Using convolutional neural networks for 3D species distribution modeling of Southern Ocean phytoplankton

Ayush Nag, Justin Chae, Lennart Wittke

University of Washington

Problem Statement and Research Question

The Southern Ocean is a large component of the global carbon cycle, and phytoplankton play a key role by converting CO_2 to organic carbon, which can be transported to the deep ocean. In the past, we have explored training a Maximum Entropy (MaxEnt) model to predict species distributions (SD) in 2D [2]. However, a drawback of this approach is that the MaxEnt modeling software only accepts 2D raster layers as features. Therefore, we aim to model phytoplankton distribution in the Southern Ocean in 3D using deep learning.

Research Questions. *Can deep learning techniques improve upon the MaxEnt SDM approach? How beneficial is it to model species distribution in 3D?*

Species Distribution Models and Datasets

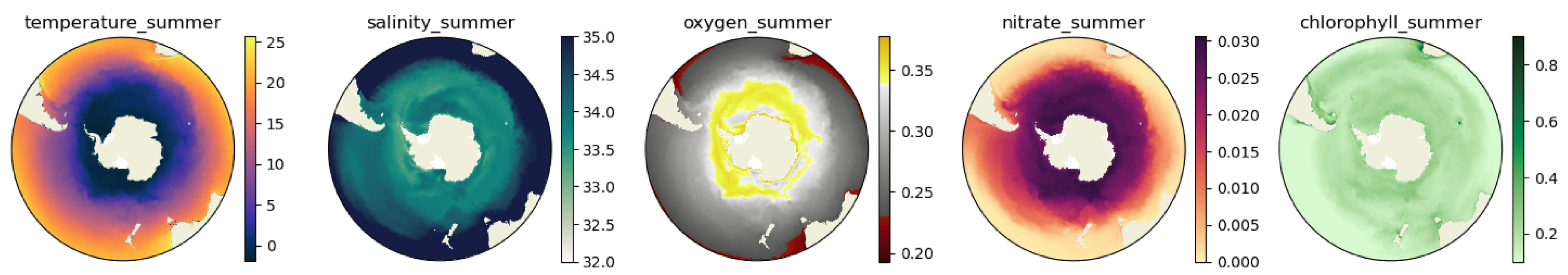
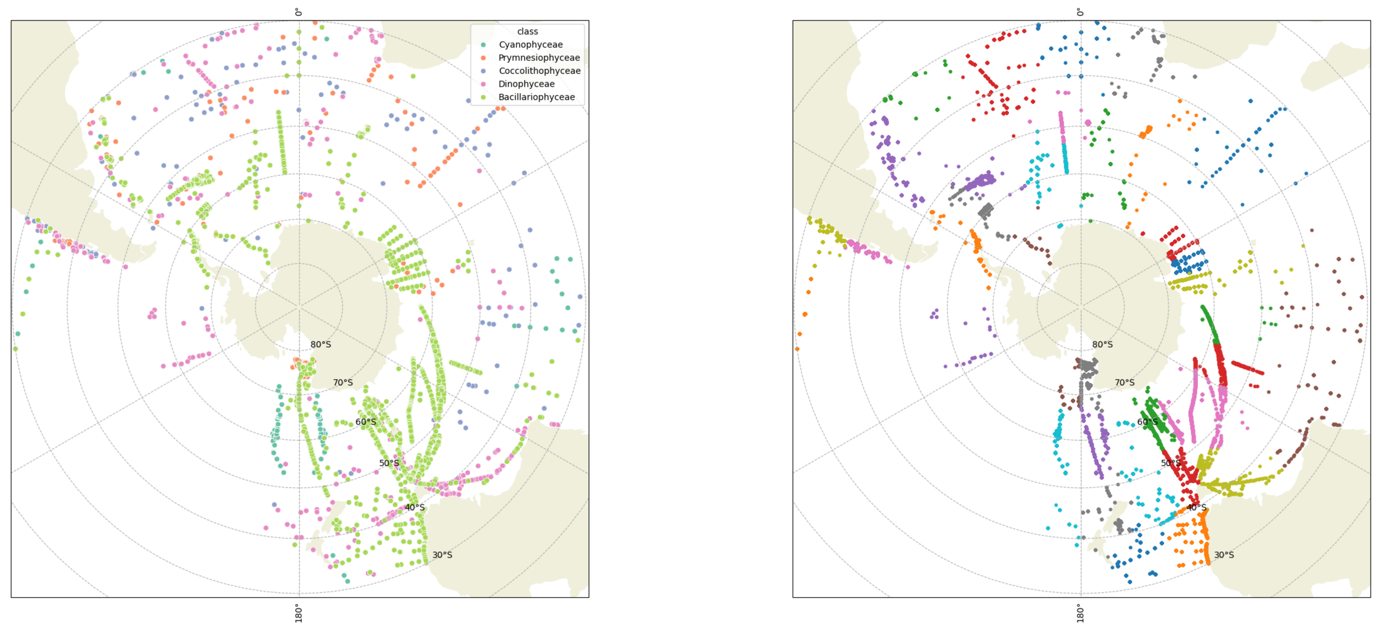


Figure: BSOSE environmental layers over summer months; Using averaged top 50 m depth

BSOSE: This dataset covers biogeochemical variables for the Southern Ocean from 2013 to 2018. The five input features include temperature, salinity, oxygen, nitrate, and chlorophyll.



(a) Species Occurence Map (b) Spatial Clustering for Validation

Figure: Data Preprocessing and Representation

PhytoBase: This dataset contains phytoplankton occurrence data. We focused on the top 5 phytoplankton classes with the largest data points, totaling 36,000 presence points. The top 5 classes and their corresponding percent composition as follows: (*Bacillariophyceae*: 54.9%, *Dinophyceae*: 22.0%, *Prymnesiophyceae*: 10.7%, *Coccolithophyceae*: 7.7%, and *Cyanophyceae*: 4.6%).

Proposed Deep Learning Method

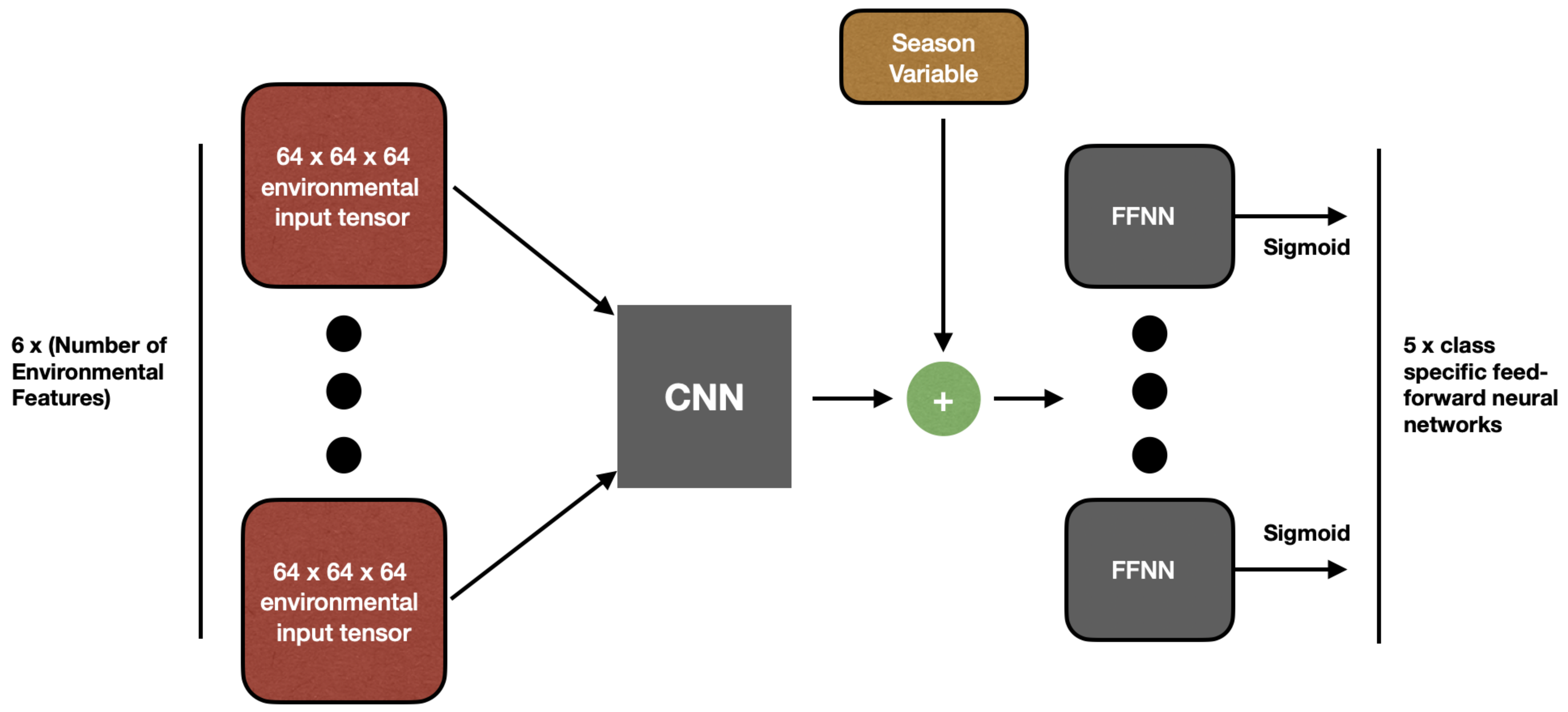


Figure: Proposed CNN architecture

Input Image. The environmental structure around a precise location also matters for accurate estimation [1]. Therefore, we use a 3D input tensor of dimension $8x8x8$ (or $16x16x16$) around each location and for each feature.

Model Architecture. We used a CNN to encode the input image into a feature vector, then applied a FFNN with a sigmoid activation for each species to learn independent class probabilities. This approach reflects the co-existence of multiple species at a given location. Additionally, we encode time (the month) as a categorical variable to account for distribution shifts throughout the year.

CNN Module. We employed two different architectures:

- a simple 3-layer CNN based on 3D convolution (1.17 million parameters)
- a much larger 18 layer ResNet3D model [3] (33.16 million parameters)

Evaluation and Baseline Models

Baselines. We compared our approach to a Random Forest (RF) model, a baseline MLP, and a biased random guess (BRG), using environmental features at specific locations as input. We also trained models on 2D data to assess the information gain from including depth.

Spatial Clustering. To address overfitting caused by spatial autocorrelation from clustered ship survey data, we avoid using random train/test splits, which can still bias results. Instead, we additionally employ geographic k-fold splits, dividing the data into spatially defined clusters for train and test sets.

Results and Discussion

Classifier	F1-score	Balanced Accuracy	Accuracy
Standard Validation (3D Data)			
ResNet3D (8x8)	0.7238	0.6398	0.7310
CNN-SDM (8x8)	0.7660	0.6716	0.7715
ResNet3D (16x16)	0.7815	**0.7196**	0.7879
CNN-SDM (16x16)	**0.7826**	0.7162	**0.7896**
MLP	0.6112	0.4858	0.6510
RF	0.7302	0.6512	0.7344
Random (Biased)	0.3534	0.2122	0.3485
Spatial Validation (3D Data)			
ResNet3D (8x8)	0.4372	**0.3584**	0.4975
CNN-SDM (8x8)	**0.4694**	0.3413	**0.4979**
MLP	0.3655	0.2673	0.4373
RF	0.3519	0.2533	0.4212
Random (Biased)	0.2782	0.2018	0.2909
Ablation Study (2D Data)			
CNN-SDM (8x8)	0.7995	0.7342	**0.8033**
RF	**0.80**	0.74	0.8010
MaxEnt	0.7487	**0.7601**	0.7511

- Our CNN-based approach significantly outperforms other baseline methods.
- As expected, the results for a spatial train-test split are considerably worse, as it is a more challenging task where the model must predict for unseen regions.
- 2D data performs slightly better, likely due to avoiding sparsity from the additional dimension.

Conclusion and Future Steps

- We showed that there is potential for species distribution modeling using CNNs. However, data scarcity still denotes are huge issue which makes the predictions ins some areas unreliable
- 3D data complicates training and may slightly reduce performance but enables insights into species evolution with ocean depth.
- With more compute and data, CNNs most likely outperform other methods by a large amount, justifying longer training times and compute expenses.
- With a high-performance CNN model in place, future research could focus on identify which environmental features most influence the presence of specific phytoplankton species.

References

[1] Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz, and Alexis Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, 17(4):e1008856, 2021.

[2] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.

[3] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.