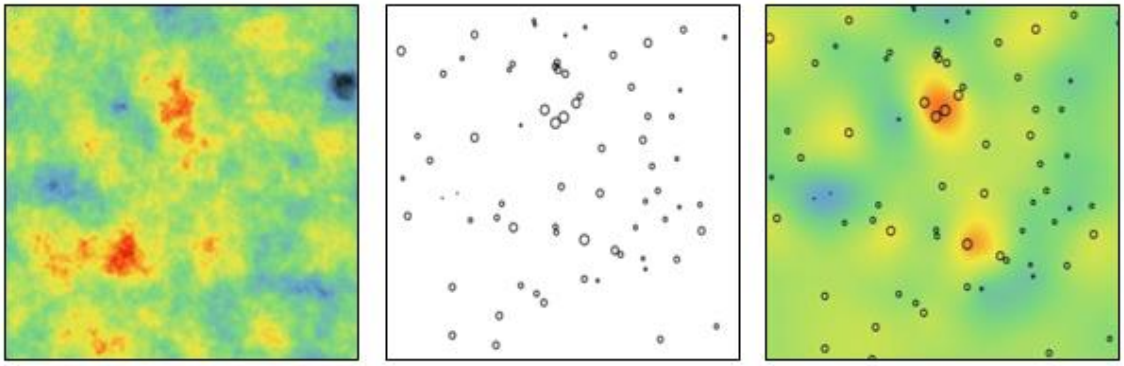


Figure 22. Exhaustive gridded data set or Reality (left) – Location of samples where Reality has been measured (Middle) – Map of the estimation carried over grid nodes or Point Estimation (Right)

3.4.4. Enhancement in Presence of Several Variables

The information on the variable of interest can sometimes be improved by the knowledge of an additional co-variable. The impact of this multivariate approach is measured in the Model, which reflects the joint spatial characteristics of all variables: this requires the calculation of the experimental simple variograms (of each variable considered separately) as well as the cross-variogram (of the pair of variables) that are all fitted by a multivariate Model.

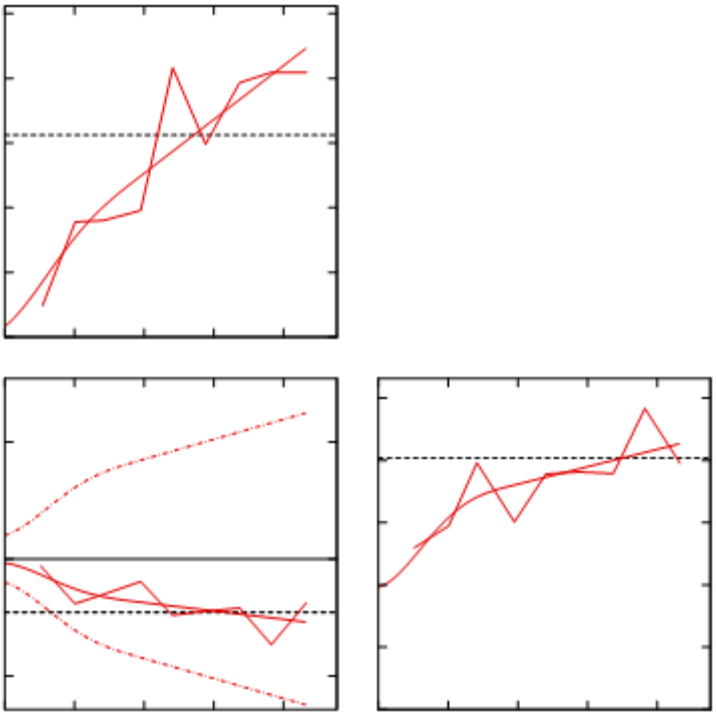


Figure 23. Case of 2 variables processed jointly: Simple variograms for each variable (top-left and bottom-right) – Cross-variogram between both variables (bottom-left)

The estimation is then performed by a Co-Kriging procedure, which takes information from both variables into account through the multivariate Model. In the next example, we can measure the difference between the estimation of the target variable by Kriging compared to the Co-Kriging result using one secondary co-variable. The Co-Kriging procedure can be extended to any number of variables treated jointly as long as they present some significant spatial correlation (at least one cross-variogram between a pair of variables is not flat).

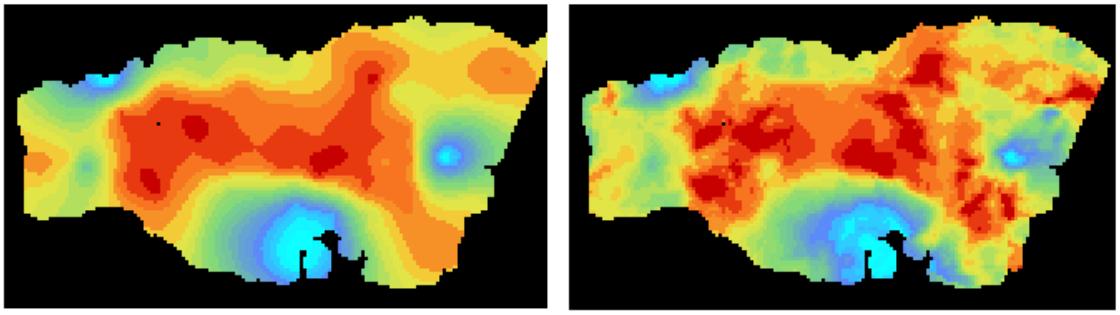


Figure 24. Estimation maps of the primary variable: Kriging (left) and Co-Kriging using the covariate (right)

3.4.5. External Drift Concept

We may sometimes consider that the target variable, only measured at a limited number of data points, is correlated with a co-variable which is known exhaustively. Then the co-variable is used more extensively and provides the shape factor to the variable of interest: this technique is known as the *Kriging with External Drift*.

3.5. Simulations

Given the set of measurements of the variable of interest and its associated Model, we recall that Kriging produces two results: the estimation on one hand and the standard deviation of estimation error on the other hand. At each target site, the first one gives the most probable value (given the data and the Model) whereas the second one provides the range of plausible values for reality. Nevertheless, this information is not sufficient for global (non-linear) criteria such as the probability that a pollutant concentration exceeds a given threshold for example.

For such problems, we are not looking for optimality anymore (which leads to results which are systematically too smooth to be compared to reality). Instead we wish to produce outcomes which have the same spatial behavior as reality. This refers to another technique called *geostatistical simulations* which produce a series of equiprobable maps, each one of them reproducing the input Model and honoring the data.

3.5.1. Several Realizations

The principle of the Simulations is to produce a large series of outcomes and to perform the calculation separately on each one of them. As all outcomes have the same likelihood, all the resulting values are equally possible.

The following demonstrative example has been carried on based on the Yeu Island (western coast of France). 40 measurements provide the elevation of the sea floor along 8 bathymetric profiles. Note that all elevations are negative (below sea level). The problem is to guess the presence of a possible island and its possible surface. The true island measures 23.32km^2.

Figure 25. Yeu Island

The first attempt is to estimate the elevation using Kriging retaining the positive results as the island: as expected the result is biased and the estimated island tends to have a too small surface (22.94km^2).

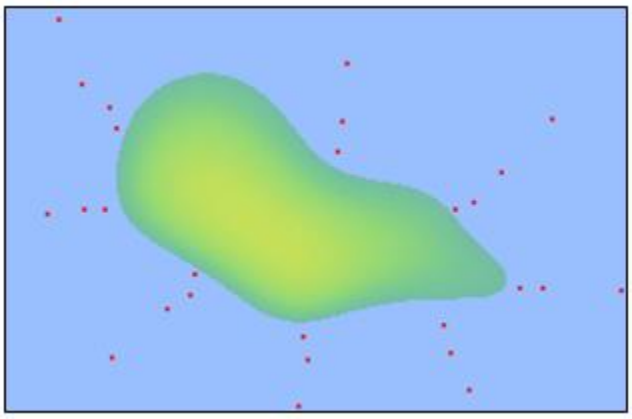


Figure 26. Kriging estimate with data along bathymetric profiles

Instead we perform a large set of simulations (some of them are represented here) and derive the surface for each outcome.

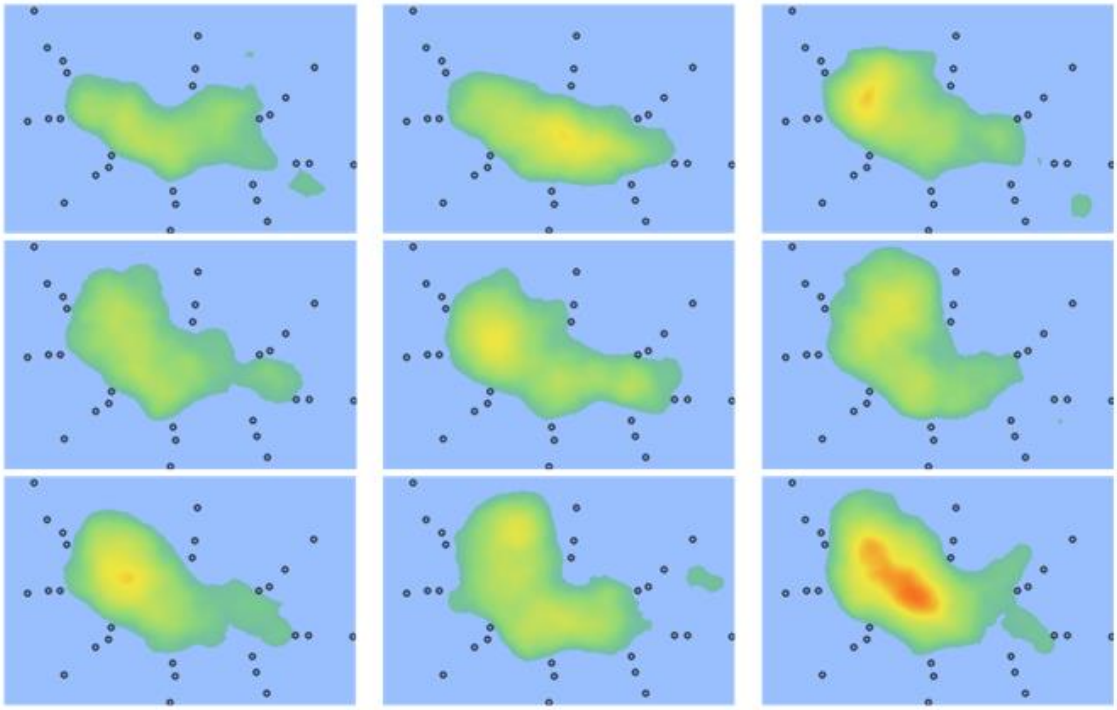


Figure 27. Few simulation outcomes

3.5.2. Risk Curve and Probabilities

The results are finally represented as a risk curve, which gives the range of possible surfaces: the average value of these surfaces is 23.17km^2.

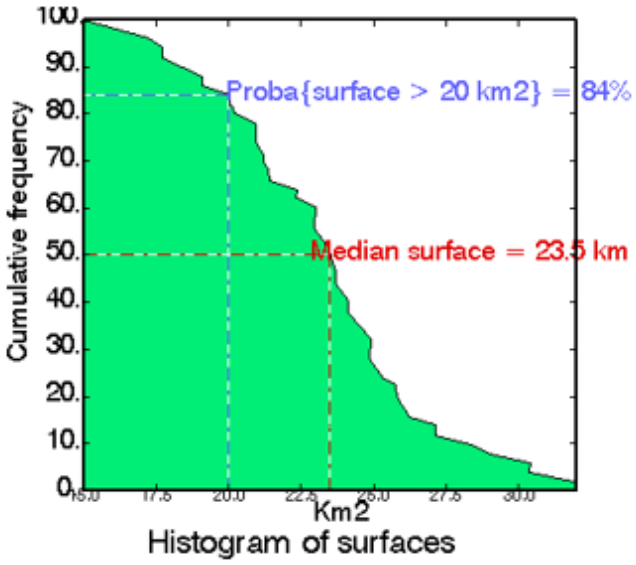


Figure 28. Risk Curve for Surfaces of Yeu Island.
It gives the probability that the actual surface exceeds a threshold surface value

These simulation outcomes can also be exploited in order to provide, at each target point, the probability that it belongs to the island.

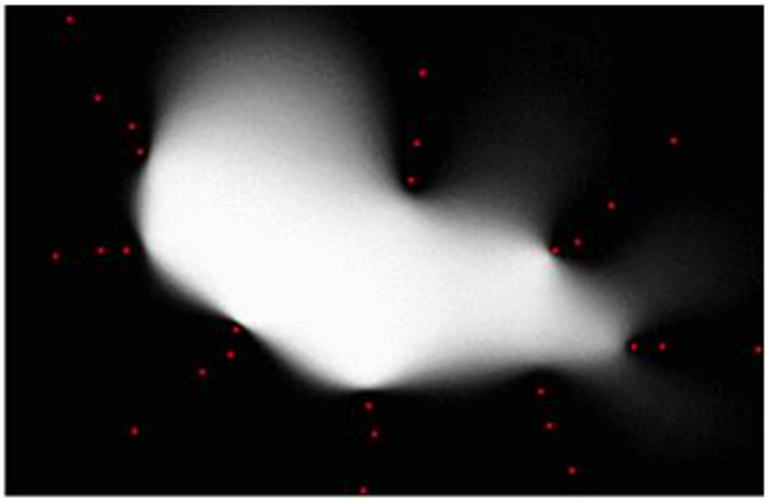


Figure 29. Probability map of the island.
The lighter the color, the higher the probability that the area belongs to the actual island.

3.6. Some previous use of Geostatistics in INTAROS application fields

3.6.1. Use cases and articles

Here are some previous use cases and associated publications of the Geostatistics Team from the Ecole des Mines de Paris in the INTAROS application fields.

- In several domains of the geosciences (e.g., geodesy, paleo-magnetism, climatology, and oceanography), the methodology must be applied to phenomena defined on a global scale such as Earth. In the following paper, a new innovative technique is proposed for applying traditional geostatistical methods, such as estimation and simulations, on spheres:

Christian Lantuéjoul, Xavier Freulon, Didier Renard. Spectral Simulation of Isotropic Gaussian Random Fields on a Sphere, Mathematical Geosciences, Springer Verlag, 2019, 51 (8), pp.999-1020. (10.1007/s11004-019-09799-4)

- This paper presents a novel application of the geostatistical multivariate method known as min–max autocorrelation factors (MAFs) for analyzing fisheries survey data in a space–time context. The method was used to map the essential fish habitats and to evaluate the variability in time of their occupancy:

  Pierre Petitgas, Didier Renard, Nicolas Desassis, Martin Huret, Jean-Baptiste Romagnan et al. Analysing Temporal Variability in Spatial Distributions Using Min–Max Autocorrelation Factors: Sardine Eggs in the Bay of Biscay, Mathematical Geosciences, Springer Verlag, 2020, (10.1007/s11004-019-09845-1)

- This paper presents a first implementation of a particle filter into a hydro-biogeochemical model for metabolism's parameter estimation. The assimilation of a 15-min “observation” dissolved oxygen data has been realized in the Seine River system on a synthetic case study:

  Shuaitao Wang, Nicolas Flipo, Thomas Romary (2019). Oxygen data assimilation for estimating micro-organism communities’ parameters in river systems, Water Research, Volume 165, 2019, 115021, ISSN 0043-1354,
  https://doi.org/10.1016/j.watres.2019.115021.

- In this paper, we proposed a methodology to optimize sampling schemes of static sensors networks. This methodology is based on the construction of a relevant objective function and its optimization. Although it has been developed for air quality problems, it is general and can be adapted and applied to various natural Arctic phenomena:

  Romary, T., Malherbe, L., & De Fouquet, C. (2014). Optimal spatial design for air quality measurement surveys. Environmetrics, 25(1), 16-28.

- This paper shows how geostatistical non-linear indicator tools can be used to define hotspots in fisheries ecology as well as how inter-annual variability can be handled with multivariate Geostatistics. Variograms and cross-variograms of indicators were used to estimate transition probabilities, which allowed to define hotspots in relative terms as the areas within which higher values occurred unpredictably:

  Petitgas P, Woillez M, Doray M, Rivoirard J (2016). A Geostatistical Definition of Hotspots for Fish Spatial Distributions. Mathematical Geosciences, Springer Verlag.

3.6.2. Handbook for fisheries and marine ecology

Several studies have been made for fisheries and marine ecology using Geostatistics through the use of the RGeostats package. A handbook has been published (2017) for helping scientists in using geostatistical methodologies for these application fields:

Petitgas, P., Woillez, M., Rivoirard, J., Renard, D., and Bez, N. 2017. Handbook of geo-statistics in R

for fisheries and marine ecology. ICES Cooperative Research Report No. 338. 177 pp. Download here: https://archimer.ifremer.fr/doc/00585/69732/67621.pdf.

The following list give an overall idea of several specific topics that have been addressed, such as:

- Demersal surveys are a typical case of sampling schemes without preferential directions. They usually follow a stratified random sampling protocol so that samples are uniformly distributed in each large stratum. A specific variogram calculation is used.
- In acoustic surveys, the sampling has usually high resolution data along parallel and regularly spaced transects (ship's sailing tracks separated by tens of nautical miles). Variograms have been computed along and across transects with different lag distances to check for structural anisotropy in the fish distribution. The (global) estimation of population abundance can be performed in one dimension. It suffices to sum fish concentrations along the transect lines and work on the one-dimensional dataset made of fish biomass per transect (Petitgas, 1993a). This technique has been applied successfully on anchovy surveys (Engraulis encrasicolus) in the Bay of Biscay.
- Transitive covariograms have been computed for studying the cephalopod concentrations in 2D. Cephalopod surveys were carried out by INRH (Institut National de Recherche Halieutique) – Casablanca – Morocco (Faraj and Bez, 2007). The data corresponds to a regular stratified sampling where one sample was taken at random in each square of a 11 x 11 nautical mile regular grid.
- The estimation variance of the mean over a domain V has been used for analyzing the concentration of herring (Clupea harengus) eggs over a spawning bed. The survey design is made of dredge hauls dispersed more or less evenly over the spawning bed. Kriging estimation for mapping has been considered with different neighborhood configurations. Two criteria have been studied: the weight of the mean, and the decrease in kriging weights with distance from the target point to be kriged. Multivariate Geostatistics, which permits studying the relationships between different regionalized variables, has been used on herring mean length and bottom depth collected at the same (trawl) stations around Shetland.

4. iAOS showcase applications

4.1. Showcase with Task 6.2 Improved Ecosystem understanding and management

The aim of this iAOS Showcase application is to generate estimation and error maps of the temperature and of the salinity in the Barents sea from temporal and spatial data measured along vertical profiles in the North Sea (CTDs coming from IMR). It is described in the INTAROS D5.6 report and no new Geostatistical analysis has been conducted since the last results. It will not be duplicated in this document.

Nevertheless, we know that the results can still be improved by building a unique 4D (3D + Time) temperature variogram model which will be valid for the whole data time interval and the whole data bounding box. Currently, for a given sub-area, a given time interval and a given depth, the variogram model is automatically fitted by using CTDs measurements that belong to the selection. That means that some severe discontinuities may show up when tiling all generated maps (horizontally, vertically or through time).

Building a global unique 4D model for temperature would require a deeper Geostatistical analysis and would require addressing the following topics:

- Modeling vertical Temperature and Salinity profile evolution by area (using Functional Geostatistics)
- Analyzing seasonal variability (climate, currents, ...)
- Modeling of a global spatial trend and local non-stationarities
- Finding local or zonal anisotropies
- Considering the distance to coast

4.2. Showcase application with Task 6.4 Natural Hazards in the Arctic

4.2.1. Main objectives

This showcase has been proposed by Roberta Pirarzzini (FMI) to ARMINES during the INTAROS General Assembly in Sopot (2019). This case study is meant to forecast the Snow Depth on a sub area of Svalbard Island in order to feed the avalanche forecast model.

The information available is:

- The Topography map of the area.
- The information provided by a few weather stations. This information is quite different depending on the type of the station. They measure several variables (such as the Temperature, the Snow Depth (few of them actually do), the Wind Speed and Direction…) almost continuously over time since 2011 (for the oldest ones).

- The maps of some attributes (such as Snow Depth, Wind Speed and Direction) which stand as indirect data resulting from the AROME Arctic model: https://www.met.no/en/projects/The-weather-model-AROME-Arctic/about.
  This information is provided on large cell grids which cover the whole area. One map is provided every 6 hours.

4.2.2. Some deeper insight on the data

The global setup is expressed in the following map which shows the global geography as well as the Isfjorden - the area of interest (blue rectangle) and the various weather stations (the SYNOP ones are in red).

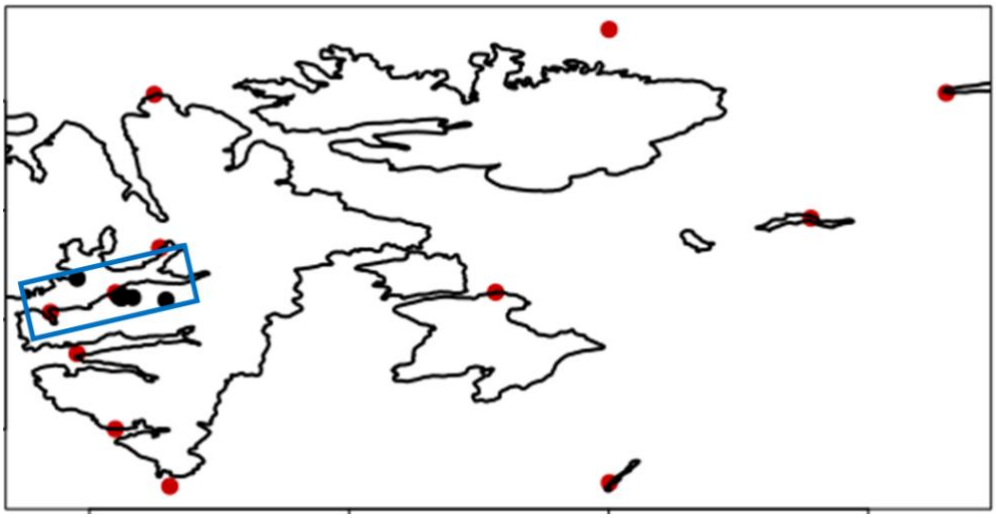


Figure 30. Svalbard showcase study area

The Topography map is provided next. It has been restricted to the grid that is provided by the AROME Arctic model. This is the grid where the snow precipitation will be provided ultimately.

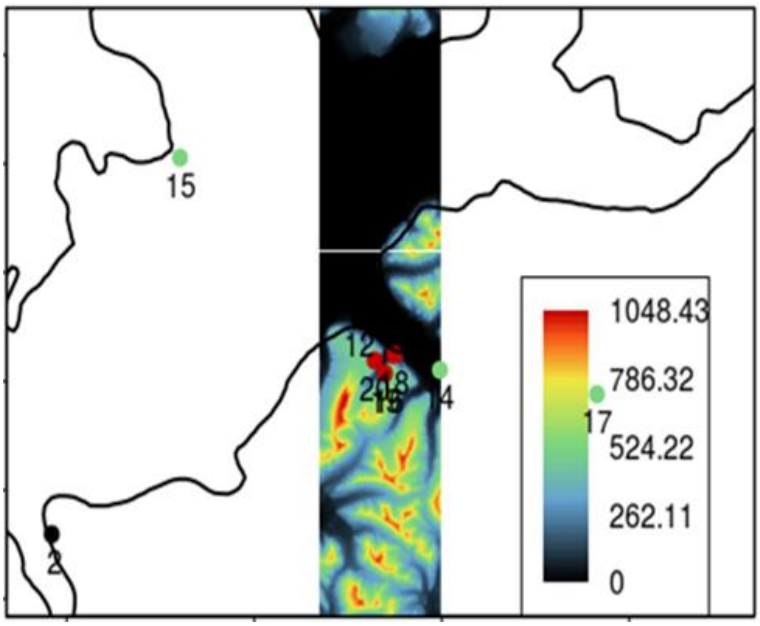


Figure 31. Topography map of Isfjorden, near Longyearbyen.

The next maps show some outcomes of the AROME Arctic model. It helps in understanding the dimensions of the mesh: Temperature (top left), Snow Depth (top right), Wind Speed (bottom left) and Wind Direction (bottom right) for November 1st of 2017.

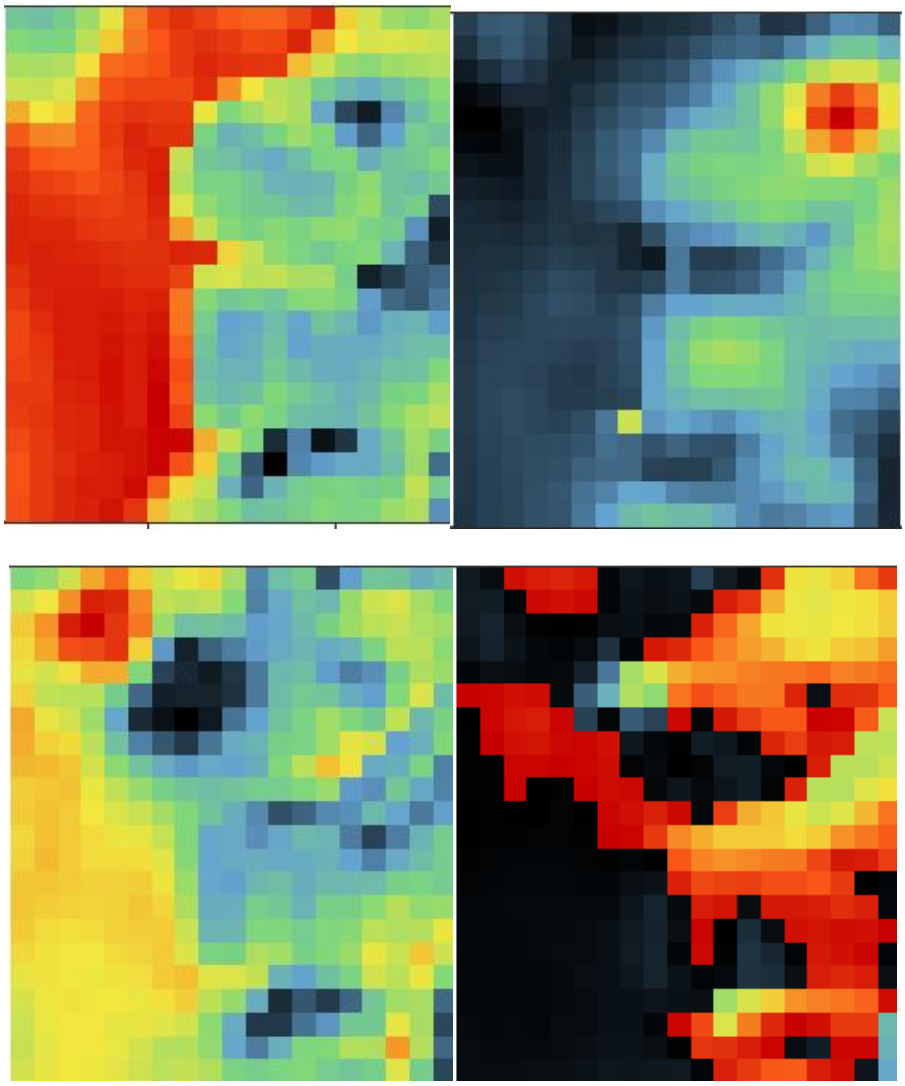


Figure 32. AROME Arctic model outputs

Similarly, we can check the evolution of these variables over time. The next figure shows the evolution of the Temperature map for the 4 time slices for a given day, i.e. T0 (top left), T6 (top right), T12 (bottom left) and T18 (bottom right).

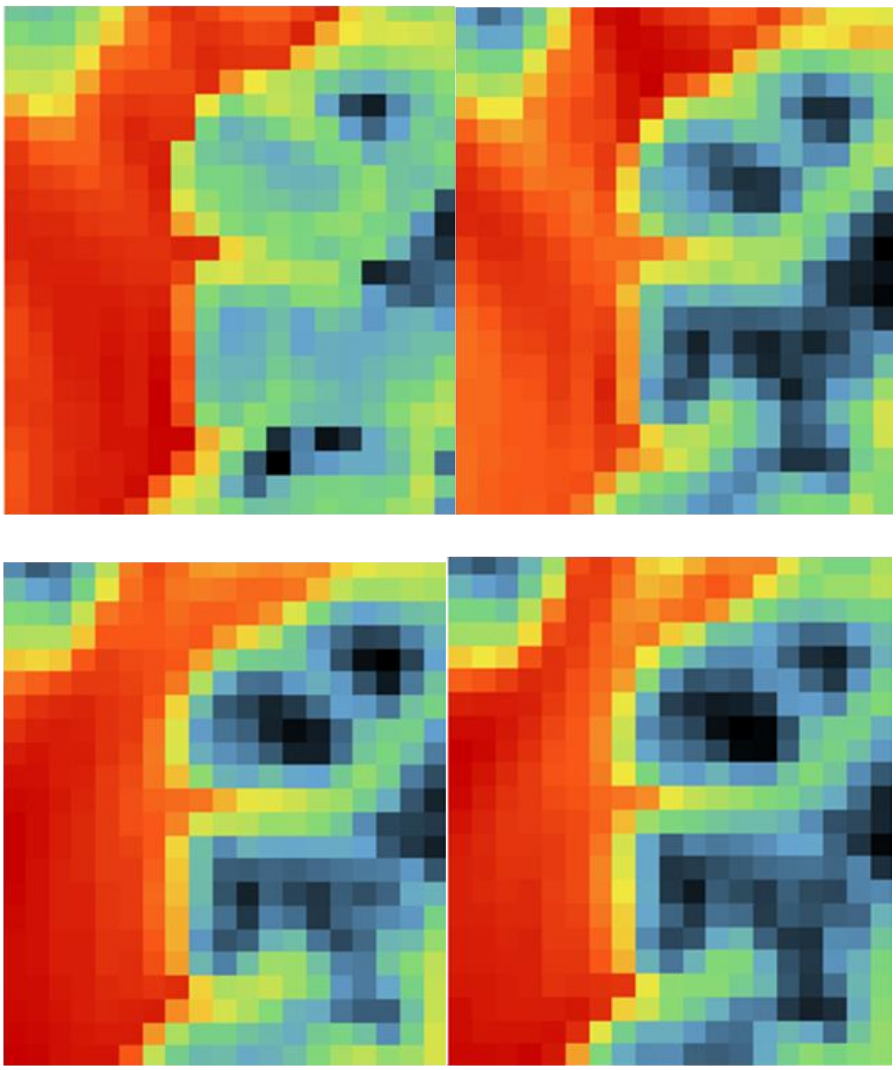


Figure 33. AROME Arctic Temperature maps along one day

Their resemblance shows that, using them among the set of covariables, will naturally ensure the continuity over time.

The measurements at weather stations are sampled almost continuously along the time dimension. Due to their scarcity (only 6 in the Fjord area), we consider all of them with no discrimination regarding neither their location nor their data quality (which is not known at our knowledge). Out of almost 1,000,000 measurements, 918,452 have recorded air Temperature, 899,145 Wind information but only 27,344 provide Snow Depth information. Moreover, this last information needs some clarification as the thickness varies from -2 to 192 and a negative thickness does not make sense.

The histograms of the different variables show large differences: Temperature (left), Wind Speed (center) and Snow Thickness (right).

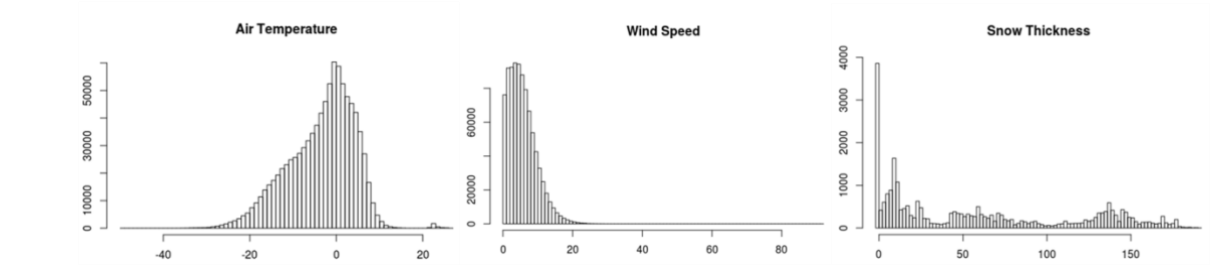


Figure 34. Histograms of weather stations data

It is also interesting to view this information as time series. The next figure represents the Snow Depth measurement recorded from 2013 to 2020, along the 6 stations where this information is recorded (one color per station).

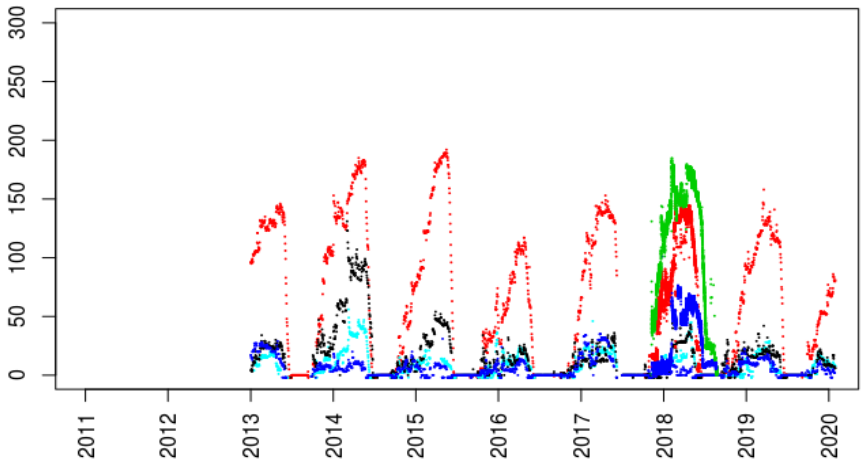


Figure 35. Weather station Snow Depth measurements number through time

We obviously notice that the different time series present the same seasonal cycles but with very different amplitudes. We can also see that not all the stations produce data along these 7 years; year 2018 is the one where most stations are active.

On the other hand, the time series for the Temperature present a very different behavior: the cyclicity representing the seasonal effect is present but the amplitudes of the various time series are similar (up to a band width of around 5 degrees).

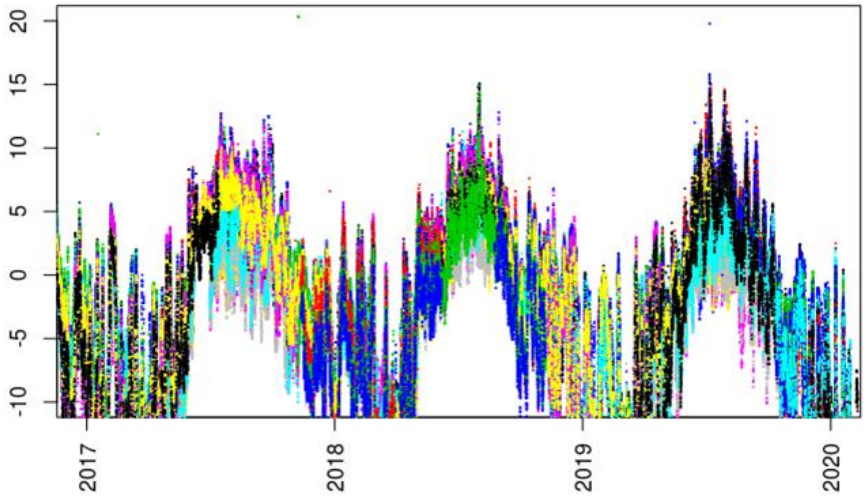


Figure 36. Temperature time series at the same weather stations

The scatter plot diagrams have been established to measure the relationship between any variable and the snow depth: i.e. Snow vs. Temperature (left) and Snow vs. Wind Speed (right).

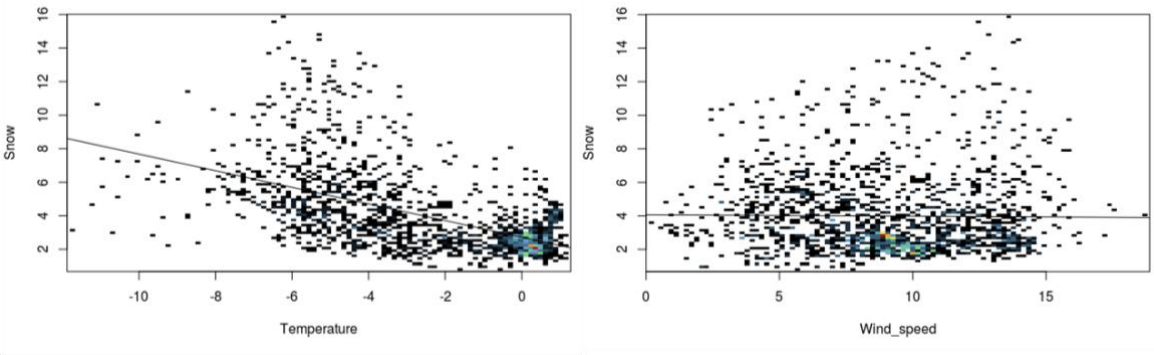


Figure 37. Weather station scatter plots

The dependency is weak which does not allow a simple conversion of the abundant Temperature information into the sparse Snow Depth information using the linear relationship for example.

4.2.3. Next steps

The idea is to provide maps on a fine grid (as the Topography grid) of the Snow precipitation. This map must be conditioned by the information measured at the stations considered as punctual. Due to the very small amount of punctual information, this is clearly insufficient to carry the lateral continuity of the variable throughout the whole fjord area. For this reason, it makes sense to consider the Snow Thickness information provided by the AROME Arctic model as a covariable which will enhance the global knowledge. Moreover, we can also assume that this gridded Snow Depth information accounts for the Topography. It still needs to be demonstrated if the Wind should still be considered as another relevant covariable or not.

But both sources of information are not provided with the same support:

- Weather stations provide a punctual information in time and a continuous information along time,

- AROME Arctic model maps provide information at the cell size of the coarse grid, at regular intervals along time. Intuitively, the value provided at a grid node can be considered as the average of the snow depth over the surface covered by the cell.

The AROME information will be used in order to provide the spatial characteristics of the Snow Depth averaged over the grid cell. Therefore, its variogram (see section 3.3) is linked to the punctual variogram of the Snow Depth variable, which is needed to perform small scale estimation. There will remain an indetermination on the small scale variation of the Snow Depth variable, i.e. from distance 0 to the distance equal to the cell dimension.

To overcome this difficulty, we unfortunately cannot use the actual weather stations as they are not located close enough. It seems difficult to convert the dense information available along time at each weather station, into a lateral (or spatial) continuity. Instead, one can think of using the spatial characteristics of the Topography (up to a scaling factor).

Finally, the weather station information is provided almost continuously along time. Nevertheless, the already gridded information is only available every 6 hours. So it seems reasonable to consider that the resulting Snow Depth maps will be provided 4 times a day and that the information at the weather stations will be integrated beforehand over 6 hours periods.

At this stage, and due to the amount of additional work needed to implement the new methodology, the maps are not produced yet.

4.3. Showcase Application with Task 6.8 Demonstrations for fisheries and environmental management agencies

4.3.1. Showcase overview

This iAOS Showcase application has been proposed by Mikael Sejr (Aarhus Univ.) to ARMINES during the INTAROS General Assembly in Sopot (2019). The aim of this showcase is to study the bottom ocean temperature (T°) in the West Arctic Ocean off Greenland by using both CTDs and Trawl measurements. Several objectives have been raised:

1. Generate maps over the whole study area that show average bottom T°
2. Generate maps of average linear trends of bottom T°
3. Produce graphs for average bottom T° over time for large Arctic sub-area

Some secondary objectives have been identified:

1. Identify where are the data gaps in order to guide the future sampling policy
2. Address the presence of fishes which live at the ocean bottom by looking for their correlation to T°

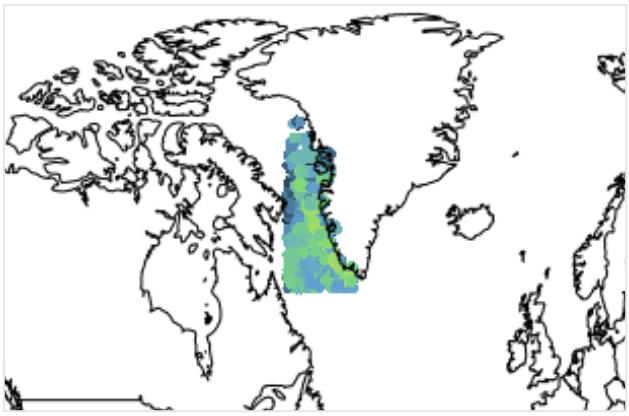


Figure 38. Overview of available measurements in the studied area

The study area bounding box is the following: Longitude: [-61°E, -46°E] Latitude: [58°N, 74°N]. Generated maps only have to cover the continental shelf area (more precisely, only where trawls data depth is shallower than 600m). One map per year of the average bottom T°C is generated. The output time interval must be as large as possible.

All input data files have been put in the following google drive:

https://drive.google.com/drive/folders/18CzDsIJCXIJtoLDI2gq7AMwZqJTOYHaj?usp=sharing

4.3.2. Data presentation and preparation

Several datasets were provided for this study:

4.3.2.1. ICES CTDs and Bottles

CTDs and Bottles are measurements obtained by lowering probes on a cable down to the seafloor from scientific vessels. For a given time and location, temperature and salinity measurements are distributed along one vertical profile.

Download from https://data.ices.dk

Data request parameters:

- Period: all data available
- Longitude Interval: [-61°E, -46°E]
- Latitude Interval: [58°N, 74°N]
- Type of data: CTDs and Temperature/Salinity (bottle)

Results:

We obtain 2 CSV files available in the google drive mentioned above, containing the following variables of interest: "yyyy*" (date), "Longitude*" (longitude), "Latitude*" (latitude), "Bot.*" (bottom depth in meter, should be close to the seafloor), "PRES*" (pressure in deci-bar assimilated to the measure depth in meter), "TEMP*" (temperature in °C) and "PSAL*" (salinity in g/kg)

- CTDs: 0428172c_CTD.csv (Number of clean samples = 2890) [1977 - 2017]
- Bottles: 0428518b_Bottle.csv (Number of clean samples = 5493) [1960 - 2017]

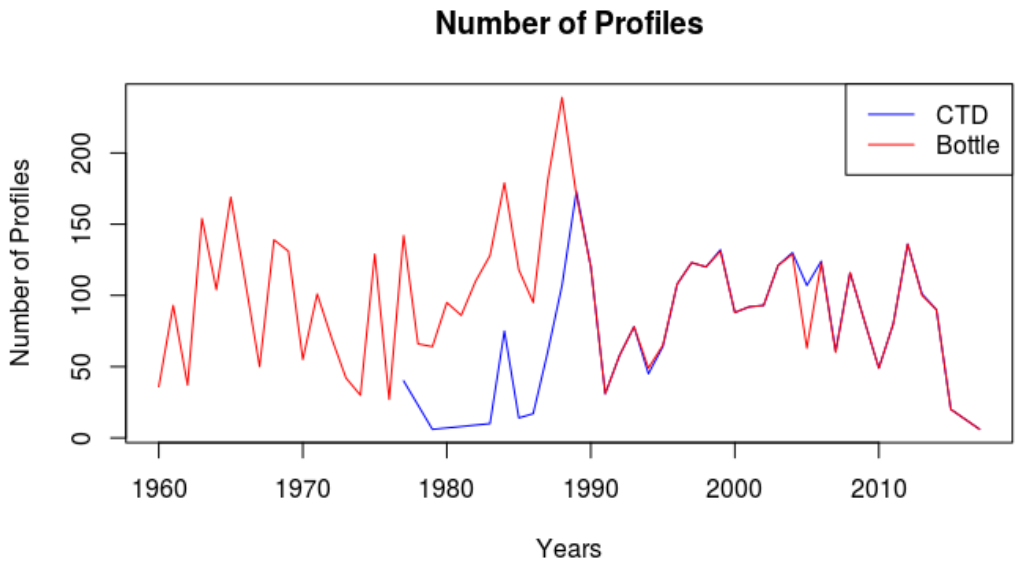


Figure 39. Number of ICES CTDs and Bottle profiles

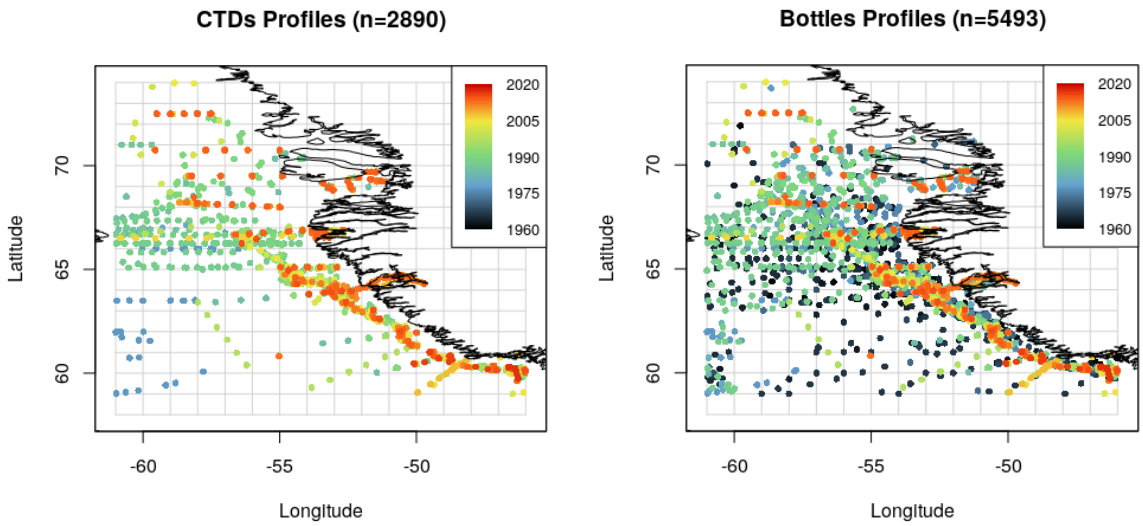


Figure 40. Basemap of ICES CTDs and Bottle profiles colored by year

Starting from 1990, we clearly see that Bottle and CTDs profiles are the same. Redundant information will be discarded (see section §4.3.2.5).

4.3.2.2. GINR Trawls

This dataset corresponds to a large number of trawl catches performed close to the seafloor. For each measurement done at a given location, time and depth, the sea water temperature and fish catch description have been recorded.

This dataset has been provided by GINR as an Excel file. In the ‘org’ sheet, several columns are available. The following variables of interest are retained: "TimeStart" (date), "lon" (longitude), "lat" (latitude), "FishingDepth2" (depth in meter), "BottomTemperature" (temperature in °C) and fish sizes and species for a later use.

The ‘org’ sheet has been exported to a CSV file available in the google drive.

Trawls: gin_fishtral2_v1.csv (Number of cleaned samples = 6421) [1995 - 2016]

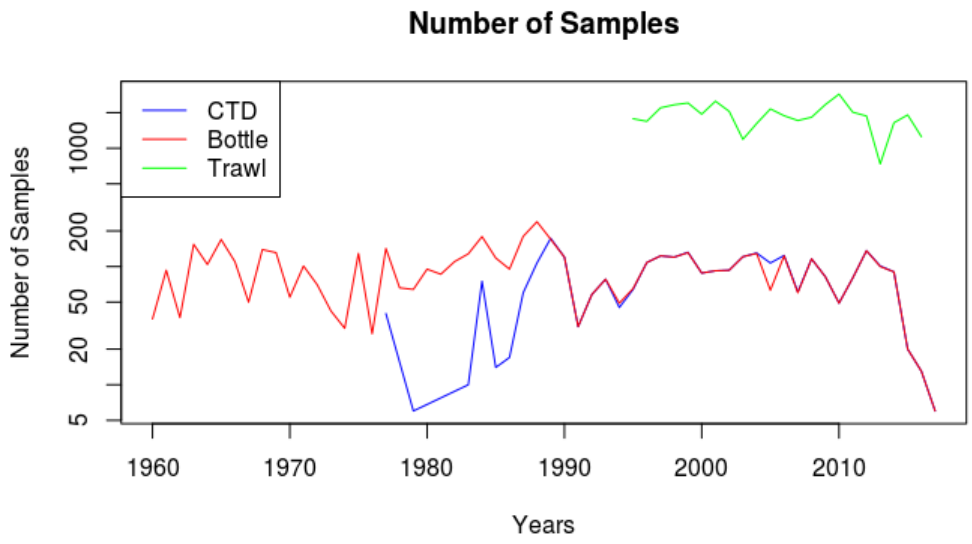


Figure 41. Number of trawls samples

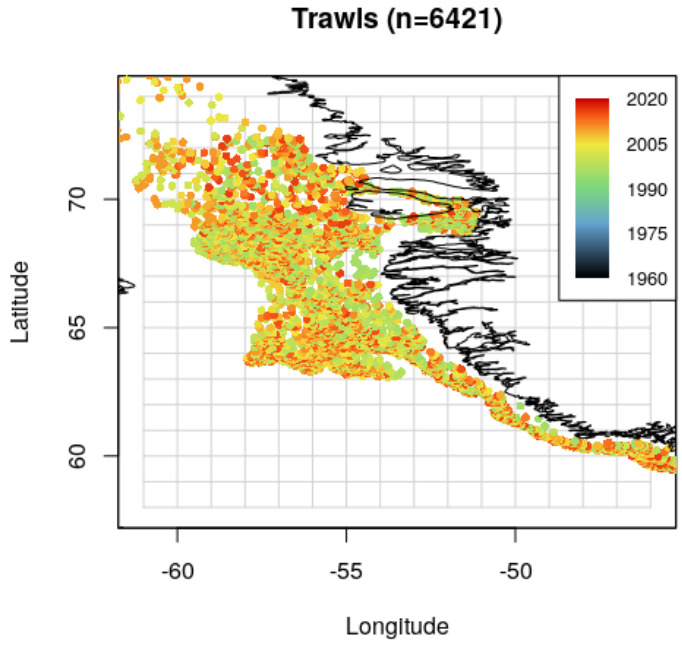


Figure 42. Basemap of Trawls colored by year

4.3.2.3. Filtering one measure per profile

This study only focuses on the seafloor water temperature. Hence, only the deepest sample of each CTD and Bottle profile is considered. The profile is accepted (in green) if the deepest sample is located at most at 25m from the sea bottom (for sea bottom shallower than 300m) and at most at 50m from the sea bottom (for sea bottom deeper 300m). The profiles with no deep enough measurements are ignored (in red).

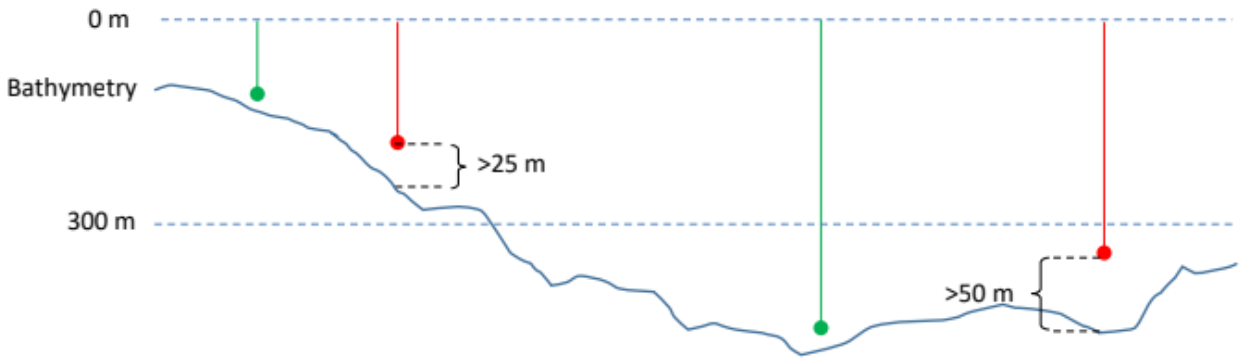


Figure 43. Profiles validation rules

4.3.2.4. GEBCO Bathymetry

In order to double check that trawls data have been collected close to the seafloor, a bathymetry dataset has been used. It has been downloaded from GEBCO website:

https://www.gebco.net/data_and_products/gridded_bathymetry_data

We obtain a gridded variable stored in a NetCDF file at a fine resolution (at 15 arc-second). Here is the bathymetry and the gridded study area overlayed.

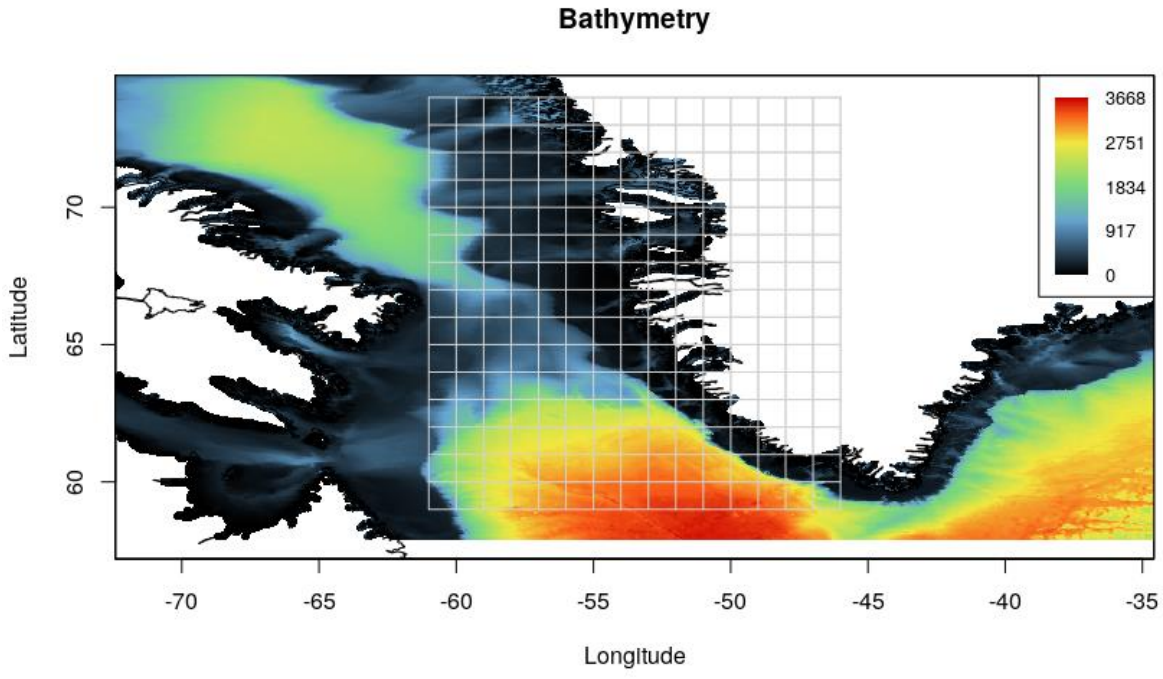


Figure 44. Bathymetry and output grid location

In the following cross-plot, we see that trawl depths are strongly correlated with bathymetry.

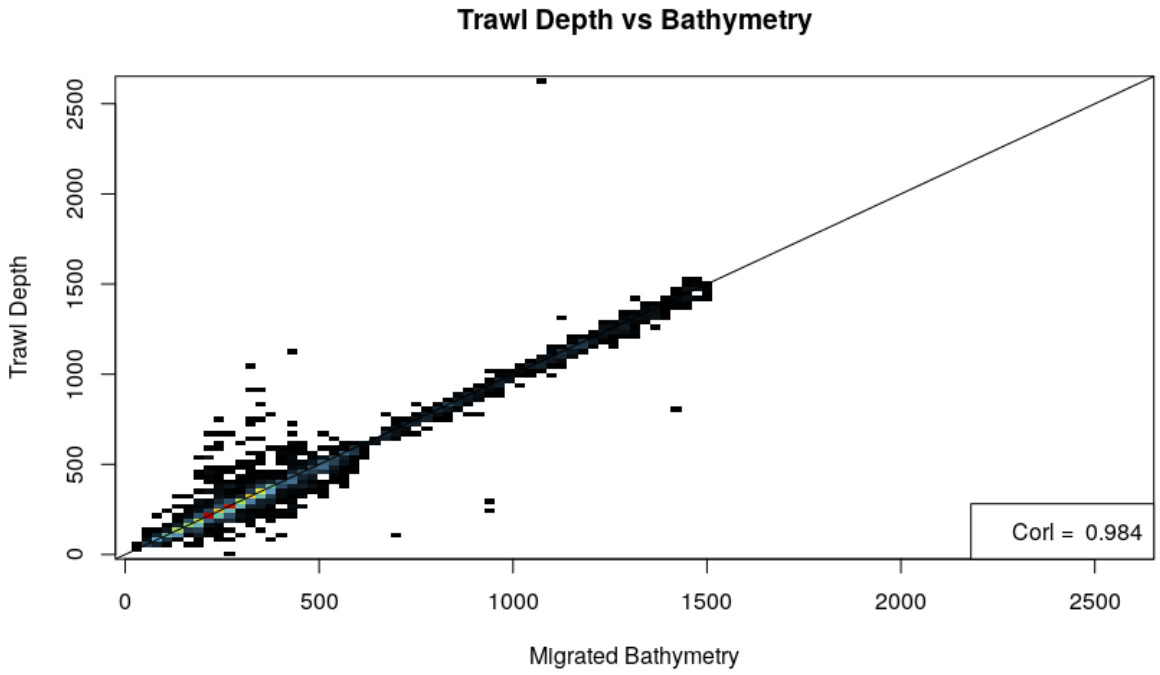


Figure 45. Scatter plot between trawl depths and bathymetry

4.3.2.5. Merging all 3 data sources

The three data sources having temperature measurements close to the seafloor, have been merged in only one RGeostats database. Duplicated information between CTDs and Bottles (in time and location) have been analyzed and removed. Because the Temperature and Salinity values are identical for all duplicates. Thus, it has been chosen to keep only CTDs samples in case of duplicates. Here is the basemap of the temperature variable (°C) for all the available data points.

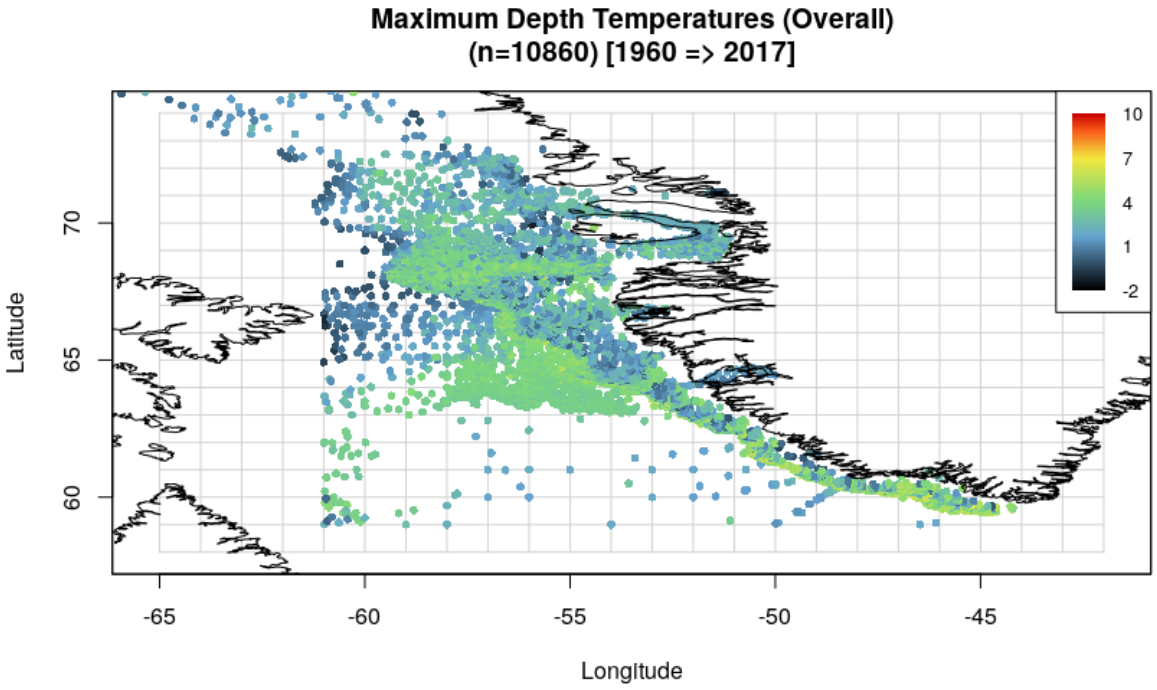


Figure 46. Basemap of all data available

4.3.2.6. Data selection on continental shelf:

Because of the discontinuity induced by the large depths encountered in the northern and southern basins, the study area (and corresponding data used) has been reduced to the continental shelf. A polygon has been digitized manually to build a selection variable named “Shelf” (red line below).

Here is the basemap for the selected Trawl data:

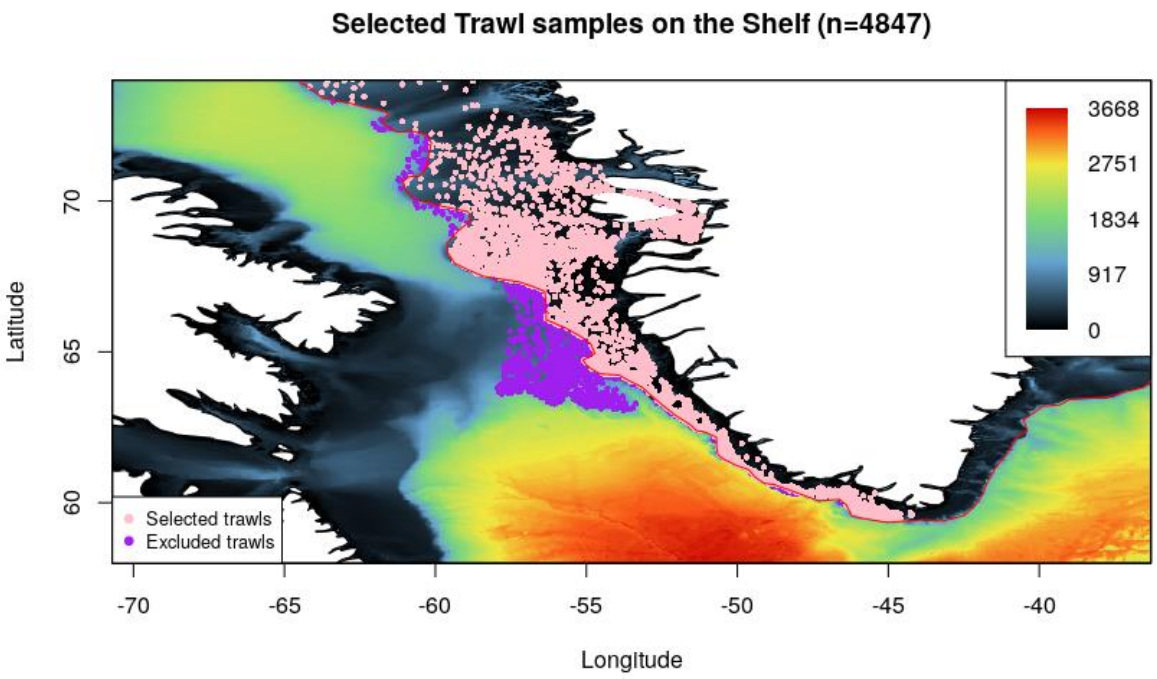


Figure 47. Shelf data selection vs depth selection

Here is the final input dataset that will be used for estimations.

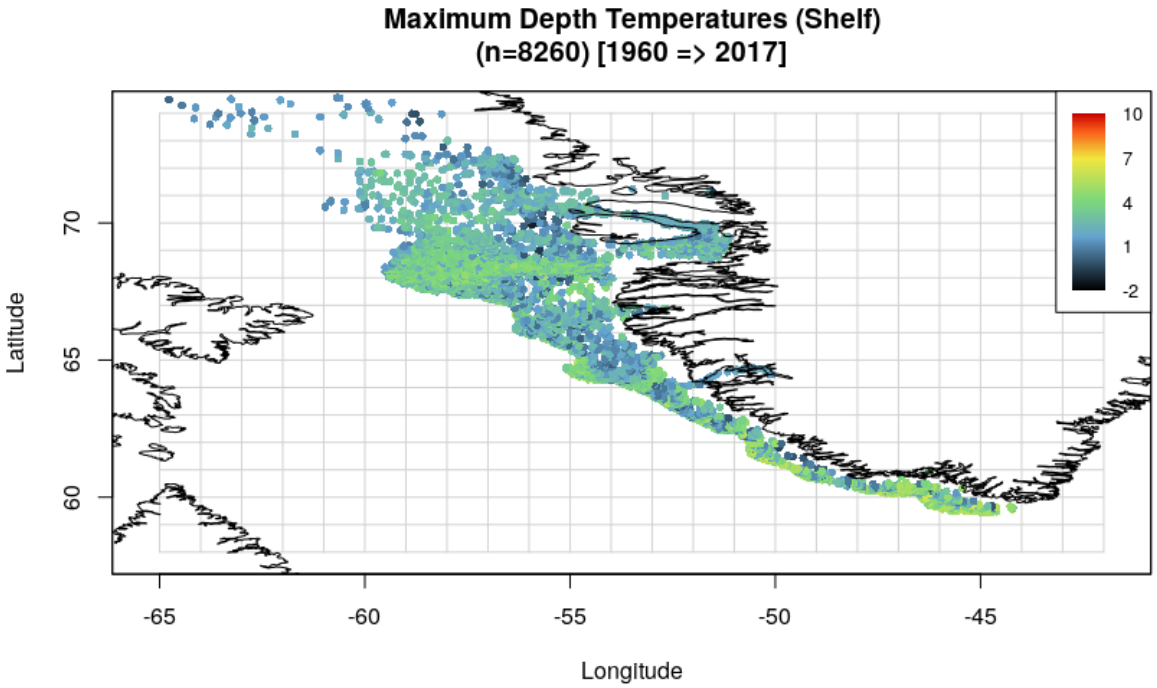


Figure 48. Data filtered on the Continental Shelf of South West Greenland

4.3.2.7. Data coordinates transformation

For computing the variograms, we need to calculate distances between samples. Thus, 2D coordinates must be homogeneous and comparable in all directions. That's why the Longitude / Latitude coordinates must be converted in a cartesian referential system. At first, a simple cartesian system, already implemented within Rgeostats, was used: the nautic miles coordinates representation. Longitude and Latitude have been converted using the following simple equations:

Xs = 60 * Longitude * 0.4
Ys = 60 * Latitude

A more appropriate coordinates system, adapted to the Arctic region (reducing distance distortion), will be used in the future (i.e. stereographic).

4.3.3. Bottom temperature estimation

4.3.3.1. Variogram map

A first 2D horizontal variogram map has been calculated in order to analyze the spatial behavior of the ocean bottom temperature variable (see section §3.3 for more details on variograms). This variogram map integrates all data from 1960 to 2010. The next figure measures the spatial variability of the bottom temperature, for all directions as a function of the distance between two samples (measured from the center). We can clearly see that the variability remains almost constant along the N160° direction: this North/South direction (up to a small tilt) coincides with the orientation of the West Greenland coast.

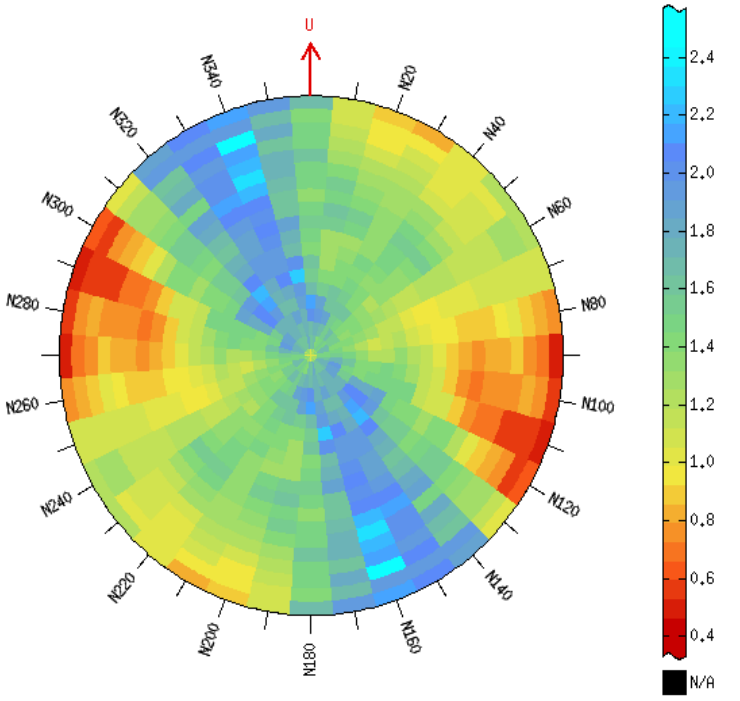


Figure 49. Bottom Temperature variogram map

4.3.3.2. Experimental variograms

Then, a 3D (2D + Time) experimental variogram has been computed from Shelf data. The next figure represents the 2D horizontal omnidirectional variogram (left), and the 1D variogram along time on the right. Plain line corresponds to the variogram model which has been automatically fitted to the experimental variograms. See next paragraph for details.

Here are the parameters for the experimental variogram calculation:

Direction 1 : N160		Direction 2 : D-90	
Height of the slicing	= 1 nmil		
Calculation lag	= 10 nmil	Calculation lag	= 1 year
Tolerance (perc. Of lag)	= 50.00 %	Tolerance (perc. Of lag)	= 50.00 %
Number of lags	= 19	Number of lags	= 17
Last Lag to be Refined	= 3		
Lag Subdivision	= 5 nmil		
Angular tolerance	= 90.000000	Angular tolerance	= 1.000000
Direction	= Azimuth=N160.00	Direction	= Vertical (Time)

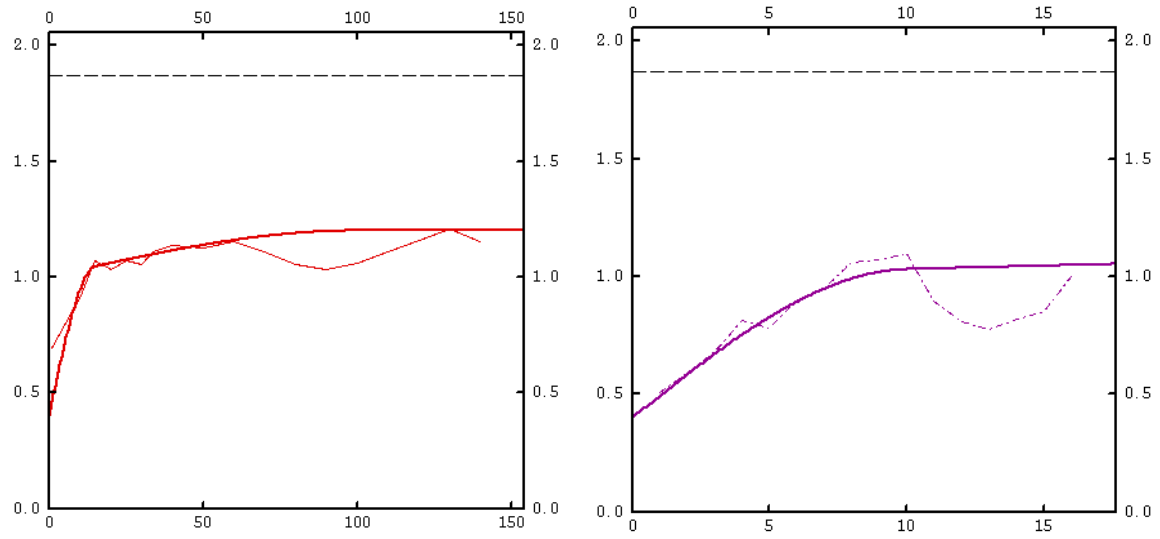


Figure 50. Bottom Temperature experimental variograms

The horizontal axis corresponds to the distance in nautical miles (nmil) (in red) and the distance in years (in purple). The black dashed line corresponds to the global temperature variance. What is important to notice is that the variability of the bottom temperature starts at 0.4 °C² for two closed samples (smallest possible distance in time or space). Then, the variability increases up to 1 °C² at a distance of 15 nmil in space and 10 years in time, and then stabilizes around a sill at 1.2 °C² at a distance of 100 nmil in space and 10 years in time.

4.3.3.3. Variogram model

The obtained variogram model is the following:

```
Model : Covariance part
=======================
Number of variables = 1
- Variable 1 : Temperature
Number of basic structures = 3
Global Rotation = Azimuth=N160.00 (Geologist Plane)

S1 : Nugget effect
Sill = 0.4
S2 : Spherical
Sill = 0.6
Ranges = 15 nmil, 15 nmil, 10 years
S3 : Spherical
Sill = 0.2
Ranges = 100 nmil, 100 nmil, 100 years
```

In the above parameters, we recognize:

- variance offset at the origin (the nugget effect with a sill to 0.4 °C²),
- small scale of 15 nmil / 10 years for the first spherical covariance (with a sill at 0.6 °C²)
- long scale 100 nmil / 100 years for the last spherical covariance (with a sill of 0.2 °C²)

4.3.3.4. Kriging of Temperature (2D + Time)

The variogram model presented above has been used in an Ordinary Kriging procedure, in order to obtain the bottom temperature block presented in the following figure. Bottom values correspond to older estimations (from 1960) whereas Top values correspond to the most recent estimations (up to 2018).

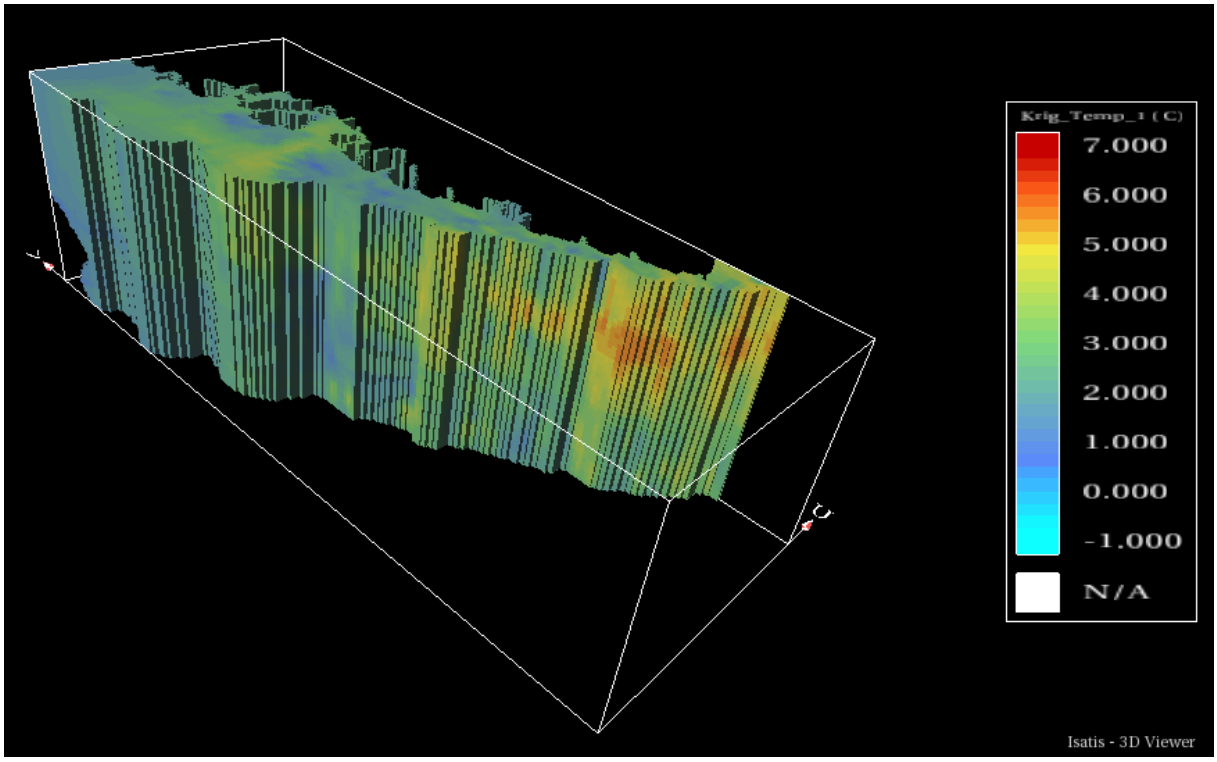


Figure 51. Bottom Temperature estimation by kriging – Continental shelf off South West Greenland

4.3.3.5. Mean annual Temperature (°C) estimated at the ocean bottom

In Figure 52, some slices are extracted from the previous block (shown in Fig. 50) for given years, and are represented sharing the same color scale. These maps show a global warming of the mean annual ocean bottom temperature through decades.

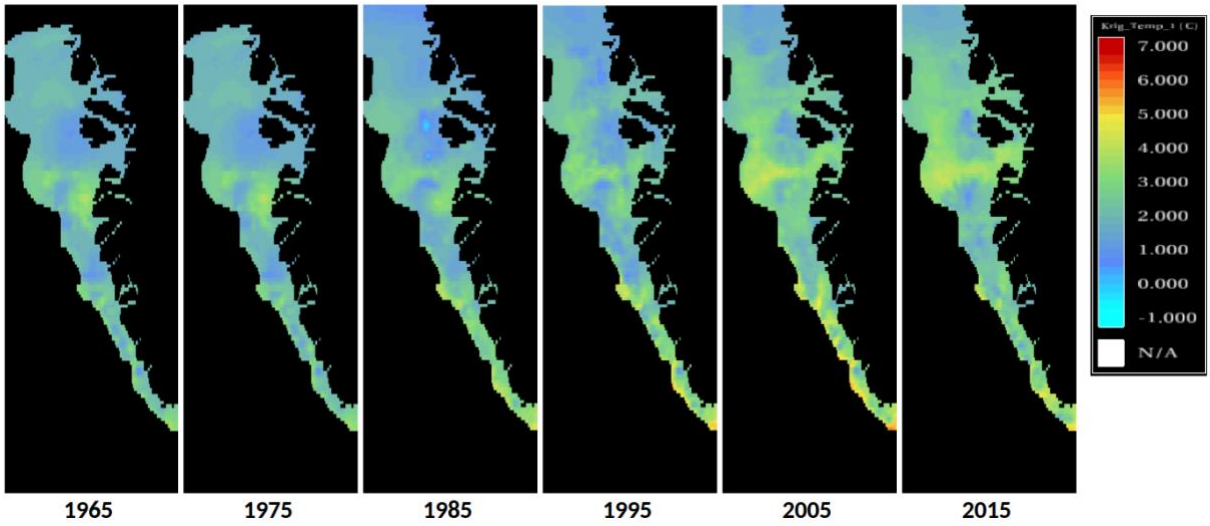


Figure 52. Extracted maps for Bottom Temperature estimation – South West Greenland

4.3.3.6. Standard Deviation of the Temperature estimation (°C)

Finally, we emphasize the estimation error (standard deviation), which is quite large in this first study (around 0.6°C on average).

The maps are rendered in nautical miles geographic coordinates, and the standard deviation values range from low values (blue colors in the legend) to rather high values for the northernmost estimations (red colors in the legend).

Note that such estimation error is logically important when the estimation is done in regions where there is a lack of measurements (which is the case for the first decades and northern regions here).

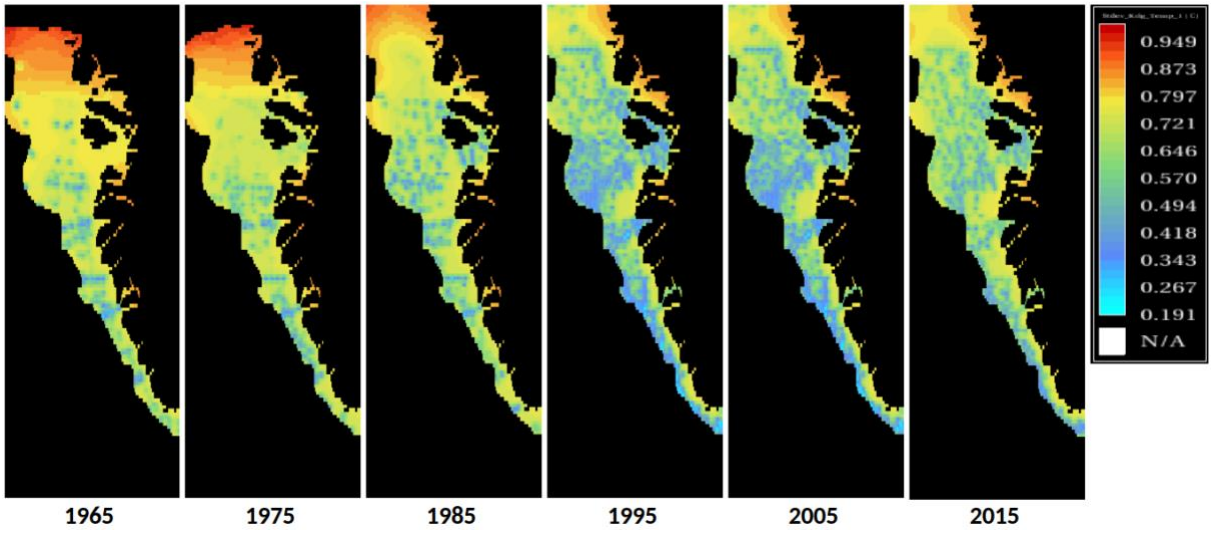


Figure 53. Extracted maps for Bottom Temperature estimation error – South West Greenland

5. Conclusion

The main Geostatistical tasks are available in RGeostats. The package RIntaros has been developed in order to offer the complementary tools adapted to the different case studies that have been addressed. Essentially, it contains the modules for reading data from the various data supports. It also provides a set of dedicated Geostatistical tools, which simplify the workflows for a non-expert user.

The different case studies which are presented in section §4 also demonstrate the versatility of RGeostats / RIntaros combination. These case studies usually refer to problems requiring some spatio-temporal models. They also show the necessity to deal with several sources of information, which may even correspond to different supports (sample size and repartition). The aim of these case studies is to find an adapted solution to make the best use of the different pieces of information.

After a deep exploratory data analysis of the IMR CTDs (see D5.6), the IMR case study workflow has been tuned to minimize the number of user input parameters customization (somehow automatic). It has been implemented into R scripts to be run as a “push button” procedure in a Web Processing Service of the Ellip Cloud platform. It is now possible to generate “on the fly” fields of the water temperature in the Barents Sea. This workflow output could be used for example, by INTAROS community in order to validate climate model projections. Curious people can have a deeper look to the RGeostats-Workshop github repository where the IMR workflow is described in detail (Jupyter notebooks).

On the other hand, the two other case-studies (Svalbard snow depth and bottom temperature off Greenland) necessitate more knowledges in geostatistics. Work produced for these two showcases demonstrates the capabilities of the Geostatistical Library v2 but no automatic procedure has been designed. There is no script delivered for these showcases as it remains important features to be implemented in the Geostatistical Library. However, the proof of concept is provided and shows that the two workflows could be used for similar problems with different data sources.

Regarding the showcase for bottom ocean temperature off Greenland, the main questions of our stakeholder have been addressed: Yes, the water bottom temperature in the Arctic has warmed through last decades and this is probably due to the global warming impact. From these results, further analysis can be done by finding correlations between interpolated temperatures and fish presence.

---------- END of DOCUMENT----------

INTAROS

This report is made under the project
Integrated Arctic Observation System (INTAROS)
funded by the European Commission Horizon 2020 program
Grant Agreement no. 727890.

Project partners: