Algal Bloom Forecasting Using Sparse Data

Einar de Gruyl E.A.deGruyl@student.tudelft.nl

Responsible Professor: dr. Jan van Gemert
Supervisors: Attila Lengyel, Robert-Jan Bruintjes

TU Delft

Research Question 1

How to forecast algal blooms using spatially and temporally sparse satellite data?

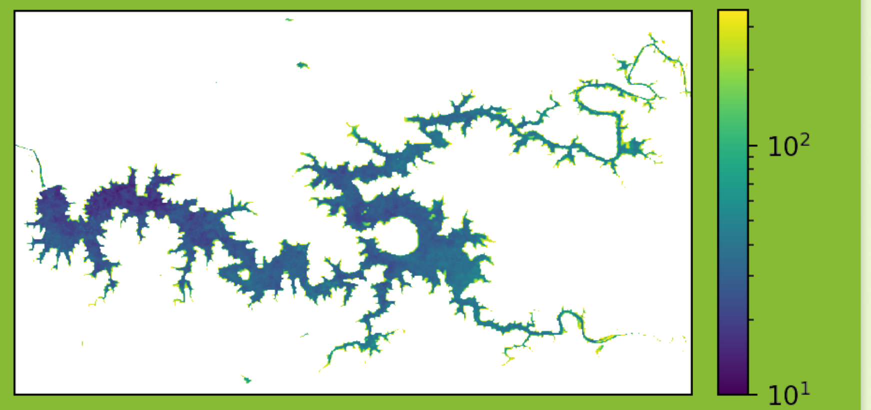
Introduction 2

- Algal blooms are an increase in the population of algae in a body of water, which can have negative impacts on water quality and ecosystems.
- Satellite data can be used to monitor and forecast algal blooms, but the data is often spatially and temporally sparse, which means it only covers a small area and is only collected at certain times.
- The use of satellite data to forecast algal blooms can provide valuable information to water managers and policymakers, helping them to take action to prevent or mitigate the negative impacts of algal blooms.

Data Analysis 3

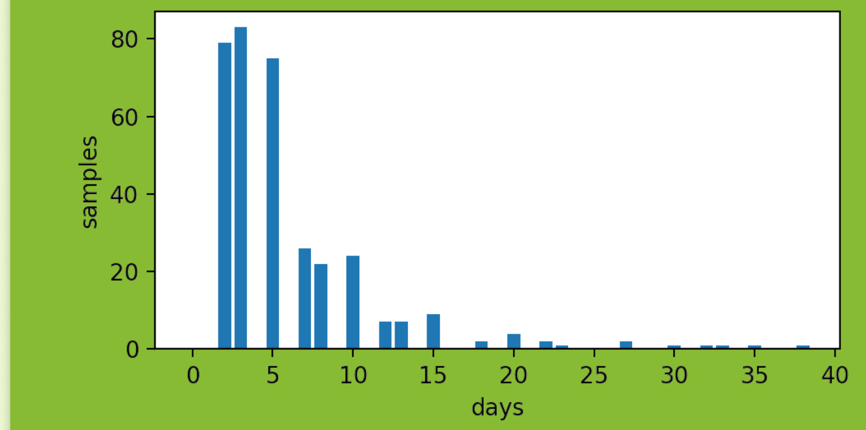
- As can be seen in figure 1, the majority of locations have only 10-20 missing samples, while some locations have over a hundred missing samples.
- As can be seen in figure 2, the majority of samples have a gap of 2 to 5 days. However a significant portion of the samples have a gap of over 10 days and in extreme cases even over 20 days. This can make the interpolation of these samples quite difficult.

Fig 1 Spatial Sparsity



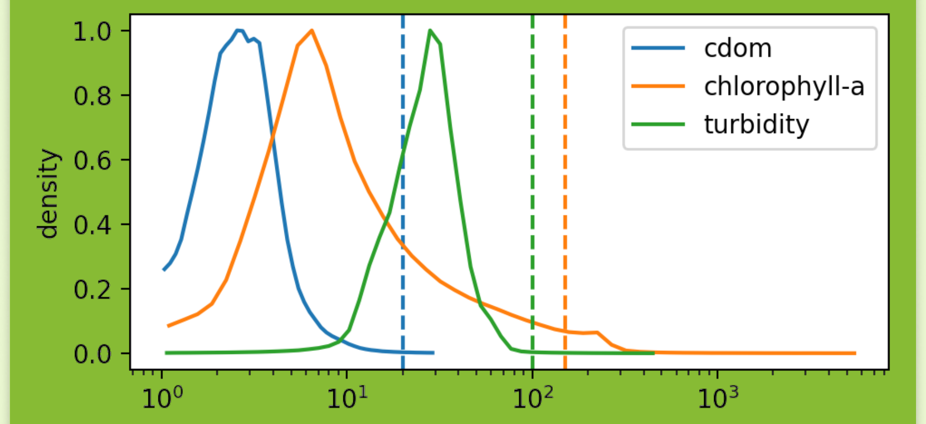
The number of missing spatial samples in the dataset for the Palmar reservoir in Uruguay

Fig 2 Temporal Sparsity



The number of days between consecutive samples in the the dataset for the Palmar reservoir in Uruguay

Fig 3 Value Distribution



Value distribution of the cdom, chlorophyll-a, and turbidity bands of the Palmar reservoir in Uruguay. The vertical lines show the chosen clip values.

Method 4

U-Net [1]

- U-Net is a type of convolutional neural network that is used for segmentation.
- U-Net can also be used for regression tasks.

Loss Functions

- Mean squared error (MSE) loss function.
- DenseWeight loss function.
  - Weighted MSE loss.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Root mean squared error (RMSE) for validation.
- Missing values will not be considered in the loss functions.

Baseline

- Good comparison for other (better) interpolation methods.
- Missing values are filled by a 30-day lookback.

Nearest Neighbour Interpolation

- Missing values are filled by the closest real value.
- Possible to fill in missing spatial values.

Linear Interpolation

- Missing values are filled by estimating values by constructing a line between two existing values.
- Possible to preserve the gradient between samples.

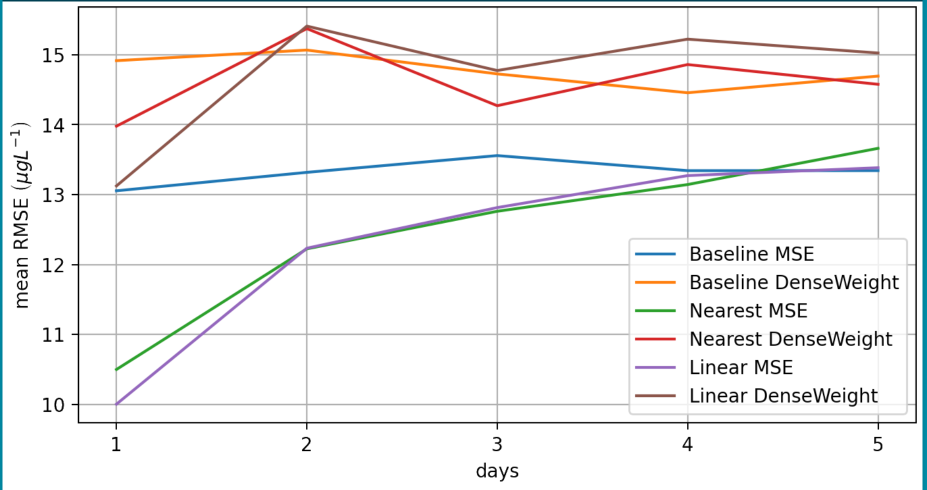
References

[1] Olaf Ronneberger and Philipp Fischer and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. CoRR, abs/1505.04597, 2015.

Results 5

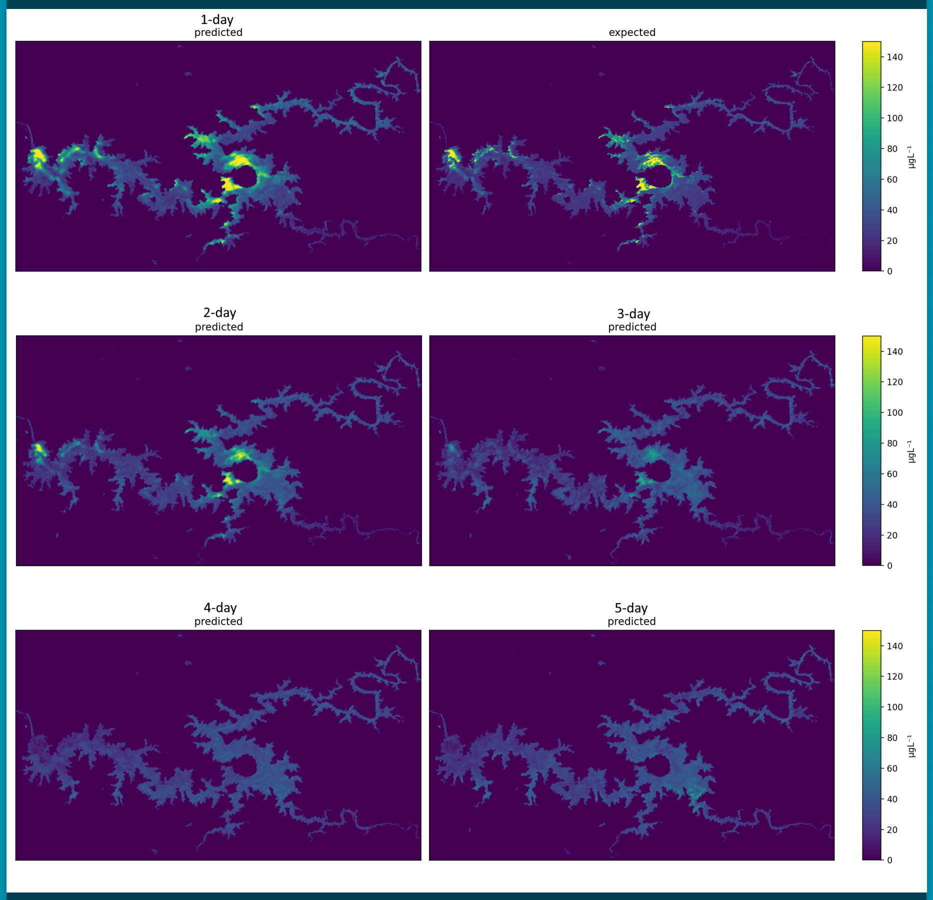
Each interpolation method was trained and tested with a prediction horizon of 1 to 5.

Fig 5 Validation Loss



The validation loss of training U-Net for 30 epochs with a window size of 5 and a prediction horizon of 1

Fig 6 Linear DenseWeight



Predicted values for the linear DenseWeight experiment with prediction horizons of 1 to 5.

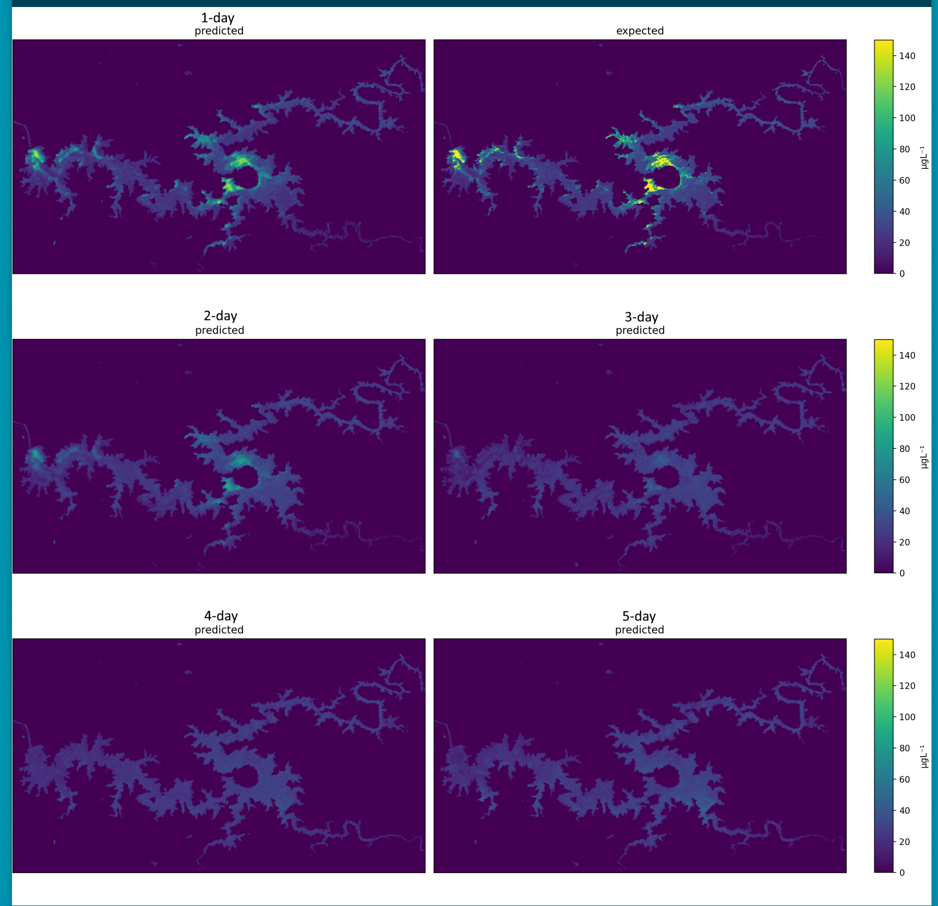
Conclusion 6

- MSE seems to perform better than DenseWeight in figure 5.
- DenseWeight over-predicts. See figure 6.
- MSE under-predicts. See figure 7.
- The baseline fails to predict anything of use.
- Both nearest neighbour and linear interpolation seem to work equally well with linear interpolation slightly outperforming nearest neighbour interpolation.

Future Work 7

- Balanced loss function between MSE and DenseWeight.
- Other neural networks such as LSTM.
- Include other data modalities.

Fig 7 Linear MSE



Predicted values for the linear MSE experiment with prediction horizons of 1 to 5.