RECONSTRUCTION OF MISSING DATA IN SATELLITE IMAGES OF THE SOUTHERN NORTH SEA USING A CONVOLUTIONAL NEURAL NETWORK (DINCAE)

Barth, A.[1], Alvera-Azcárate, A.[1], Troupin, C.[1], Beckers, J.-M.[1], Van der Zande, D.[2]

[1] GHER, University of Liège, Liège, Belgium
[2] RBINS, Direction Natural Environment, Brussels, Belgium

ABSTRACT

A neural network with the architecture of a convolutional autoencoder is used to reconstruct missing data in satellite images of the Southern North Sea. The technique is applied to a multi-satellite data product of chlorophyll-a and total suspended particulate matter (SPM) concentration (representing 20 years of data). The presence of clouds significantly reduces the extent of the ocean that can be measured by satellite sensors using the visible or infrared spectrum. The accuracy of the reconstruction is assessed using cross-validation (i.e. increasing the actual extent of the cloud coverage). The results of the neural network compare favourably the data withheld for cross-validation.

Index Terms— Satellite data reconstruction, convolutional auto-encoder, neural network, machine learning, DINCAE

1. INTRODUCTION

On average, clouds cover 75% of the ocean surface [1] and are opaque for oceanographic remote sensing instruments working in the visible and infrared spectrum. This leads to a significant loss of data. Several techniques have been proposed in the scientific literature to interpolate and extrapolate satellite observations to reconstruct data under clouds. A common technique is the optimal interpolation [2], which updates a first guess (based on e.g. a long-term seasonal average) with satellite observations. The first guess and the satellite observations are assumed to be unbiased and their error covariance should be known. In practice, the error covariance is often parameterized as gaussian radial functions where the essential parameter is the correlation length. This procedure has several important consequences: spatial scales smaller than the correlation length are filtered out and regions of missing data with a spatial extent larger than the chosen correlation length are essentially filled with the background value. In practice, it is also difficult to determine a reasonable correlation length as multiple spatial scales might be present in the dataset. The use of Empirical Orthogonal Functions (also called principal component analysis) allows for a different approach where the variability that can be represented is not specified a priori but learned from a dataset. For oceanic satellite observations it has been shown repeatedly that a relatively small number of EOFs (10-50) can represent a significant fraction of the total variability for regional applications of ocean tracers (like temperature, salinity, chlorophyll-a concentration and turbidity). This suggests that the effective degrees of freedom are relatively few (compared for example to the number of grid points). This property is the motivation for reconstruction methods like DINEOF (Data Interpolating Empirical Orthogonal Functions, [3]).

Empirical Orthogonal Functions can only represent linear relationships between the values at different grid points. This limitation can be alleviated to some degree by using a non-linear transformation of the satellite observations (Gaussian Anamorphosis). But still the expressiveness for non-linear relationships is limited for such approaches. Another issue of DINEOF and optimal interpolation is the difficulty to provide a reliable estimate of the error of the reconstructions. For products derived from optimal interpolation, this error is generally computed but is only indicative of the expected error standard deviation. The flexibility of the neural network to learn from observations and to represent non-linear relationships offers some interesting perspectives to address these issues.

2. DATA

The domain considered is the Southern part of the North Sea, more specifically, ranging from 0.71°E to 3.75°E and from 50.85°N to 51.70°N The dynamics of the domain is strongly influenced by tides and riverine inputs. For this study we consider the concentration of suspended matter (1998 - 2017, 4690 images) and chlorophyll-a (1998 to 2017, 4998 images). The image sequence has a daily time resolution but images without any data have been removed from the time series.

The coherent multi-algorithm satellite-based chlorophyll-a (CHL) product used in this study was generated following the JMP-EUNOSAT approach [4, 5] using publicly accessible data available from the Copernicus Marine Environment Monitoring Service (CMEMS), European Space Agency

(i.e. ODESA) and other data providers (i.e. IFREMER). The product consists of a combination of quality controlled chlorophyll-a algorithms (e.g. blue-green ratio, red-nir ratio and Neural-Net approaches) based on the best suited algorithm/water type combination (e.g. turbid waters, clear waters, CDOM-dominated waters). The Suspended Particulate Matter (SPM) products were generated by applying the approach of Nechad et al. [6] to the OC-CCI Remote Sensing Reflectance (Rrs) product obtained from CMEMS (OCEANCOLOUR_ATL_OPTICS_L3_REP_OBSERVATIONS_009_066, CMEMS data portal). This data is a subset of the dataset in the Belgian region that is reconstructed by the DINEOF technique [7] which is also used in [8].

To estimate the accuracy of the reconstruction some additional pixels of the satellite images are marked as missing (i.e. marked as clouded). Starting with 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. The procedure is repeated using the images with the second, third, ... lowest number of clouds until the total number of grid points masked this way reaches 10% of the number of clear pixels of the total dataset. The procedure ensures that the cloud withheld for validation has a realistic spatial extent and shape.

3. METHOD

The method used in this work is an improvement of the data-interpolating convolutional auto-encoder (DINCAE, [9]). The challenge is to train a neural network with a significant number of missing data. The input of the neural network is:

- the remote sensed quantity divided by the expected error variance (previous image, current image, and next image)
- the inverse of the expected error variance (previous image, current image, and next image)
- longitude
- latitude
- $\cos(2\pi\, t/T)$ and $\sin(2\pi\, t/T)$ where t is the day of the year and $T = 365.25$ days.

In total there are thus 10 input variables (also called features) provided to the neural network. For missing data (under clouds or on land), the expected error variance is infinitely large and the corresponding inverse of the variance is set to zero. Where data is available, the expected error variance is a constant value (depending on the considered variable).

The neural network is composed of an encoder and a decoder part. The encoder part is composed of 3×3 convolutional layers followed by 2×2 average pooling. There are 5 pairs of convolutional layers (with filter size of 16, 30, 58, 110 and 209) and average pooling layers. The decoder part is essentially the reverse of the encoder part with the exception that the pooling layer is replaced by an upsampling layer (here interpolation to the nearest neighbor) and the final convolutional layer provides 2 output features which are related to the reconstruction value and the uncertainty.

The network also contains sum-skip connections where the intermediate arrays in the encoder part are added to the corresponding arrays in the decoder part of the network. These skip connections prevent excessive smoothing of the results.

The neural network is trained by mini batches of 50 instances chosen at random from the complete time series over 1000 epocs. During training, some additional data points (pixels) are withheld and the loss function is computed over the withheld data to train the network explicitly to reconstruct missing data. The loss function uses the negative log likelihood of the training data [9]. The neural network provides an estimate over the full grid including land points, but only the grid points corresponding to the sea are used in the loss function and are presented in the following.

For the chlorophyll-a reconstruction we used this auto-encoder in two stages. A first encoder/decoder pair followed by a second encoder/decoder pair. The loss function is a weighted combination of the output of the final decoder and the output of the intermediate decoder with a weight of 0.7 and 0.3 respectively. This architecture is motivated by the GoogLeNet model [10] where it has been shown that the approach allows training more efficiently a deeper network. For SPM this refinement step did not result in an improvement of the result. DINCAE has been implemented in the Julia programming language [11] using the neural network package Knet [12] striking both a remarkable balance between ease-of-use, flexibility and performance.

4. RESULTS

Two sample results are shown in Figure 1 for Chlorophyll-a concentration and Figure 3 for SPM concentration. This image is chosen because the original data has a particularly low cloud coverage for these dates. The network is only trained where additional data are masked as missing (as shown in panel b for these figures). The reconstruction agrees reasonably well with the original data for both parameters. The Chlorophyll-a concentration is usually high near the Rhine-Meuse-Scheldt delta. The high concentration area (above 10 $\mathrm{mg\,m^{-3}}$) is well represented in the reconstruction despite it not being directly measured for the particular day considered here. The uncertainty predicted by the neural network is also relatively high near the coast and in the delta. This is to be expected because the variability there is also very high.

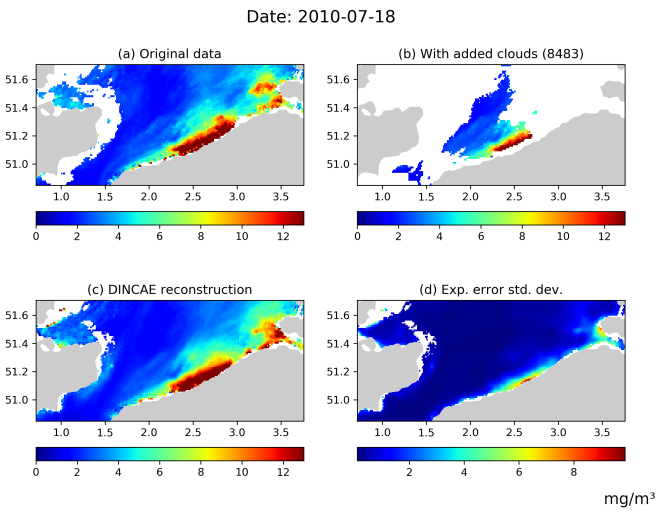

Fig. 1. Chlorophyll-a reconstruction for the date 2010-07-18. Panel (a) shows the original data. Panel (b) By imposing the cloud mask from a different time instance chosen at random, 6868 additional pixels are masked in this image. Panel (c) is the DINCAE reconstruction using the data from panel (b) among others. The expected error standard deviation is in panel (d). All units are in $\mathrm{mg\,m^{-3}}$.

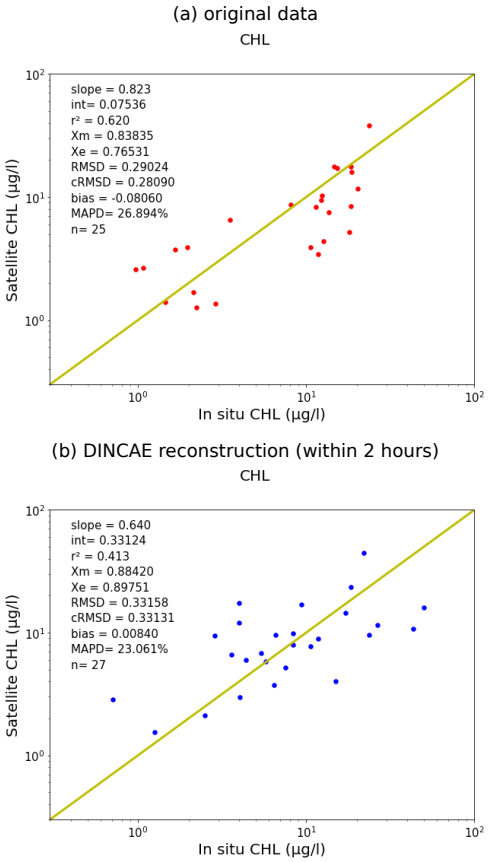

Fig. 2. Validation of chlorophyll-a original satellite data (panel (a)) and the reconstructed DINCAE data (panel (b)) with in situ observations. Slope and int are the slope and intercept of the regression line, r^2 is the determination coefficient, X_m (resp. X_e) is the mean of the log (base 10) transformed chlorophyll-a in situ data (resp. Satellite data), RMSD (resp. CRMSD) is the (resp. centred) root mean square difference, MAPD is the median absolute percentage difference and n is the sample size.

Chlorophyll-a fields are validated using ship-based chlorophyll data obtained from the Belgian Marine Data Centre (http://www.bmdc.be/NODC/index.xhtml) for the considered period (Figure 2). Only surface observations were retained (0-3 m depth). Data of deeper waters were discarded since the matchup analysis focuses on concentrations in the upper mixed layer that can be observed by satellite. Given the strong tidal dynamics, only data within a time difference of two hours were considered for the DINCAE validation. The restriction was relaxed to 24 hours for the original data to increase the sample size. Overall the RMSD between in situ observations and satellite data is 0.29 (log transformed of concentration in mg/m^3). The reconstructed satellite data has an RMS difference of 0.33 when considering only data within 2 hours of the satellite pass. If the validation is expanded to all match-ups within 24 hours the RMS difference is 0.39.

The SPM reconstructions (Figure 3) show that small scales in the center of the English channel present in the input image of the neural network (panel b) are retained in the reconstruction. It is surprising that the network also correctly represents the region with higher SPM concentration west of the Rhine-Meuse-Scheldt delta along the coastline, albeit with a somewhat under-estimated concentration.

Overall the RMS error of the Chlorophyll-a concentration is 4.0 mg/m^3 and for the SPM is 0.24 (log of concentration expressed in g/l).

5. CONCLUSIONS

Convolutional auto-encoders constitute 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 of this approach to the North Sea have been presented. The neural network is able to 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 (only 17% of the sea points in the chlorophyll-a training dataset and only 26% of the sea point of in the SPM training dataset have measurements). The chlorophyll-a reconstructions have also been validated against in situ measurements. The RMS difference between

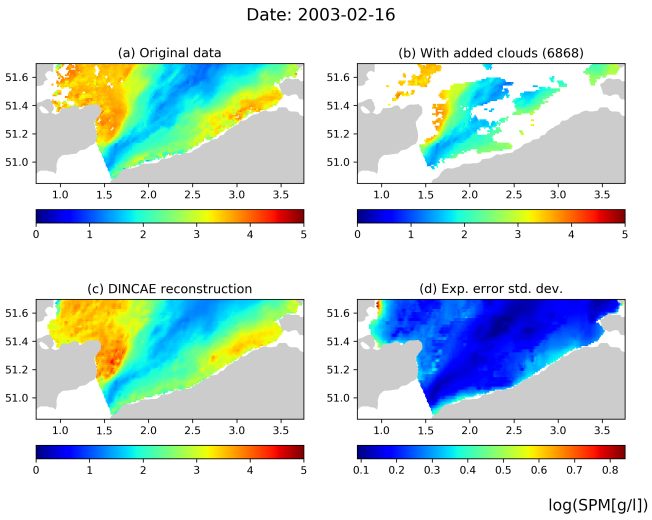

Fig. 3. SPM reconstruction (logarithm of the concentration expressed in g/l) for the date 2003-02-16. Panels are the same as in Figure 1.

the reconstruction and in situ observations (after log transformation if concentration is expressed in mg/m^3) is 0.33 (resp. 0.39) when considering matchups within 2 hours (resp. 24 hours) of the satellite pass.

6. ACKNOWLEDGEMENTS

The F.R.S.-FNRS (Fonds de la Recherche Scientifique de Belgique) is acknowledged for funding the position of Alexander Barth. This research was partly performed with funding from the Belgian Science Policy Office (BELSPO) STEREO III program in the framework of the MULTI-SYNC project (contract SR/00/359). Computational resources have been provided in part by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the F.R.S.-FNRS under Grant No. 2.5020.11 and by the Walloon Region.

7. REFERENCES

[1] D. Wylie, D. L. Jackson, W. P. Menzel, and J. J. Bates, "Trends in Global Cloud Cover in Two Decades of HIRS Observations," *Journal of Climate*, vol. 18, no. 15, pp. 3021–3031, 2005.

[2] B. Buongiorno Nardelli, S. Colella, R. Santoleri, M. Guarracino, and A. Kholod, "A re-analysis of black sea surface temperature," *Journal of Marine Systems*, vol. 79, no. 1–2, pp. 50–64, 2010.

[3] A. Alvera-Azcárate, Q. Vanhellemont, K. Ruddick, A. Barth, and J.-M. Beckers, "Analysis of high frequency geostationary ocean colour data using DINEOF," *Estuarine, Coastal and Shelf Science*, vol. 159, pp. 28 – 36, 2015.

[4] D. Van der Zande, H. Lavigne, A. Blauw, T. Prins, X. Desmit, M. Eleveld, F. Gohin, S. Pardo, G. Tilstone, and J. F. Cardoso Dos Santos, "Enhance coherence in eutrophication assessments based on chlorophyll, using satellite data as part of the EU project 'Joint monitoring programme of the eutrophication of the North Sea with satellite data'," Tech. Rep., Royal Belgian Institute of Natural Sciences, 2019, Activity 2 Report, 106 pp.

[5] H. Lavigne, D. Van der Zande, K. Ruddick, J. F. Cardoso Dos Santos, F. Gohin, V. Brotas, and S. Kratzer, "Quality-control tests for OC4, OC5 and NIR-red satellite chlorophyll-a algorithms applied to coastal waters," *Remote Sensing of Environment*, 2021, In press.

[6] B. Nechad, K. Ruddick, and Y. Park, "Calibration and validation of a generic multisensor algorithm for mapping of total suspended matter in turbid waters," *Remote Sensing of Environment*, vol. 114, no. 4, pp. 854 – 866, 2010.

[7] A. Alvera-Azcárate, D. Van der Zande, A. Barth, C. Troupin, S. Martin, and J.-M. Beckers, "Analysis of 23 years of daily cloud-free chlorophyll and suspended particulate matter in the Greater North Sea," *Frontiers in Marine Science*, 2021, submitted.

[8] A. Alvera-Azcárate, A. Barth, C. Troupin, J.-M. Beckers, and D. V. der Zande, "Creation of high resolution suspended particulate matter data in the North Sea from Sentinel-2 and Sentinel-3 data," in *IEEE International Symposium on Geoscience and Remote Sensing IGARSS*. 2021, p. 4, IEEE, submitted.

[9] A. Barth, A. Alvera-Azcárate, M. Licer, and J.-M. Beckers, "DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations," *Geoscientific Model Development*, vol. 13, no. 3, pp. 1609–1622, 2020.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM review*, vol. 59, no. 1, pp. 65–98, 2017.

[12] D. Yuret, "Knet: beginning deep learning with 100 lines of julia," in *Machine Learning Systems Workshop at NIPS*, 2016.