

# Variational Autoencoder Framework for Hyperspectral Retrievals (Hyper-VAE) of Phytoplankton Absorption and Chlorophyll a in Coastal Waters for NASA's EMIT and PACE Missions

Jiadong Lou, Bingqing Liu\*, Yuanheng Xiong, Xiaodong Zhang, and Xu Yuan\*

**Abstract**—Phytoplankton absorb and scatter light in unique ways, subtly altering the color of water—changes that are often minor for human eyes to detect but can be captured by sensitive ocean color instruments onboard satellites from space. Hyperspectral sensors, paired with advanced algorithms, are expected to significantly enhance the characterization of phytoplankton community composition (PCC), especially in coastal waters where ocean color remote sensing applications have historically encountered significant challenges. This study presents novel machine learning-based solutions for NASA's hyperspectral missions including EMIT (7nm), and PACE (2.5 nm), tackling high-fidelity retrievals of phytoplankton absorption coefficient ( $a_{phy}$ ) and chlorophyll a (Chl-a) from their hyperspectral remote sensing reflectance ( $R_{rs}$ ). Given that a single  $R_{rs}$  spectrum may correspond to varied combinations of inherent optical properties (IOPs) and associated concentrations (e.g., Chl-a), the Variational Autoencoder (VAE) is used as a backbone in this study to handle such multi-distribution prediction problems. We first time tailor the VAE model with innovative designs to achieve hyperspectral retrievals of  $a_{phy}$  and of Chl-a from hyperspectral  $R_{rs}$  in optically complex estuarine-coastal waters. Validation with extensive experimental observation demonstrates superior performance of the VAE models with high precision and low bias. The in-depth analysis of VAE's advanced model structures and learning designs highlights the improvement and advantages of VAE-based solutions over the mixture density network (MDN) approach, particularly on high-dimensional data, such as PACE. Our study provides strong evidence that current EMIT and PACE hyperspectral data as well as the upcoming Surface Biology Geology (SBG) mission will open new pathways toward a better understanding of phytoplankton community dynamics in aquatic ecosystems when integrated with AI technologies.

**Index Terms**—Hyperspectral remote sensing; Machine learning; Variational autoencoder; Phytoplankton absorption; Phytoplankton community composition; Chlorophyll a; Coastal waters

This work was supported in part by NASA EMIT program through Grant 80NSSC24K0865, by NOAA NCCOS Hypoxia program under Grant NA23NOS4780285, by NSF under Grants 2425811, 2425812, and 2019511. (Corresponding author: Bingqing Liu and Xu Yuan).

Jiadong Lou is with Department of Computer & Information Sciences, University of Delaware, Newark, DE 19716 USA (e-mail: loujd@udel.edu).

Bingqing Liu is with School of Geosciences, University of Louisiana at Lafayette, Lafayette, LA 70504 USA (e-mail: bingqing.liu@louisiana.edu).

Yuanheng Xiong is with Department of Ocean Science and Engineering, The University of Southern Mississippi, Stennis Space Center, MS 39525 USA (e-mail: yuanheng.xiong@usm.edu).

Xiaodong Zhang is with School of Ocean Science and Engineering, The University of Southern Mississippi, Stennis Space Center, MS 39525 USA (e-mail: xiaodong.zhang@usm.edu).

Xu Yuan is with Department of Computer & Information Sciences, University of Delaware, Newark, DE 19716 USA (e-mail: xyuan@udel.edu).

## I. INTRODUCTION

Phytoplankton is an extremely diverse set of microorganisms, varying in cell morphologies, biogeochemical functions, and physiological responses to environmental disturbances [21]. As the primary producer of ocean's food web, phytoplankton produces approximately 50 percent of Earth's oxygen, regulate the global carbon cycle and climate, and support various ecosystem services, such as fisheries, water quality, and biodiversity [7]. Therefore, knowledge of phytoplankton biomass and their community composition is critical to understanding the food web structure, higher trophic level production (e.g., fisheries), and biological shifts among other complex Earth Science questions, especially in the context of degraded water quality (e.g., eutrophication) and climate change (e.g., warming temperatures), demanding attention at local, regional, and global scales [9]. As such, there is an increasing interdisciplinary interest in studying phytoplankton community dynamics in estuarine-coastal waters, where massive riverine inputs of nutrient-rich freshwaters often lead to eutrophication, harmful algal blooms (HABs), and the annual recurrence of bottom-water hypoxia events, which cause widespread and severe impacts on the aquatic ecosystem [3], [43], [57], [23], [54].

In the field of ocean color remote sensing, the concentration of chlorophyll a (Chl-a) and phytoplankton absorption properties ( $a_{phy}$ ) are two of the most commonly used metrics for assessing phytoplankton abundance and diversity in aquatic environments [41], [56]. Those phytoplankton-related satellite algorithms are rooted in the physical principle that remote sensing reflectance ( $R_{rs}$ ,  $sr^{-1}$ ), the ratio of water-leaving radiance to the total downwelling irradiance just above water, is determined by the inherent optical properties (IOPs), most importantly the total backscattering coefficient ( $b_{total}$ ,  $m^{-1}$ ) and the total absorption coefficient ( $a_{total}$ ,  $m^{-1}$ ) [38]. These components are interrelated through the radiative transfer theory, presented here in its simplified form [38]:

$$R_{rs}(\lambda) = c \times \left( \frac{bb_{total}}{a_{total} + bb_{total}} \right), \quad (1)$$

where  $R_{rs}(\lambda)$  represents remote sensing reflectance at wavelength  $\lambda$ ;  $c$  is the proportionality constant that accounts for factors such as the geometry of the observation and specific conditions of the water body. Further,  $a_{total}$  can be separated into absorption by four major constituents (Roesler and Perry,

1995) including water ( $a_w$ ), colored dissolved organic matter ( $a_{CDOM}$ ), non-algal particle ( $a_{NAP}$ ) and phytoplankton ( $a_{phy}$ ), i.e.,

$$a_{total} = a_w + a_{CDOM} + a_{NAP} + a_{phy} . \quad (2)$$

Recognizing this physical principle, Chl-a, a key indicator of phytoplankton biomass, has been widely estimated from multi-spectral satellite data using semi-analytical approaches that firstly infer  $a_{phy}$  peaks within the red portions of the spectrum from  $R_{rs}$  and then estimate Chl-a [18], [29], [34], [50]. Alternatively, empirical methods that utilize band-ratio metrics, have a long history of being employed to estimate Chl-a directly from  $R_{rs}$  [35], [24], [49]. Band ratios, particularly those in the red and near-infrared (NIR) regions [17], as well as fluorescence line-height (FLH) approaches that utilize three bands in either red or red-NIR bands to compute characteristic spectral reflectance peaks, are commonly used for Chl-a estimation in estuarine and coastal waters [20], [31]. Additionally, recent phytoplankton-related algorithms have increasingly favored machine learning techniques, such as Mixture Density Networks (MDNs) [42], [51], to enhance the accuracy and efficiency based on large amounts of historical data, such as GLORIA [30]. Pahlevan et al. (2021) [41] not only applied MDNs—originally developed for multispectral sensors (e.g., Sentinel-3 OLCI) to enhance the estimation of Chl-a and, more critically,  $a_{phy}$ , but also adapted this approach for the Hyperspectral Imager for the Coastal Ocean (HICO), the first spaceborne hyperspectral imaging spectrometer designed specifically for sampling coastal oceans. This adaptation paves new possibilities for ocean color algorithms, especially in the context of the increasing availability of hyperspectral missions, such as PACE (Fig. 1).

Hyperspectral sensors, with their narrow spectral bands from ultraviolet to NIR, capture far more spectral details than traditional multispectral sensors and, when paired with advanced retrieval algorithms, can extend beyond phytoplankton abundance (e.g., Chl-a) to quantify certain aspects of phytoplankton community composition (PCC). Hyperspectral  $a_{phy}$  is a crucial optical proxy for characterizing PCC as the variability in phytoplankton cellular pigment compositions significantly affects the magnitude and spectral shape of  $a_{phy}$ , which is further reflected in  $R_{rs}$  [5], [11], [25], [33], [60], [59]. Semi-analytical models [29], [46], and machine learning approaches [41], [2], [56] have been popular methods for solving the inverse problem of estimating IOPs (e.g.,  $a_{phy}$ ) from  $R_{rs}$ , most of which generally perform well in the open ocean and are well-suited to multispectral satellites. Studies focusing on the retrieval of  $a_{phy}$  in bio-geo-chemically and optically complex estuarine and coastal waters are comparatively sparse. This is because coastal waters, typically rich in colored dissolved organic matter (CDOM) and/or non-algal particles (NAP), can overwhelm phytoplankton optical signals [32], [15], [16], which further leads to uncertainty regarding the degree to which PCC information (e.g., taxa, function, or size) can be accurately gleaned from satellite data in coastal waters and the levels of confidence. Last but not least, different combinations of IOPs and associated concentrations that could yield similar or identical  $R_{rs}$  spectra [41]

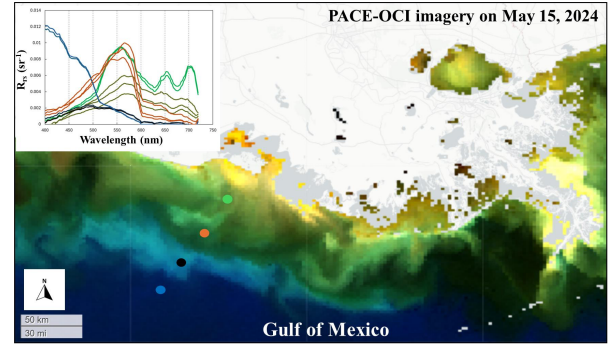


Fig. 1. NASA's PACE-OCI Level 2 AOP data obtained on May 15, 2024, covers diverse water types in the northern Gulf of Mexico. The data is displayed using a band combination of remote sensing reflectance ( $R_{rs}$ ) at 440, 560, and 650 nm and processed using the HyperCoast open-source tool (<https://hypercoast.org/>).

are further exacerbated by the scarcity of spectral bands available in current multispectral satellite sensors, complicating the inversion of IOPs (e.g.,  $a_{phy}$ ) from  $R_{rs}$  spectra. MDN-based machine learning approaches [41] have been proposed, bringing progress toward addressing this “one to many” inversion challenge. While MDN’s performance in predicting IOPs—compared to heritage inversion algorithms such as the Quasi Analytical Algorithm (QAA) [29] and Generalized IOP Inversion (GIOP) [58], demonstrate their potential in coastal waters, the “one-to-many” issue still requires further exploration from both algorithmic and data perspectives. This is because the GLORIA dataset lacks true “one-to-many” features, and thus, capabilities of learning models designed to handle one-to-many inversion problems have yet to be fully explored. Nevertheless, the global Chl-a and  $a_{phy}$  algorithms in coastal waters, especially hyperspectral retrievals of  $a_{phy}$ , are still in the early stages of development.

With the launch of new hyperspectral sensors like NASA’s Ocean Color Instrument (OCI) on the Plankton, Aerosol, Cloud, Ocean Ecosystem (PACE) mission (Fig. 1), which features narrow spectral bands at 2.5 nm spanning ultraviolet (350 nm) to NIR (885 nm) and provides 2-day global coverage [8], [13], there is an urgent need for global coastal algorithms to retrieve inherent IOPs and biogeochemical metrics, such as Chl-a and  $a_{phy}$ , in optically complex waters. Further, the Earth Surface Mineral Dust Source Investigation (EMIT) instrument, a precursor to the Surface Biology Geology (SBG) mission, offers both high spectral (380-2500 nm, 7.4 nm resolution) and much higher spatial (60 m) resolutions, focusing on coastal oceans [19], [53]. Both PACE and EMIT represent critical advancements for monitoring water quality in coastal zones. However, to fully utilize their potential, this is an urgent need for globally applicable algorithms for the retrieval of IOPs, tailored to these new hyperspectral missions.

To this end, this study aims to propose novel machine learning-based solutions, notably Variational Autoencoders (VAE) [22], [26], [37] for NASA’s hyperspectral missions, e.g., PACE and EMIT, to achieve high fidelity retrievals of  $a_{phy}$  and Chl-a from the hyperspectral  $R_{rs}$ . Given the complexity of a single  $R_{rs}$  spectrum, which may correspond to varied com-

binations of IOPs and concentrations, particularly in optically complex waters, the VAE model is ideally suited as a backbone to manage such multi-distribution prediction problems because it captures the essential and high-level features of the input data, thereby effectively learning and characterizing the most important patterns. VAEs excel in balancing uncertainty and reliability for multi-value prediction tasks by effectively mitigating overfitting caused by noise in the training data, and learning latent patterns through a lower-dimensional latent space [10], [27], [52], [61], thereby providing robust high-dimensional predictions. This manuscript will introduce the innovative design of VAE-based models tailored for PACE and EMIT, and their applications in predicting  $a_{phy}$  and Chl-a. The performance is evaluated using the dataset presented in [41], [40] along with bio-optical datasets collected in Galveston Bay, TX, USA [33]. The advantages of the VAE-based solution, including its model structure and learning designs, for hyper-retrieval predictions are also discussed.

## II. DATASET

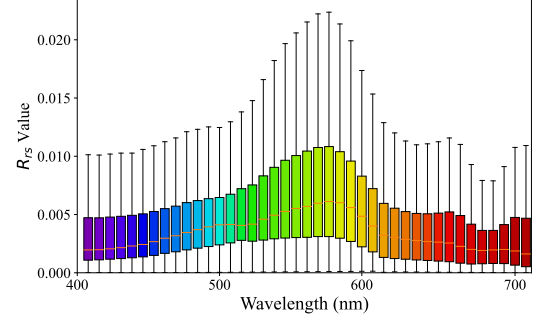
### A. Development Dataset

For the retrievals of  $a_{phy}$ , we utilized paired  $R_{rs}$ - $a_{phy}$  data from optically complex waters in the U.S. from the SeaWiFS Bio-optical Archive and Storage System (SeaBASS) dataset in conjunction with data from lakes in Asia, all of which were compiled and shared by Pahlevan et al. (2021) [41] with additional data added from [40]. It was documented that the majority of  $R_{rs}$  spectra were acquired using the HyperOCR® radiometers, while a small portion was obtained through a handheld radiometer (e.g., ASD FieldSpec®) [39]. All  $a_{phy}$  spectra were measured using the standard quantitative filter pad technique (QFT) [45]. In comparison to paired  $R_{rs}$ - $a_{phy}$  dataset, the  $R_{rs}$ -Chl-a covers a markedly wider range of Chl-a concentrations, better representing global coastal water quality conditions [41].

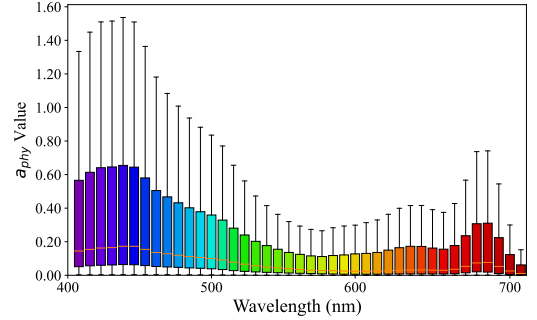
### B. Data Preprocessing

Data quality control was conducted to exclude outliers (e.g. the negative values) and invalid values (e.g., NaN) from both  $R_{rs}$ - $a_{phy}$  and  $R_{rs}$ -Chl-a datasets. The  $a_{phy}$  spectra that are very noisy or exhibit zigzag patterns, were removed, and the same criteria were applied to the  $R_{rs}$  spectra. This quality control resulted in 2114 and 6111 pairs of valid  $R_{rs}$ - $a_{phy}$  and  $R_{rs}$ -Chl-a datasets, respectively. We resampled hyperspectral  $R_{rs}$ - $a_{phy}$  data to the wavelength settings corresponding to NASA's latest hyperspectral missions, including PACE (2.5 nm) and EMIT (7.4 nm) within the 400-700 nm range. For  $R_{rs}$ - $a_{phy}$  prediction, the VAE- $a_{phy}$  models take 141, and 41 bands of  $R_{rs}$  and  $a_{phy}$  as both input and output data for PACE, and EMIT, respectively. For  $R_{rs}$ -Chl-a prediction, the input data for the VAE-Chl-a model consists of 141, and 41 bands of  $R_{rs}$  vectors for PACE, and EMIT, all with one-element output for the Chl-a value. Fig. 2 shows the spectral distribution of the  $R_{rs}$ - $a_{phy}$  dataset corresponding to EMIT wavelengths. For the predictions of  $a_{phy}$  and Chl-a concentrations, 70% of preprocessed data is used for training, while the remaining 30% is reserved for testing. Moreover,

for  $R_{rs}$ - $a_{phy}$  data, the min-max normalization with empirical parameters from training data is applied to each  $R_{rs}$  data sample, i.e., the  $R_{rs}$  spectra are scaled to fit into the designated value range, to account for the wide range of values across different wavelengths. As Chl-a is log-normally distributed in the natural environment, a logarithmic scale is performed on Chl-a to account for its natural variability.



(a) Spectral distribution of the  $R_{rs}$



(b) Spectral distribution of the  $a_{phy}$

Fig. 2. Spectral distribution of the  $R_{rs}$ - $a_{phy}$  dataset, with bars denoting the minimum and maximum values of (a)  $R_{rs}$  ( $sr^{-1}$ ) and (b)  $a_{phy}$  ( $m^{-1}$ ) using EMIT spectral setting.

### C. In-situ Hyperspectral Measurements

Additional hyperspectral  $R_{rs}$ - $a_{phy}$  field data are used to test the VAE's robustness and generalizability. Phytoplankton absorption coefficient ( $a_{phy}$ ) was obtained from surface water samples collected during two surveys conducted in September and October of 2017, in Galveston Bay, TX, USA. The water samples were filtered through 0.7  $\mu m$  GF/F filters for the measurement of absorption by particles and non-algal particles via a Perkin-Elmer Lambda 850 spectrophotometer fitted with a 15-cm diameter integrating sphere [33] using the internally mounted sample in the IS-mode [45]. Phytoplankton absorption coefficient ( $a_{phy}$ ) was then calculated as the difference between the adsorption of particles and non-algal particles [34]. The  $R_{rs}$  values were obtained using a GER 1500 512iHR spectroradiometer across the 350–1050 nm spectral range. At each station, sky radiance, plate radiance, and water radiance were recorded (each repeated three times) and processed to derive the above-water remote sensing reflectance [33].

#### D. Revisiting MDN-based Methods

Mixture Density Networks (MDN)-based methods have been developed for IOP retrievals (e.g., Chl-a and  $a_{phy}$ ) in ocean color remote sensing and tailored for applications to diverse hyperspectral missions, such as HICO and PRISMA [42], [2], [40], [51], [41]. MDN is a class of neural networks designed to model complex conditional probability distributions. In contrast to traditional deterministic models, which predict a single output for a given input, MDNs learn a mixture of Gaussian distributions to capture the multi-modal nature of the target variable [42]. By estimating both the mean and variance of multiple Gaussian components, MDNs can in principle include uncertainty in regression tasks. Due to this capability, MDNs have been utilized to tackle the one-to-many problem in IOPs retrieval [40], [42], [41], where the same  $R_{rs}$  can correspond to multiple combinations of IOPs and associated concentrations. Despite this potential, the existing MDN implementations [40], [42], [41] either select the mean of the highest-weighted Gaussian component or compute a weighted average of the means of all components as the final predictions, which renders the inference process entirely deterministic. Once trained, the model consistently produces the same prediction for a given  $R_{rs}$ , regardless of how many times it is executed. While this approach ensures stability, it inherently collapses the multimodal nature of the problem to a single-point estimate, limiting MDNs' ability to address the one-to-many challenge. Further, upon analyzing the GLORIA dataset used in the existing MDN work, it became evident that the dataset does not contain one-to-many features—no different Chl-a values or  $a_{phy}$  spectra correspond to the same or similar  $R_{rs}$  spectra. This explains why the current implementations of MDNs as reported in [40], [42], [41] perform well when tested on the GLORIA and additional supplementary datasets. However, in real-world scenarios, the one-to-many problem can still arise, highlighting the need for developing methods that can more effectively address this challenge.

To fully utilize the intrinsic capability of MDNs for true one-to-many mapping, rather than directly selecting the Gaussian mean with the highest weight or weighted average of the means of all components as the output as reported in previous studies, a feasible solution is to sample  $a_{phy}$  or Chl-a from the learned mixture of Gaussian distributions. This modification ensures that the uncertainty captured in the learned distribution is effectively incorporated in MDN predictions for the same  $R_{rs}$  input. However, this modification also increases variability in predictions due to the inherent stochastic nature of the sampling process. In Section IV, we implement this modified MDN approach, referred to as M-MDN, and evaluate its performance in addressing one-to-many problems.

### III. HYPER-VAE

#### A. Variational Autoencoder

In this study, Variational Autoencoder (VAE) was introduced and for the first time applied in the ocean color remote sensing field for inversion retrieval of IOPs and concentrations. As a category of generative models, VAEs were originally

designed to generate new data that resemble the original input. The encoding phase aims to capture the essential high-level features of the input data and compress them into a low-dimensional representation of the input data, known as a latent vector, which characterizes the most important patterns in the data. In the subsequent decoding phase, this latent vector is decoded back into the original data space, enabling the reconstruction of the input data from the distributions learned in the latent space. This encoder-decoder process allows the VAE to generate new data that closely resemble the original input, preserving their key features while introducing variations. The VAE consists of two core components: an encoder and a decoder, each implemented as a neural network (Fig. 3).

Mathematically, the encoder maps the input data, denoted as  $\mathbf{x}$ , to a latent vector  $\mathbf{z}$ , where each element in  $\mathbf{z}$  follows a Gaussian distribution, yielding

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu(\mathbf{x}), \sigma^2(\mathbf{x})) , \quad (3)$$

where  $\mu(\mathbf{x})$  and  $\sigma^2(\mathbf{x})$  represent the mean and the variance of Gaussian distributions, learned from the input data  $\mathbf{x}$ .

The key feature of the VAE that allows it to address the one-to-many prediction problem is the generation of the latent vector, where stochasticity is introduced through its latent space. Mathematically, to enable backpropagation calculation [48] during the model training process, the reparameterization trick [26] is typically used to sample each element  $z$  in the latent vector  $\mathbf{z}$ , expressed as:

$$z = \mu(x) + \sigma(x) \odot \epsilon , \quad (4)$$

where  $\epsilon \sim \mathcal{N}(0, I)$  is a standard normal noise, and  $\odot$  denotes the element-wise multiplication. Due to the incorporation of standard normal noise, VAE models can effectively learn the uncertainty mapping between the  $R_{rs}$  and the target IOP. This enables the model to represent the inherent ambiguity of the inverse problem, where the same  $R_{rs}$  can correspond to multiple valid IOP combinations. Furthermore, during the inference stage, the reparameterization trick is also adopted to ensure that even for the same input  $R_{rs}$ , different latent vectors can be generated. This crucial property allows the VAE to produce diverse IOP predictions from the same input, effectively addressing the one-to-many mapping problem.

To obtain a single output, the decoder takes the latent vector  $\mathbf{z}$  as input and then outputs the reconstructed data, denoted as  $\mathbf{x}'$ . Mathematically, the generation of  $\mathbf{x}'$  involves sampling from a conditional distribution  $p(\mathbf{x}|\mathbf{z})$ . In practice, this conditional distribution is assumed to be a Gaussian distribution with a fixed variance  $\sigma^2$ , while the mean is provided by the decoder output, yielding:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{x}', \sigma^2) , \quad (5)$$

where  $\sigma^2$  is the fixed variance.

While the parameterization step introduces noise, the decoder learns a structured mapping from  $p(\mathbf{x}|\mathbf{z})$  to the desired IOP that ensures the predictions remain consistent and stable. Essentially, the decoder translates variations in the latent space into realistic, physically meaningful predictions, rather than erratic fluctuations.

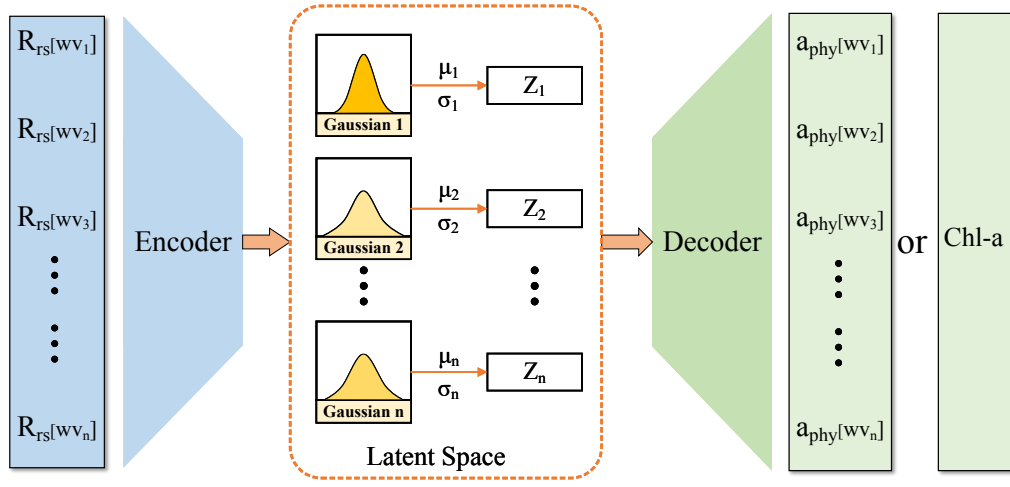


Fig. 3. The structure of VAE for Predicting  $a_{phy}$  and Chl-a.

Furthermore, the training objective of VAE is to minimize the loss calculated from two aspects, i.e.,

$$\mathcal{L} = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] + \lambda \cdot KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) , \quad (6)$$

where  $\lambda$  is the weight to balance the two terms. The loss function aligns with the standard VAE loss introduced in [26] in their seminal work on VAEs and has been extensively used in many VAE-based studies [22]. Here, the first term characterizes the reconstruction loss, which is practically calculated as the difference between the reconstructed data  $\mathbf{x}'$  and original data  $\mathbf{x}$ . The second term is the KL divergence to regularize the latent space to be close to the prior distribution  $p(\mathbf{z})$ , which is usually chosen to be a standard normal distribution. Assigning a weight to the KL divergence term and combining two terms together allow for a trade-off between accurate reconstruction and meaningful latent space structure. In our study, we carefully tuned this  $\lambda$  to capture the diversity of input-output relationships while avoiding over-regularization.

### B. VAE-based Model for Predicting $a_{phy}$ and Chl-a

While the VAE was originally designed for data generation and reconstruction, it can also be applied for prediction tasks. In this study, we tailor basic VAE structures into two models, VAE- $a_{phy}$  and VAE-Chl-a, to predict  $a_{phy}$  and Chl-a, separately. Since this is the first application of VAE-based models for IOP retrievals, we intentionally designed separate models for each parameter to fully explore the core capability of VAEs in one-to-many mapping. Notably, the model structure and design logic implemented in this study can be readily modified for the simultaneous prediction of multiple parameters, which is discussed in Section V-E.

In both VAE- $a_{phy}$  and VAE-Chl-a models, the input is  $R_{rs}$ , and the outputs are  $a_{phy}$  and Chl-a, respectively. The encoders map the  $R_{rs}$  into the latent vectors  $z_{a_{phy}}$  and  $z_{Chla}$ , respectively, i.e.,

$$q(z_{a_{phy}}|R_{rs}) = \mathcal{N}(z_{a_{phy}}; \mu(R_{rs}), \sigma^2(R_{rs})) , \quad (7)$$

$$q(z_{Chla}|R_{rs}) = \mathcal{N}(z_{Chla}; \mu(R_{rs}), \sigma^2(R_{rs})) . \quad (8)$$

The decoders of VAE- $a_{phy}$  and VAE-Chl-a then map their latent vectors back to the  $a_{phy}$  and Chl-a data space for reconstructing these data, denoted as  $a'_{phy}$  and  $Chla'$ , yielding

$$p(a_{phy}|z_{a_{phy}}) = \mathcal{N}(a_{phy}; a'_{phy}, \sigma^2) , \quad (9)$$

$$p(Chla|z_{Chla}) = \mathcal{N}(Chla; Chla', \sigma^2) . \quad (10)$$

Hence, the two VAE models will output the reconstructed  $a'_{phy}$  and  $Chla'$ , respectively, which are able to further predict the  $a_{phy}$  and Chl-a values from new  $R_{rs}$  data. During the model training process, the reconstruction loss is tailored to be the L-1 loss between the predicted value  $a'_{phy}$  ( $Chla'$ ) and its ground truth value. The overall training loss for each VAE model is expressed as:

$$\mathcal{L}_{a_{phy}} = \|a'_{phy} - a_{phy}\|_1 + \lambda \cdot KL(q(z_{a_{phy}}|R_{rs})||p(z_{a_{phy}})) , \quad (11)$$

$$\mathcal{L}_{Chla} = \|Chla' - Chla\|_1 + \lambda \cdot KL(q(z_{Chla}|R_{rs})||p(z_{Chla})) , \quad (12)$$

where  $p(z_{a_{phy}})$  and  $p(z_{Chla})$  are considered as the standard normal distribution.

By adapting the VAE models in this way, we harness its capability to learn meaningful latent representations, enabling effective prediction of  $a_{phy}$  and Chl-a.

### C. VAE- $a_{phy}$ and VAE-Chl-a Models Configurations

To accommodate different data dimension requirements, the configuration of VAE- $a_{phy}$  and VAE-Chl-a models vary, with their parameters listed in Tables I and II.

For VAE- $a_{phy}$ , the encoder consists of two fully connected layers, each followed by Batch Normalization and a LeakyReLU activation function. The dimension of the input layer is indicated with a variable input\_dim, which can be flexibly adjusted according to the dimension of  $R_{rs}$ . The first layer maps the input data dimension to 512 dimensions, and the second layer maps from 512 to 1024 dimensions. Two parallel fully connected layers then generate a 256-dimensional latent vector, producing the means and variances for 256 Gaussian distributions by mapping the 1024-dimensional output of the encoder to the 256 dimensions. The decoder consists of three fully connected layers. The first two layers, each followed by Batch Normalization and a LeakyReLU activation function,



map from 256 dimensions to 512 dimensions and from 512 dimensions to 1024 dimensions, respectively. The last layer maps from 1024 dimensions to the output dimensions, indicated as `output_dim`, which are the dimensions of predicted  $a_{phy}$  vectors. Here, the Softplus activation function is applied to ensure that the value is positive.

Similarly, for VAE-Chl-a, the encoder consists of two fully connected layers, each followed by Batch Normalization and LeakyReLU activation function. The input layer's dimension is also indicated with a variable `input_dim`. The first layer maps the input data dimensions to 256 dimensions, and the second layer maps from 256 dimensions to 128 dimensions. For the 64-dimensional latent vector generation, two parallel fully connected layers produce the means and variances for 64 Gaussian distributions by mapping the 128-dimensional output of the encoder to 64 dimensions. The decoder consists of three fully connected layers. The first two layers, each followed by Batch Normalization and a LeakyReLU activation function, both map from 64 dimensions to 64 dimensions. The last layer maps from 64 dimensions to the output dimension 1, which is the predicted Chl-a value.

#### D. Performance Metrics

Following previous studies [40], [42], we adopt eight metrics to comprehensively evaluate the prediction performance of our proposed VAE models. For the following metric formulas, we collectively denote  $E$  as the predicted value of our model and  $M$  as the actual value. Additionally, since  $a_{phy}$  and Chl-a values span a large range, we follow the approach in [40], [42] by employing a logarithmic transformation for calculating several metrics, where 'log' represents  $\log_{10}$ . These evaluation metrics can be categorized into three groups.

First, we use Mean Absolute Log-Error (MALE), Root Mean Square Error (RMSE), and Root Mean Square Log-Error (RMSLE) to assess prediction errors, i.e., the magnitude difference between the predicted values and the actual values. These metrics are calculated as below,

$$\text{MALE} = 10^Y, Y = \frac{\sum_{i=1}^N |\log(E_i) - \log(M_i)|}{n}, \quad (13)$$

$$\text{RMSE} = \left[ \frac{\sum_{i=1}^N (E_i - M_i)^2}{n} \right]^{\frac{1}{2}}, \quad (14)$$

$$\text{RMSLE} = \left[ \frac{\sum_{i=1}^N (\log(E_i) - \log(M_i))^2}{n} \right]^{\frac{1}{2}}. \quad (15)$$

The three metrics collectively provide a comprehensive evaluation of a model's performance by capturing different aspects of the error distribution. Specifically, MALE calculates the average magnitude of errors, providing the most straightforward measure of overall prediction accuracy. RMSE emphasizes larger errors by squaring them before averaging, highlighting whether a model suffers from larger deviations in some  $a_{phy}$  and Chl-a predictions. Furthermore, RMSLE captures the logarithmic differences between predicted and actual values, making it particularly useful for the wide range of  $a_{phy}$  and Chl-a values. By applying a logarithmic transformation,

RMSLE reduces the influence of large errors relative to small ones, providing a more balanced assessment of prediction accuracy. The closer these three metrics are to zero, the better the model's performance, indicating higher precision in predictions.

Second, we adopt another two metrics, i.e., Log-Bias and Slope, to evaluate whether there is a systematic deviation between the predicted values and true values. The Log-Bias is calculated as follows:

$$\text{Log-Bias} = 10^Z, Z = \frac{\sum_{i=1}^N (\log(E_i) - \log(M_i))}{n}. \quad (16)$$

It reflects the overall direction of prediction bias in VAE models. Additionally, *Slope* is determined as the slope of the line derived from the linear regression of predicted values versus actual values, expressed as

$$\text{Slope} = \frac{\sum_{i=1}^N (M_i - \bar{M})(E_i - \bar{E})}{\sum_{i=1}^N (M_i - \bar{M})^2}, \quad (17)$$

where  $\bar{M}$  and  $\bar{E}$  respectively represent the means of the actual and predicted values. The closer the two metrics are to 1, the better the model's performance, indicating that the model does not systematically overestimate or underestimate the target values.

Third, we also follow the work [40] to report three median value-oriented metrics, *MAPE*,  $\epsilon$ , and  $\beta$ , as calculated below.

$$\text{MAPE} = 100 \times \text{md} \left( \frac{|E - M|}{M} \right), \quad (18)$$

$$\epsilon = 100 \times \left( 10^{\text{md}(|\log(\frac{E}{M})|)} - 1 \right), \quad (19)$$

$$\beta = 100 \times \text{sign}(Z)(10^{|Z|} - 1), Z = \text{md} \left( \log \frac{E}{M} \right), \quad (20)$$

where  $\text{md}(\cdot)$  represents the median value. These three metrics assess the prediction error and bias of the median value, aiming to minimize the impact of prediction outliers. However, it is worth noting that outliers in the prediction of  $a_{phy}$  and Chl-a are important, as they sometimes reflect extreme water conditions, such as algal blooms. Therefore, relying solely on the median value cannot provide a comprehensive assessment of model performance.

## IV. RESULTS

### A. MDN Baseline Settings

To ensure a comprehensive comparison with prior work, we train the MDN model following the same approach as described in [40]. Specifically, we reproduce their model based on the reported settings, including layer parameters, the number of Gaussian distributions, and the selection strategy of the highest-weighted Gaussian mean instead of sampling. The model consists of five layers, each with 100 neurons and five Gaussian distribution components. However, differences in dataset preprocessing and hardware environment compared to prior work necessitate adjustments to the learning rate and training epochs to achieve optimal performance. The learning rate is set to  $1e^{-3}$  with Adam optimizer and the training epoch is 2000, where the model with the best performance

TABLE I  
MODEL STRUCTURE FOR VAE- $a_{phy}$

Layer Type	Input Dimension	Output Dimension	Additional Operations
<b>Encoder</b>			
Fully Connected	input_dim	512	Batch Normalization, LeakyReLU(0.2)
Fully Connected	512	1024	Batch Normalization, LeakyReLU(0.2)
Fully Connected	1024	256	Output Mean
Fully Connected	1024	256	Output Log Variance
<b>Decoder</b>			
Fully Connected	256	512	Batch Normalization, LeakyReLU(0.2)
Fully Connected	512	1024	Batch Normalization, LeakyReLU(0.2)
Fully Connected	1024	output_dim	Softplus Activation

TABLE II  
MODEL STRUCTURE FOR VAE-CHL-A

Layer Type	Input Dimension	Output Dimension	Additional Operations
<b>Encoder</b>			
Fully Connected	input_dim	256	Batch Normalization, LeakyReLU(0.2)
Fully Connected	256	128	Batch Normalization, LeakyReLU(0.2)
Fully Connected	128	64	Output Mean
Fully Connected	128	64	Output Log Variance
<b>Decoder</b>			
Fully Connected	64	64	Batch Normalization, LeakyReLU(0.2)
Fully Connected	64	64	Batch Normalization, LeakyReLU(0.2)
Fully Connected	64	1	

among these 2000 epochs is stored for use in prediction tasks. This setup ensures a fair and direct comparison between both MDN implementations under consistent conditions. Moreover, MDN involves matrix calculations where the full covariance matrix becomes increasingly unstable and difficult to optimize as the input dimensionality grows. Specifically, the number of parameters in a full covariance matrix grows quadratically with the input dimensionality. Consequently, under PACE spectral settings, the covariance matrix might become ill-conditioned, leading to numerical instability. This can cause gradient explosion or vanishing gradients during optimization, leading to poor convergence. To address these issues, we replace the full covariance matrix with a diagonal covariance matrix approximation, which significantly reduces computational complexity, improves numerical stability, and facilitates more reliable optimization, ultimately enhancing the performance of MDN in high-dimensional settings. The prediction results from the modified version of MDN, i.e., M-MDN, as we have presented in Section II-D, are also considered to assess the MDN-based approach’s capability for one-to-many mapping.

### B. Evaluations of VAE- $a_{phy}$ Predictions

We have conducted experiments using VAE- $a_{phy}$  for predicting  $a_{phy}$  at PACE and EMIT spectral settings in the range of 400-700 nm, comparing the results with those from MDN models [40]. Regarding predictions of  $a_{phy}$  for VAE and MDN models, we present three important wavelengths, including 440 nm at Chl-a blue absorption, 670 nm at Chl-a red absorption, and 620 nm at phycocyanin absorption.

Figure 4 illustrates the performance of VAE- $a_{phy}$  and MDN models across eight metrics, in predicting  $a_{phy}$  at PACE and EMIT wavelengths closest to 440 nm, 620 nm, and 670 nm. The results demonstrate that VAE- $a_{phy}$  model performs better than MDN across most metrics for both PACE and EMIT, with lower errors and improved agreement with actual  $a_{phy}$  values. Specifically, for PACE wavelength, the VAE model achieves lower MALE values across all these three bands, with 1.32 at 440 nm, 1.36 at 620 nm, and 1.32 at 670 nm, while the MDN model yields comparatively higher values for PACE, at 1.43, 1.56, and 1.54, respectively. Besides, regarding the other two prediction error metrics, RMSE and RMSLE, VAE- $a_{phy}$  also exhibits smaller values, reinforcing its stable predictive accuracy. Beyond prediction errors, VAE- $a_{phy}$  also achieves better Log-Bias and slope values, remaining closer to 1, which suggests a more accurate and reliable fit between predicted and actual values. In contrast, MDN displays greater variability and deviation at extreme values, such as overestimation of  $a_{phy}$  (670) for PACE, as shown in Figure 4(j)). Furthermore, for the three median-value oriented metrics, VAE- $a_{phy}$  still exhibits superior performance, except for  $\beta$  for PACE at 670 nm. The stability and accuracy of VAE in hyperspectral  $a_{phy}$  retrievals suggest that it is a suitable machine learning framework for one-to-many inversion problems in ocean color remote sensing, highlighting its potential for future hyperspectral applications, such as GLIMR and SBG.

To comprehensively assess the prediction performance across the 400-700 nm wavelength range, we compared three evaluation metrics, including RMSE, Log-Bias, and  $\beta$ , between VAE and MDN models across all PACE and EMIT

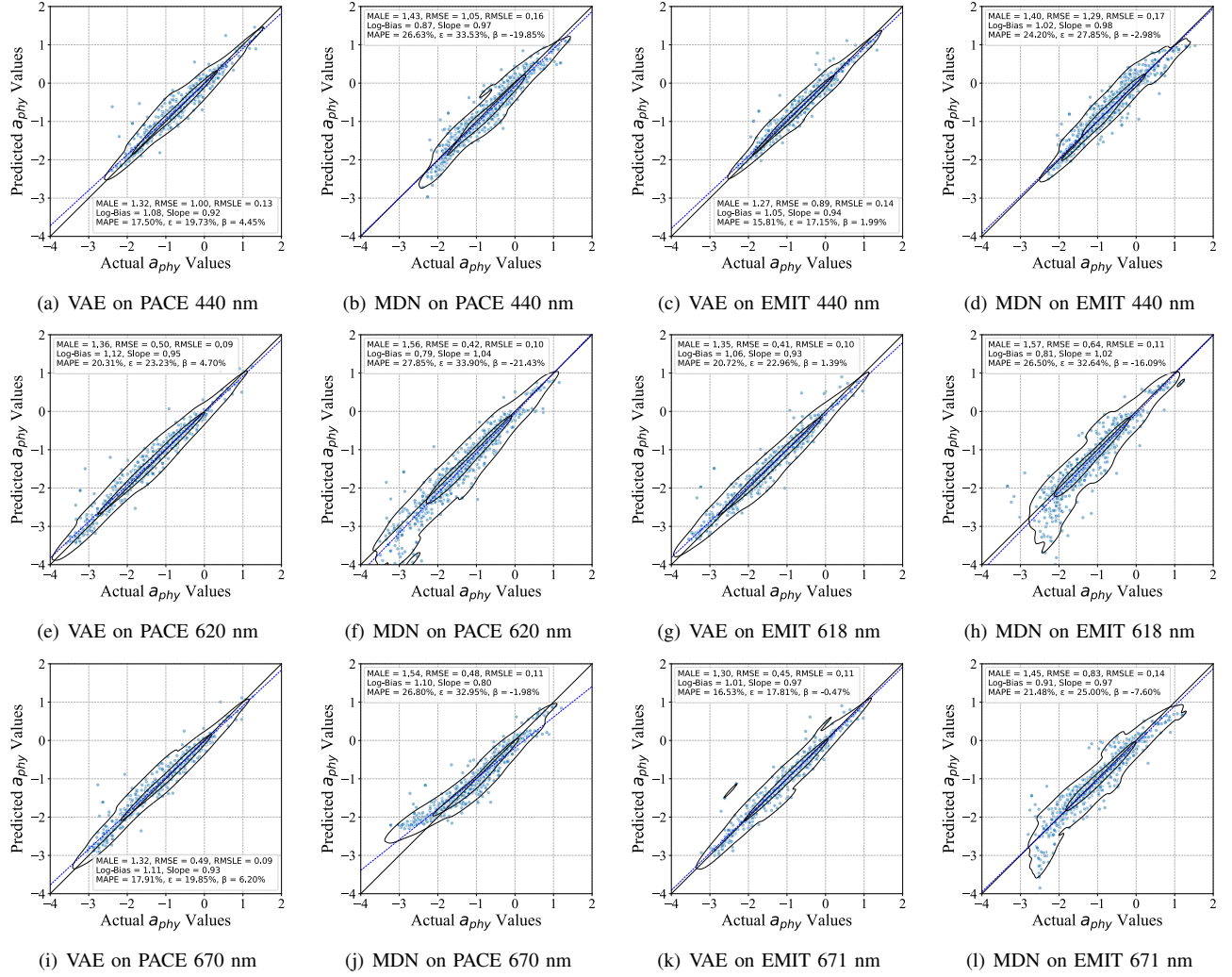


Fig. 4. Scatter plots and evaluation metrics for VAE and MDN predictions of  $a_{phy}$  at three representative wavelengths for PACE and EMIT. (a)–(c) VAE for PACE and EMIT at 440 nm, (d)–(f) MDN for PACE and EMIT at 440 nm, (g)–(i) VAE for PACE and EMIT at 620 nm, and (j)–(l) MDN for PACE and EMIT at 670 nm.

wavelengths, with results showing in Figure 5 and Figure 6. Examining RMSE trends for PACE, which are shown in Figs. 5(a) and 5(d), both VAE and MDN models show comparable performance across most wavelengths. While MDN exhibits slightly higher RMSE values in the shorter wavelength range (400–500 nm), within the 600–700 nm range, RMSE values for both models remain stable. In contrast, for EMIT, VAE maintains an RMSE under 1.0, whereas MDN peaks at values slightly above 2.0 at shorter wavelengths, indicating higher prediction errors for MDN in these regions. At longer wavelengths (600–700 nm), VAE maintains a smooth trend, with RMSE values stabilizing below 0.5, whereas MDN exhibits mild fluctuations, remaining below 0.8. In terms of Log-Bias, VAE- $a_{phy}$  also highlights the stability of the model for both PACE and EMIT. For instance, the VAE model maintains Log-Bias values close to the ideal value of 1 across the entire wavelength range (400–700 nm), demonstrating its ability to produce unbiased estimates. In contrast, the MDN model exhibits certain fluctuations, particularly at longer wavelengths (600–700 nm), where deviations from 1 become increasingly

pronounced. Lastly, the VAE model exhibits more stable and lower  $\beta$  values compared to MDN across both PACE and EMIT. For VAE,  $\beta$  fluctuations remain within  $\pm 10\%$  across the spectrum for both sensors, demonstrating its robust predictive capability. In contrast, MDN experiences substantial variations, exceeding  $\pm 50\%$  at several wavelengths, with a notable increase in deviation in longer wavelengths (600–700 nm). When comparing PACE and EMIT, MDN exhibits larger  $\beta$  deviations for PACE, particularly at wavelengths beyond 600 nm. This suggests that MDN struggles more with PACE's spectral characteristics, likely due to its higher spectral resolution, leading to greater instability in its predictions. Additionally, while EMIT shows moderate fluctuations in MDN's  $\beta$  values at shorter wavelengths (400–500 nm), PACE exhibits more pronounced variations across the spectrum, reinforcing the idea that MDN is less stable for PACE than EMIT. Despite these differences, VAE maintains consistently low fluctuations across both datasets, which is crucial in hyperspectral applications where each wavelength provides critical input to construct the spectra. Collectively, these results demonstrate that



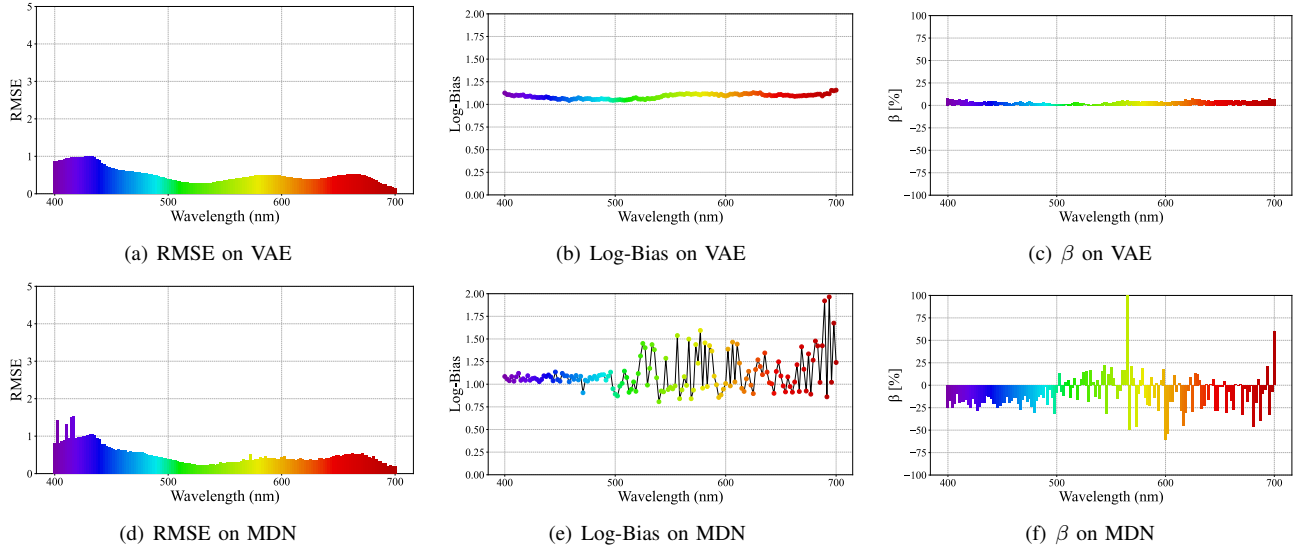


Fig. 5. Performance of VAE in terms of RMSE, Log-Bias, and  $\beta$  for  $a_{phy}$  prediction on PACE wavelengths across 400nm-700nm.

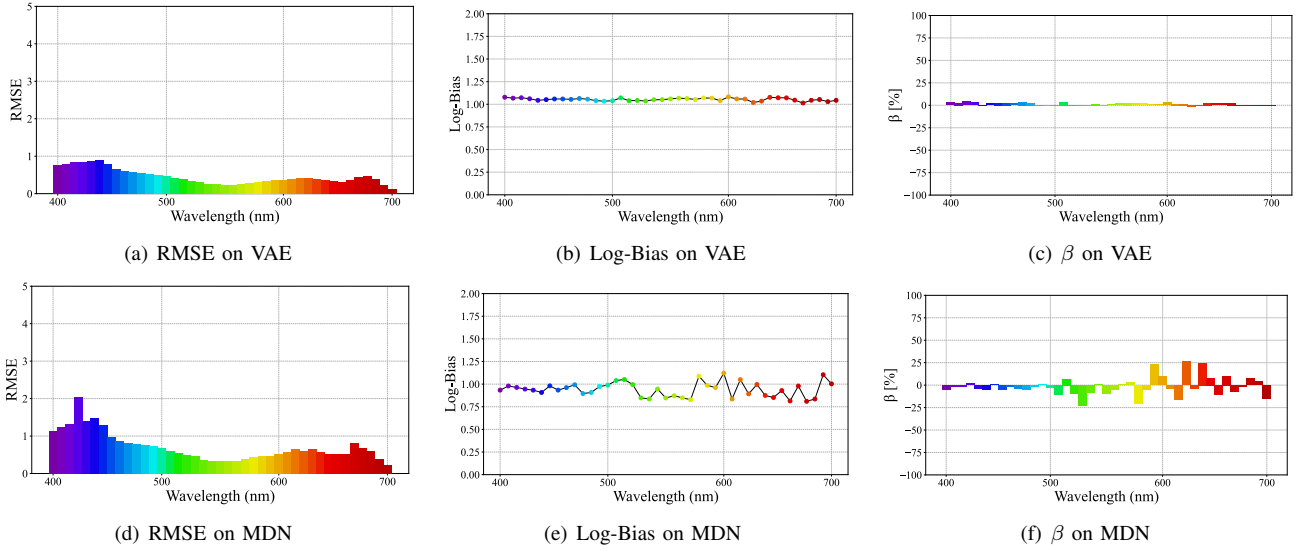


Fig. 6. Performance of VAE in terms of RMSE, Log-Bias, and  $\beta$  for  $a_{phy}$  prediction with EMIT spectral setting across 400nm-700nm.

VAE is a suitable framework for PACE and EMIT, as it offers a more stable, precise, and unbiased estimation of  $a_{phy}$  values across their spectral settings. Accuracy across wavelengths is crucial because detailed and precise  $a_{phy}$  spectral predictions are key to accurately determining phytoplankton community composition (PCC). The ability of VAE to maintain consistent Log-Bias and low  $\beta$  variations also underscores its potential for future hyperspectral remote sensing missions, such as GLIMR, and SBG, where VAE's spectral settings can be easily tuned and adapted to meet mission-specific configurations.

Moreover, both VAE and MDN models are also applied to the same randomly selected testing  $Rrs_{a_{phy}}$  spectra, resampled at PACE and EMIT wavelength settings, to compare the predicted and actual  $a_{phy}$  spectra. We also include a comparison with M-MDN for  $a_{phy}$  predictions, which samples  $a_{phy}$  from the learned mixture of Gaussian distributions instead of selecting the Gaussian mean with the highest weight as the output to handle one-to-many estimation problems. Fig. 7

presents the actual and predicted spectra within the 400-700 nm range with Figures 7(a)-(c) and Figures 7(d)-(f) at PACE and EMIT spectral settings, respectively.

The results demonstrate that the VAE- $a_{phy}$  (Figures 7(a) and (d)) provide superior predictions, as the estimated  $a_{phy}$  curves (blue) closely align with the actual values (red) across the entire spectral range. In contrast, while the MDN model (Figures 7(b) and (e)) preserves general spectral shape, it exhibits slight deviations between the predicted and actual spectra, which is most evident at shorter wavelengths in the blue range (400-500 nm). The M-MDN model (Figures 7(c) and (f)) exhibits the highest level of instability, with substantial fluctuations and increased prediction noise across all wavelengths, particularly where high-frequency noise is more pronounced. These results suggest that M-MDN struggles to generalize effectively, introducing large uncertainties in  $a_{phy}$  predictions when sampling  $a_{phy}$  from the learned mixture of Gaussian distributions to address one-to-many issues. The

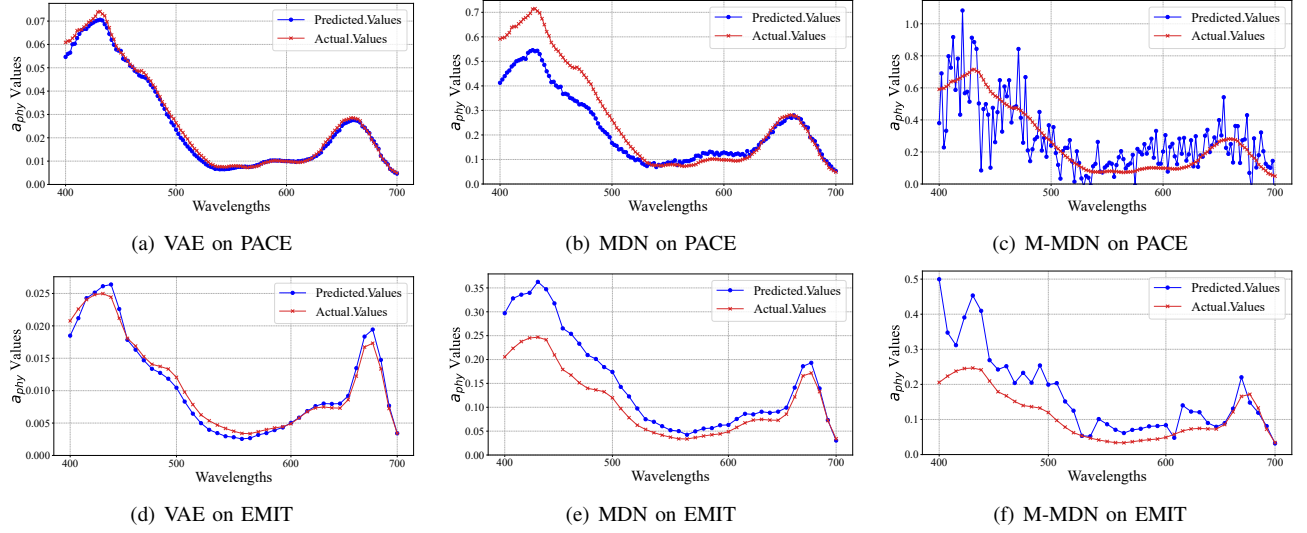


Fig. 7. Comparison of actual and predicted  $a_{phy}$  spectra in the 400–700 nm range: (a)–(c) show results for VAE, MDN, and M-MDN using PACE spectral setting, while (d)–(f) present corresponding results using EMIT wavelengths.

deterioration in  $a_{phy}$  predictions from M-MDN is particularly pronounced at PACE wavelengths, which have a higher spectral resolution (2.5 nm) and consequently higher input and output dimensions. M-MDN struggles with high-dimensional data, whereas VAE- $a_{phy}$  effectively maps high-dimensional data into a lower-dimensional latent variable space, leading to a better performance on PACE wavelengths. This result underscores the reliability and accuracy of VAE- $a_{phy}$  model across varying wavelength conditions, highlighting its novelty in the applications of high-dimensional data predictions.

### C. Evaluations of Chl-a Predictions

In this section, we present the performance of VAE-Chl-a and MDN in predicting Chl-a using the  $R_{rs}$  PACE and EMIT spectral settings, respectively, in Fig. 8. It is observed that the VAE-Chl-a model slightly outperforms the MDN model across all evaluation metrics. First, VAE-Chl-a model exhibits lower prediction errors, as evidenced by overall lower values in MALE, RMSE, and RMSLE. For example, as shown in Figs 8(a), which are based on  $R_{rs}$  at PACE wavelengths, the VAE-Chl-a model shows a MALE of 1.47, an RMSE of 105.28, and an RMSLE of 0.20, while the MDN model yields an MALE of 1.52, an RMSE of 114.54, and an RMSLE of 0.22. Additionally, the prediction deviation of VAE-Chl-a model is better than that of the MDN model, as rendering Log-bias and slope values are closer to 1. However, unlike the substantial advantage of VAE for  $a_{phy}$  predictions, VAE-Chl-a model achieves only minor improvements over the MDN model. This is because the prediction for Chl-a is a single-value prediction whereas  $a_{phy}$  predictions involve high-dimensional spectral vectors. Consequently, the advantage of the VAE in handling high-dimensional prediction is not as evident in this VAE-Chl-a case. Despite this, the slight improvement still demonstrates the robustness of the VAE model as an ideal framework for adoption in the ocean color field for the inversion retrievals of IOPs and optical active biogeochemical parameters.

### D. Comparisons of Generality for $a_{phy}$ Predictions

We further assess the robustness and generality of the trained VAE- $a_{phy}$  and M-MDN models by applying them to a separate in-situ dataset from Galveston Bay, TX, USA, which was not included in the training process. This assessment verifies the models' ability to generalize  $a_{phy}$  estimation to unseen data. This in-situ dataset from Galveston Bay was collected one and two months after Hurricane Harvey in 2017, during which the optical properties were highly complex. The one-month post-hurricane data represents turbid waters, while the two-month dataset captures productive waters with high Chl-a concentrations [33]. Figs. 9 and 10 present three examples of actual and predicted  $a_{phy}$  spectra across the 400–700 nm range at PACE and EMIT spectral settings, evaluating the ability of VAE and M-MDN models to generalize across different unseen geographic locations. We consistently observe robust and stable estimations from VAE- $a_{phy}$  model in Figures 9 and 10. In addition to capturing the two pronounced Chl-a absorption peaks at 440 nm and 670 nm, the model effectively captures subtle variations in the range of 450–500 nm, a critical spectral region for differentiating phytoplankton groups. This also holds true for EMIT, which has a slightly lower spectral resolution (7.4 nm). Moreover, minor variations along the  $a_{phy}$  spectra, such as those in the 480–500 nm and 600–620 nm ranges, were accurately predicted (Figure 10). However, predictions for EMIT show slightly greater deviations in the 500–600 nm range compared to PACE. Overall, the  $a_{phy}$  predictions from the VAE- $a_{phy}$  model closely align well with the actual values across all wavelengths, whereas M-MDN predictions exhibit zigzag patterns, indicating instability. This difference in performance stems from the VAE's learning structure, which does not directly fit the  $a_{phy}$  distribution from the training data but instead learns a Gaussian distribution to capture the high-level patterns in the  $R_{rs}$ - $a_{phy}$  relationship for predictions. These experimental results highlight the practical advantages of using VAE, such as improved generality to unseen data and greater stability across wavelengths, in

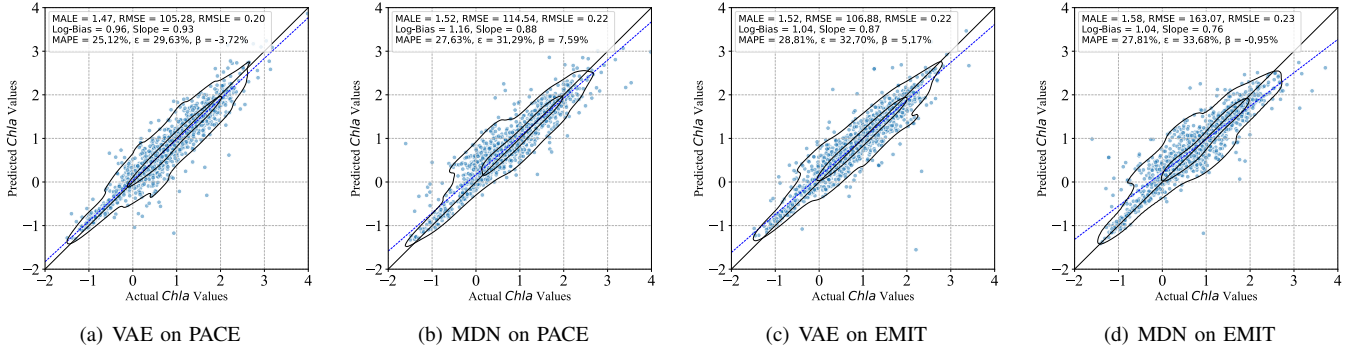


Fig. 8. Evaluation metrics for VAE and MDN predictions of Chl-a using PACE and EMIT spectral settings. (a), (c) VAE for PACE and EMIT, and (b), (d) MDN for PACE and EMIT

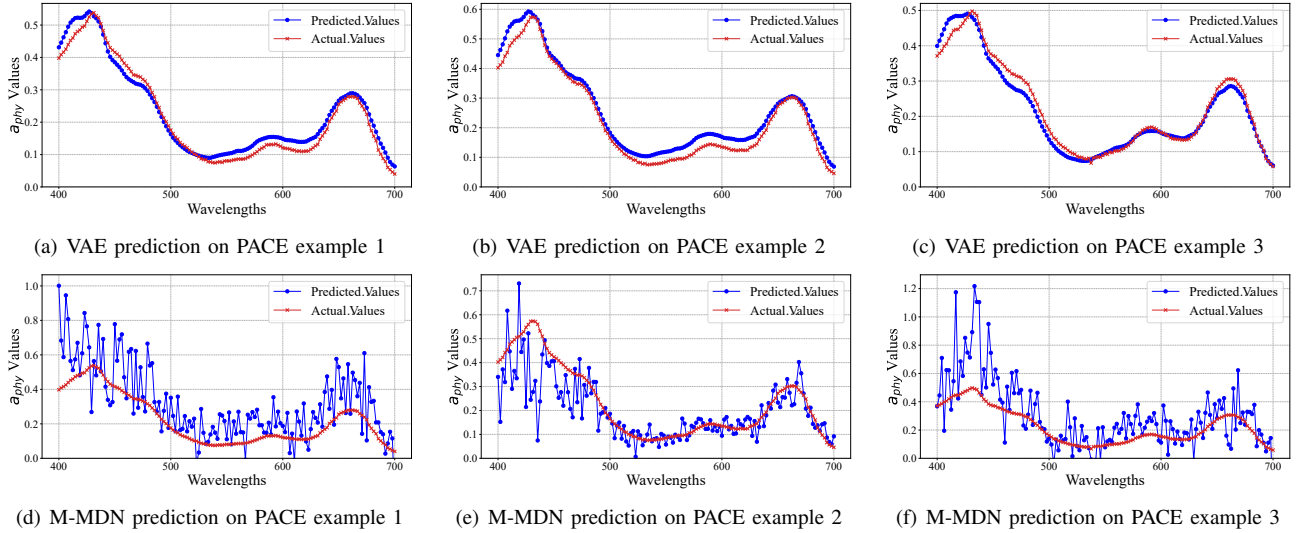


Fig. 9. Comparison of actual and predicted  $a_{phy}$  spectra at PACE wavelength setting in the range of 400–700 nm, using (a)–(c) VAE and (d)–(f) M-MDN..

addressing one-to-many inversion challenges in ocean color applications.

## V. DISCUSSION AND FUTURE WORK

In this section, we discuss the advantages and reasoning behind employing the VAE structure for predicting  $a_{phy}$  and Chl-a, as well as its potential for future applications in ocean color remote sensing.

### A. Balancing Uncertainty and Reliability for Multi-value Prediction

Predicting  $a_{phy}$  and Chl-a through hyperspectral  $R_{rs}$  data involves addressing one-to-many problems, which stems from the fact that for a given  $R_{rs}$  spectrum, the inverse function of  $R_{rs}$  is not unique as a single  $R_{rs}$  may correspond to multiple combinations of IOPs and associated concentrations [12], [55], especially when the spectral resolution is limited. While MDN-based approaches can be rectified to support one-to-many predictions, such as the M-MDN method described in Section II-D, obtaining a specific prediction requires random sampling from the learned distribution. This process inherently introduces additional uncertainty. For instance, when sampling from a mixture of Gaussian distributions with multiple

peaks, the resulting predictions can vary significantly. This variability makes it challenging to achieve consistent and reliable estimates for  $a_{phy}$  and Chl-a, as different samples may lead to fluctuating predicted values, reducing the model's overall stability. In contrast, VAEs can solve the one-to-many prediction problem and avoid the uncertainty associated with the distribution nature. This can be explained from two perspectives. First, the encoder learns the latent vectors from the probability distribution learned from the input data  $R_{rs}$  and introduces randomness through the reparameterization operation [44], [6], [14], as shown in Eq. (4), enabling the VAE to generate different outputs from the same input. Second, the decoder directly learns the mapping between the learned variable latent vectors and the derived prediction data, allowing VAEs to generate accurate predictions from the latent distribution. This approach reduces the randomness and uncertainty associated with the sampling process, resulting in more reliable predictions.

### B. Handling Complex Data Distributions

Phytoplankton absorption coefficient  $a_{phy}$  is one of the IOPs with the most complex distributions, often varying significantly across geographic locations. Figures 11(a)–(b)

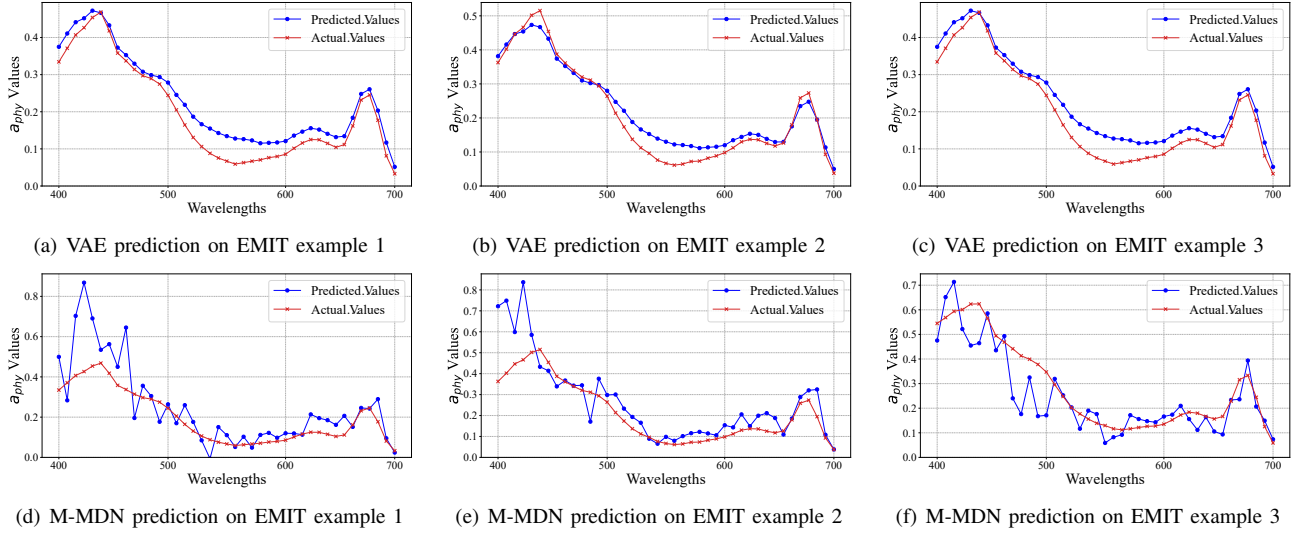


Fig. 10. Comparison of actual and predicted  $a_{phy}$  spectra at EMIT wavelength setting in the range of 400–700 nm, using (a)–(c) VAE and (d)–(f) M-MDN.

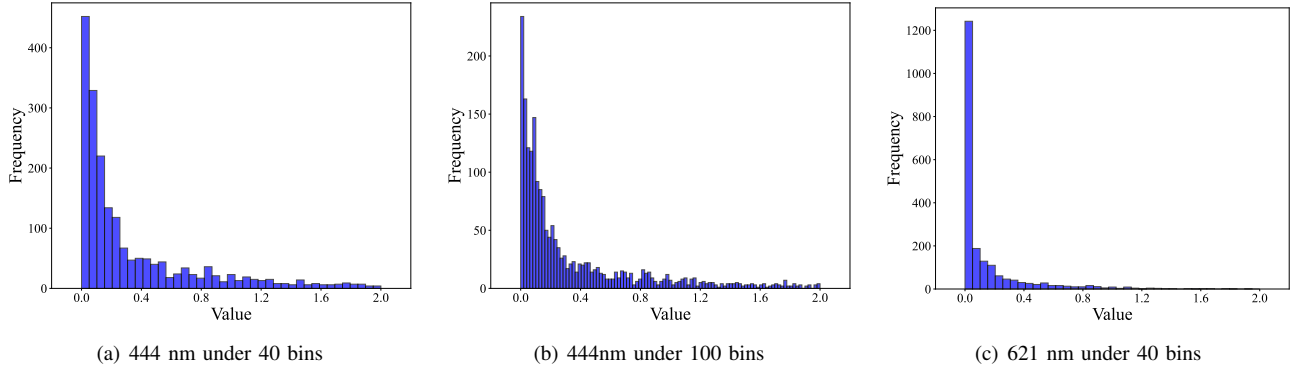


Fig. 11. The  $a_{phy}$  ( $m^{-1}$ ) distribution at different wavelengths.

display the  $a_{phy}$  data distributions at 444 nm under different bin settings, where larger bin values indicate finer resolution of data variability, capturing multimodal distributions and avoiding oversimplification. When comparing Figs. 11(a) and 11(b), it is found that higher data precision leads to more peaks in the  $a_{phy}$  distribution, suggesting that fitting the  $a_{phy}$  data distribution at different scenarios may require different numbers of Gaussian components. Moreover, it is observed from Figs. 11(a) and 11(c) that the  $a_{phy}$  distribution also varies significantly across different wavelengths due to absorption peaks of specific pigments. As such, the MDN, which aims to fit  $a_{phy}$  distributions at different wavelengths and Chl-a distributions, with a finite and fixed number of Gaussian distributions, may fail to produce accurate predictions in practice. In contrast, VAEs use Gaussian distributions to represent each element in the latent vector, aiming at capturing high-level patterns in the  $R_{rs}$  spectra for  $a_{phy}$  and Chl-a prediction. Despite the complex distributions of  $a_{phy}$  spectra, the underlying latent patterns can be more reasonably assumed to follow Gaussian distributions [4], [22]. By learning these latent patterns, VAE effectively captures the high-level inherent relationship between  $R_{rs}$  and  $a_{phy}$  or Chl-a, eliminating the need to assume and fit mixed Gaussian distributions for

target prediction. This enables VAEs to model intricate data distributions more efficiently and flexibly. In addition, it also supports VAE models in maintaining acceptable performance in prediction tasks on unseen data, even when the distributions differ from those of the training data, as shown in Figs. 9 and 10.

### C. Robustness to Noise

The training data, i.e.,  $R_{rs}$ ,  $a_{phy}$ , and Chl-a typically contain a certain level of noise. While data quality control and preprocessing help filter out clearly abnormal values, a certain degree of noise could remain inevitable due to in-situ sampling variability. MDN models the relationship between  $R_{rs}$  and  $a_{phy}$ /Chl-a by fitting a mixture of Gaussian distributions to the training data. However, when noise is present, MDN tends to fit Gaussian components to all observed variations, including those introduced by noise [1], [47]. Since the model does not inherently distinguish between meaningful trends and random fluctuations, it may allocate Gaussian components to spurious variations rather than true signal patterns. This results in overfitting, where the learned distribution reflects noise-induced variability instead of capturing the underlying physical relationships. Consequently, MDN predictions can become

unstable, especially when applied to new data with different noise characteristics. In contrast, VAE-based approaches are more robust to noisy data for two reasons. First, KL divergence loss, represented as the second term in Eqn. (11), served as a regularization term. It constrains the learned latent distribution to remain close to a prior distribution, which helps prevent the latent vector from overfitting the noise [4]. Additionally, the reparameterization trick naturally introduces noise to the latent vector, as shown in Eq. (4), enabling the VAEs to inherently generate accurate predictions under noisy training data [6], [14]. These mechanisms help VAEs maintain stability and robustness, producing more consistent predictions for  $a_{phy}$  and Chl-a even in noisy environments.

#### D. Generalizing to High-Dimensional Data

In this study, VAE- and MDN-based learning frameworks are applied for  $a_{phy}$  predictions for both the PACE and EMIT missions. EMIT has a lower-dimensional input space ( $N=41$ ) compared to PACE, which has 141 spectral bands, making  $a_{phy}$  predictions the highest-dimensional to date. This further emphasizes the need for robust learning frameworks capable of handling such high-dimensional data. Since PACE was just launched on February 8, 2024, no prior studies have yet been published exploring high-dimensional  $a_{phy}$  predictions using learning frameworks. For example, in previous studies [41], [42], [40], hyperspectral data from HICO and PRISMA required at most 50-dimensional inputs for  $a_{phy}$  prediction tasks and 67-dimensional inputs for phycocyanin prediction tasks. With the higher spectral resolution of PACE-OCI, VAEs, and MDNs are now expected to handle even higher-dimensional data input, enabling accurate predictions across finer wavelength intervals. However, MDNs encounter computational challenges in this setting. Since MDNs model the output as a mixture of multivariate Gaussian distributions, they need to learn a full covariance matrix for each Gaussian component. Specifically, for the input of  $N$  dimensions with  $M$  Gaussian components, the MDN must estimate an  $M \times N \times \text{times}$  covariance matrix, meaning the number of parameters grows quadratically with respect to  $N$ . This will incur computational inefficiencies and challenges in model stability, primarily due to two key factors. First, as the dimensionality increases, the covariance matrix often becomes ill-conditioned or nearly singular, making the model highly sensitive to small numerical errors. These small numerical errors can result in inconsistent and unreliable parameter updates, leading to exploding or vanishing gradients during training. Second, the full covariance matrix introduces  $O(N^2)$  parameters per component, significantly increasing the complexity of the loss function. This expansive parameter space makes it challenging for gradient-based optimizers to find stable updates, often resulting in slow convergence or divergence. In contrast, VAEs mitigate these challenges by mapping high-dimensional data into a latent space, where each dimension follows a simple Gaussian distribution. Rather than modeling full covariance matrices, VAEs learn independent latent features that capture essential spectral patterns, allowing for effective feature extraction and dimensionality reduction while preserving crucial spectral

information. This structure significantly reduces the number of parameters to optimize and eliminates the numerical instability issues associated with MDNs. In our experiments, conducted on a workstation equipped with an Intel Core i9-13900K CPU and an NVIDIA RTX 4090 GPU (24GB VRAM), training the VAE for 2000 epochs took approximately 30 seconds, compared to roughly 500 seconds for MDN model with full covariance matrix. Given these advantages, VAEs present a more scalable alternative for hyperspectral inversion retrieval tasks, such as  $a_{phy}$  and Chl-a prediction explored in this study.

#### E. Forward Path

Moving forward, the VAE framework introduced in this study will be applied to PACE and EMIT hyperspectral imagery; however, evaluating its performance on these satellite datasets remains an ongoing effort. PACE, launched on February 8, 2024, provides daily global hyperspectral imagery, though corresponding field validation is still in progress. EMIT, on the other hand, has conducted irregular sampling across estuaries worldwide, making consistent match-up sampling a challenge—not only for this study but on a global scale. As these missions are still in their early stages, extensive field validation is required for both PACE and EMIT to better understand the uncertainties involved in estimating  $a_{phy}$  and Chl-a from satellite-derived  $R_{rs}$  using learning-based models. Importantly, these challenges do not conflict with the current work presented in this study; rather, they represent parallel and complementary efforts. As satellite  $R_{rs}$  products continue to be calibrated and validated, the proposed algorithms can be more effectively applied, particularly as more paired *in situ*  $R_{rs}$ - $a_{phy}$  datasets along with other IOPs become available on the same day as satellite overpasses (e.g., from PACE and EMIT).

On the other hand, while machine learning approaches are increasingly favored in hyperspectral ocean color remote sensing, there is a persistent need for large training datasets encompassing diverse water quality conditions. Consequently, there is a growing demand for more coincident reflectance-biophysical parameter pairs to be acquired across different water types, environmental conditions, and geographic locations. Unfortunately, the currently available in-situ  $R_{rs}$ - $a_{phy}$  and  $R_{rs}$ -Chl-a datasets are limited and contain uncertainties. This is because the data have been collected by various groups using different instruments and calibration or correction approaches, all of which contribute to uncertainties in machine learning predictions. Additionally, current available global coastal bio-optical datasets—such as  $R_{rs}$ - $a_{phy}$  and  $R_{rs}$ -Chl-a in GLORIA—lack true one-to-many features. To further enhance our VAE model's generality across different geo-locations and water types, two lines of direction can be explored. First, considering the difficulty of enriching the real field data samples, the radiative transfer model (RTM), traditionally used in ocean color remote sensing to simulate apparent optical properties (AOPs), such as  $R_{rs}$ , based on precise characterization of IOPs, offers a method to generate a large dataset to complement the real data. The simulated  $R_{rs}$  over global coastal waters using RTM, such as HydroLight,



employing hyperspectral  $a_{phy}$ , and scattering phase functions measured from diverse phytoplankton groups/genus [28], [36] to represent the dynamic and varied phytoplankton community composition, along with varying concentrations of CDOM and NAP independent of phytoplankton in a wide range of oceanic and coastal waters, is expected to capture detailed spectral signatures in response to phytoplankton community variability in biogeochemically dynamic and optically complex coastal-estuarine environments. With a large and high-quality simulation dataset, a plausible solution is to first use it to pre-train the VAE model, allowing it to learn the abundant patterns, and then fine-tune the model using real data samples. This approach is expected to improve both model prediction accuracy and generalizability. Secondly, the simulated data would further support the extension of VAE-based models into a multi-head framework, enabling simultaneous prediction of multiple parameters, similar to the prior work [40]. A complete set of IOPs and AOPs would allow multiple decoders to be trained concurrently, with each decoder dedicated to predicting a single IOP. Such a design preserves the benefits of multi-head prediction, allowing a single encoder to capture relationships among different IOPs while minimizing error propagation through independent decoders. Nonetheless, further research is needed from both data and algorithm perspectives for future exploration. to refine the VAE architecture used in this study.

## VI. CONCLUSION

In this paper, VAEs—originally designed for data generation and reconstruction—were tailored for the first time to prediction tasks in ocean color remote sensing. Two VAE-based models, VAE- $a_{phy}$  for high-dimension predictions and VAE-Chl-a for single-value predictions in optically-complex waters, targeting two key indicators of phytoplankton, i.e., Chl-a for abundance and  $a_{phy}$  for diversity. The dataset used by [41] and [40] for the MDN model was utilized for VAE model development, validation, and testing in this study. These models were applied to support NASA’s hyperspectral missions, including PACE and EMIT. The configurations of both the VAE and MDN models were provided in this study to benefit the broader community. The performance of the VAE and MDN models was evaluated using eight metrics—MALE, RMSE, RMSLE, Log-Bias, Slope, and three median value-oriented metrics—to provide a comprehensive assessment of prediction accuracy. Overall, the VAE presented greater stability and robustness in predicting  $a_{phy}$  across the PACE and EMIT spectral settings, demonstrating higher precision and lower bias in all eight metrics. The benefits of applying VAEs to hyperspectral data for high-dimensional predictions, such as PACE-predicted  $a_{phy}$ , were discussed in depth across three aspects: uncertainty and reliability, robustness to noise, and generalization to high-dimensional data. Additionally, a new dataset collected post-Hurricane Harvey in 2017, which presented different conditions and distributions compared to the training data, was used as a test dataset. Both the VAE and MDN models perform well in the predictions of  $a_{phy}$  at longer wavelengths, with the VAE-based solution demonstrating better robustness in the range of 400-500 nm when

applied to unseen data collected in Galveston Bay. Moving forward, using simulated data to expand the training dataset, representing diverse coastal waters would be one of the key solutions to overcoming data scarcity issues and enhancing model generalization for  $a_{phy}$  predictions. This approach will also further enhance support for the multi-head predictions. This study advances hyperspectral retrievals of phytoplankton metrics, paving the way for our understanding of phytoplankton community dynamics in optically complex coastal waters.

## ACKNOWLEDGMENTS

The authors would like to acknowledge the valuable comments provided by two anonymous reviewers, which greatly helped improve the quality of this manuscript. We also extend our gratitude to Dr. Ryan O’Shea for graciously sharing the data that made this work possible. All source code associated with this study is publicly available on GitHub at <https://github.com/CASELabCode/Jiadong-HyperVAE.git>. We encourage the community to explore, utilize, and contribute to the codebase to further advance research in this area.

## REFERENCES

- [1] Sundarabalan V Balasubramanian, Ryan E O’Shea, Arun M Saranathan, Christopher C Begeman, Daniela Gurlin, Caren Binding, Claudia Giardino, Michelle C Tomlinson, Krista Alikas, Kersti Kangro, et al. Mixture density networks for re-constructing historical ocean-color products over inland and coastal waters: demonstration and validation. *Frontiers in Remote Sensing*, 6:1488565, 2025.
- [2] Sundarabalan V Balasubramanian, Nima Pahlevan, Brandon Smith, Caren Binding, John Schalles, Hubert Loisel, Daniela Gurlin, Steven Greb, Krista Alikas, Mirjam Randla, et al. Robust algorithm for estimating total suspended solids (tss) in inland and nearshore coastal waters. *Remote Sensing of Environment*, 246:111768, 2020.
- [3] Thomas S Bianchi, Steven F DiMarco, JH Cowan Jr, Robert D Hetland, P Chapman, JW Day, and MA Allison. The science of hypoxia in the northern gulf of mexico: a review. *Science of the Total Environment*, 408(7):1471–1484, 2010.
- [4] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.
- [5] Annick Bricaud, Hervé Claustre, Josephine Ras, and Kadija Oubelkheir. Natural variability of phytoplanktonic absorption in oceanic waters: Influence of the size structure of algal populations. *Journal of Geophysical Research: Oceans*, 109(C11), 2004.
- [6] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.
- [7] BB Cael, Alison Chase, and Emmanuel Boss. Information content of absorption spectra and implications for ocean color inversion. *Applied Optics*, 59(13):3971–3984, 2020.
- [8] Ivona Cetinić, Cecile S Rousseaux, Ian T Carroll, Alison P Chase, Sasha J Kramer, P Jeremy Werdell, David A Siegel, Heidi M Dierssen, Dylan Catlett, Aimee Neeley, et al. Phytoplankton composition from space: Requirements, opportunities, and challenges. *Remote Sensing of Environment*, 302:113964, 2024.
- [9] Alison P Chase, Sasha J Kramer, Nils Haëntjens, Emmanuel S Boss, Lee Karp-Boss, Mimi Edmondson, and Jason R Graff. Evaluation of diagnostic pigments to estimate phytoplankton size classes. *Limnology and Oceanography: Methods*, 18(10):570–584, 2020.
- [10] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.
- [11] Aurea M Ciotti, Marlon R Lewis, and John J Cullen. Assessment of the relationships between dominant cell size in natural phytoplankton communities and the spectral shape of the absorption coefficient. *Limnology and Oceanography*, 47(2):404–417, 2002.
- [12] Michael Defoin-Platel and Malik Chami. How ambiguous is the inverse problem of ocean color in coastal waters? *Journal of Geophysical Research: Oceans*, 112(C3), 2007.

- [13] Heidi M Dierssen, Steven G Ackleson, Karen E Joyce, Erin L Hestir, Alexandre Castagna, Samantha Lavender, and Margaret A McManus. Living up to the hype of hyperspectral aquatic remote sensing: science, resources and outlook. *Frontiers in Environmental Science*, 9:649528, 2021.
- [14] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
- [15] Eurico J D'Sa, Ishan Joshi, and Bingqing Liu. Galveston bay and coastal ocean optical-geochemical response to hurricane harvey from viirs ocean color. *Geophysical Research Letters*, 45(19):10–579, 2018.
- [16] Eurico J D'Sa, Ishan D Joshi, Bingqing Liu, Dong S Ko, Christopher L Osburn, and Thomas S Bianchi. Biogeochemical response of apalachicola bay and the shelf waters to hurricane michael using ocean color semi-analytic/inversion and hydrodynamic models. *Frontiers in Marine Science*, 6:523, 2019.
- [17] Anatoly Gitelson. The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration. *International Journal of Remote Sensing*, 13(17):3367–3373, 1992.
- [18] Herman J Gons, Machteld Rijkeboer, and Kevin G Ruddick. A chlorophyll-retrieval algorithm for satellite imagery (medium resolution imaging spectrometer) of inland and coastal waters. *Journal of Plankton Research*, 24(9):947–951, 2002.
- [19] Robert Green. The nasa earth surface mineral dust source investigation. *43rd COSPAR Scientific Assembly. Held 28 January-4 February*, 43:37, 2021.
- [20] Daniela Gurlin, Anatoly A Gitelson, and Wesley J Moses. Remote estimation of chl-a concentration in turbid productive waters—return to a simple two-band nir-red model? *Remote Sensing of Environment*, 115(12):3479–3490, 2011.
- [21] Stephanie A Henson, BB Cael, Stephanie R Allen, and Stephanie Dutkiewicz. Future phytoplankton diversity in a changing climate. *Nature communications*, 12(1):5372, 2021.
- [22] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.
- [23] Paul R Hill, Anurag Kumar, Marouane Temimi, and David R Bull. Habnet: Machine learning, remote sensing-based detection of harmful algal blooms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3229–3239, 2020.
- [24] Alia L Khan, Heidi M Dierssen, Ted A Scambos, Juan Höfer, and Raul R Cordero. Spectral characterization, radiative forcing and pigment content of coastal antarctic snow algae: approaches to spectrally discriminate red and green communities and their impact on snowmelt. *The Cryosphere*, 15(1):133–148, 2021.
- [25] Tyler King, Carl Legleiter, Natalie Hall, Adam Mumford, Kurt Carpenter, and E Slonecker. Spectral mixture analysis for surveillance of harmful algal blooms (smash): A hybrid laboratory-and satellite-based approach to identify algal taxa in inland waterbodies via remote sensing. In *AGU Fall Meeting Abstracts*, volume 2021, pages GH15E–0628, 2021.
- [26] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.
- [27] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.
- [28] Lisl Robertson Lain, Jeremy Kravitz, Mark Matthews, and Stewart Bernard. Simulated inherent optical properties of aquatic particles using the equivalent algal populations (eap) model. *Scientific Data*, 10(1):412, 2023.
- [29] ZhongPing Lee, Kendall L Carder, and Robert A Arnone. Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters. *Applied optics*, 41(27):5755–5772, 2002.
- [30] Moritz K Lehmann, Daniela Gurlin, Nima Pahlevan, Krista Alikas, Ted Conroy, Janet Anstee, Sundarabalan V Balasubramanian, Cláudio CF Barbosa, Caren Binding, Astrid Bracher, et al. Gloria-a globally representative hyperspectral in situ dataset for optical sensing of water quality. *Scientific data*, 10(1):100, 2023.
- [31] Yunmei Li, Qiao Wang, Chuangqing Wu, Shaohua Zhao, Xing Xu, Yanfei Wang, and Changchun Huang. Estimation of chlorophyll a concentration using nir/red bands of meris and classification procedure in inland turbid water. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3):988–997, 2011.
- [32] Bingqing Liu, Eurico J D'Sa, and Ishan Joshi. Multi-decadal trends and influences on dissolved organic carbon distribution in the barataria basin, louisiana from in-situ and landsat/modis observations. *Remote Sensing of Environment*, 228:183–202, 2019.
- [33] Bingqing Liu, Eurico J D'Sa, and Ishan D Joshi. Floodwater impact on galveston bay phytoplankton taxonomy, pigment composition and photo-physiological state following hurricane harvey from field and ocean color (sentinel-3a olci) observations. *Biogeosciences*, 16(9):1975–2001, 2019.
- [34] Bingqing Liu, Eurico J D'Sa, Kanchan Maiti, Victor H Rivera-Monroy, and Zuo Xue. Biogeographical trends in phytoplankton community size structure using adaptive sentinel 3-olci chlorophyll a and spectral empirical orthogonal functions in the estuarine-shelf waters of the northern gulf of mexico. *Remote Sensing of Environment*, 252:112154, 2021.
- [35] Bingqing Liu, Eurico J D'Sa, Francesca Messina, Melissa M Baustian, Kanchan Maiti, Victor H Rivera-Monroy, Wei Huang, and Ioannis Y Georgiou. Dissolved organic carbon dynamics and fluxes in mississippi-atcatalaya deltaic system impacted by an extreme flood event and hurricanes: a multi-satellite approach using sentinel-2/3 and landsat-8/9 data. *Frontiers in Marine Science*, 10:1159367, 2023.
- [36] Michael W Lomas, Aimee R Neeley, Ryan Vandermeulen, Antonio Mannino, Crystal Thomas, Michael G Novak, and Scott A Freeman. Phytoplankton optical fingerprint libraries for development of phytoplankton ocean color satellite products. *Scientific Data*, 11(1):168, 2024.
- [37] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.
- [38] Curtis D Mobley. Light and water: radiative transfer in natural waters. (No Title), 1994.
- [39] James L Mueller, A Morel, R Frouin, C Davis, R Arnone, K Carder, ZP Lee, RG Steward, S Hooker, CD Mobley, et al. Ocean optics protocols for satellite ocean color sensor validation, revision 4. volume iii: Radiometric measurements and data analysis protocols. 2003.
- [40] Ryan E O'Shea, Nima Pahlevan, Brandon Smith, Emmanuel Boss, Daniela Gurlin, Krista Alikas, Kersti Kangro, Raphael M Kudela, and Diana Vaičiūtė. A hyperspectral inversion framework for estimating absorbing inherent optical properties and biogeochemical parameters in inland and coastal waters. *Remote Sensing of Environment*, 295:113706, 2023.
- [41] Nima Pahlevan, Brandon Smith, Caren Binding, Daniela Gurlin, Lin Li, Mariano Bresciani, and Claudia Giardino. Hyperspectral retrievals of phytoplankton absorption and chlorophyll-a in inland and nearshore coastal waters. *Remote Sensing of Environment*, 253:112200, 2021.
- [42] Nima Pahlevan, Brandon Smith, John Schalles, Caren Binding, Zhi-gang Cao, Ronghua Ma, Krista Alikas, Kersti Kangro, Daniela Gurlin, Nguyen H, et al. Seamless retrievals of chlorophyll-a from sentinel-2 (msi) and sentinel-3 (olci) in inland and coastal waters: A machine-learning approach. *Remote Sensing of Environment*, 240:111604, 2020.
- [43] Nancy N Rabalais, R Eugene Turner, and William J Wiseman Jr. Hypoxia in the gulf of mexico. *Journal of environmental quality*, 30(2):320–329, 2001.
- [44] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.
- [45] Collin Roesler, Dariusz Stramski, Eurico D'Sa, Rüdiger Röttgers, and Rick A Reynolds. Spectrophotometric measurements of particulate absorption using filter pads. *Washington, DC: IOCCG*, 2018.
- [46] Collin S Roesler and Emmanuel Boss. Spectral beam attenuation coefficient retrieved from ocean color inversion. *Geophysical Research Letters*, 30(9), 2003.
- [47] Jonas Rothfuss, Fabio Ferreira, Simon Boehm, Simon Walther, Maxim Ulrich, Tamim Asfour, and Andreas Krause. Noise regularization for conditional density estimation. *arXiv preprint arXiv:1907.08982*, 2019.
- [48] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [49] GZ Sass, IF Creed, SE Bayley, and KJ Devito. Understanding variation in trophic status of lakes on the boreal plain: A 20 year retrospective using landsat tm imagery. *Remote Sensing of Environment*, 109(2):127–141, 2007.
- [50] Th Schroeder, M Schaale, and J Fischer. Retrieval of atmospheric and oceanic properties from meris measurements: A new case-2 water processor for beam. *International Journal of Remote Sensing*, 28(24):5627–5632, 2007.
- [51] Brandon Smith, Nima Pahlevan, John Schalles, Steve Ruberg, Reagan Errera, Ronghua Ma, Claudia Giardino, Mariano Bresciani, Claudio

Barbosa, Tim Moore, et al. A chlorophyll-a algorithm for landsat-8 based on mixture density networks. *Frontiers in Remote Sensing*, 1:623678, 2021.

- [52] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.
- [53] Daniel Sousa and Christopher Small. Topological generality and spectral dimensionality in the earth mineral dust source investigation (emit) using joint characterization and the spectral mixture residual. *Remote Sensing*, 15(9):2295, 2023.
- [54] Deyong Sun, Yunmei Li, Qiao Wang, Jay Gao, Chengfeng Le, Changchun Huang, and Shaoqi Gong. Hyperspectral remote sensing of the pigment c-phyocyanin in turbid inland waters, based on optical classification. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7):3871–3884, 2013.
- [55] Michael Sydor, Richard W Gould, Robert A Arnone, Vladimir I Haltrin, and Wesley Goode. Uniqueness in remote sensing of the inherent optical properties of ocean water. *Applied optics*, 43(10):2156–2162, 2004.
- [56] Surya Prakash Tiwari and Palanisamy Shanmugam. An evaluation of models for the satellite-estimation of phytoplankton absorption coefficients in coastal/oceanic waters. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(1):364–371, 2013.
- [57] Weijia Wang, Kun Shi, Yibo Zhang, Na Li, Xiao Sun, Dong Zhang, Yunlin Zhang, Boqiang Qin, and Guangwei Zhu. A ground-based remote sensing system for high-frequency and real-time monitoring of phytoplankton blooms. *Journal of Hazardous Materials*, 439:129623, 2022.
- [58] P Jeremy Werdell, Bryan A Franz, Sean W Bailey, Gene C Feldman, Emmanuel Boss, Vittorio E Brando, Mark Dowell, Takafumi Hirata, Samantha J Lavender, ZhongPing Lee, et al. Generalized ocean color inversion model for retrieving marine inherent optical properties. *Applied optics*, 52(10):2019–2037, 2013.
- [59] Hongyan Xi, Martin Hieronymi, Hajo Krasemann, and Rüdiger Röttgers. Phytoplankton group identification using simulated and in situ hyperspectral remote sensing reflectance. *Frontiers in Marine Science*, 4:272, 2017.
- [60] Hongyan Xi, Martin Hieronymi, Rüdiger Röttgers, Hajo Krasemann, and Zhongfeng Qiu. Hyperspectral differentiation of phytoplankton taxonomic groups: a comparison between using remote sensing reflectance and absorption spectra. *Remote Sensing*, 7(11):14781–14805, 2015.
- [61] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.



**Jiadong Lou** (S'20) received the B.S. degree in Internet of Things from the School of Computer Science and Technology, Harbin Institute of Technology, China, in 2018, the M.S. degree in Computer Science from the University of Louisiana at Lafayette, USA, in 2023, and the Ph.D. degree from the Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA. He was the recipient of IEEE/IFIP DSN 2023 Best Paper Award and Distinguished Paper Award, and IEEE CNS 2024 Best Paper Runner-up Award. His

research interests include cybersecurity, Machine Learning, Networking, and data privacy.



**Bingqing Liu** earned her B.S. in Marine Science from Shanghai Ocean University and her Ph.D. in Oceanography from Louisiana State University in 2020. She is currently an Assistant Professor and the SL Mid-Winter Fair Endowed Professor in Renewable Resources at the University of Louisiana at Lafayette. As a satellite oceanographer, her research focuses on coastal carbon cycling and its responses to climate change, extreme events, and human activities. She develops remote sensing algorithms and integrates satellite observations with models to investigate coastal ocean processes. Dr. Liu is a member of NASA's PACE and EMIT science teams, contributing to advancements in ocean color remote sensing.

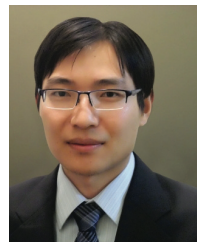


**Yuanheng Xiong** received a B.S. degree in biological science from Shenzhen University, Shenzhen, China, master's degrees in environmental consultancy from Newcastle University, U.K., and in ethics and public policy from Suffolk University, Boston, MA, USA. He received the Ph.D. degree in Earth system science and policy from the University of North Dakota, Grand Forks, ND, USA, in 2023, studying the size distribution and optical properties of oceanic submicron particles. He is currently a Research Associate II with the Division of Marine Science, University of Southern Mississippi, Long Beach, MS, USA, and serves as an early career representative on the 2026 Ocean Sciences Meeting planning committee. His research interests include ocean optics, submicron particle characterization, and radiative transfer simulations in water. He has participated in major field campaigns, including NASA EXPORTS, the NSERC Lake Pulse Network, and Canada's Line P program.



**Xiaodong Zhang** earned a B.S. degree in Computer Science from Nanjing University, China and received a Ph.D. degree in Oceanography from Dalhousie University, Halifax, Canada in 2002. He is currently an Endowed Chair and Professor in Marine Science with the School of Ocean Science and Engineering at The University of Southern Mississippi. He has authored and co-authored nearly 100 peer-reviewed journal articles. His main research interests are in the field of ocean optics, investigating interaction of light and particles to characterize various

water constituents from optical observations. He led the development of the molecular scattering model for seawater that is widely used globally. He is a member of many science teams dedicated to ocean color remote sensing and its applications.



**Xu Yuan** (M'16-SM'22) received the B.S. degree from the College of Information Technology, Nankai University, Tianjin, China, in 2009, and the Ph.D. degree from the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA, in 2016. From 2016 to 2017, he was a Post-Doctoral Fellow of Electrical and Computer Engineering at the University of Toronto, Toronto, ON, Canada. He is currently an Associate Professor in the Department of Computer and Information Sciences at the University of Delaware, Newark, DE, USA. He was the recipient of NSF CRII Award, NSF CAREER Award, IEEE/IFIP DSN 2023 Best Paper Award and Distinguished Paper Award, and IEEE CNS 2024 Best Paper Runner-up Award. His research interest focuses on machine learning and AI, AI for sciences, and cybersecurity.