Baltic Sea cyanobacteria blooms - Machine Learning-based short-term prediction (Algae Storm)

Tobias Edman, RISE Research Institutes of Sweden
tobias.edman@ri.se

Smita Chakraborty, Bengt Karlson, Lars Arneborg, Jenny Carlstedt, Jesper Dejby, Leon Sütfeld, Martin Willbo

DestinE userXchange
16th October 2024

Algae Storm Description and Objective

Overall objectives

- Improve short term forecasts of cyanobacteria blooms in the Baltic Sea using an ML/DL-based approach
- Evaluate the availability of data in DestinE Data lake for this purpose
- Check if the DestinE infrastructure is fit for this purpose
- Improve the Baltic Algae Watch System

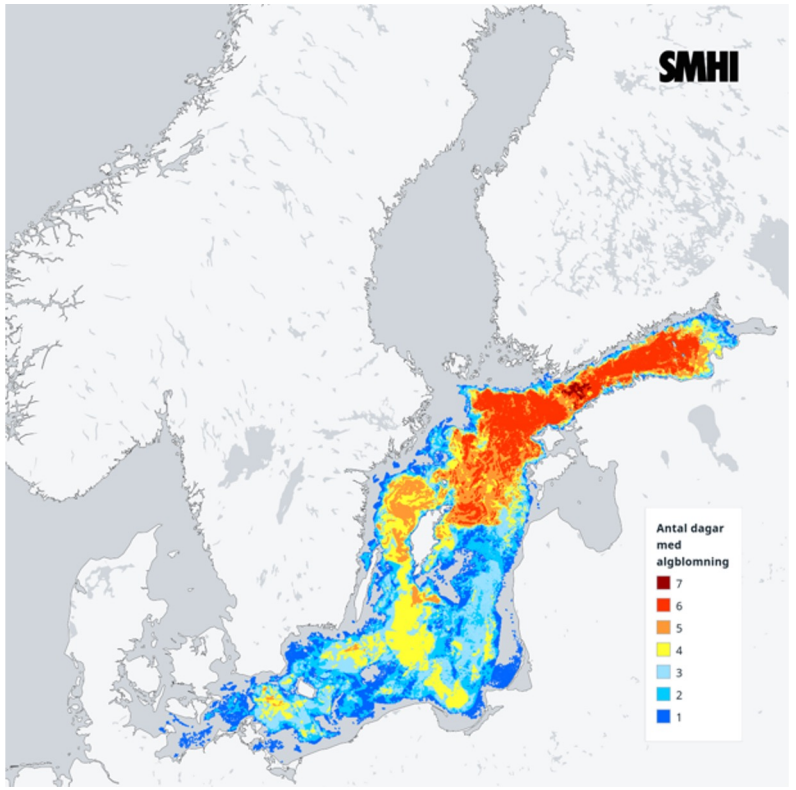
Photo by Swedish Coast Guard, Air Patrol

Algae Storm Description and Objective

Baltic Sea cyanobacteria blooms

- Cyanobacteria blooms in the brackish water Baltic Sea is a recurrent phenomenon
- The cyanobacteria are nitrogen fixers - they contribute to the nitrogen pool
- Some of the cyanobacteria produce cyanotoxins that may accumulate in food chains
  - Humans may be affected indirectly or directly
  - The Baltic Algae Watch system has been in operation since 2002

30 June, one week summary



Baltic Algae Watch System*

Data presented daily in summer
https://www.smhi.se/data/oceanografi/algsituationen

*Öberg and Karlson, 2014, Öberg et al. 2022

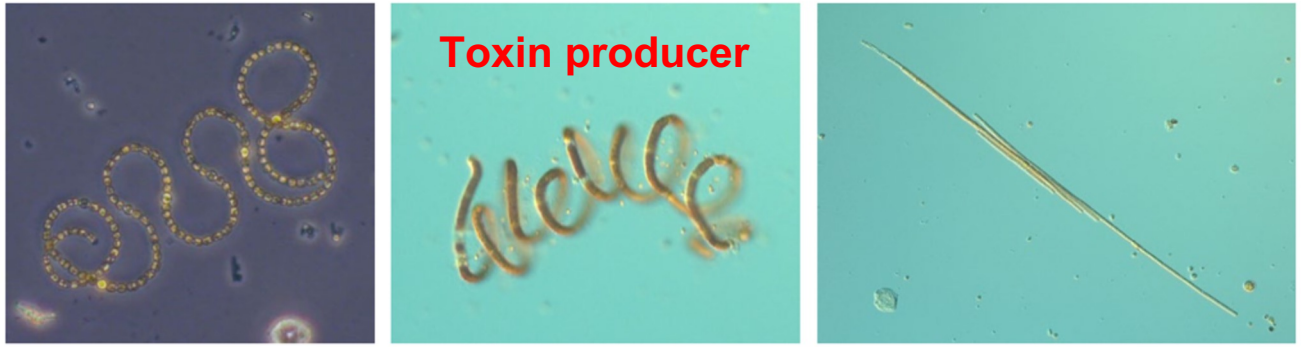
Funded by the European Union

Algae Storm Description and Objective

More on cyanobacteria

- The cyanobacteria accumulate near the sea surface during calm conditions in summer
- Cyanobacteria sometimes end up on beaches, a nuisance for tourism
- Desalination plants may be affected as well

Karlson et al. 2021



Dolichospermum sp., *Nodularia spumigena* and *Aphanizomenon flosaquae*

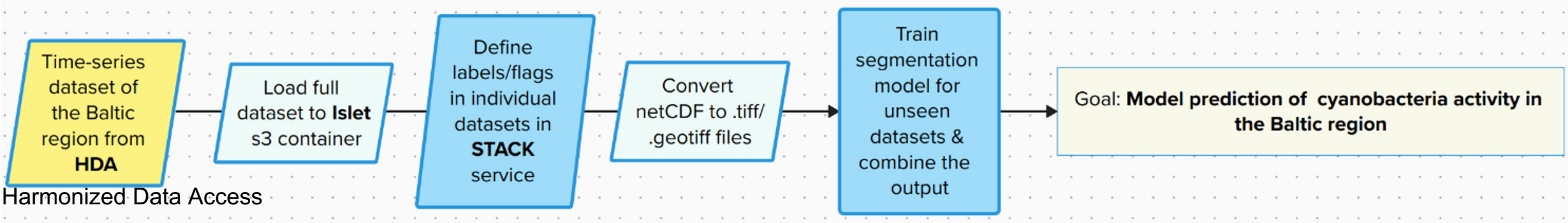
Accumulations of cyanobacteria in the Baltic Proper at Norr Tyrislöt in St. Annas archipelago, photo Peter Weyderd, source: Information Centre for the Baltic Proper

Funded by the European Union

Deep learning model for Algae Storm

- Time series dataset, with marked areas of interests (polygons) and labels/flags (from *Water Quality and Science Flags (WQSF)* files)

- ***Task: to predict the abundance of cyanobacteria (label) for future time-series for the combined dataset***
  - Disclaimer: A version of this already exists in *snap*



Harmonized Data Access

Objectives

- Finding and downloading data from the following physical properties:
  - Sea surface colour (bands with high reflectance from surface cyanobacteria)*
  - Sea wind speed
  - Sea surface temperature
  - Irradiation (Photosynthetically Active Radiation)

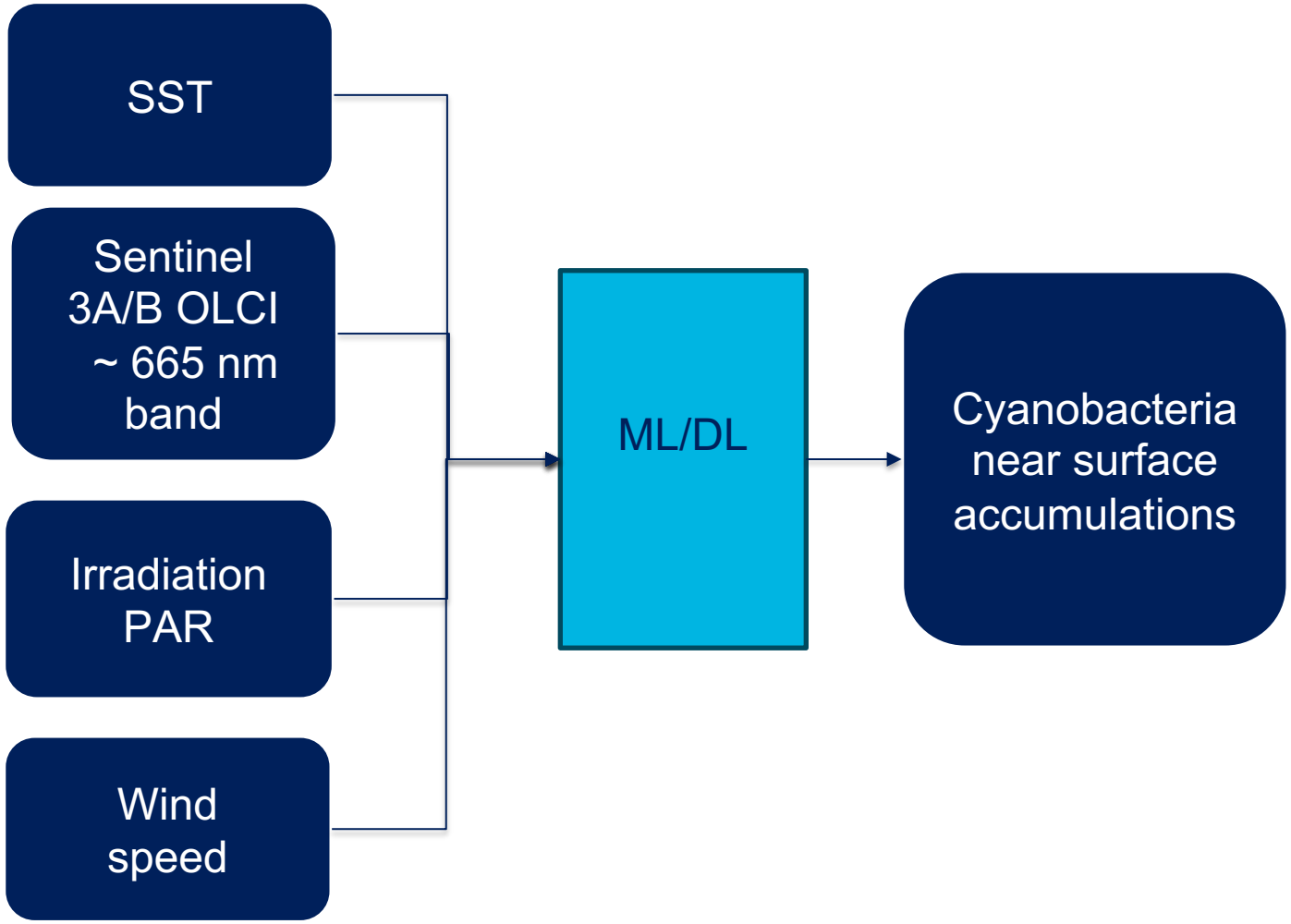
- Process these datasets to:
  - Cover only the Baltic sea region as defined by a bounding box
  - Have the same projection
  - Have the same resolution (resampling)
  - Merged into a single GeoTiff file with multiple bands that can then be used for

*Kahru 1997, Kahru et al. 2007, 2014, Karlson et al. 2022

DestinE Data Lake (DEDL) datasets that we considered

Notes from meeting 4 June 2024

- Sentinel-3
  - OLCI Level 2 Ocean Colour Full Resolution
  - SLSTR Level 2 Sea Surface Temperature
- ODYSSEA Global Ocean
  - Sea Surface Temperature L3 Observations
- FAPAR
  - Fraction of Absorbed Photosynthetically Active Radiation
- Gridded products
  - Global Ocean Daily Gridded Sea Surface Winds from Scatterometer



DEDL Services used in Algae Storm Implementation

- [x] Harmonised Data Access service
- [x] ISLET service
- [x] S3 storage
- [x] OpenStack
- [ ] Kubernetes
- [x] STACK service
- [x] JupyterLab
- [ ] DASK
- [ ] HOOK service
- [ ] Existing processors (data-harvest, sen2cor, etc.)
- [ ] User workflows

Deep Learning

1. Oa08 band indicates surface accumulations of cyanobacteria

2. Oa06 band indicates sub-surface abundance of cyanobacteria
   a. Threshold value: Oa08 = 0.0012 x π(band for 665 nm)
   b. Oa06 = 0.00435 x π x .9 (band for 560 nm)

3. Considering the data in June 2019 for this model:

   a. We train on values more than zero in a given tile, also can subtract the respective threshold values

   b. Predict a similar set of tiles and compare that with a ‘new’ dataset

Ref: Bengt Karlson, Lars Arneborg, et al. A suggested climate service for cyanobacteria blooms in the Baltic Sea – Comparing three monitoring methods, Harmful Algae, Volume 118, 2022, 102291,https://doi.org/10.1016/j.hal.2022.102291.

Funded by the European Union

Technical development of the Deep Learning model

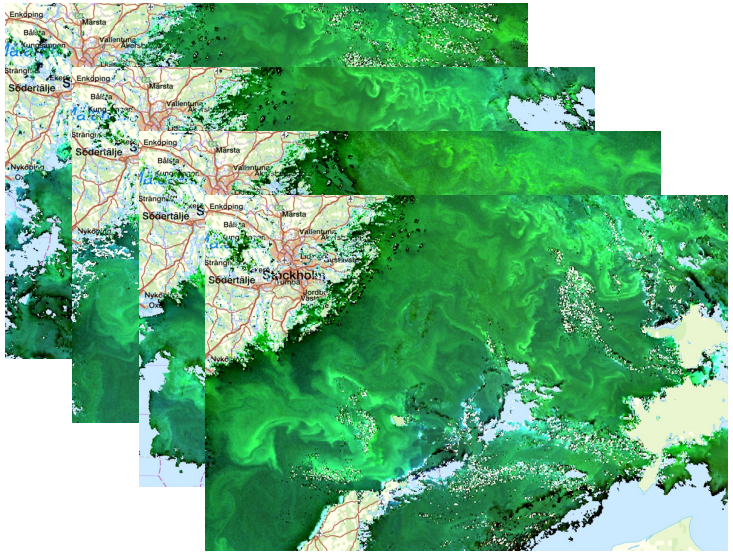
1. Input tensor to feed into a convolutional 2D layer to learn the features with a batch size=n (image height, image width = 1000, 1000)

2. Compare the output tensor to a ‘unseen’ or new tensor, particularly from a different day
   a. Note: can be a new tile on the same day as well, the ML model does not distinguish between spatial or temporal data prediction

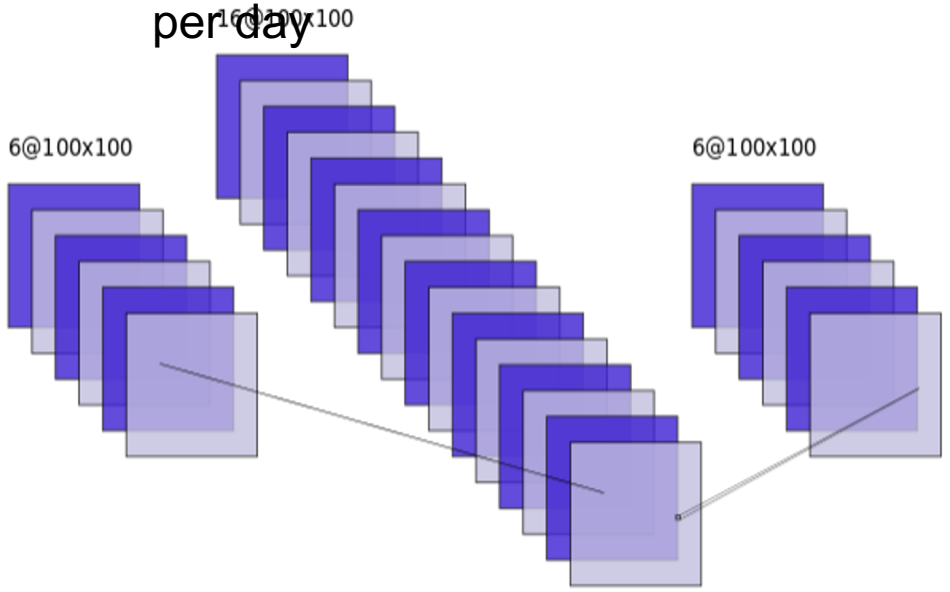
3. Output layer: Conv2D to decode the processed information reshape to specified output tensor shape

Deep learning model workflow

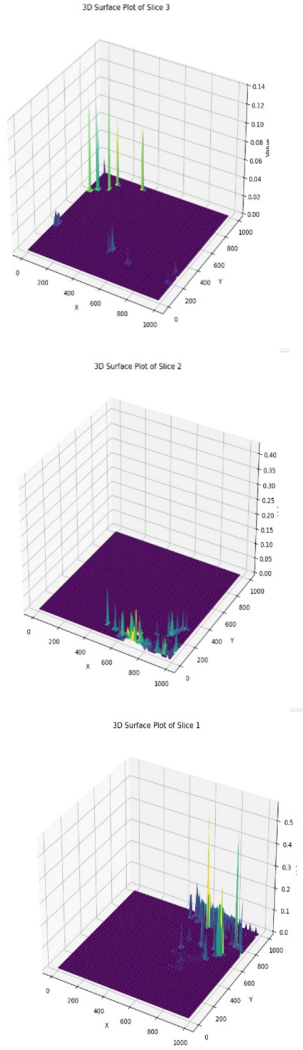
Collect data:
Sentinel 3 spectral information



Create tensors:
Based on datasets per day



Calculate output:
Based on tensors and the trained conv2d modell



Reshape

Conv2D

Train modell:
Conv2d modell training to learn to predict the third day

Funded by the European Union

Challenges from the Deep Learning perspective

- Missing continuous and same resolution data
- Non-standardized datasets for multiple products for ML processing, e.g. resolution, size of file
- Feedback: ML libraries needed to be installed separately in DEDL, not pre-installed
- Missing ‘analysis-ready’ data with labels/annotations:
  - Preprocessing and processing from scratch was required to build the dataset
  - Dataset created was not large enough to train deeper models, such as vision transformers or LSTM or U-Net. Further processing required to create whole dataset per day (e.g with mosaicking of data per day)
  - Use of GIS tools or others were not part of the DEDL library or references - proves to be a steep learning curve, needs further GIS/satellite image processing expertise
  - Sentinel 3a/3b dataset is ‘quite new’ - both useful and challenging to adapt to the data structure during processing
  - Need for marked ‘categories’ in data without the need for preprocessing - so far present in other netCDF products, but not training-ready for ML

Way forward

- Based on the learnings form the project we want to continue working with the forecasting and include more elements of the DestinE data lake and utilise more data from the DestinE engine outputs making it possible to work better with scenarios for the cyanobacteria blooms in a changing climate.

Thanks!

Tobias Edman, RISE Research Institutes of Sweden
tobias.edman@ri.se

Bengt Karlson
Oceanographic Research
Swedish Meteorological and Hyrdological Institute, Sweden
bengt.karlson@smhi.se

RI. SE
SMHI