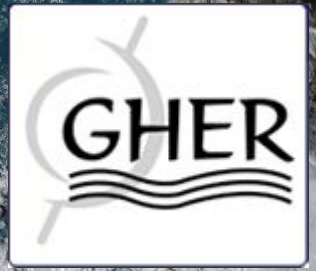
# Data analysis for ocean satellite data

Aida Alvera-Azcárate & Alexander Barth

GHER, University of Liège
Belgium



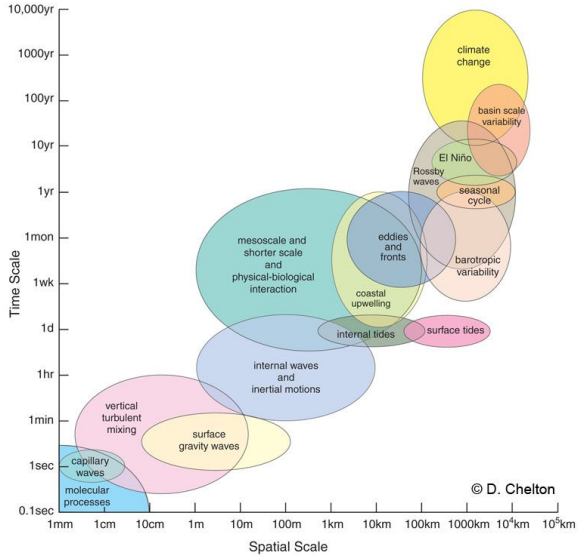



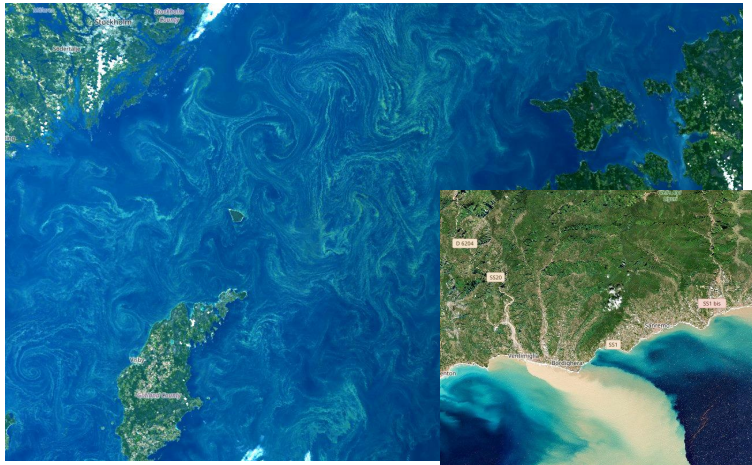
# Measuring the ocean

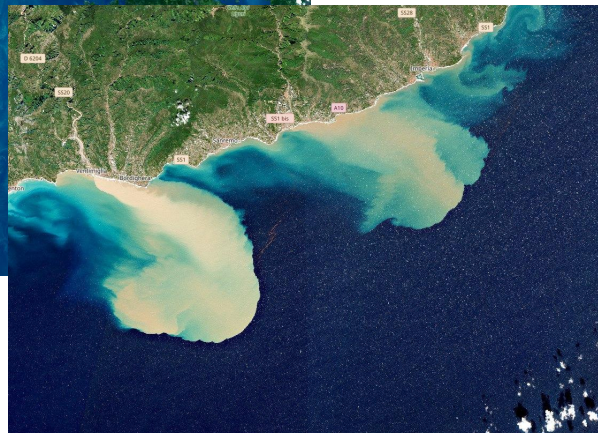
Large spatial and temporal variability

Climate change trends, extreme events

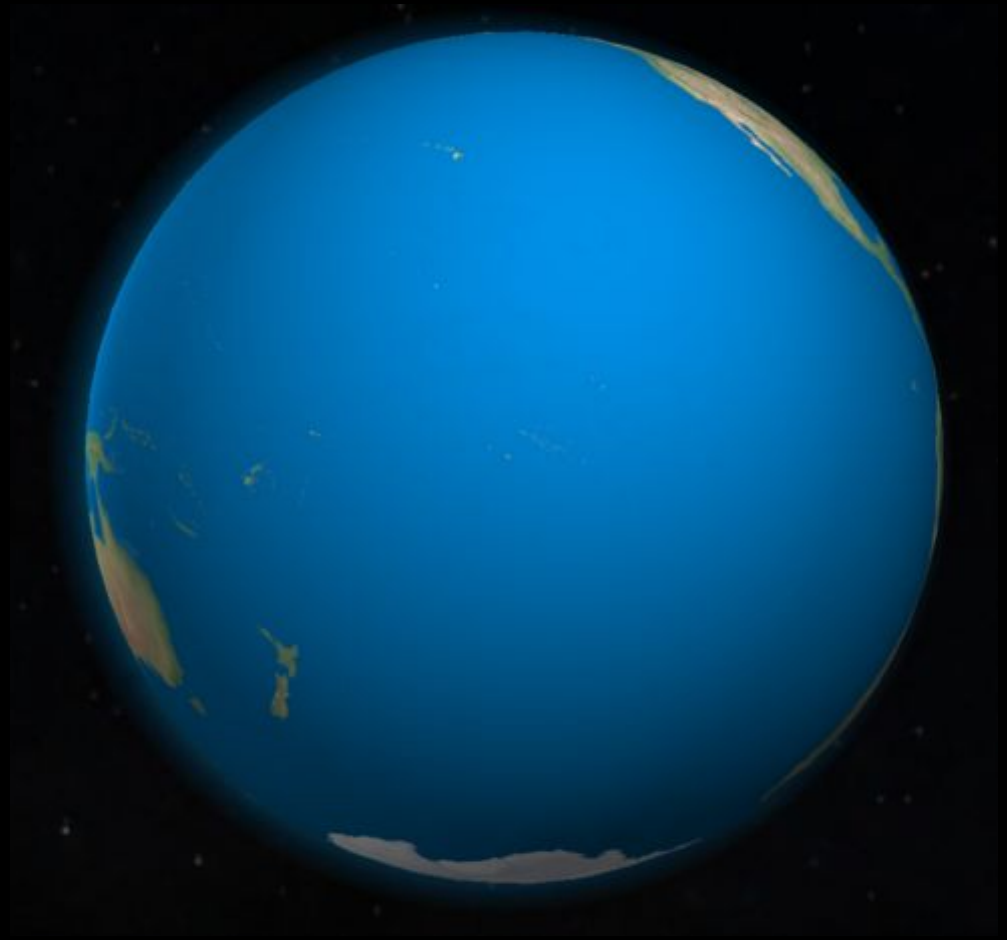
Pollution, changes in water quality





Impossible to measure adequately to resolve all scales



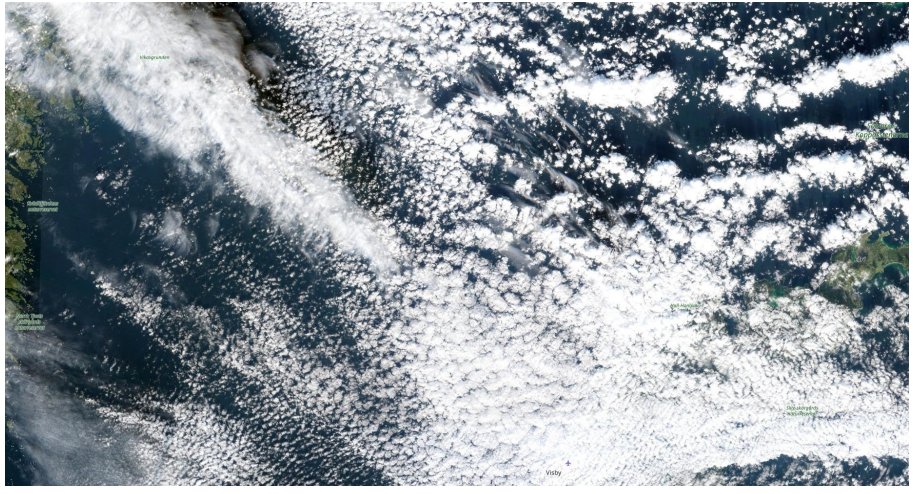
# Measuring the ocean: satellite data

- Satellites offer the best spatial and temporal coverage
    - Global daily coverage at < 1km
- Only surface
- Problem (in visible and infrared wavebands): clouds, dust…
    - LOTS of missing data

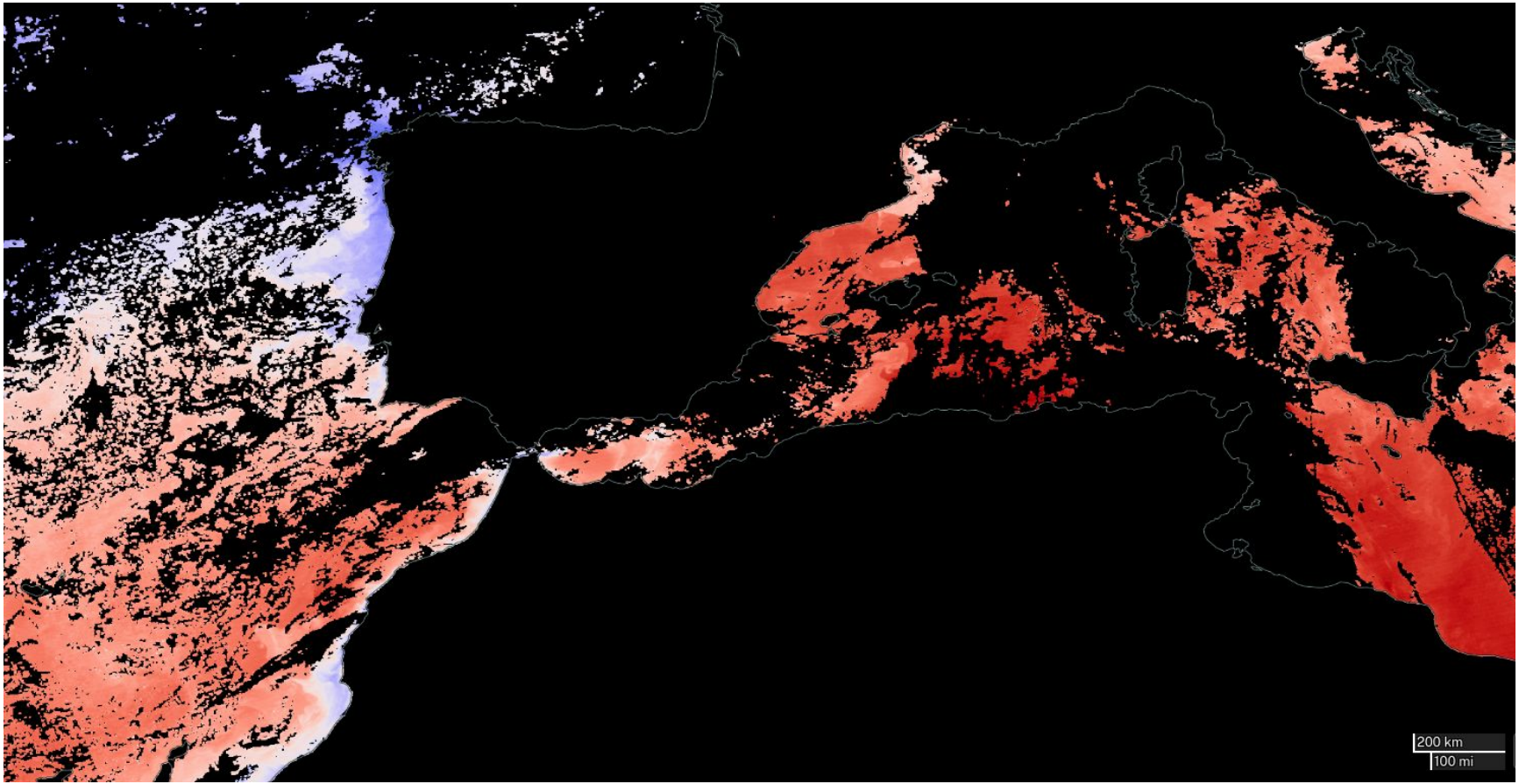
What you asked for…



… what you get



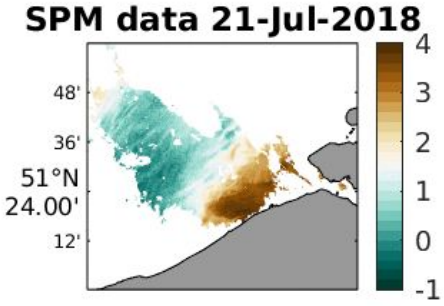
# Sea Surface Temperature (Suomi-NPP)



# Interpolating missing data in satellite datasets

- Clouds (etc) have **always** been a problem

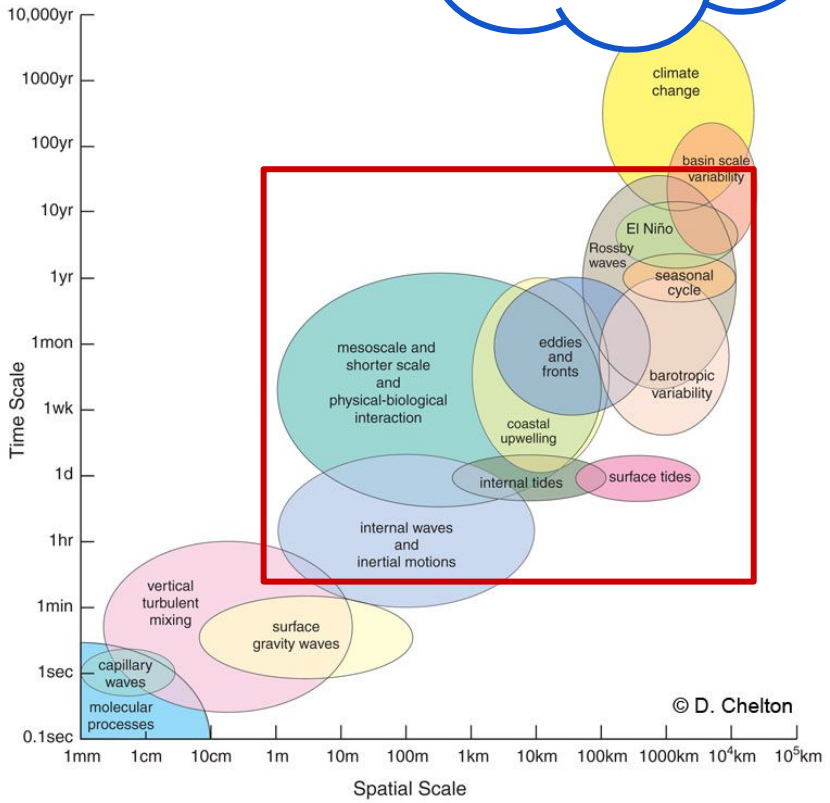
- Luckily they move around
  → spatio-temporal variability




Suspended Particulate Matter
in the Belgian coast

- Continuous measurements since:
  1998 for ocean colour
  1979 for sea surface temperature

**Large potential for data-driven analysis techniques**



# Interpolating missing data in satellite datasets

**Several approaches exist**

- **Empirical Orthogonal Functions (EOFs):**

  - Data compression, extraction of patterns, mains modes of variability…
  - Cannot handle missing data
  - However, we can use EOFs to **estimate missing information from the EOF basis calculated from the data → DINEOF**

- **Deep learning techniques**

  - Convolutional Autoencoder → **DINCAE**
  - Denoising Diffusion Probabilistic Model → **DINDiff**

# DINEOF (Data Interpolating Empirical Orthogonal Functions)

Technique to **fill in missing data** in geophysical data sets, based on a EOF decomposition

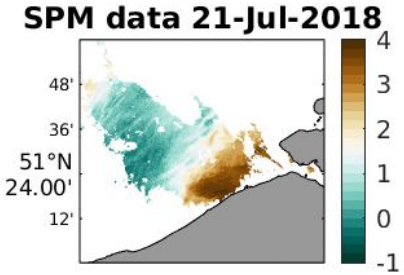
- Missing data? They get initialised to the mean value (and anomalies calculated)

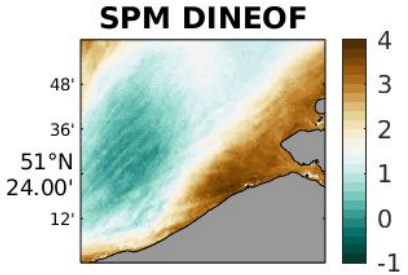
  First guess has low accuracy
  Iterative & incremental calculation of EOF modes

- **Truncated EOF basis** to calculate missing data
  - EOFs extract main patterns of variability
  - Reduced noise
  - Downside: reduced variability as well

- Optimal number of EOFs?
  - Reconstruction error by cross-validation:

    2-3% of valid data set aside
    Comparison at each converged EOF



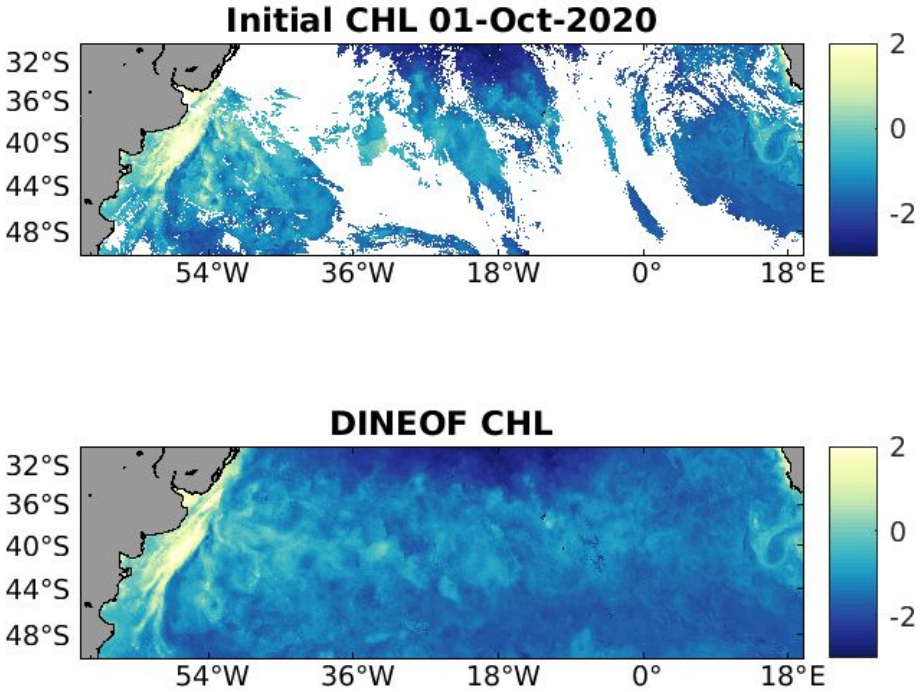

Beckers & Rixen (2003), https://doi.org/cc5kjc
Alvera-Azcárate et al (2005), https://doi.org/f7dr9p

# DINEOF (Data Interpolating Empirical Orthogonal Functions)

- Uses EOF basis to infer missing data: **non-parametric**

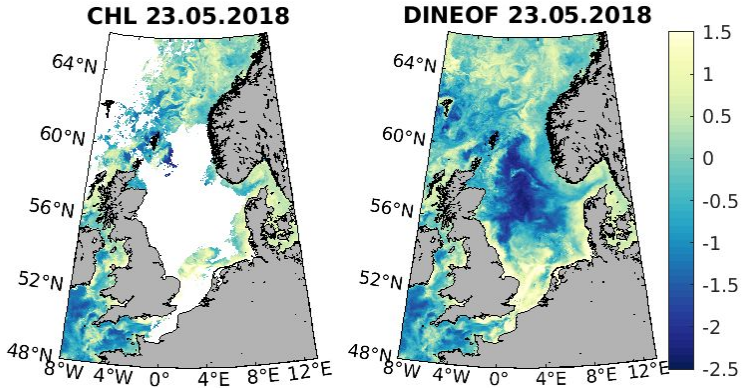
- **No need of a priori information** (correlation length, covariance function…)

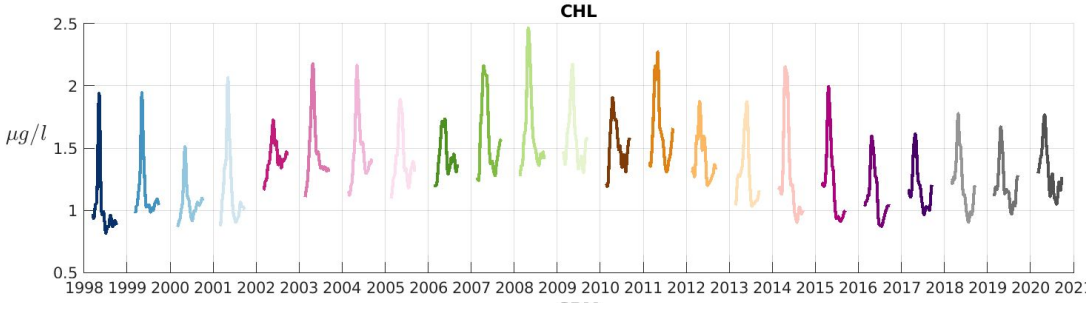
- Spatio-temporal coherence exploited to calculate missing values



# DINEOF example applications

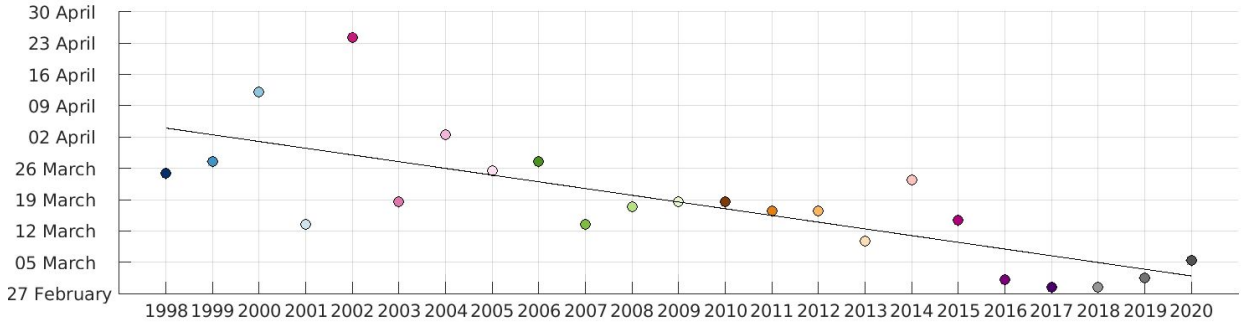
**Full fields are needed to study mean state, trends…**




## Spring bloom onset

Timing of spring bloom has decreased by 1.5 days $yr^{-1}$ from 1998 to 2020

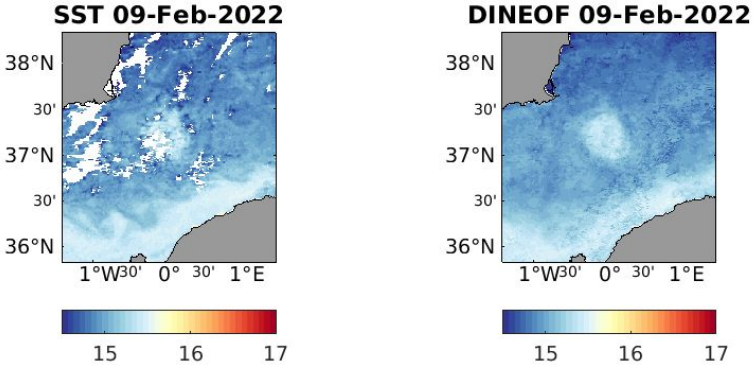


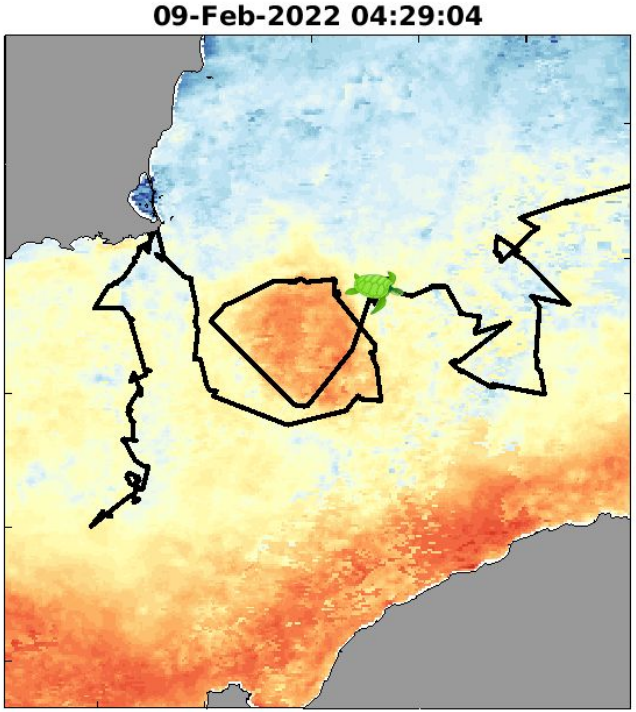

Alvera-Azcárate et al (2021), https://doi.org/rn3f

# DINEOF example applications

**Full fields are needed to track features**

Eddy variability, influence in ecosystem

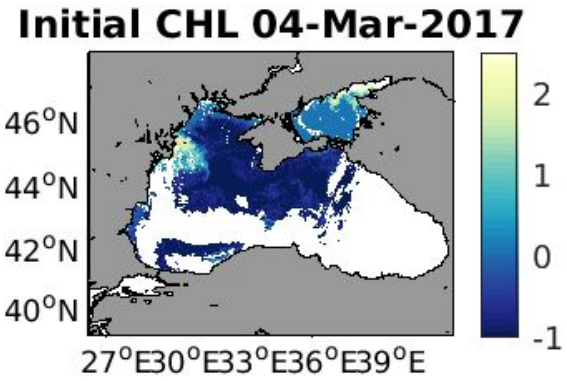


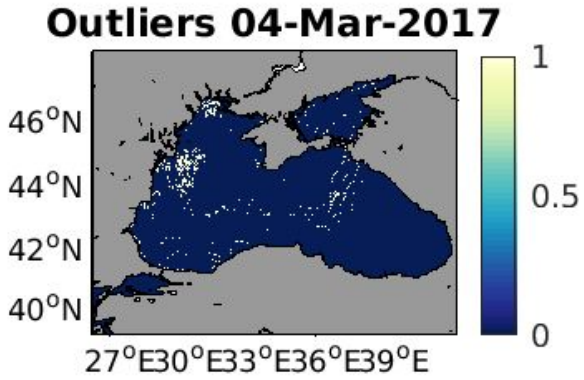



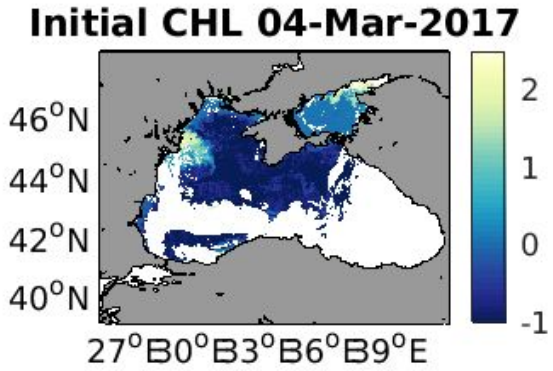
# DINEOF extensions

- **Detection of outliers** (Alvera-Azcárate et al, 2012; Alvera-Azcárate et al, 2015)

  - DINEOF run
  - Combined test (depart from EOF basis, from local median, proximity test) to detect suspect data
  - Once removed, a normal DINEOF reconstruction can be done



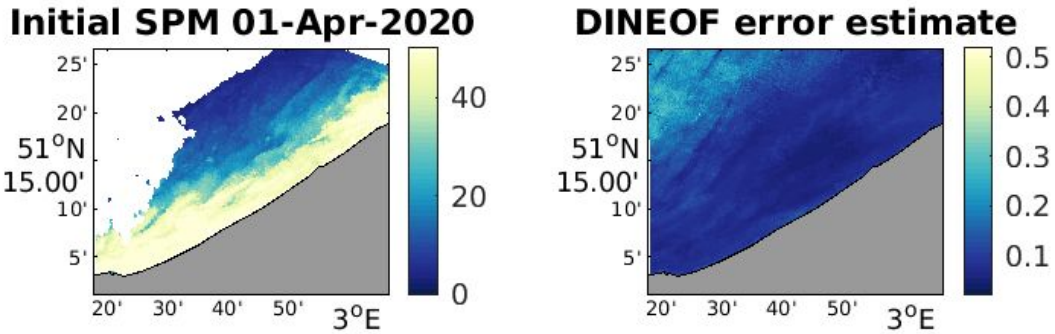
# DINEOF extensions

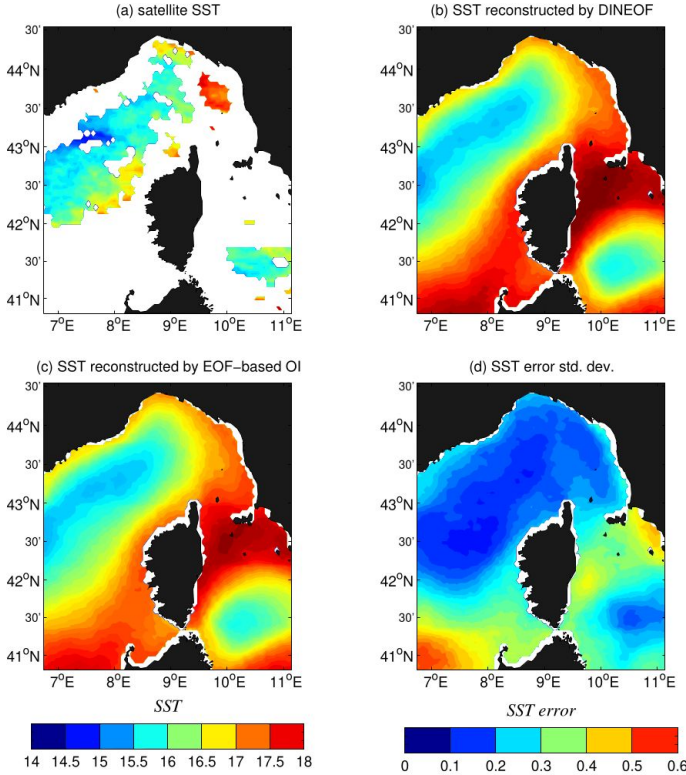
- **3D estimation of reconstruction error** (Beckers et al, 2006)

  - Exploiting the information rejected by the EOF basis

The observational error variance is taken to be the variance not retained within the EOF

An estimation of the correlation length of the error is needed

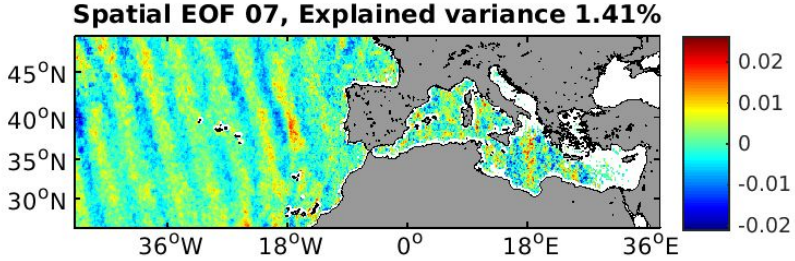


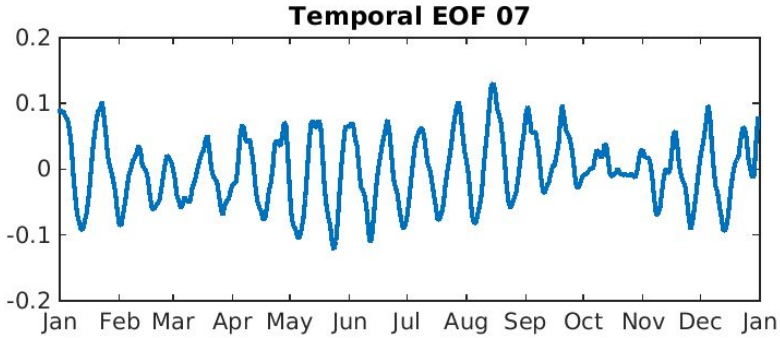



# DINEOF extensions

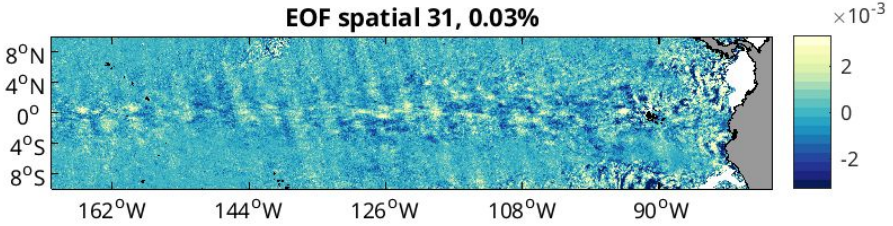
- **Removal of non-physical patterns** (e.g. swath edges) (Alvera-Azcárate et al, 2016)

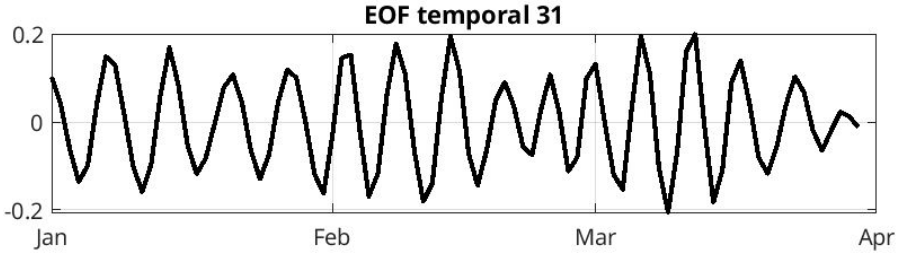
Sea Surface Salinity



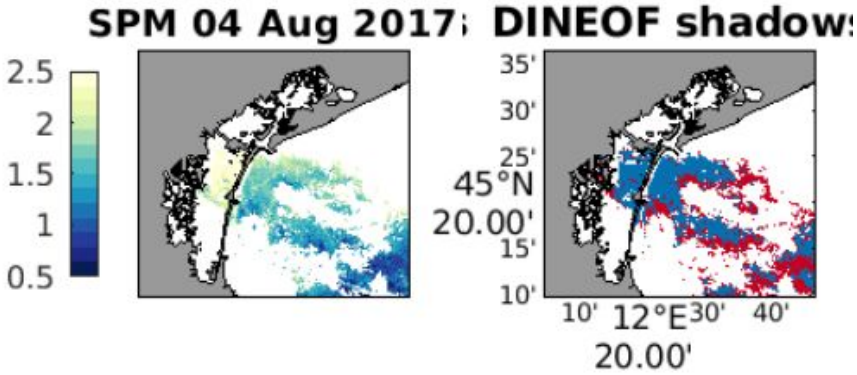

Chlorophyll-a concentration

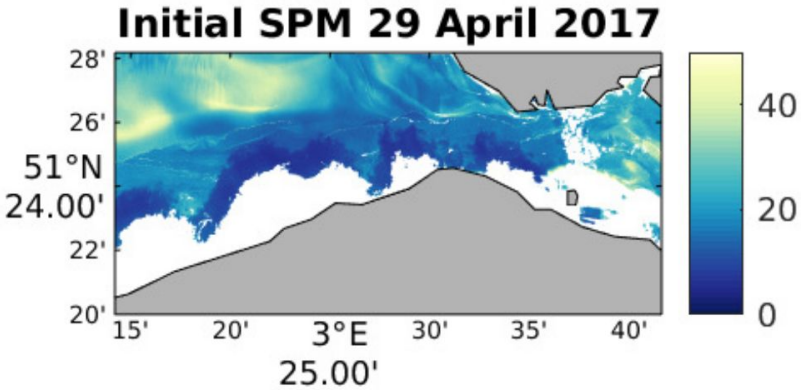


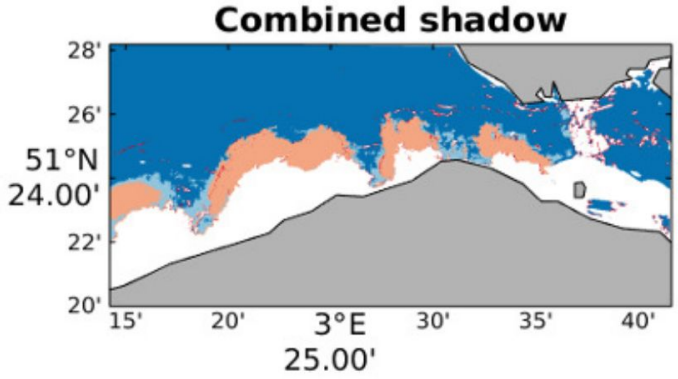



# DINEOF extensions

- Removal of cloud shadows (Alvera-Azcárate et al, 2021)



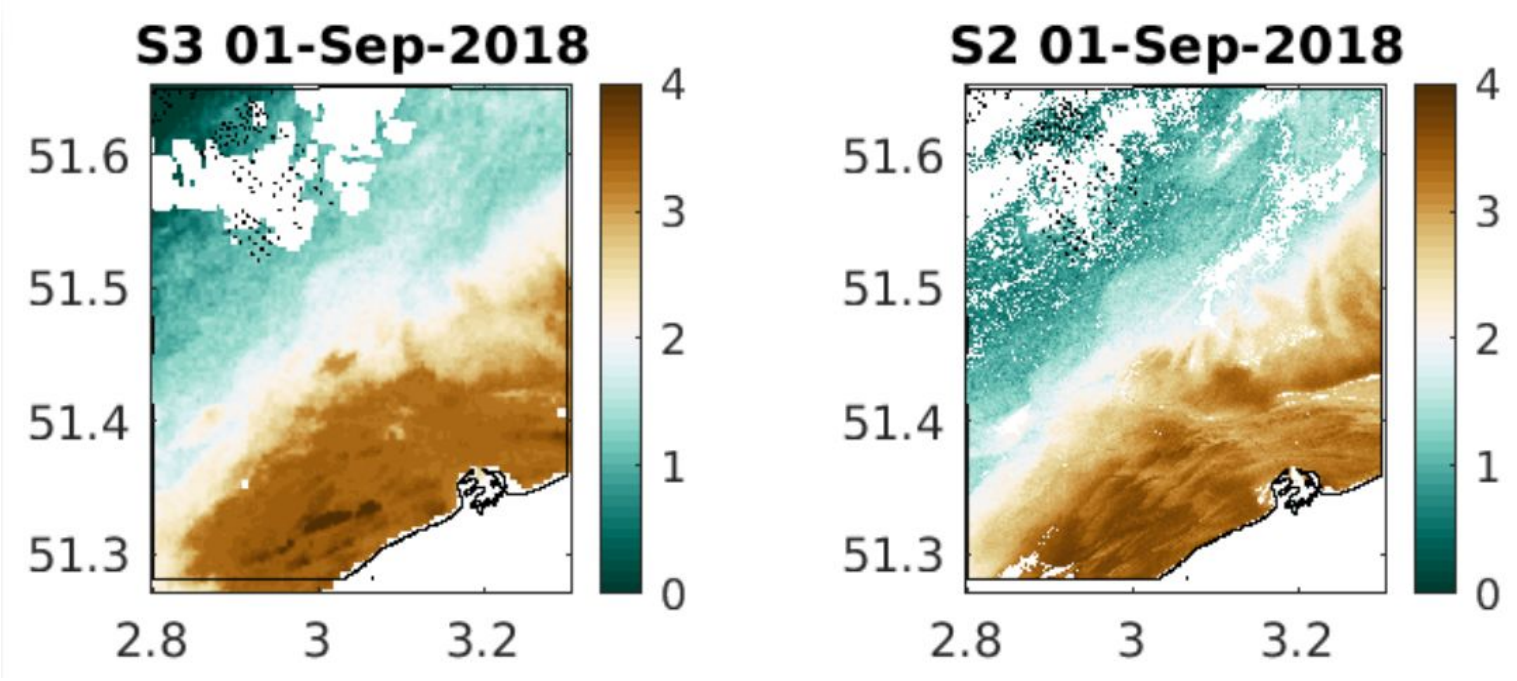




A combination of tests:

- EOF-basis departure
- Proximity to cloud
- Threshold value
- Ray tracing

# DINEOF extensions: synergy between datasets

- Different satellites measure similar properties: potential for synergy analyses



# DINEOF extensions: synergy between datasets

- Coastal ocean areas → very dynamic regions subject to strong anthropogenic pressure

- Satellite data → unique tool to study and monitor these areas at a unique spatial and temporal resolution.

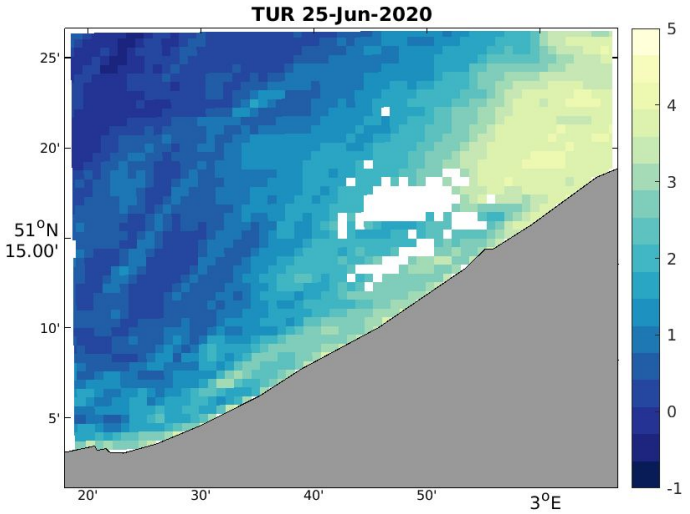
- Sentinel-2 data have high spatial resolution (10-20-60m) but 3-5 day temporal resolution

- Sentinel-3 data have medium spatial resolution (300 m) and daily temporal resolution

- Both affected by clouds

**Enhancing spatial resolution** (Alvera-Azcárate et al, 2025)

Data matrix is formed by a combination of 2 data streams:
Sentinel-2 if available (here at 60 m)
Sentinel-3 if no Sentinel-2 data (interpolated at 60 m)
Both cloudy (if more than 98% clouds no data for that day)



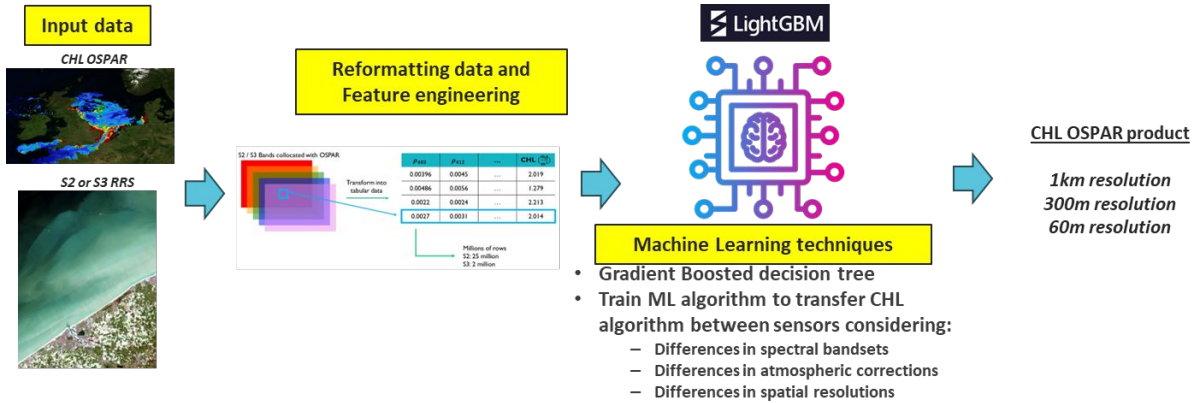
# DINEOF extensions: synergy between datasets

## 1. Processing chain

**Generation of coherent Sentinel-2/MSI and Sentinel-3/OLCI ocean color products**

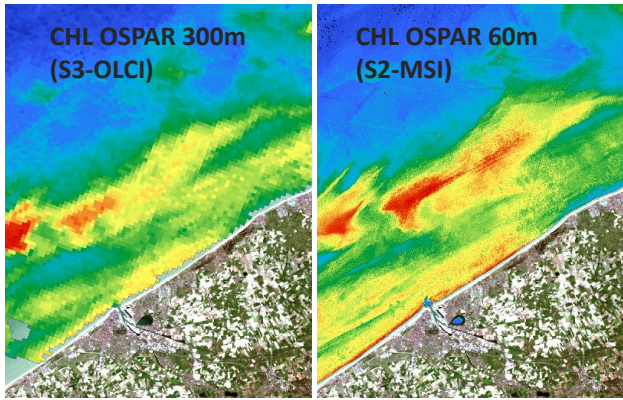
- Multi-algorithm approach guided by optical water types
- Generated products: Rrs, CHL, SPM, TUR

## 2. Increase coherency between Sentinel-2 and Sentinel-3

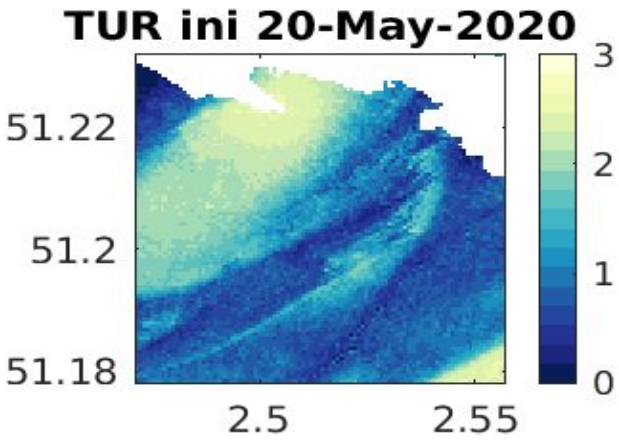


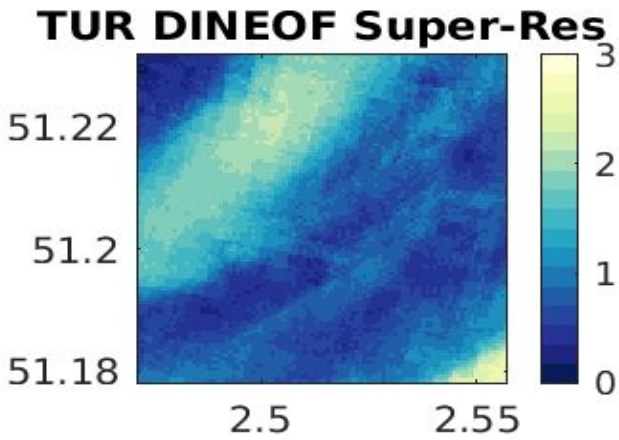

- Gradient Boosted decision tree
- Train ML algorithm to transfer CHL algorithm between sensors considering:
  - Differences in spectral bandsets
  - Differences in atmospheric corrections
  - Differences in spatial resolutions

Sentinel-2 RGB 20200405 (Oostende, BE)

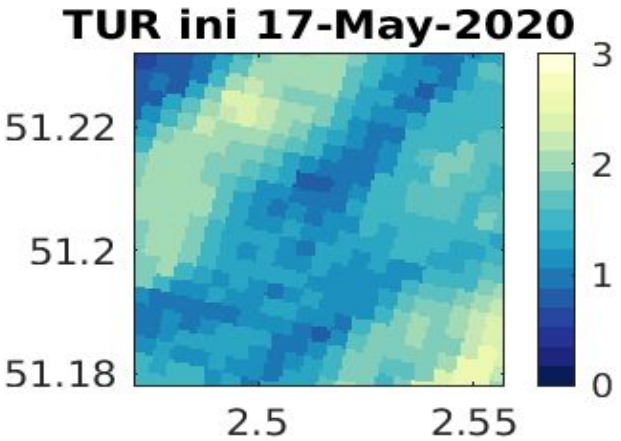


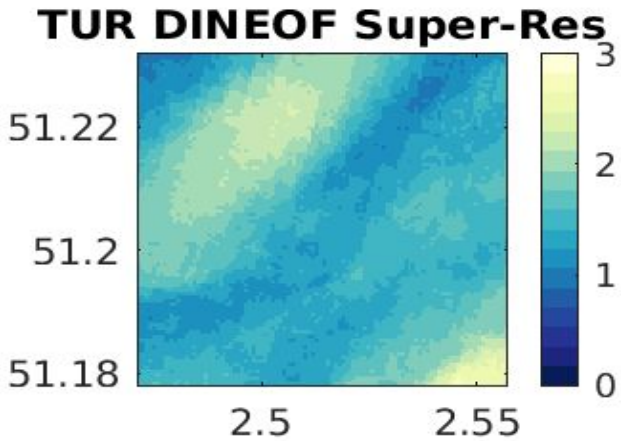
# An increase in spatial resolution




**Initially Sentinel-2**

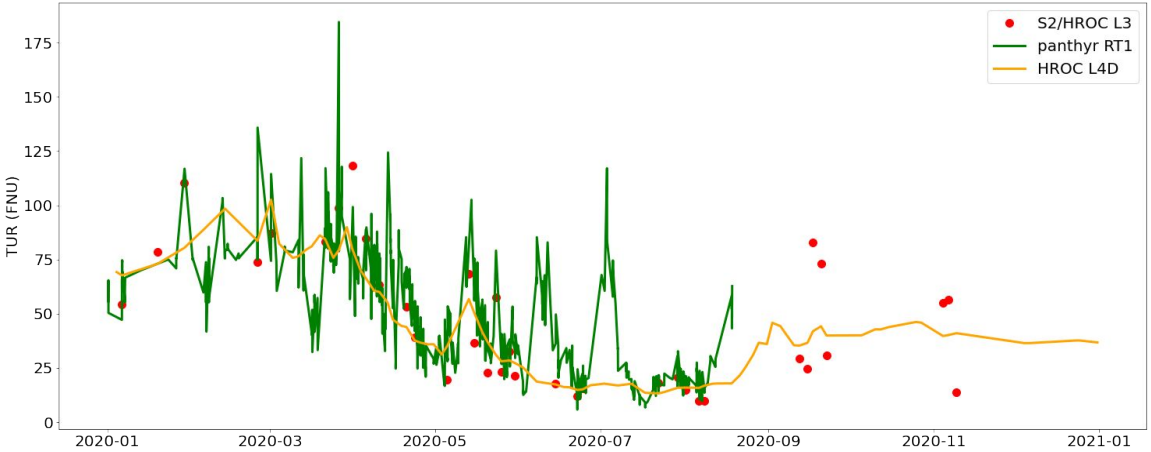


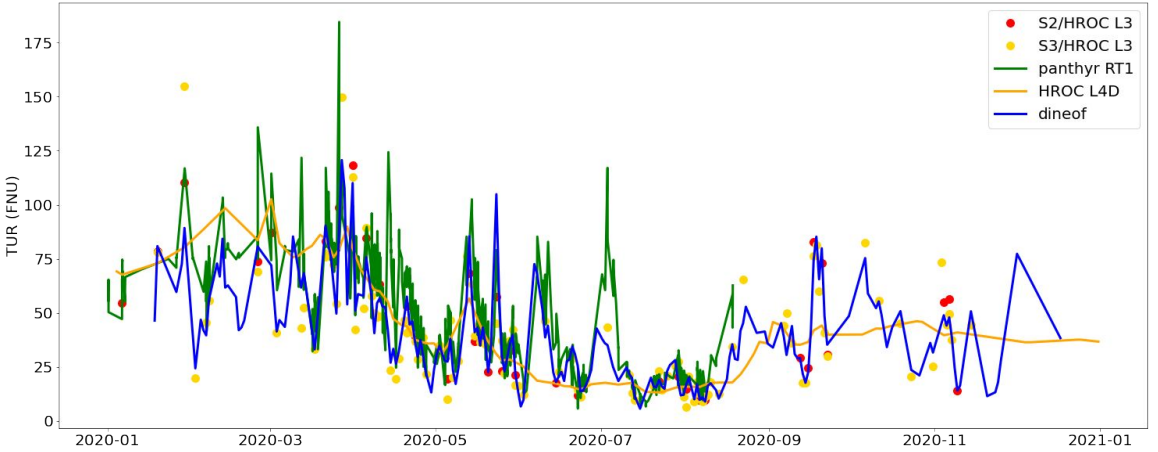




**Initially Sentinel-3**

(Alvera-Azcárate et al, 2025)
https://doi.org/rn3j

# And an increase in temporal resolution



# DINEOF extensions




or




?

# Good old classics vs new approaches?

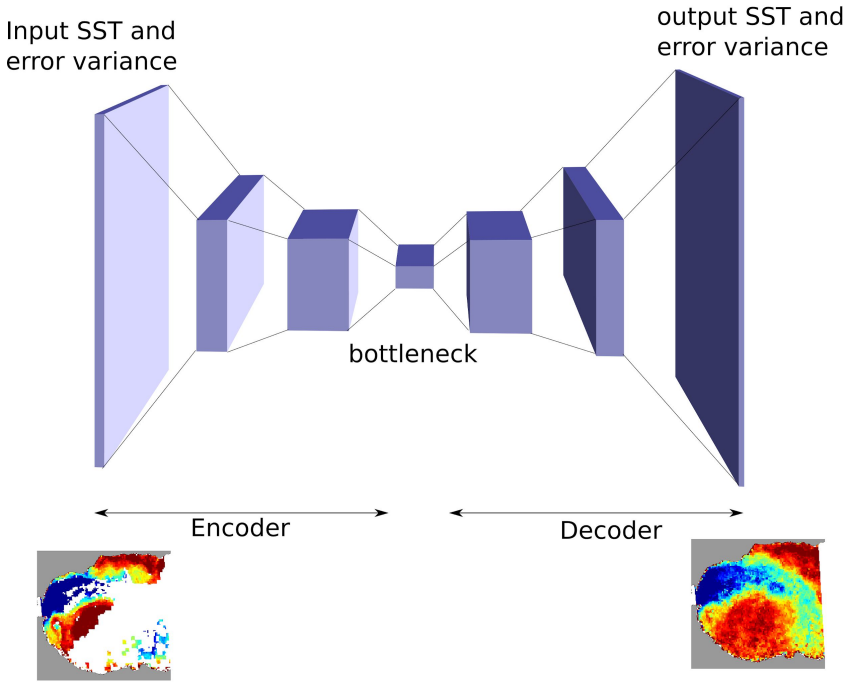
**DINEOF:**

- Robust and fast reconstruction of missing data
- Ease of use
- Explainable, multiple applications
- Too smooth? (only linear relationships are taken into account)

**Deep learning-based approaches:**

- Fast development of new architectures
- Good results, promising avenue
- Steep learning curve, GPUs needed

# DINCAE (Data Interpolating Convolutional Autoencoder)




**Auto-Encoder:** used to efficiently compress/decompress data, by extracting main patterns of variability

- Similarity to EOFs (auto-encoder with 1 encoding/decoding layer and no activation function)

**Convolutional:** works on subsets of data, i.e. trains on local features

Missing data handled as data with different initial errors

- If **missing, error variance** ($\sigma^2$) **tends to infinity**

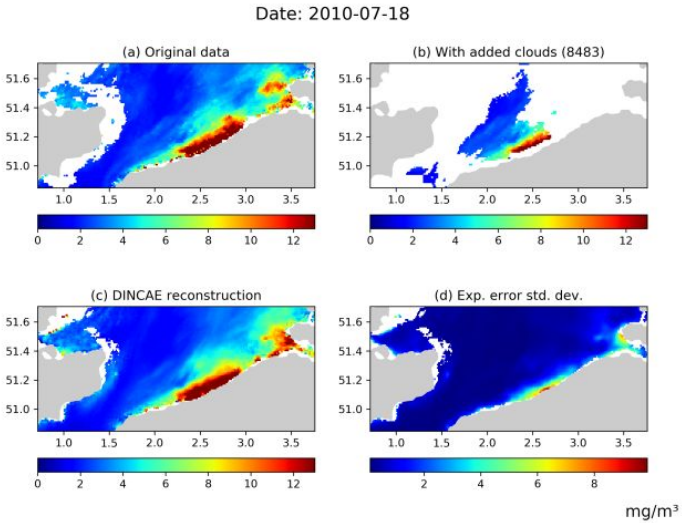
Skip connections

Multivariate input data

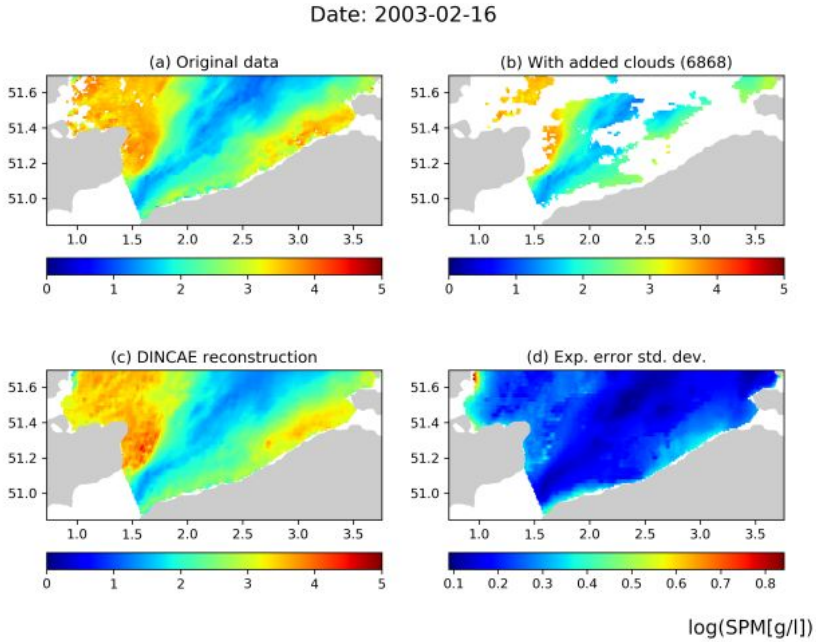
Barth et al. (2022), https://doi.org/hvzj

# DINCAE examples

## Chlorophyll




## SPM

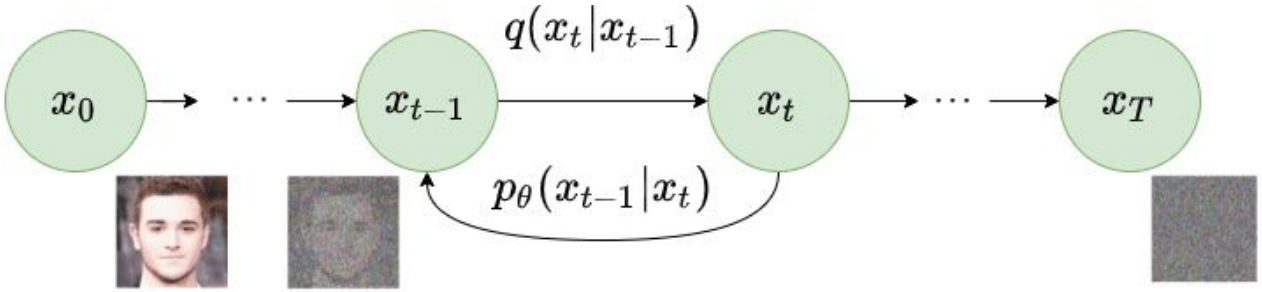



Barth et al. (2021) https://doi.org/g6stkt

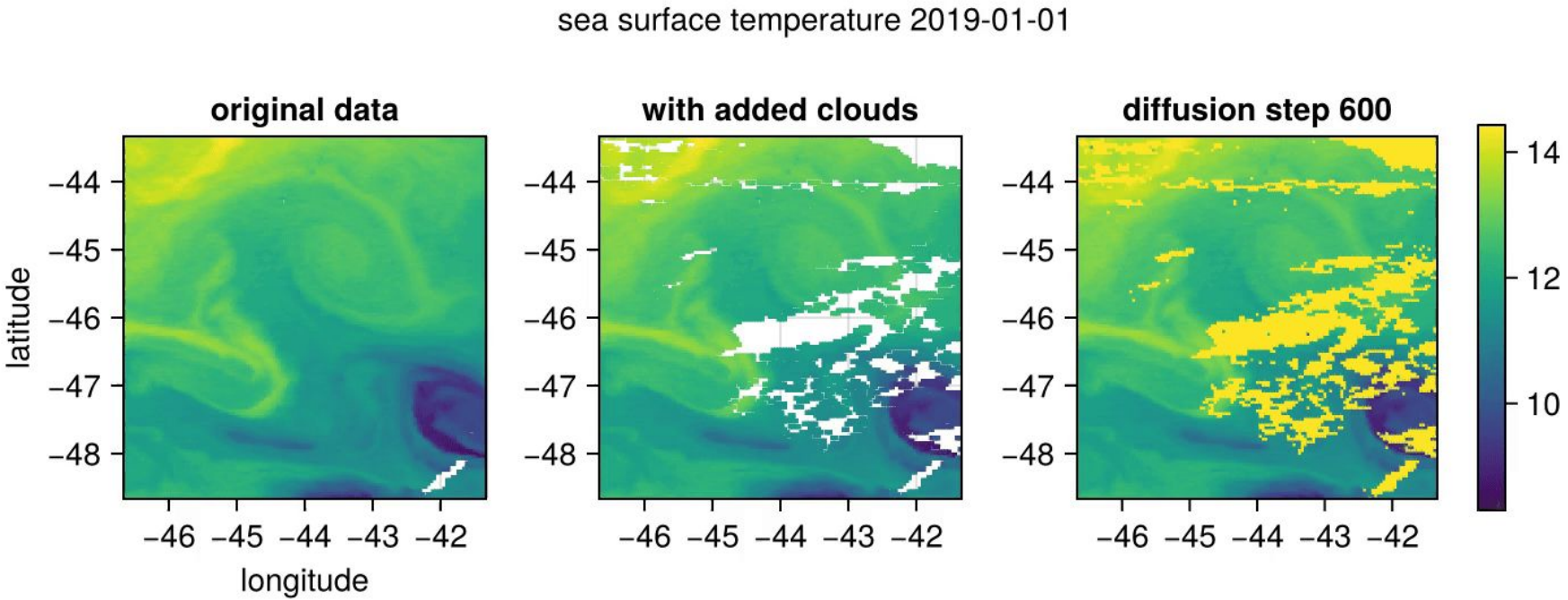
# DINDiff: denoising diffusion model

## Training a diffusion model

- Repeatedly add noise to initial image or field
- Sum of two Gaussian distributed random variables → another Gaussian distributed random variable
- To generate a degraded state $\mathbf{x}_t$ at stage t: no need to make all intermediate $\mathbf{x}_2$, $\mathbf{x}_3$,… $\mathbf{x}_{t-1}$
- Train a neural network to predict the noise at stage t
- The reverse process is also stochastic



# DINDiff: denoising diffusion model

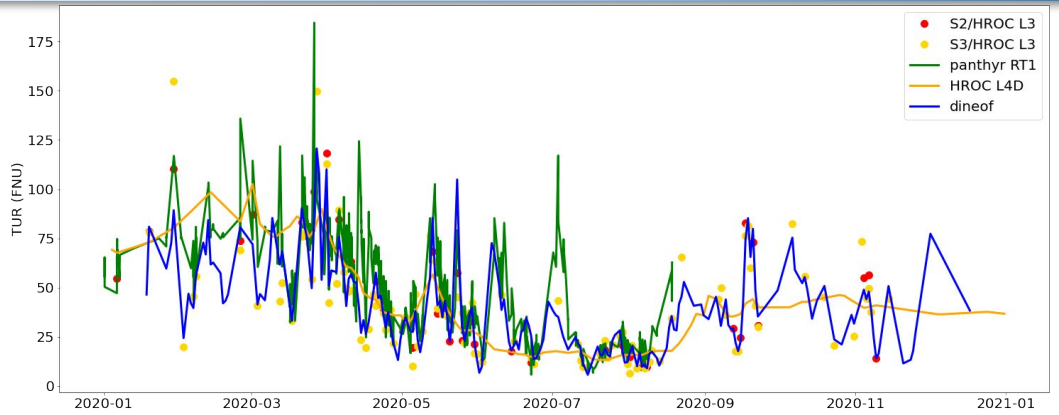



Barth et al. (2024), https://doi.org/pj7c
https://github.com/gher-uliege/DINDiff.jl

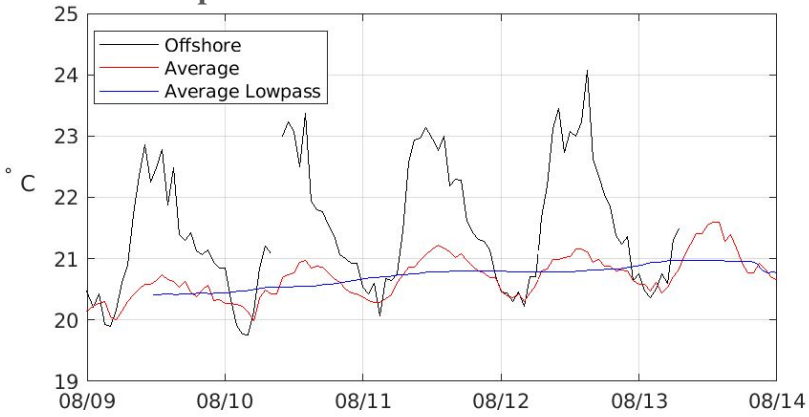
# Some data musings

# Heterogeneity of data sources

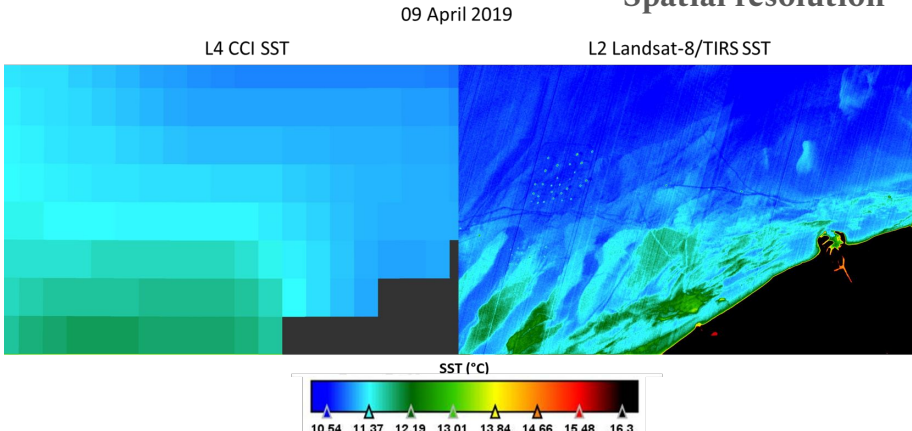
## In situ vs satellite



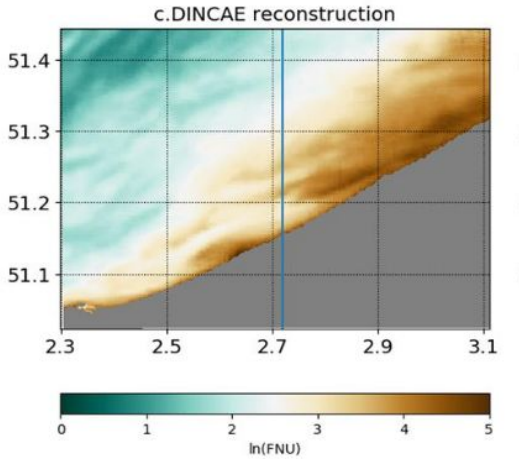

## Temporal resolution + time of measurement



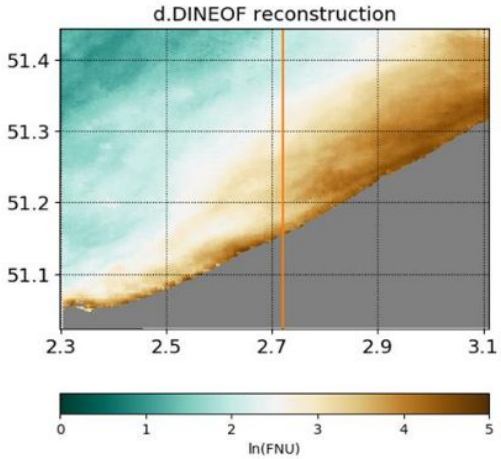

## Spatial resolution



# Diversity of methods






- The state of the ocean is **unknown**, and the ocean is hugely undermeasured

- Several interpolated fields can be considered *correct*:
    - because they are produced for different purposes
    - because of a similar error estimate: **validation**

# Best practices in deep learning for oceanography

When developing a new methodology / application for data interpolation:

- Compare with the best available, state of the art methodologies
- Careful partition of datasets (training, validation, testing)
- Check the independence of the data, clarify their status
- Provide a measure of uncertainty
- Beware of overfitting
- Use benchmarks
- Avoid black-box use

**Best practices in deep learning for oceanography**

Aida Alvera-Azcárate[1], Alexander Barth[1], Julien Brajard[2], Rachel Furner[3], Michael Gray[4], Matjaž Ličer[5,6], Redouane Lguensat[7], Aneesh C. Subramanian[8], Ali Aydogdu[9], Andrea Storto[10], and Maike Sonnewald[11]

[1]GeoHydrodynamics and Environment Research (GHER), FOCUS, University of Liège, Liège, Belgium
[2]Nansen Center; Bergen, Norway
[3]European Center for Medium Range Weather Forecasting (ECMWF), Reading, UK
[4]Fathom Science Inc, Raleigh, North Carolina, United States of America
[5]Slovenian Environment Agency, Ljubljana, Slovenia
[6]National Institute of Biology, Marine Biology Station, Piran, Slovenia
[7]Institut Pierre-Simon Laplace, IRD; Sorbonne Université, Paris, France
[8]Atmospheric and Oceanic Sciences, University of Colorado Boulder
[9]CMCC Foundation, Italy
[10]Institute of Marine Sciences (ISMAR), National Research Council (CNR), Rome, Italy

**Correspondence:** Aida Alvera-Azcárate (A.Alvera@uliege.be)

**To be submitted soon to Ocean Science** (https://www.ocean-science.net)

# Conclusions

Several data-driven techniques available for missing data reconstruction

EOFs are a powerful tool, with diverse applications in ocean remote sensing
- Explainability

Deep learning techniques developing fast & with great potential
- Validate, validate, validate
- Code sharing may boost your approach

https://github.com/gher-uliege/DINCAE.jl

https://github.com/gher-uliege/DINDiff.jl

https://github.com/aida-alvera/DINEOF

