Filling gaps in ocean satellite data

Aida Alvera-Azcárate, Alexander Barth
GHER, University of Liège
Belgium

Objective: give you an overview of data-driven gap-filling techniques for ocean satellite data

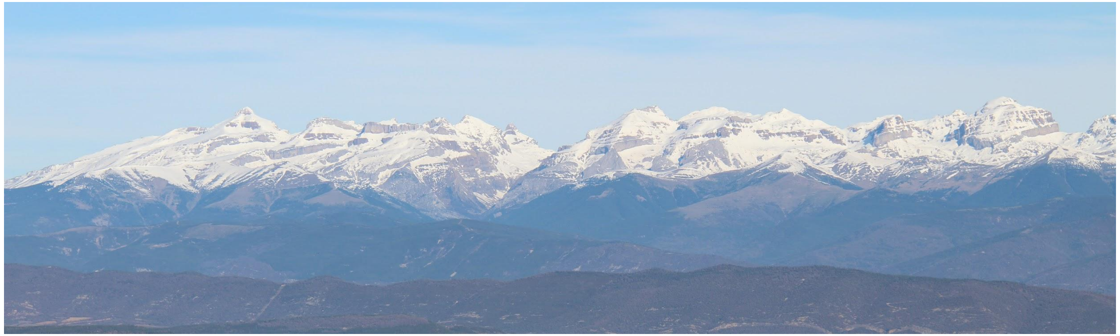
This webinar will be divided in three parts

- Description of DINEOF (statistical method)
- Description of DINCAE (neural network method)
- Demonstration exercice (materials will be provided for you to try later)

First let me know you

Q1: little questionnaire

An oceanographer in the mountains...



Career:

- 1995-2000: MSc in Marine Sciences (University of Las Palmas de Gran Canaria, Spain).

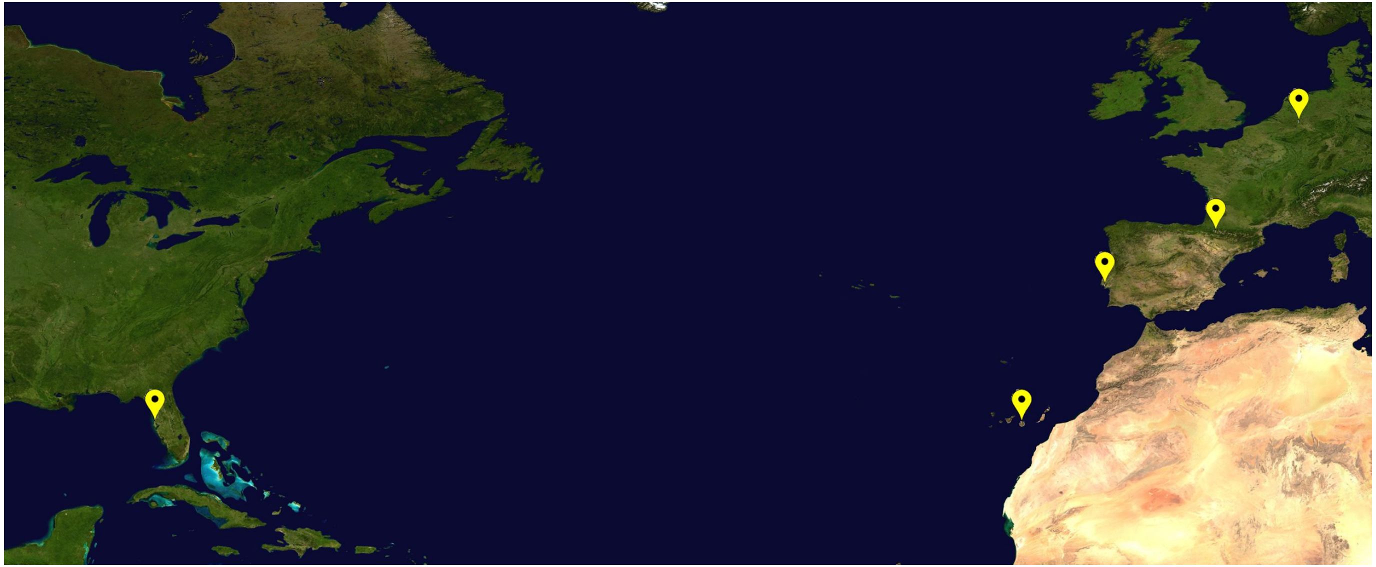
- 2001: Master in Oceanography: University of Liège (Belgium) and New University of Lisbon (Portugal).

- 2001 - 2004: PhD in Oceanography, University of Liège, Belgium.

- 2004 - 2007: Research Associate at the College of Marine Science, University of South Florida (US).

- 2007 - 2012: Chargé de Recherches FRS-FNRS (Fonds de la Recherche Scientifique) at the University of Liège.

- 2012-present: Researcher at the University of Liège.



The GHER

Physical oceanography group at the University of Liège (Belgium)

Main research activities

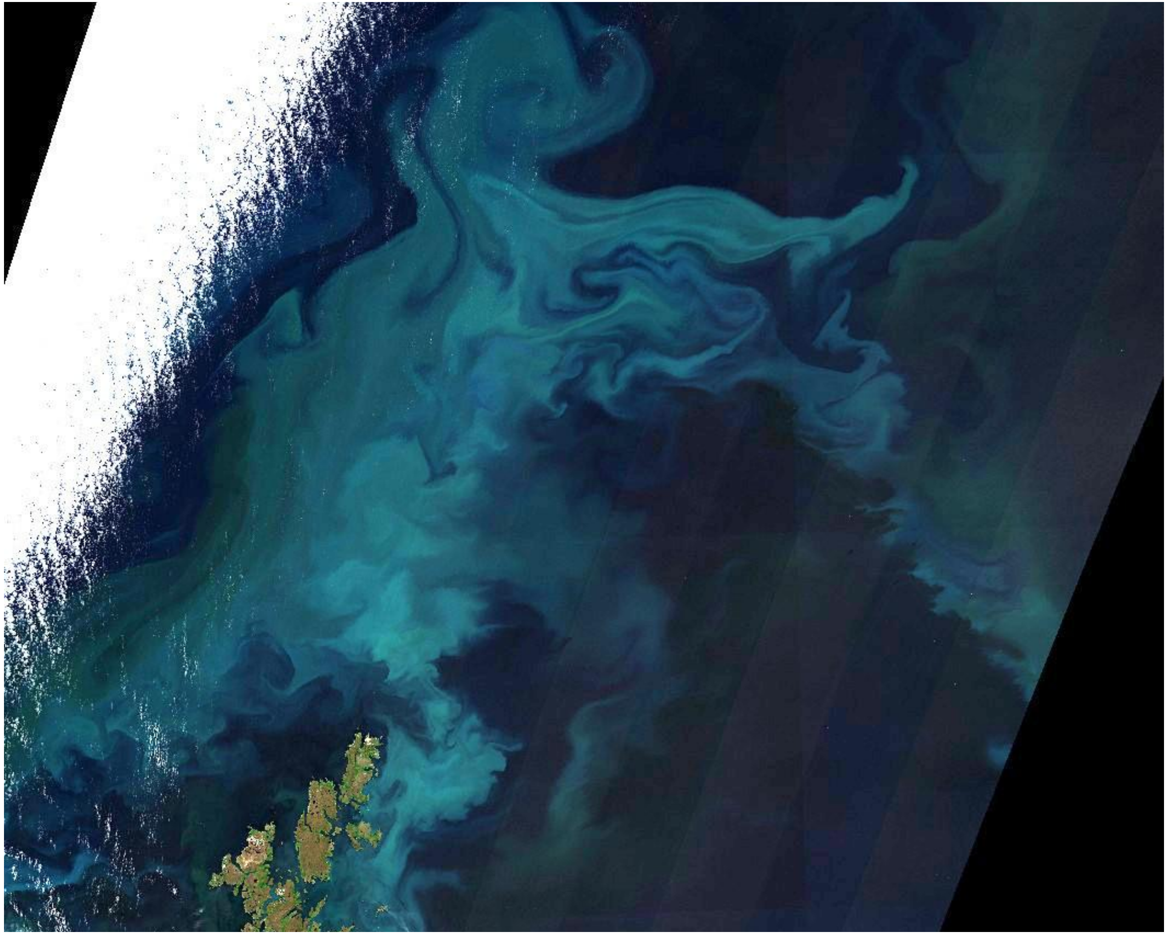
- Ocean modelling
- Data assimilation
- Development & application of data analysis techniques
  - DIVA, DIVAnd
  - DINEOF
  - DINCAE

Master in Oceanography, Erasmus+ Master MER2030

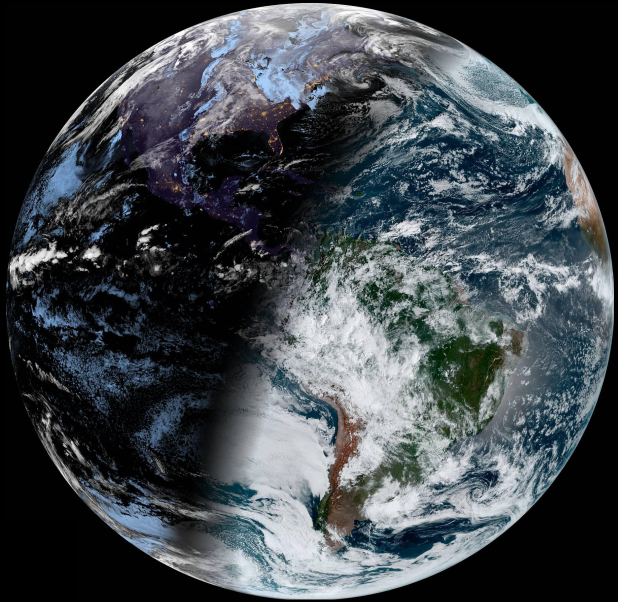
Organizers of the Liège Colloquium in Ocean Dynamics

GeoHydrodynamics and Environment Research

No “sea view” from my window, but amazing views from my computer screen



GOES-East



Meteosat

Himawari-8

Do you see beautiful marble pictures of the Earth?

… I see clouds

The problem

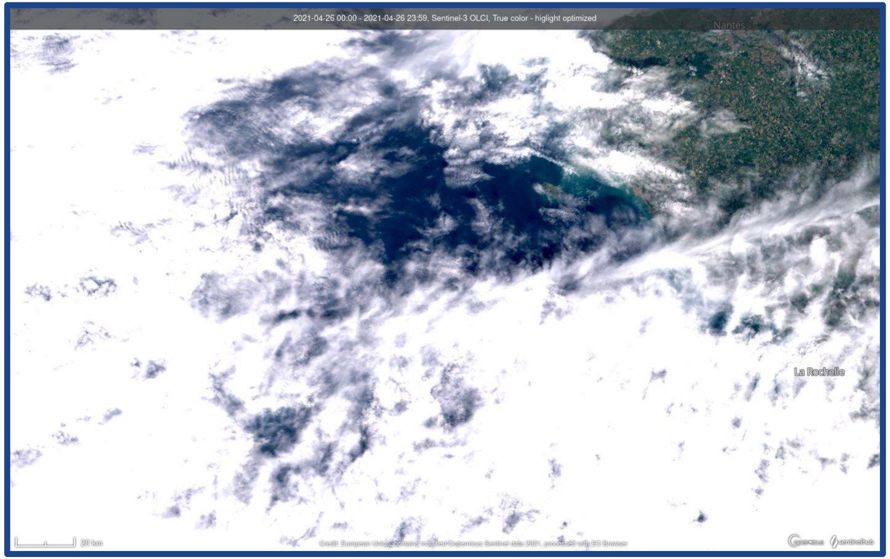
Satellite sensors measuring in the visible and infrared wavelengths can't "see" through clouds, dust, haze...

As a result, satellite data for variables like sea surface temperature, chlorophyll concentration, suspended sediments, etc, are heavily affected by missing data

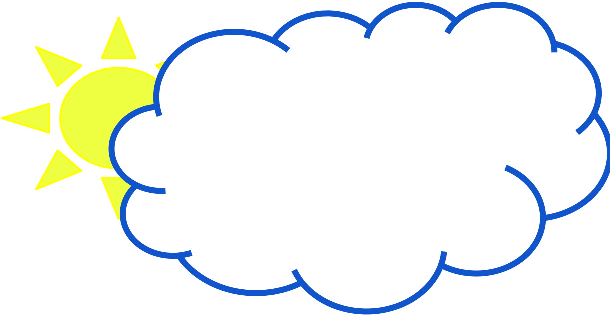
- Latitudinal and seasonal variability in the % of missing data

What you asked for...

... what you get



Interpolating missing data in satellite datasets

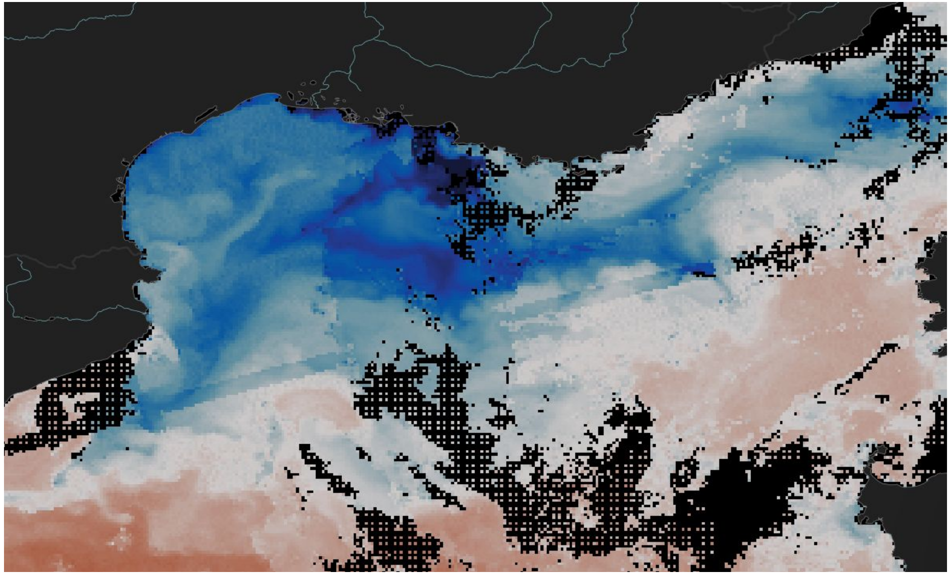


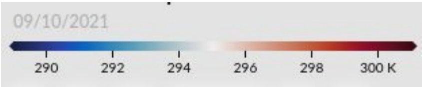
- Clouds have been **always** a problem

- Luckily they move around: spatio-temporal analyses can help

- Several approaches have been used to remove or minimise the effect of clouds, e.g. :

    1 - Compositing (loss of spatial/temporal resolution, biases, artefacts)





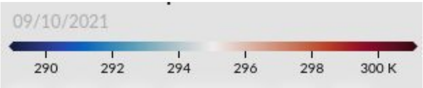
2 - Interpolation techniques (e.g. Optimal Interpolation)

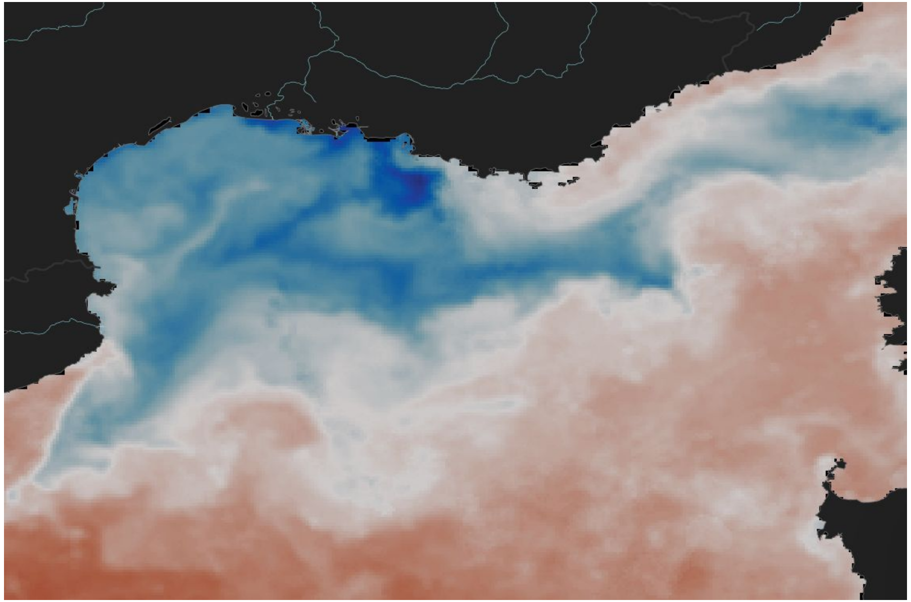
Gridded field = First guess + weighted sum of observations

- Typically previous knowledge of the characteristics of the interpolated variable are needed

    - Correlation length (how far do observations influence the final product)

    This leads to **subjectivity** and **local** analyses (i.e. teleconnections not taken into account)





3 - Data-driven approaches, e.g. DINEOF

Beckers & Rixen (2003) develop a method to **estimate missing information from the EOF basis calculated from the data**

- EOFs provide a series of main modes of variability, classified by importance
- Uses an SVD method to calculate the EOFs (provides best truncated EOF matrix)
- For a data matrix X → $X = USV^T$

EOFs should **not** be calculated with missing data

- SVD assumes data matrix X is perfectly and completely known
- If covariance matrix ($C = X^TX$) is only calculated on available data:
    - C no longer semi positive definite (eigen-values not positive or zero)
    - Eigenvalues can be negative: classification of EOFs by their importance no longer possible

In short: we're calculating EOFs (that shouldn't be used when missing data) to find the values of the missing data

How does that work??

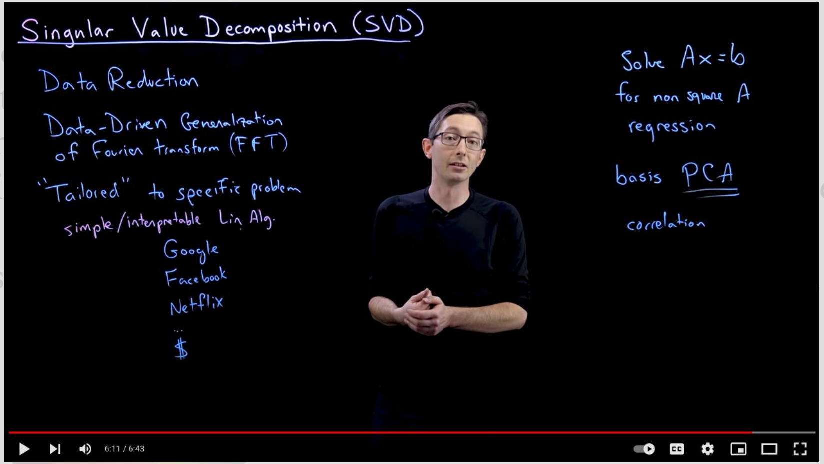
EOFs, PCAs, SVDs....

This all goes back to the algebra lessons we have had in high school

Lots of great resources out there to get the basics, refresh or deepen your knowledge

Q2: have you worked with EOFs before?

"Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control" by Brunton and Kutz

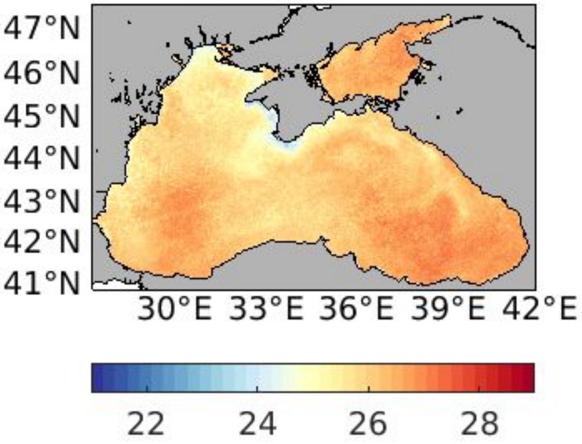


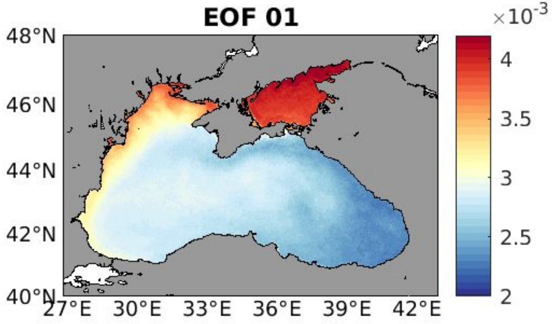
https://www.youtube.com/watch?v=gXbThCXjZFM

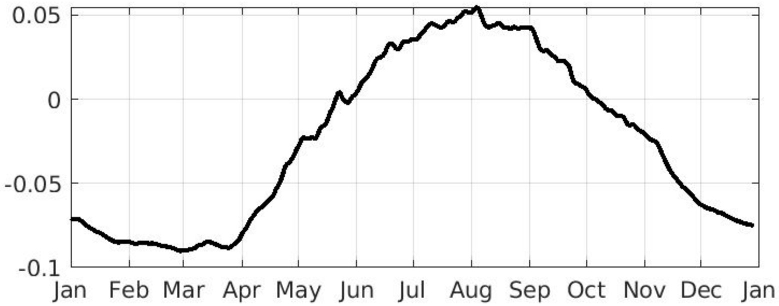
Empirical Orthogonal Functions (EOFs)

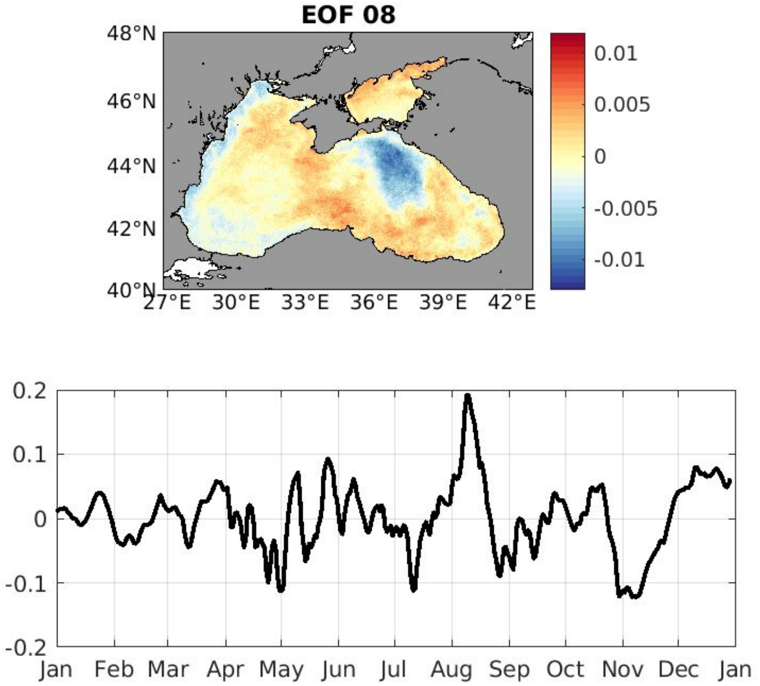
- They are a compact way of representing main modes of variability in a dataset
- Each EOF mode consist of a spatial field and time series, that together sum X% of total variability
- EOFs are ordered in decreasing variability order

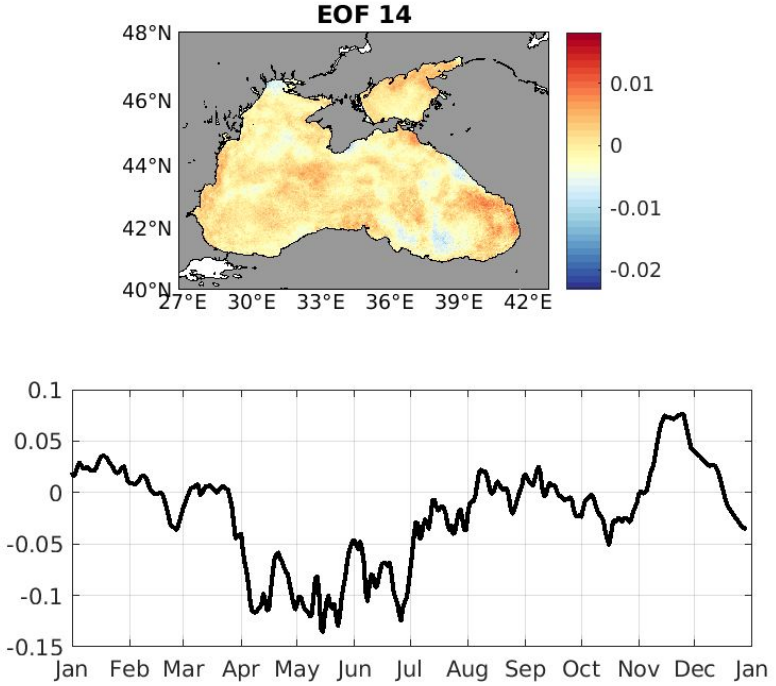
A varying field of daily Sea Surface Temperature











EOF 01: 96.68% variability
EOF 08: 0.03% variability
EOF 14: 0.01% variability

3 - Data-driven approaches, e.g. DINEOF

Beckers & Rixen (2003) develop a method to **estimate missing information from the EOF basis calculated from the data**

- EOFs provide a series of main modes of variability, classified by importance
- Uses an SVD method to calculate the EOFs (provides best truncated EOF matrix)
- For a data matrix X → $X = USV^T$

EOFs should **not** be calculated with missing data

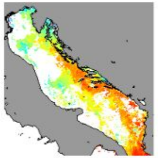
- SVD assumes data matrix X is perfectly and completely known
- If covariance matrix ($C = X^TX$) is only calculated on available data:
    - C no longer semi positive definite (eigen-values not positive or zero)
    - Eigenvalues can be negative: classification of EOFs by their importance no longer possible

In short: we're calculating EOFs (that shouldn't be used when missing data) to find the values of the missing data

How does that work??

DINEOF

DINEOF (Data Interpolating Empirical Orthogonal Functions)



1st: Demeaned matrix: missing data flagged and set to zero

Some data are set aside for cross-validation

2nd: EOF decomposition with N=1 EOF

Calculate missing values:

$$\boxed{X_{i,j} = \sum_{p=1}^{k} \rho_p (u_p)_i (v_p^T)_j}$$

Improved guess for missing values

Convergence: $\begin{cases} \text{best value for missing data with 1 EOF} \\ \text{cross validation: error} \end{cases}$

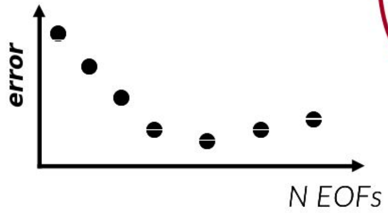
EOF decomposition with N=2 EOFs

Calculate missing values

Improved guess for missing values

Then we repeat with N= 3 EOFs

and so on...



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

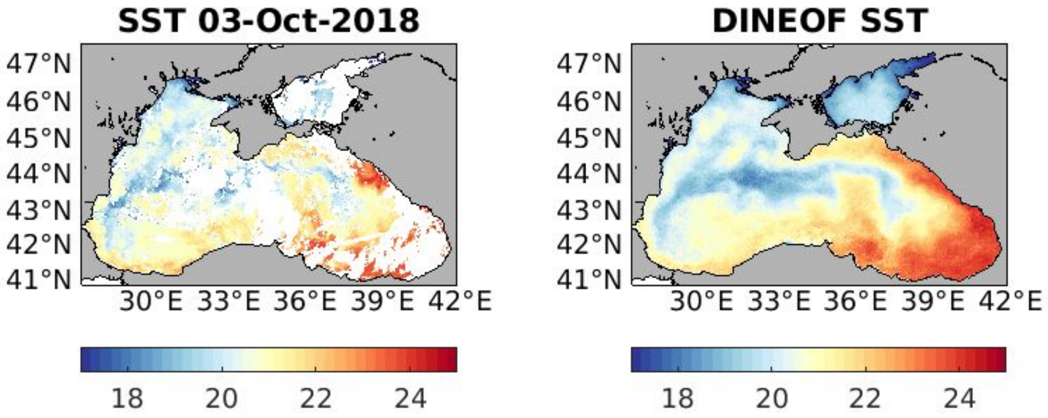
  Incremental & iterative calculation of EOF modes

- **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:

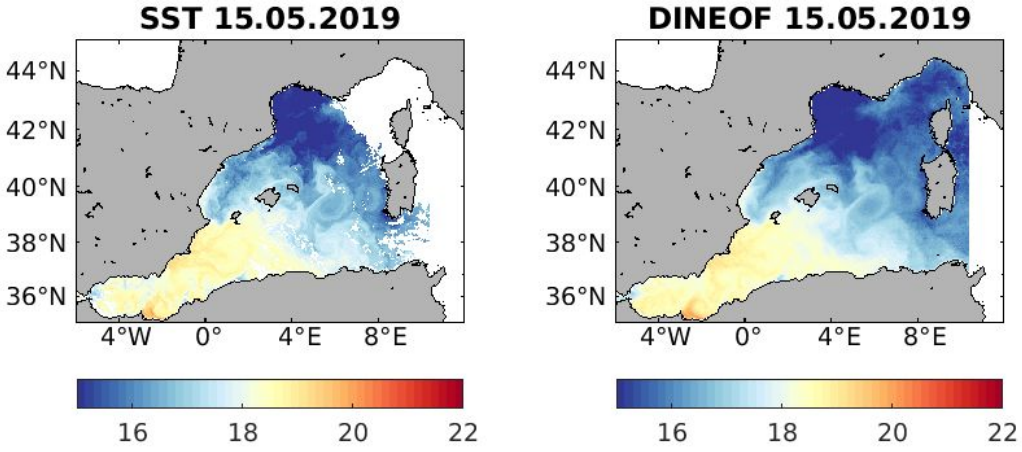
  2-3% of valid data set aside

  Comparison at each converged EOF

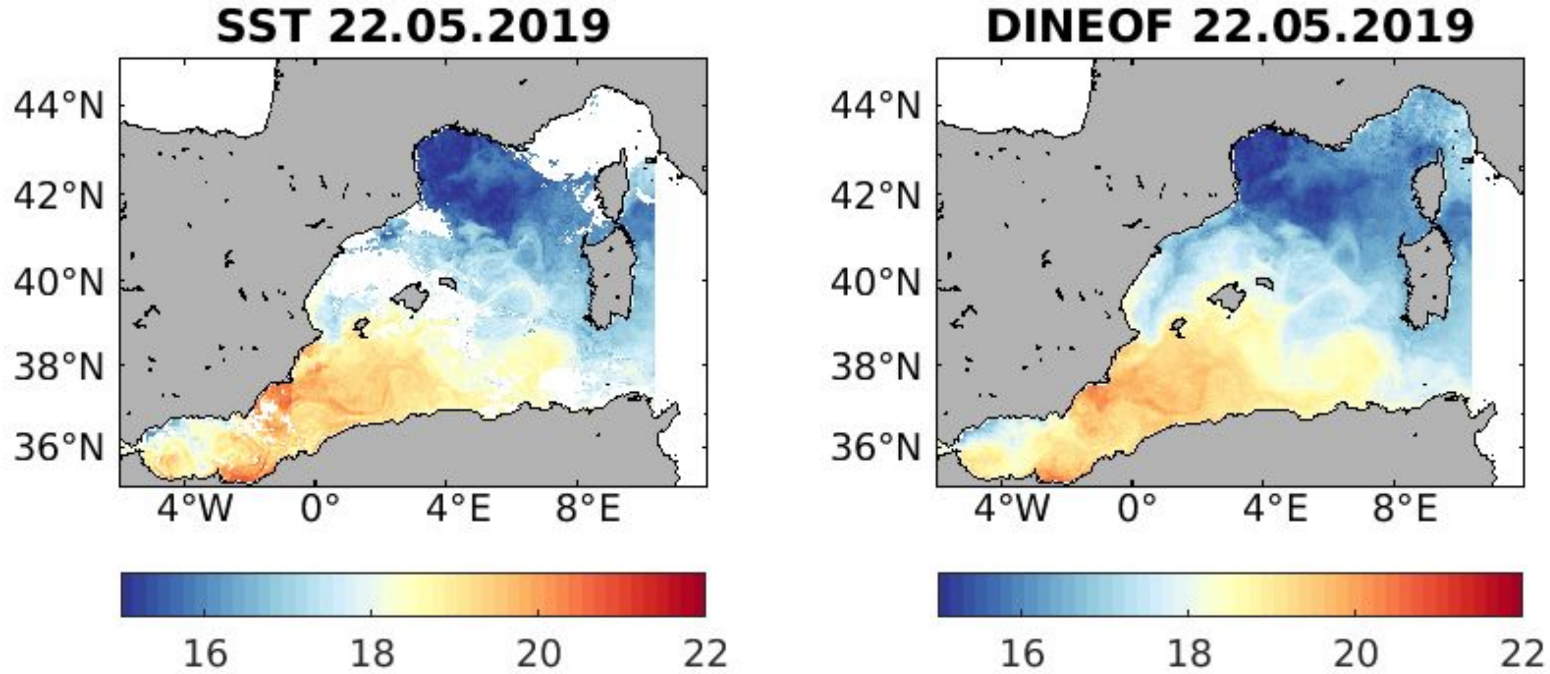


- Uses EOF basis to infer missing data:
    - **non-parametric, data-based**
    - No need of a priori information (correlation length, covariance function...)

- The spatio-temporal coherence present in the data is used to calculate missing values.
    → Three-dimensional data are used. Correlated information in space and time is used to infer missing data values.

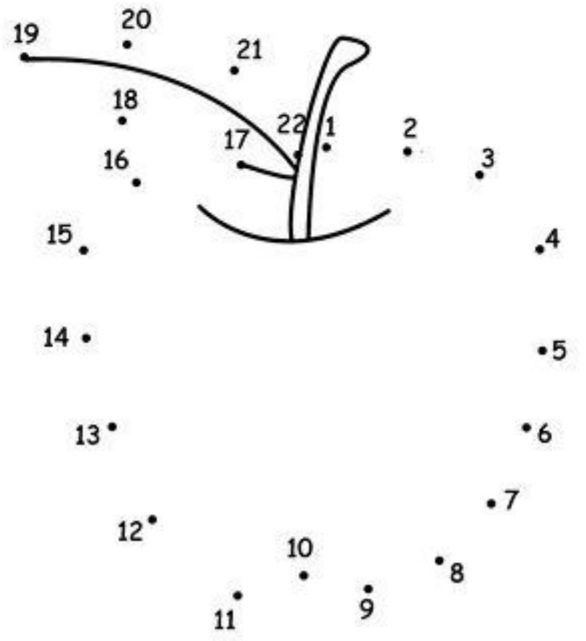


Mesoscale information in SST

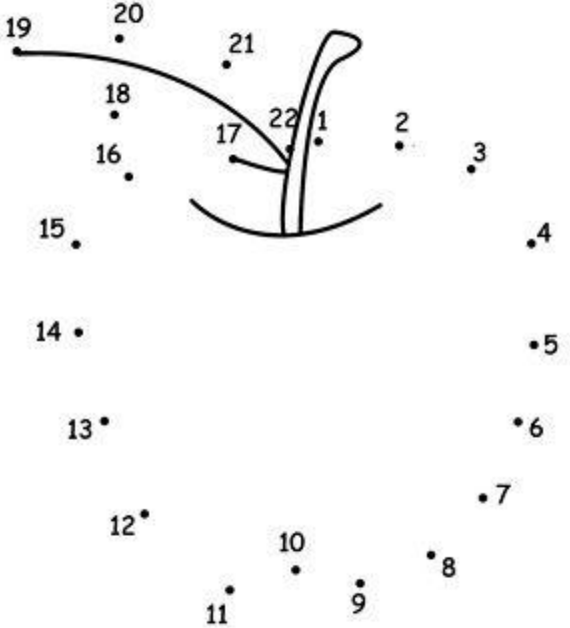


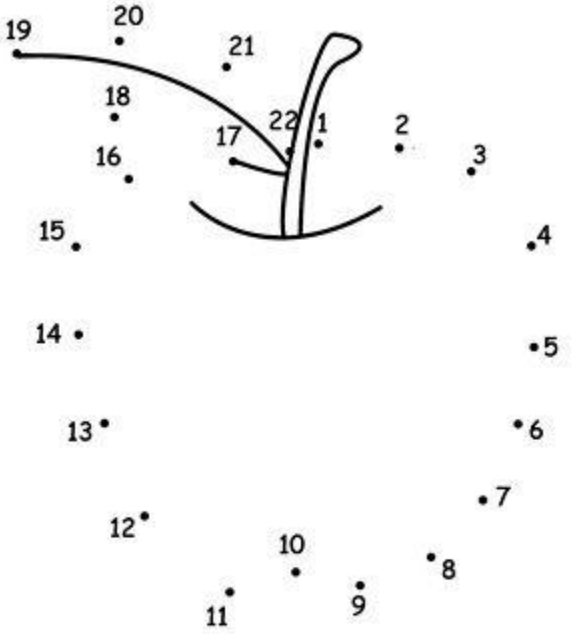
A word on resolved scales and their reconstruction

Let's play "join the dots!": what is the hidden image?

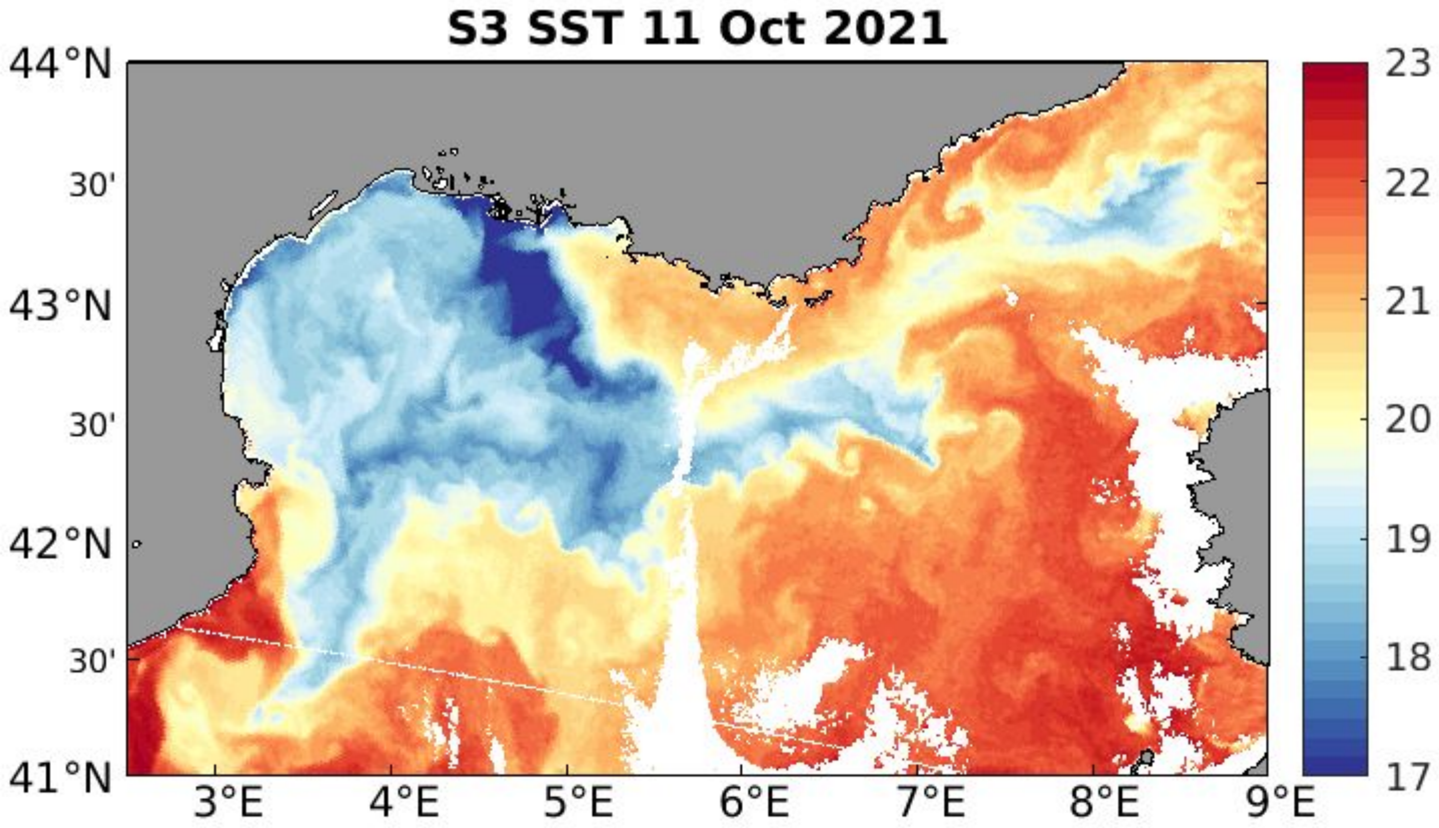


Q2: what is the hidden figure?


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22



We're only able to retain in the final reconstructed dataset those scales that are sufficiently represented in the initial dataset


S3 SST 11 Oct 2021
44°N
30'
43°N
30'
42°N
30'
41°N
3°E
4°E
5°E
6°E
7°E
8°E
9°E
23
22
21
20
19
18
17

Additions made to DINEOF

The “vanilla” method is, as said, parameter-free.

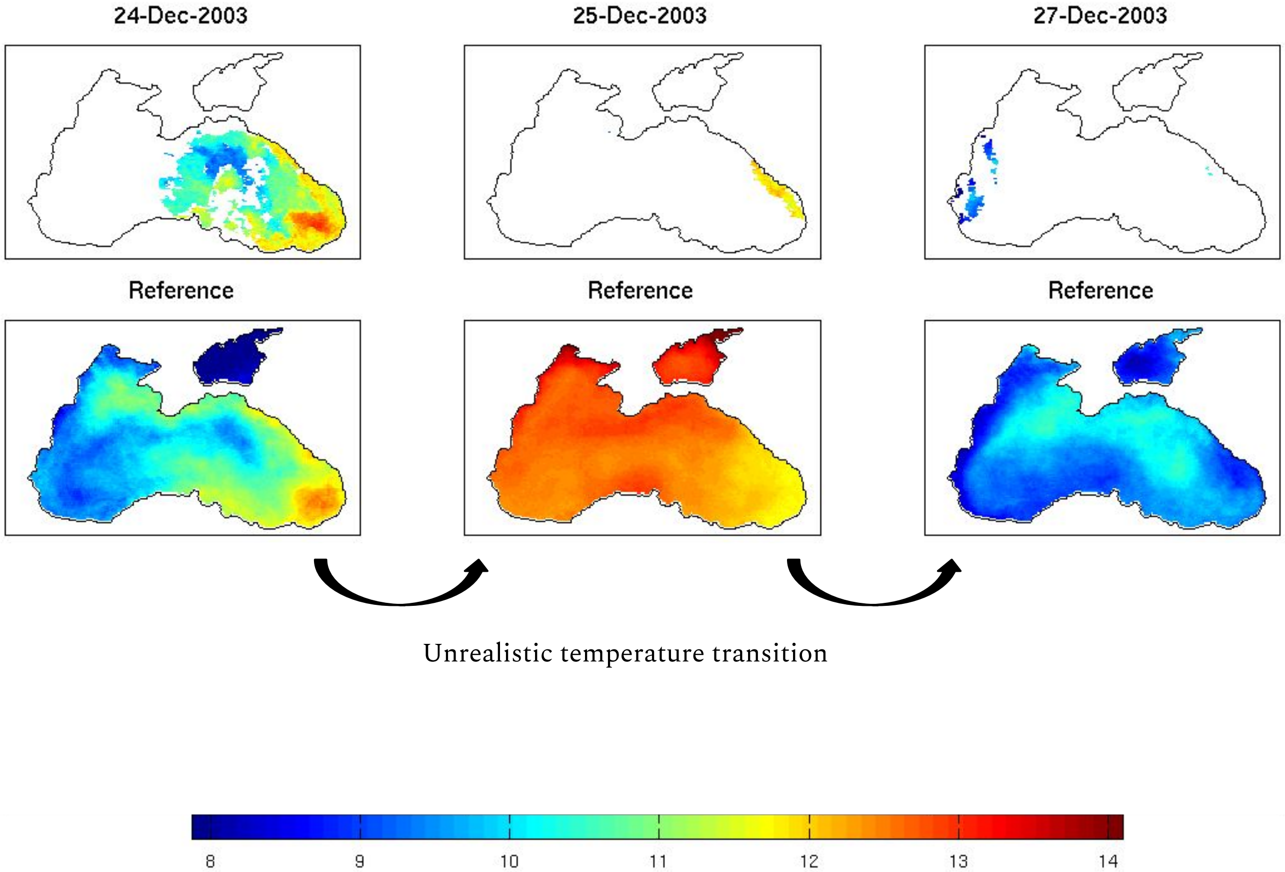
However, some small additions (with some parameterisation) lead to better results

We will briefly see those now:

- Temporal coherence of the reconstruction
- Outlier & shadow detection

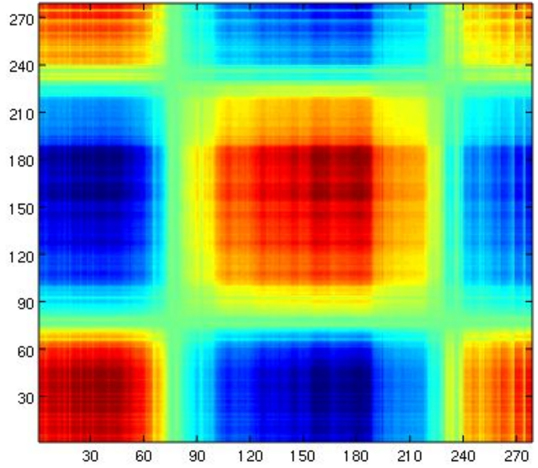
Enhancement of temporal coherence in DINEOF reconstructions

When too few data are present: temporal EOFs poorly constrained: unrealistic discontinuities

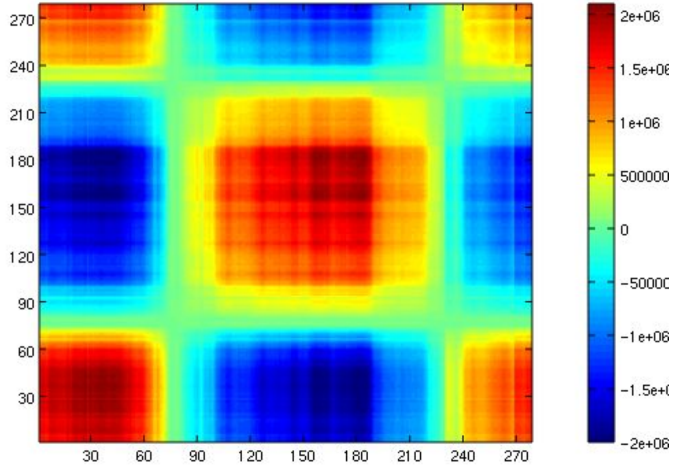


Unrealistic transitions are reflected in the covariance matrix ($\mathbf{C} = \mathbf{X}^T\mathbf{X}$)

→ filter to the temporal covariance matrix to reduce this



$$\rightarrow \qquad \mathbf{C}' = \mathbf{F}^T\mathbf{C}\mathbf{F} \qquad \rightarrow$$

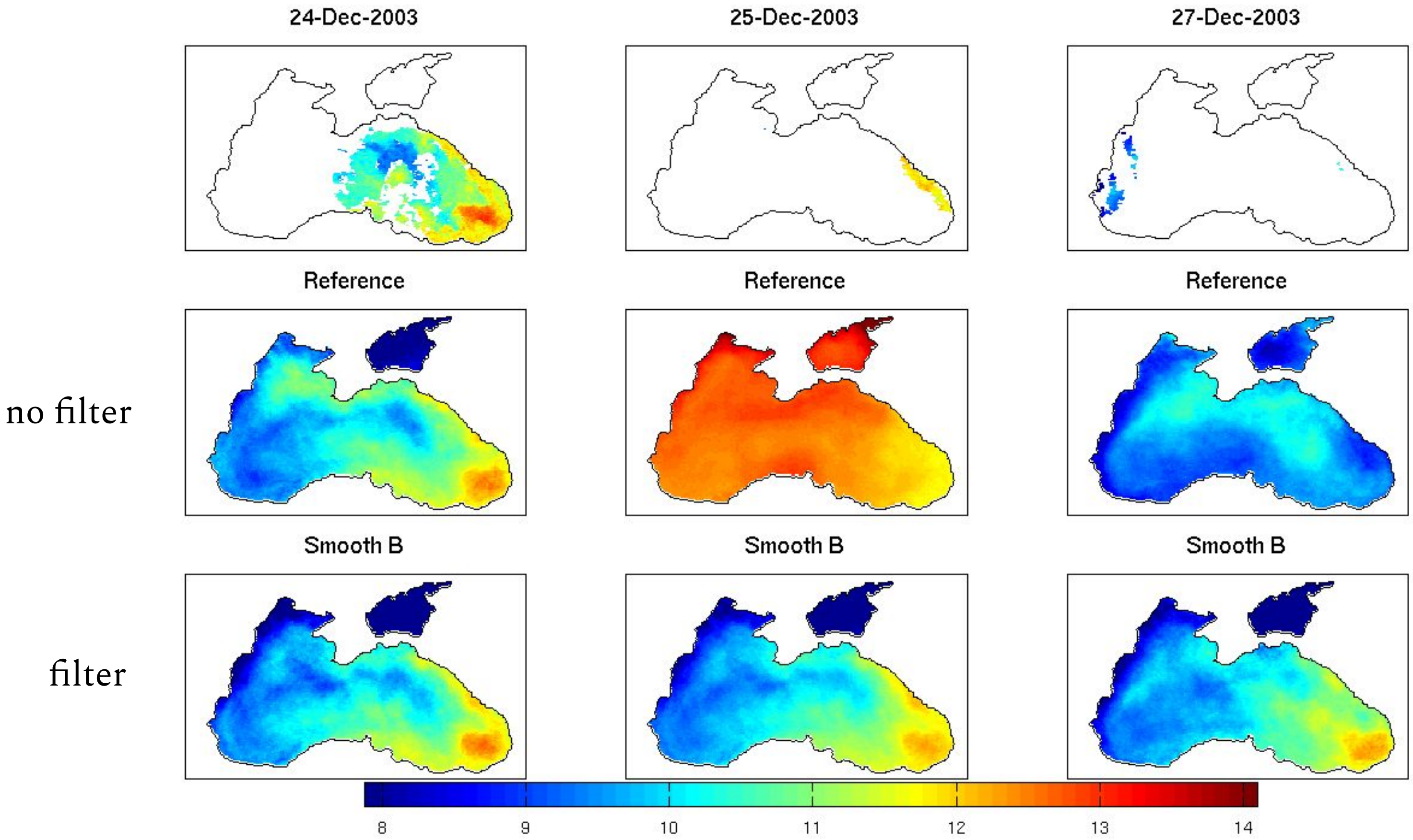


- F is a Laplacian filter
- Filter on **C** instead of **X**: **C** is much smaller and less sensitive to missing data
- Filter applied iteratively: more iterations, further reach of the filter

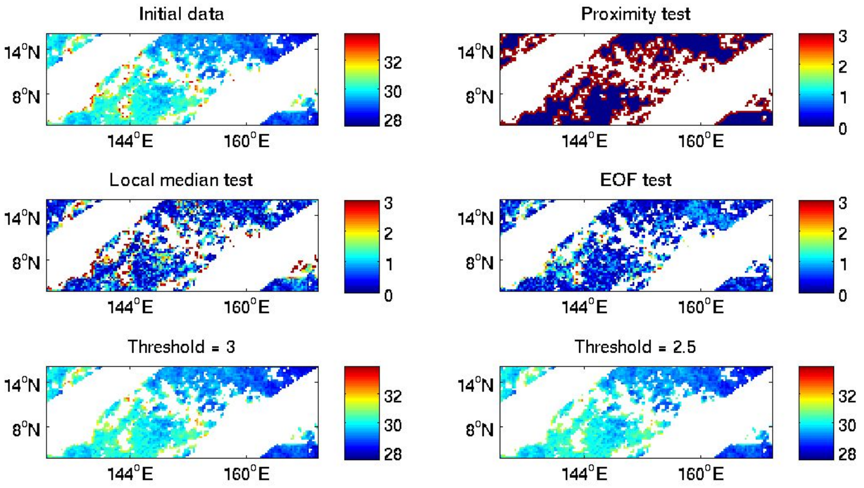
Alvera-Azcárate et al, 2009

Unrealistic transitions are removed efficiently using this filter

(in this case, the length of the filter was 1.1 days)



Other developments



Outlier detection

Based on EOF basis + median test + proximity tests

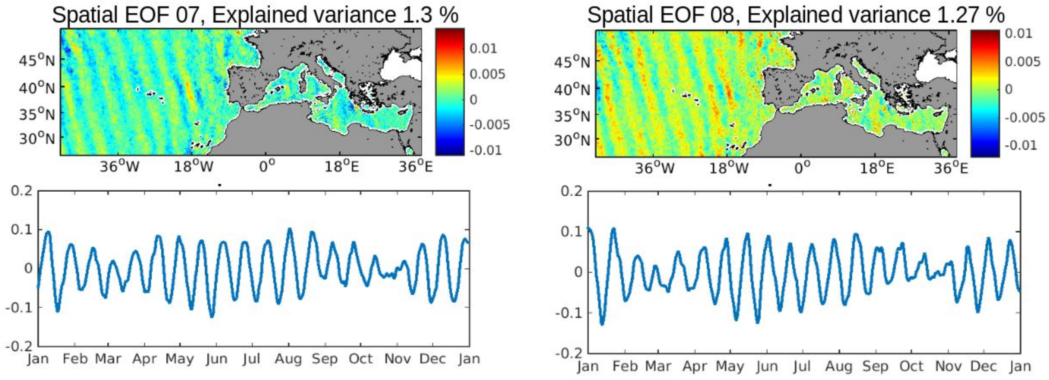
Allows for threshold decision on outliers

Removal of non-physical signals

If consistent biases present, EOFs can detect those (e.g. seasonal biases)

Removal of those EOFs improves quality of data

SMOS L2 data, biases at swath edges picked by repeat cycle



Shadow detection

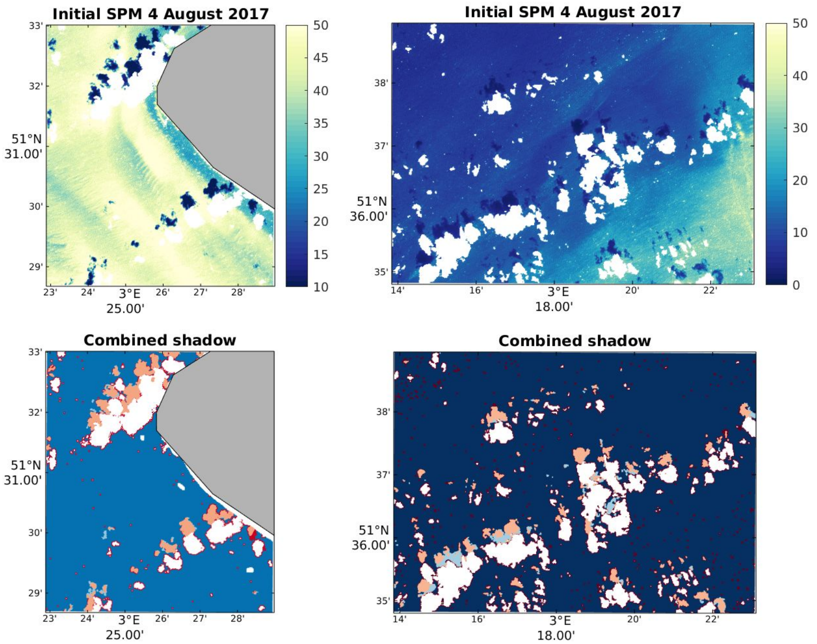
High resolution satellite data (e.g. Sentinel-2 with 10 m resolution) resolve cloud shadows

Difficult to remove because pixels have a “correct” spectral information

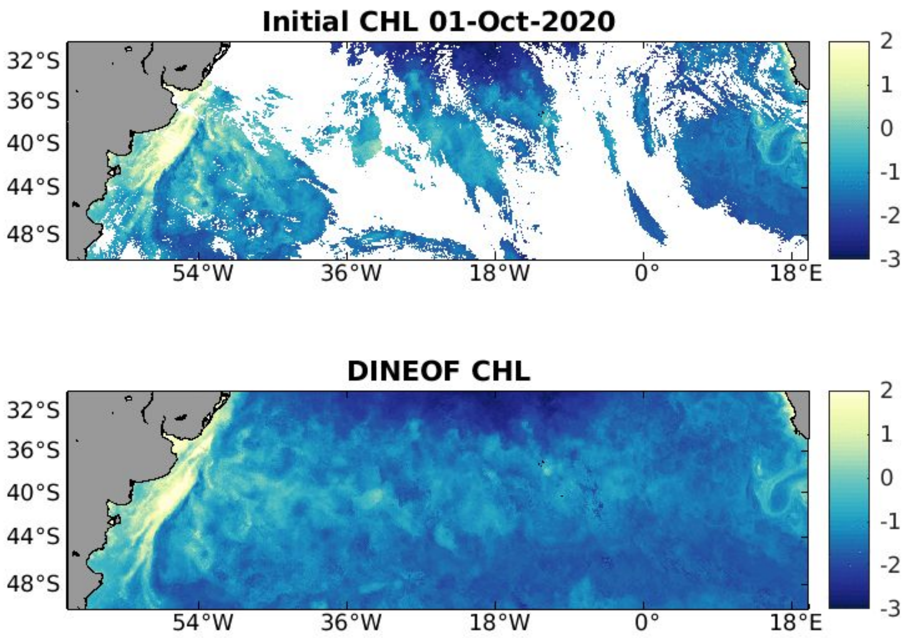
EOF basis can be used to detect and remove cloud shadows

Additional tests:

- Low values penalised
- Departure from median
- Ray tracing



Alvera-Azcárate et al, 2021



In short...

DINEOF is a reliable method for filling missing data.
It's been used, developed & improved for many years.
Several applications for data quality improvement have been developed from DINEOF

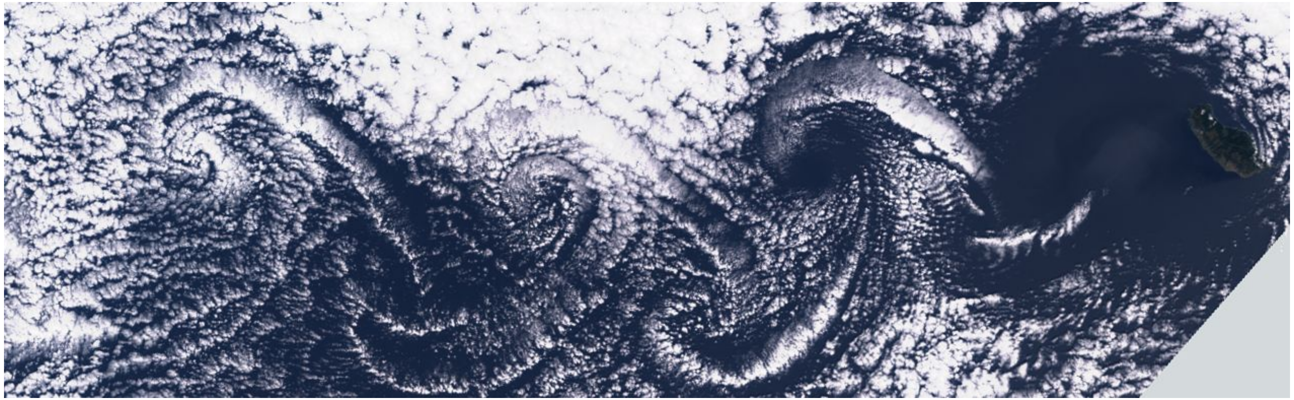
Data-Interpolating Convolutional Auto-Encoder (DINCAE)

Q4: have you worked with neural networks before?

Objectives

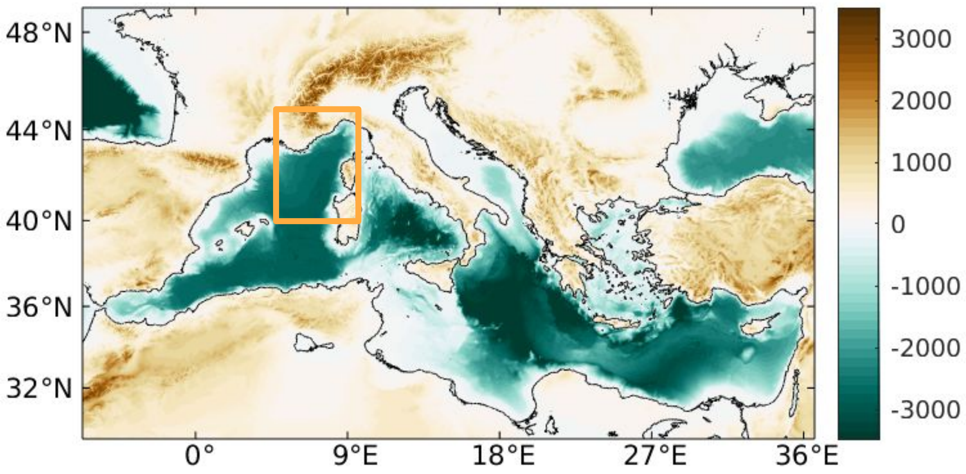
- To derive a methodology to reconstruct missing information in satellite data
    - Based on **neural networks**
    - Making use of ~four decades of sea surface temperature measurements
    - Able to **retain small scale variability**

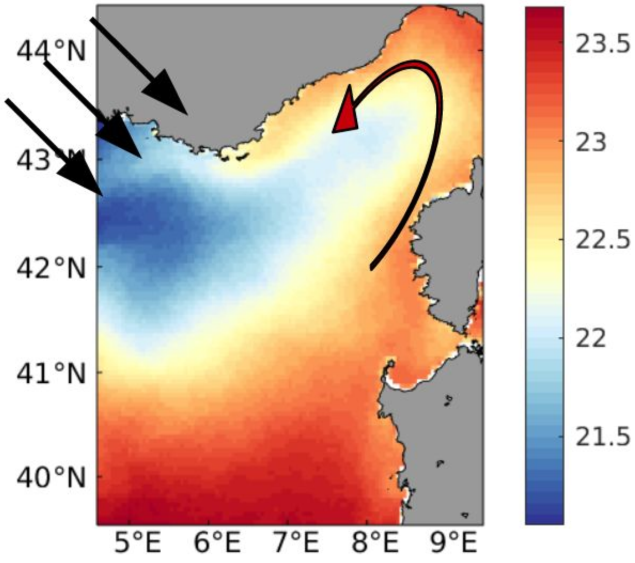
- To assess the benefit of using neural networks in comparison with other state-of-the-art methodologies
    - DINEOF (Data Interpolating Empirical Orthogonal Functions)



Data used

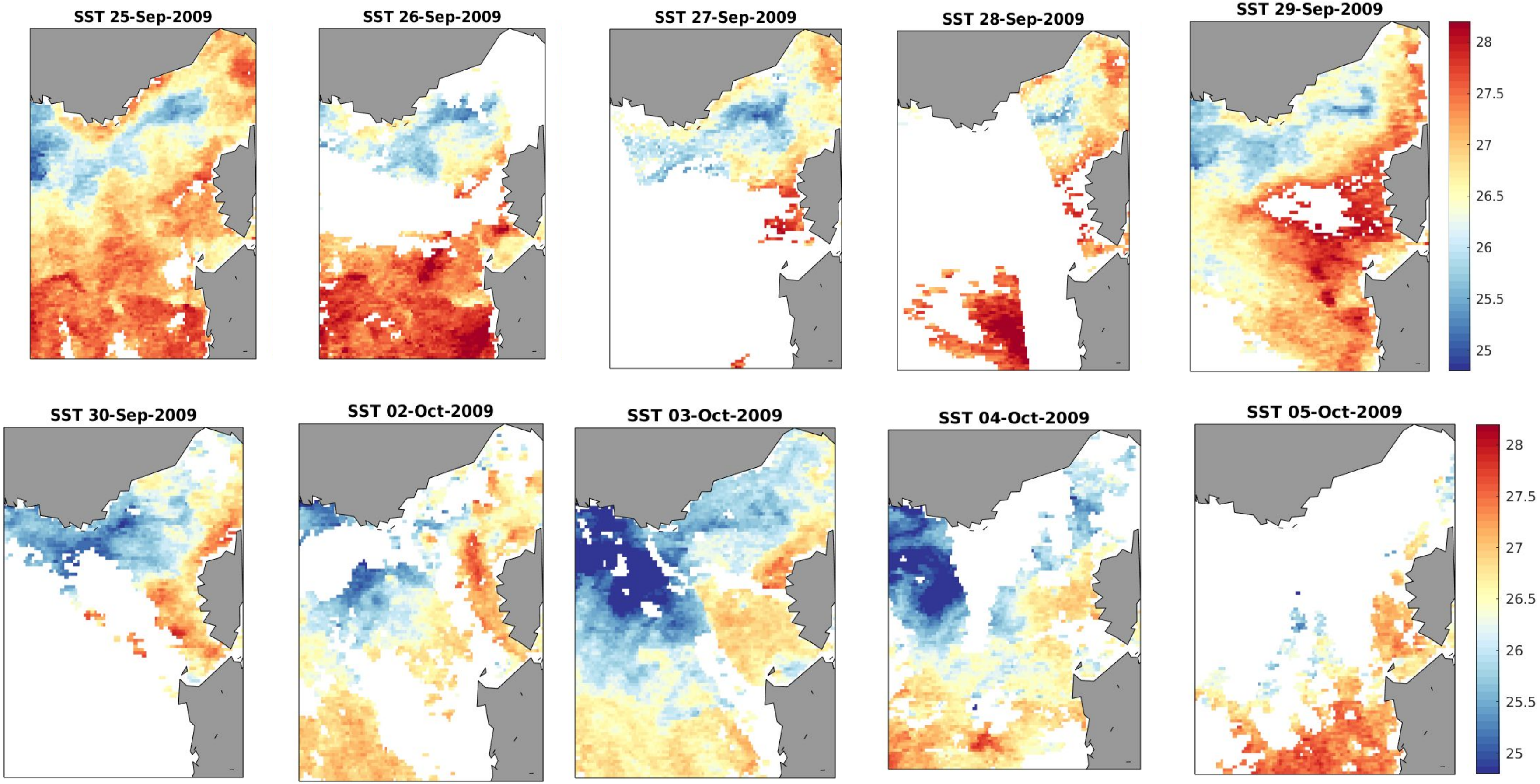
- Daily Advanced Very High Resolution Radiometer (AVHRR) Sea Surface Temperature (SST) data
- **4 km spatial resolution**
- **Liguro-Provençal basin** (western Mediterranean Sea)
- 1 April 1985 to 31 December 2009 (**25 years**) -> **longest homogenous time series**
- **47 % of missing data**





Average SST (°C)

Challenge: **training on gappy data (lots of gaps!)**



The Bayes' rule or how to handle information of different accuracy

For **Gaussian-distributed errors**:

- prior: $\mathcal{N}(x^f, \sigma^f)$
- observations: $\mathcal{N}(y^o, \sigma^o)$
- posterior: $\mathcal{N}(x^a, \sigma^a)$

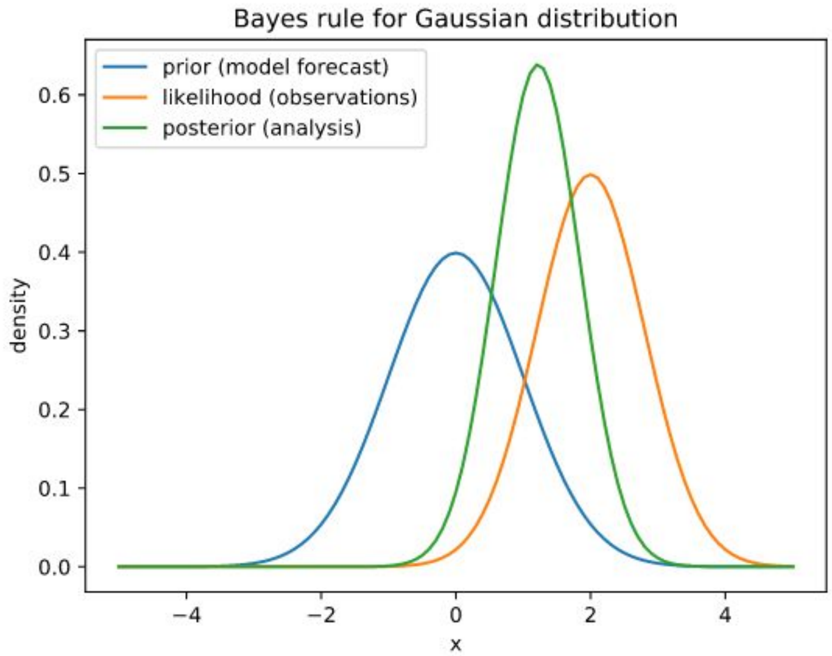
Bayes' rule:

$$p(x|y^o) = \frac{p(x)p(y^o|x)}{p(y^o)}$$

- Mean and variance of posterior given by:

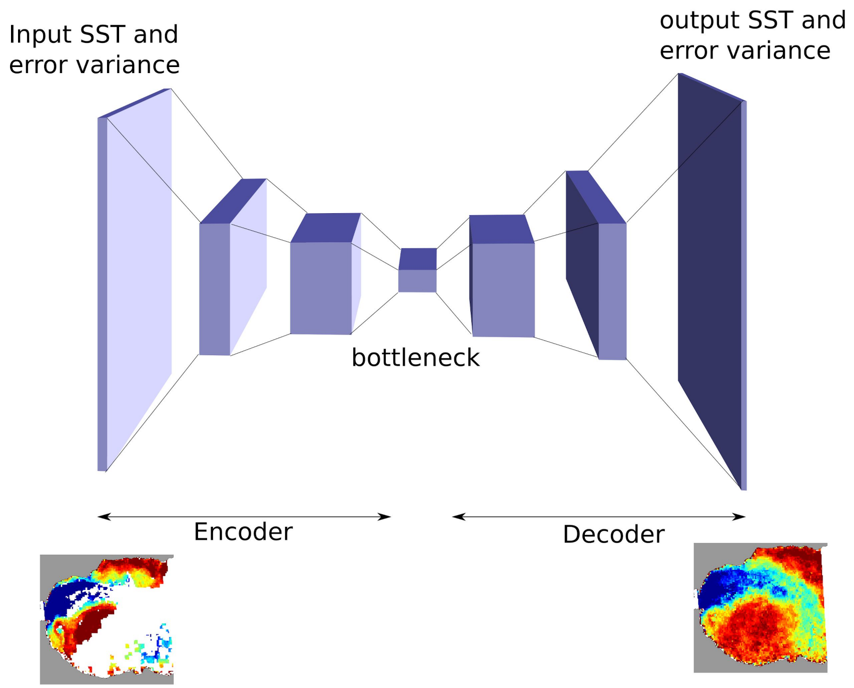
$$\sigma^{a^{-2}} x^a = \sigma^{f^{-2}} x^f + \sigma^{o^{-2}} y^o$$
$$\sigma^{a^{-2}} = \sigma^{f^{-2}} + \sigma^{o^{-2}}$$

- **Inverse of the variance are simply added linearly**



Methodology

DINCAE: Data-Interpolating Convolutional Auto-Encoder



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)

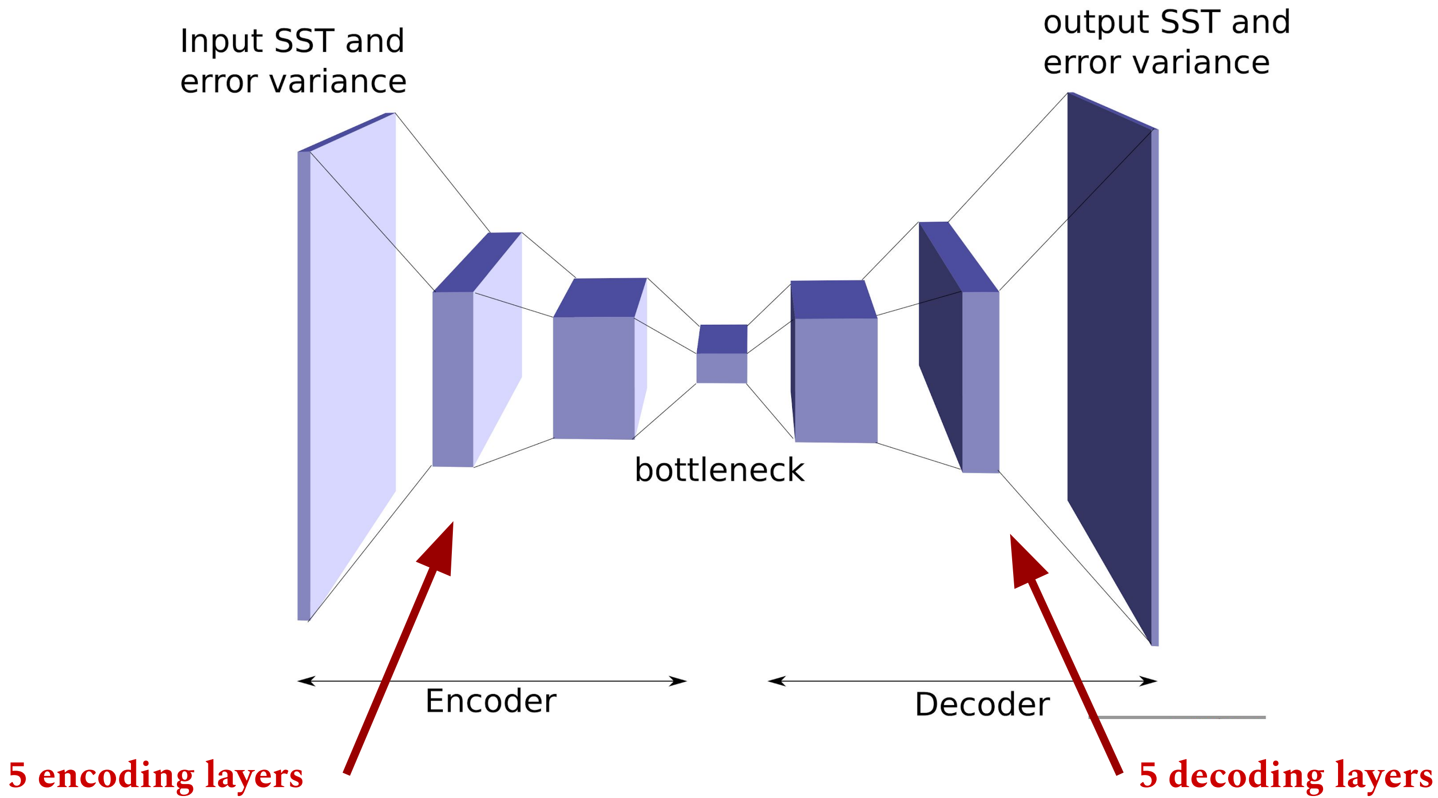
Convolutional: works on subsets of data, i.e. trains on local features

Missing data handled as data with different initial errors

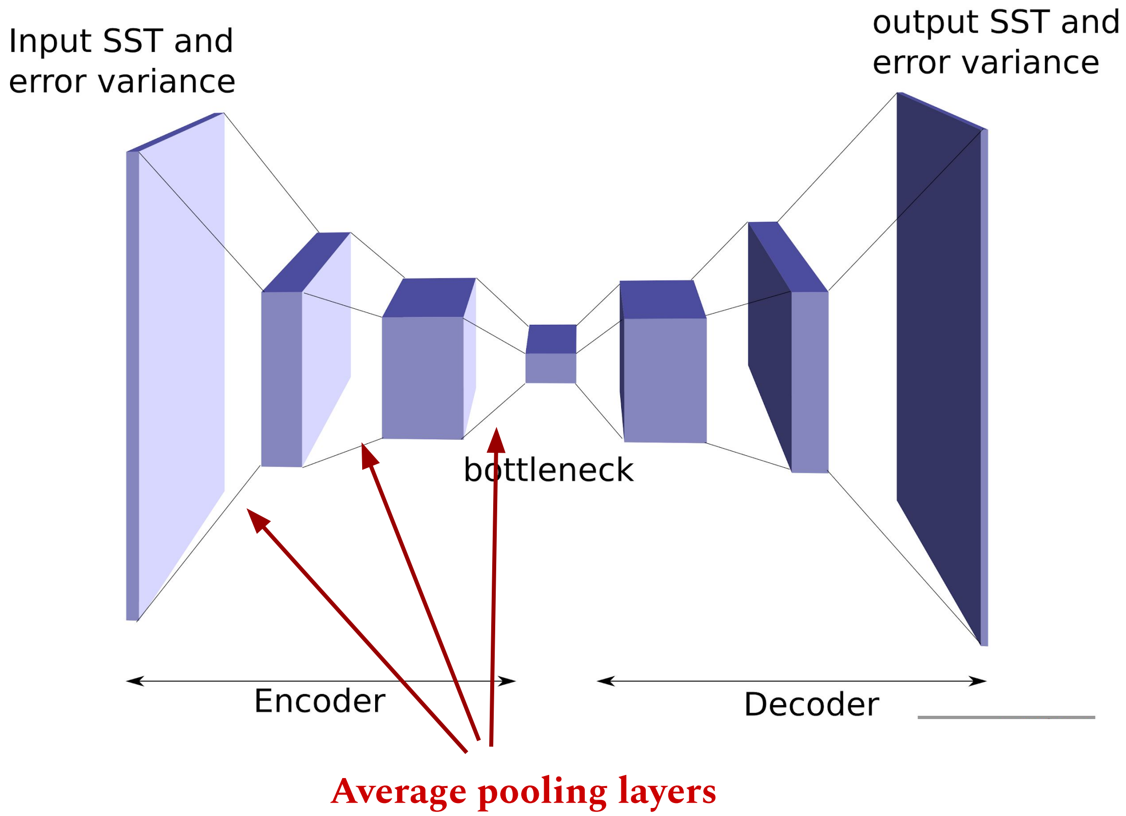
- If **missing, error variance** ($\boldsymbol{\sigma^2}$) **tends to infinity**

Input data:

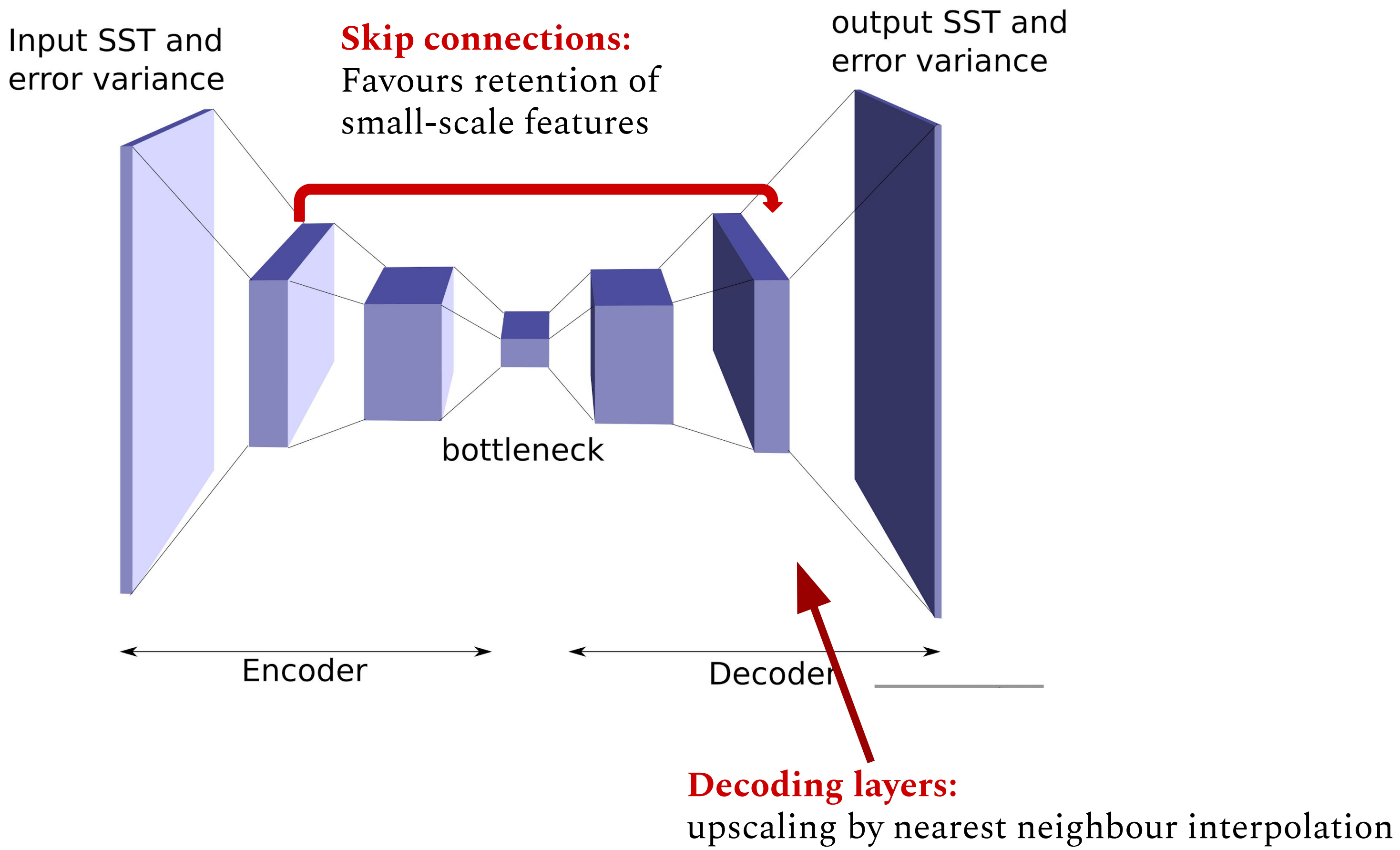
- SST/σ^2 (previous day, current day, following day)
- 1/σ^2 (previous day, current day, following day)
- Longitude
- Latitude
- Time (cosine and sine of the year-day/365.25)



3x3 convolutional filters applied at each layer



Reduce size by retaining the average value on 2x2 boxes


Input SST and error variance
Skip connections:
Favours retention of small-scale features
output SST and error variance
bottleneck
Encoder
Decoder
Decoding layers:
upscaling by nearest neighbour interpolation

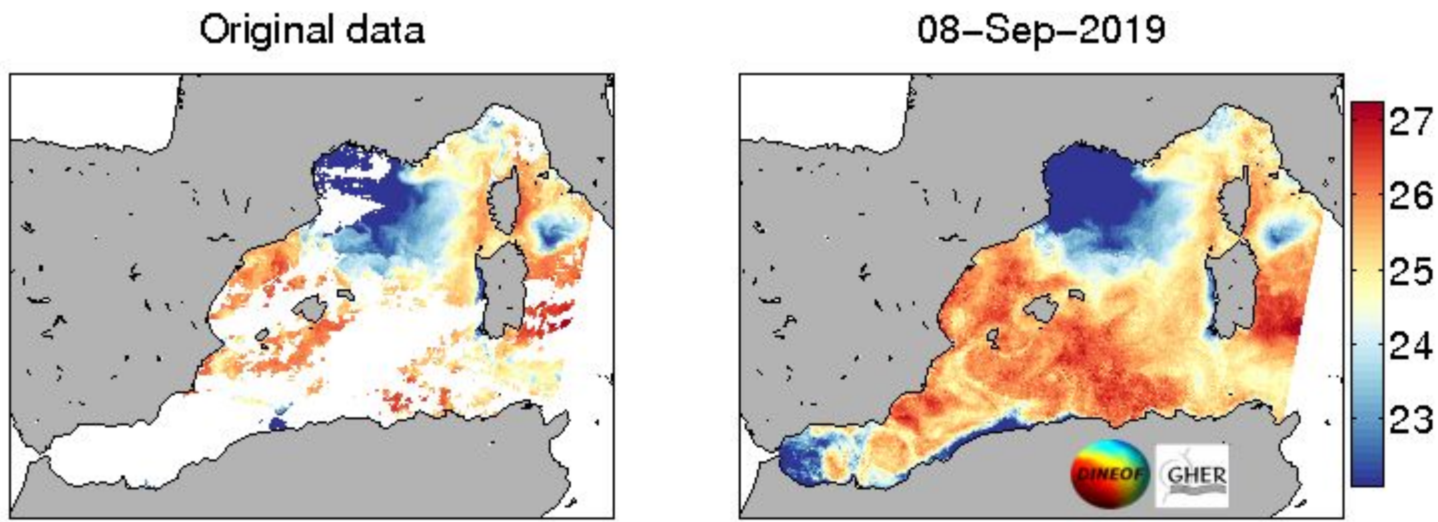
Baseline method to be improved

DINEOF (Data Interpolating Empirical Orthogonal Functions)

A reconstruction method based on the EOF basis from the dataset

~15 years of development & improvements

http://www.dineof.net/DINEOF/



Training

- Partitioned into so-called **mini-batches** of 50 images
- The entire dataset is used **multiple times (epochs)**
- For every input image, **more data points were masked** (in addition to the cross-validation) by using a **randomly chosen cloud mask during training** (data set augmentation).
- The output of the neural network (for every single grid point i,j) is a **Gaussian probability distribution** function characterized by a mean $\hat{y}_{ij}$ and a standard deviation $\hat{\sigma}_{ij}$.

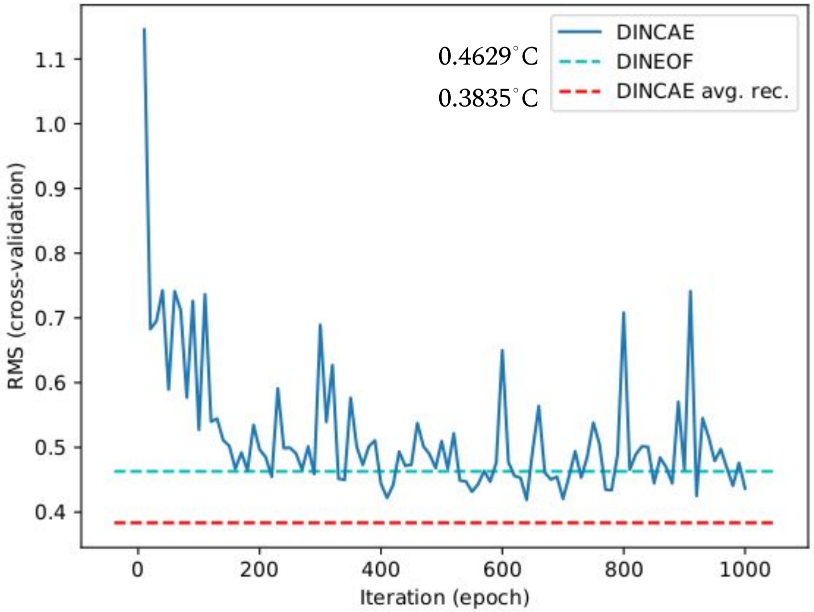
$$J(\hat{y}_{ij}, \hat{\sigma}_{ij}) = \frac{1}{2N} \sum_{ij} \left[\left(\frac{y_{ij} - \hat{y}_{ij}}{\hat{\sigma}_{ij}} \right)^2 + \log(\hat{\sigma}_{ij}^2) + 2\log(\sqrt{2\pi}) \right]$$

- The first term: mean square error, but scaled by the estimated error standard deviation.
- The second term: penalizes any over-estimation of the error standard deviation.

Results

Cross-validation: data removed from the last **50 images of the times series** (with cloud mask from first 50 images)

Averaging epochs 200 to 100 improved DINCAE results



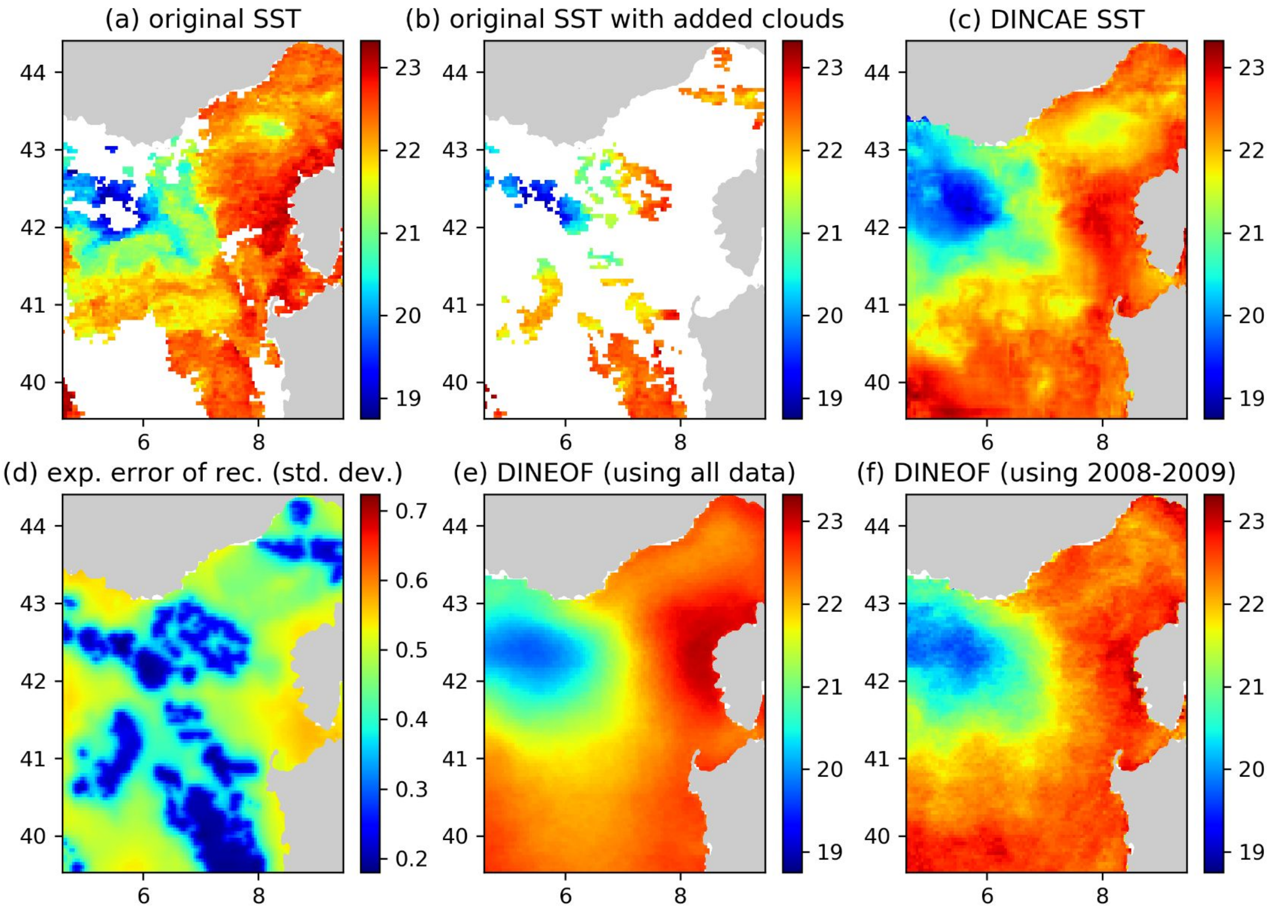
Reconstruction results -full time series- compared to WOD in situ data (under clouds)

RMS (DINEOF) 1.1676°C

RMS (DINCAE) 1.1362°C

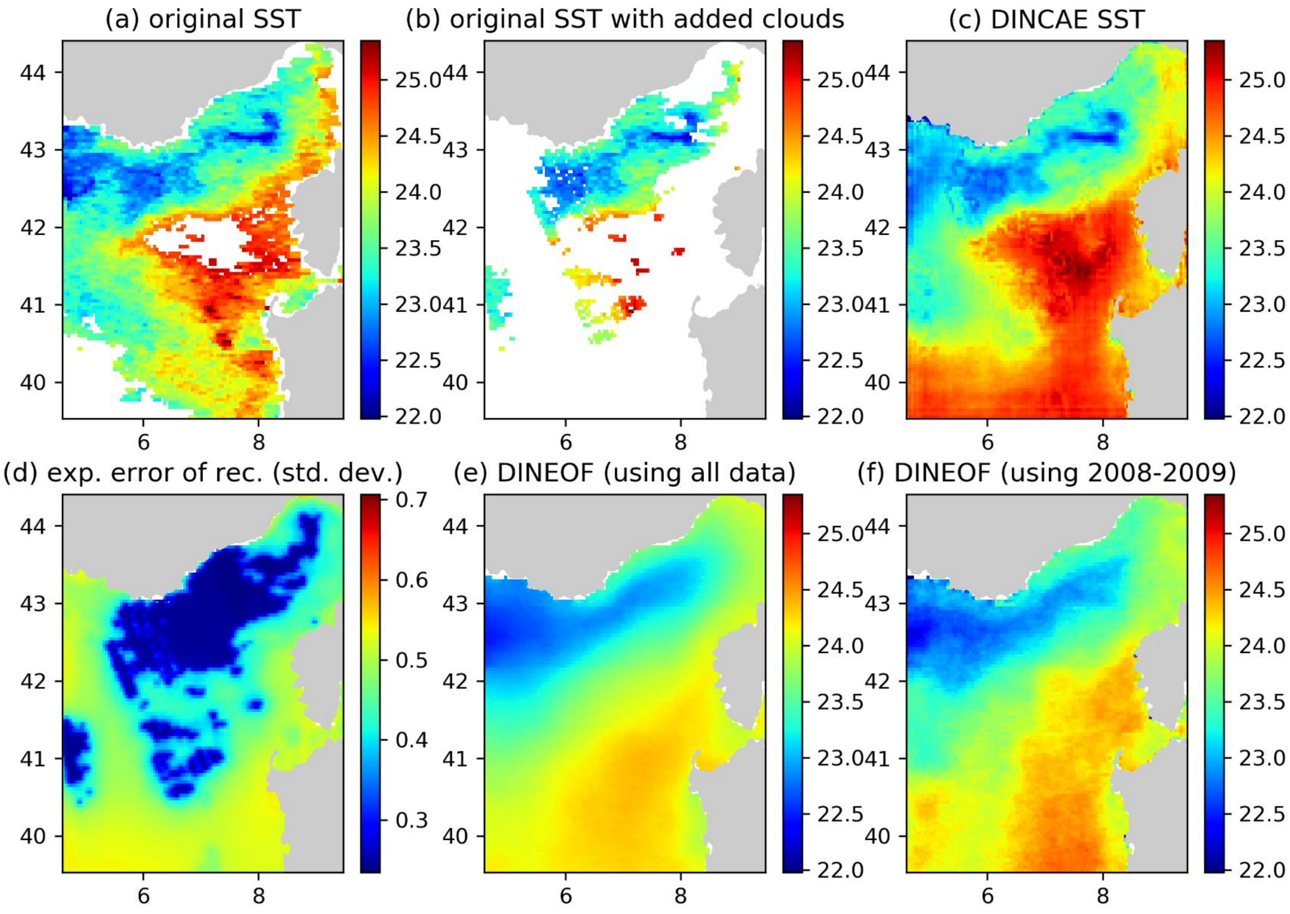
Results

Reconstruction examples



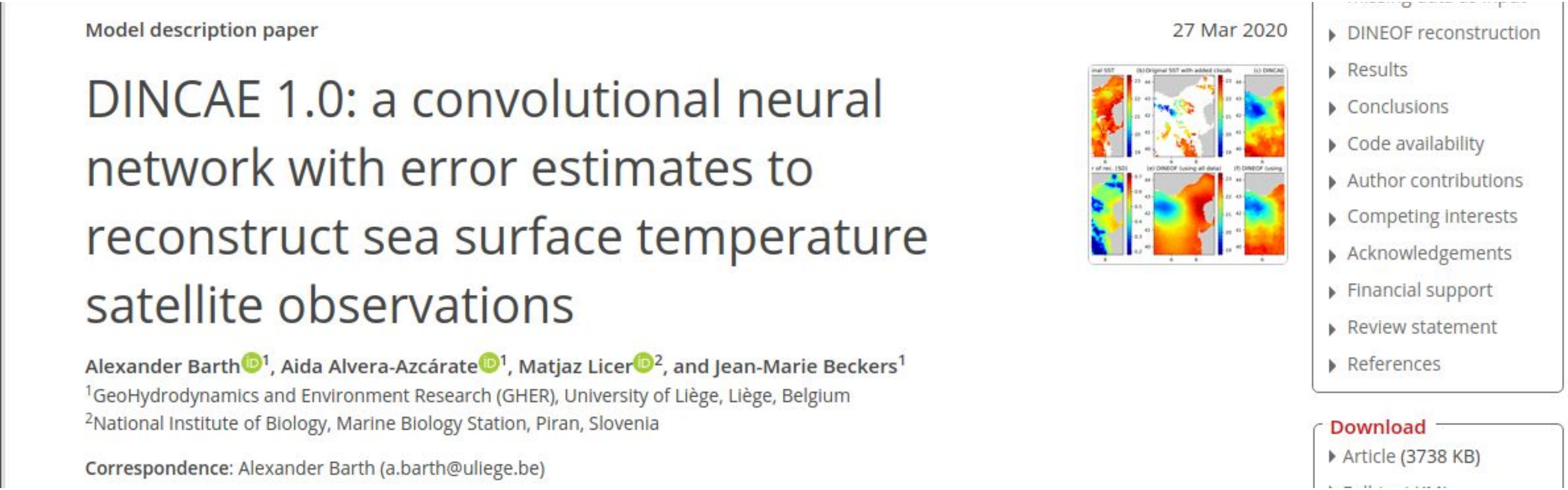
Results

Reconstruction examples



If you want to know more...

- Manuscript in GMD: https://doi.org/ghf3cd



Model description paper

27 Mar 2020

DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations

Alexander Barth[1], Aida Alvera-Azcárate[1], Matjaz Licer[2], and Jean-Marie Beckers[1]

[1]GeoHydrodynamics and Environment Research (GHER), University of Liège, Liège, Belgium

[2]National Institute of Biology, Marine Biology Station, Piran, Slovenia

Correspondence: Alexander Barth (a.barth@uliege.be)

Download
- Article (3738 KB)

- Python Code available at:

https://github.com/gher-ulg/DINCAE (currently rewritten in Julia)

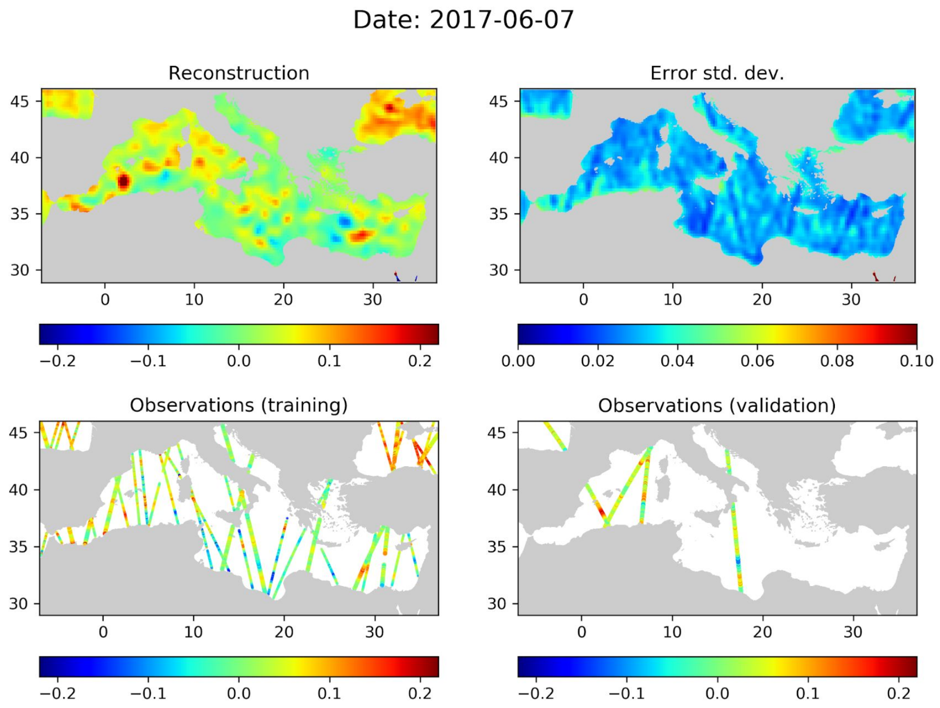
Unstructured data

Altimetry data from 1993-01-01 to 2019-05-13 from CMEMS

Multiple satellites missions

- 70% training data (determine weight of the networks)
- 20% developpement data (determine structure of the network,...)
- 10% test data (independent validation)

Structure of the network determined by Bayesian optimization



Validation

Reasonable **good match** with the validation data

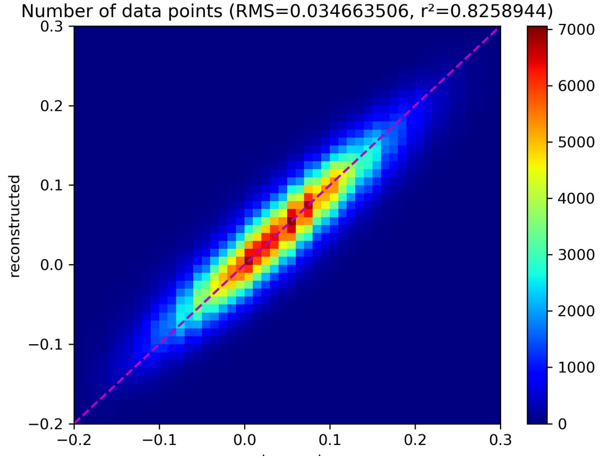
Reliable expected reconstruction errors are notoriously hard to obtain from methods like optimal interpolation

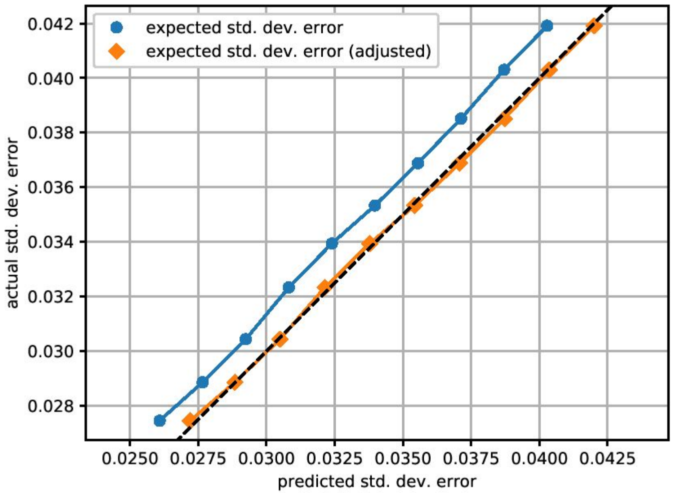
DINCAE also provide as expected error of the reconstruction (per pixel)

The validation data has been **grouped into bins** using the expected error

For every bin the **standard deviation of the actual error** has been computed

The predicted error underestimates the actual error only by 4%



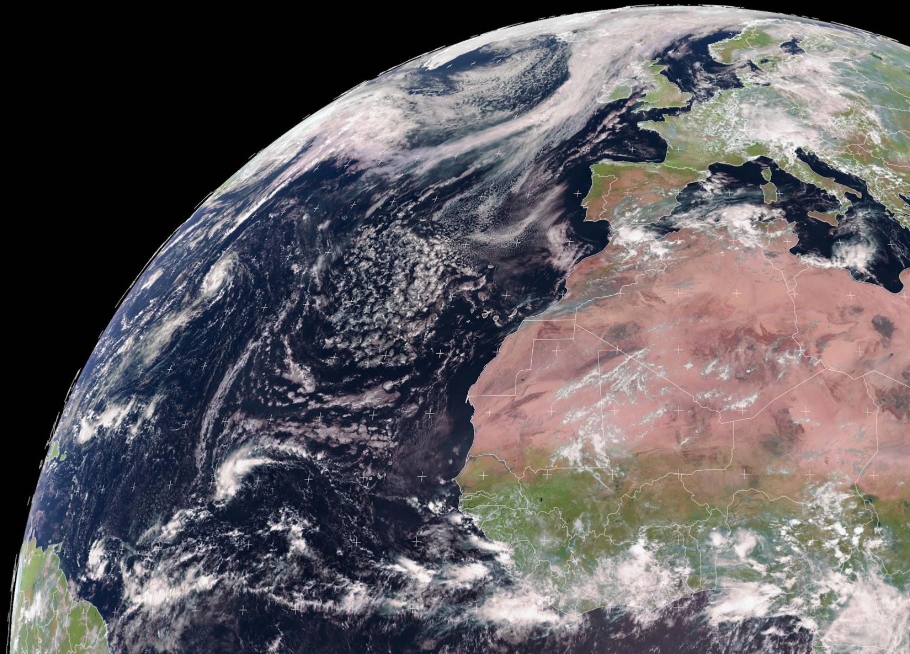


Conclusions

DINEOF:
- A reliable method for filling missing data.
- It's been used, developed & improved for many years.
- Several applications for data quality improvement have been developed from DINEOF
  Outlier detection, temporal filter, shadow detection…

DINCAE:
- A convolutional Autoencoder approach to reconstruct missing data
- Missing data handled by including expected error variance in the input data
- Estimation of missing data + estimation of error of the reconstruction obtained

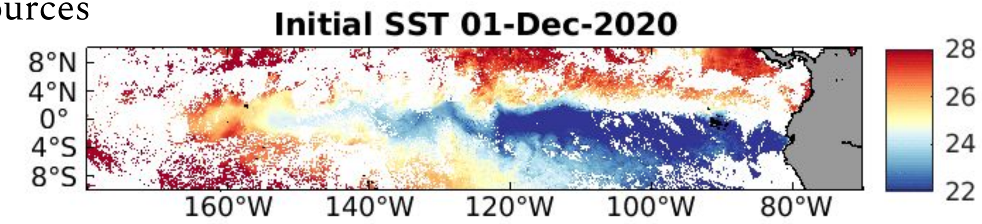
Both methods aim to compress the data into a low dimensional subspace and they reconstruct a full field from this compressed representation.



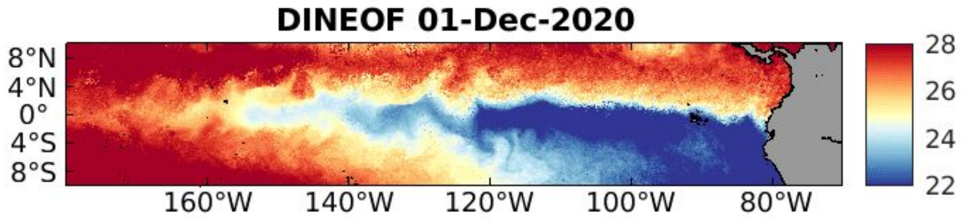
DINEOF or DINCAE???

DINCAE
- Shows promise and adaptability to new challenges (e.g. unstructured data)
- Needs a powerful computer with GPU(s)
- Only on very small regions, very long time series needed for training

DINEOF
- Fast, reliable, 20-yrs development behind
- Can do large regions with few computing resources
- Short time series are enough



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



Jupyter notebooks for a test:

http://dineof.net/Temp/DINEOF_test.zip

Data for the exercices: multisensor L3 SST from CMEMS

https://marine.copernicus.eu/

Global or European regions:

A 3D dataset (lon,lat,time) extracted from here should work directly in the example notebooks