27-29 May 2024

International conference on Marine Data and Information Systems

SeaDataNet AISBL · Havforskningsinstituttet · Ifremer · MARIS · MEDIN · National Oceanography Centre · INGV · iMagine · ICDE · eosc Blue-Cloud2026

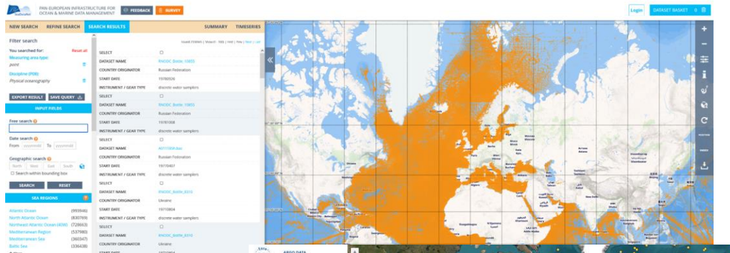
Beacon update

Next-Generation Data Lake and Subsetting Solution

Peter Thijsse, Robin Kooyman, Dick Schaap, Tjerk Krijger (all MARIS)

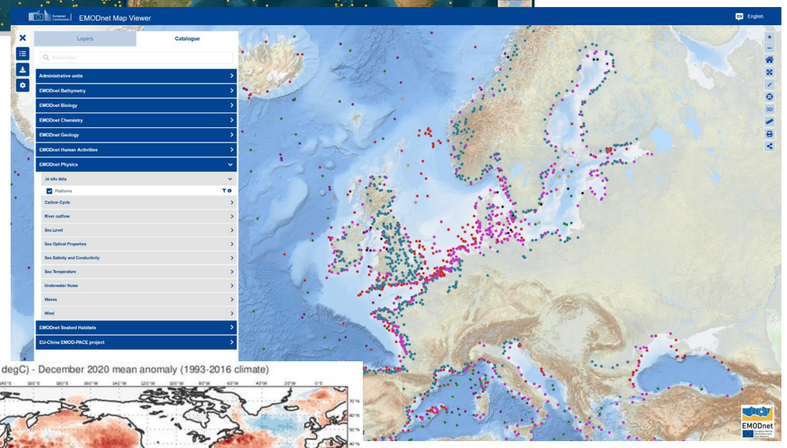
Challenge

- Data organised in millions of files (e.g. >2.8M for SeaDataNet) with metadata
- Human user interfaces exist to query metadata, select files, order and download.
- **But how to optimise systems like SeaDataNet CDI, ARGO, .. for Machine2Machine access to subsets, for access by VRE's, Jupyter Notebooks, by other applications?**
- **How to go from files to serving exactly data as needed, on the fly, in one file, as a true "Data lake" component?**

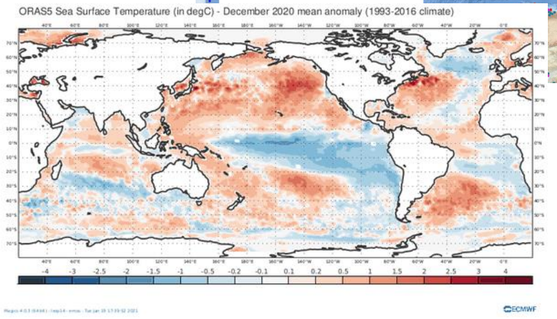


SeaDataNet CDI

Euro-Argo



Emodnet Physics



ORAS5

Example user query

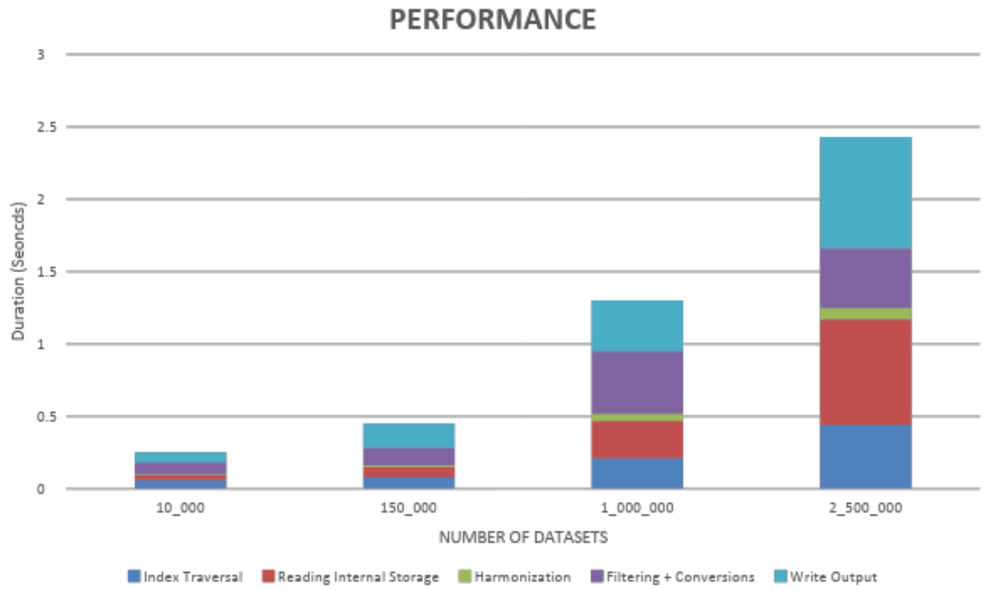
- Request:
  Give me all the temperature data in the North Sea, from 2010-2012, in degrees celsius, at a depth from 0-50 m.

- Response:
  One NetCDF file containing exactly this data. On the fly, directly usable in a Jupyter notebook and for HPC, with **all original metadata accessible**.

=> In these cases **Beacon** provides the solution

Beacon in a nutshell

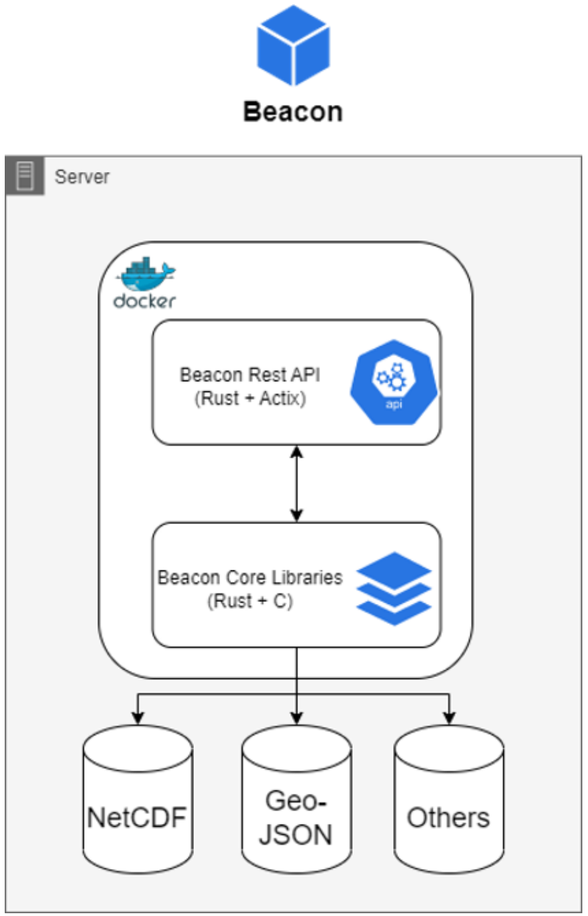
- High Performance Data Lake
- Written in Rust + C from scratch
- Easy to deploy (Docker)
- On the fly subsetting out of millions of datasets
- Harmonized single file output (Eg. 1 netcdf file)
- Powerful query capabilities
  - Filter on:
    - Ranges
    - Polygons
    - Metadata



Beacon performance for extracting dataset of approx. 10M datapoints (max), from increasing number of datasets, containing billions of datapoints.

How does Beacon work?

- Build indexes
- Reading only the necessary data
- Harmonization of files types
- Filtering, calls to the API
- Conversions of units, aggregation into parameters
- Writing output to various formats (a.o. NetCDF, Beacon binary format)



Beacon architecture

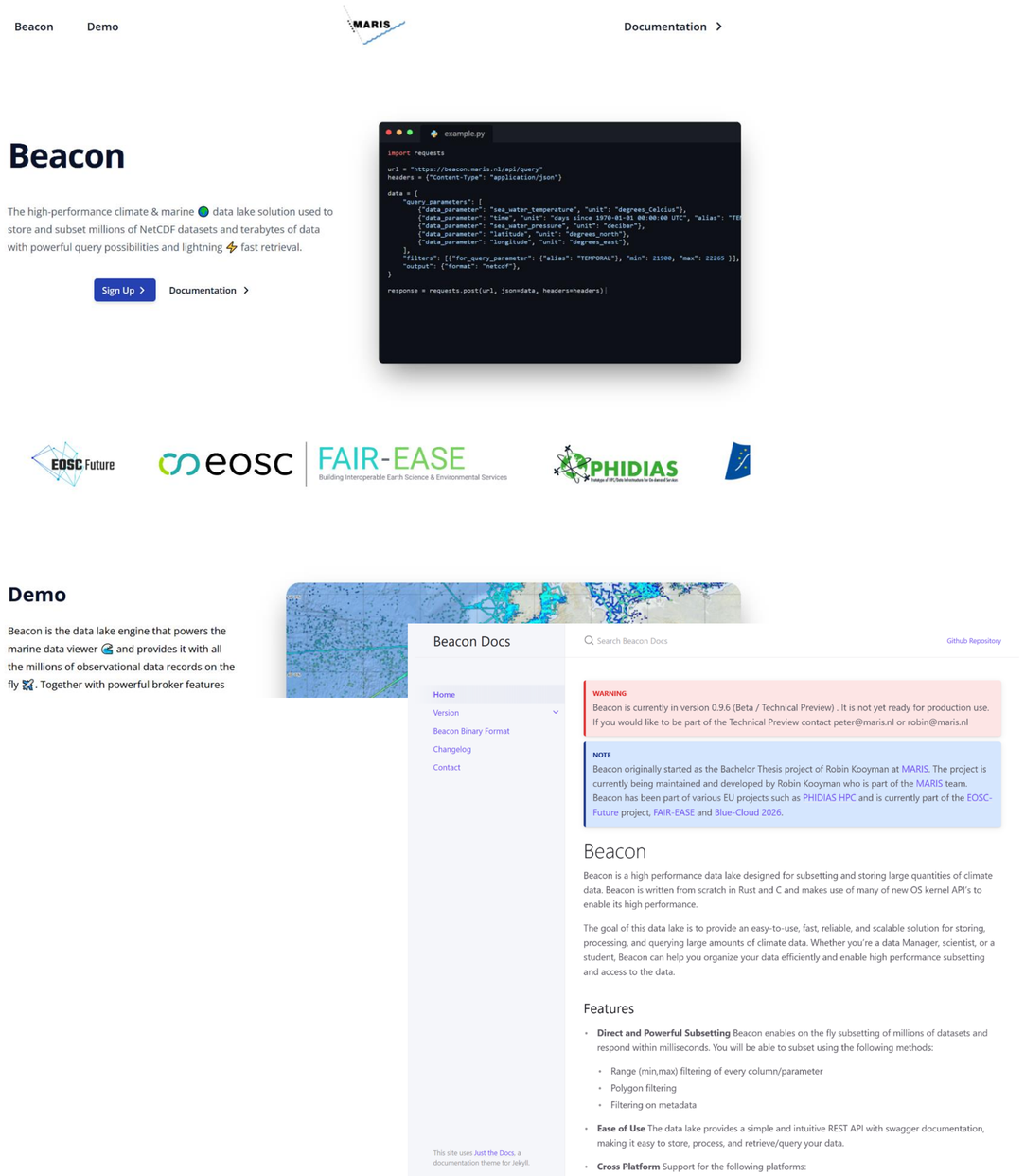
Current State

- Beacon **pre-**registration available at:
  - https://beacon.maris.nl/
  - Github available for community: sharing notebooks, python scripts, drivers

- Several Beacon deployments running for tests as part of BlueCloud2026
  - Argo
  - SeaDataNet CDI
  - ERA5/ORAS5 (sample set)
  - EMODnet Physics (operational)
  - ..

- Example notebooks created, and more will follow

- EOSC-Future demonstrator for subsetting CDI and ARGO, co-location with CMEMS



Usage example in Notebook: Creating an ARGO product in Pluto

Perform request and write into netCDF file

- If the file is already there, the Beacon API is not called.
- The file name is created according to the region name, the period of interest and the depth range.

```
begin
    filename = joinpath(datadir,
    "Argo_$(parameter)_$(regionname)_$(Dates.format(datestart, "yyyymmdd"))-
    $(Dates.format(dateend, "yyyymmdd"))_$(Int(mindepth))-$(Int(maxdepth))m.nc");

    @info("Data will be written in file:\n$(filename)")

    if isfile(filename)
        @info("File already downloaded")
    else
        @time open(filename, "w") do io
            r = HTTP.request("POST", "$(beaconURL)/api/query",
            ["Content-Type" => "application/json"], query,
            response_stream=io);
            @info(r.status)
        end

    end;

    @info("NetCDF file size: $(round(filesize(filename)/1000^2, digits=1))M")
end;
```

```
Data will be written in file:
../data/Argo_sea_water_temperature_Canary_Islands_20000101-20101231_0-400m.nc
200
NetCDF file size: 1.8M
  2.249019 seconds (2.16 k allocations: 2.017 MiB)
```

2.2 s

Read netCDF content

```
begin
    @time obslon, obslat, obsdepth, obsdates, obsval =  read_netcdf(filename);
    coords = [ [obslon[i], obslat[i]] for i in 1:length(obslon) ];
    unique!(coords);
    @info("Found $(length(coords)) unique coordinates")
end
```

```
Found 949 unique coordinates
```

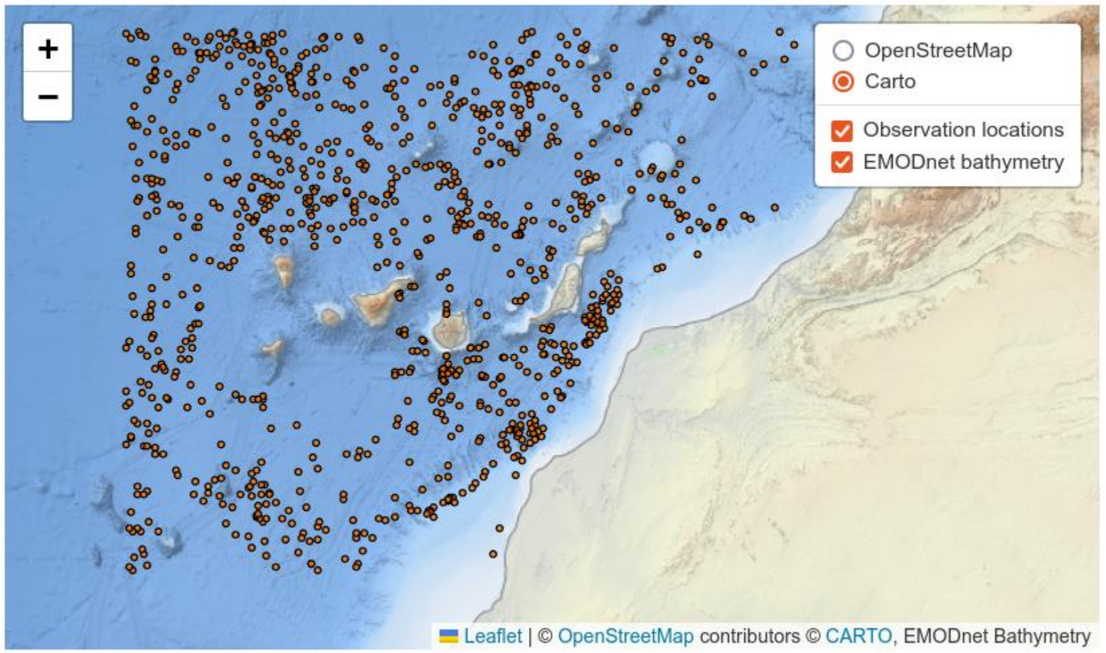
```
  0.002420 seconds (415 allocations: 1.705 MiB)
```

Interactive map (Leaflet)

Leaflet

The coordinates obtained from the input file (netCDF) are directly fed into Leaflet.

Changing the region or the period of interest will automatically update the interactive map.

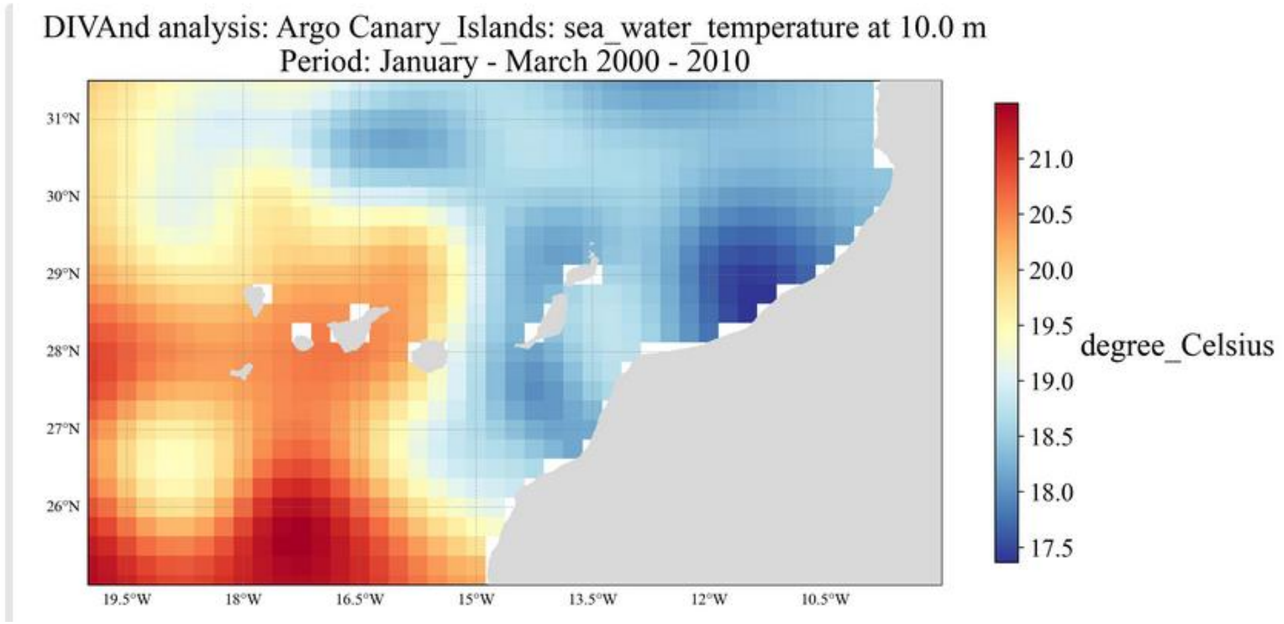


Notebook output: Creating an ARGO product in Pluto

Create the plot

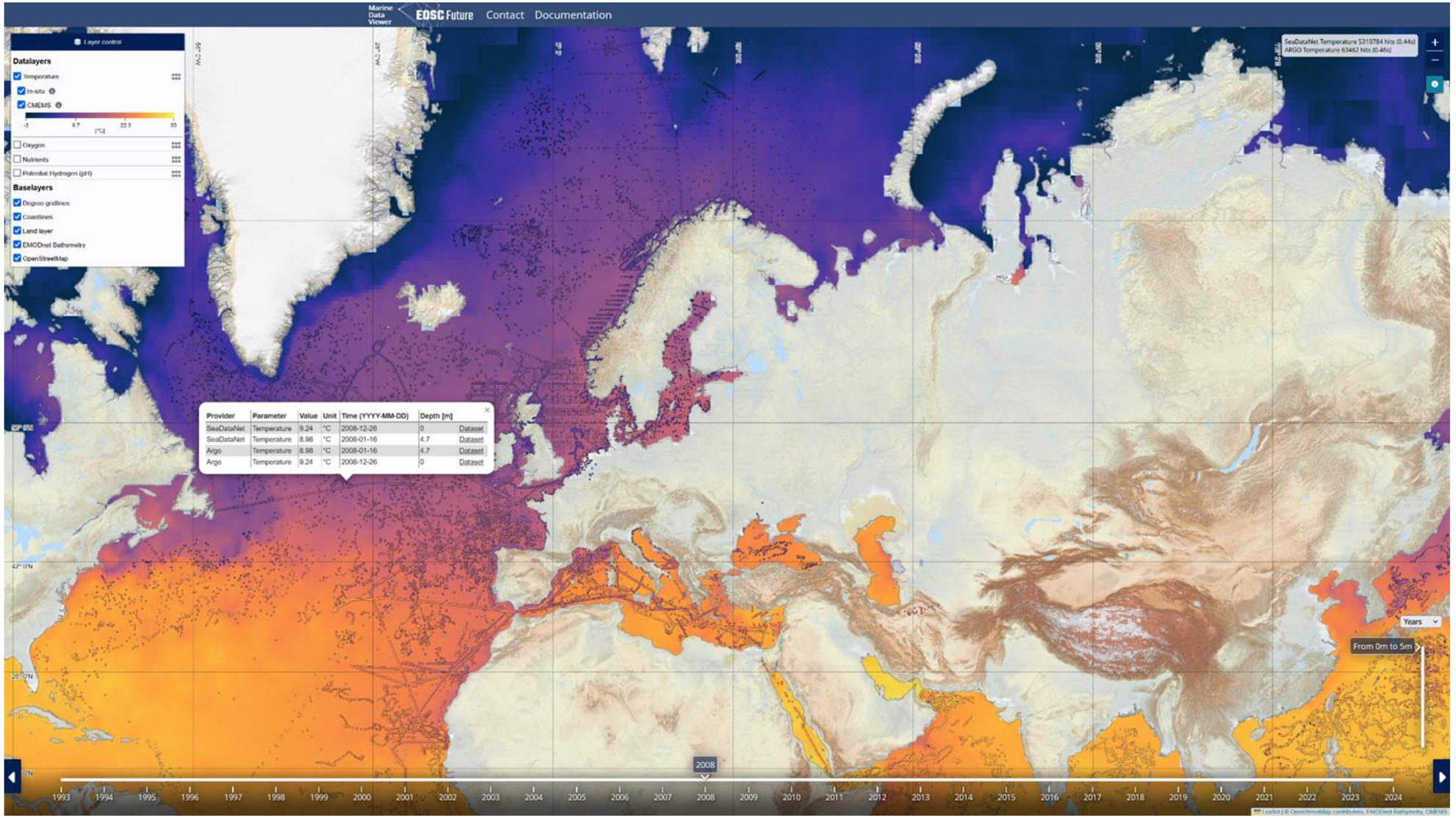
The figure title is set according to the depth and the time period

```
field2D = @view field[:,:,depthindex,timeindex];
```



Thanks Charles!

EOSC-Future demo (Beacon as back-end and supporting server side drawing)



https://eosc-future.maris.nl/

Current release (V0.9)

- Supports ODV pipeline for QC for the Workbenches
- Improved additional metadata handling and query options
- New Beacon feature: Dynamic Chunking
  - supports high-performance queries on large datafiles (remote sensing, products, etc).
  - Can extract e.g. timeseries on certain location
  - Like ZARR but more powerful💪 and faster ⚡
- Improved Stability of the Beacon Platform

What’s next?

- Deployment of Beacon for various data infrastructures under Blue-Cloud2026
- Integration and use into BC2026 and Fair-Ease VRE’s
- Towards first stable release:
  - Available for academic use
  - Open source available: API, Python scripts, Notebooks.
- Slowly rolling out the software into the community, and building a community

Questions and further information?

- Email to:
  - dick@maris.nl
  - peter@maris.nl
  - robin@maris.nl

- Join and connect to us via https://beacon.maris.nl

27-29 May 2024

International conference on Marine Data and Information Systems

SeaDataNet AISBL · Havforskningsinstituttet · Ifremer · MARIS · MEDIN · National Oceanography Centre · INGV · iMagine · ICDE · eosc Blue-Cloud2026