On the optimal data processing of the Soil Moisture and Ocean Salinity measurements

A. García-Espriu [1,2], E. Olmedo [1,2], V. González-Gambau [1,2], C. González-Haro [1,2], A. Turiel [1,2]

(1) Institue of Marine Sicences (CSIC) P. Marítim 37-49, 08003 Barcelona, Spain
(2) Barcelona Expert Center (BEC), P. Marítim 37-49, 08003 Barcelona, Spain

Contact: ainagarcia@icm.csic.es

BEC Barcelona Expert Center

Motivation

With more than 12 years of SMOS data, the complexity and cost of the processing of the complete dataset increases over time. Not only the size of the data enlarges but also enhanced algorithms, that require more resources and processing time, have been proposed during those years.

Processing such amount of data is expensive regarding computational time, disk space and hardware resources. This requires an optimization of the full process as well as the application of some data science techniques.

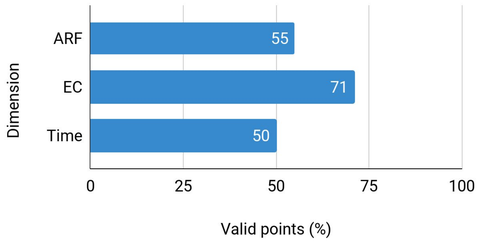
Work on a new salinity processor for SMOS data is presented here. This processor starts from L1B data and produces L2 and L3 Sea Surface Salinity (SSS) products. It is designed to apply data and resources optimizations as well as incorporating improved algorithms.

Extract, Transform, Load (ETL)

It is important to follow the *ETL methodology* from the start of the process

- **Extract:** Get data from source. In our case L1B products.
- **Transform:** data manipulation and data cleansing techniques
  - **Contextualizing data:** data is in Antenna Reference Frame (ARF), knowing its projection in Earth Coordinates (EC) is essential in order to apply SSS retrieval techniques
  - **Removal of invalid points:** we only keep EAF-FOV and OCEAN points. Time dimensionality is reduced by 50% by using half first Stokes (Tx+Ty)/2
- **Load:** load the reduced and contextualized dataset for further processing

Valid points (%) for each dimension



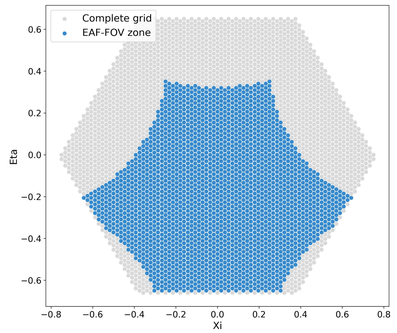
Data is reduced to a **19.5%** of the input size. Only this # of measurements will be processed.

Adding geolocation metadata results in a dataset that weighs **~60%** of the input one

Introduction to SMOS data

Input processor data [1]

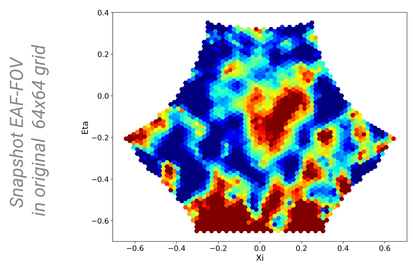
- 14.39 orbits per day
- One data snapshot every 1.2s
- 4096 (64x64) Brightness Temperatures (TB) per snapshot forming an hexagonal shape
- Full polarimetric data blocks: Tx, Ty, Txy
- The EAF-FOV zone is a subset of the complete grid which can be used to retrieve SSS.

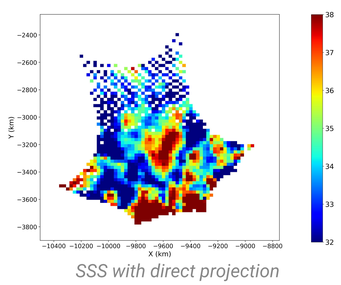


SMOS antena grid

Input data has total of 1.076e11 TB measures per year

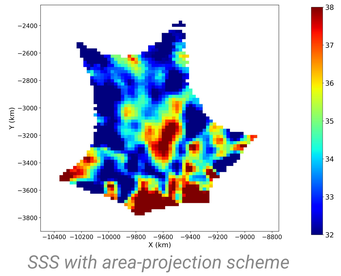
SSS Retrieval in xi-eta coordinates





SSS with direct projection

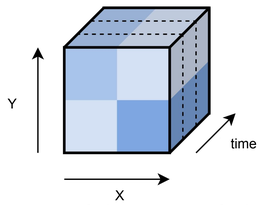
- Salinity retrieval is done in xi-eta coordinates to reduce the number of interpolations
- The hexagonal pixel is projected over the Earth with an enhanced area-preserving interpolation scheme, giving more homogeneous coverage
- Projecting hexagons is computationally demanding. We use a precomputed snapshot projection which is translated at any earth location



SSS with area-projection scheme

Acknowledgements: This work has been carried out as part of the SO-FRESH project (AO/1-10461/20/I-NB) and by means of the contract SMOS ESL L2OS both funded by the European Space Agency. It has also been supported in part by the Spanish R&D project INTERACT (PID2020-114623RB-C31), which is funded by MCIN/AEI/10.13039/501100011033. We also received funding from the Spanish government through the "Severo Ochoa Centre of Excellence" accreditation (CEX2019-000928-S). This work is a contribution to the CSIC Thematic Interdisciplinary Platform Teledetect.

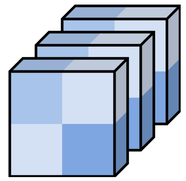
Data distribution techniques



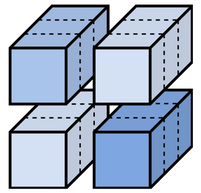
Our data has always **X**, **Y** and **time** dimension

- X dimension can be *xi*, *longitude* or *geoX*
- Y dimension can be *eta*, *latitude* or *geoY*
- Time dimension can be *snapshot time, semi-orbit* or *date*

Data at semi-orbit level



- Until the retrieval, data is distributed by time. Each semi-orbit can be processed independently
- Data is **not** referenced to a **common grid**



Aggregated maps generation

- Final product grid is set at this point
- Empirical SSS corrections are derived by using large temporal series
- Area projection enlarges the original dataset (x20)

Data is distributed in X,Y to have enough hardware resources to process the entire map

Summary and conclusions

	Current*	New	Current*	New
Geolocation	2.2 TB/y	0.5 TB/y	14.2 d/y	3.55 d/y
Inversion	1.4 TB/y	0.2 TB/y	3.55 d/y	2.13 d/y
Pre-projection	-	1.5 TB/y [T]	-	4.06 d/y
L3 maps	6 GB/y	6 GB/y	11.34 d/y	3.58 d/y
Total	3.6 TB/y	0.7 TB/y 2.2 TB/y [T]	29 days	13.3 days

Estimation computed using 1y of global data. Using 80GB RAM and 80 threads
**Comparison with respect to the current BEC processor*

54% time reduction
39% space reduction (with temporal [T] files)
80.5% space reduction (no temporal [T] files)

- Data rearrangement to enable temporal and spatial parallelism
- Multi-thread and multi-process parallelization. Scalable in multi-process
- Minimum hardware requirements reduced to 16GB RAM and 8 threads
- **Preliminary results show data quality and resolution improvements**

References:

[1] M-Brown et al. 2008. "SMOS: The Payload" IEEE Transactions on Geoscience and remote Sensing 1-2, 11