Reconstruction of missing data in satellite images of the Southern North Sea using a convolutional neural network (DINCAE)

Alexander Barth, Aida Alvera-Azcárate, Charles Troupin,
Jean-Marie Beckers, Dimitry Van der Zande

GHER, University of Liège, Belgium
RBINS, Belgium

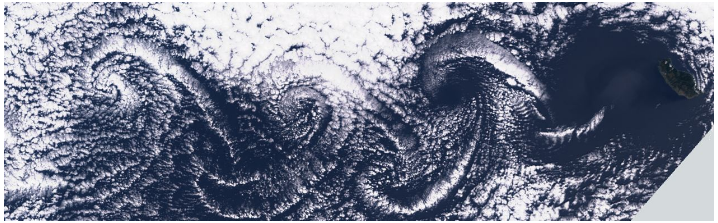
Objectives

- **Training a neural network to infer missing data in satellite observations**

- Training a neural network
    - From **model data** (complete; but affected by errors and biases)
    - From **observations** (**incomplete**; still possibly affected by errors and biases; but to a lesser degree)

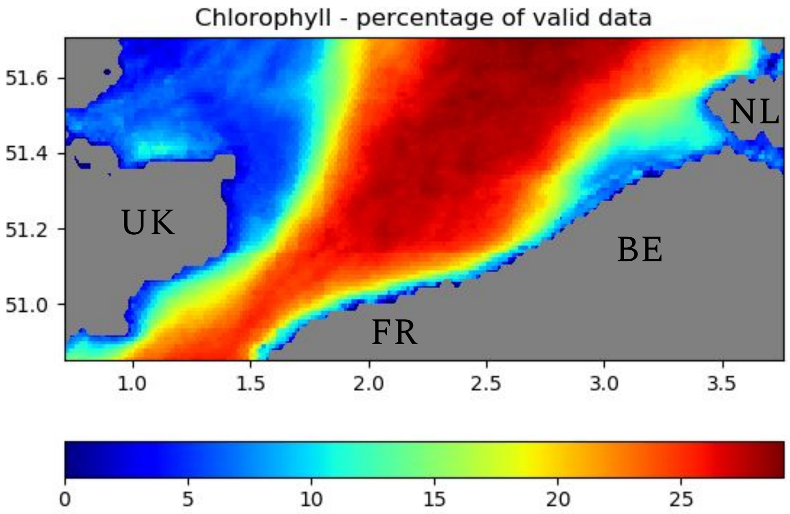
- Different approaches. Neural network is either
    - the **method** to created a complete field (input: present data)
    - a representation of the the **field** (input coordinates; see e.g. physics-informed neural networks)

- Neural network should be able to provide a complete field ("analyse") based on satellite data
        - Able to **retain small scale variability**
        - Just the surface here

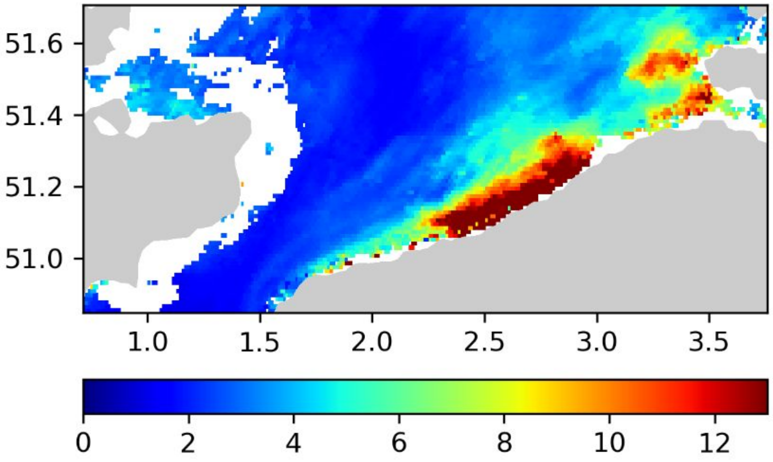


Data used

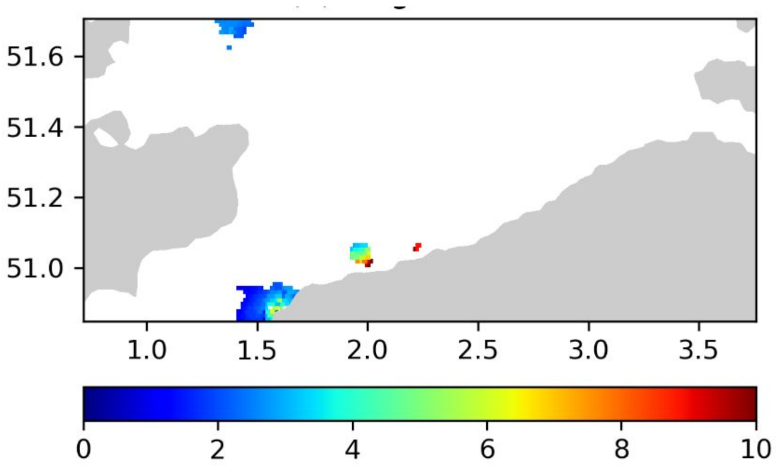
- Southern part of the North Sea
- Dynamics are strongly influenced by tides and riverine inputs
- **Chlorophyll-a** (20 years, 1998 to 2017, 4998 images). Daily time resolution.
- In total, 19 % of valid data (for sea points)



2010-07-17 (exceptional coverage)



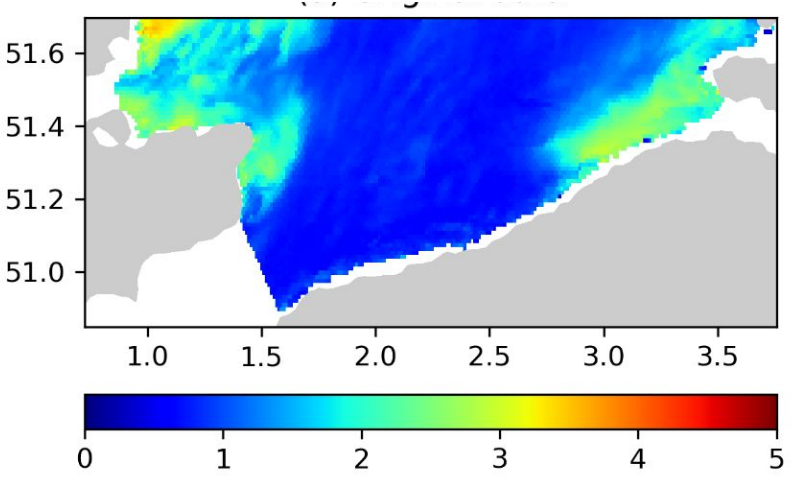
2005-07-17 (typical coverage)

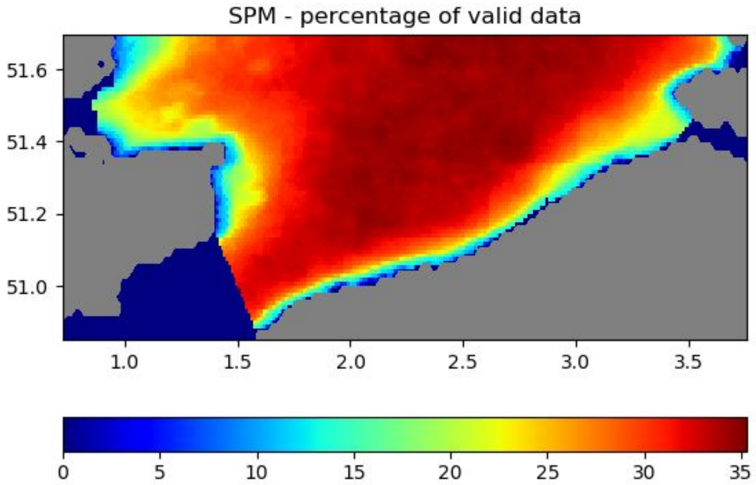


Data used

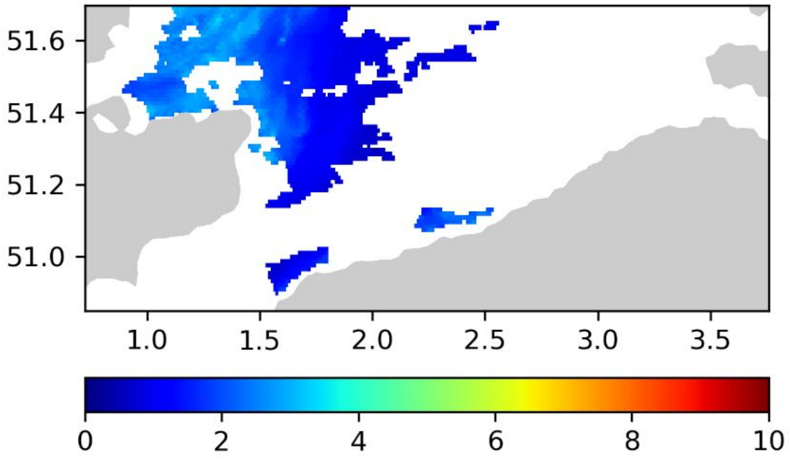
- Concentration of suspended matter (1998 - 2017, 4690 images)
- Daily time resolution.
- In total, 29 % of valid data (for sea points)

2002-05-16 (exceptional coverage)



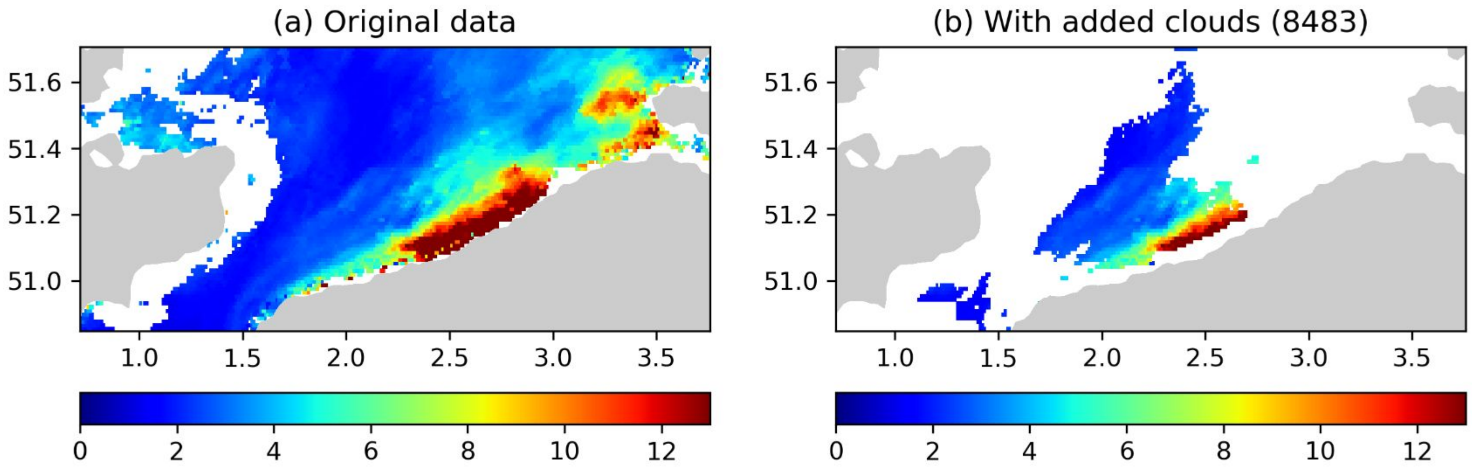


2017-09-15 (typical coverage)



Cross-validation data used

- To estimate the accuracy of the reconstruction:
  - Some additional pixels of the satellite images are **marked as missing**
  - On the **image with the lowest number of clouded pixels**, the cloud **mask from a different day** (chosen at random) is used to mark additional grid points as missing.
  - Data withheld for validation has a **realistic spatial extent and shape**.



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: $\mathcal{N}(x^a, \sigma^a)$

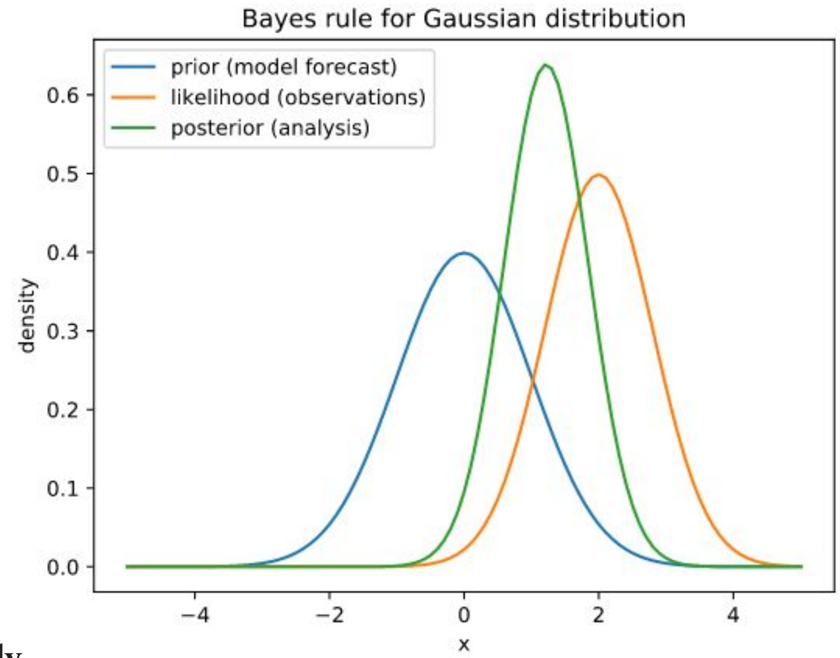
Bayes' rule:

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

- Mean and variance of posterior given by:

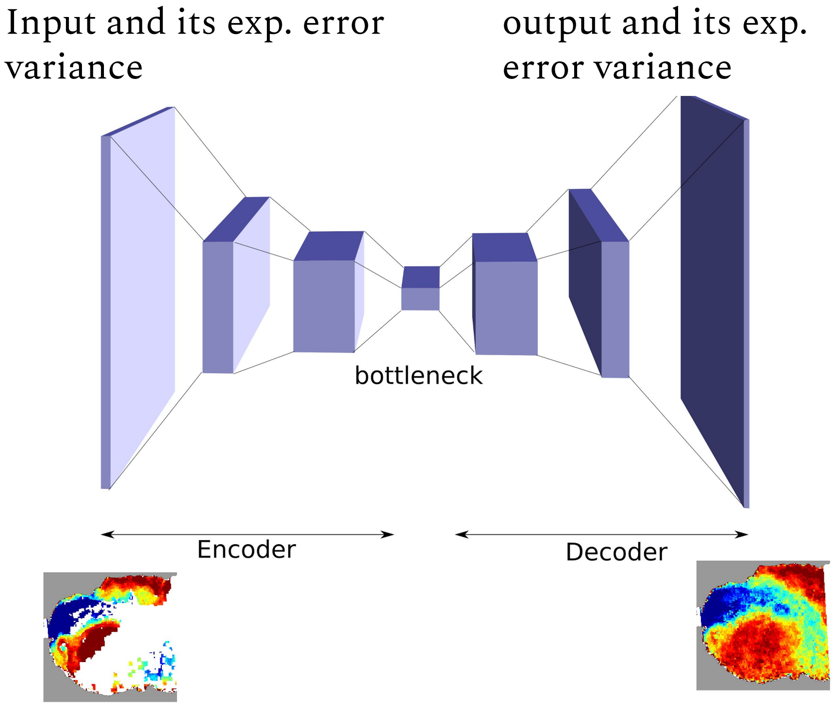
$$\sigma^{a^{-2}} x^a = \sigma^{f^{-2}} x^f + \sigma^{o^{-2}} y^o$$
$$\sigma^{a^{-2}} = \sigma^{f^{-2}} + \sigma^{o^{-2}}$$

- **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**

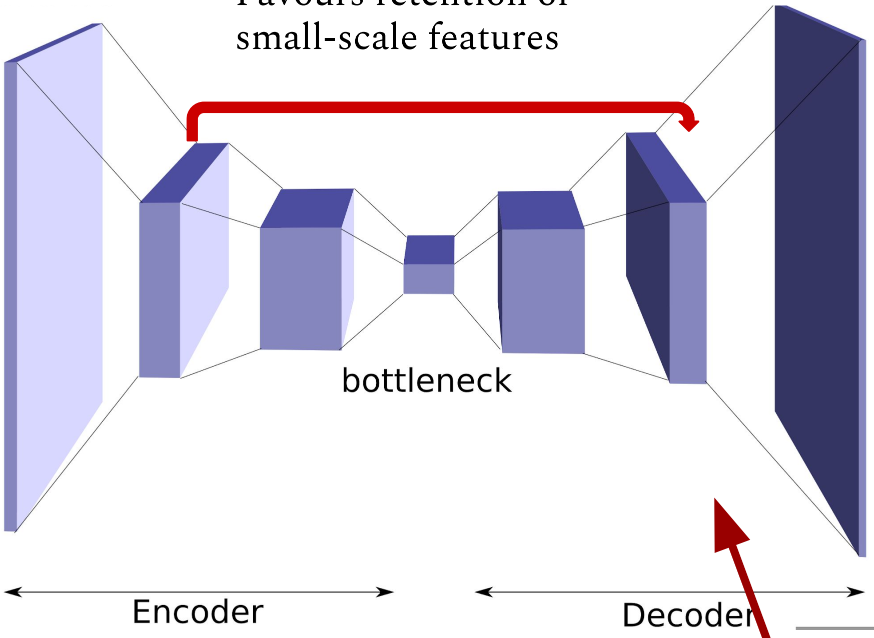
Input data:

- obs./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)

Input and its exp. error variance

Skip connections:
Favours retention of small-scale features

output and its exp. error variance



Decoding layers:
upscaling by nearest neighbour interpolation

Convolutional layer: linear transformation working a small (e.g. 3 by 3) patch of images
Pooling: degrading the resolution (here by a factor of 2) by averaging or computing the maximum on 2 by 2 patches

Training

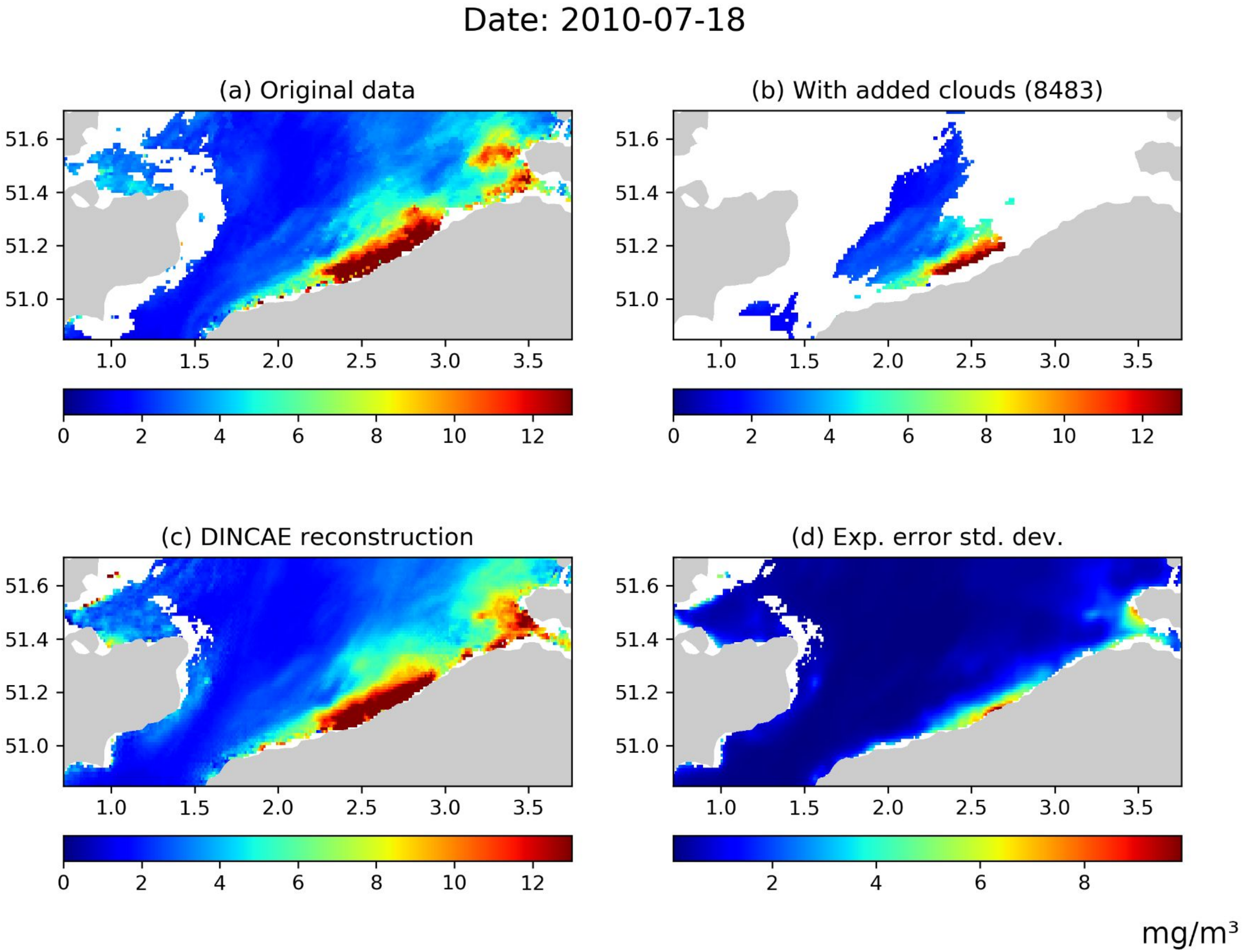
- The output of the neural network (for every single grid point i,j) is a **Gaussian probability distribution** function characterized by a mean $\hat{y}_{ij}$ and a standard deviation $\hat{\sigma}_{ij}$.

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

- 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.
- **Gradient** of the cost function is computed relative to all parameters of the neural network
- 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).

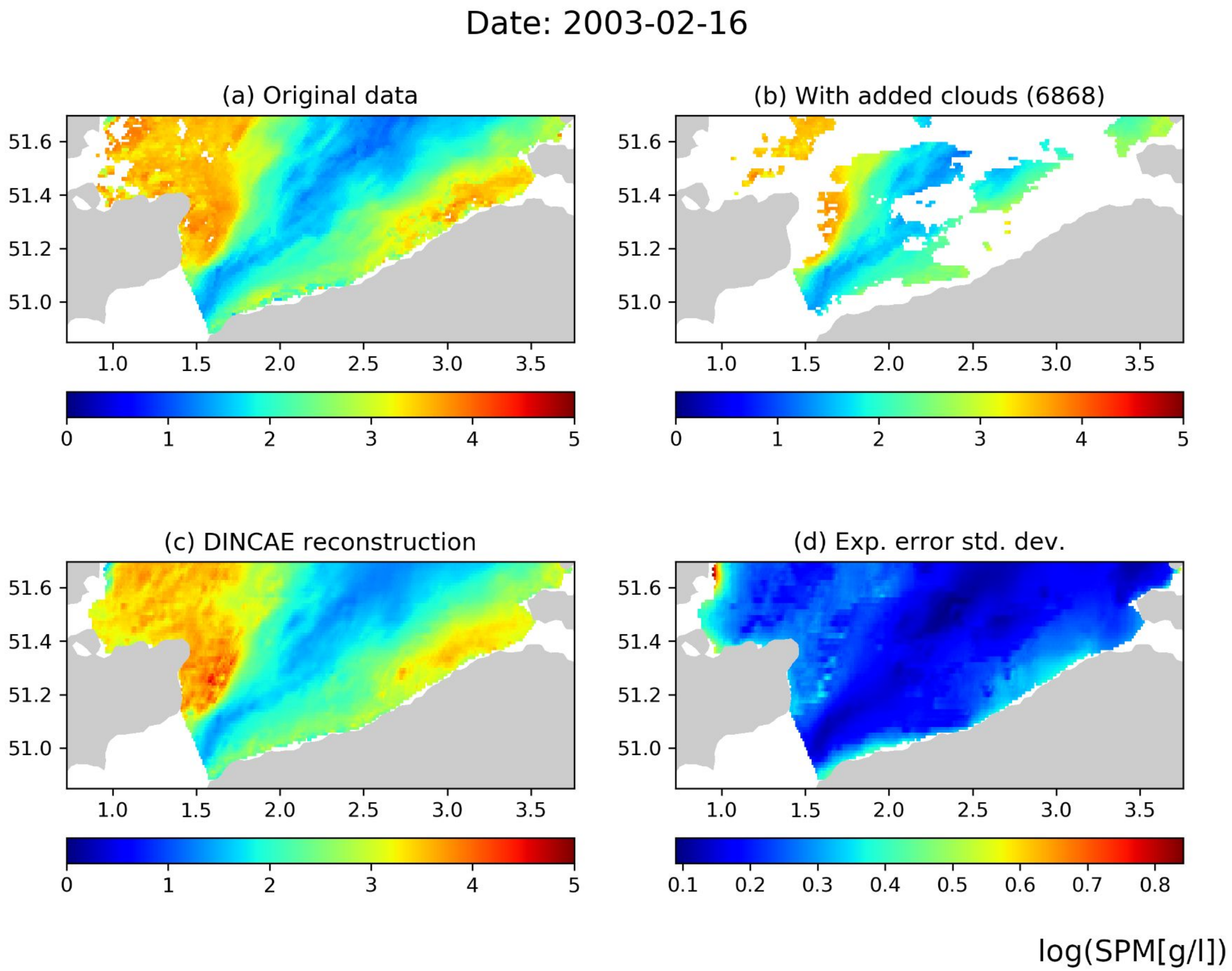
Results

Chlorophyll



Results

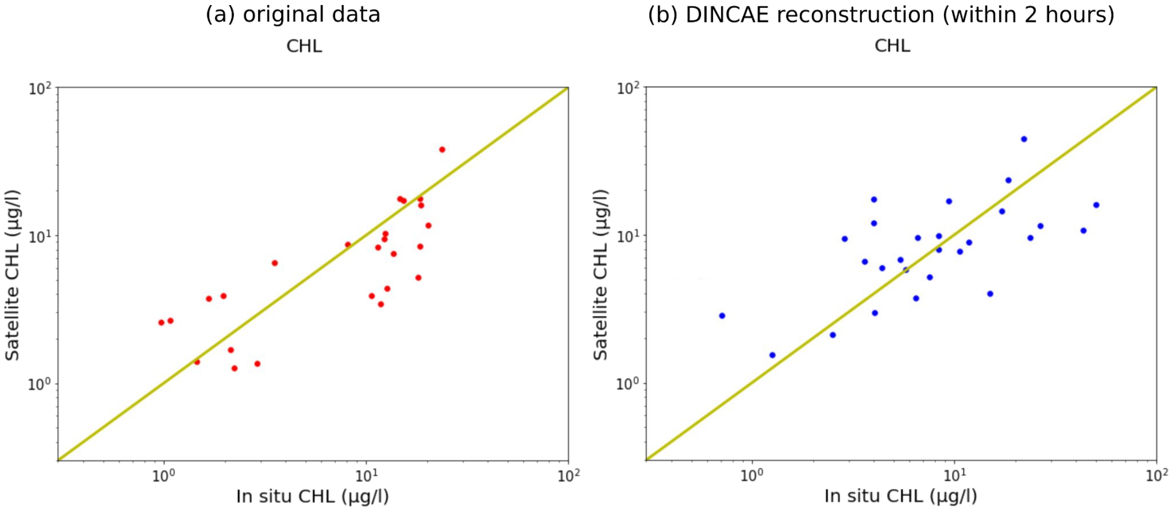
SPM



Date: 2003-02-16
(a) Original data
(b) With added clouds (6868)
(c) DINCAE reconstruction
(d) Exp. error std. dev.
51.6
51.4
51.2
51.0
1.0
1.5
2.0
2.5
3.0
3.5
0
1
2
3
4
5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
log(SPM[g/l])

In situ validation

- Chlorophyll-a fields are validated using ship-based chlorophyll (Belgian Marine Data Centre)
- Only surface observations (0-3 m depth).
- Two hours (time difference) were considered for the DINCAE validation.
- The restriction was relaxed to 24 hours for the original data to increase the sample size.
- Orig. data: RMS diff. = **0.29**
- Rec. data: RMS diff. = **0.33**



Units: log transformed of concentration in mg/m^3

Conclusions

- **Convolutional auto-encoders**: a very promising approach to reconstruct missing data in satellite images.
- The neural network DINCAE was **originally tested with sea-surface temperature**. In this work, two new applications with **chlorophyll and total suspended matter**
- **Recover spatial structures** partially or fully covered by clouds for structures that have been consistently observed in the training dataset even if the number of missing data is very high
- The chlorophyll-a reconstructions have also been **validated against in situ measurements**. The RMS difference between the reconstruction and in situ observations (after log transformation if concentration is expressed in mg/m^3) is 0.33 when considering matchups within 2 hours of the satellite pass.
- Contact: A.Barth@uliege.be