Neural Emulator based on Physical fields for Accelerating the Simulation of Surface Chlorophyll in an Earth System Model

Bizhi Wu[a,1], Shiyao Zheng[a,1], Shasha Li[a], Shanlin Wang[a,*]

[a] *State Key Laboratory of Marine Environmental Science & College of Ocean and Earth Sciences, Xiamen University, 361102, Xiamen, China*

1. Introduction

This is the supplementary for the paper of ”Neural Emulator based on Physical fields for Accelerating the Simulation of Surface Chlorophyll in an Earth System Model”

Appendix A. Evaluation of Additional Feature Combinations Using Divergence of UVEL and VVEL

Table A.1: Performance metrics for various combinations of input features incorporating the divergence of UVEL and VVEL. Metrics include MAE, MAPE, R^2, PSNR, and RMSE.

Combination	**MAE**	**RMSE**	**MAPE**	R^2	**PSNR**
TEMP QSW dx+dy	7.07×10^{-3}	8.22×10^{-3}	2.96×10^{-2}	-14.47	19.22
TEMP QSW dx	8.70×10^{-3}	9.91×10^{-3}	3.70×10^{-2}	-18.15	18.42
TEMP QSW dy	1.16×10^{-2}	1.30×10^{-2}	4.99×10^{-2}	-28.30	18.61
TEMP dx dy	1.02×10^{-2}	1.14×10^{-2}	4.35×10^{-2}	-13.87	17.65
QSW dx dy	3.55×10^{-2}	3.61×10^{-2}	1.71×10^{-1}	-381.63	18.21
dx dy dx+dy	3.54×10^{-2}	3.56×10^{-2}	1.27×10^{-1}	-231.93	26.04

[*] Corresponding author

Email address: shlwang@xmu.edu.cn (Shanlin Wang)

[1] These authors contributed equally to this work.

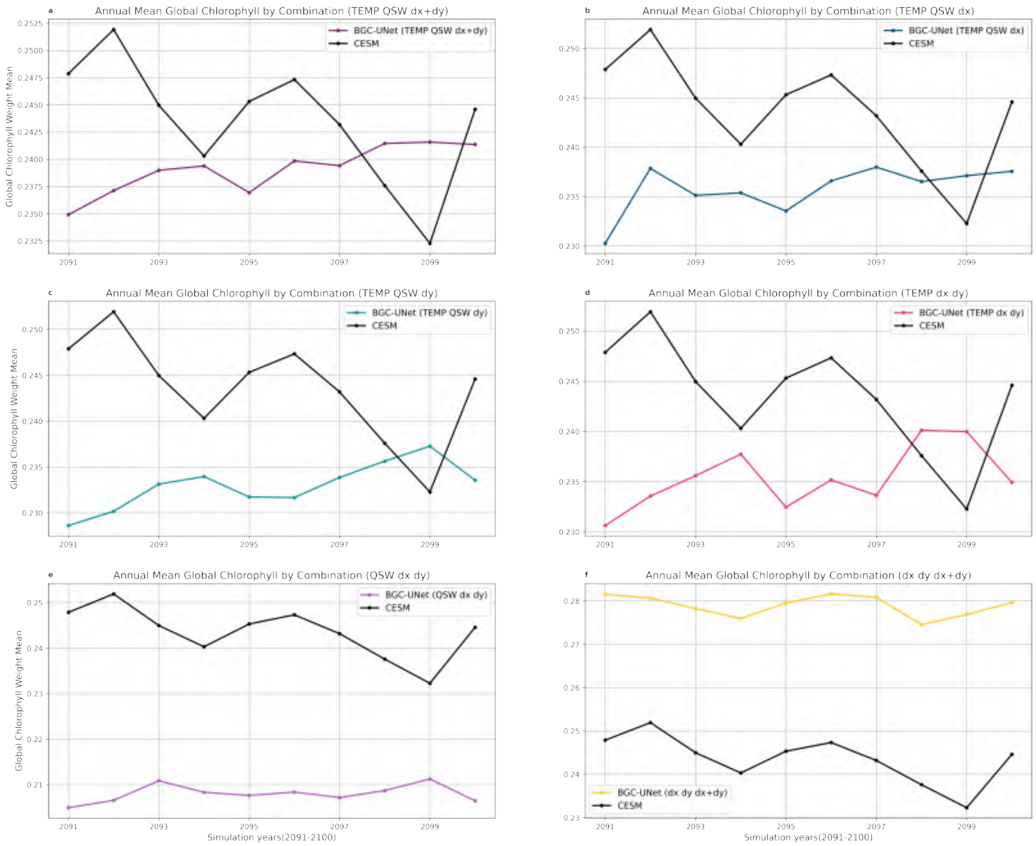

Figure A.1: Annual time series of chlorophyll concentrations predictions using different combinations of input features: (Top Left) TEMP QSW dx+dy, (Top Right) TEMP QSW dx, (Middle Left) TEMP QSW dy, (Middle Right) TEMP dx dy, (Bottom Left) QSW dx dy, (Bottom Right) dx dy dx+dy. The plots compare the BGC-UNet model outputs (colored lines) against the CESM-LE data (black line) to assess the impact of incorporating the divergence of UVEL and VVEL (dx, dy) on model performance.

Appendix B. Patch-Based BGC-UNet for High-Resolution Biogeochemical Data

Patch processing may serve as a viable strategy for managing high-resolution data, especially when models are trained on datasets of lower resolution. This method, which segments large datasets for individual processing before reintegration, could potentially refine the output quality.

In defining the dimensions of input data as $\alpha \times \text{lon}$ by $\beta \times \text{lat}$, where α

and β represents the number of submaps along longitude and latitude (the dilation factor) respectively. lon and lat denote the latitude and longitude of the biogeochemical data. we propose the following steps for potential high-resolution output generation:

1. Dividing and Extracting
   For a 4D tensor T with shape $(time, features, \alpha \times lon, \beta \times lat)$, we aim to extract sub-tensors $ST_{i,j}$ of shape $(time, features, lon, lat)$ using dilation factors α and β for the longitude and latitude dimensions, respectively.
   Let's denote:
   T as the input tensor.
   $ST_{i,j}$ as the sub-tensor at spatial position (i, j).
   α and β as the dilation factors for lon and lat dimensions, respectively.
   Each element in $ST_{i,j}$ is defined by:
   $$ST_{i,j}(t, f, x, y) = T(t, f, \min(\lceil i + x \cdot \alpha \rceil, \text{rows} - 1), \min(\lceil j + y \cdot \beta \rceil, \text{cols} - 1)) \tag{B.1}$$
   Where:
   t and f are the indices for the time and features dimensions, respectively.
   x, y are the coordinates within the spatial dimensions of $ST_{i,j}$.
   $\lceil \cdot \rceil$ represents the ceiling function, mapping the indices to the nearest integer upwards.
   min() ensures that the indices do not exceed the original tensor's dimensions, particularly near the edges.
2. Processing
   Each patch is fed through the BGC-UNet, which processes the patch as if it were an independent tensor of the same shape, lon × lat.
3. Rebuilding
   After processing each sub-tensor $ST_{i,j}$ to obtain modified sub-tensors $ST'_{i,j}$, the goal is to mapping the output tensor from these modified sub-tensors. The features of $ST'_{i,j}$ is f', represents the numbers of biogeochemical variables which desired output.
   The mapped output tensor, denoted as R, is defined by inversing the

dilated sampling process:

$$R(t, f', \min(\lceil i+x\cdot\alpha\rceil, \mathrm{rows}-1), \min(\lceil j+y\cdot\beta\rceil, \mathrm{cols}-1)) = ST'_{i,j}(t, f', x, y) \tag{B.2}$$

Where:
R is the rebuilt tensor.
$ST'_{i,j}$ is the processed sub-tensor.
t, f', i, j, x, y maintain their previous definitions but for the feature dimension.
This process places each element from the processed sub-tensors into their corresponding positions in the larger tensor R, effectively generating the output tensor with the modified values.

The approach outlined provides a systematic method to manage and manipulate 4D tensors, particularly large ones, by dividing them into smaller, more manageable sub-tensors using dilated sampling. This method is especially useful when direct processing of the entire tensor is impractical or impossible due to resource constraints.

Appendix C. Preliminary Findings on High-Resolution Data Reconstruction

In our preliminary studies, we utilize data from the International Laboratory for High-Resolution Earth System Prediction (iHESP) to evaluate our methods for reconstructing high-resolution simulations. Although the reconstructed biogeochemical data exhibits some edge noise, this exploratory work points to the potential of methods like BGC-UNet in this context.

In our study, we utilize data from the International Laboratory for High-Resolution Earth System Prediction (iHESP) to validate our high-resolution reconstruction methods. IHESP, a collaborative initiative, is dedicated to advancing modeling frameworks for high-resolution, multiscale Earth System predictions. It strives to deliver reliable information across global and regional scales by harnessing the collective expertise of three leading institutions. Notably, iHESP is renowned for its comprehensive high-resolution, coupled climate simulations that span the entire globe and provide regionally downscaled simulations for specific areas of interest, such as the Gulf of Mexico.

For our reconstruction, we select four variables: UVEL, VVEL, SHF-QSW, and TEMP, analogous to those used in BGC-UNet. However, the dimensions we work with are $(2400, 3600)$, as opposed to the $(384, 320)$ typically use, resulting in non-integer scaling factors of $\alpha = 6.25$ and $\beta = 11.25$. This discrepancy introduces the potential for edge effects due to the non-integral dilation factor, meaning higher resolution inputs can not be segmented into non-overlapping patches. Consequently, this leads to multiple outputs for the same locations and subsequently results in the reconstructing biogeochemical data containing considerable edge noise.

To our knowledge, no higher resolution data exists that precisely aligns with the $(320, 384)$ dimensions of the CESM-LE dataset. Despite the challenges pose by the floating dilation sampling, iHESP remains the most suitable dataset available to us. Figure 7 showcases the reconstructed high-resolution global chlorophyll distribution for January 2006.

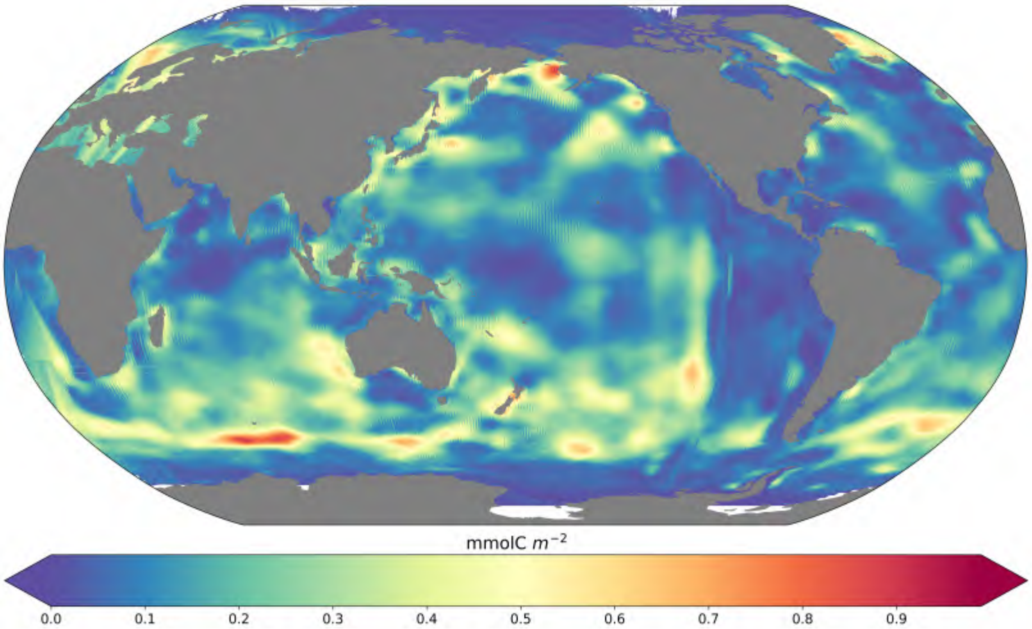

Figure C.2: High-resolution (2400, 3600) chlorophyll data based on BGC-UNet with iH-ESP inputs in Jan 2006 (minmax scaler).

Appendix D. Reflection on High-Resolution Reconstruction Attempts

Generating high-resolution biogeochemical data is a significant aspect of BGC-UNet's capabilities with patch processing. However, the current unavailability of aligned, high-resolution physical and biogeochemical datasets necessary for a comprehensive validation of the model's output remains a critical limitation. This highlights a crucial gap in the field and underscores the importance of future efforts to collect and integrate high-resolution environmental data.

Furthermore, while high-resolution satellite remote sensing data are available, these datasets often serve as validation tools rather than direct inputs for biogeochemical modeling. There is a need for methodological advancements to align and integrate remote sensing data with model outputs effectively. This alignment is not straightforward due to differences in the spatial and temporal resolutions and the diverse nature of the data captured by remote sensing technologies compared to the outputs of biogeochemical models.

As for the availability of high-resolution physical data, it opens the possibility of constructing high-resolution mappings from low-resolution biogeochemical models. This can potentially bridges the gap between the resolutions; however, the integration of such mappings with tools like BGC-UNet poses another set of challenges. For instance, the scaling of biogeochemical processes from low to high resolution is not linear and may involve complex interactions that are spatially and temporally variable.

Appendix E. Time-series of chlorophyll at 14 sample sites

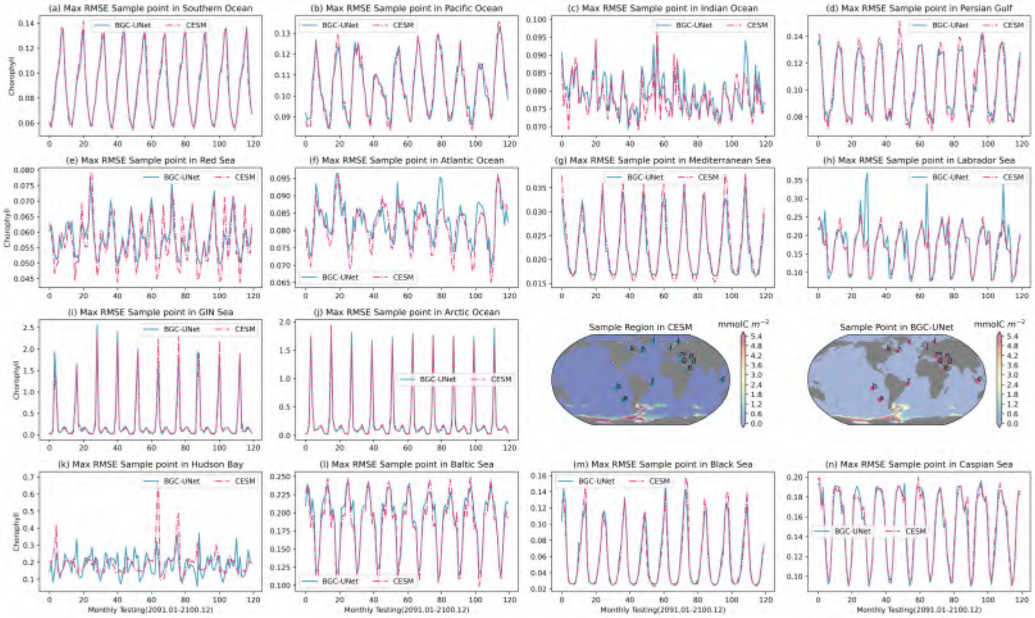
(a) Max RMSE Sample point in Southern Ocean
(b) Max RMSE Sample point in Pacific Ocean
(c) Max RMSE Sample point in Indian Ocean
(d) Max RMSE Sample point in Persian Gulf
(e) Max RMSE Sample point in Red Sea
(f) Max RMSE Sample point in Atlantic Ocean
(g) Max RMSE Sample point in Mediterranean Sea
(h) Max RMSE Sample point in Labrador Sea
(i) Max RMSE Sample point in GIN Sea
(j) Max RMSE Sample point in Arctic Ocean
Sample Region in CESM
Sample Point in BGC-UNet
(k) Max RMSE Sample point in Hudson Bay
(l) Max RMSE Sample point in Baltic Sea
(m) Max RMSE Sample point in Black Sea
(n) Max RMSE Sample point in Caspian Sea

Figure E.3: Time-series of chlorophyll at 14 sample sites using BGC-UNet. These sites correspond to regions with the largest RMSE between the model predictions and the CESM-LE (member 01) data.

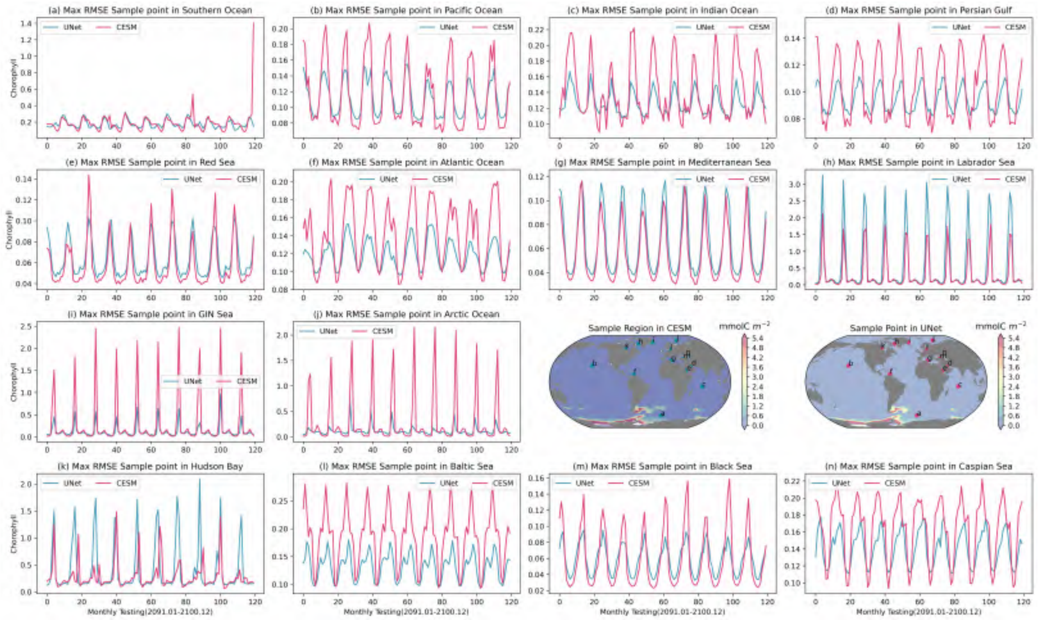

Figure E.4: Time-series of chlorophyll at 14 sample sites using UNet. These sites correspond to regions with the largest RMSE between the model predictions and the CESM-LE (member 01) data.

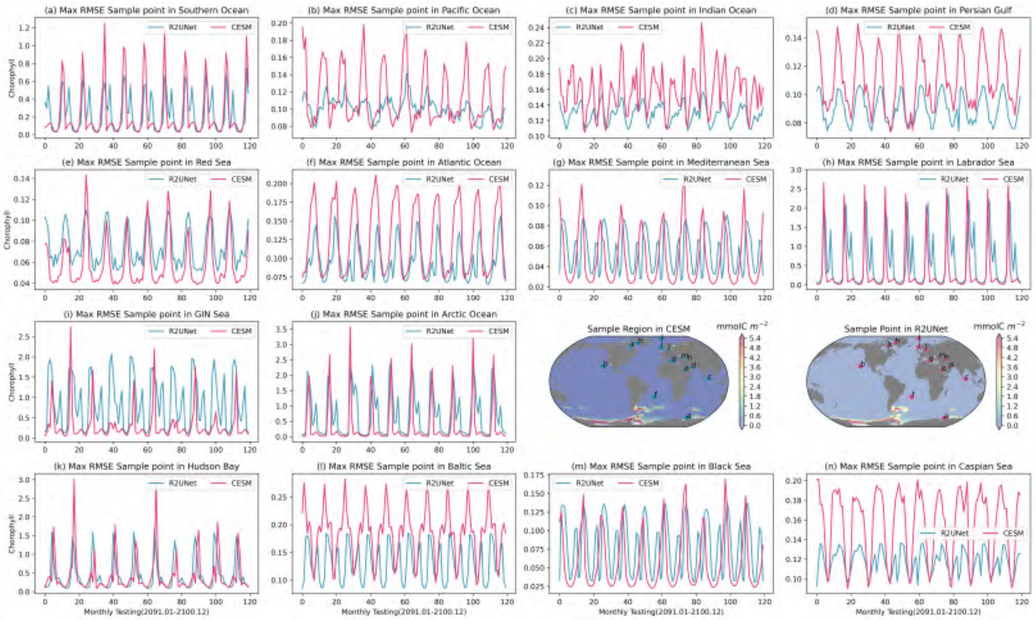

Figure E.5: Time-series of chlorophyll at 14 sample sites using R2UNet. These sites correspond to regions with the largest RMSE between the model predictions and the CESM-LE (member 01) data.

Appendix F. Basin-wide Monthly Anomalies

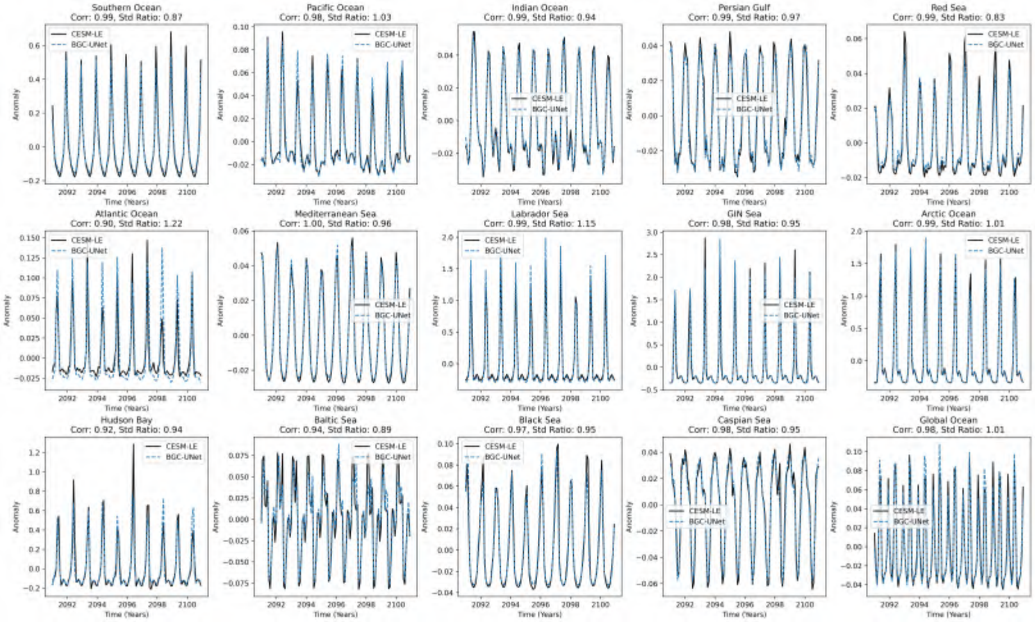

Figure F.6: Comparison of chlorophyll anomalies between BGC-UNet and CESM-LE across all 14 ocean basins and the global ocean. The correlation coefficients (Corr.) and standard deviation ratios (Std Ratio) are displayed for each basin. This figure provides a comprehensive analysis of how well BGC-UNet captures the temporal variability of chlorophyll concentrations across diverse oceanic regions, including smaller and more complex basins not highlighted in the main text.

Appendix G. Basin-wide evaluation metrics

Table G.2: Evaluation metrics for BGC-UNet, UNet, and R2UNet across 8 ocean basins. The metrics include RMSE, Reliability Index (RI), Average Error (AE), Average Absolute Error (AAE), and Modeling Efficiency (MEF).

Basin	Model	RMSE	RI	AE	AAE	MEF
Southern Ocean	BGC-UNet	0.3011	1.448	0.0129	0.0800	0.7364
	UNet	0.5697	4.487	-0.1439	0.2231	0.0565
	R2UNet	0.5810	4.413	-0.1494	0.2259	0.0186
Pacific Ocean	BGC-UNet	0.1304	1.179	-0.0024	0.0229	0.7598
	UNet	0.3554	3.244	0.1282	0.2240	-0.7853
	R2UNet	0.3547	3.215	0.1258	0.2214	-0.7782
Indian Ocean	BGC-UNet	0.0375	1.128	-0.0022	0.0112	0.5439
	UNet	0.2545	2.440	0.1703	0.1888	-19.9778
	R2UNet	0.2464	2.369	0.1674	0.1841	-18.6563
Persian Gulf	BGC-UNet	0.0124	1.115	0.0003	0.0089	0.8205
	UNet	0.3395	3.092	0.2520	0.2560	-133.3716
	R2UNet	0.3097	2.930	0.2335	0.2350	-110.8377
Red Sea	BGC-UNet	0.0127	1.189	0.0009	0.0074	0.6647
	UNet	0.1010	2.460	0.0468	0.0626	-20.1692
	R2UNet	0.0874	2.000	0.0434	0.0531	-14.8407
Atlantic Ocean	BGC-UNet	0.1433	1.308	-0.0086	0.0292	0.5890
	UNet	0.3155	3.364	0.0703	0.1954	-0.9939
	R2UNet	0.3058	3.280	0.0696	0.1918	-0.8733
Mediterranean Sea	BGC-UNet	0.0171	1.202	-0.0005	0.0078	0.8247
	UNet	0.1832	3.026	0.1097	0.1170	-19.0257
	R2UNet	0.1749	2.949	0.1096	0.1141	-17.2478
Labrador Sea	BGC-UNet	0.3004	1.655	-0.0357	0.1214	0.8192
	UNet	0.6717	6.970	-0.3056	0.3065	0.0957
	R2UNet	0.6641	6.598	-0.3017	0.3022	0.1159

Table G.3: Evaluation metrics for BGC-UNet, UNet, and R2UNet across the other 6 ocean basins and the global ocean. The metrics include RMSE, Reliability Index (RI), Average Error (AE), Average Absolute Error (AAE), and Modeling Efficiency (MEF).

Basin	**Model**	**RMSE**	**RI**	**AE**	**AAE**	**MEF**
GIN Sea	BGC-UNet	0.3562	1.967	-0.0015	0.1356	0.7963
	UNet	0.7684	8.230	-0.3183	0.3185	0.0524
	R2UNet	0.7640	7.645	-0.3156	0.3158	0.0630
Arctic Ocean	BGC-UNet	0.3074	1.676	-0.0100	0.1143	0.8488
	UNet	0.7385	9.118	-0.3007	0.3012	0.1271
	R2UNet	0.7402	8.313	-0.3021	0.3027	0.1231
Hudson Bay	BGC-UNet	0.2876	1.793	0.0227	0.1311	0.5951
	UNet	0.4398	4.969	-0.2303	0.2317	0.0534
	R2UNet	0.4296	4.964	-0.2248	0.2264	0.0970
Baltic Sea	BGC-UNet	0.0457	1.262	0.0029	0.0282	0.4049
	UNet	0.1596	2.444	0.0474	0.1192	-6.2495
	R2UNet	0.1483	2.369	0.0355	0.1103	-5.2616
Black Sea	BGC-UNet	0.0165	1.243	0.0011	0.0101	0.8258
	UNet	0.0752	2.230	0.0373	0.0485	-2.6057
	R2UNet	0.0739	1.954	0.0403	0.0471	-2.4842
Caspian Sea	BGC-UNet	0.0157	1.102	0.0028	0.0112	0.8092
	UNet	0.3772	2.976	0.3029	0.3163	-108.4636
	R2UNet	0.3586	2.794	0.2824	0.2950	-97.9681
Global Ocean	BGC-UNet	0.1732	1.348	-0.0018	0.0343	0.7822
	UNet	0.3776	4.811	0.0066	0.1569	-0.0348
	R2UNet	0.3787	4.131	0.0051	0.1558	-0.0410