Team: Burning Hydrogen

Team Members:

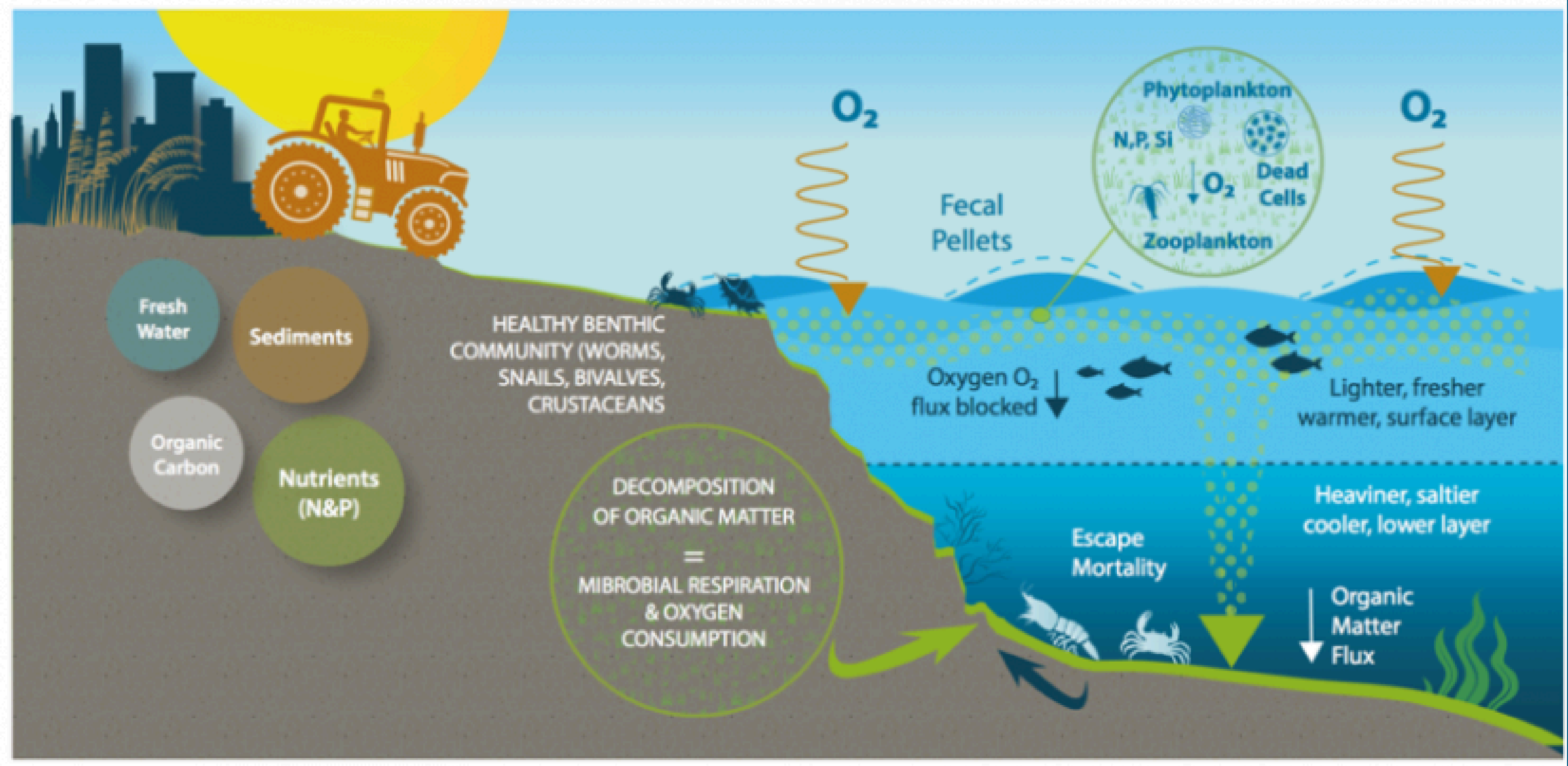
- Touseef Haider
- Sanjay Kumar
- Afsin Ozdemir
- Alex Rasmussen
- Rafatu Salis

What are we looking for?

- What are the main factors that indicate an upcoming drop in oxygen concentration in a region?
- Can we predict the deoxygenation of the oceans in the long term?

NEW
ATLANTIS

Hypothesis



Reference:
Impacts of eutrophication on ocean oxygen. Figure from Global Ocean Oxygen Network, Breitburg, D., M. Gregoire, K. Isensee (eds.) 2018. The ocean is losing its breath: Declining oxygen in the world's ocean and coastal waters. IOC-UNESCO, IOC Technical Series, No. 137 40pp. (IOC/2018/TS/137)

What are the primary signs of a decrease in oxygen levels?

Possible physical indicators:

- Salinity
- Temperature
- Phosphate
- Nitrate
- Silicate

Possible complex indicators:

- Chlorophyll-a
- Phytoplankton
- Marine Life

NEW ATLANTIS

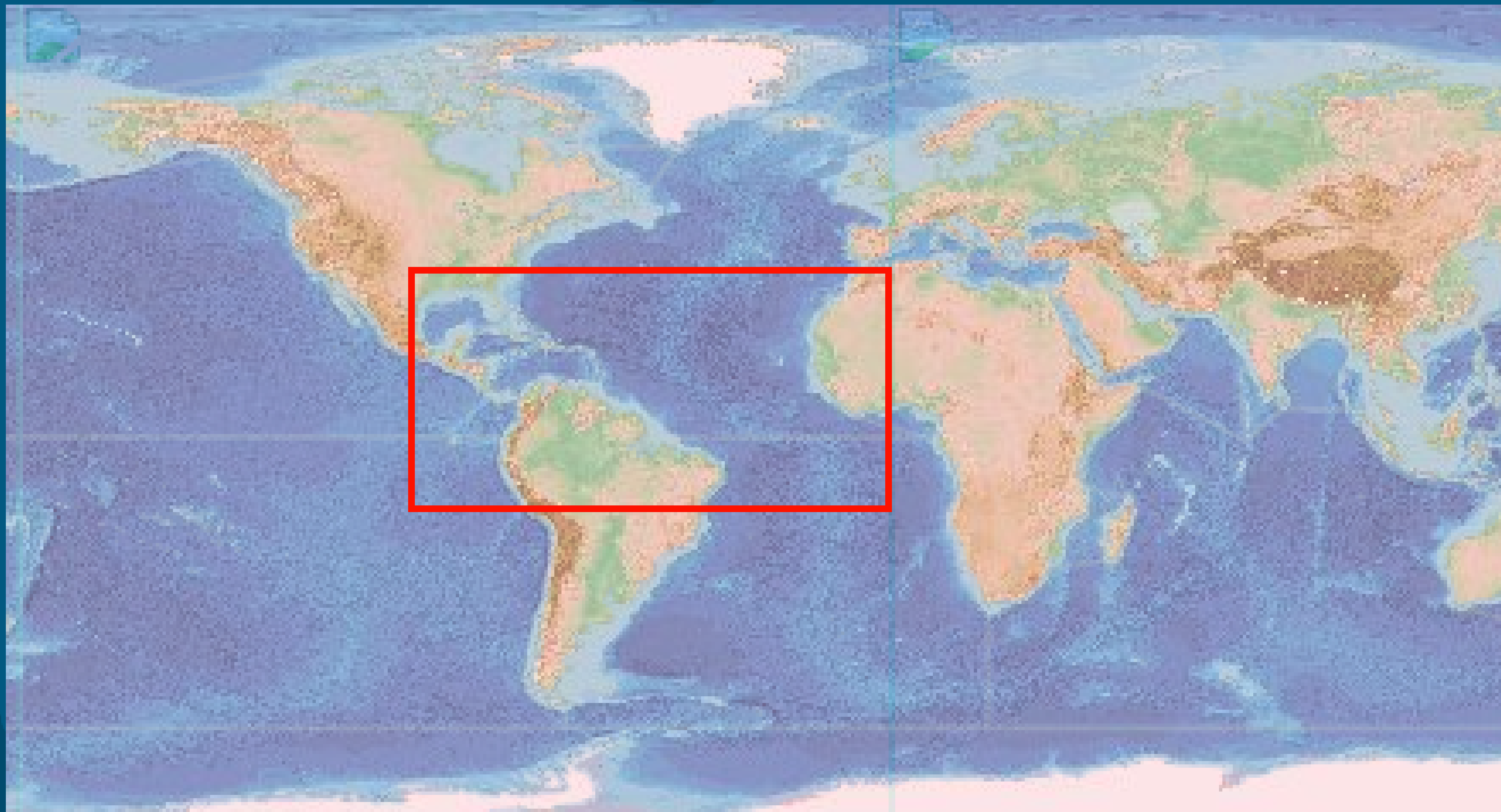
Data Sets:

Region:
(-100, 0) Longitude
(-15,35) Latitude
with 20x10 grid

Time Intervals:
- 1965-2022

For XGBoost:

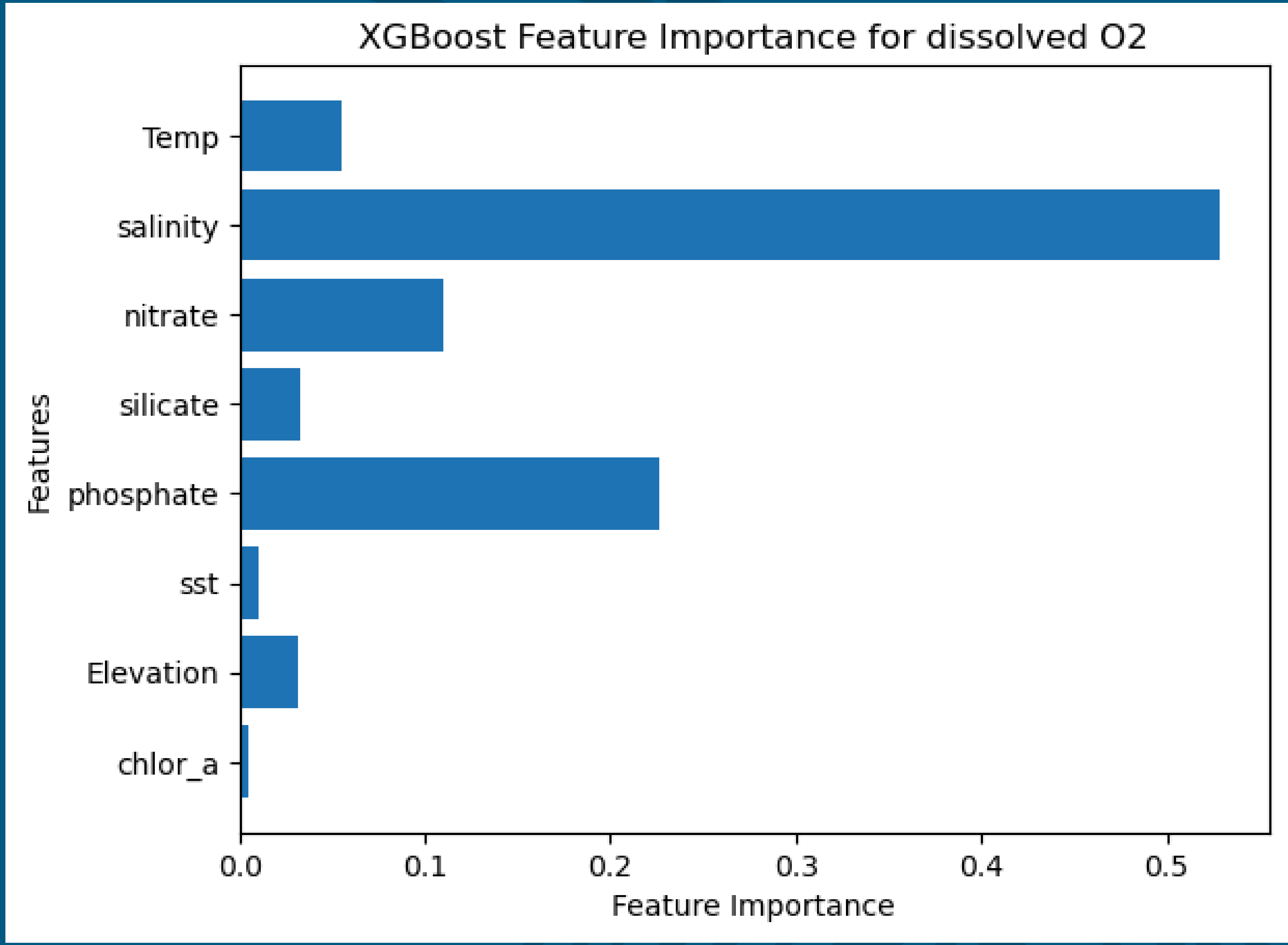
- World Ocean Atlas
- World Ocean Database
- GEBCO Gridded Bathymetry
- NASA Earth Observations, Modis chlorophyll-a data
- NOAA Physical Science Laboratory, SST
- Copernicus Marine data



Predicting O2 from other variables

XGBoost Regression:

Feature Importance



Model Parameters:

- max_depth=10,
- learning_rate=0.1,
- n_estimators=200,
- subsample=0.8,
- colsample_bytree=0.8,
- gamma=0,
- min_child_weight=1

Statistical mean of all datasets between 1965-2022:

- Phosphate
- Silicate
- Nitrate
- Salinity
- Temperature

Time series modeling with neural networks

- Goal: forecast future daily O2 concentration in our region based on past ocean variables (oxygen, temperature, salinity, depth, nitrate, phosphate, and silicate).
- Used Copernicus dataset to create **dataframe** with 10 days of lag features and 30 days of lead features

	date	lon_rounded_up	lat_rounded_up	...	salinity_lead_30	temp_lead_30	o2_lead_30
0	2021-12-10	-95	-10	...	35.680194	22.169635	224.147746
1	2021-12-10	-95	-5	...	35.260607	23.094601	221.890783
2	2021-12-10	-95	0	...	34.729936	21.073147	199.361494
3	2021-12-10	-95	5	...	33.862983	24.679508	190.822396
4	2021-12-10	-95	10	...	33.840988	23.992337	179.579970
...	...	...	...	...	...	...	...
162595	2024-03-01	0	15	...	NaN	NaN	NaN
162596	2024-03-01	0	20	...	NaN	NaN	NaN
162597	2024-03-01	0	25	...	NaN	NaN	NaN
162598	2024-03-01	0	30	...	NaN	NaN	NaN
162599	2024-03-01	0	35	...	NaN	NaN	NaN

A baseline model

- Just forecast that nothing changes! Predict oxygen to stay the same in the future as on the current day.

```
class NoModel(nn.Module):
    def __init__(self):
        super(NoModel, self).__init__()

    def forward(self, x, future_steps=30, hidden_state=None):
        return torch.stack(future_steps * [x[:, -1, :, :, :]], 1)
```

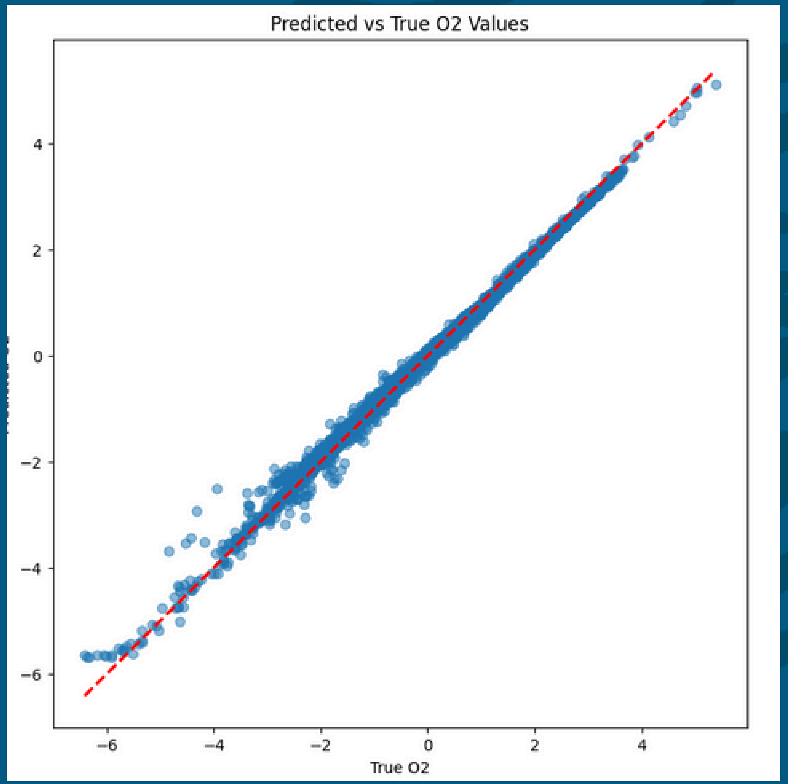
- This model does pretty well! MSE = 0.06396 for predicting (normalized) O2 concentration on 30 consecutive days in the future.

LSTM Model

- Fed 7 days and 14 days sequential data
- Dropout layer: 20% of Neuron
- Learning Rate Scheduler: Patience= 5, factor = 0.5
- Predicted 10, 30, and 60 days oxygen value in the future

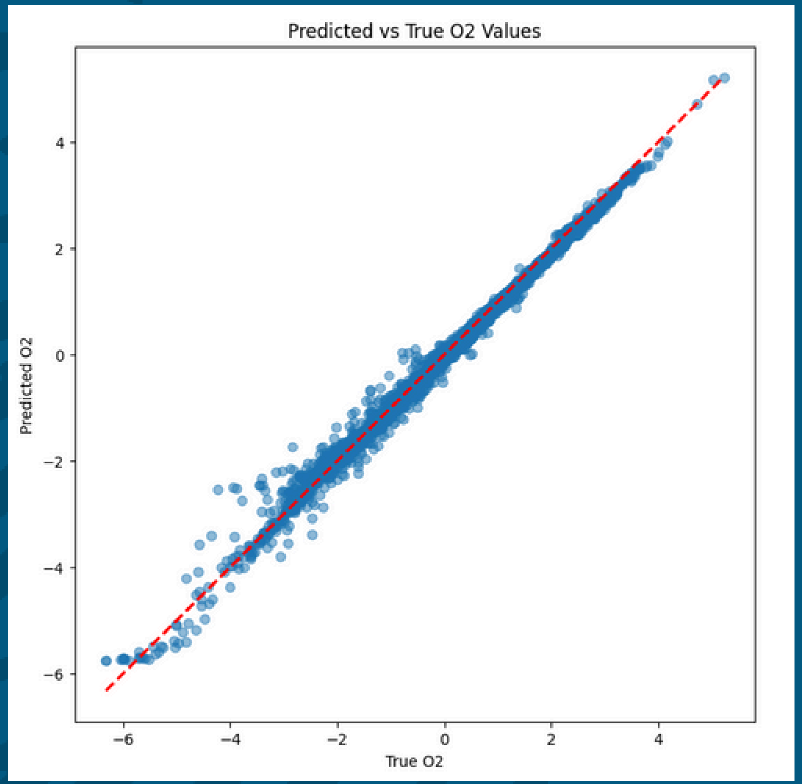
Input: 7-day sequential data
Output: 10 days prediction

mse: 0.0311



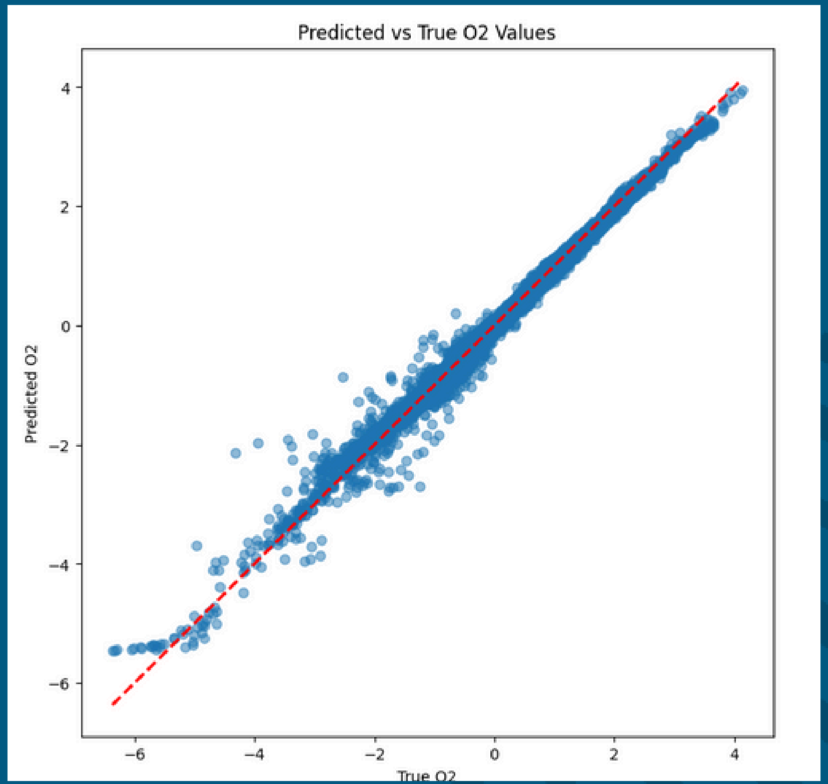
Input: 7-day sequential data
Output: 30 days prediction

mse: 0.0526



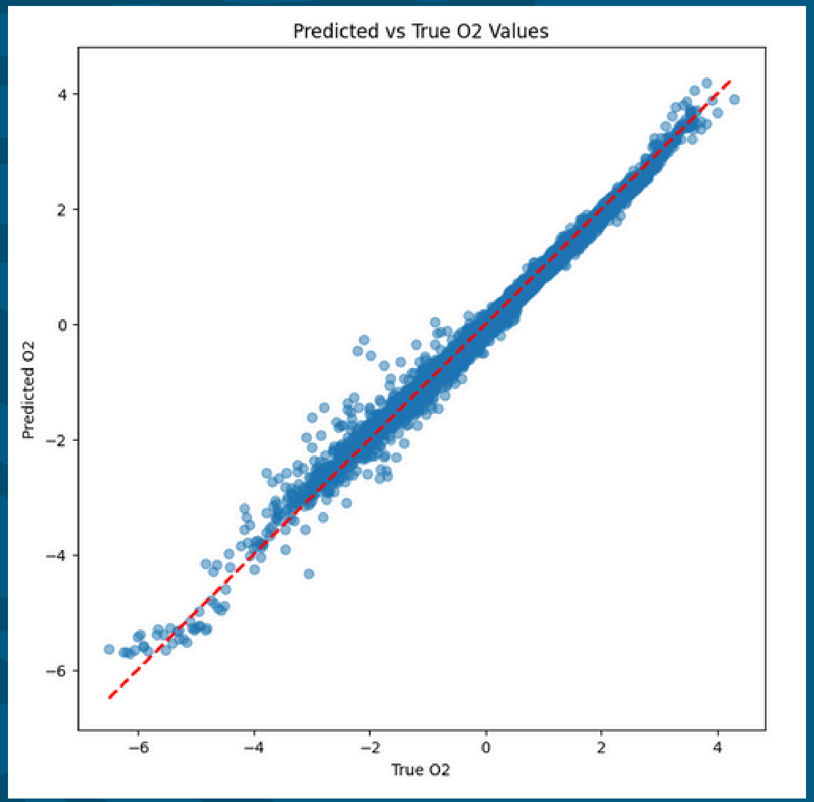
Input: 7-day sequential data
Output: 60 days prediction

mse: 0.0863



Input: 14-day sequential data
Output: 60 days prediction

mse: 0.0676



NEW ATLANTIS

Time series modeling with neural networks

- Transformed dataframe into a **numpy array** of shape num_dates x num_days x channels x num_lat x num_lon

num_dates = 813 = number of total dates **num_days** = 41 = lag + lead + current day

channels = 8 = total number of variables measured **num_lat** = 10 **num_lon** = 20

- Think of: dataset.npy = list of videos, each video with 41 frames, each frame with 8 channels

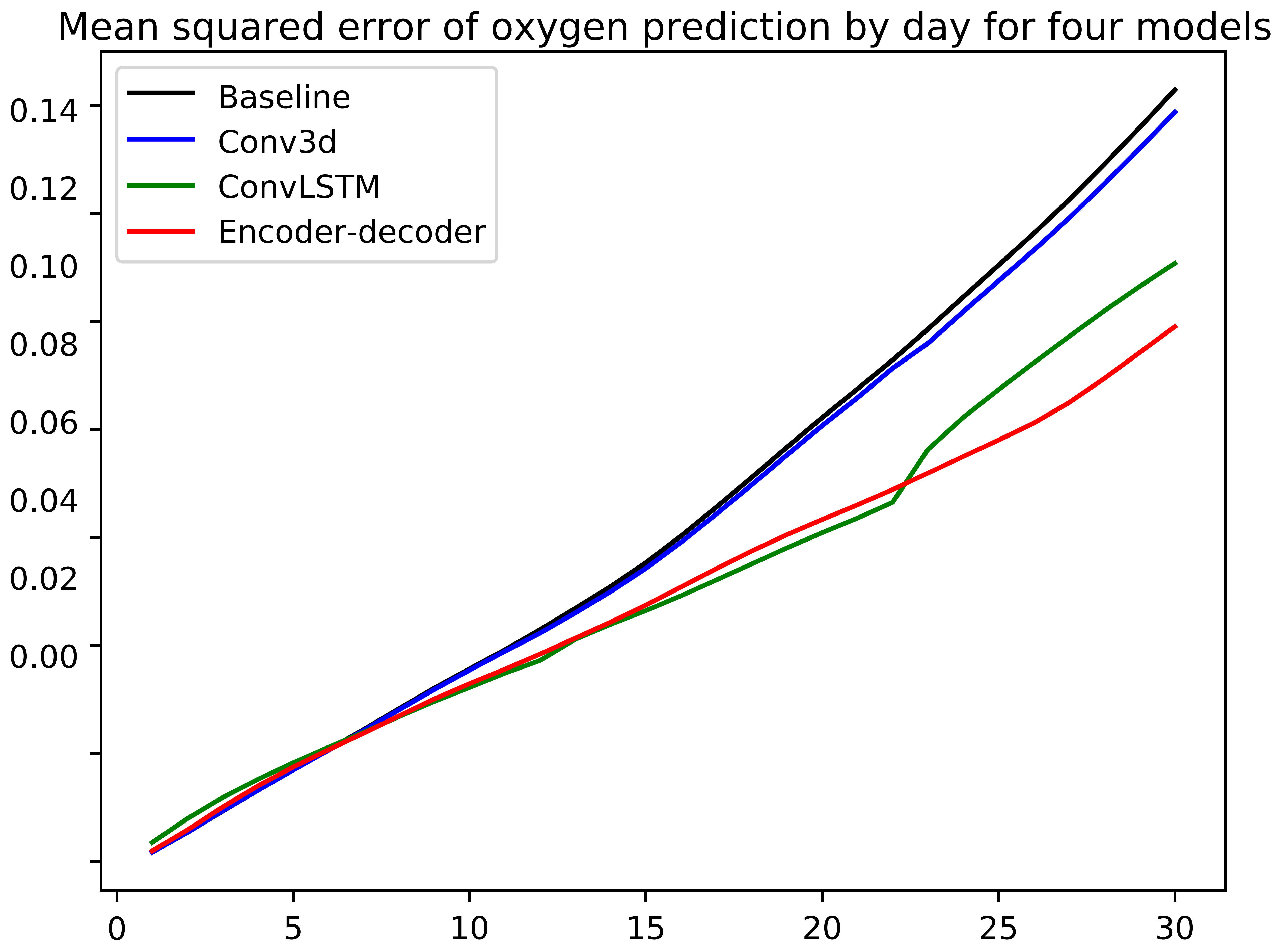
ConvLSTM models

- Convolutional LSTM (ConvLSTM) cells introduced in “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting”
- ConvLSTM cells mimic architecture of LSTM cells while using convolutions instead of matrix multiplication
- Used implementations of ConvLSTM cells and architectures based on them modified from models available freely on GitHub
- Two models:
  - ConvLSTM stacks ConvLSTM cells and concatenates their hidden states to form output
  - EncoderDecoderConvLSTM has an encoder and decoder made of ConvLSTM cells
    - decoder receives hidden state of encoder and produces output
- Added dropout and batch normalization to all models
- Each model receives one day of data at a time, updating its hidden state
- Outputs 30 days of O2 predictions (normalized)

NEW ATLANTIS

Conv3d models

- Alternatively we can use 3-dimensional convolutions to process videos
- Conv3dModel is based on residual layers using Conv3d cells
- Many of these residual layers are stacked sequentially

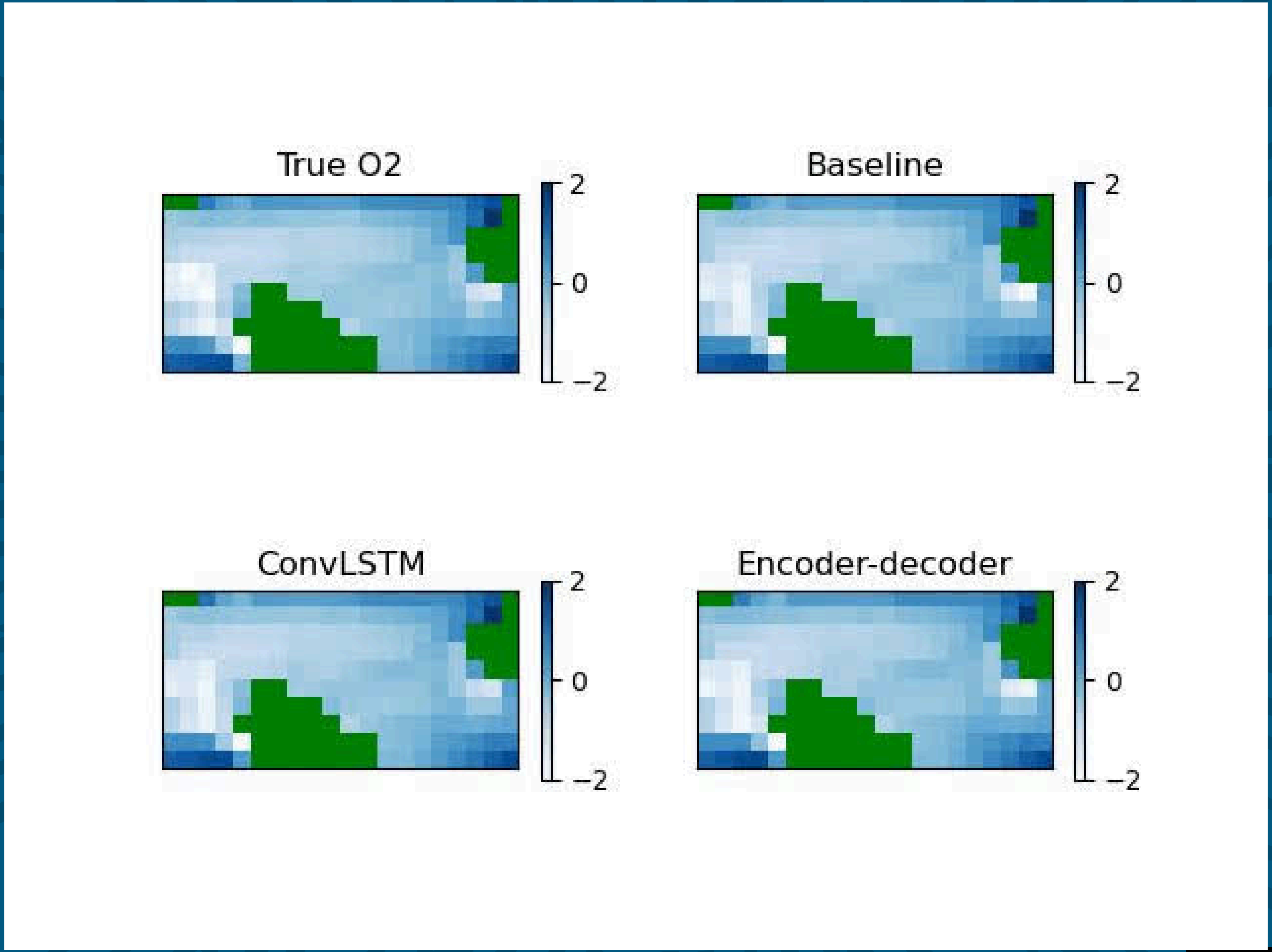

Mean squared error of oxygen prediction by day for four models
Baseline
Conv3d
ConvLSTM
Encoder-decoder
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30

Comparison

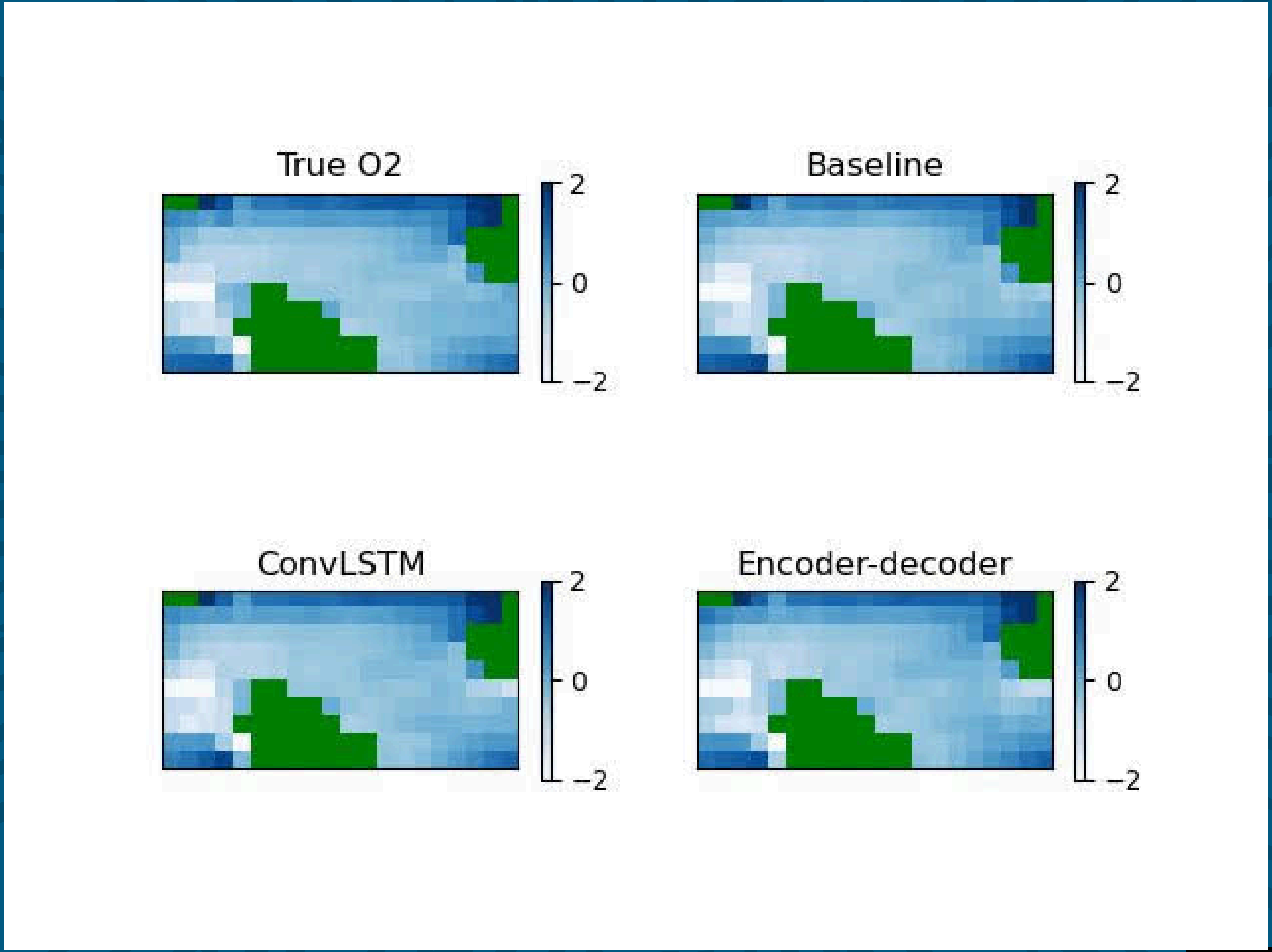
- Overall MSE
  - for baseline = 0.06396
  - for Conv3d = 0.06258
  - for ConvLSTM = 0.05177
  - for Encoder-Decoder = 0.04950
- Conv3d barely outperforms the baseline
- ConvLSTM and Encoder-Decoder outperform baseline significantly

NEW
ATLANTIS

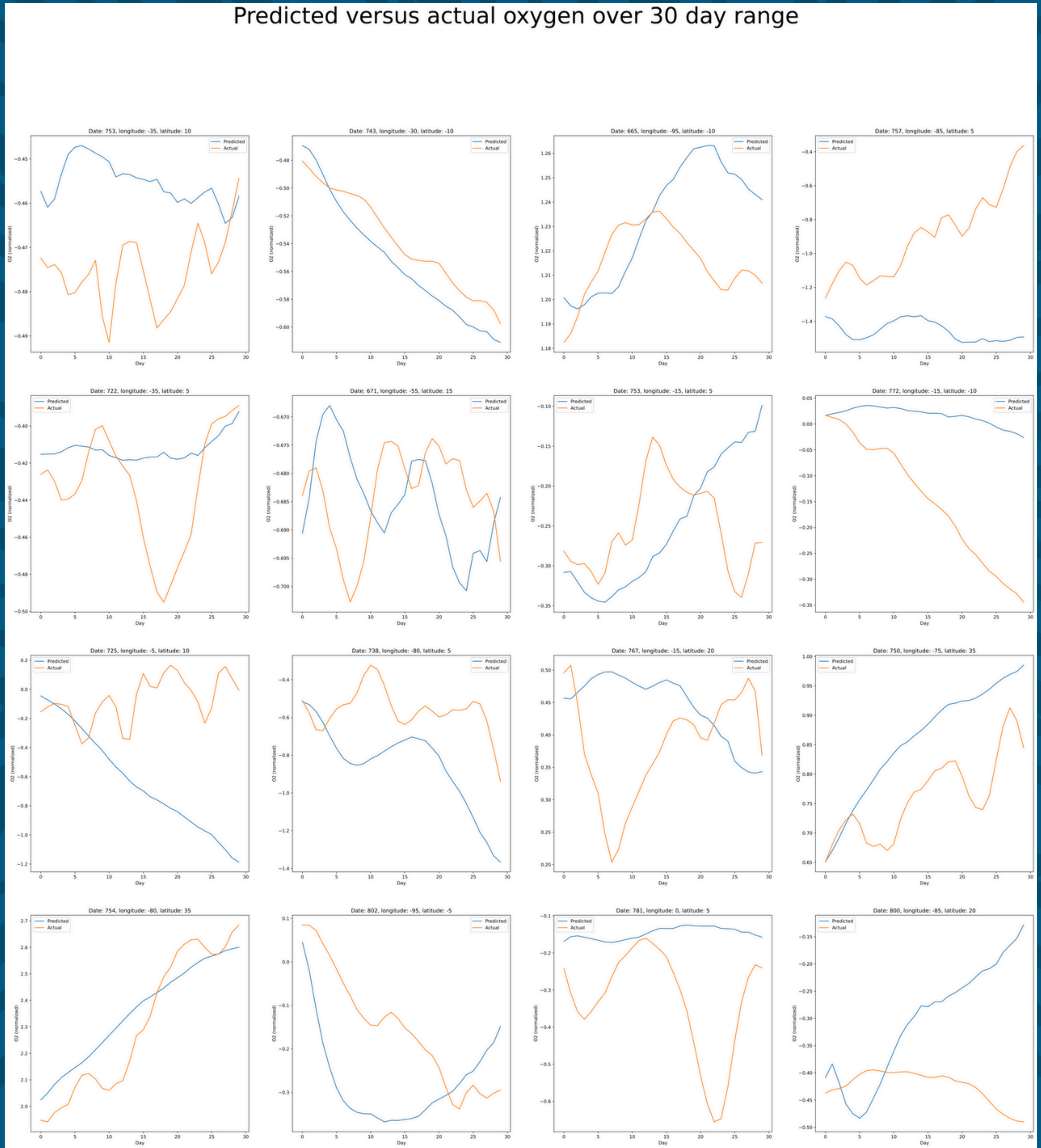
Visualizing predictions



Visualizing predictions

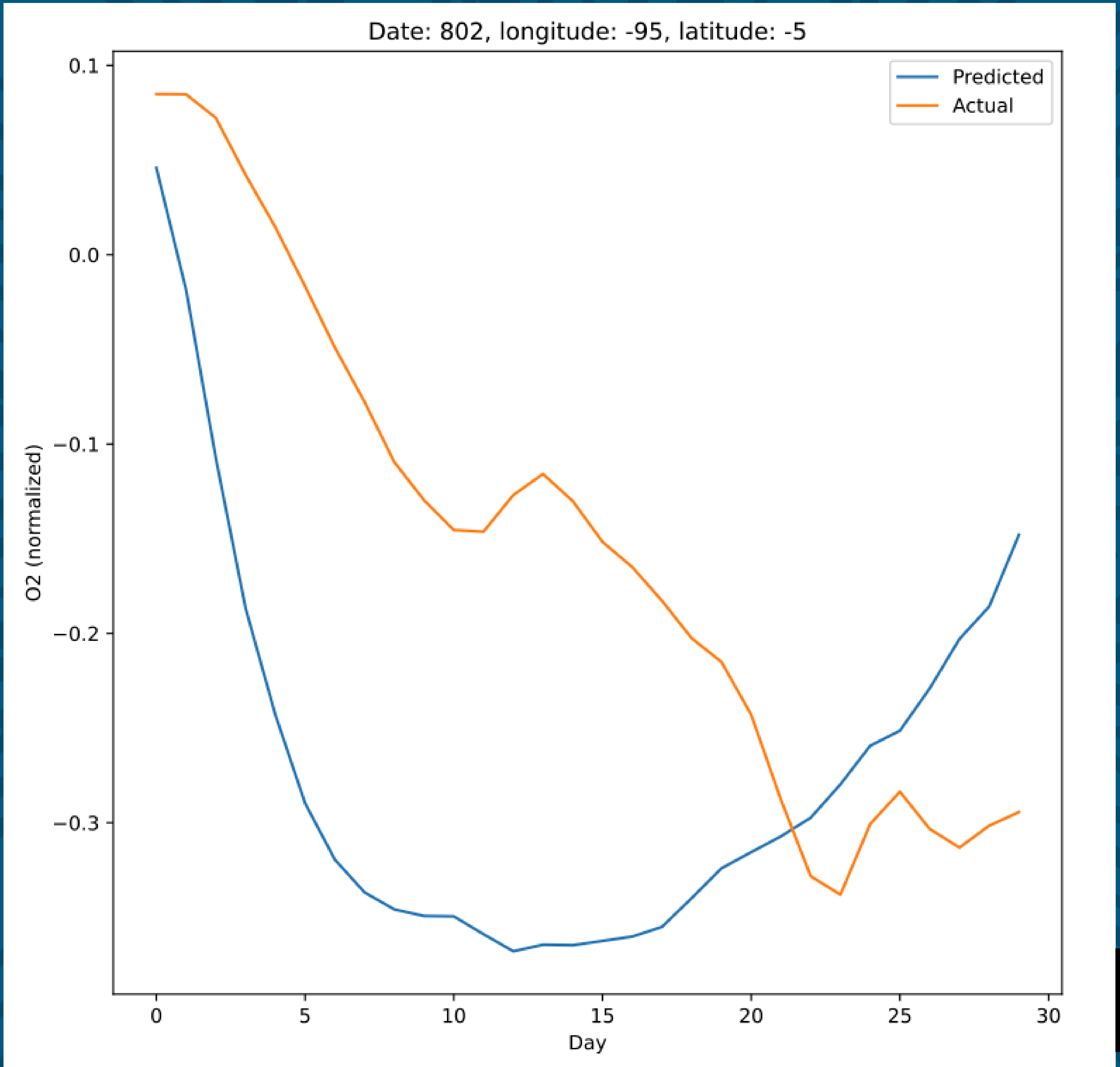


Visualizing predictions (ConvLSTM model)



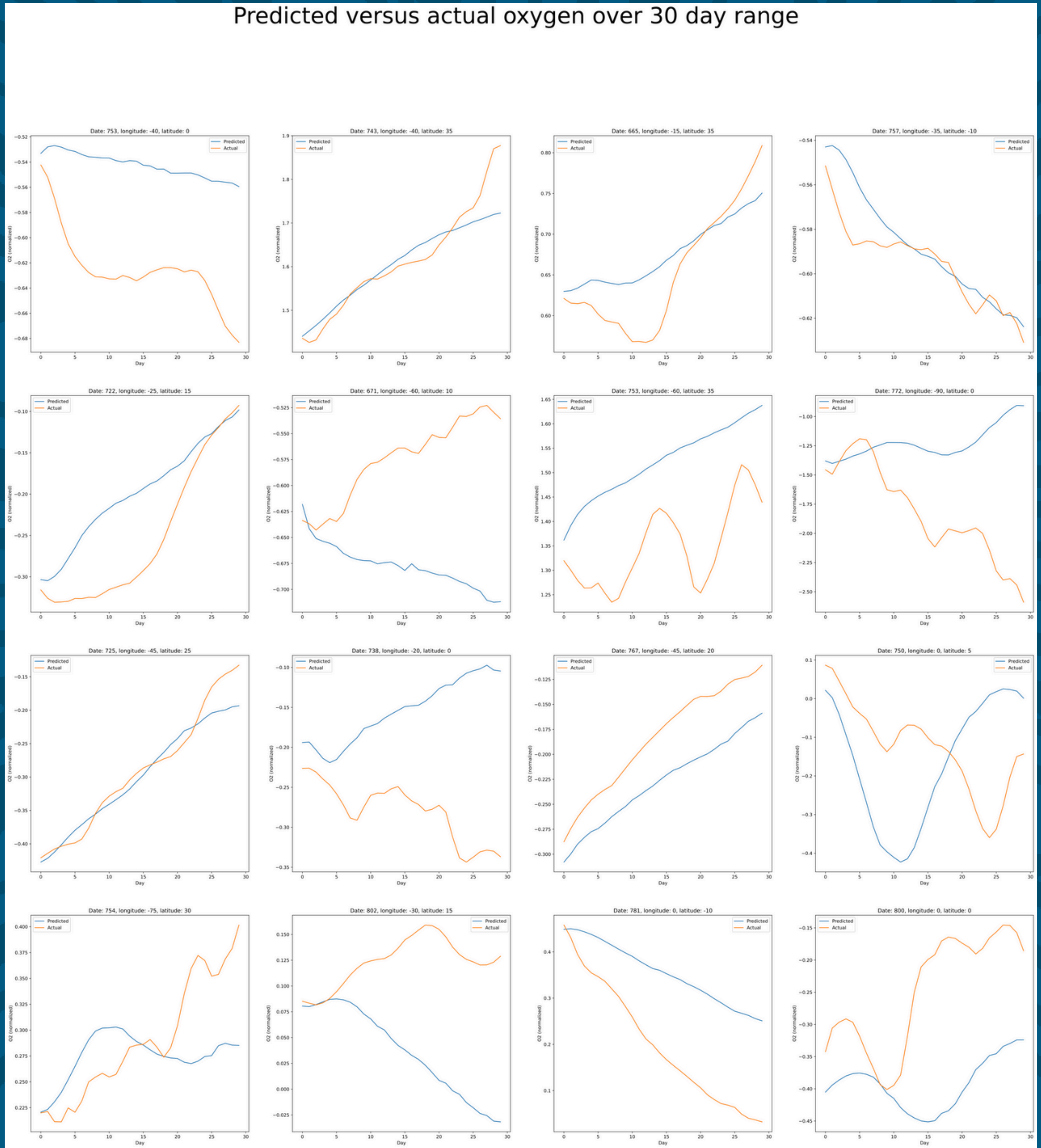
NEW ATLANTIS

Visualizing predictions (ConvLSTM model)



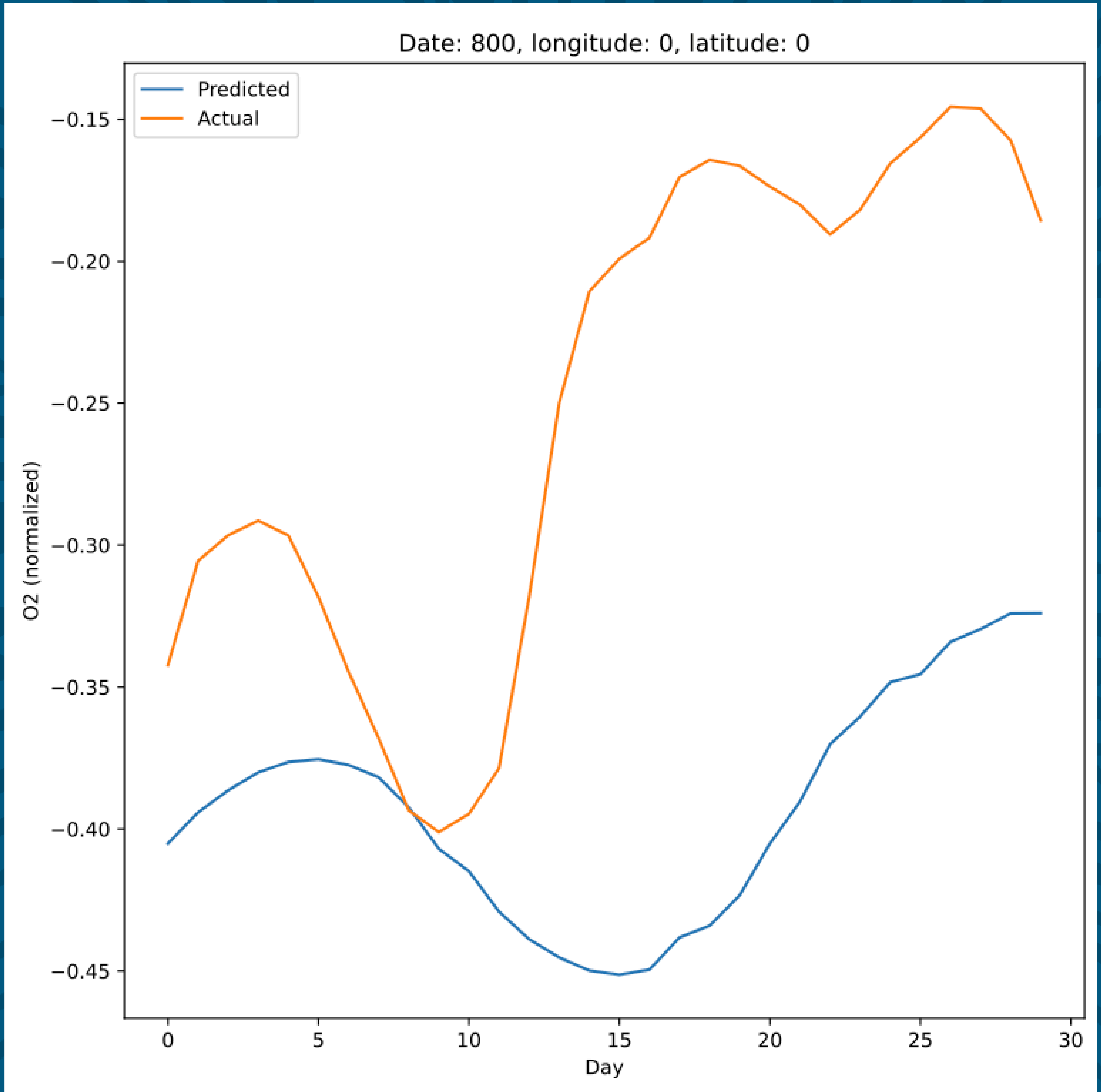
NEW ATLANTIS

Visualizing predictions (Encoder-decoder model)



NEW ATLANTIS

Visualizing predictions (Encoder-decoder model)



NEW ATLANTIS

Conclusion

- A high correlation between temperature, salinity, and O2 levels
- Possible organic confounding variables
- Neural network models can predict O2 significantly better than the baseline on long (30 day) time scales

Future directions

- Confounding variable search
- Further investigation into hyperparameter tuning
- Create automation of the model using continuously updated datasets
- Exploring other datasets and features
- Extending the analysis to a wider region
- Fine-tuning pre-trained models built for similar tasks

NEW ATLANTIS

Acknowledgments

- GEBSCO
- Copernicus Marine Service
- New Atlantis mentors (Stanley, Justin, Malik, Ollie)
- Erdős Institute
- https://github.com/ndrplz/
- https://github.com/holmdk/