

PhyOceanCast: Global Ocean Forecasting with Physics-Informed Diffusion

Supplementary Material

7. Dataset

7.1. GLORYS12

GLORYS12V1 (Global Ocean Reanalysis and Simulation, Version 1) is a state-of-the-art eddy-resolving global ocean reanalysis dataset developed by Mercator Ocean International within the Copernicus Marine Environment Monitoring Service (CMEMS) framework [21]. This dataset provides continuous global ocean state estimates with comprehensive coverage of physical oceanographic variables essential for spatiotemporal ocean forecasting tasks.

Spatiotemporal Coverage. The dataset spans from January 1, 1993 to present, coinciding with the start of the satellite altimetry era. The reanalysis provides outputs at both daily mean and monthly mean temporal resolutions, with operational updates extending coverage to within approximately two months of present time through near-real-time assimilation.

Spatial Resolution. GLORYS12 employs the ORCA12 tripolar curvilinear grid configuration with $1/12^\circ$ horizontal resolution, corresponding to approximately 9.25 km at the equator and 4.5 km at subpolar latitudes. The grid dimensions encompass 4,320 points zonally and 2,041 points meridionally, covering from 77°S to 90°N . The vertical discretization utilizes 50 z-coordinate levels with partial bottom steps, featuring 22 levels concentrated in the top 100 meters with approximately 1-meter spacing near the surface, progressively decreasing to 450-meter spacing at 5,000-meter depth.

Physical Variables. The dataset provides three-dimensional fields of core oceanographic variables:

- *Temperature and Salinity:* Sea water potential temperature (θ_{tao}) and salinity (so) across all 50 vertical levels
- *Ocean Currents:* Eastward (uo) and northward (vo) velocity components in m/s at all depths
- *Sea Surface Height:* Absolute dynamic topography (zos) above geoid in meters
- *Mixed Layer Depth:* Ocean mixed layer thickness ($mlotst$) diagnosed using density threshold criterion
- *Sea Ice Variables:* Sea ice concentration ($siconc$), thickness ($sithick$), and velocity components (siu , siv)

Data Assimilation. The reanalysis integrates observations through the SAM2 (Système d'Assimilation Mercator version 2) system, utilizing a reduced-order Kalman filter based on the SEEK (Singular Evolutive Extended Kalman) formulation. The system assimilates four primary observation types: (1) along-track satellite altimetry from multiple missions including TOPEX-Poseidon, Jason series, and

Sentinel-3A/B, (2) satellite sea surface temperature from NOAA AVHRR at 0.25° resolution, (3) sea ice concentration from CERSAT passive microwave retrievals, and (4) in-situ temperature and salinity profiles from the CORA database, incorporating Argo floats, moored buoys, CTDs, and XBTs. **Configuration.** GLORYS12 employs the NEMO (Nucleus for European Modelling of the Ocean) version 3.1 ocean model coupled with the LIM2-EVP sea ice component. The model operates with a 24-minute timestep and implements TKE 1.5 closure for vertical mixing, Laplacian lateral isopycnal diffusion, and TVD advection schemes. Atmospheric forcing derives from ECMWF reanalysis products (ERA-Interim through 2018, ERA5 from 2019) with 3-hourly sampling.

Data Format. Products are distributed in NetCDF-4 format following CF-1.6/1.7 Climate and Forecast Metadata Conventions with compression enabled. Files are organized by variable type and temporal resolution, facilitating efficient access for large-scale spatiotemporal analysis. The dataset is identified as GLOBAL_MULTIYEAR_PHY_001_030 in the Copernicus Marine Data Store.

7.2. Data Preprocessing

Spatial Downsampling. While GLORYS12V1 provides resolution at $1/12^\circ$, we downsample the data to 1° horizontal resolution for computational tractability and to align with common climate model grids. This resolution reduction, from $4,320 \times 2,041$ to 360×171 grid points, reduces memory requirements by a factor of approximately 144 while retaining large-scale ocean dynamics essential for medium-range forecasting.

Interpolation Method. We employ bilinear interpolation to resample from the native grid to a regular 1° longitude-latitude grid. Given the high-resolution source data at coordinates (lon_{hr}, lat_{hr}) with values f_{hr} , we compute the coarse-resolution field f_{lr} at target coordinates (lon_{lr}, lat_{lr}) through:

$$f_{lr}(lon, lat) = \sum_{i,j} w_{ij} \cdot f_{hr}(lon_i, lat_j), \quad (17)$$

where w_{ij} represents the bilinear weights determined by the relative distances to the four nearest high-resolution grid points. This approach preserves smoothness in the interpolated fields while avoiding aliasing artifacts that could arise from simple nearest-neighbor or area-weighted methods.

Implementation Details. The downsampling pipeline processes the entire GLORYS12 archive from 1993 to 2020, maintaining the original directory structure organized by year and month. For each NetCDF file, we:

Table 2. Variables used from GLORYS12 dataset. Vertical variables are sampled at 36 depth levels ranging from 0.494m to 1062.44m. All variables except static and clock inputs serve dual roles as both inputs and prediction targets.

Type	Variable Name	Short Name	Depth Levels	Role
Vertical	Sea water potential temperature	thetao	36	Input/Predicted
	Sea water salinity	so	36	Input/Predicted
	Eastward velocity	uo	36	Input/Predicted
	Northward velocity	vo	36	Input/Predicted
Surface	Sea surface height	zos	n/a	Input/Predicted
Static	Land-sea mask	mask	36*	Input only
	Latitude	-	n/a	Input only
	Longitude	-	n/a	Input only
Clock	Elapsed year progress	-	n/a	Input only

*Land-sea mask is provided at each depth level to account for bathymetry.

Table 3. Vertical variables and corresponding depth levels (meters)

Depth Variable	Input/Predicted Depth Levels (m)
Sea water potential temperature (thetao)	0.494025, 1.541375, 2.645669, 3.819495, 5.078224, 6.440614, 7.92956, 9.572997, 11.405, 13.46714, 15.81007, 18.49556, 21.59882, 25.21141, 29.44473, 34.43415, 40.34405, 47.37369, 55.76429, 65.80727, 77.85385, 92.32607, 109.7293, 130.666, 155.8507, 186.1256, 222.4752, 266.0403, 318.1274, 380.213, 453.9377, 541.0889, 643.5668, 763.3331, 902.3393, 1062.44
Sea water salinity (so)	0.494025, 1.541375, 2.645669, 3.819495, 5.078224, 6.440614, 7.92956, 9.572997, 11.405, 13.46714, 15.81007, 18.49556, 21.59882, 25.21141, 29.44473, 34.43415, 40.34405, 47.37369, 55.76429, 65.80727, 77.85385, 92.32607, 109.7293, 130.666, 155.8507, 186.1256, 222.4752, 266.0403, 318.1274, 380.213, 453.9377, 541.0889, 643.5668, 763.3331, 902.3393, 1062.44
Eastward velocity (uo)	0.494025, 1.541375, 2.645669, 3.819495, 5.078224, 6.440614, 7.92956, 9.572997, 11.405, 13.46714, 15.81007, 18.49556, 21.59882, 25.21141, 29.44473, 34.43415, 40.34405, 47.37369, 55.76429, 65.80727, 77.85385, 92.32607, 109.7293, 130.666, 155.8507, 186.1256, 222.4752, 266.0403, 318.1274, 380.213, 453.9377, 541.0889, 643.5668, 763.3331, 902.3393, 1062.44
Northward velocity (vo)	0.494025, 1.541375, 2.645669, 3.819495, 5.078224, 6.440614, 7.92956, 9.572997, 11.405, 13.46714, 15.81007, 18.49556, 21.59882, 25.21141, 29.44473, 34.43415, 40.34405, 47.37369, 55.76429, 65.80727, 77.85385, 92.32607, 109.7293, 130.666, 155.8507, 186.1256, 222.4752, 266.0403, 318.1274, 380.213, 453.9377, 541.0889, 643.5668, 763.3331, 902.3393, 1062.44

- Load the high resolution data using xarray with parallel processing enabled for efficient I/O.
- Define the target 1° grid spanning the original geographic extent: longitude from $[lon_{min}]$ to $[lon_{max}]$ and latitude from $[lat_{min}]$ to $[lat_{max}]$.
- Apply scipy's linear interpolation through xarray's interp method, which handles missing values and main-

tains coordinate metadata.

- Preserve all 36 vertical levels and physical variables without modification.
- Store the downsampled data in compressed NetCDF-4 format.

Temporal Processing. We maintain the original daily temporal resolution to capture synoptic-scale variability and

diurnal cycles. The temporal sequence is organized as consecutive daily snapshots at 00:00 UTC, providing consistent sampling for the autoregressive forecasting framework. No temporal interpolation or averaging is applied to preserve the native temporal dynamics of the reanalysis.

7.3. variables used in our datasets

We utilize a subset of GLORYS12 variables for ocean forecasting, comprising 145 total channels: 144 ocean state variables (4 physical quantities \times 36 depth levels) plus sea surface height. Table 2 summarizes the input and output variables employed in our experiments.

Depth Discretization. The 36 selected depth levels follow a non-uniform vertical grid with higher resolution near the surface where ocean-atmosphere interactions are most pronounced. The depth levels (in meters) are: 0.494025, 1.541375, 2.645669, 3.819495, 5.078224, 6.440614, 7.92956, 9.572997, 11.405, 13.46714, 15.81007, 18.49556, 21.59882, 25.21141, 29.44473, 34.43415, 40.34405, 47.37369, 55.76429, 65.80727, 77.85385, 92.32607, 109.7293, 130.666, 155.8507, 186.1256, 222.4752, 266.0403, 318.1274, 380.213, 453.9377, 541.0889, 643.5668, 763.3331, 902.3393, and 1062.44. This vertical discretization captures the essential dynamics of the upper ocean (0-1000m) where the majority of oceanic heat content and kinetic energy resides.

7.4. Data Availability

The GLORYS12 reanalysis dataset is a publicly available datasets, which can be downloaded from https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_001_030/services.

8. Training details

PhyOceanCast employs a physics-informed diffusion framework that predicts ocean state evolution through residual learning with a second order Markov roll-out strategy. The model integrates two complementary modules: SGAN-MOC (Spherical Graph Attention Network for Multi-Scale Ocean Coupling) that preserves spherical topology while enabling cross-variable interactions, and PWTC (Physics-Informed Wavelet Temporal Coherence) that decomposes ocean dynamics across multiple temporal scales with advection-diffusion constraints.

Given previous ocean states \mathbf{X}_{t-2} and \mathbf{X}_{t-1} , the model predicts the residual r_t through an iterative diffusion refinement process. The temporal position encoding employs separated sine and cosine components, $\tau_t = [\sin(2\pi t/T), \cos(2\pi t/T)]$, to capture intra-annual periodicity and seasonal variations in ocean dynamics.

The reverse denoising process, conditioned on the historical states \mathbf{X}_{t-2} and \mathbf{X}_{t-1} , learns to recover the clean

residual through:

$$\hat{r}_t^{(i-1)} = D_\theta(\hat{r}_t^{(i)}, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathcal{F}, \tau_{t-1}, \tau_{t-2}, \sigma^{(i)}) \quad (18)$$

where i denotes the diffusion step, $\sigma^{(i)}$ represents the noise level at step i , $\hat{r}_t^{(i)}$ and $\hat{r}_t^{(i-1)}$ denote the results before and after the current denoising step respectively, and D_θ is the denoising network parameterized by θ .

The denoising network processes the multi-scale spatiotemporal features through the SGAN-MOC module, which heterogeneously encodes each variable while incorporating spherical geometry information. Combined with PWTC, the network decomposes ocean dynamics across different scales and enforces diffusion-advection constraints. The final prediction is obtained by adding the denoised residual to the previous state: $\hat{\mathbf{X}}_t = \mathbf{X}_{t-1} + \hat{r}_t^{(0)}$. For long-term forecasting, this process is applied autoregressively, using predicted states as conditions for subsequent time steps.

8.1. Training objective

We train the model to predict denoised residual by minimizing the following mean squared error objective, weighted per depth (vertical) level and by latitude–longitude cell area:

$$\mathbb{E}_{\sigma, \epsilon} \left[\lambda(\sigma) \sum_{v \in V} \sum_{i \in G} w_{v, d(v)} \cdot a_i \cdot \left\| D_\theta(\tilde{r}_t; \mathbf{X}_{t-2}, \mathbf{X}_{t-1}, \mathcal{F}, \tau_{t-2}, \tau_{t-1}, \sigma) - r_t \right\|^2 \right], \quad (19)$$

where

- τ_{t-i} denotes the temporal encoding for the input historical state \mathbf{X}_{t-i} , with the computational logic detailed in Algorithm 1.
- $\lambda(\sigma)$ is the per-noise-level loss weight from [23].
- \tilde{r}_t represents r_t with added noise $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.
- As shown in Fig. 6, $w_{v, d(v)}$ represents depth-dependent weights for variable v at depth $d(v)$ following oceanographic principles.
- a_i denotes the area weight for grid cell i accounting for latitude-dependent grid distortion, with the computational logic detailed in Algorithm 2.

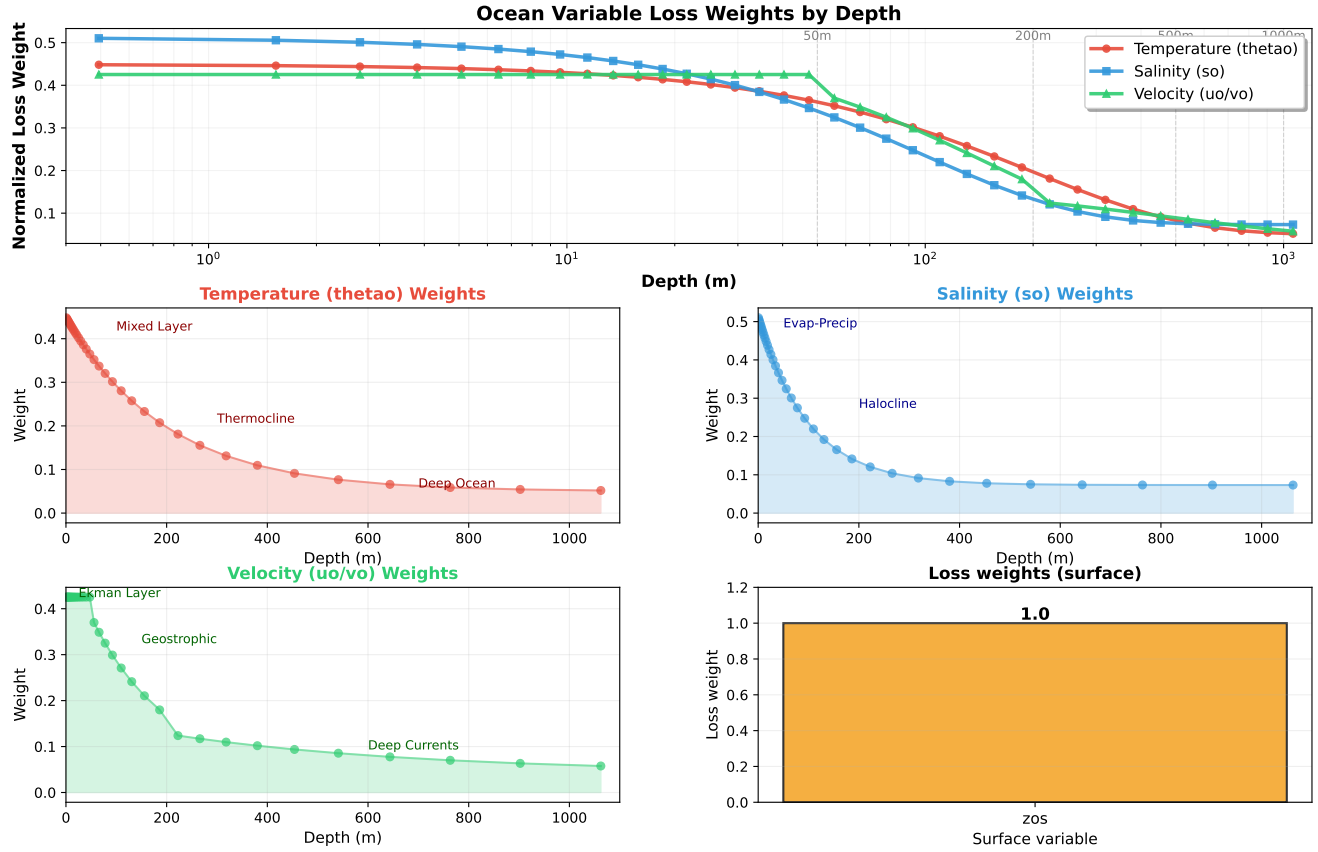
9. Evaluation Metrics

We evaluate PhyOceanCast’s forecasting performance using five complementary metrics that assess different aspects of ocean prediction quality. All metrics are computed separately for each variable $v \in V$ at each depth level $d(v)$, where V denotes the set of output variables.

9.1. Notation and Common Elements

For a particular variable v , depth level $d(v)$, and lead time τ :

Ocean Model Variable Weighting Scheme for Loss Calculation



Weights are designed based on physical oceanography principles: thermocline/halocline dynamics, Ekman layer theory, and ocean energy distribution

Figure 6. Depth-dependent loss weighting system for ocean variables. The weighting scheme assigns differential importance to sea water potential temperature (thetao), salinity (so), and velocity (uo/vo) across 36 depth levels based on oceanographic principles. Temperature weights decay exponentially with depth reflecting reduced variability in deep waters. Salinity weights emphasize near-surface layers where evaporation-precipitation drives strong gradients. Velocity weights distinguish between surface Ekman dynamics, geostrophic flows, and deep currents. Sea surface height (zos) maintains unit weight following established practices. This physics-informed weighting ensures the loss function prioritizes dynamically active regions while accounting for observational density and physical variability at each depth.

Algorithm 1 Time Encoding: PyTorch-style Pseudocode

```
# Constants
AVG_DAY_PER_YEAR = 365.24219
# d: day of year
# pos: temporal position encoding
def time_position_encoding(d):
    theta = 2 * pi * d / AVG_DAY_PER_YEAR
    sin_enc = sin(theta)
    cos_enc = cos(theta)
    pos = [sin_enc, cos_enc]
    return pos
```

- $\hat{\mathbf{X}}_t^{v,d}$ denotes the predicted ocean state at time t for variable v at depth d
- $\mathbf{X}_t^{v,d}$ denotes the corresponding ground truth verifica-

tion target

- $M^{v,d} \in \{0,1\}^{H \times W}$ denotes the ocean-land binary mask, where $M_{i,j}^{v,d} = 1$ indicates valid ocean grid cells and $M_{i,j}^{v,d} = 0$ indicates land or missing data
- $i \in G$ indexes the latitude-longitude grid cells, where G represents the global grid
- a_i denotes the area weight of grid cell i , accounting for latitude-dependent grid distortion (computed via Algorithm 2)

Ensemble Forecasting. For probabilistic ensemble predictions with M members, we denote:

- $\hat{\mathbf{X}}_t^{v,d,m}$ as the m -th ensemble member prediction, $m \in \{1, \dots, M\}$
- $\hat{\mathbf{X}}_t^{v,d} = \frac{1}{M} \sum_{m=1}^M \hat{\mathbf{X}}_t^{v,d,m}$ as the ensemble mean pre-

Algorithm 2 Normalized Latitude Weighting: PyTorch-style Pseudocode

```
# Input: lat - latitude vector
# Output: weights - normalized area weights
def normalized_latitude_weights(lat):
    # Check if poles (±90°) are included
    if any(abs(lat) ≈ 90):
        weights = weight_with_poles(lat)
    else:
        weights = weight_without_poles(lat)
    return weights / weights.mean()
def weight_without_poles(lat):
    # Latitudes: [90-Δ/2, ..., -90+Δ/2]
    delta = check_uniform_spacing(lat)
    # Weight ∝ cos(lat)
    return cos(deg2rad(lat))
def weight_with_poles(lat):
    # Latitudes: [90, ..., -90]
    delta = check_uniform_spacing(lat)
    # Non-pole weights
    w = cos(deg2rad(lat)) *
    sin(deg2rad(delta/2))
    # Special weights for pole points
    w[0], w[-1] = sin(deg2rad(delta/4)) ^ 2
    return w
def check_uniform_spacing(vec):
    return vec[1] - vec[0]
```

diction

RMSE, MAE, ACC_{lat} are computed in the original physical space after denormalizing the model outputs using the Min-Max scaling parameters:

$$\hat{\mathbf{X}}_{t,\text{physical}}^{v,d} = \hat{\mathbf{X}}_{t,\text{normalized}}^{v,d} \cdot (\max_v - \min_v) + \min_v. \quad (20)$$

9.2. RMSE

RMSE measures the average magnitude of prediction errors in physical units, computed as:

$$\text{RMSE}^{v,d} = \sqrt{\frac{1}{N_{\text{valid}}} \sum_{i \in G} M_i^{v,d} \cdot (\bar{\mathbf{X}}_{t,i}^{v,d} - \mathbf{X}_{t,i}^{v,d})^2}, \quad (21)$$

where $N_{\text{valid}} = \sum_{i \in G} M_i^{v,d}$ is the total number of valid grid cells. The mask $M^{v,d}$ ensures errors are only computed over ocean regions.

Physical Significance. RMSE penalizes large errors quadratically, making it sensitive to outliers and extreme events (e.g., marine heatwaves, eddy-induced anomalies). For ocean temperature, RMSE is reported in Celsius; for salinity in PSU (Practical Salinity Units); for currents in m/s; and for sea surface height in meters.

9.3. MAE

MAE provides a robust alternative to RMSE by using absolute errors:

$$\text{MAE}^{v,d} = \frac{1}{N_{\text{valid}}} \sum_{i \in G} M_i^{v,d} \cdot |\bar{\mathbf{X}}_{t,i}^{v,d} - \mathbf{X}_{t,i}^{v,d}|, \quad (22)$$

where $N_{\text{valid}} = \sum_{i \in G} M_i^{v,d}$ is the total number of valid grid cells.

Physical Significance. Unlike RMSE, MAE treats all errors equally and is less sensitive to outliers. This is particularly valuable for ocean forecasting where reanalysis data may contain spurious extremes due to sparse observational coverage in deep ocean or polar regions. MAE provides interpretable error magnitudes that directly reflect typical prediction deviations.

9.4. ACC_{lat}

ACC_{lat} measures the pattern correlation between predicted and observed anomaly fields, with latitude weighting to account for spherical geometry:

$$\text{ACC}_{\text{lat}}^{v,d} = \frac{\sum_{i \in G} M_i^{v,d} \cdot w_{\text{lat}}(i) \cdot \Delta \bar{\mathbf{X}}_{t,i}^{v,d} \cdot \Delta \mathbf{X}_{t,i}^{v,d}}{\sqrt{\sum_{i \in G} M_i^{v,d} \cdot w_{\text{lat}}(i) \cdot (\Delta \bar{\mathbf{X}}_{t,i}^{v,d})^2}} \times \frac{1}{\sqrt{\sum_{i \in G} M_i^{v,d} \cdot w_{\text{lat}}(i) \cdot (\Delta \mathbf{X}_{t,i}^{v,d})^2}}, \quad (23)$$

where the anomaly fields are computed by removing the spatial mean:

$$\Delta \bar{\mathbf{X}}_t^{v,d} = \bar{\mathbf{X}}_t^{v,d} - \frac{1}{N_{\text{valid}}} \sum_{i \in G} M_i^{v,d} \cdot w_{\text{lat}}(i) \cdot \bar{\mathbf{X}}_{t,i}^{v,d}, \quad (24)$$

$$\Delta \mathbf{X}_t^{v,d} = \mathbf{X}_t^{v,d} - \frac{1}{N_{\text{valid}}} \sum_{i \in G} M_i^{v,d} \cdot w_{\text{lat}}(i) \cdot \mathbf{X}_{t,i}^{v,d}. \quad (25)$$

Algorithm 2 details the calculation logic of latitude weights.

Physical Significance. It emphasizes spatial pattern fidelity rather than absolute magnitude, making it crucial for capturing ocean circulation structures. The latitude weighting prevents high latitude grid cells from dominating the metric due to meridional convergence.

9.5. Pearson Correlation Coefficient

Pearson correlation measures linear relationship strength between predictions and observations:

$$\text{Pearson}^{v,d} = \frac{\text{Cov}(\bar{\mathbf{X}}_t^{v,d}, \mathbf{X}_t^{v,d})}{\sigma_{\bar{\mathbf{X}}_t^{v,d}} \cdot \sigma_{\mathbf{X}_t^{v,d}}}, \quad (26)$$

where covariance and standard deviations are computed only over valid ocean grid cells defined by $M^{v,d}$:

$$\text{Cov}(\bar{\mathbf{X}}_t^{v,d}, \mathbf{X}_t^{v,d}) = \frac{1}{N_{\text{valid}}} \sum_{i \in G} M_i^{v,d} \times (\bar{\mathbf{X}}_{t,i}^{v,d} - \mu_{\bar{\mathbf{X}}}) \cdot (\mathbf{X}_{t,i}^{v,d} - \mu_{\mathbf{X}}), \quad (27)$$

$$\sigma_{\bar{\mathbf{X}}_t^{v,d}} = \sqrt{\frac{1}{N_{\text{valid}}} \sum_{i \in G} M_i^{v,d} \cdot (\bar{\mathbf{X}}_{t,i}^{v,d} - \mu_{\bar{\mathbf{X}}})^2}. \quad (28)$$

Physical Significance. Unlike ACC_{lat} , Pearson correlation does not apply latitude weighting and uses normalized data space, providing a complementary view of pattern similarity. Values near 1 indicate strong linear predictability, which is expected for large-scale ocean features like basin-wide temperature gradients and major current systems. The metric is computed using `scipy.stats.pearsonr`.

9.6. SSIM

SSIM assesses perceptual similarity by comparing local patterns of luminance, contrast, and structure:

$$\text{SSIM}^{v,d} = \frac{(2\mu_{\bar{\mathbf{x}}}\mu_{\mathbf{x}} + C_1)(2\sigma_{\bar{\mathbf{x}}\mathbf{x}} + C_2)}{(\mu_{\bar{\mathbf{x}}}^2 + \mu_{\mathbf{x}}^2 + C_1)(\sigma_{\bar{\mathbf{x}}}^2 + \sigma_{\mathbf{x}}^2 + C_2)}, \quad (29)$$

where μ and σ denote local means and standard deviations computed over 11×11 Gaussian windows with $\sigma = 1.5$, and C_1, C_2 are small constants for numerical stability. The implementation uses `skimage.metrics.structural_similarity` with:

- `data_range=1.0`
- `win_size=11`
- `gaussian_weights=True`
- `use_sample_covariance=False`

Physical Significance. For ocean forecasting, SSIM is particularly valuable for assessing mesoscale eddy coherence by evaluating whether predicted eddies maintain realistic size and shape, frontal structure preservation through the detection of sharp gradients in temperature and salinity at ocean fronts, and spatial texture fidelity by examining fine-scale variability patterns in currents and mixed layer depth. Unlike pixel-wise metrics (RMSE, MAE), SSIM tolerates small spatial shifts, which is crucial since ocean features may be predicted with correct structure but slight phase errors. The metric operates on normalized data to ensure consistent comparison across variables with different physical ranges.

10. Implementation Details

To assess the effectiveness of each method, we initially identified the best hyperparameters with a single seed applied to the validation set. Subsequently, each method was trained with the selected hyperparameters across five different random seeds. We train each model justly on 6 NVIDIA H800 GPUs with a total batch size of 6 for 1K epochs. We employ the AdamW optimizer [30] with an initial learning rate of $1e^{-3}$ and weight decay of $1e^{-2}$. The learning rate uses warmup followed by square root decay. We apply dropout with probability 0.13 and maintain an EMA for stable inference. All experiments are implemented in PyTorch with DDP training, which takes 700–1500 total GPU-hours for the global models.

Table 4. Ensemble forecasting hyperparameter and ablation analysis at 7 day lead time. Best results in **bold**.

Method	RMSE ↓	SSIM ↑
<i>Ensemble Size Analysis</i>		
Single member	0.9381	0.9430
3 members	0.7829	0.9745
10 members	0.7203	0.9767
<i>Ablation Study (3 members)</i>		
Full model	0.7829	0.9745
w/o SGAN-MOC & PWTC	0.8883	0.9521
w/o SGAN-MOC	0.7932	0.9600
w/o PWTC	0.7832	0.9668

11. Hyperparameters and Ablation Studies

Extensive ablation studies validate the effectiveness and complementary nature of new components in PhyOceanCast. Furthermore, we performed a hyperparameter study across different ensemble sizes to assess the effectiveness of the probabilistic ensemble forecasting approach. Tab. 4 shows the quantitative comparison results.

Effect of ensemble forecasts Changes in the ocean state are driven by many factors, including temperature, salinity, and gravitational forcing, while limited initial conditions do not provide sufficient information for a model to fully capture stochastic natural processes. Consequently, ensemble forecasting with probabilistic models employs multiple, independently sampled members to broaden coverage of plausible state evolutions, which explains why a 3 members ensemble achieves significantly better performance than a single-member forecast.

Effect of SGAN-MOC Spherical domain processing yields significant performance gains, and when coupled with heterogeneous variable encoding—enables the model to effectively capture cross-variable feature interactions in multivariate forecasting of ocean state variables.

Effect of PWTC By imposing multiscale decomposition and physics based constraints on ocean state variables, PWTC substantially improves the model’s forecasting accuracy and the physical consistency of its outputs, and it effectively characterizes the dynamical variability of ocean circulation.

12. Comparison of probabilistic metrics & inference efficiency

As shown in Tab. 5, for the lead time 10 day forecasting task, we evaluated probabilistic ensemble forecasting metrics for diffusion models and compared the single-step inference time and peak memory usage against existing methods. Compared to DiffCast, PhyOceanCast achieved the best CRPS ($\downarrow 11.8\%$) and SSR ($\uparrow 4.6\%$) under 10 members. Furthermore,

benefiting from the adoption of ADM as backbone, our inference time is reduced by 5% compared to DiffCast utilizing the DDIM strategy. Although PhyOceanCast incurs a higher memory cost during inference, with the iteration of forecast steps (15–30 days), the error growth is reduced by 15.8% (Tab. 1 in paper, compare to the second best model). Moreover, following the forecasting paradigm of GenCast [34], we can implement parallel ensemble operational forecasting.

Table 5. Inference efficiency and probabilistic metrics comparison.

Method	Memory Cost (GB)	CRPS ↓	SSR (Ideal=1)	Inference Time (s/step) ↓
SimVP (CVPR’22)	2.21	-	-	0.6
GraphCast (Science’23)	4.78	-	-	3.2
DiffCast (CVPR’25)	3.56	0.2874	0.8706	142.5
PhyOceanCast (Ours)	9.89	0.2533	0.9111	134.7

13. Limitations and Discussion

13.1. Limitations

Discontinuous Ocean Domain Constrains Model Performance. Unlike atmospheric forecasting over continuous global grids, ocean prediction faces inherent spatial discontinuities due to complex coastline geometries and irregular bathymetry. The land-ocean mask $M^{v,d}$ creates fragmented prediction domains that prevent seamless information propagation across continental boundaries. For instance, the Atlantic and Pacific basins remain largely isolated in our graph construction despite physical connections through the Southern Ocean and Arctic passages. This geographic fragmentation particularly degrades forecasting skill in marginal seas, where limited spatial extent reduces the effective receptive field of graph attention mechanisms. Future work should explore ocean-aware graph partitioning strategies that respect basin connectivity and incorporate explicit inter-basin teleconnection patterns (e.g., ENSO-IOD linkages) to mitigate these geometric constraints.

Slow-Scale Ocean Dynamics Prolong Training Convergence. Ocean circulation exhibits multi-decadal variability driven by thermohaline processes, with characteristic timescales far exceeding our training sequences. This temporal scale mismatch creates two challenges: (1) The model cannot fully learn low-frequency modes such as Atlantic Meridional Overturning Circulation variations and Pacific Decadal Oscillation, limiting skill for climate-scale predictions. (2) Training convergence is slower compared to atmospheric models, as ocean diffusion processes require longer temporal context to establish stable gradient signals. But fundamental improvements necessitate: (a) incorporating centennial scale ocean reanalysis products when available, (b) developing scale adaptive training logic that progressively emphasize slower dynamics, and (c) investigating physics-informed pretraining on idealized ocean simulations with known long term solutions.

Clarification on density-driven mechanism coupling. Our model does not explicitly model the coupling between velocity and temperature, salinity. However, to capture vertical water mixing, as illustrated in Fig. 4a, we employ a depth-coupled attention mechanism to capture cross-layer interactions.

13.2. Discussion

Why Does Velocity Forecasting Show Smaller Improvements Compared to Thermohaline Variables? As shown in Fig. 3 (c-d), PhyOceanCast’s performance gains for u_o/v_o are less pronounced than for sea water potential temperature and sea water salinity, despite substantial RMSE reductions. This phenomenon stems from two factors:

(1) Scale of Daily Variability: Ocean currents exhibit smaller day-to-day changes compared to temperature fluctuations or salinity variations. In normalized space, velocity gradients provide weaker training signals, making it harder for diffusion models to distinguish subtle flow modifications from noise. The RMSE improvements are therefore compressed within a narrower dynamic range.

(2) Observational Sparsity: GLORYS12 assimilates abundant satellite SST and altimetry data, providing dense constraints on temperature and sea surface height. In contrast, subsurface velocity observations rely primarily on sparse Argo float trajectories and moored current meters, leading to higher uncertainty in the reanalysis target itself. This introduces a practical ceiling on achievable velocity forecast skill.

14. Extended Experimental Results

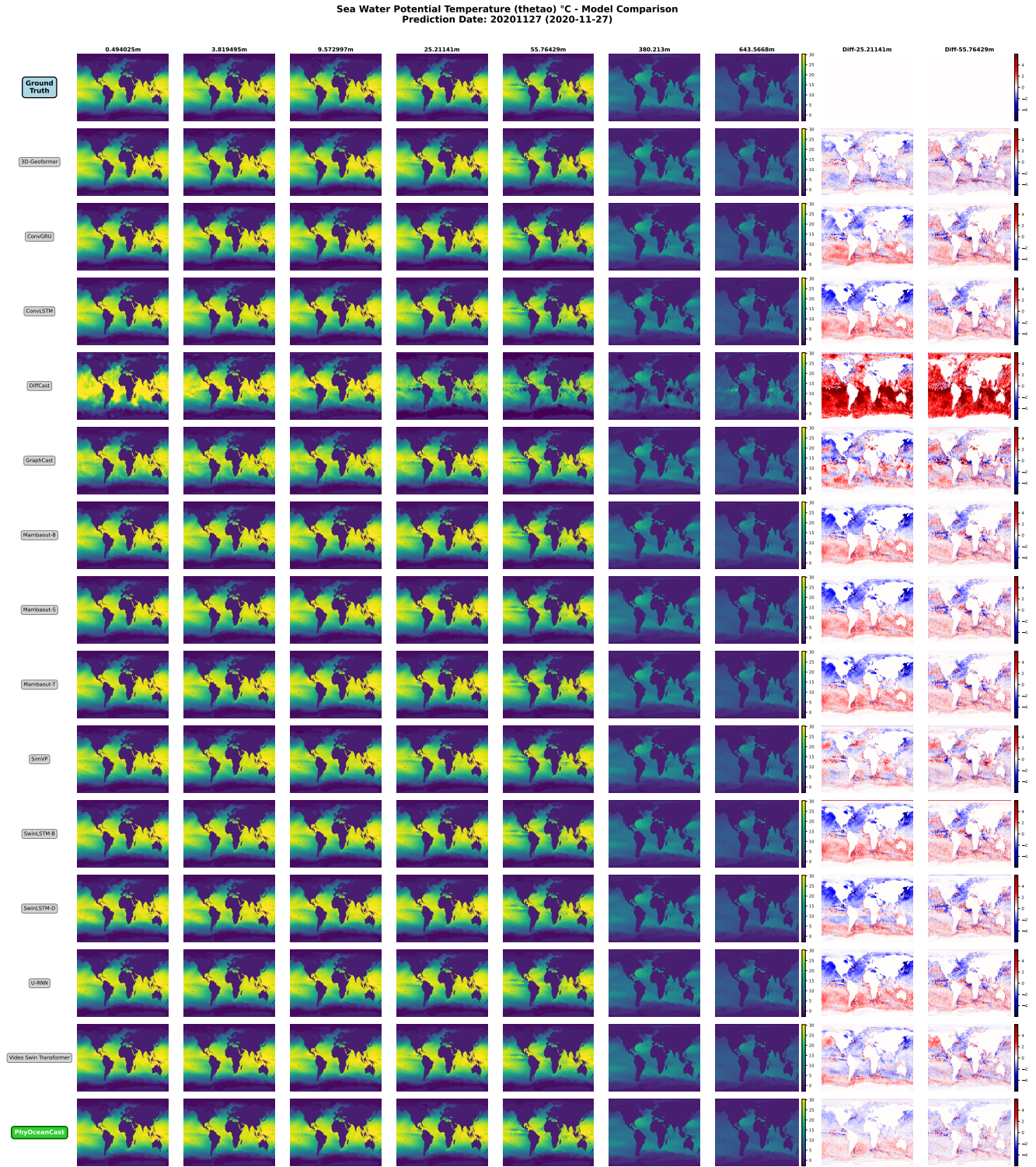


Figure 7. Comparison of qualitative results of different models at lead time 30 days.

Sea Water Potential Temperature (thetao) °C - Ensemble Forecasting Single member

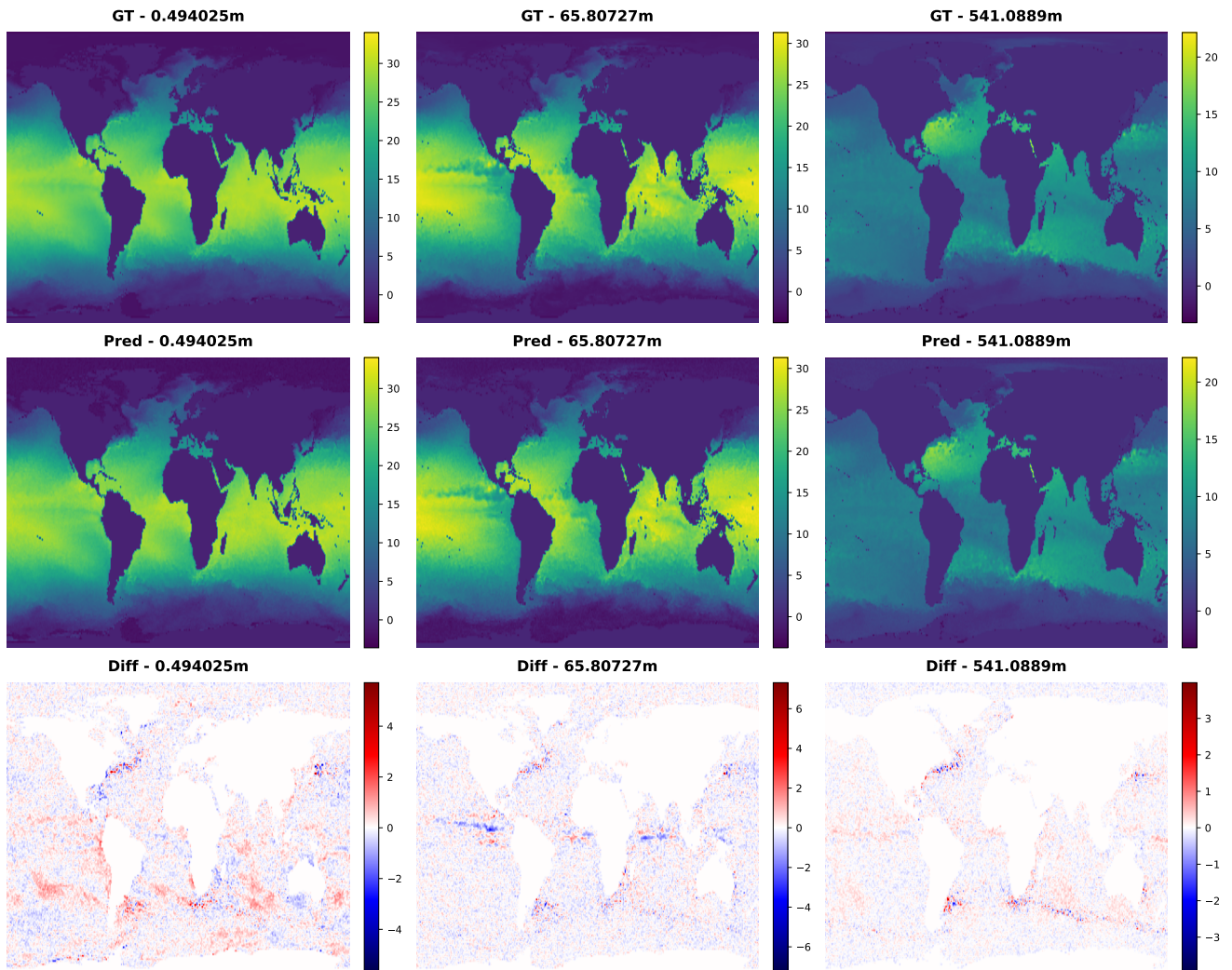


Figure 8. Example of a single member forecast for thetiao.

Sea Water Salinity (so) PSU - Ensemble Forecasting Single member

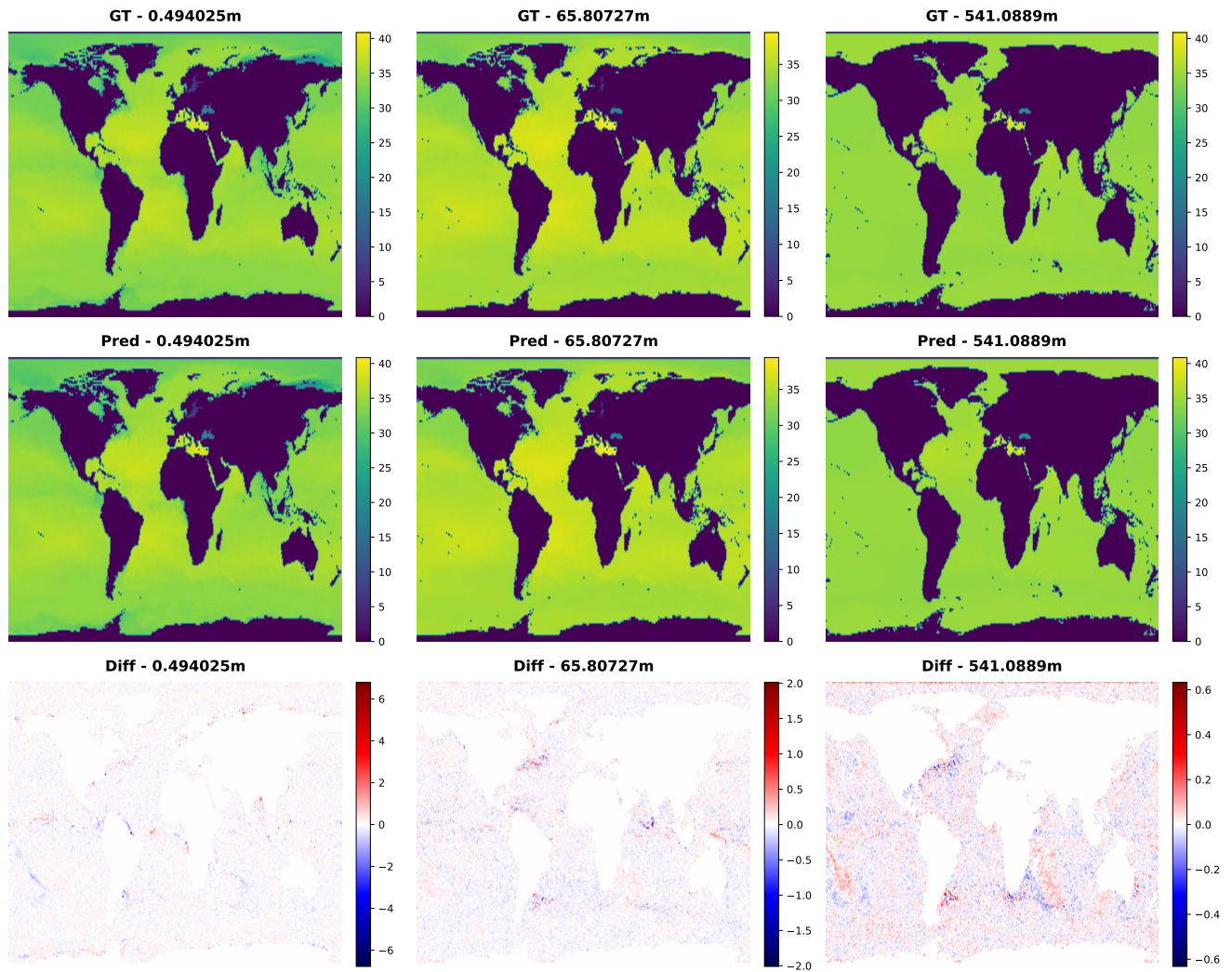


Figure 9. Example of a single member forecast for so.

Eastward Sea Water Velocity (uo) m/s - Ensemble Forecasting Single member

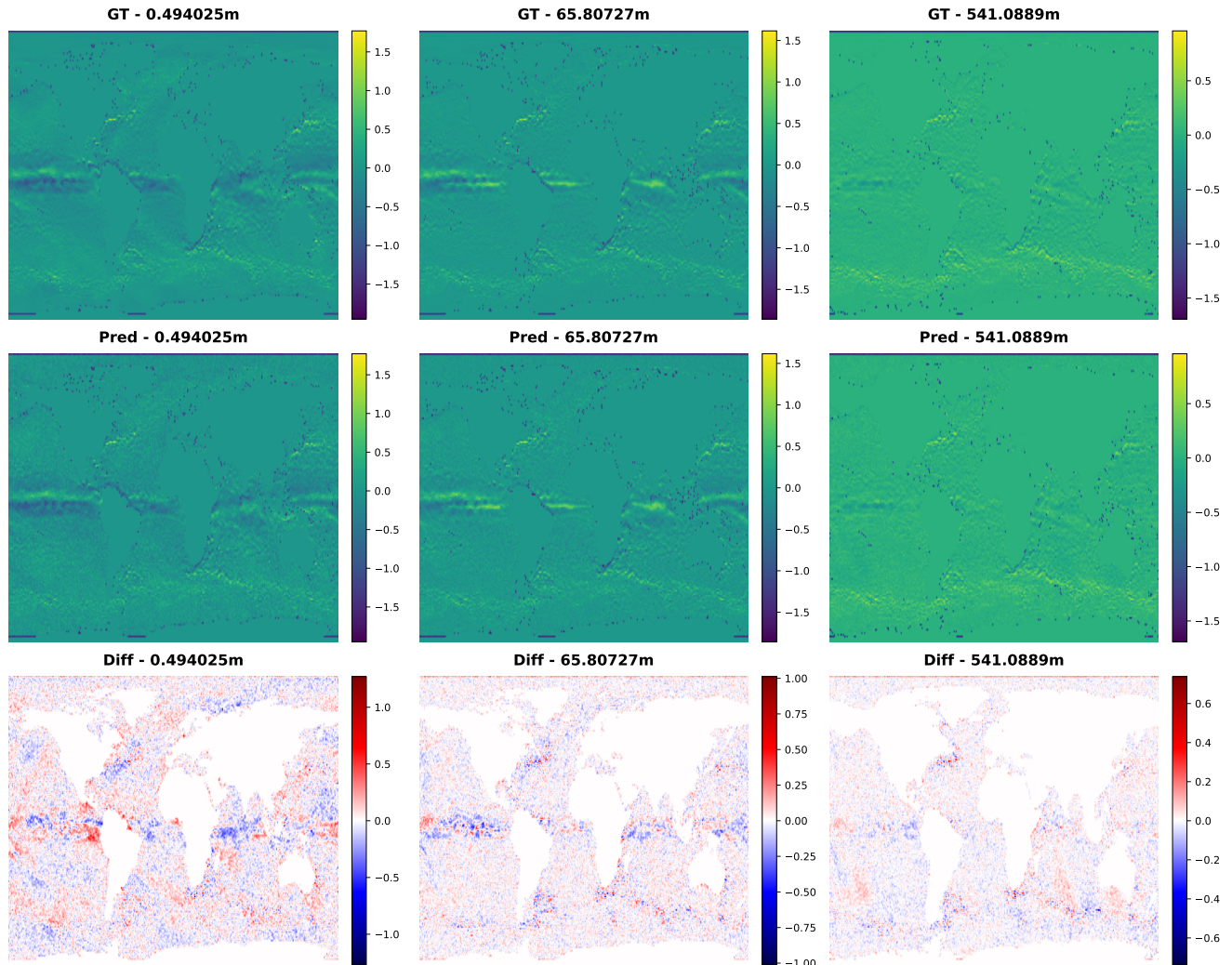


Figure 10. Example of a single member forecast for uo.

Northward Sea Water Velocity (vo) m/s - Ensemble Forecasting Single member

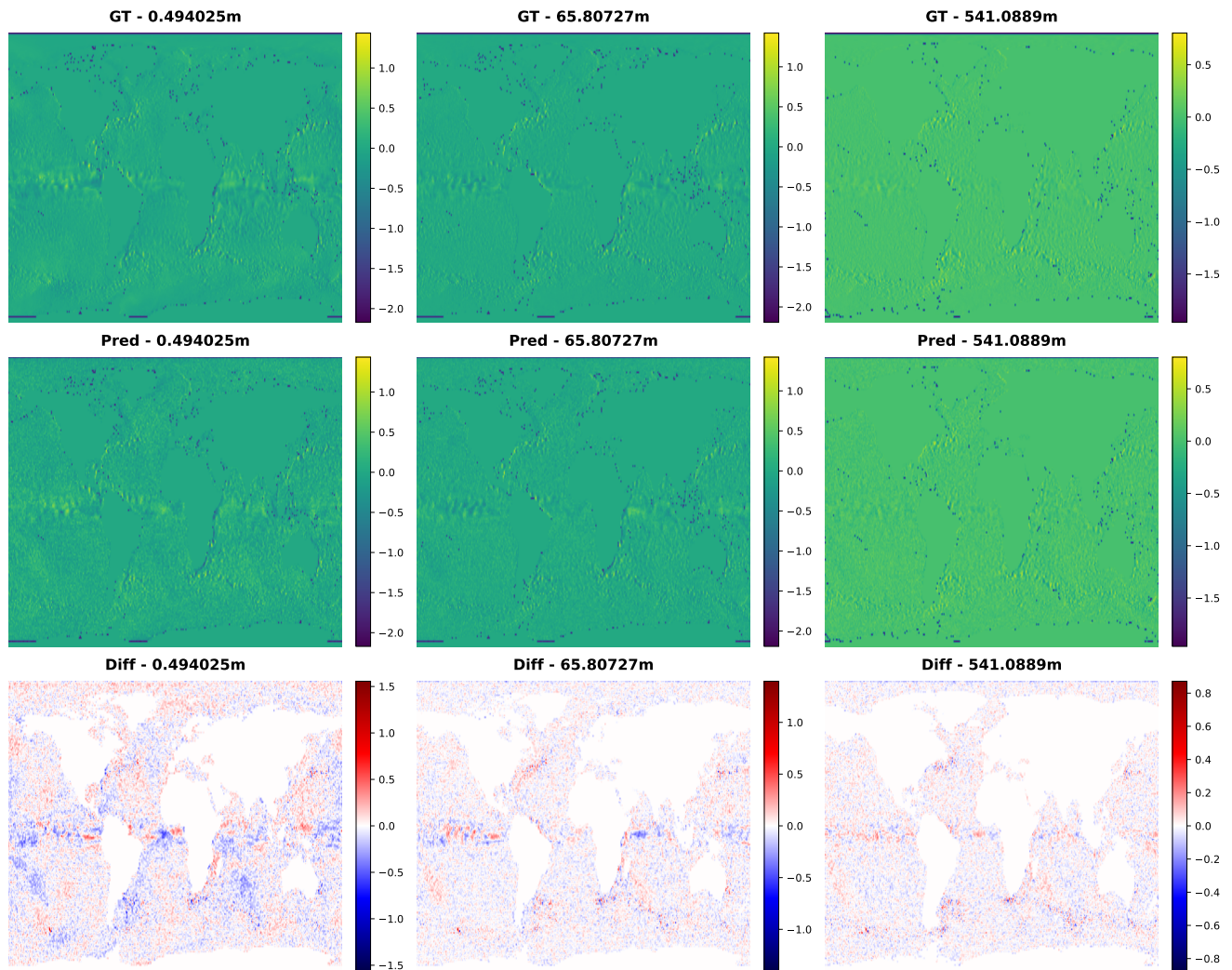


Figure 11. Example of a single member forecast for vo.

Sea Surface Height (zos) m - Ensemble Forecasting Single member

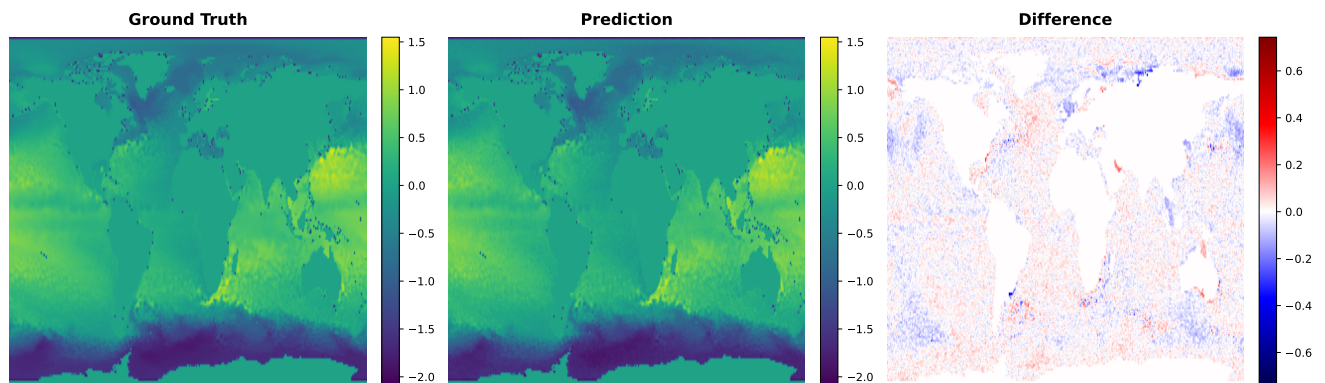


Figure 12. Example of a single member forecast for zos.

Sea Water Potential Temperature (thetao) °C - Ensemble Forecasting 3 members

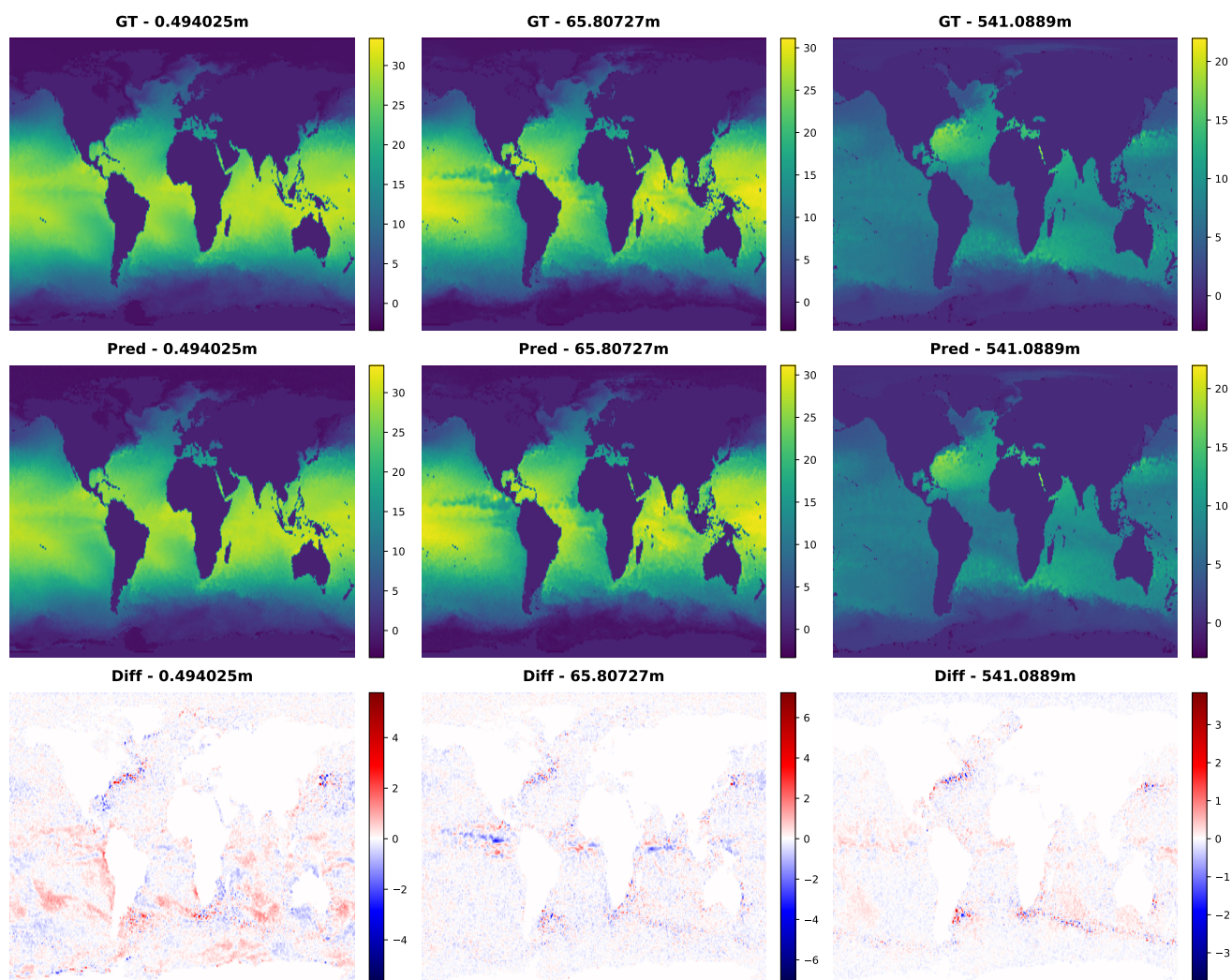


Figure 13. Example of 3 members forecast for thetiao.

Sea Water Salinity (so) PSU - Ensemble Forecasting 3 members

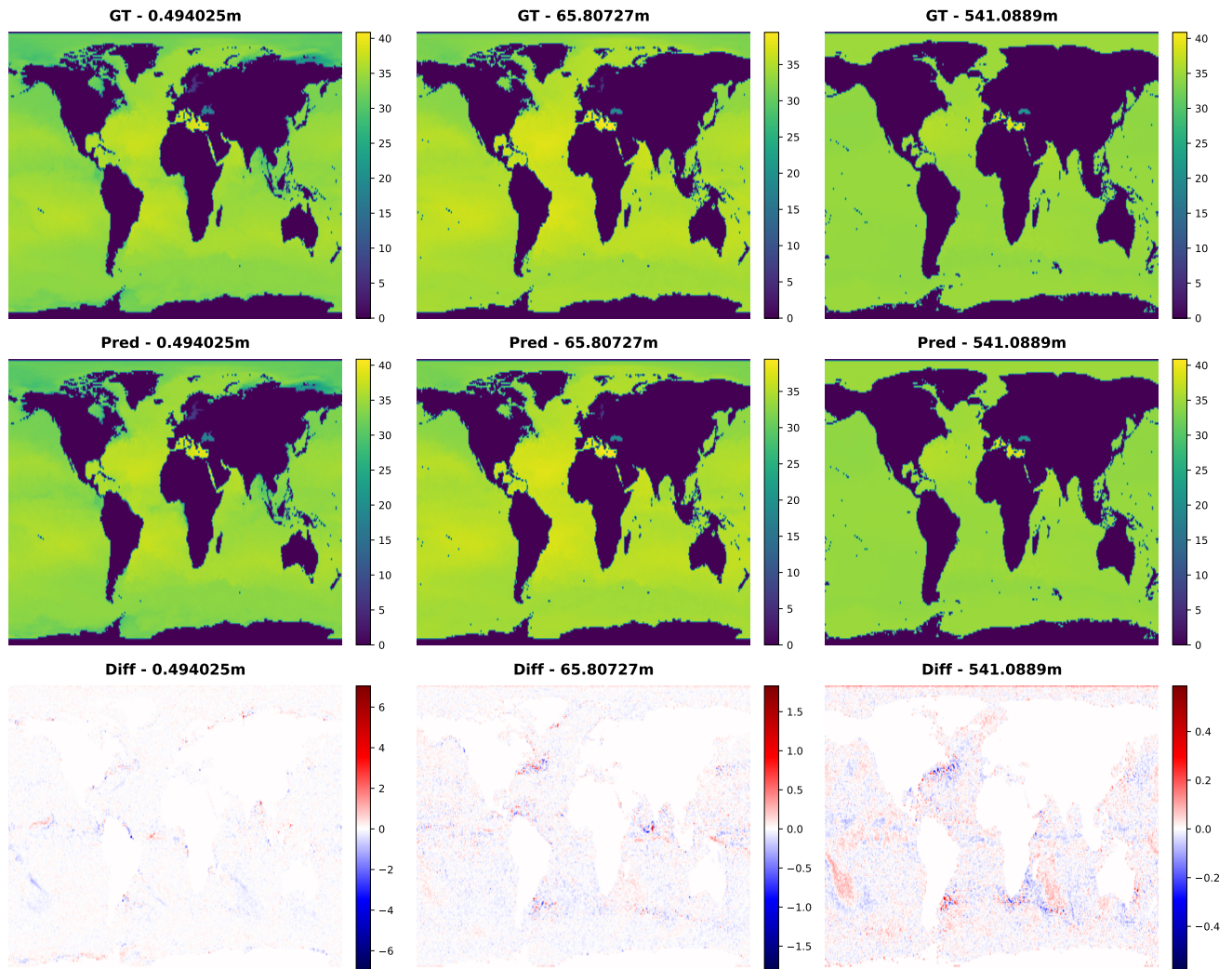


Figure 14. Example of 3 members forecast for so.

Eastward Sea Water Velocity (uo) m/s - Ensemble Forecasting 3 members

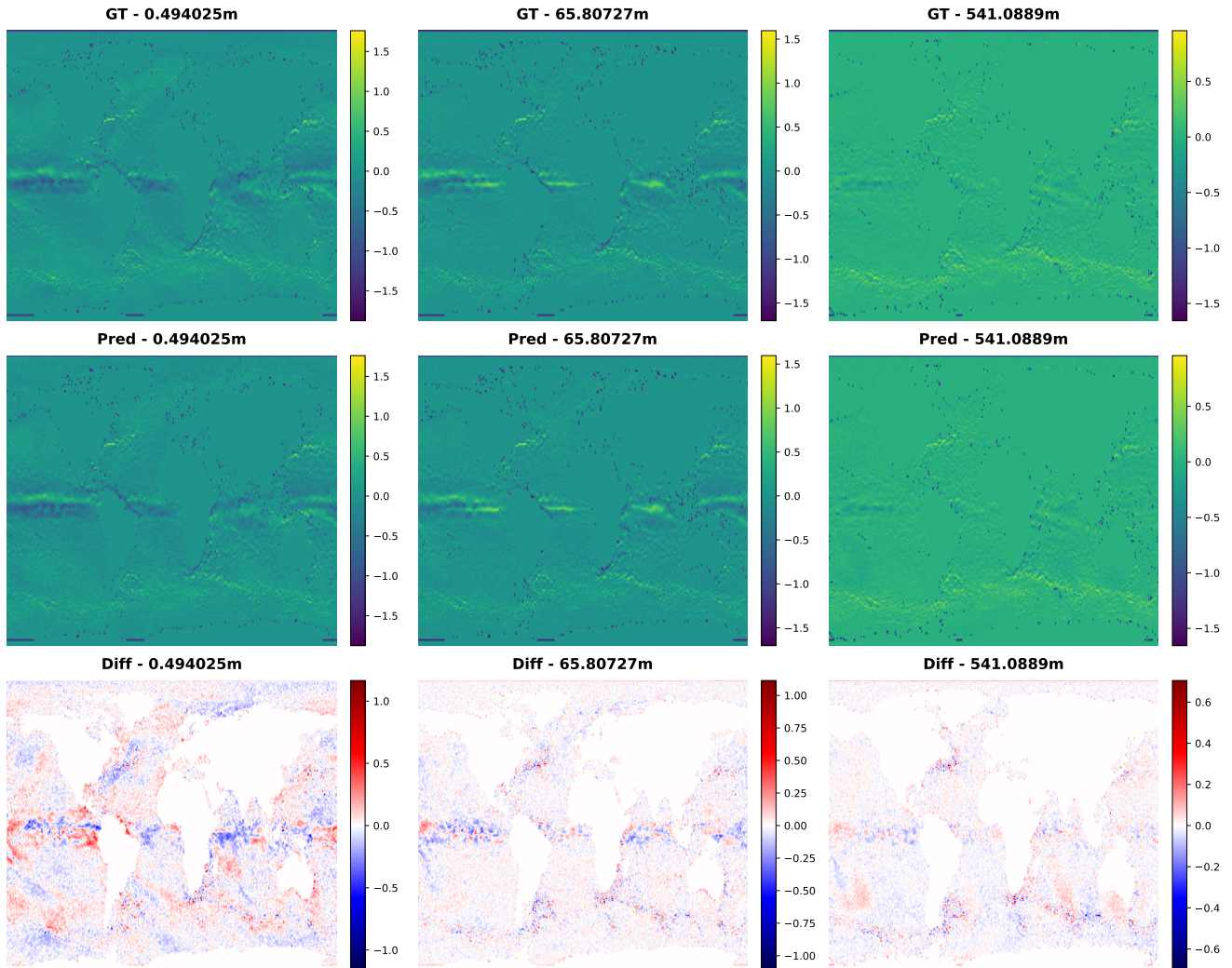


Figure 15. Example of 3 members forecast for uo.

Northward Sea Water Velocity (vo) m/s - Ensemble Forecasting 3 members

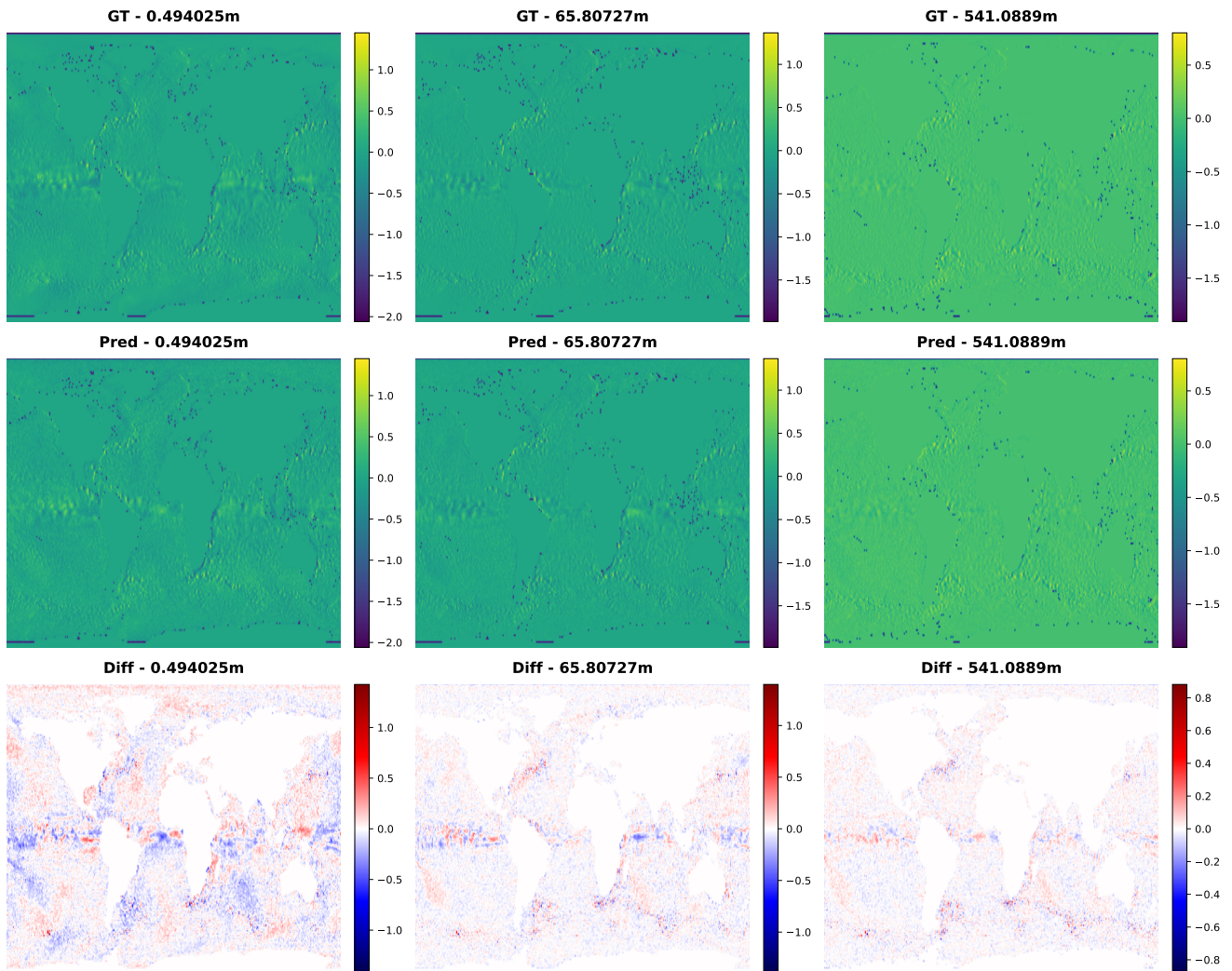


Figure 16. Example of 3 members forecast for vo.

Sea Surface Height (zos) m - Ensemble Forecasting 3 members

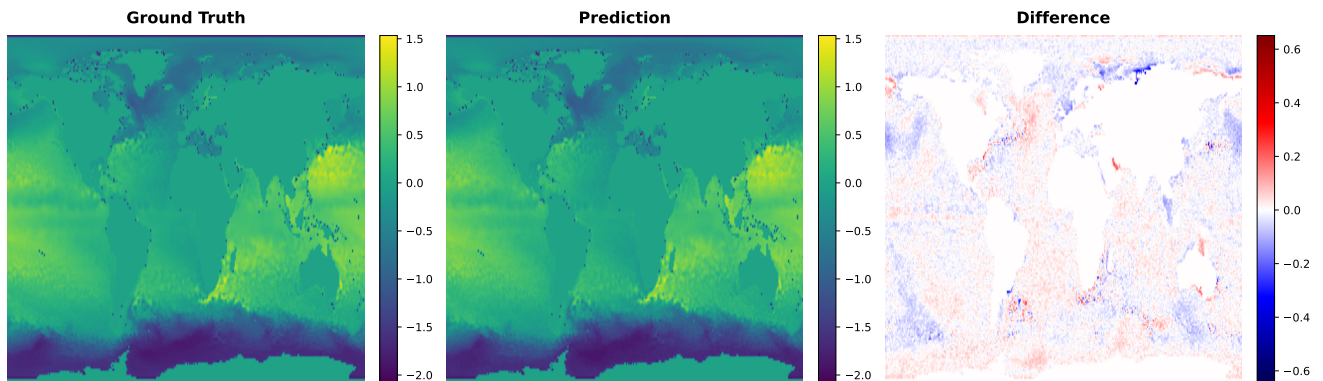


Figure 17. Example of 3 members forecast for zos.

Sea Water Potential Temperature (thetao) °C - Ensemble Forecasting 10 members

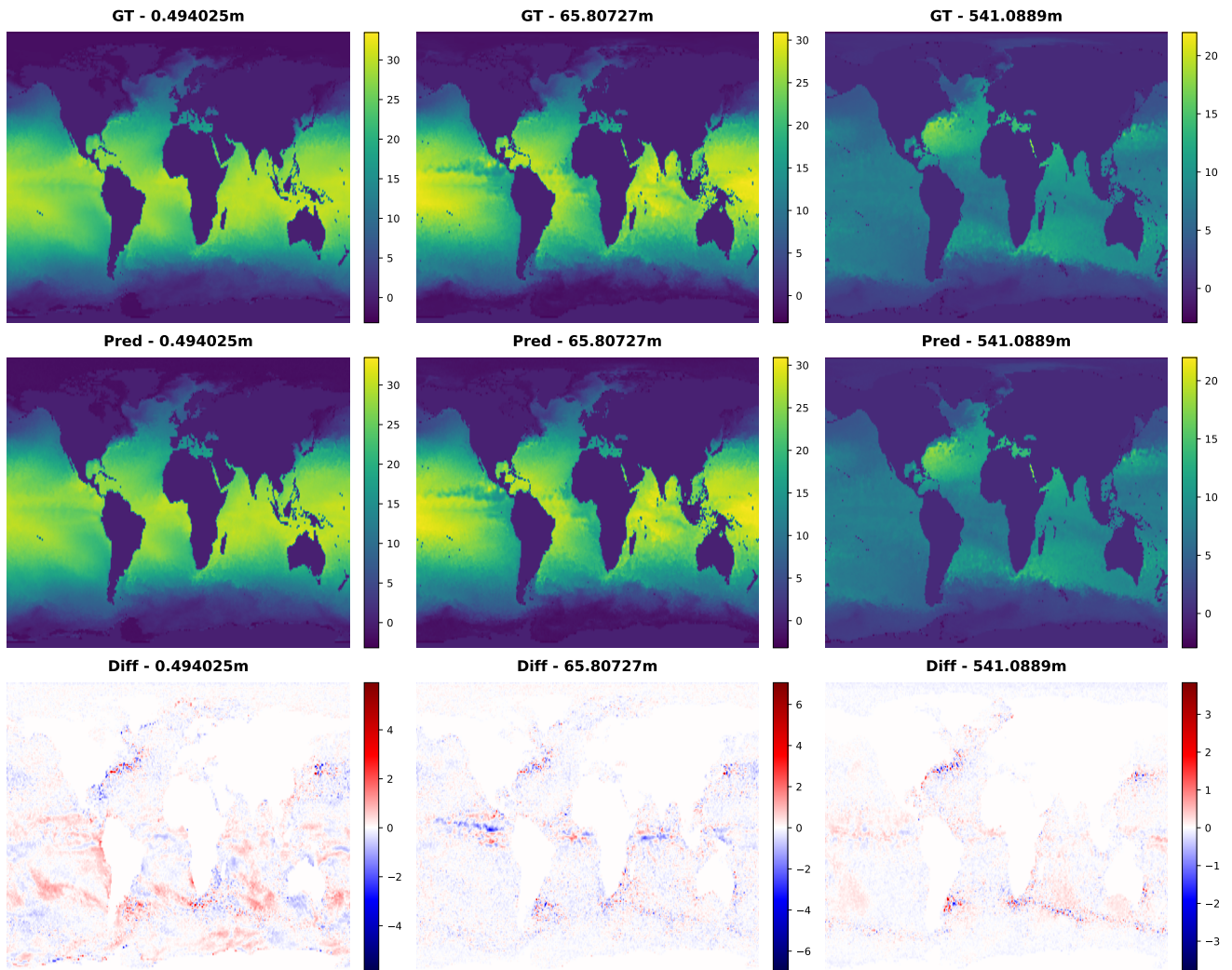


Figure 18. Example of 10 members forecast for thetiao.

Sea Water Salinity (so) PSU - Ensemble Forecasting 10 members

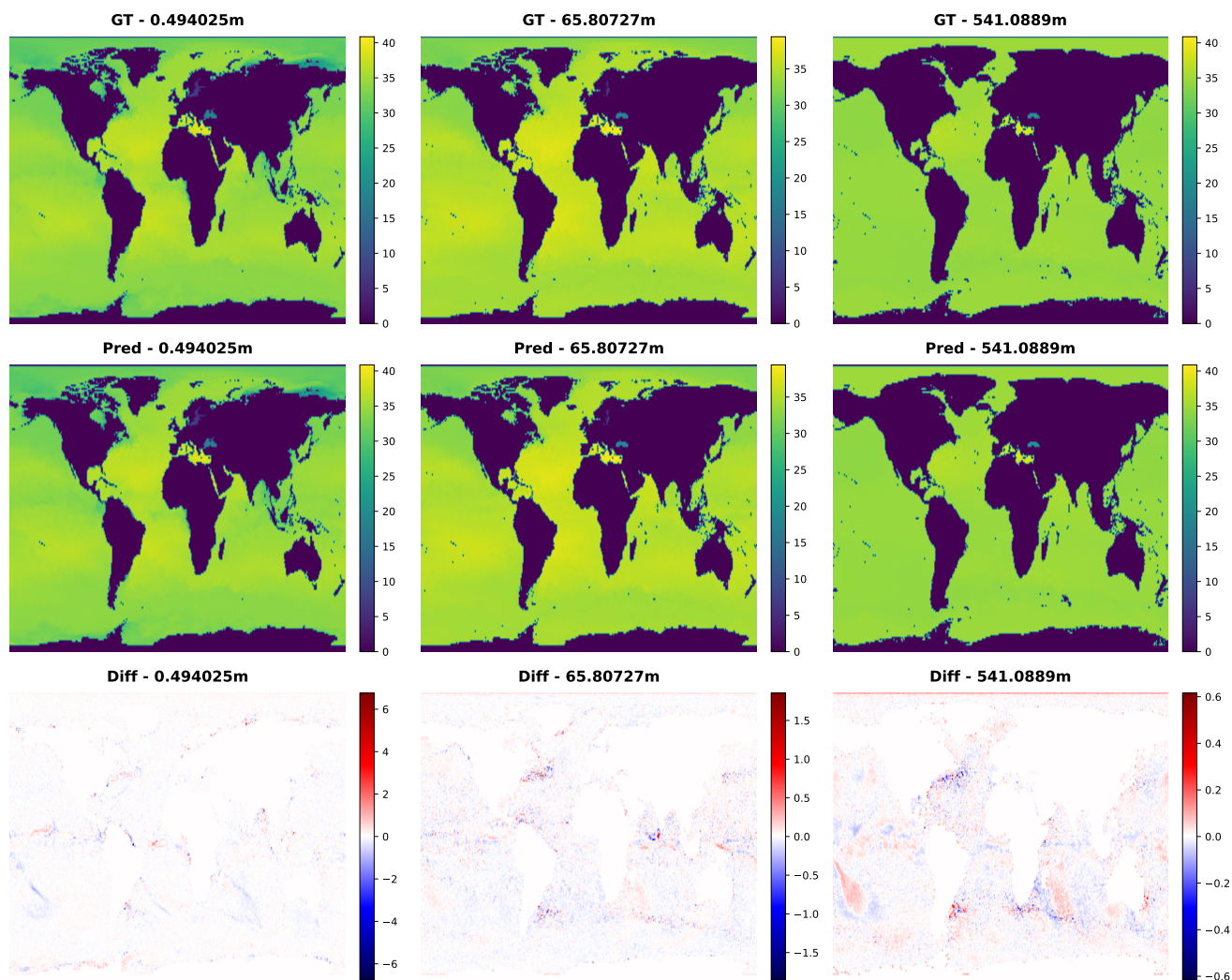


Figure 19. Example of 10 members forecast for so.

Eastward Sea Water Velocity (uo) m/s - Ensemble Forecasting 10 members

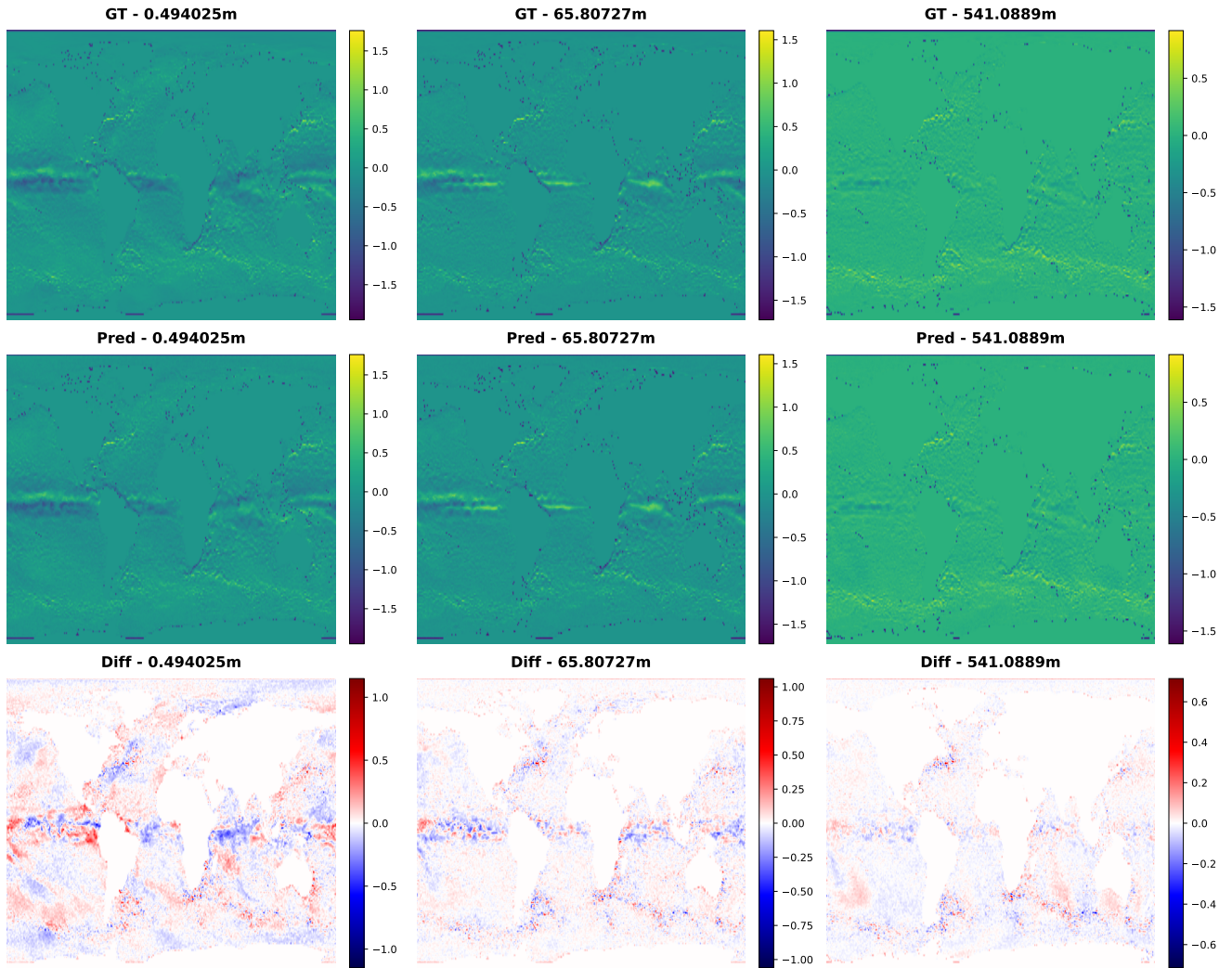


Figure 20. Example of 10 members forecast for uo.

Northward Sea Water Velocity (vo) m/s - Ensemble Forecasting 10 members

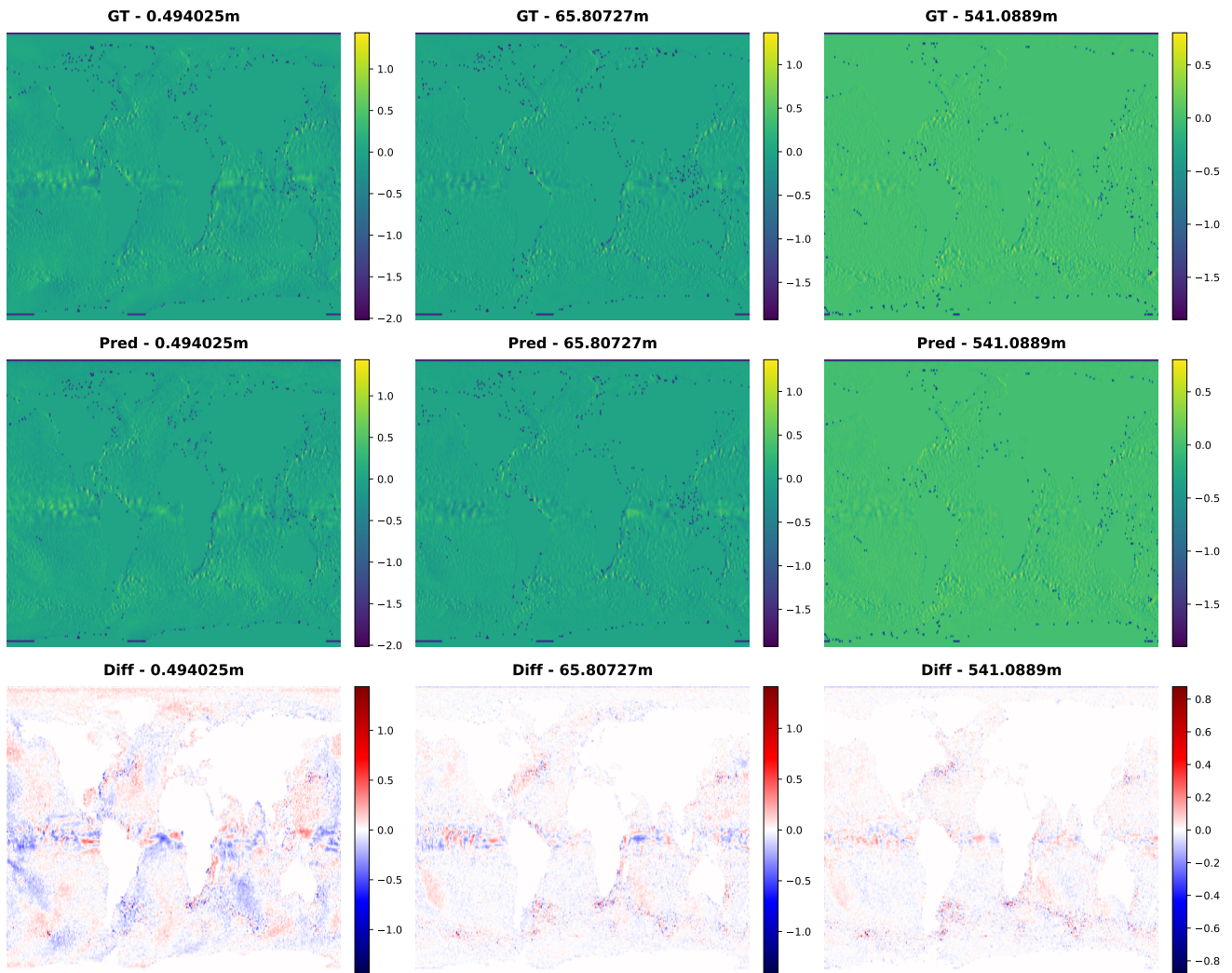


Figure 21. Example of 10 members forecast for vo.

Sea Surface Height (zos) m - Ensemble Forecasting 10 members

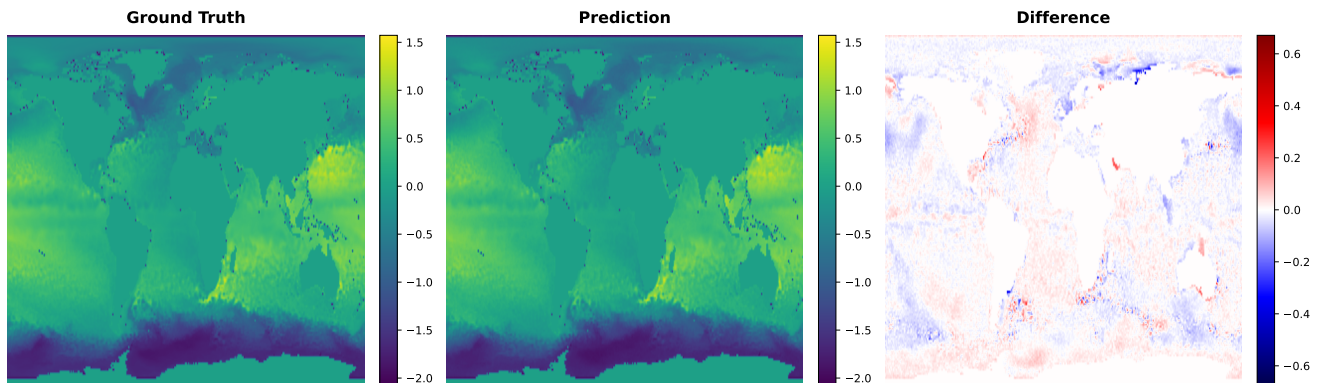


Figure 22. Example of 10 members forecast for zos.