JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH000851

Key Points:

- An AI-based physically consistent long-term regional emulator has been developed for the Gulf of Mexico region
- A deterministic and stochastic downscaling model has been developed to super-resolve and bias-correct the low-resolution predictions to high resolution
- The performance of the framework has been compared with AI-based and high-resolution physics-based baselines on short- and long-term metrics

Correspondence to:

A. Chattopadhyay,
aschatto@ucsc.edu

Citation:

Lupin-Jimenez, L., Darman, M., Hazarika, S., Wu, T., Gray, M., He, R., et al. (2025). Simultaneous emulation and downscaling with physically consistent deep learning-based regional ocean emulators. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2025JH000851. https://doi.org/10.1029/2025JH000851

Received 26 JUN 2025
Accepted 30 JUL 2025

Author Contributions:

Conceptualization: Leonard Lupin-Jimenez, Ashesh Chattopadhyay
Funding acquisition: Leonard Lupin-Jimenez, Ashesh Chattopadhyay
Methodology: Leonard Lupin-Jimenez, Ashesh Chattopadhyay
Project administration: Leonard Lupin-Jimenez, Ashesh Chattopadhyay
Resources: Leonard Lupin-Jimenez, Ashesh Chattopadhyay
Supervision: Leonard Lupin-Jimenez, Ashesh Chattopadhyay
Writing – original draft: Leonard Lupin-Jimenez, Ashesh Chattopadhyay

Simultaneous Emulation and Downscaling With Physically Consistent Deep Learning-Based Regional Ocean Emulators

Leonard Lupin-Jimenez[1], Moein Darman[1], Subhashis Hazarika[2], Tianning Wu[3], Michael Gray[3], Ruyoing He[3], Anthony Wong[2], and Ashesh Chattopadhyay[1]

[1]Department of Applied Mathematics, University of California, Santa Cruz, Santa Cruz, CA, USA, [2]Converging Technologies Laboratory, Fujitsu Research of America, Santa Clara, CA, USA, [3]North Carolina State University, Marine, Earth, and Atmospheric Sciences, Raleigh, NC, USA

Abstract Building upon recent advancements in AI-driven atmospheric emulation, we present a novel framework for AI-based ocean emulation, downscaling, and bias correction, with a specific focus on high-resolution modeling of the regional ocean in the Gulf of Mexico. Emulating regional ocean dynamics poses distinct challenges due to intricate bathymetry, complex lateral boundary conditions, and inherent limitations of deep learning models, including instability and the potential for hallucinations. In this study, we introduce a deep learning framework that autoregressively integrates ocean surface variables at 8 km spatial resolution over the Gulf of Mexico, maintaining physical consistency over decadal time scales. Simultaneously, the framework downscales and bias-corrects the outputs to 4 km resolution using a physics-informed generative model. Our approach demonstrates short-term predictive skill comparable to high-resolution physics-based simulations, while also accurately capturing long-term statistical properties, including temporal mean and variability.

Plain Language Summary Recent advances in artificial intelligence (AI) have shown that deep learning models can effectively reproduce complex atmospheric behavior. In this study, we apply similar techniques to simulate ocean dynamics in the Gulf of Mexico, a region that is especially difficult to model due to its complex seafloor topography (bathymetry) and the influence of surrounding land and open ocean boundaries. We developed an AI-based method that predicts how ocean surface conditions change over time at high spatial resolution (8 km) while also correcting for biases and enhancing resolution to 4 km using physics-informed techniques. Unlike many AI models that can become unstable or produce unrealistic outputs over time, our approach remains physically consistent even over decades. It achieves short-term accuracy comparable to traditional high-resolution simulations and also captures long-term patterns such as average conditions and variability, offering a more efficient way to study regional ocean systems.

1. Introduction

The North Atlantic Ocean's western boundary current system (WBC), including the Loop Current (LC), Gulf Stream (GS), and Gulf Stream meander (GSM) plays a significant role in controlling the Earth's ocean circulation by transporting heat, salt, nutrients, and strongly influencing the global weather and climate system, including marine ecology. Modeling the regional ocean, for example, in the Gulf of Mexico region (GoM), involves several challenges, starting with complex land boundaries, incorporation of lateral boundary conditions, which is usually computed from a global ocean model, and also eddy shedding events which are caused due to the interactions between the cool subpolar circulation from the North and warm subtropical circulation from the South. Accurately resolving the eddy shedding process in the GoM region has been a challenge for even high-resolution numerical ocean models in the past (Chassignet & Marshall, 2008; Dengo, 1993; Ezer, 2016).

Recent years have seen widespread success in machine learning (ML)-based data-driven emulation of atmospheric dynamics (Bi et al., 2023; Guan et al., 2024; Lam et al., 2022; Pathak et al., 2022), where the weather forecasting accuracy of such ML models has surpassed that of numerical weather prediction models while being several thousand times faster. While several of these AI weather models such as FourCastNet (Pathak et al., 2022), GraphCast (Lam et al., 2022), and Pangu (Bi et al., 2023) eventually become unstable or unphysical, a few studies focusing on long-term stability have successfully demonstrated ML emulators capable of producing stable atmospheric behavior over climate time scales with accurate climatology and variability (Chattopadhyay & Hassanzadeh, 2023; Guan et al., 2024; Watt-Meyer et al., 2024). To extend such ML-based emulation capabilities toward modeling the full Earth system, the development of ocean emulators is a critical next step. Unlike the

Writing – review & editing: Leonard Lupin-Jimenez, Ashesh Chattopadhyay

atmosphere, the ocean presents unique challenges due to longer memory, nonlinear coupling with atmospheric forcing, and complex boundary conditions. However, progress in this area has been limited, with most prior work focusing on low-dimensional models that capture large-scale ocean patterns over short time scales (Agarwal et al., 2021; J. L. Wang et al., 2019). Recently, we have demonstrated success in data-driven regional ocean modeling at very high resolution (4 km) near the Gulf of Mexico (GoM) and Gulf Stream (GS) regions, where OceanNet (Chattopadhyay et al., 2024; Gray et al., 2024; Lowe et al., 2025) achieved short-term prediction performance exceeding that of traditional numerical ocean models such as the Regional Ocean Modeling Systems (ROMS) while maintaining long-term physical consistency. Similarly, Subel et al. and others (Dheeshjith, Subel, Adcroft, et al., 2024; Dheeshjith, Subel, Gupta, et al., 2024; X. Wang et al., 2024) have shown promising progress in emulating the global ocean under varying CO_2 forcings.

In this paper, we extend the recent success of data-driven autoregressive ocean forecasting demonstrated by OceanNet (Chattopadhyay et al., 2024; Gray et al., 2024; Lowe et al., 2025) to simultaneously model multiple surface variables over climate time scales. Furthermore, we explore deep learning-based downscaling and bias correction as a means to enhance the spatial resolution, accuracy, and physical consistency of the emulated Gulf Stream produced by the forecasting model. Downscaling and super resolution have been very popular in the weather and climate community (Harris et al., 2022) where the focus has been primarily on increasing the fidelity of the forecasts from numerical models. Recently, Mardani et al. (2024) developed a generative model-based residual correction algorithm to downscale coarse-grained 25 Km reanalysis to 2 Km-scale fields (using observations) over Taiwan, wherein fine-scale convective structures were recovered. Such efforts in ocean dynamics have been largely absent. Here, we demonstrate both short- and long-term emulation of the surface ocean dynamics near the GoM region along with simultaneous downscaling and bias correction at a higher resolution. We highlight a few key important features in our framework. Unlike most downscaling work that super-resolves reanalysis products into higher resolution observations, we autoregressively predict the surface ocean dynamics with an ML-based forecasting model while downscaling and bias-correcting the emulated fields to a higher resolution reanalysis product derived from a different physical model as well as parameterization schemes. This is particularly difficult since autoregressive models have limited prediction skills, instability issues (Chattopadhyay & Hassanzadeh, 2023; Chattopadhyay et al., 2024), and a tendency to become unphysical at long time scales. To account for such drifts of the autoregressive models, the downscaling framework is further fine-tuned on the difference between the emulation and the target high-resolution reanalysis data.

Our ML-based prediction and downscaling framework has the following features:

- A long-term stable and physically consistent data-driven regional ocean emulator, that is, a forecasting model (FC) trained on sea-surface height (SSH), sea-surface zonal and meridional velocities (SSU and SSV), and sea-surface kinetic energy (SSKE) from low-resolution (LR) GLORYS reanalysis data (Garric et al., 2018).
- A deterministic and generative downscaling framework, DS model, that super resolves and bias corrects the predicted fields from the FC model to high-resolution (HR) CNAPS reanalysis fields (He et al., 2025).
- Finally, an online fine-tuning phase where the DS model undergoes transfer learning to learn the drift between the emulation and CNAPS data starting from a GLORYS initial condition in the training data set.

In the remainder of the paper, we present the two data sets used for training the forecasting (FC) and downscaling (DS) models in the $FCDS$ framework, detail the training and downscaling methodologies, and describe the principled structures in the machine learning models designed to enforce physical consistency, particularly in preserving the energy spectrum. Finally, we evaluate the short- and long-term performance of the downscaled fields using held-out high-resolution CNAPS reanalysis data.

2. Data Sets

We used two reanalysis data sets in this paper. For training the autoregressive forecasting model, FC, we utilize LR data corresponding to SSH, SSU, SSV, and SSKE fields from the global ocean reanalysis, GLORYS (Garric et al., 2018), at $\frac{1}{12}$ degree resolution. We downscale the autoregressively emulated fields from FC to a HR $\frac{1}{24}$ degree resolution regional reanalysis product, CNAPS (He et al., 2025). Further details about the CNAPS reanalysis product can be obtained in Chattopadhyay et al. (2024) and He et al. (2025). The goal of the FCDS framework is to emulate the ocean starting from an initial condition in the GLORYS data set while simultaneously downscaling and bias-correcting the emulation to the CNAPS data set. In this paper, the autoregressive model

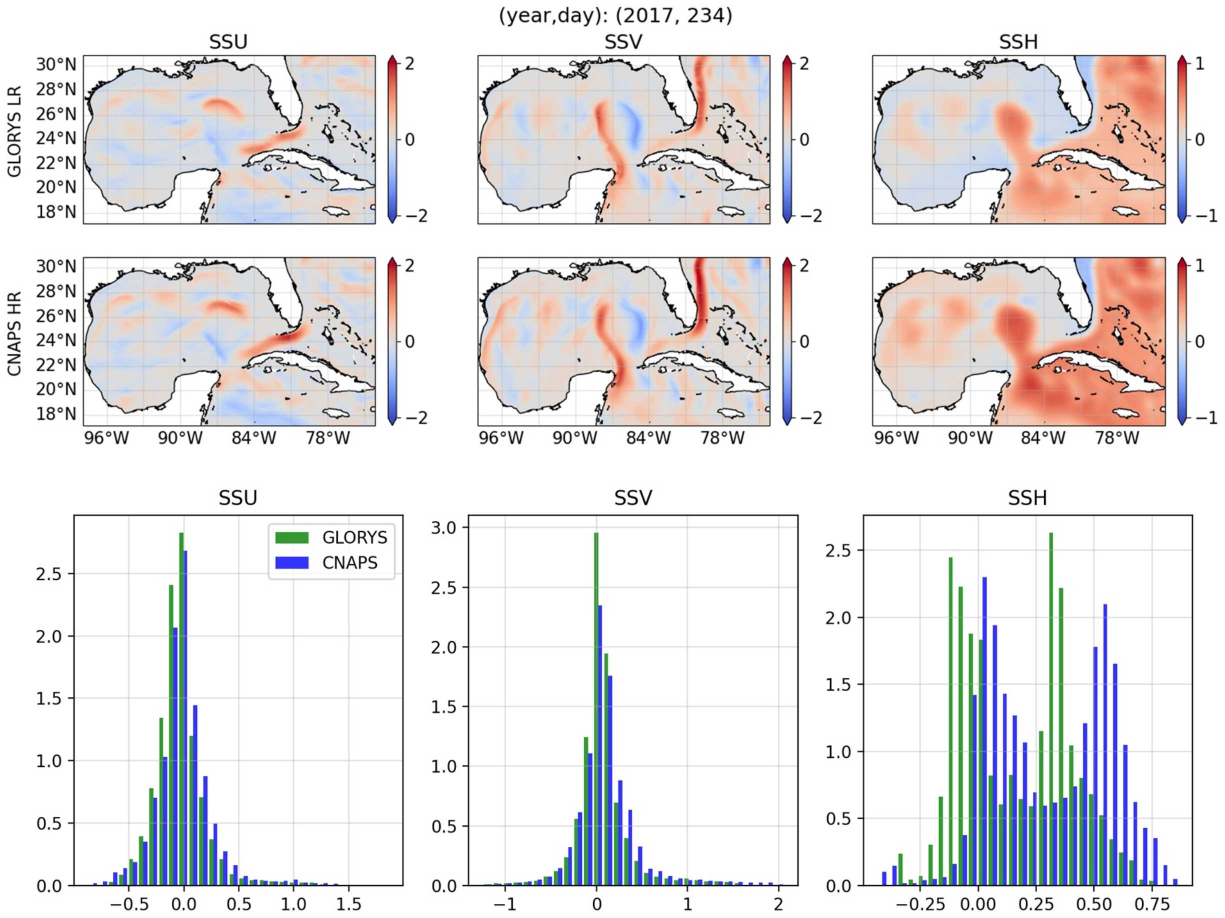

Figure 1. Example snapshot for GLORYS low-resolution and CNAPS high-resolution states, for SSH, SSU, and SSV, for a single time. Large scale structures are similar, but there are differences between the fields between each of the reanalysis products. Below, distributions for GLORYS LR and CNAPS HR data sets, for SSH, SSU, and SSV.

integrates the surface ocean dynamics at a daily time scale. Unlike other atmospheric emulators, which are typically integrated at 6 hourly temporal resolution, we integrated our ocean emulator daily, owing to the longer time scales of the oceanic processes. Figure 1 shows the SSU, SSV, and SSH fields in GLORYS and CNAPS. In our forecasting framework, we include SSKE as a prognostic variable although in reality it is a diagnosed quantity. However, incorporating it in the loss function improves FCDS's prediction and physical consistency of the other prognostic SSU, SSV, and SSH fields and their Fourier spectrums. Furthermore, using SSKE as a prognostic variable helps FCDS to maintain stable and energy-consistent outputs in the forecast.

3. Methodology

In the following sections, we outline the details of the FC and the DS model in our proposed framework as shown in Figure 3. Instead of forecasting at high resolution by training on CNAPS data, as we had done in Chattopadhyay et al. (2024), Gray et al. (2024), and Lowe et al. (2025), the FC model is trained on lower resolution GLORYS regional data. The choice of using a low-resolution forecasting model circumvents the cost of computational memory when using 4 prognostic variables during training and also reduces spectral bias (Chattopadhyay & Hassanzadeh, 2023), thereby promoting stability of the FC model. Then, a physics-constrained generative (as well as a deterministic) downscaling model is trained to resolve higher resolution features in the predicted variables from the FC model to the CNAPS fields. This downscaling model, not only increases the resolution of the predicted fields but also corrects the bias between CNAPS and GLORYS reanalysis data. The overall FCDS framework is illustrated in Figure 2. For the FC component, we employ a two-dimensional Fourier Neural

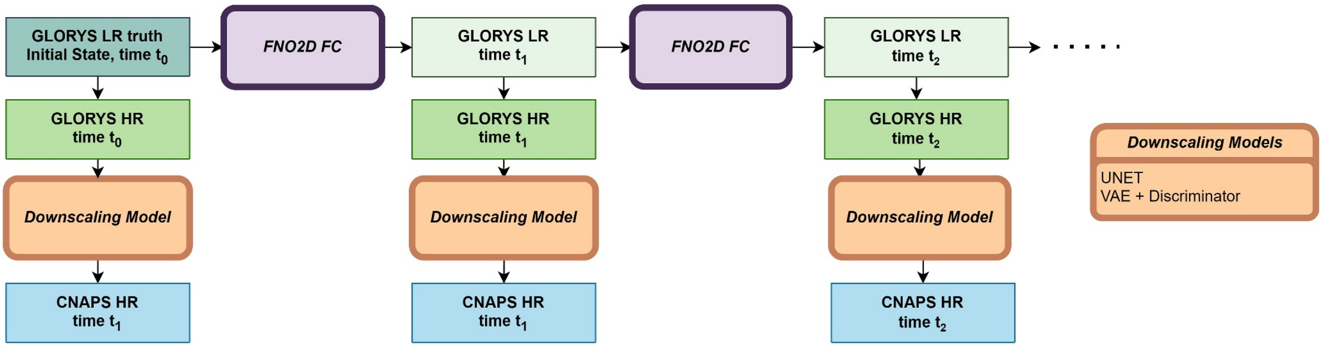

Figure 2. Framework of FC (left to right) and DS (top to bottom). Forecast inference is done from an initial low resolution GLORYS field state (GLORYS LR) at t_0. Before downscaling using our data driven models (top to bottom), the low resolution state is linearly interpolated to the high resolution CNAPS spatial grid.

Operator (FNO), which enables efficient learning of spatiotemporal dynamics. The DS component is implemented using two architectures: a modified variational autoencoder (VAE) enhanced with a PatchGAN-style discriminator and a baseline UNet architecture for comparison. A spectral regularization technique is employed as a physical constraint in both the DS and FC models to enhance the representation of high-wavenumber variability. This regularizer has been demonstrated to improve the long-term stability of data-driven emulators for both atmospheric and oceanic dynamics (Chattopadhyay & Hassanzadeh, 2023; Chattopadhyay et al., 2024). Furthermore, the DS model is fine-tuned in an online transfer learning setting, allowing it to adaptively correct for drift between the FC model's predictions starting from an initial condition in GLORYS from the training data set and the ground-truth CNAPS data, thereby ensuring physical consistency in long-term emulations.

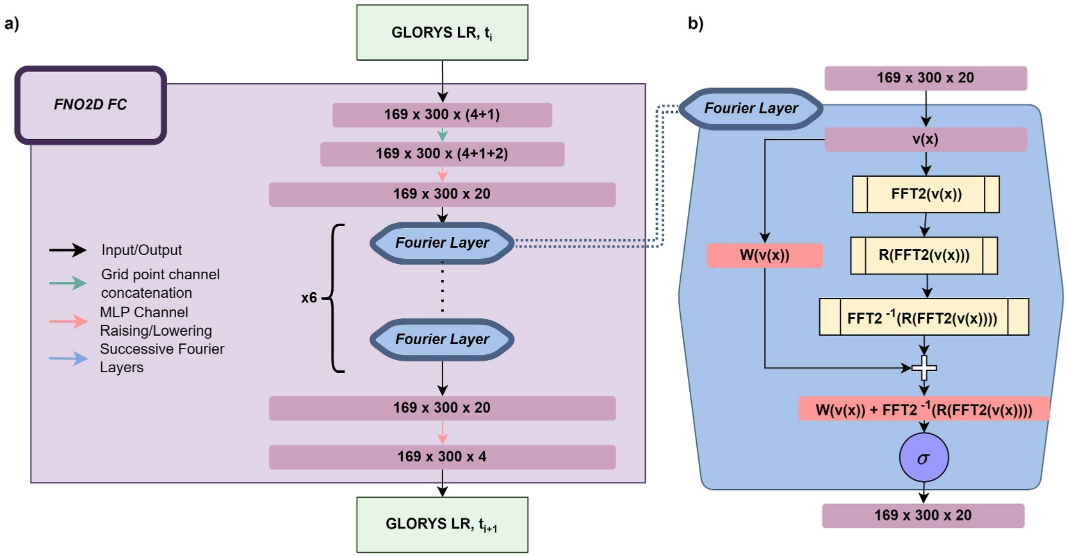

Figure 3. (a) A diagram of the FNO architecture, with channel raising, Fourier layers, and channel lowering to the original dimensionality (Li et al., 2020). Data are inputted with dimensionality (4 + 1), indicating the four channels and a single boolean 1/0 mask representing land or ocean. Then, the latitude and longitude coordinates are concatenated, giving (4 + 1 + 2) channels. The channels are raised, and passed through 6 Fourier layers. Finally, the channels are lowered to the original 4-channel space, and the loss is computed from this output, show in Equations 1–4. (b) An individual Fourier layer. The input data pass through two separate channels: one, which performs a linear transformation $W(v(x))$ on the input, and another which performs a 2-D Fourier transform on the data FFT2$(v(x))$. In the second Fourier layer pipeline on the right of the diagram, the Fourier amplitudes are truncated to remove higher wavenumber modes. A linear transform R is then applied to this truncated form of the 2-D Fourier data, and then an inverse transform is applied. The linear transformation tensor is added to the Fourier operated tensor and is passed through an activation function.

3.1. Loss Functions

The FC and DS models are trained using a composite loss function that incorporates both spatial and spectral information to ensure accurate and physically consistent predictions. Specifically, the total loss is defined as a weighted sum of the mean squared error (MSE) in grid space and the absolute spectral MSE computed along both latitude and longitude directions.

The grid-space loss is defined as

$$L_{\text{grid}} = \frac{1}{lmn}\sum_{c=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{t;cij} - y_{p;cij}\right)^2, \tag{1}$$

where y_t and y_p denote the ground truth and model predictions in grid space, respectively. Here, c is the channel index and i and j correspond to the spatial indices for latitude and longitude.

The spectral losses along the longitude and latitude dimensions are defined as

$$L_{\text{spectral, lon}} = \frac{1}{lm\hat{n}}\sum_{c=1}^{l}\sum_{i=1}^{m}\sum_{\hat{j}=1}^{\hat{n}}\left(\hat{y}_{t;ci\hat{j}} - \hat{y}_{p;ci\hat{j}}\right)^2, \tag{2}$$

$$L_{\text{spectral, lat}} = \frac{1}{l\hat{m}n}\sum_{c=1}^{l}\sum_{\hat{i}=1}^{\hat{m}}\sum_{j=1}^{n}\left(\hat{y}_{t;c\hat{i}j} - \hat{y}_{p;c\hat{i}j}\right)^2, \tag{3}$$

where $\hat{y}_t$ and $\hat{y}_p$ are the absolute values of the Fourier coefficients of the ground truth and predictions, respectively, obtained via discrete Fourier transforms (DFTs) computed along the corresponding spatial dimensions. The indices $\hat{m}$ and $\hat{n}$ denote the number of wavenumbers in the latitude and longitude directions. Since the spatial domains are non-periodic, we mirror each field along the relevant dimension before applying the DFT to enforce periodicity.

The total loss is expressed as

$$L_{\text{total}} = (1-\lambda)L_{\text{grid}} + \lambda\cdot\frac{L_{\text{spectral, lon}} + L_{\text{spectral, lat}}}{2}, \tag{4}$$

where the parameter $\lambda \in [0,\ 1]$ controls the relative contribution of the spectral loss components. This formulation helps mitigate spectral bias and reduces the accumulation of unphysical energy during long-term autoregressive inference (Chattopadhyay & Hassanzadeh, 2023). In our experiments, we find that $\lambda = 0.2$ provides a robust balance for both forecasting and downscaling applications.

For all loss computations, land regions are masked by setting the values of y_t and y_p to zero while ocean regions retain their original values. This masking ensures the models focus training only on oceanic dynamics.

3.2. FC Model Training and Testing

The proposed FC model is trained on LR data from the GLORYS reanalysis product. The input at day t_i consists of SSH, SSU, SSV, and SSKE while the corresponding fields at day t_{i+1} serve as the training targets. The SSKE field is derived from SSU and SSV. A static land-sea mask is also included as an additional input channel to the model. During training, we enforce climatological lateral boundary conditions derived from the training data across all four prognostic variables. The FC model used within the FCDS framework is implemented as a two-dimensional FNO (Li et al., 2020). It comprises six Fourier layers, each retaining 64 spectral modes. The model is optimized using the ADAM optimizer and trained with the total loss function defined in Equation 4. The training data set spans the years 1992–2018, and the autoregressive forecasting performance of the trained FC model is evaluated using 50 distinct initial conditions drawn from the 2019–2020 period. These evaluations assess the model's ability to generate stable and accurate multi-day predictions in an out-of-sample setting.

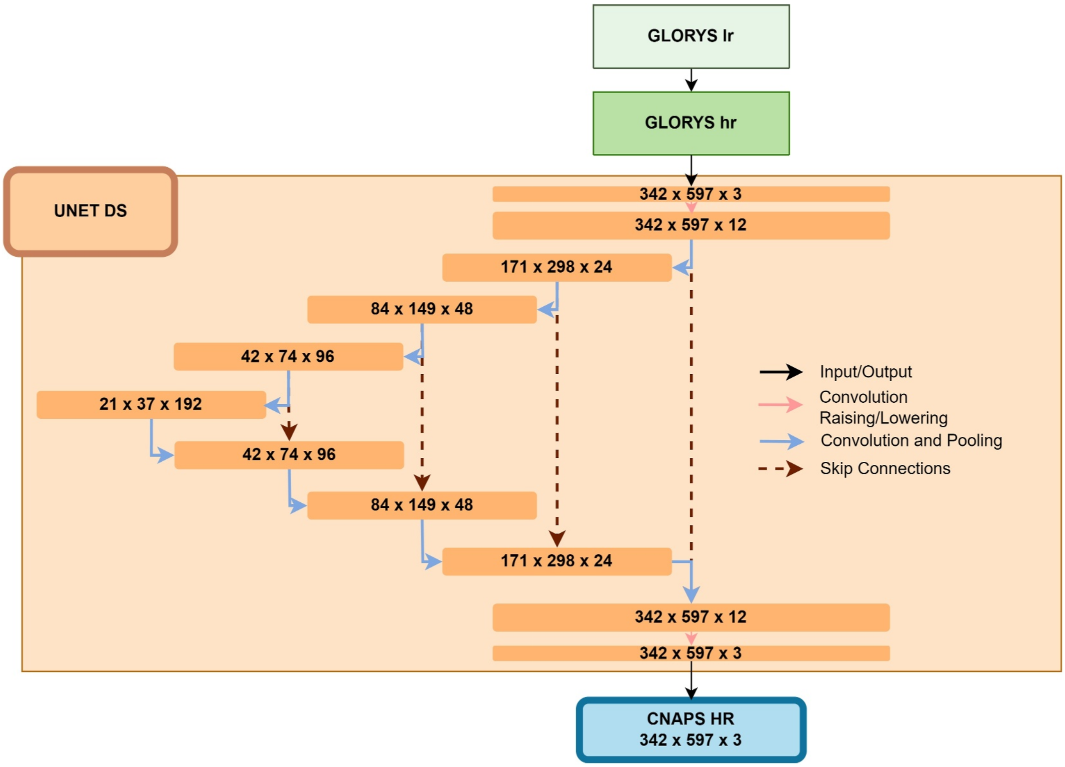

Figure 4. Diagram of UNET architecture used; successive layers shown top to bottom. The x and y sizes of each layer, as well as the number of channels c, are shown in the form in the form $x \times y \times c$. There are for contraction, which perform 3×3 kernel convolutions with ReLU activation, along with a max-pooling layer. Then, 4 layers of expansion are done, with skip connection concatenation up-convolution 2×2 kernel convolution.

3.3. Downscaling Architectures

We compare two architectures for the DS model: a standard UNET with and without the spectral regularizer and a modified variational autoencoder (VAE) with the spectral regularizer (VAE does not show major degradation in performance without spectral regularizer) in the loss functions described in Equations 1–4. The DS model functions as a super-resolution operator that maps low-resolution forecasts from the FC model to high-resolution fields consistent with CNAPS reanalysis data. It is trained offline to learn a deterministic mapping from GLORYS prognostic variables to their CNAPS counterparts, effectively capturing both the fine-scale spatial structure and the systematic bias between the two data sets. Once trained, the DS model is applied to the outputs of the FC model to generate high-resolution predictions.

3.3.1. UNET

The UNET architecture, illustrated in Figure 4, is originally designed for image segmentation, but it has also shown significant promise in various other tasks, including image classification, regression, and downscaling.

The network is structured in a U-shape, with an encoder-decoder configuration. It begins with a channel-raising convolutional layer, followed by four down-sampling layers consisting of convolutional operations and pooling layers, which progressively reduce the spatial resolution while increasing the number of feature channels. This encoder portion captures the high-level features of the input data by compressing spatial information into a compact latent representation.

The decoder portion of the network consists of four up-sampling layers with transposed convolutions, which progressively reconstruct the spatial resolution of the data. These layers use learned filters to expand the feature maps, enabling the network to reconstruct the output image with high spatial fidelity. The key feature of the

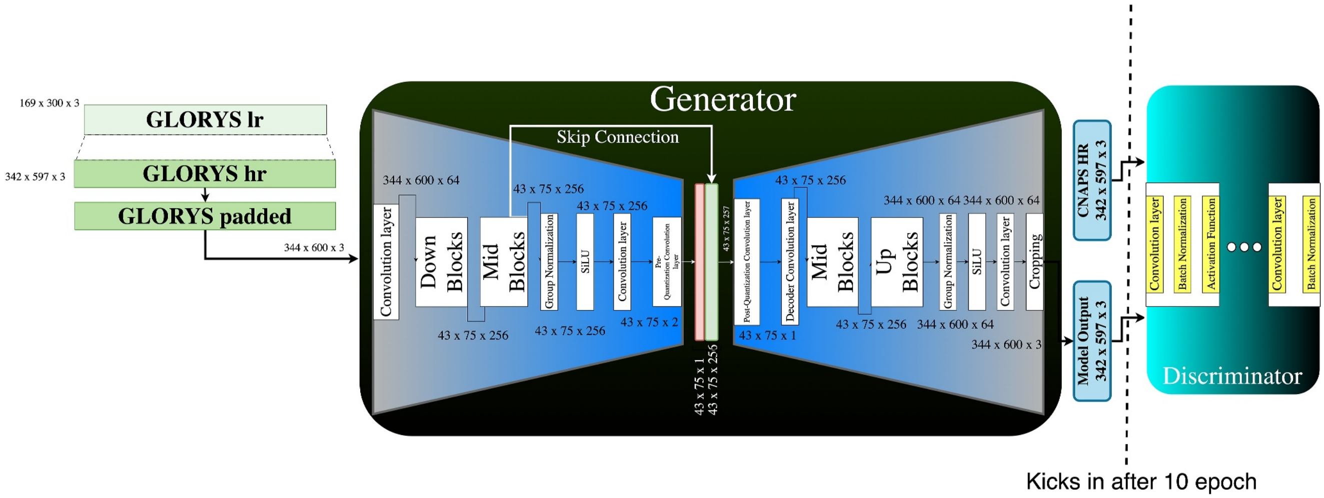

Figure 5. Diagram of the VAE architecture, illustrating the transformation from low-resolution GLORYS data input to high-resolution CNAPS output. The encoder maps the input to a latent space characterized by the mean (μ) and log variance $\left(\log\sigma^2\right)$, with the reparameterization trick enabling differentiable sampling. Skip connections retain critical spatial features to aid the decoder, which upsamples and reconstructs the high-resolution outputs. The PatchGAN discriminator evaluates the realism of reconstructions at the patch level. Training is guided by loss function comprising reconstruction, KL divergence, and adversarial components, producing high-quality realistic reconstructions, and robust latent space learning.

UNET architecture is the use of skip connections between the corresponding layers of the encoder and decoder. These skip connections allow the model to retain fine-grained spatial information that might otherwise be lost during down-sampling. By directly passing feature maps from the encoder to the decoder, these connections help refine the output and prevent the loss of important structural details.

In our specific application, the UNET framework leverages its latent space representation and skip connections to refine low-resolution input data. By utilizing the detailed feature maps passed through the skip connections, the network can correct discrepancies and adjust for the differences between the low-resolution and high-resolution data. This ensures that the model is better equipped to map the low-resolution input to a high-resolution output, aligning the predictions more closely with the characteristics of the high-resolution data. This approach enhances the accuracy and detail of the downscaled output, making it more suitable for applications requiring fine spatial resolution.

3.3.2. Variational Autoencoder With Adversarial Training

The Variational Autoencoder (VAE) is designed to reconstruct high-resolution CNAPS data from low-resolution GLORYS data while learning a meaningful latent representation. The architecture consists of an encoder, a decoder, and a PatchGAN discriminator, trained in an adversarial framework as shown in Figure 5.

The encoder processes the input through an initial convolutional layer, followed by several downsampling layers (*DownBlocks*), which reduce spatial dimensions while increasing feature channels. The bottleneck layers (*MidBlocks*) refine the feature maps before generating the latent distribution, represented by the mean μ and log variance $\log\sigma^2$. The latent sample z is computed using the reparameterization trick:

$$z = \mu + \sigma \cdot \epsilon, \quad \sigma = \exp\left(0.5 \cdot \log\sigma^2\right), \quad \epsilon \sim \mathcal{N}(0,1), \tag{5}$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution. Sampling the latent variable z using Equation 5 ensures differentiability while mapping inputs to latent space, thus improving latent representation for reconstruction. Encoded feature maps are preserved as skip connections to enhance the decoder's reconstruction ability.

The decoder concatenates the latent sample with the skip-connected features and upsamples the spatial dimensions through *UpBlocks*, mirroring the encoder's downsampling operations. Additional *MidBlocks* further refine the reconstructed feature maps. A final convolutional layer produces the high-resolution output, cropped to remove padding applied during the encoder's input processing.

The discriminator is a PatchGAN, designed to evaluate the realism of reconstructed images at the patch level. It processes inputs through sequential convolutional layers, progressively reducing spatial dimensions while predicting a grid of values. Each cell in the grid corresponds to a patch of the input, with higher values indicating greater realism. The architecture employs LeakyReLU activation and batch normalization, with the final layer producing the grid of predictions.

The training combines multiple loss functions to optimize the VAE and discriminator, represented by $\mathcal{V}$ and $\mathcal{D}$, respectively. The reconstruction loss incorporates spatial and spectral components and is defined as

$$L_{\text{recon}} = L_{\text{total}} = (1-\lambda)L_{\text{grid}} + \lambda\frac{L_{\text{spectral, lon}} + L_{\text{spectral, lat}}}{2}. \tag{6}$$

Latent space regularization is achieved via the Kullback-Leibler (KL) divergence:

$$L_{\text{KL}} = -\frac{1}{2}\sum\left(1 + \log\sigma^2 - \mu^2 - \sigma^2\right). \tag{7}$$

The adversarial loss encourages the generator to produce realistic outputs classified as real by the discriminator:

$$L_{\text{adv}} = \frac{1}{N}\sum\left(\mathcal{D}(\mathcal{V}(X(t))) - 1\right)^2, \tag{8}$$

where $\mathcal{D}(\circ)$ represents the discriminator's output, with values in (0, 1); 0 indicates generated samples and 1 indicates real samples.

The total generator loss L_{gen} combines the reconstruction loss L_{recon}, the Kullback-Leibler divergence L_{KL} weighted by β_{KL}, and the adversarial loss L_{adv} weighted by λ_{adv}, where β_{KL} and λ_{adv} control the relative contributions of regularization and adversarial terms.

$$L_{\text{gen}} = L_{\text{recon}} + \beta_{\text{KL}} \cdot L_{\text{KL}} + \lambda_{\text{adv}} \cdot L_{\text{adv}}. \tag{9}$$

The discriminator minimizes the following objective:

$$L_{\text{disc}} = \frac{1}{2}\left(L_{\text{real}} + L_{\mathcal{V}}\right), \tag{10}$$

where:

$$L_{\text{real}} = \frac{1}{N_{\text{real}}}\sum_{i=1}^{N_{\text{real}}}\left(\mathcal{D}(X_{\text{hr},i}(t)) - 1\right)^2, \tag{11}$$

$$L_{\mathcal{V}} = \frac{1}{N_{\text{lr}}}\sum_{i=1}^{N_{\text{lr}}}\left(\mathcal{D}(\mathcal{V}(X_{\text{lr},i}(t))) - 0\right)^2. \tag{12}$$

Here, $X_{\text{hr},i}(t)$ denotes the high-resolution CNAPS ground truth, and $\mathcal{V}(X_{\text{lr},i}(t))$ represents low-resolution VAE-reconstructed data.

Training alternates between optimizing the generator and the discriminator. The generator minimizes L_{gen} via gradient accumulation, while the discriminator minimizes L_{disc}. The dynamic adjustment of the learning rates for both networks ensures stability, preventing either model from overpowering the other. This dynamic adjustment ensures convergence and produces high-quality realistic reconstructions.

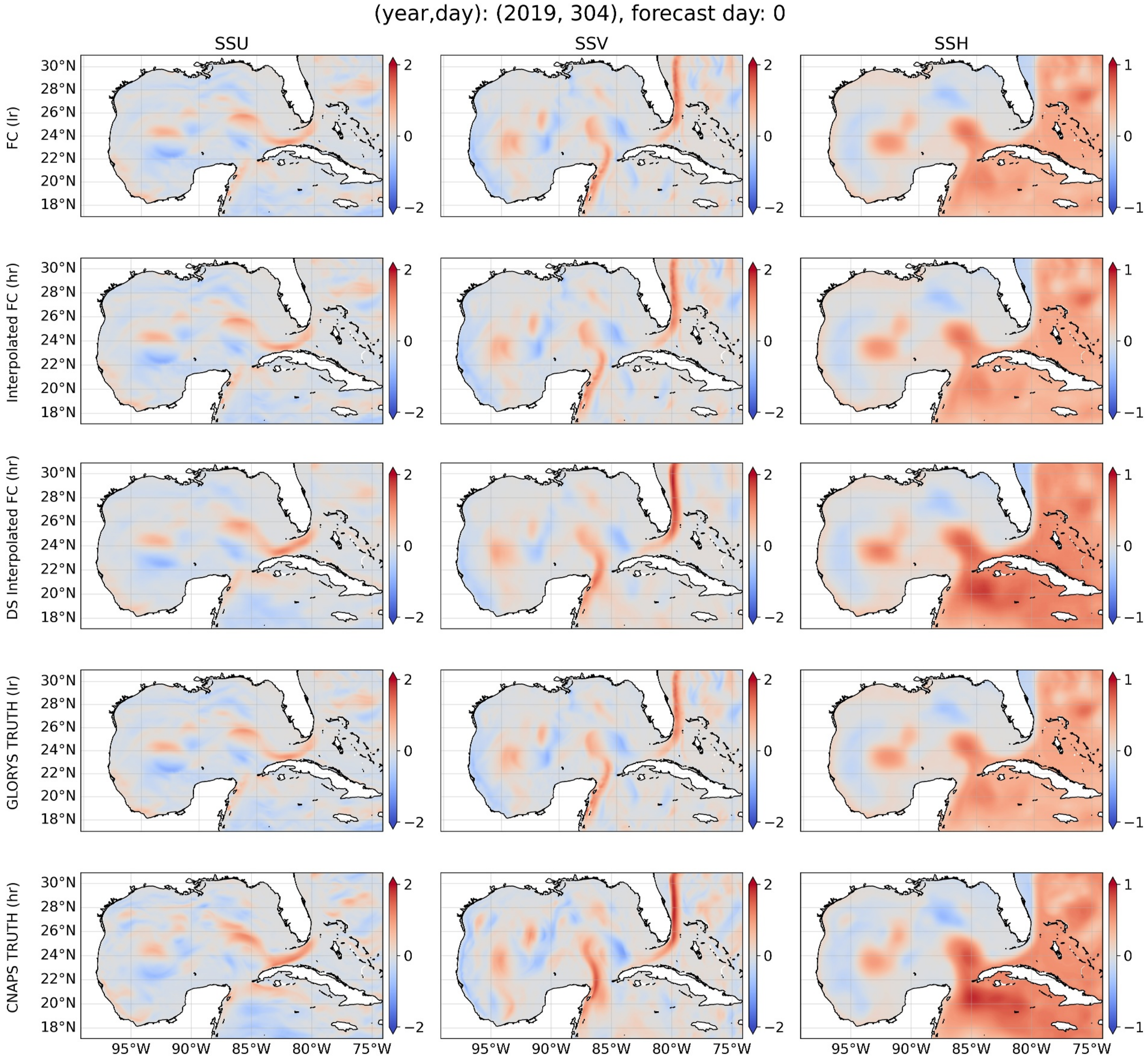

Figure 6. Snapshot of FCDS framework outputs initialized from a single initial condition. The first three rows depict model-generated forecasts and downscaled fields: (1) the low-resolution forecast from the FNO-based autoregressive model, (2) a naive high-resolution bilinear interpolation of the forecast, and (3) the downscaled output using the online fine-tuned UNET-based DS model (FCDS outputs). The fourth and fifth rows show reference data: (4) the low-resolution GLORYS field interpolated to the CNAPS grid and (5) the corresponding high-resolution CNAPS ground truth. Columns represent the three prognostic variables—SSH, SSU, and SSV, respectively.

3.3.3. Online Fine-Tuning of the DS Model for Bias Correction

Unlike conventional downscaling or super-resolution tasks, our approach addresses the challenge of downscaling the outputs of an autoregressive emulator, which itself accumulates model error over time. As the FC model evolves forward, its predictions gradually diverge from the true LR GLORYS fields due to a combination of model approximation error and sensitivity to initial conditions. To address the growing discrepancy between the emulator's outputs and the high-resolution CNAPS data, we apply an online fine-tuning procedure to both DS architectures using the combined grid and spectral loss.

This fine-tuning is performed on training data from the year 2018, where the FC model is evolved autoregressively from 50 different initial conditions over a 30-day forecast horizon. The offline-trained DS models are then refined to learn the residual bias between the FC predictions and the corresponding CNAPS fields.

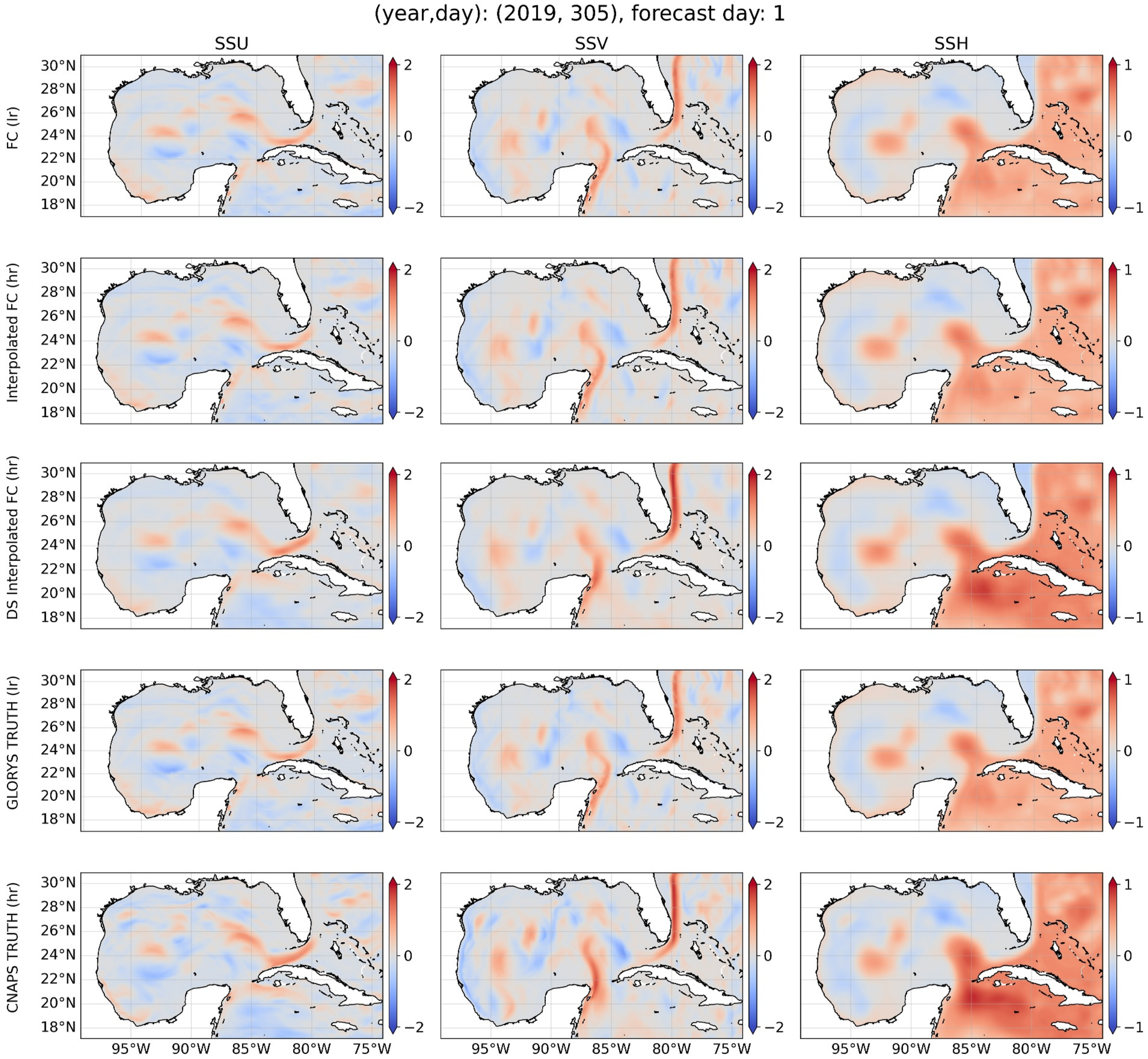

Figure 7. 1 day prediction with the FCDS framework.

A key advantage of this online fine-tuning approach is its ability to compensate for the growing difference between FC and GLORYS during autoregressive emulation and correcting for this additional bias (this bias is different from the inherent bias between GLORYS and CNAPS due to different underlying physical models, parameterization schemes, etc.). By explicitly adapting the DS model to the 30-day autoregressive error of the FC emulator, we improve the physical consistency (Fourier spectrum) and stability of the final high-resolution outputs.

4. Results

In this section, we evaluate the performance of the proposed FCDS framework across a comprehensive set of baselines. Specifically, we compare six configurations of the DS model using the UNET and VAE architectures, each evaluated with and without spectral loss, and with and without online fine-tuning. In total, this yields *six baselines*. Additionally, we include a physics-agnostic baseline using a naive interpolation-based downscaling method, as well as a high-resolution, physics-based benchmark that employs the ROMS model initialized with GLORYS data from 2019. This diverse set of comparisons enables a systematic assessment of the impact of

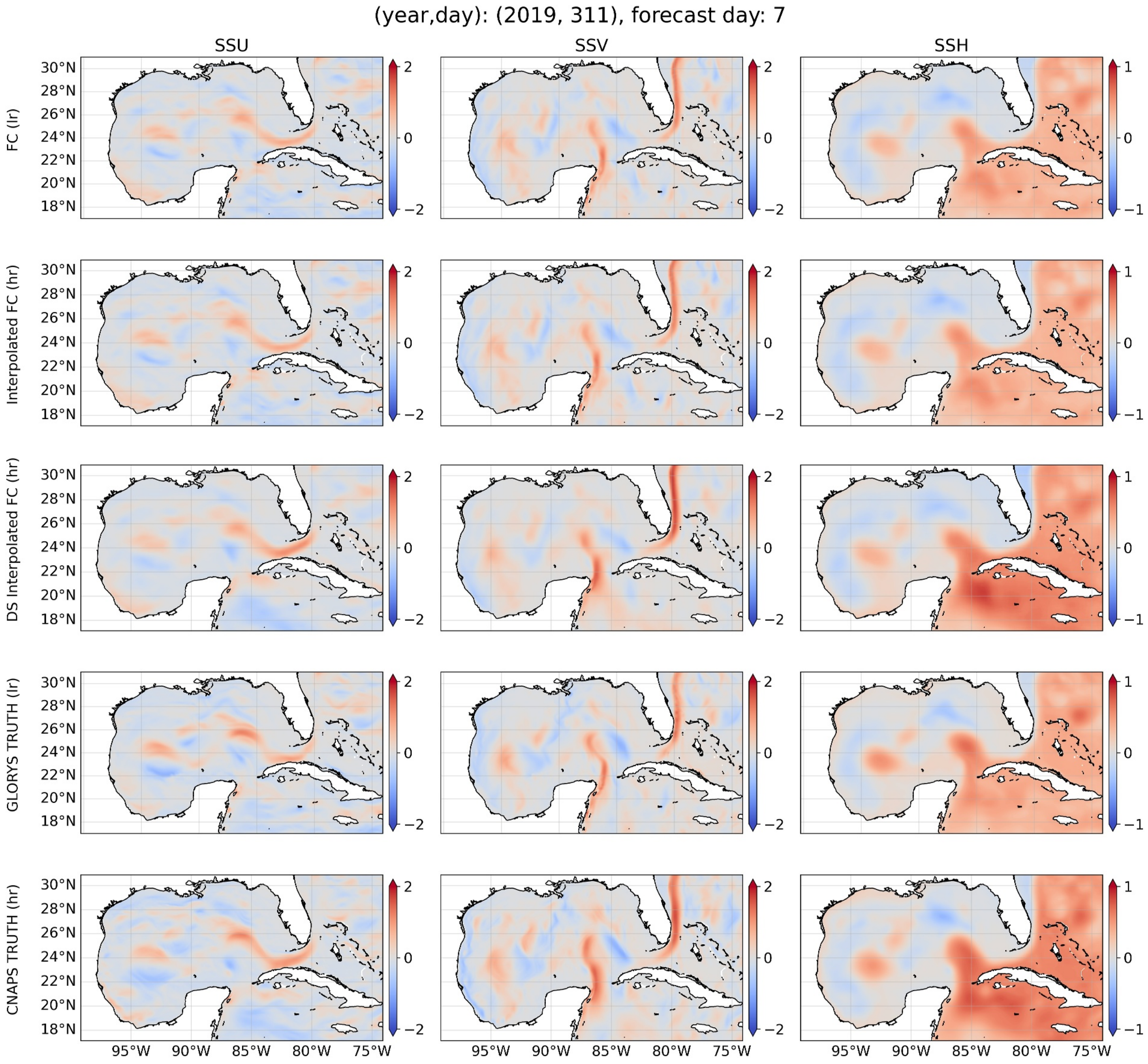

Figure 8. 7 days prediction with the FCDS framework.

architectural design, spectral regularization, and fine-tuning strategies on the performance of the FCDS framework.

4.1. Short-Term Skills of the FCDS Framework

In this section, we present visualizations of the emulated and downscaled SSH, SSU, SSV fields produced by the FCDS framework (Figures 6–10). Figure 6 shows the initial condition while Figure 10 displays the output after 1 year of autoregressive emulation and downscaling. While the FCDS outputs are not expected to perfectly match the high-resolution CNAPS fields due to chaos (sensitivity to initial conditions) after a full year of integration, they remain physically consistent and free from numerical instabilities. Each panel in Figures 6–10 includes, from top to bottom: (a) the low-resolution fields emulated by the FNO-based FC model, (b) a naive bilinear interpolation of those fields to the CNAPS grid, (c) the FCDS downscaled output, (d) the corresponding low-resolution GLORYS data, and (e) the true high-resolution CNAPS fields. The forecast quality of the FC model can be qualitatively assessed by comparing panels one and four while the accuracy of the downscaling procedure can be

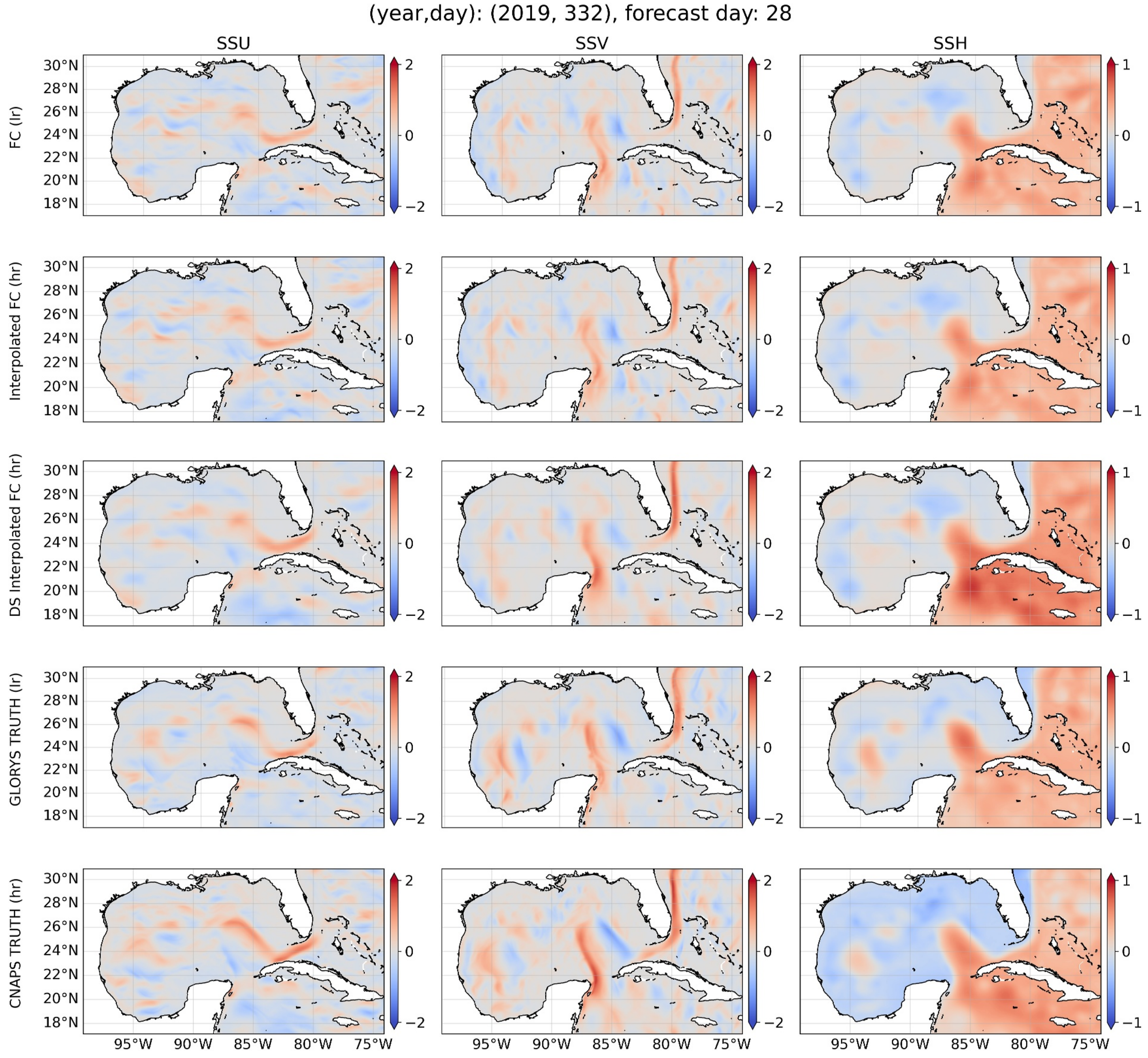

Figure 9. 28 days prediction with the FCDS framework.

evaluated by comparing panels three and five. Although these visualizations are derived from a single initial condition, a broader evaluation is performed using multiple initializations (discussed in the next paragraph). The figures shown correspond to the best-performing configuration of FCDS, which uses the UNET-based DS model with online fine-tuning. As evident in the SSH panels, the FCDS framework captures the intensity of GS more faithfully than the naive interpolation baseline.

In Figure 11, we present a comparative evaluation of several short-term performance metrics over a 30-day forecast horizon to assess the effectiveness of the FCDS framework. Specifically, we analyze the Pearson correlation coefficient, anomaly correlation coefficient (ACC), structural similarity index measure (SSIM), and root mean squared error (RMSE) across all different FCDS configurations, a naive interpolation model, and *a high-resolution physics-based ROMS model initialized from GLORYS*. Figure 11 clearly illustrates that all configurations of the FCDS framework outperform the naive interpolation baseline and remain competitive with ROMS. Notably, the online fine-tuned UNET with spectral loss configuration achieves the highest correlations, SSIM scores, RMSE, and outperforms ROMS. The SSIM metric, which effectively captures the structural fidelity of eddy formations and their shedding events, reveals significant instability and uncertainty associated with the

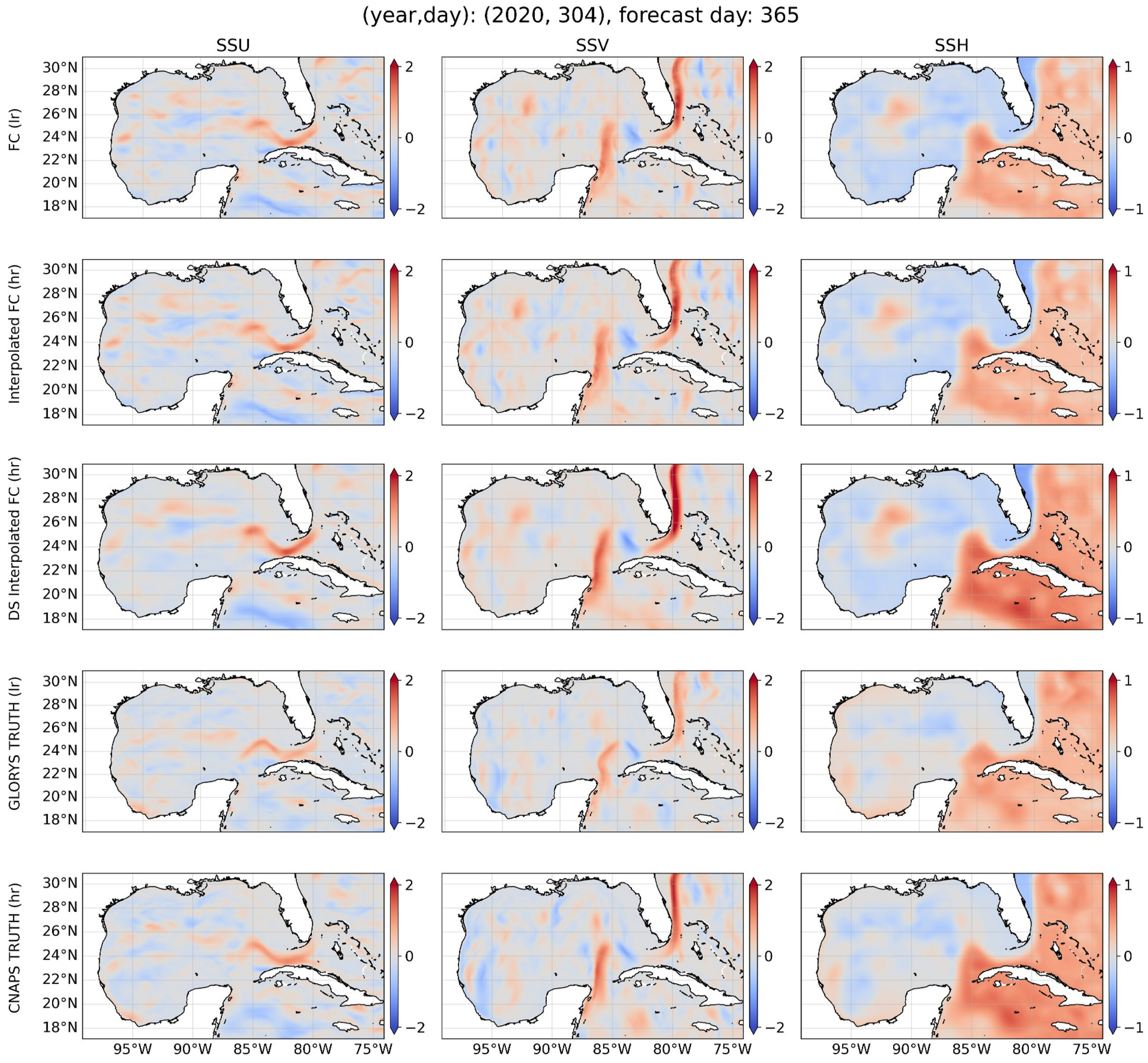

Figure 10. 365 days prediction with the FCDS framework.

naive interpolation-based approach across various initial conditions. Conversely, all FCDS configurations exhibit robust and consistent performance, with comparatively smaller uncertainties. It is important to emphasize that, unlike traditional prediction tasks without downscaling, the ACC metric does not start from an ideal value of 1.0 at the initial condition. This discrepancy occurs because the initial conditions originate from the low-resolution GLORYS data set, whereas the true high-resolution snapshots come from the CNAPS reanalysis data set. Such a scenario represents a realistic forecasting and downscaling challenge, where the training data and the initial conditions can originate from different reanalysis or observational products.

4.2. Power Spectrum of the FCDS Framework

In this section, we move beyond conventional short-term statistical evaluations and investigate physics-informed metrics, specifically focusing on the power spectra of SSKE and SSH. We compare these spectral metrics derived from the FCDS framework against reference high-resolution CNAPS spectra to evaluate the fidelity of physically relevant oceanic structures captured by the models.

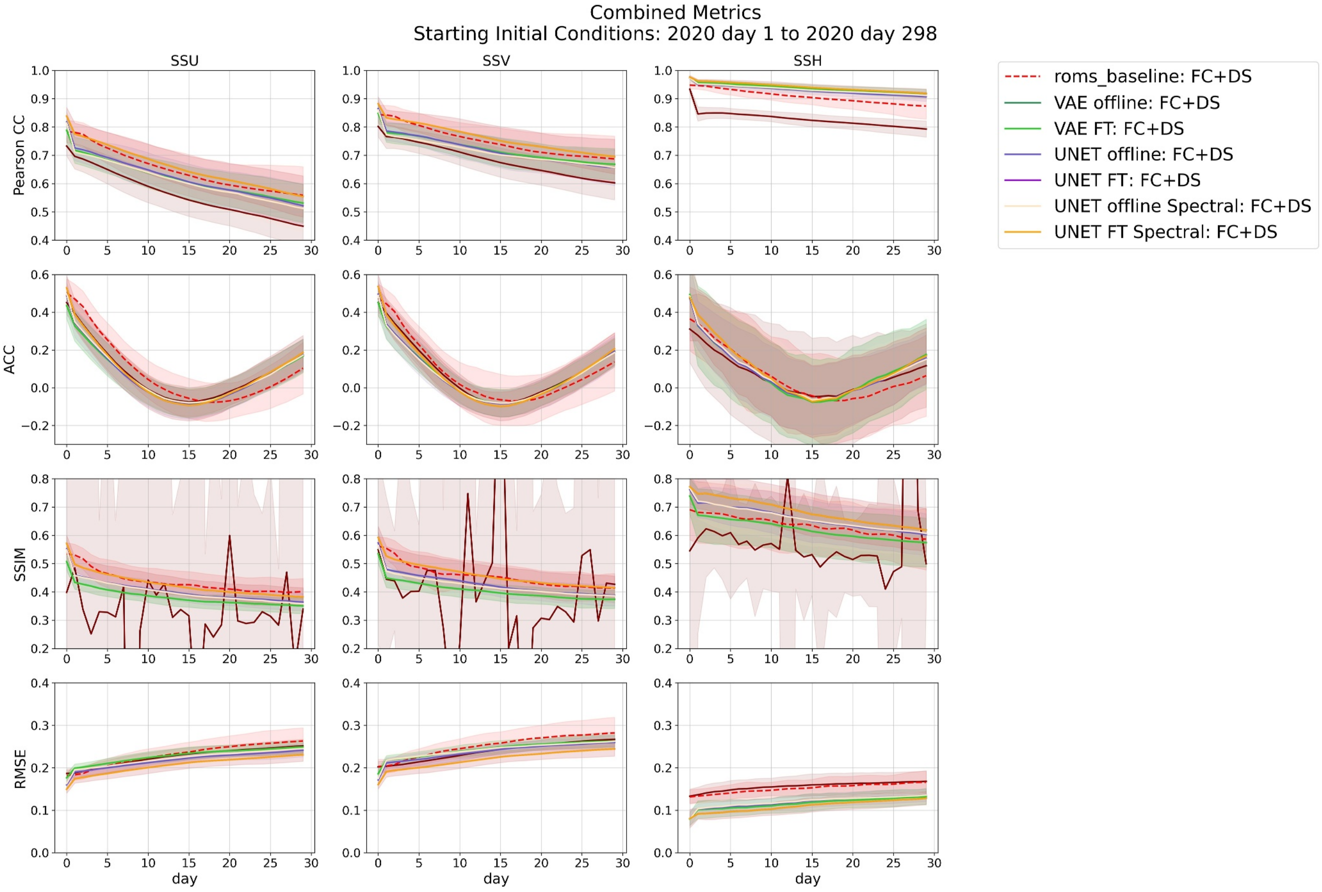

Figure 11. Comparison of FCDS framework performance over 50 initial conditions from the test data set against the high-resolution CNAPS ground truth. Forecasts are generated using an FNO model trained with a spectral loss term ($\lambda = 0.2$), followed by application of various downscaling models. For each forecast, either an offline-trained DS model or a model additionally fine-tuned online ("offline" and "FT" in the legend, respectively) with or without spectral regularizer ("Spectral" in the legend) is used to generate high-resolution outputs. These outputs are evaluated against CNAPS using four metrics: Pearson correlation coefficient, anomaly correlation coefficient (ACC), structural similarity index (SSIM), and root mean squared error (RMSE). All FCDS configurations consistently outperform the naive linear interpolation baseline across all evaluation metrics and either outperforms or remains competitive with high-resolution ROMS initialized from GLORYS ("roms_baseline" in the legend). The uncertainty across 50 initial conditions are provided in the shading as ± std.

First, we consider a non-autoregressive evaluation (Figure 12), in which the downscaling models directly map GLORYS fields to high-resolution CNAPS fields without iterative forecasting. In this offline scenario, the generative VAE-based DS model more effectively captures the SSKE spectrum relative to the deterministic UNET-based DS model, particularly at intermediate wavenumbers. Nevertheless, both architectures demonstrate limitations at higher wavenumbers, indicating difficulties in fully resolving fine-scale ocean dynamics. Conversely, all FCDS configurations accurately reproduce the SSH power spectrum in the non-autoregressive scenario, highlighting their effectiveness in capturing large-scale oceanic variability. It should also be noted that LR's (GLORYS) SSH spectrum is significantly different from HR (CNAPS).

Subsequently, we assess these models under an autoregressive prediction framework (Figure 13), where prediction errors accumulate over iterative forecasts. In this online mode, the previously observed performance advantage of the VAE relative to the UNET model diminishes, with both DS architectures showing comparable performance for both SSKE and SSH spectra. This shift in relative model performance from offline to online settings underscores the known discrepancy between offline evaluation metrics and online predictive skill, a phenomenon previously documented in subgrid-scale modeling and climate parameterization studies (Lin et al., 2023).

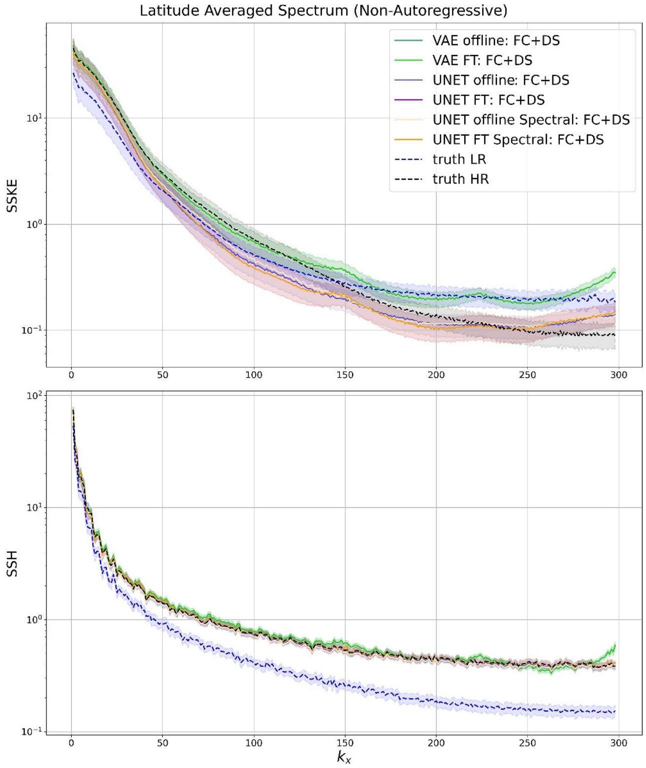

Figure 12. Non-autoregressive mean power spectra of SSKE and SSH over a 30-day period, with shaded regions indicating the standard deviation across 50 initial conditions. The downscaling models used here are trained independently of any autoregressive forecasting, that is, they operate directly on low-resolution GLORYS inputs without iterative time evolution. The VAE-based downscaling model more accurately reconstructs the zonal SSKE spectrum compared with the UNET-based model while both perform similarly in reproducing the SSH spectrum.

Importantly, in the autoregressive prediction mode (Figure 13), the FCDS framework consistently outperforms the naive interpolation baseline, successfully preserving physically realistic high-wavenumber spectral structures. Thus, the DS models within the FCDS framework effectively mitigate spectral inaccuracies inherent to naive interpolation, ensuring stable and realistic fine-scale oceanic features throughout extended autoregressive forecasting.

4.3. Long-Term Stability, Mean, and Variability

In this section, we conduct a comprehensive evaluation of the long-term stability and physical consistency of the FCDS framework by extending its emulation period to 10 years. Such extended integrations are essential for detecting potential numerical instabilities, biases, or unphysical drifts in deep learning-based ocean emulators. Figures 14 and 15 illustrate snapshots of the FCDS framework outputs after approximately 100 and 4,000 days of continuous emulation, respectively.

Inspection of the power spectra for the sea surface zonal velocity (SSU) and meridional velocity (SSV) fields reveals that their spectral distributions remain broadly consistent with the high-resolution CNAPS reference data, although minor artifacts and deviations are noticeable at higher wavenumbers, indicating persistent challenges in accurately capturing the smallest-scale structures. In contrast, the SSH power spectrum is exceptionally well-preserved over the entire 10-year integration period, suggesting that the FCDS framework robustly maintains the large-scale oceanic dynamics over multi-year timescales.

To quantitatively assess the long-term performance, we compute the temporal mean and standard deviation of SSH, SSU, and SSV fields produced by the FCDS framework (Figure 16). These metrics are directly compared against the corresponding statistics from the high-resolution CNAPS data set. As demonstrated by Figure 16, the FCDS framework accurately reproduces both the climatological mean state and the associated variability of the

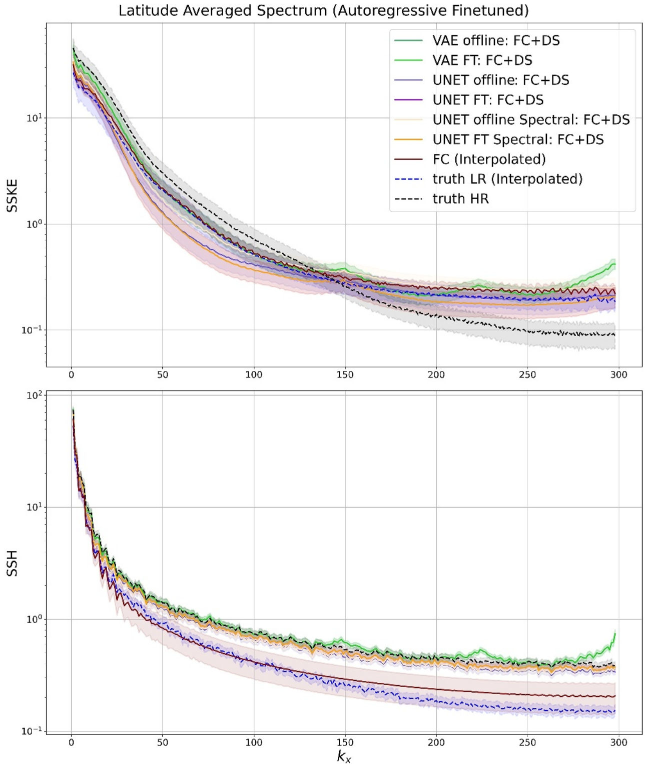

Figure 13. Mean power spectra of SSKE and SSH over a 30-day autoregressive emulation window, with shaded regions indicating the standard deviation across an ensemble of 50 initial conditions. The downscaling models were initially trained in an offline setting using low-resolution to high-resolution mappings, and subsequently fine-tuned in an online mode using multi-step forecasts from the FNO-based forecasting model. This autoregressive fine-tuning allows the DS models to adapt to the accumulated errors and biases present during iterative forecasting, improving spectral consistency over extended time horizons.

CNAPS data. This result confirms the model's capability to maintain a physically realistic ocean climate throughout extended emulation periods without significant drift or instability, highlighting the potential of the FCDS framework for stable, realistic, and physically consistent long-term ocean predictions.

5. Discussion

In this paper, we introduce FCDS—a framework for autoregressive emulation and simultaneous downscaling and bias correction for the region of GoM. We have considered only the ocean surface variables in this paper, noting that the framework is easily extendable to other sub-surface variables. The framework is purely data-driven and hence can be executed at orders of magnitude faster runtime than physics-based modeling and downscaling frameworks. Both the autoregressive model and the downscaling models are equipped with a physics-inspired spectral loss function that remedies the adverse effects of spectral bias which leads to instability and unphysical drifts (Bonavita, 2023; Chattopadhyay & Hassanzadeh, 2023).

In this work, instead of taking physics-based forecasts and downscaling them to high resolution, we develop an autoregressive emulator that is orders of magnitude faster than a physics-based model and downscale the emulated fields to higher resolution. Consequently, the autoregressive model deviates from the true low-resolution fields due to compounding model error and chaos. Therefore, the downscaling model is fine-tuned to perform bias correction as well. The bias correction strategy accounts for three sources of error: the deviation of the emulator, the discrepancy in resolutions between the GLORYS and CNAPS data, and the underlying differences in the physical parameterizations used by the numerical models underlying CNAPS and GLORYS.

In this work, we emulate low-resolution fields instead of high-resolution fields. This choice stems from the role of spectral bias—a fundamental tendency of deep neural networks to under-represent high-wavenumber features—

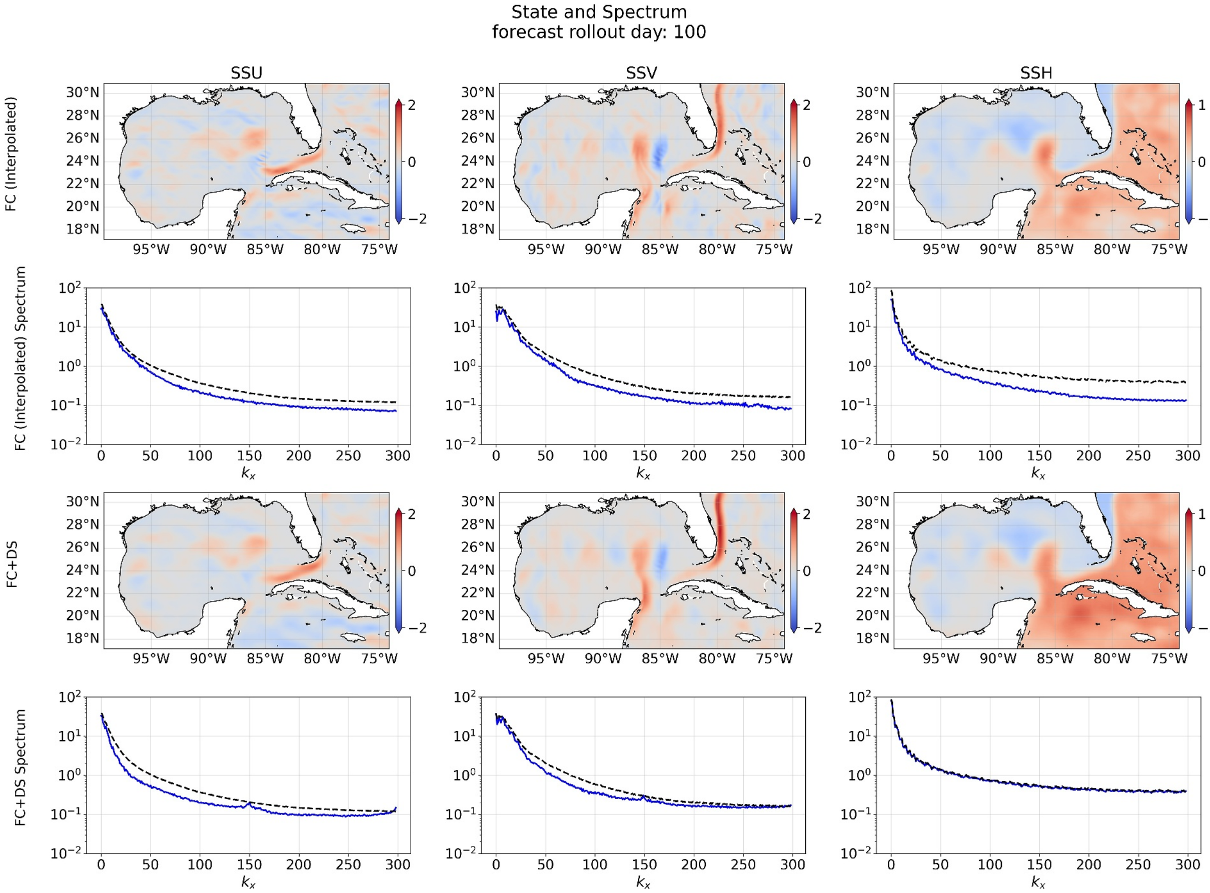

Figure 14. Snapshot of the FCDS framework outputs after 100 days of autoregressive emulation, using the online fine-tuned UNET-based downscaling model. The top row displays the low-resolution FC forecast interpolated onto the high-resolution CNAPS grid, accompanied by its power spectrum with wavenumber annotations. The bottom row presents the corresponding downscaled fields from the UNet-based DS model and their associated spectra. In the spectral plots, the dotted black line indicates the reference high-resolution CNAPS spectrum, whereas the solid blue lines represent the spectra from the interpolated FC output (top) and the FCDS output (bottom). Shaded regions around each line indicate the standard deviation across multiple ensemble realizations, highlighting uncertainties. This visualization demonstrates that the FCDS framework significantly improves spectral fidelity at higher wavenumbers compared to simple interpolation, thereby effectively resolving physically realistic small-scale oceanic features.

as it manifests during autoregressive integration. In general, for weather and climate emulators that have gained popularity in the climate sciences, high-resolution emulators often evolve at much lower effective resolution due to spectral bias (Bonavita, 2023; Chattopadhyay & Hassanzadeh, 2023). This is a major contributor to unphysical drifts and instabilities in such models. However, spectral bias is significantly reduced when the target field is already low resolution, making the model less susceptible to instability or drift. Therefore, we focus on low-resolution autoregressive emulation and use a dedicated downscaling model to project the predicted fields to higher resolution. Since both components of the FCDS framework are data-driven, they are highly efficient at inference time, and the final high-resolution outputs are produced at a fraction of the computational cost of traditional physics-based models.

Unlike conventional super-resolution tasks, where low-resolution fields are downscaled to high-resolution fields within the same data distribution (e.g., downscaling low-resolution ERA5 data to high-resolution ERA5), our work addresses a more complex setting: downscaling from one data distribution (GLORYS) to another (CNAPS). These data sets differ not only in resolution but also in their underlying physical models, parameterizations, and data assimilation strategies. This setup more accurately reflects real-world forecasting scenarios, where models may be initialized using one reanalysis product but evaluated against another. In our case, the emulator is initialized with surface fields from GLORYS while the final downscaled outputs are compared against CNAPS. Although this is a more challenging task, the combination of offline downscaling and online fine-tuning enables

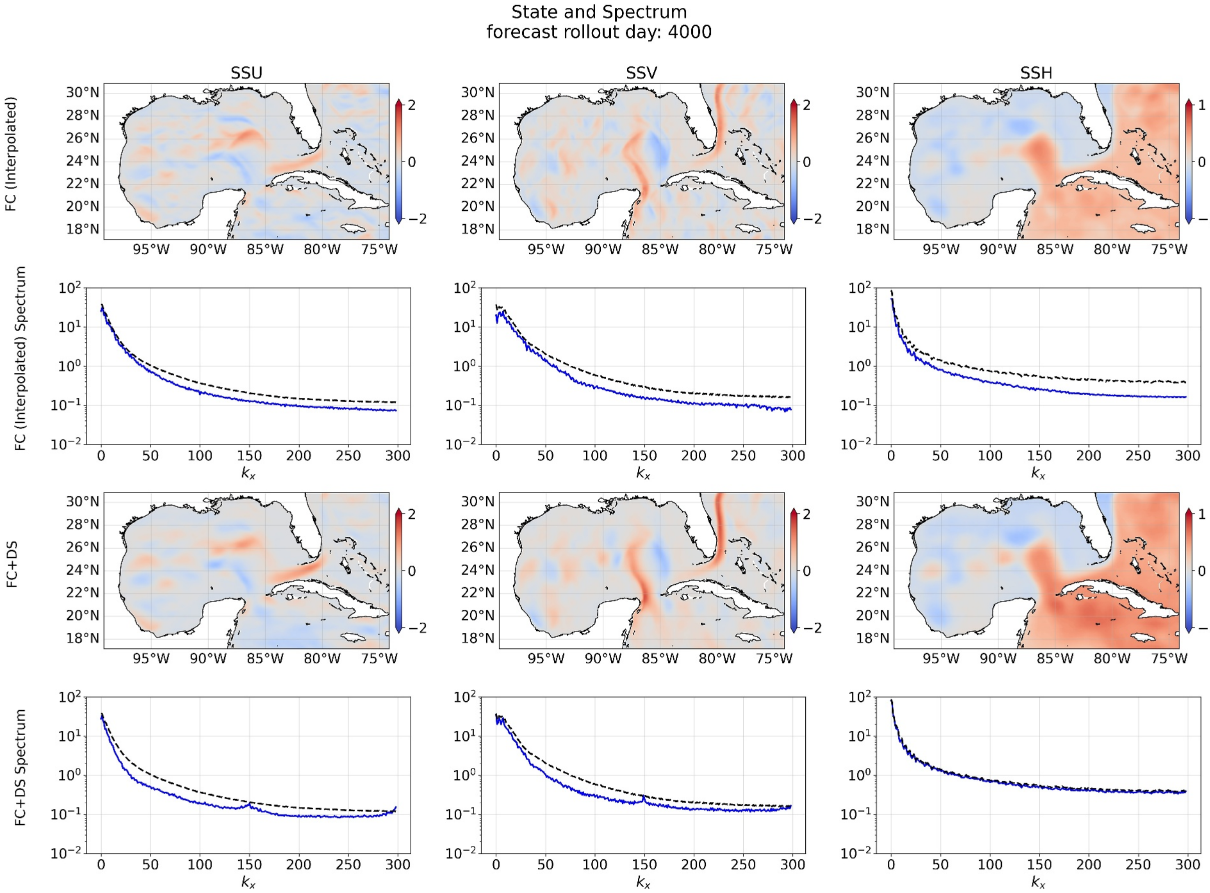

Figure 15. Same as Figure 14 but for FCDS emulation at day 4,000.

the FCDS framework to produce accurate short-term forecasts while also preserving long-term statistical properties that are physically consistent with the reference data.

A major challenge with our framework is the abundance and availability of gridded reanalysis data. However, in reality the observations are available through satellites, ocean buoys, and drifters which are noisy and sparse. Either training with observations or an efficient data assimilation strategy is essential for frameworks such as FCDS to be operational. In atmospheric dynamics, end-to-end emulation framework initialized with observations have seen nominal success (Allen et al., 2025). Our current endeavors are focused on developing generative data assimilation frameworks for surface ocean dynamics, early success for which has been demonstrated in Martin et al. (2025).

While the FCDS framework uses surface ocean variables and remains stable at decadal time scales, there are some key components missing in the autoregressive model. To begin with, the model does not consider any atmospheric forcing which can only be incorporated with a separate atmospheric emulator. We are currently working towards a regional coupled emulator at high resolution. Furthermore, while we can successfully emulate the control climate, we cannot estimate the regional response of the ocean to CO_2 forcings. In future work, we would focus on developing the emulator with sub-surface ocean variables as well as radiative forcing to study the climate change impacts on the GoM region.

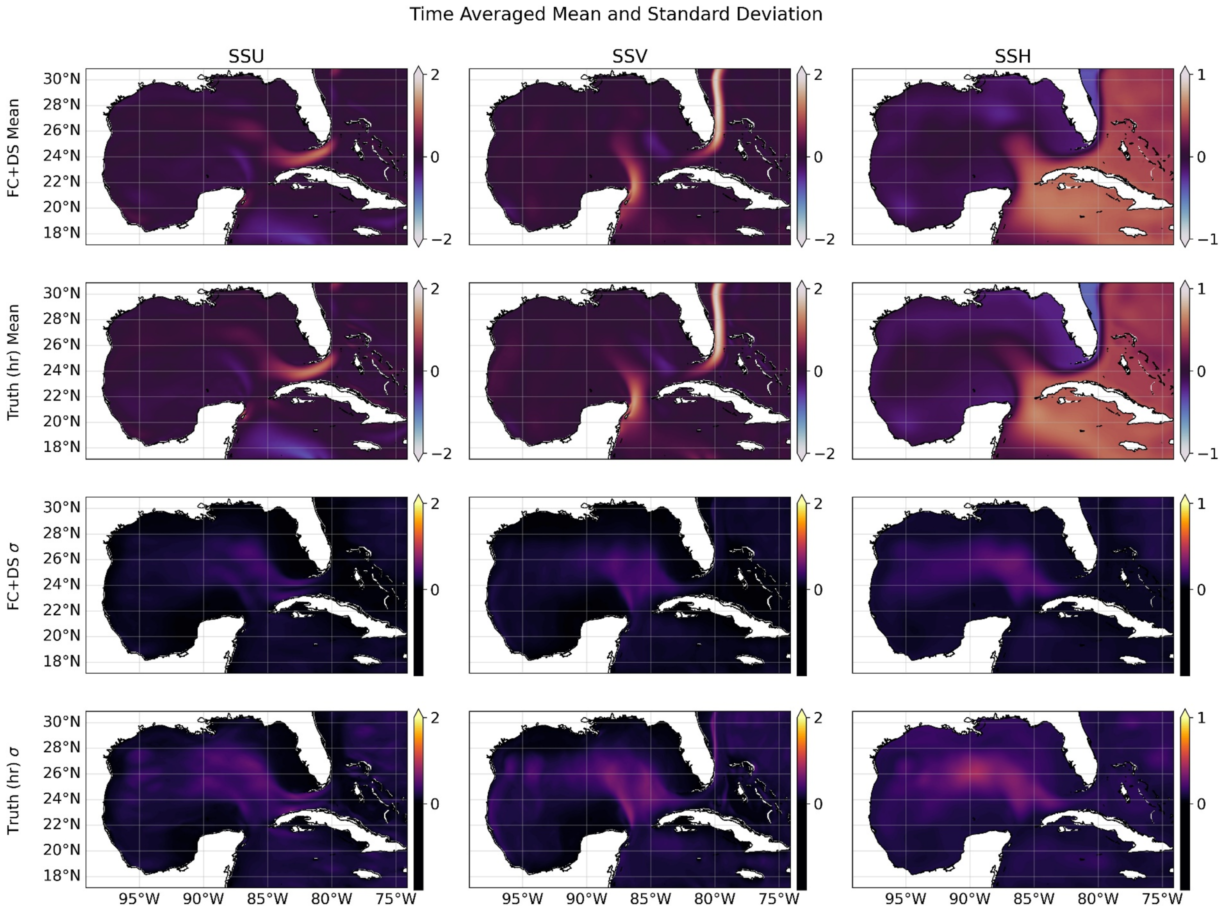

Figure 16. 10-year mean and standard deviation of fields for FCDS prediction and truth. The first two rows are the 10-year mean of FCDS and CNAPS (truth) and the last two rows represent the standard deviation (σ).

Data Availability Statement

The codes and accompanying data for this paper are available at Chattopadhyay et al. (2025).

Acknowledgments

AC and SH designed the study. LJ wrote the codes and conducted the analysis. MD wrote the VAE used in the study. TW, MG, and RH provided the CNAPS data. AC, SH, LJ, and MD analyzed the results and discussed the conclusions. AC, LJ, and MD wrote the manuscript. All authors edited the manuscript. AC, MD, and LJ were supported by the National Science Foundation (Grant 2425667) and computational support from NSF ACCESS MTH240019 and NCAR CISL UCSC0008, and UCSC0009. A part of this study was conducted when LJ was an internee at Fujitsu Research of America under the supervision of SH and AW.

References

Agarwal, N., Kondrashov, D., Dueben, P., Ryzhov, E., & Berloff, P. (2021). A comparison of data-driven approaches to build low-dimensional ocean models. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002537. https://doi.org/10.1029/2021ms002537

Allen, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W. P., Andersson, T. R., et al. (2025). End-to-end data-driven weather prediction. *Nature*, *641*(8065), 1172–1179. https://doi.org/10.1038/s41586-025-08897-0

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, *619*(7970), 1–6. https://doi.org/10.1038/s41586-023-06185-3

Bonavita, M. (2023). On the limitations of data-driven weather forecasting models. arXiv preprint arXiv:2309.08473.

Chassignet, E., & Marshall, D. (2008). Gulf Stream separation in numerical ocean models. *American Geophysical Union*, 39–61. https://doi.org/10.1029/177gm05

Chattopadhyay, A., Gray, M., Wu, T., Lowe, A. B., & He, R. (2024). OceanNet: A principled neural operator-based digital twin for regional oceans. *Scientific Reports*, *14*(1), 21181. https://doi.org/10.1038/s41598-024-72145-0

Chattopadhyay, A., Gray, M., Wu, T., Lowe, A. B., & He, R. (2025). Simultaneous emulation and downscaling with physically-consistent deep learning–based regional ocean emulators [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.14607130

Chattopadhyay, A., & Hassanzadeh, P. (2023). Long-term instabilities of deep learning-based digital twins of the climate system: The cause and a solution. arXiv preprint arXiv:2304.07029.

Dengo, J. (1993). The problem of Gulf Stream separation: A barotropic approach. *Journal of Physical Oceanography*, *23*(10), 2182–2200. https://doi.org/10.1175/1520-0485(1993)023<2182:tpogss>2.0.co;2

Dheeshjith, S., Subel, A., Adcroft, A., Busecke, J., Fernandez-Granda, C., Gupta, S., & Zanna, L. (2024). Samudra: An AI global ocean emulator for climate. arXiv preprint arXiv:2412.03795.

Dheeshjith, S., Subel, A., Gupta, S., Adcroft, A., Fernandez-Granda, C., Busecke, J., & Zanna, L. (2024). Transfer learning for emulating ocean climate variability across CO_2 forcing. arXiv preprint arXiv:2405.18585.

Ezer, T. (2016). Revisiting the problem of the Gulf Stream separation: On the representation of topography in ocean models with different types of vertical grids. *Ocean Modelling*, *104*, 15–27. https://doi.org/10.1016/j.ocemod.2016.05.008

Garric, G., Parent, L., Greiner, E., Drévillon, M., Hamon, M., Lellouche, J., et al. (2018). Performance and quality assessment of the global ocean eddy-permitting physical reanalysis GLORYS2V4. In *Proceedings of the Eight EuroGOOS International Conference, 3–5 October 2017, Bergen, Norway* (pp. 215–222).

Gray, M. A., Chattopadhyay, A., Wu, T., Lowe, A., & He, R. (2024). Long-term prediction of the Gulf Stream meander using OceanNet: A principled neural operator-based digital twin. *EGUsphere*, *2024*, 1–23.

Guan, H., Arcomano, T., Chattopadhyay, A., & Maulik, R. (2024). Lucie: A lightweight uncoupled climate emulator with long-term stability and physical consistency for o (1000)-member ensembles. arXiv preprint arXiv:2405.16297.

Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS003120. https://doi.org/10.1029/2022ms003120

He, R., Wu, T., Mao, S., Zong, H., Zambon, J., Warrillow, J., et al. (2025). Advanced ocean reanalysis of the northwestern Atlantic: 1993-2022. arXiv preprint arXiv:2503.06907.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., et al. (2022). Graphcast: Learning skillful medium-range global weather forecasting. arXiv preprint arXiv:2212.12794.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2020). Fourier neural operator for parametric partial differential equations. arXiv preprint arXiv:2010.08895.

Lin, J., Yu, S., Beucler, T., Gentine, P., Walling, D., & Pritchard, M. (2023). Systematic sampling and validation of machine learning-parameterizations in climate models. arXiv preprint arXiv:2309.16177.

Lowe, A. B., Gray, M., Chattopadhyay, A., Wu, T., & He, R. (2025). Long-term predictions of loop current eddy evolutions using oceannet: A Fourier neural operator-based data-driven ocean emulator. *Artificial Intelligence for the Earth Systems*, *4*(3). https://doi.org/10.1175/aies-d-24-0039.1

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., et al. (2024). Residual diffusion modeling for km-scale atmospheric downscaling.

Martin, S. A., Manucharyan, G., & Klein, P. (2025). Generative data assimilation for surface ocean state estimation from multi-modal satellite observations. *Journal of Advances in Modeling Earth Systems*, *17*(8), e2025MS005063. https://doi.org/10.1029/2025ms005063

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv preprint arXiv:2202.11214.

Wang, J. L., Zhuang, H., Chérubin, L. M., Ibrahim, A. K., & Muhamed Ali, A. (2019). Medium-term forecasting of loop current eddy Cameron and eddy Darwin formation in the Gulf of Mexico with a divide-and-conquer machine learning approach. *Journal of Geophysical Research: Oceans*, *124*(8), 5586–5606. https://doi.org/10.1029/2019jc015172

Wang, X., Wang, R., Hu, N., Wang, P., Huo, P., Wang, G., et al. (2024). Xihe: A data-driven model for global ocean eddy-resolving forecasting. arXiv preprint arXiv:2402.02995.

Watt-Meyer, O., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., et al. (2024). Ace2: Accurately learning subseasonal to decadal atmospheric variability and forced responses. arXiv preprint arXiv:2411.11268.