Table 1. Details of the Feature Data.

Variable	Description	Spatial Resolution	Temporal Resolution	Temporal Coverage	Data Source
Temperature (°C)	Seawater temperature	0.5° × 0.5°; 187 standard depth levels (surface–5500 m)	Monthly	Jan 1960–Jun 2024	Szekely et al. (2025)
Salinity	Seawater salinity	0.5° × 0.5°; 187 standard depth levels (surface–5500 m)	Monthly	Jan 1960–Jun 2024	
U ($m\,s^{-1}$)	U-wind vector component at 10 m	0.25° × 0.25°	Monthly	Jan 1993–~~Aug 2023~~ TS1	Mears et al. (2022)
V ($m\,s^{-1}$)	V-wind vector component at 10 m				
MLD (m)	Ocean mixed layer depth	0.25° × 0.25°	Monthly	Jan 1993–Dec 2022	Guinehut et al. (2012)
DIC ($\mu mol\,kg^{-1}$)	Surface ocean dissolved inorganic carbon	0.25° × 0.25°	Monthly	Jan 1985–~~Dec 2023~~ TS2	Chau et al. (2022, 2024)
pH	Surface pH on total scale				
pCO_2 (µatm)	Surface aqueous partial pressure of CO_2				
CO_2 flux ($mol\,m^{-2}\,yr^{-1}$)	Surface downward flux of total CO_2				
Alkalinity ($\mu mol\,kg^{-1}$)	Total alkalinity in surface seawater				
PAR ($mol\,m^{-2}\,d^{-1}$)	Photosynthetically available radiation	4 km/9 km	Monthly	Oct 1997–Feb 2025	NASA Ocean Biology Processing Group (2018)
Chl-a ($mg\,m^{-3}$)	Mass concentration of chlorophyll in surface water				
SSH (m)	Sea surface height above geoid	0.25° × 0.25°	Monthly	Jan 1993–Aug 2023	Hauser et al. (2020)
EKE ($cm^2\,s^{-2}$)	Surface averaged eddy kinetic energy				

(CCMP) product (Mears et al., 2022). Mixed-layer depth (MLD) is obtained from the CMEMS Multi-Observation Global Ocean 3D product (Guinehut et al., 2012). Dynamical variables include sea surface height (SSH) and eddy kinetic energy (EKE), both derived from AVISO satellite altimetry (Hauser et al., 2020). Bio-optical variables comprise photosynthetically active radiation (PAR) and chlorophyll a (Chl-a) from NASA Level-3/Level-4 ocean-color products (NASA Ocean Biology Processing Group, 2018). Carbon-chemistry variables include dissolved inorganic carbon (DIC), total alkalinity, pH, sea surface partial pressure of CO_2 (pCO_2), and CO_2 flux, all obtained from the CMEMS Surface Ocean Carbon Fields product (Chau et al., 2022, 2024). All feature variables, last accessed in March 2025, were standardized onto a uniform monthly 0.5° grid to maintain spatial consistency across the reconstruction.

3 Method

The overall workflow for constructing the GEOXYGEN dataset is illustrated in Fig. 2, comprising data collection and preprocessing, heterogeneity-based partitioning, and