BRIDGECAST: BRIDGING OCEAN WAVE FORECASTS TO REANALYSIS VIA FLOW MATCHING WITH EXOGENOUS VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Ocean wave forecasting is essential for maritime safety, offshore operations, and coastal resilience, yet remains challenging due to systematic biases in physics-based models. Physical models, while widely used, rely on approximations and parameterizations that limit their accuracy under complex ocean-atmosphere conditions. To enhance ocean wave forecasting, we propose BRIDGECAST, within a physics-AI hybrid framework for bias correction. BRIDGECAST is a probabilistic model based on conditional flow matching (CFM) that learns to transform physical model forecasts into reanalysis-like fields. It treats physical forecasts as corrupted observations and employs a continuous-time generative process to bridge their distribution toward that of reanalysis data. BRIDGECAST is parameterized by a Transformer-based architecture that enables spatiotemporal modeling, incorporation of exogenous atmospheric variables, and flexible inference via both ordinary and stochastic differential equation formulations. Extensive experiments on real-world datasets demonstrate that BRIDGECAST consistently outperforms state-of-the-art baselines across regions and forecast lead times.

1 INTRODUCTION

Ocean wave forecasting is critical for maritime safety, offshore operations, energy infrastructure, hazard mitigation, and coastal resilience (Sun et al., 2022; Zhang et al., 2024; Fang et al., 2026). At present, operational wave forecasts are predominantly generated by physics-based numerical models, such as the spectral wave model `WaveWatch-III` (`WW3`) (Tolman, 2009), which is often forced by atmospheric forecasts from the Global Forecast System (`GFS`) operated by the National Centers for Environmental Prediction (NCEP), and the `WaveModel` (`WAM`) (ECMWF, 2023) used within the Integrated Forecasting System (`IFS`) operated by the European Centre for Medium-Range Weather Forecasts (ECMWF). These models simulate the evolution of the wave energy spectrum through governing equations that account for wind forcing, nonlinear wave-wave interactions, and dissipation processes. While grounded in well-established physical principles and widely used in practice, they inevitably rely on approximations, parameterizations, and imperfect boundary conditions, leading to systematic biases and reduced accuracy under complex or rapidly changing ocean-atmosphere conditions such as storms. For example, Fig. 1 (left column) shows the forecast errors of `GFS-WW3` in the Gulf of Mexico and North Sea in 2025 (lead-time: 30 h), varying within $\pm$0.8m (RMSE: 0.29) and $\pm$2.0m (RMSE: 0.67). These errors are substantial and may lead to critical misjudgments, posing safety risks for maritime and coastal operations.

To correct the biases of the physical forecasters, we introduce a physics-AI hybrid framework that improves their outputs using a novel deep learning model, BRIDGECAST. BRIDGECAST is a probabilistic predictor designed to bridge the distribution of physical model forecasts to that of reanalysis data using conditional flow matching (CFM) (Lipman et al., 2023; Tong et al., 2024). Reanalysis data, such as those provided by the ECMWF Reanalysis 5th Generation (`ERA5`) (Hersbach et al., 2020), are generated through data assimilation that integrates observational data and are widely used as a proxy for ground truth in evaluating oceanic and atmospheric forecasts due to their consistency and reliability (Hersbach et al., 2020; Sharmar & Markina, 2020; Zhou et al., 2026). BRIDGECAST treats physical model forecasts as "corrupted images", which serve as conditions for a generative model based on continuous-time stochastic processes, enabling the generation of more

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

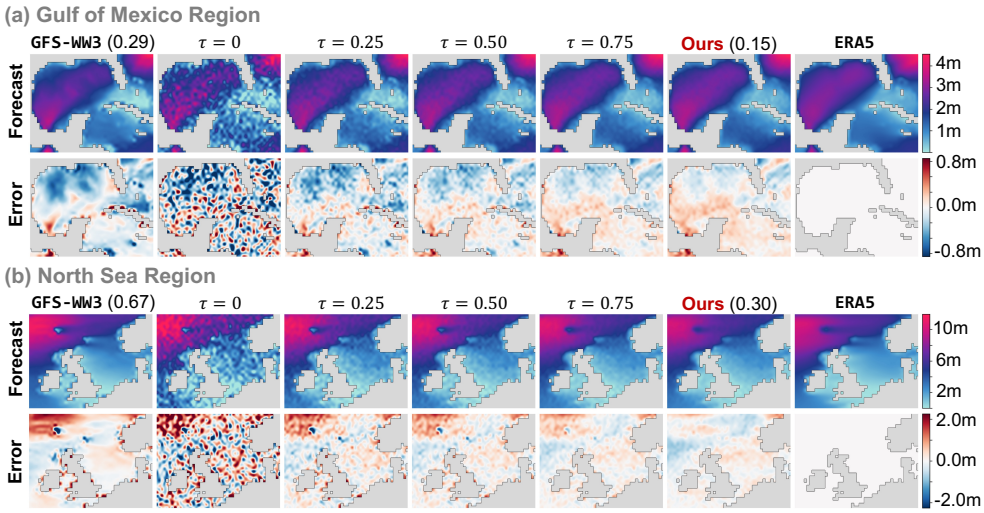


Figure 1: An illustration of BRIDGECAST's process of bridging GFS-WW3's forecasts to the reanalysis data of ERA5 in (a) Gulf of Mexico, and (b) North Sea. The numbers in brackets are RMSEs. τ is flow time. "Ours" is BRIDGECAST, which resembles ERA5 with less errors than GFS-WW3.

accurate reanalysis-like outputs. Fig. 1 illustrates the transport from GFS-WW3 forecasts to the ERA5 distribution by BRIDGECAST, where τ is the flow time, "Ours" refers to bias-corrected forecasts, and ERA5 represents the reanalysis data. BRIDGECAST progressively refines a perturbed forecast ($\tau = 0$) and ultimately reduces forecasting errors, producing outputs that closely resemble reanalysis fields.

To our knowledge, BRIDGECAST is among the first to apply CFM for bias correction in physical models for ocean wave forecasting, and can serve as a plug-and-play module for existing operational systems (*e.g.*, GFS-WW3, IFS-WAM) within a physics-AI hybrid framework that inherits the advantages of physical modeling while mitigating their limitations through "learning to bridge" distributions. Additionally, BRIDGECAST is parameterized by a novel two-head Diffusion Transformer (DiT (Peebles & Xie, 2023)) encoder-decoder capable of (1) spatiotemporal modeling of the geoscientific data; and (2) incorporating of exogenous variables such as atmospheric forcing factors that have a critical impact on ocean dynamics. Moreover, BRIDGECAST supports inferences using either an ordinary differential equation (ODE) formulation for efficient computation or a stochastic differential equation (SDE) formulation for uncertainty-aware forecasting, providing flexibility depending on practical requirements. In summary, our contributions are as follows.

- We investigate a physics-AI hybrid framework and study the problem of correcting biases in physical models for ocean wave forecasting using deep learning-based CFM.
- We propose BRIDGECAST, a novel CFM-based probabilistic model that bridges the distribution of forecasts from physical models and that of reanalysis data. BRIDGECAST adopts a Transformer-based architecture supporting spatiotemporal modeling, exogenous variables, and flexible inference.
- We conduct comprehensive experiments using GFS, IFS, and ERA5 data, and compare BRIDGECAST with state-of-the-art (SOTA) deterministic and probabilistic baselines, demonstrating its superior performance across different regions and lead times.

2 RELATED WORK

Bias Correction for Ocean Wave Forecasting. Bias correction and statistical post-processing have become increasingly important in ocean wave forecasting. Early studies focused on statistical and operational correction techniques (Vannitsem et al., 2021). Later works extended these ideas through spatiotemporal correction frameworks and climate-aware calibration methods (Ai et al., 2022; Lira Loarca et al., 2023). Recent advances increasingly employ deep learning models (Sun et al., 2022; Zhang et al., 2024; Fang et al., 2026). For example, (Sun et al., 2022) introduced a BU-Net architecture for correcting GFS-WW3 forecasts; (Zhang et al., 2024) utilized TrajGRU (Shi et al., 2017) with wind-wave coupled inputs for real-time correction; (Lira Loarca et al., 2023) applied

U-Net and Transformer components to improve medium- and long-range wave forecasts. Related studies have further explored physics-informed corrections and upstream atmospheric adjustments (Meisner et al., 2023; Peduto et al., 2025). Despite these advances, most existing learning-based approaches (Sun et al., 2022; Zhang et al., 2024; Fang et al., 2026) rely on deterministic regression objectives and do not explicitly model conditional forecast distributions or transport dynamics, leaving generative approaches such as flow matching underexplored for ocean wave forecast bias correction.

Earth System Predictions. Recent data-driven Earth models are mostly for weather and climate forecasting. Representative global forecasting systems include deterministic models (Lam et al., 2023; Pathak et al., 2022) that aim to emulate or replace weather prediction at global scales. Precipitation nowcasting models (Shi et al., 2015; Sønderby et al., 2020; Ribeiro & Pucer, 2026) focus on rapid prediction of weather events. Foundation-style Earth system models such as Earthformer (Gao et al., 2022b), ClimaX (Nguyen et al., 2023), and NeuralGCM (Kochkov et al., 2024) extend these capabilities toward multi-variable, multi-task modeling of the Earth system. Emerging probabilistic (Gong et al., 2024; Gao et al., 2025), latent-variable (Leinonen et al., 2023; Wu et al., 2024), and diffusion-based approaches (Yu et al., 2024; Gao et al., 2023) further explore generative formulations. However, these models are primarily designed for global weather or Earth system variables and are not tailored for ocean wave forecasting.

Flow-Based Models. Traditional continuous normalizing flows (CNFs) (Chen et al., 2018) model data via neural ODEs, but require costly numerical integration, which can be prohibitive for high-dimensional spatiotemporal applications. CFM (Lipman et al., 2023; Tong et al., 2024) and Rectified Flow (Liu et al., 2023) methods address this limitation by directly regressing a time-dependent vector field against a conditional probability path, avoiding the need for solving ODEs during training. Unlike diffusion models, which rely on stochastic denoising processes that typically require many iterative sampling steps, CFM enables learning ODE trajectories that transport noise to data more efficiently. FlowCast (Ribeiro & Pucer, 2026) is a recent example that applies CFM to precipitation nowcasting. More recently, stochastic interpolants (Albergo et al., 2023; 2025a) have been proposed as a unified framework connecting diffusion models and flow-based formulations. To the best of our knowledge, our work is the first to adapt CFM-based modeling to ocean wave forecast correction with explicit spatiotemporal modeling and exogenous variable conditioning.

3 The Proposed BridgeCast Model

Problem Statement. Bias correction for ocean wave forecasts is framed as an image sequence correction task with exogenous variables. In this task, a physical model generates a forecast of wave variables — which are *endougenous variables* — $\mathcal{X} = [\mathbf{X}_1, ..., \mathbf{X}_T]$ within a forecast horizon of length T, where $\mathbf{X}_t \in \mathbb{R}^{H \times W \times C}$ $(1 \leq t \leq T)$ is a forecasted image with C channels (*i.e.*, variables) at time step t. Additionally, we are given observed atmospheric variables — which are *exogenous variables* — $\mathcal{Z} = [\mathbf{Z}_1, ..., \mathbf{Z}_{T'}]$ within a time window of length T', where $\mathbf{Z}_{t'} \in \mathbb{R}^{H' \times W' \times C'}$ $(1 \leq t \leq T')$ is an image with C' channels at time step t'. It is noteworthy that $\mathcal{X}$ and $\mathcal{Z}$ can have different lengths and sizes. The goal is to predict wave variables within the physical model's forecasting horizon $\hat{\mathcal{Y}} = [\hat{\mathbf{Y}}_1, ..., \hat{\mathbf{Y}}_T]$ where $\hat{\mathbf{Y}}_t \in \mathbb{R}^{H \times W \times C}$, such that $\hat{\mathcal{Y}}$ is closer to the reanalysis data $\mathcal{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_T]$ than the physical model's raw output $\mathcal{X}$.

Preliminaries. BridgeCast is built upon CFM (Albergo et al., 2025b), which defines a conditional distribution $p_\tau(x_\tau|x_0, x_1)$, typically a Gaussian distribution with mean $(1-\tau)x_0 + \tau x_1$ and time-dependent standard deviation σ_τ, where $0 \leq \tau \leq 1$. This distribution establishes a bridge that interpolates between a source data sample $x_0 \sim p_0$ and a target data sample $x_1 \sim p_1$. It basically defines a stochastic process $x_\tau = (1-\tau)x_0 + \tau x_1 + \sigma_\tau \epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, that can bridge two arbitrary distributions p_0 and p_1 (Albergo et al., 2025a). CFM trains a velocity field $v_{\boldsymbol{\theta}}(x, \tau)$ by regressing it against the velocity target $\frac{dx_\tau}{d\tau}$. Additionally, the framework trains a denoiser function $s_{\boldsymbol{\theta}}(x, \tau)$ by regressing it against the noise-based score $\nabla_{x_\tau} \log p(x_\tau|x_0)$. Next, we specify the proposed BridgeCast under the CFM framework (§3.1), and introduce our parameterization of $v_{\boldsymbol{\theta}}(x, \tau)$ and $s_{\boldsymbol{\theta}}(x, \tau)$ (§3.2).

3.1 BridgeCast Formulation

Given a physical model's forecast $\mathcal{X}$, we aim to bridge its underlying distribution with that of reanalysis data $\mathcal{Y}$ during model training. Following the CFM framework, we define a stochastic

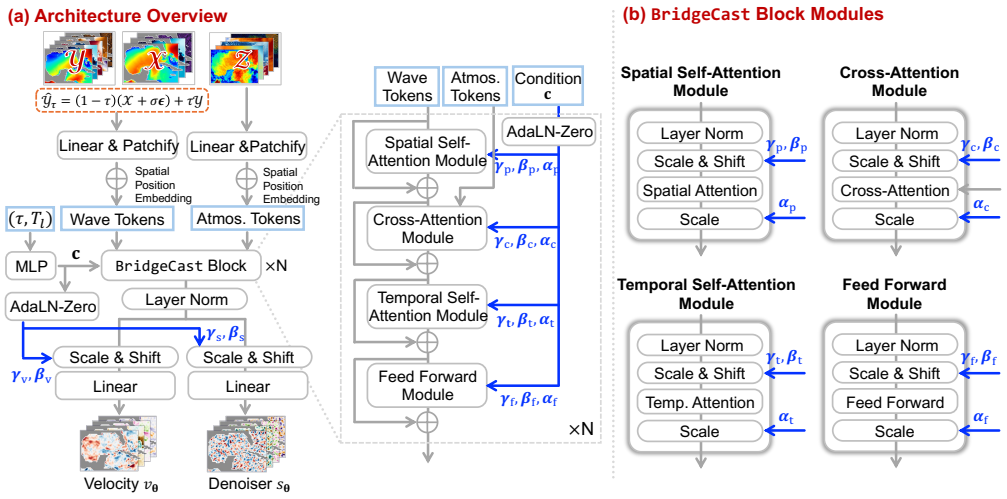


Figure 2: An illustration of the BRIDGECAST Architecture. (a) is an overview of the architecture with a zoom-in of the BRIDGECAST block. (b) illustrates the four modules in the BRIDGECAST block. τ is the flow time. T_l is the lead time. $\boldsymbol{\alpha}$'s, $\boldsymbol{\beta}$'s, $\boldsymbol{\gamma}$'s are scaling and shifting parameters.

interpolant (Albergo et al., 2025a)

$$\hat{\mathcal{Y}}_\tau = (1-\tau)\mathcal{X} + \tau\mathcal{Y} + \sigma_\tau \boldsymbol{\epsilon}, \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{1}$$

which is a flow between $\mathcal{X}$ and $\mathcal{Y}$. Specifically, we use $\sigma_\tau = (1-\tau)\sigma$ where σ is a hyperparameter so that the noise $\sigma_\tau \boldsymbol{\epsilon}$ gradually vanishes as τ moves from 0 to 1. Thus, Eq. (1) can be rewritten as

$$\hat{\mathcal{Y}}_\tau = (1-\tau)(\mathcal{X} + \sigma\boldsymbol{\epsilon}) + \tau\mathcal{Y}, \tag{2}$$

Eq. (2) can be interpreted as a flow between a perturbed source $\mathcal{X} + \sigma\boldsymbol{\epsilon}$ and the clean target $\mathcal{Y}$. This corresponds to a one-sided noisy interpolation (Albergo et al., 2025a). Empirically, it transports a smoothed version of the source sample that makes the training process more stable. Consequently, our conditional distribution $p_\tau(\hat{\mathcal{Y}}_\tau | \mathcal{X}, \mathcal{Y})$ is a Gaussian distribution $\mathcal{N}\big((1-\tau)\mathcal{X} + \tau\mathcal{Y}, (1-\tau)^2\sigma^2\mathbf{I}\big)$.

Training Objective. Starting from an SDE-based formulation that defines a stochastic probability path, we derive a hybrid training objective combining CFM and diffusion-style score matching under a unified interpolant framework. Using SDE explicitly equips BRIDGECAST with the capability of representing the probabilistic structures of ocean and atmospheric data including uncertainty and stochastic variability through an explicit noise process. In contrast, the counterpart ODE formulation provides a deterministic flow that represents the same marginal distributions but does not explicitly model intermediate stochasticity along trajectories.

Inspired by (Albergo et al., 2025b), to approximate the decomposed velocity with a parameterized velocity field $v_{\boldsymbol{\theta}}(\cdot)$ and a parameterized denoiser function $s_{\boldsymbol{\theta}}(\cdot)$, we derive a joint training objective (the full derivation process is deferred to Appendix B)

$$\ell(\boldsymbol{\theta}) = \mathbb{E}_{\substack{(\mathcal{X},\mathcal{Z},\mathcal{Y})\sim\mathcal{D}_{\text{tr}}, \\ \tau\sim\mathcal{U}(0,1), \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}} \left[\|v_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau) - (\mathcal{Y} - \mathcal{X})\|^2 + \|s_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau) - \boldsymbol{\epsilon}\|^2\right] \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathcal{D}_{\text{tr}}$ is the training dataset. The first term in Eq. (3) is a flow-style regression of velocity field. The second term is a denoising score matching that empirically alleviates overfitting of model parameters $\boldsymbol{\theta}$ during training.

3.2 BRIDGECAST ARCHITECTURE

To parameterize $v_{\boldsymbol{\theta}}(\cdot)$ and $s_{\boldsymbol{\theta}}(\cdot)$ with $\boldsymbol{\theta}$, we propose a two-head `DiT`-alike architecture that is equipped with the capability of (1) spatiotemporal encoding, and (2) exogenous variable encoding. Fig. 2 illustrates BRIDGECAST's architecture.

BRIDGECAST extends Diffusion Transformer backbone (Peebles & Xie, 2023) for spatiotemporal and exogenous modeling. First, each image in the interpolant wave variables, *i.e.*, $\hat{\mathcal{Y}}_\tau$, which is of

size $H \times W \times C$, is patchified via a 2×2 convolution into $N = HW/4$ patch tokens. A linear layer then transforms each token from C channels to d dimensions. All of the T images in $\hat{\mathcal{Y}}_\tau$ are then transformed to a tensor of size (T, N, d). Similarly, the atmospheric variables $\mathcal{Z}$ are projected to a tensor of size (T', N', d) where $N' = H'W'/4$. Additionally, each token is added with a spatial position embedding $\mathbf{P}_\text{s}$ based on sinusoidal function (Vaswani et al., 2017). The final input wave and atmospheric tokens are $\tilde{\mathbf{Y}}_\tau \in \mathbb{R}^{T \times N \times d}$ and $\tilde{\mathbf{Z}} \in \mathbb{R}^{T' \times N' \times d}$.

Conditioning via `AdaLN-Zero`. As illustrated by Fig. 2(a), the wave and atmospheric tokens are fed to the proposed BRIDGECAST block, which also takes the flow time τ in Eq. (1) and the physical model's forecasting lead time T_l as the condition, enabling the model to be aware of the stochastic process during inference and the difficulty of the forecasting task. Inspired by `DiT` (Peebles & Xie, 2023) and `FiLM` (Perez et al., 2018), we employ adaptive layer norm with zero initialization (`AdaLN-Zero`) to encode the condition: τ and T_l are mapped through sinusoidal embeddings, concatenated, and passed through a 2-layer `SiLU MLP` to produce scales (*i.e.*, $\boldsymbol{\gamma}$'s in Fig. 2), shifts (*i.e.*, $\boldsymbol{\beta}$'s in Fig. 2), and gates (*i.e.*, $\boldsymbol{\alpha}$'s in Fig. 2). Particularly, the `MLP` is initialized to output the zero-vector for all gates α's, ensuring every module in Fig. 2(b) acts as the identity function initially for stable optimization of deep stacks, resembling the skip connections in `ResNet` (He et al., 2016).

Factorized Attention Modules. Each BRIDGECAST block applies three attention modules and a feed forward module in sequence:

$$
\begin{aligned}
\tilde{\mathbf{Y}}^{l,1} &= \tilde{\mathbf{Y}}^{l} + \boldsymbol{\alpha}_\text{p} \circ \texttt{SpatialAttention}(\boldsymbol{\gamma}_\text{p} \circ \texttt{LayerNorm}(\tilde{\mathbf{Y}}^{l}) + \boldsymbol{\beta}_\text{p}) \\
\tilde{\mathbf{Y}}^{l,2} &= \tilde{\mathbf{Y}}^{l,1} + \boldsymbol{\alpha}_\text{c} \circ \texttt{CrossAttention}(\boldsymbol{\gamma}_\text{c} \circ \texttt{LayerNorm}(\tilde{\mathbf{Y}}^{l,1}) + \boldsymbol{\beta}_\text{c}, \tilde{\mathbf{Z}}) + \mathbf{P}_\text{t} \\
\tilde{\mathbf{Y}}^{l,3} &= \tilde{\mathbf{Y}}^{l,2} + \boldsymbol{\alpha}_\text{t} \circ \texttt{TemporalAttention}(\boldsymbol{\gamma}_\text{t} \circ \texttt{LayerNorm}(\tilde{\mathbf{Y}}^{l,2}) + \boldsymbol{\beta}_\text{t}) \\
\tilde{\mathbf{Y}}^{l+1} &= \tilde{\mathbf{Y}}^{l,3} + \boldsymbol{\alpha}_\text{f} \circ \texttt{FeedForward}(\boldsymbol{\gamma}_\text{f} \circ \texttt{LayerNorm}(\tilde{\mathbf{Y}}^{l,3}) + \boldsymbol{\beta}_\text{f})
\end{aligned}
\tag{4}
$$

where $l \in \{0, ..., L-1\}$ denotes the l-th BRIDGECAST block, and $\tilde{\mathbf{Y}}^0 = \tilde{\mathbf{Y}}_\tau$. $\mathbf{P}_\text{t}$ is a temporal position embedding based on sinusoidal function. $\texttt{SpatialAttention}(\cdot)$ applies self-attention to the N patch tokens in $\tilde{\mathbf{Y}}^l \in \mathbb{R}^{T \times N \times d}$ at each time step t ($1 \le t \le T$). $\texttt{TemporalAttention}(\cdot)$ applies self-attention to the T patch tokens at the same location across T time steps in parallel. $\texttt{CrossAttention}(\cdot)$ is applied to all patch tokens at each individual time step t, which uses scaled/shifted wave patch tokens in $\tilde{\mathbf{Y}}^{l,1}$ as the Query, and uses atmospheric patch tokens $\tilde{\mathbf{Z}}$ as the Key and Value. By factorizing spatiotemporal attention, BRIDGECAST block avoids $O(N^2T^2)$ cost.

Two-Head Output. The architecture terminates with two independent projection heads for realizing $v_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau)$ and $s_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau)$ and generating

$$
\begin{aligned}
\mathcal{V} &= v_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau) = \texttt{Reshape}\big(\texttt{Linear}(\boldsymbol{\gamma}_\text{v} \circ \texttt{LayerNorm}(\tilde{\mathbf{Y}}^L) + \boldsymbol{\beta}_\text{v})\big) \in \mathbb{R}^{T \times H \times W \times C} \\
\mathcal{S} &= s_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau) = \texttt{Reshape}\big(\texttt{Linear}(\boldsymbol{\gamma}_\text{s} \circ \texttt{LayerNorm}(\tilde{\mathbf{Y}}^L) + \boldsymbol{\beta}_\text{s})\big) \in \mathbb{R}^{T \times H \times W \times C}
\end{aligned}
\tag{5}
$$

which are used in Eq. (3) for model training.

3.3 TRAINING AND INFERENCE ALGORITHMS

The BRIDGECAST model learns parameter $\boldsymbol{\theta}$ via Algorithm 1 by minimizing the joint velocity and score loss in Eq (3). In the inference process, we use Euler-Maruyama method (Kloeden & Platen, 1992) to derive the forward process for SDE, leading to Algorithm 2. For brevity, the full derivation process (involving the Tweedie score) is deferred to Appendix B. It is noteworthy that when the SDE noise $\varepsilon_\text{SDE} = 0$, Algorithm 2 is equivalent to ODE. Finally, following (Lipman et al., 2023; Ribeiro & Pucer, 2026; Tong et al., 2024; Albergo et al., 2025a), due to the randomness of SDE, we generate an ensemble of forecasts to achieve robust probabilistic forecasting.

4 EXPERIMENTS

Datasets. We evaluate BRIDGECAST using public real-world oceanic and atmospheric data, including (1) `ERA5` (ECMWF ReAnalysis, 5th generation) (Hersbach et al., 2020); (2) `GFS` (Global Forecast System) (Zhou et al., 2019); and (3) `IFS` (Integrated Forecasting System) (Richardson, 2000). `ERA5` provides reanalysis data about ocean and atmosphere in different regions, which are often considered as accurate and used as ground truth for assessing different forecasting models (Hersbach et al., 2020;

Algorithm 1: BRIDGECAST Training Process

Input: (1) Training dataset $\mathcal{D}_{\text{tr}} = \{(\mathcal{X}, \mathcal{Y}, \mathcal{Z})\}$, (2) Standard deviation σ
Output: Trained model parameter $\boldsymbol{\theta}$

Initialize model parameter $\boldsymbol{\theta}$
while *not converged* **do**
  Sample a batch of data $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) \sim \mathcal{D}_{\text{tr}}$
  Sample $\tau \sim \mathcal{U}(0, 1)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\hat{\mathcal{Y}}_\tau \leftarrow (1-\tau)(\mathcal{X} + \sigma\boldsymbol{\epsilon}) + \tau\mathcal{Y}$ // Stochastic interpolant in Eq. (2)
  $\mathcal{V} \leftarrow v_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau), \mathcal{S} \leftarrow s_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau)$ // Forward pass through BRIDGECAST (Fig.2(a))
  $\ell \leftarrow \|\mathcal{V} - (\mathcal{Y} - \mathcal{X})\|^2 + \lambda\|\mathcal{S} - \boldsymbol{\epsilon}\|^2$ // Compute loss in Eq. (3)
  $\boldsymbol{\theta} \leftarrow \text{AdamW}(\ell, \boldsymbol{\theta})$ // Gradient step
end
return $\boldsymbol{\theta}$

Algorithm 2: BRIDGECAST Inference Process

Input: (1) Model parameter $\boldsymbol{\theta}$, (2) Sample $(\mathcal{X}, \mathcal{Z})$, (3) Standard deviation σ, (4) Number of SDE steps S, (5) Ensemble size K, (6) SDE noise level ε_{SDE}
Output: Ensemble of forecasts $\{\hat{\mathcal{Y}}^{(k)}\}_{k=1}^K$

$\Delta\tau \leftarrow 1/S$ // Set step size
for $k = 1, \ldots, K$ **do**
  Sample $\boldsymbol{\epsilon}^k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ // Perturbation noise
  $\hat{\mathcal{Y}}_\tau \leftarrow \mathcal{X} + \sigma\boldsymbol{\epsilon}^{(m)}$ // Initially perturbed wave variables
  for $i = 0, \ldots, S-1$ **do**
    $\tau_i \leftarrow i \cdot \Delta\tau$ // Set current step
    $\mathcal{V} \leftarrow v_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau_i), \mathcal{S} \leftarrow s_{\boldsymbol{\theta}}(\hat{\mathcal{Y}}_\tau, \mathcal{Z}, \tau_i)$ // Forward pass in Fig.2(a)
    Tweedie score: $\mathcal{S}_{\text{Tweedie}} \leftarrow -\frac{\mathcal{S}}{(1-\tau_i)\sigma}$ // Tweedie score, see Appendix B.2
    $\hat{\mathcal{Y}}_\tau \leftarrow \hat{\mathcal{Y}}_\tau + (\mathcal{V} + \varepsilon_{\text{SDE}} \cdot \mathcal{S}_{\text{Tweedie}})\,\Delta\tau$ // Euler-Maruyama drift step
    If $i < S-1$: Sample $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\hat{\mathcal{Y}}_\tau \leftarrow \hat{\mathcal{Y}}_\tau + \sqrt{2\,\varepsilon_{\text{SDE}}\,\Delta\tau}\,\boldsymbol{\xi}$
  end
  $\hat{\mathcal{Y}}^k \leftarrow \hat{\mathcal{Y}}_\tau$ // Collect ensemble forecasts
end
return $\{\hat{\mathcal{Y}}^k\}_{k=1}^K$

Sharmar & Markina, 2020; Zhou et al., 2026) (note: ERA5 does not perform forecasting). Both GFS and IFS provide wave and atmospheric forecasts by physical models for different regions, but they differ in physical models.

Dataset	Region	N_{train}	N_{val}	N_{test}	Dimension
GFS	Gulf of Mexico	193,944	2,904	2,904	33×47
	North Sea	193,896	2,904	2,904	32×48
	California Coast	192,240	2,904	2,904	35×41
IFS	Gulf of Mexico	29,520	17,376	17,184	33×47
	North Sea	29,520	17,376	17,184	32×48
	California Coast	29,520	17,376	17,184	35×41

The collected GFS data is available between 2021 and 2025 (3hr per frame), which is split into train (2021-2023)/validation (2024)/test (2025) sets. The IFS data is available between 2024 and 2025 (3hr per frame), which is split into train (2024)/validation (2025/1-2025/6)/test (2025/7-2025/12) sets. For both, there are 4 wave variables (*e.g.*, wave height, direction) and 11 atmospheric variables (*e.g.*, winds, pressure). Samples are generated by segmenting the data into 24-frame windows (*i.e.*, 72hr) with stride 2 for GFS and stride 4 for IFS. GFS's validation and tests use stride 24 to compose an evaluation of non-overlapped sample windows. It is noteworthy that each sample may be forecasted by a physical model T_l' hrs ahead ($3 \leq T_l' \leq 168$), leading to a lead-time $T_l \in [3, 240]$ for individual frames in a sample window. We categorize the lead-time into Short-Term ($3 \leq T_l \leq 24$), Mid-Term ($24 < T_l \leq 72$), and Long-Term ($72 < T_l \leq 240$) for evaluation in three regions – Gulf of Mexico, North Sea, and California Coast.

The ERA5 data covers these periods and are used for model training and evaluation. The resolution of each frame is $0.5°$. The Table above summarizes the datasets, where $N_{\text{train}}/N_{\text{val}}/N_{\text{test}}$ are the number of frames in train/validation/test set. More details about the datasets can be found in Appendix C.3.

The Compared Methods. We compare BRIDGECAST with the SOTA methods, including the vanilla physical models: (1) GFS-WW3 (Tolman, 2009), (2) IFS-WAM (ECMWF, 2023); deterministic models: (3) U-Net (Ronneberger et al., 2015), (4) EarthFormer (Gao et al., 2022b), (5) EarthFarseer (Wu et al., 2024), (6) SimVP (Gao et al., 2022a), (7) ANFO (Guibas et al., 2022), (8) TimeXer (Wang et al., 2024); and probabilistic models: (9) FlowMatch (Lipman et al., 2023), (10) DiffCast (Yu et al., 2024), (11) FlowCast (Ribeiro & Pucer, 2026). Among them, TimeXer is a recent time series model that can explicitly encode both endogenous variables and exogenous variables. For other baselines, we concatenate the wave and atmospheric variables in their inputs. We trained the baselines to correct physical model' biases as our method does except for FlowMatch which was trained to reconstruct ERA5 data without conditioning on physical forecasters. Comparing with it can highlight the importance of bridging physical models to reanalysis.

For our BRIDGECAST, we present its performance using both SDE ($\varepsilon_{\text{SDE}} \neq 0$ in Algorithm 2) and ODE ($\varepsilon_{\text{SDE}} = 0$) inferences, leading to two variants BRIDGECAST-s (*i.e.*, SDE) and BRIDGECAST-o (*i.e.*, ODE). The parameter σ is grid-searched within $[0, 1]$, ε_{SDE} is grid-searched within $(0, 0.1]$ (for a small noise) using validation set. More details about the methods and their parameter settings can be found in Appendix C.2. The implementation details are deferred to Appendix C.5.

Evaluation. Following (Ribeiro & Pucer, 2026), we evaluate wave forecasting by widely used metrics, including Root Mean Square Error (RMSE), Continuous Ranked Probability Score (CRPS) (Gneiting & Raftery, 2007), threshold-weighted CRPS (twCRPS) (Fricker et al., 2013), Critical Success Index (CSI) (Schaefer, 1990), and Heidke Skill Score (HSS) (Heidke, 1926). RMSE and CRPS measure forecasting errors, thus a small value is preferred. CSI and HSS measure hit ratio of correct predictions, thus a larger value is better. They are particularly useful for detecting extreme events (*e.g.*, storm). twCRPS is the threshold-weighted CRPS at the 95th percentile that also fits event detection. Details about the evaluation metrics can be found in Appendix C.1.

4.1 EXPERIMENTAL RESULTS

Table 1 summarizes the bias correction performance of the compared methods on GFS's wave heights across three regions and three forecasting terms, along with an overall performance without differentiating forecasting terms. For brevity, the results on other variables are deferred to Appendix C.4.2, and the results on IFS's datasets are deferred to Appendix C.4.1, which show similar insights. In Table 1, several key insights emerge: (1) All models perform better for shorter lead-times; (2) Compared with GFS-WW3, the improvements of other models suggests applying an AI-based model for correcting the physical model's forecasting bias is useful; (3) the inferior performance of FlowMatch highlights the importance to bridge physical forecasts to ERA5; (4) Deterministic models are generally inferior to probabilistic models in terms of CRPS and twCRPS, which assess their ability in modeling uncertainty; (5) Notably, BRIDGECAST, which is theoretically designed for bias correction and architecturally devised for spatiotemporal modeling with exogenous variables, outperforms the baselines in most cases, confirming its effectiveness; and (6) While BRIDGECAST-o performs best in most metrics, BRIDGECAST-s achieves the best CRPS and twCRPS, which is further confirmed by its consistent superiority in CRPS in Fig. 3, demonstrating BRIDGECAST can be easily switched by setting $\varepsilon_{\text{SDE}} = 0$ (BRIDGECAST-o) and $\varepsilon_{\text{SDE}} \neq 0$ (BRIDGECAST-s) at inference time (without retraining) based on the requirement for uncertainty-aware forecasting.

4.2 ABLATION ANALYSIS

We validate the design of BRIDGECAST through ablation studies using GFS's dataset in Gulf of Mexico. Table 2 summarizes the analysis of BRIDGECAST-o (BRIDGECAST-s uses the same trained parameter $\boldsymbol{\theta}$ thus is omitted for brevity): (a) removes exogenous atmospheric variables $\mathcal{Z}$, along with the associated cross-attention module in Fig. 2; (b) removes the spatial attention in Fig. 2; (c) removes temporal attention ; (d) removes the lead-time T_l from the condition input in Fig. 2(a); (e) removes the denoiser head $s_{\boldsymbol{\theta}}(\cdot)$ in Fig. 2(a); (f) removes the perturbation $\boldsymbol{\epsilon}$ on $\mathcal{X}$ in Eq. (2) by setting $\sigma = 0$; and (g) switches the flow matching model by diffusion model DDIM (Song et al., 2021a).

Table 2 reveals key insights into BRIDGECAST's design. (a) suggests the usefulness of atmospheric variables, though slight as they have already been incorporated by physical models. (b)(c) indicate the

Table 1: Performance comparison on GFS's dataset. ↑ (↓) indicates a larger (smaller) value is better. **Red** (Blue) values indicate the best (second-best) values per forecasting term.

		Gulf of Mexico					North Sea					California Coast				
	Model	RMSE↓	CRPS↓	CSI↑	HSS↑	twCRPS↓	RMSE↓	CRPS↓	CSI↑	HSS↑	twCRPS↓	RMSE↓	CRPS↓	CSI↑	HSS↑	twCRPS↓
Short-Term	GFS-WW3	0.2351	0.1583	0.6950	0.8096	0.0130	0.4657	0.2851	0.7457	0.8404	0.0434	0.2300	0.1600	0.6510	0.7695	0.0149
	FlowMatch	0.3025	0.1404	0.3776	0.5229	0.0261	0.7739	0.4334	0.3358	0.4287	0.0830	0.2966	0.1553	0.4571	0.5881	0.0148
	U-Net	0.1688	0.1101	0.6986	0.8127	0.0156	0.3563	0.2656	0.7297	0.8299	0.0401	0.2576	0.1946	0.6049	0.7407	0.0179
	EarthFormer	0.1939	0.1173	0.6692	0.7913	0.0187	0.3410	0.2353	0.7092	0.8118	0.0437	0.2314	0.1793	0.6044	0.7298	0.0161
	EarthFarseer	0.1734	0.1096	0.6889	0.8061	0.0164	0.2728	0.1998	0.7775	0.8633	0.0304	0.2128	0.1609	0.6281	0.7515	0.0145
	SimVP	0.1570	0.1086	0.7280	0.8341	0.0123	0.3138	0.2266	0.7494	0.8439	0.0349	0.2361	0.1788	0.5353	0.6444	0.0170
	AFNO	0.1692	0.1104	0.7015	0.8160	0.0141	0.2987	0.2076	0.7475	0.8419	0.0351	0.2074	0.1571	0.6647	0.7838	0.0129
	TimeXer	0.2008	0.1334	0.6651	0.7886	0.0144	0.4080	0.2631	0.7907	0.8735	0.0288	0.2187	0.1581	0.6731	0.7897	0.0133
	DiffCast	0.1659	0.0935	0.7074	0.8190	0.0103	0.2903	0.1617	0.7828	0.8673	0.0240	0.2184	0.1226	**0.7027**	**0.8120**	**0.0081**
	FlowCast	0.2608	0.1425	0.5689	0.7073	0.0197	0.3337	0.1789	0.7487	0.8436	0.0250	0.2224	0.1212	0.6483	0.7691	0.0106
	BridgeCast-o	**0.1348**	**0.0678**	**0.7620**	**0.8574**	**0.0087**	**0.2513**	**0.1340**	0.7915	0.8738	**0.0206**	**0.1934**	0.1134	0.6972	0.8083	0.0090
	BridgeCast-s	0.1359	0.0681	0.7591	0.8554	**0.0087**	0.2526	0.1350	**0.7937**	**0.8755**	0.0207	0.1942	**0.1070**	0.6900	0.8026	0.0084
Mid-Term	GFS-WW3	0.2554	0.1783	0.5815	0.7218	0.0121	0.5459	0.3516	0.6551	0.7743	0.0465	0.2625	0.1789	0.6488	0.7778	0.0142
	FlowMatch	0.2651	0.1382	0.2995	0.4234	0.0154	0.7462	0.4195	0.2631	0.3542	0.0617	0.3131	0.1624	0.3536	0.4735	0.0142
	U-Net	0.1626	0.1133	0.6255	0.7576	0.0105	0.4313	0.3117	0.6212	0.7462	0.0406	0.2490	0.1889	0.5857	0.7239	0.0145
	EarthFormer	0.1706	0.1175	0.5701	0.7104	0.0119	0.4190	0.2921	0.5788	0.7038	0.0430	0.2414	0.1819	0.5985	0.7384	0.0129
	EarthFarseer	0.1669	0.1156	0.6122	0.7481	0.0109	0.3943	0.2739	0.6161	0.7394	0.0390	0.2364	0.1736	0.5695	0.7088	0.0145
	SimVP	0.1738	0.1219	0.6139	0.7490	0.0107	0.4113	0.2885	0.6067	0.7310	0.0407	0.2502	0.1890	0.5898	0.7312	0.0129
	AFNO	0.1713	0.1183	0.5817	0.7220	0.0111	0.3886	0.2694	0.6317	0.7539	0.0365	0.2272	0.1708	0.6651	0.7908	0.0110
	TimeXer	0.2098	0.1447	0.5487	0.6943	0.0120	0.4966	0.3357	0.6342	0.7574	0.0369	0.2451	0.1763	0.6436	0.7745	0.0113
	DiffCast	0.1890	0.1068	0.5762	0.7181	0.0100	0.3900	0.2098	**0.6814**	**0.7953**	0.0269	0.2447	0.1359	0.6777	0.7999	0.0084
	FlowCast	0.2367	0.1362	0.5545	0.6966	0.0134	0.4040	0.2090	0.6507	0.7708	0.0268	0.2416	0.1278	0.6290	0.7617	0.0103
	BridgeCast-o	**0.1551**	0.0797	**0.6610**	**0.7860**	0.0076	**0.3531**	**0.1784**	0.6702	0.7857	**0.0266**	**0.2146**	0.1251	**0.6788**	**0.8011**	0.0089
	BridgeCast-s	0.1564	**0.0795**	0.6593	0.7847	**0.0073**	0.3538	**0.1784**	0.6696	0.7852	0.0267	0.2171	**0.1170**	0.6712	0.7956	**0.0082**
Long-Term	GFS-WW3	0.3617	0.2434	0.4250	0.5681	0.0280	0.7411	0.4903	0.3295	0.4702	0.0312	0.4293	0.2692	0.4243	0.5662	0.0353
	FlowMatch	0.3402	0.1718	0.2792	0.3896	0.0240	0.7522	0.4291	0.1340	0.2090	0.0206	0.4155	0.2089	0.2793	0.3909	0.0163
	U-Net	0.2928	0.1817	**0.4502**	**0.5906**	0.0273	0.6224	0.4229	0.2939	0.4233	0.0236	0.3850	0.2615	0.4358	0.5847	0.0180
	EarthFormer	0.2963	0.1828	0.4124	0.5508	0.0265	**0.5735**	0.3843	0.2589	0.3679	0.0209	0.3788	0.2646	0.3989	0.5273	0.0243
	EarthFarseer	0.2932	0.1836	0.4211	0.5630	0.0266	0.5789	0.3849	0.2619	0.3738	0.0216	0.3862	0.2630	0.4150	0.5573	0.0266
	SimVP	0.3021	0.1893	0.4209	0.5611	0.0270	0.5837	0.3925	0.2680	0.3872	0.0208	0.3831	0.2655	0.4108	0.5467	0.0260
	AFNO	0.2968	0.1846	0.4178	0.5610	0.0257	0.5818	0.3807	0.2924	0.4213	0.0217	0.3827	0.2605	**0.4480**	0.5917	0.0263
	TimeXer	0.3201	0.2077	0.3922	0.5354	0.0262	0.6650	0.4527	0.3131	0.4528	0.0214	0.4010	0.2683	0.3887	0.5215	0.0279
	DiffCast	0.3131	0.1712	0.4216	0.5647	0.0256	0.5942	0.3081	**0.3368**	**0.4786**	0.0195	0.4012	0.2160	0.4401	0.5831	0.0247
	FlowCast	0.3452	0.1753	0.4266	0.5656	0.0264	0.6132	0.3075	0.3129	0.4477	0.0188	0.3969	0.2084	**0.4480**	**0.5929**	0.0244
	BridgeCast-o	**0.2916**	0.1460	0.4433	0.5864	0.0227	0.5743	0.2919	0.2915	0.4209	0.0180	**0.3755**	0.2071	0.4445	0.5880	0.0245
	BridgeCast-s	0.2924	**0.1388**	0.4421	0.5851	**0.0214**	0.5753	**0.2867**	0.2949	0.4247	**0.0175**	0.3788	**0.1881**	0.4406	0.5838	**0.0226**
Overall	GFS-WW3	0.2993	0.2013	0.5244	0.6701	0.0184	0.6402	0.4119	0.5713	0.7105	0.0386	0.3334	0.2114	0.5446	0.6870	0.0225
	FlowMatch	0.3072	0.1551	0.2918	0.4128	0.0201	0.7593	0.4304	0.2288	0.3238	0.0411	0.3869	0.1933	0.3269	0.4486	0.0160
	U-Net	0.2228	0.1394	0.5512	0.6923	0.0176	0.5260	0.3609	0.5500	0.6926	0.0323	0.3976	0.2814	0.4744	0.6205	0.0288
	EarthFormer	0.2301	0.1429	0.5087	0.6531	0.0184	0.4927	0.3309	0.5166	0.6583	0.0322	0.3005	0.2133	0.5116	0.6544	0.0178
	EarthFarseer	0.2253	0.1414	0.5316	0.6767	0.0176	0.4834	0.3204	0.5508	0.6923	0.0296	0.3006	0.2065	0.5097	0.6515	0.0193
	SimVP	0.2308	0.1466	0.5393	0.6833	0.0172	0.4943	0.3327	0.5432	0.6856	0.0303	0.3067	0.2172	0.5046	0.6479	0.0186
	AFNO	0.2282	0.1432	0.5217	0.6680	0.0171	0.4839	0.3169	0.5574	0.6986	0.0290	0.2949	0.2036	0.5641	0.7050	0.0172
	TimeXer	0.2575	0.1679	0.4929	0.6423	0.0178	0.5764	0.3851	0.5720	0.7128	0.0285	0.3120	0.2096	0.5274	0.6718	0.0181
	DiffCast	0.2430	0.1304	0.5187	0.6654	0.0161	0.4915	0.2525	**0.6062**	**0.7410**	0.0230	0.3116	0.1650	0.5687	0.7090	0.0147
	FlowCast	0.2863	0.1521	0.4980	0.6426	0.0191	0.5098	0.2534	0.5730	0.7126	0.0227	0.3085	0.1578	0.5470	0.6897	0.0158
	BridgeCast-o	**0.2169**	0.1041	**0.5727**	**0.7132**	0.0136	**0.4654**	0.2286	0.5915	0.7291	0.0217	**0.2853**	0.1553	**0.5696**	**0.7098**	0.0150
	BridgeCast-s	0.2185	**0.1027**	0.5697	0.7106	**0.0131**	**0.4654**	**0.2274**	0.5917	0.7292	**0.0214**	0.2872	**0.1436**	0.5650	0.7058	**0.0139**

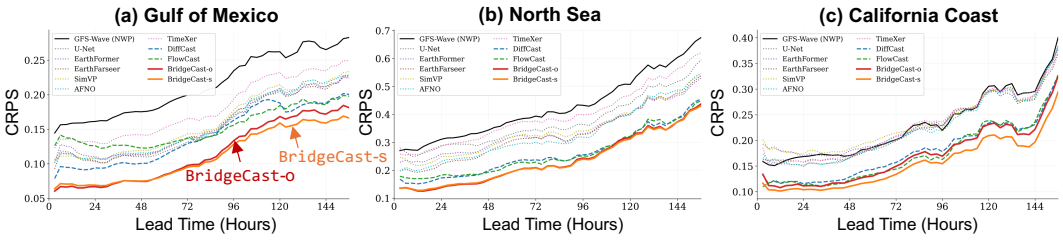


Figure 3: CRPS comparison *w.r.t.* different lead times across different regions.

importance of spatiotemporal modeling, while (d) shows conditioning on T_l enables BRIDGECAST to be aware of lead-time. The denoiser $s_\theta(\cdot)$ is beneficial in most cases as a regularizer of the model. (f) and (g) indicate perturbation and flow matching are critical to the performance of BRIDGECAST.

4.3 PERFORMANCE ANALYSIS

Qualitative Results. Fig. 4 qualitatively compares BRIDGECAST-s (BRIDGECAST-o is similar thus omitted for brevity) with the baselines on GFS's data in North Sea region (lead time 24h) when there was a storm event. FlowMatch is excluded due to its inferior performance. BRIDGECAST-s produces realistic forecasts that align better with the ERA5 reanalysis data. Other models mostly

Table 2: Ablation Analysis of BRIDGECAST-o on GFS's dataset in Gulf of Mexico. ↑ (↓) indicates a larger (smaller) value is better. "Improvement" of each ablation is relative to BRIDGECAST-o.

Model	Short-Term				Mid-Term				Long-Term			
	RMSE↓	CRPS↓	CSI↑	HSS↑	RMSE↓	CRPS↓	CSI↑	HSS↑	RMSE↓	CRPS↓	CSI↑	HSS↑
BRIDGECAST-o	0.1348	0.0678	0.7620	0.8574	0.1551	0.0797	0.6610	0.7860	0.2916	0.1460	0.4433	0.5864
(a) - Exogenous variables $\mathcal{Z}$	+0.1403	0.0689	0.7514	0.8523	0.1558	0.0819	0.6602	0.7880	0.2908	0.1466	0.4522	0.6013
Improvement	-4.08%	-1.62%	-1.39%	-0.59%	-0.44%	-2.76%	-0.13%	+0.25%	+0.26%	-0.41%	+2.03%	+2.54%
(b) - Spatial attention module	0.1517	0.0732	0.7451	0.8479	0.1603	0.0774	0.6459	0.7768	0.3019	0.1372	0.4204	0.5667
Improvement	-12.52%	-7.93%	-2.22%	-1.11%	-3.31%	+2.89%	-2.29%	-1.17%	-3.56%	+6.04%	-5.17%	-3.35%
(c) - Temporal attention	0.1433	0.0773	0.7391	0.8435	0.1673	0.0918	0.6531	0.7822	0.3104	0.1641	0.4606	0.6084
Improvement	-6.30%	-14.01%	-3.01%	-1.62%	-7.87%	-15.18%	-1.20%	-0.48%	-6.45%	-12.40%	+3.90%	+3.75%
(d) - Lead-time T_l in condition	0.1517	0.0793	0.7358	0.8417	0.1703	0.0926	0.6276	0.7631	0.2970	0.1568	0.4390	0.5876
Improvement	-12.51%	-17.04%	-3.44%	-1.83%	-9.77%	-16.22%	-5.06%	-2.92%	-1.87%	-7.38%	-0.98%	+0.20%
(e) - Denoiser head $s_\theta(\cdot)$	0.1360	0.0695	0.7568	0.8557	0.1574	0.0830	0.6618	0.7892	0.2906	0.1497	0.4464	0.5935
Improvement	-0.83%	-2.52%	-0.69%	-0.20%	-1.46%	-4.11%	+0.12%	+0.41%	+0.33%	-2.57%	+0.72%	+1.22%
(f) - Perturbation ($\sigma = 0$ in Eq. (2))	0.1340	0.0911	0.7525	0.8530	0.1660	0.1120	0.6351	0.7685	0.3191	0.1934	0.4218	0.5686
Improvement	+0.61%	-34.47%	-1.25%	-0.51%	-6.99%	-40.54%	-3.93%	-2.23%	-9.43%	-32.47%	-4.84%	-3.03%
(g) Flow matching→DDIM	0.6307	0.3463	0.2559	0.3715	0.7780	0.4486	0.1629	0.2364	0.8693	0.4976	0.1461	0.2010
Improvement	-367.71%	-411.04%	-66.43%	-56.66%	-401.42%	-462.96%	-75.36%	-69.93%	-198.13%	-240.83%	-67.04%	-65.73%

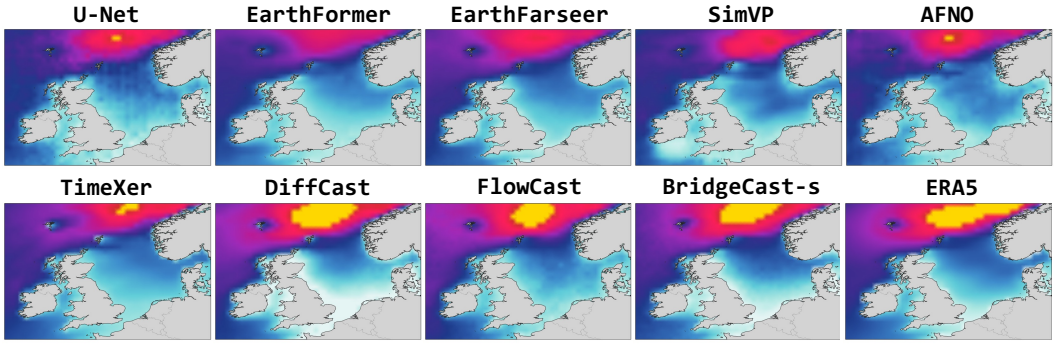


Figure 4: Qualitative comparison of BRIDGECAST and other baselines on GFS's data in North Sea region (lead time 24h) when there was a storm event.

under-forecast the wave during storm. The DiffCast's forecast, though appears reasonable, actually shifted when comparing with ERA5, leading to a higher CRPS (0.39) than BRIDGECAST-s (0.22).

Inference Time. Table 3 summarizes the averaged per-sample (a 24-frames window) inference time of BRIDGECAST across three regions for both GFS and IFS data. BRIDGECAST-s is slightly slower than BRIDGECAST-o. However, both models' inference time are relatively marginal *w.r.t.* the update time of physical models GFS-WW3 and IFS-WAM, which are both 3h, confirming the practical feasibility of BRIDGECAST as a plug-and-play module for operational systems for ocean wave forecasting.

Table 3: Inference Time of BRIDGECAST.

Data	Model	Gulf (s)	NS (s)	Cali (s)
GFS	BRIDGECAST-o	0.714	0.681	0.636
	BRIDGECAST-s	0.743	0.710	0.665
IFS	BRIDGECAST-o	0.951	0.914	0.840
	BRIDGECAST-s	0.980	0.942	0.874

Parameter Analysis. We have evaluated the impacts of parameters σ and ε_{SDE}. For brevity, these results are deferred to Appendix C.4.4, which demonstrate BRIDGECAST is not sensitive to them.

5 CONCLUSION

In this work, we proposed BRIDGECAST within a physics-AI framework for ocean wave forecasting. BRIDGECAST learns to bridge physics-based forecasts toward that of reanalysis data through a CFM process. The proposed Transformer-based architecture supports spatiotemporal modeling, exogenous variable encoding, and flexible ODE/SDE inference. Extensive experiments on GFS, IFS, and ERA5 datasets demonstrate that BRIDGECAST outperforms SOTA deterministic and probabilistic baselines across different lead times and different regions. These results highlight the potential of BRIDGECAST for improving operational geoscientific forecasting systems as a flexible plug-and-play module. Additionally, we defer our discussion of limitation and future directions to Appendix A.

AI USE STATEMENT

In this work, we have not used generative AI tools for the tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

In this paper, §3 describes the framework of BRIDGECAST. §4 describes the implementation and configuration of BRIDGECAST. We are committed to releasing the code used in our experiments to facilitate the reproducibility of this work.

REFERENCES

Bo Ai, Mengchao Yu, Jingtian Guo, Wei Zhang, Tao Jiang, Aichao Liu, Lianjie Wen, and Wenbo Li. A spatiotemporal interactive processing bias correction method for operational ocean wave forecasts. *Journal of Ocean University of China*, 21(2):277–290, 2022.

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025a.

Michael S. Albergo, Mark Goldstein, Nicholas M. Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. *arXiv preprint arXiv:2310.03725*, 2023.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *JMLR*, 26:1–80, 2025b.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

ECMWF. *IFS Documentation CY48R1, Part VII: ECMWF Wave Model*. European Centre for Medium-Range Weather Forecasts, 2023.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Can Fang, Longyu Jiang, Zeyu Wang, Quan Jin, Xingjie Jiang, and Feng Hua. Waveuformer: a bias correction model for gwsm4c wave forecasting. *Frontiers in Marine Science*, 13:1732870, 2026.

Thomas E. Fricker, Christopher A. T. Ferro, and David B. Stephenson. Three recommendations for evaluating climate predictions. *Meteorological Applications*, 20(2):246–255, 2013.

Yuan Gao, Hao Wu, Ruiqi Shu, Huanshuo Dong, Fan Xu, Rui Ray Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, et al. Oneforecast: A universal framework for global and regional weather forecasting. In *International Conference on Machine Learning*, pp. 18658–18697. PMLR, 2025.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction. In *CVPR*, 2022a.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022b.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:78621–78656, 2023.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. In *International Conference on Machine Learning*, pp. 15809–15822. PMLR, 2024.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive Fourier neural operators: Efficient token mixers for transformers. In *ICLR*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Paul Heidke. Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst. *Geografiska Annaler*, 8:301–349, 1926.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23 of *Applications of Mathematics*. Springer-Verlag, Berlin, 1992.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Andrea Lira Loarca, Peter Berg, Asuncion Baquerizo, and Giovanni Besio. On the role of wave climate temporal variability in bias correction of gcm-rcm wave simulations. *Climate Dynamics*, 61(7):3541–3568, 2023.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025. doi: 10.1038/s43247-025-02042-5. URL https://doi.org/10.1038/s43247-025-02042-5.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

Eytan Meisner, Mariano Galvagno, David Andrade, Dan Liberzon, and Raphael Stuhlmeier. Wave-by-wave forecasts in directional seas using nonlinear dispersion corrections. *Physics of Fluids*, 35 (6), 2023.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning*, pp. 25904–25938. PMLR, 2023.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Matteo Peduto, Qidong Yang, Jonathan Giezendanner, Devis Tuia, and Sherrie Wang. Observation-driven correction of numerical weather prediction for marine winds. *arXiv preprint arXiv:2512.03606*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, volume 32, 2018.

Bernardo Perrone Ribeiro and Jana Faganeli Pucer. FlowCast: Advancing precipitation nowcasting with conditional flow matching. In *ICLR*, 2026. arXiv:2511.09731.

David S Richardson. Skill and relative economic value of the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563):649–667, 2000.

Herbert Robbins. An empirical bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–163, 1956.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Joseph T. Schaefer. The critical success index as an indicator of warning skill. *Weather and Forecasting*, 5(4):570–575, 1990.

V Sharmar and M Markina. Validation of global wind wave hindcasts using era5, merra2, era-interim and cfsrv2 reanalyzes. In *IOP Conference Series: Earth and Environmental Science*, volume 606, pp. 012056. IOP Publishing, 2020.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30, 2017.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Danyi Sun, Wenyu Huang, Yong Luo, Jingjia Luo, Jonathon S Wright, Haohuan Fu, and Bin Wang. A deep learning-based bias correction method for predicting ocean surface waves in the northwest pacific ocean. *Geophysical Research Letters*, 49(23):e2022GL100916, 2022.

Hendrik L. Tolman. User manual and system documentation of WAVEWATCH III version 3.14. *NOAA/NWS/NCEP/MMAB Technical Note*, (276), 2009.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Stéphane Vannitsem, John Bjørnar Bremnes, Jonathan Demaeyer, Gavin R Evans, Jonathan Flowerdew, Stephan Hemri, Sebastian Lerch, Nigel Roberts, Susanne Theis, Aitor Atencia, et al. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):E681–E699, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *NeurIPS*, 2024.

Hao Wu, Yuxuan Liang, Wei Xiong, Zhengyang Zhou, Wei Huang, Shilong Wang, and Kun Wang. Earthfarseer: Versatile spatio-temporal dynamical systems modeling in one model. In *AAAI*, 2024.

Demin Yu, Xutao Li, Yunming Ye, Bowen Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. DiffCast: A unified framework via residual diffusion for precipitation nowcasting. In *CVPR*, 2024.

Wei Zhang, Yu Sun, Yapeng Wu, Junyu Dong, Xiaojiang Song, Zhiyi Gao, Renbo Pang, and Boyu Guoan. A deep-learning real-time bias correction method for significant wave height forecasts in the western north pacific. *Ocean Modelling*, 187:102289, 2024.

Linjiong Zhou, Shian-Jiann Lin, Jan-Huey Chen, Lucas M Harris, Xi Chen, and Shannon L Rees. Toward convective-scale prediction within the next generation global prediction system. *Bulletin of the American Meteorological Society*, 100(7):1225–1243, 2019.

Xiaobing Zhou, Frank Colberg, Debra Hudson, Yonghong Yin, Griffith Young, Christopher Bladwell, and Catherine Deburgh-Day. Comparing ocean forecasts driven with machine learning-based and physics-based atmospheric forcings. *arXiv preprint arXiv:2604.07861*, 2026.

APPENDIX

A LIMITATIONS

Two Inference Modes. The proposed BRIDGECAST has two inference methods, including SDE-based inference ($\varepsilon_{\text{SDE}} \neq 0$), leading to BRIDGECAST-s, and ODE-based inference ($\varepsilon_{\text{SDE}} = 0$), leading to BRIDGECAST-o. BRIDGECAST-s is suitable for probabilistic forecasting but is less efficient than BRIDGECAST-o, which also supports probabilistic forecasting through random initial perturbations. While the proposed model support the two approaches based on application requirements, it cannot provide the advantage of both methods simultaneously. A future work is to investigate the trade-off between efficiency and uncertainty modeling for a balanced development.

Inadequate Modeling of Look-Back Windows. The proposed BRIDGECAST correct biases of physical models instantly. The current design has not yet incorporated the historical wave and atmospheric data as look-back windows for forecasting-style bias correction. Due to the temporal dependency such as periodicity in the data, modeling look-back windows together with physical model outputs and exogenous variables may provide further improvement, which remains open for further exploration.

B THEORETICAL DEVELOPMENT

This appendix derives the joint training objective in Eq. (3) and the inference algorithm in Algorithm 2 from the Stochastic Interpolant (SI) framework of (Albergo et al., 2025b), which unifies flow matching (Lipman et al., 2023) and score-based diffusion (Song et al., 2021b; Ho et al., 2020). We instantiate BRIDGECAST as an SI with data-dependent coupling (§B.1), derive the velocity loss (§B.1), recover the score from a denoiser via Tweedie's identity (§B.2), assemble the joint objective (§B.2), and derive the SDE sampler (§B.3).

B.1 BRIDGECAST AS A STOCHASTIC INTERPOLANT

We instantiate BRIDGECAST as a stochastic interpolant (Albergo et al., 2025b, Definition 1): a stochastic process $\{x_\tau\}_{\tau\in[0,1]}$ of the form $x_\tau = I(\tau, x_0, x_1)+\gamma(\tau)z$ with $(x_0, x_1) \sim \nu$, $z \sim \mathcal{N}(0, I)$, boundary conditions $I(0) = x_0$, $I(1) = x_1$, and $\gamma(0) = \gamma(1) = 0$.

BRIDGECAST instantiation. We choose

$$I(\tau, \tilde{x}_0, x_1) = (1-\tau)\tilde{x}_0 + \tau x_1, \quad \gamma \equiv 0, \quad \tilde{x}_0 = \mathcal{X} + \sigma\boldsymbol{\epsilon}, \quad x_1 = \mathcal{Y}, \tag{6}$$

which yields the BRIDGECAST interpolant

$$\hat{\mathcal{Y}}_\tau = (1-\tau)(\mathcal{X} + \sigma\boldsymbol{\epsilon}) + \tau\mathcal{Y}, \tag{7}$$

under the data dependent coupling (Albergo et al., 2023)

$$\nu(d\tilde{x}_0, dx_1) = \int \mathcal{N}(\tilde{x}_0; \mathcal{X}, \sigma^2 I)\, \pi(d\mathcal{X}, d\mathcal{Y}), \tag{8}$$

where π is the empirical joint of paired forecast–reanalysis fields. The induced marginals are $\rho_0 = \pi_{\mathcal{X}} * \mathcal{N}(0, \sigma^2 I)$ and $\rho_1 = \pi_{\mathcal{Y}}$. The Gaussian smoothed source ρ_0 satisfies the regularity condition required for SI admissibility (Albergo et al., 2025b, Remark 11), so all results we invoke from (Albergo et al., 2025b) apply. In particular, conditional on $(\mathcal{X}, \mathcal{Y})$,

$$\hat{\mathcal{Y}}_\tau \mid \mathcal{X}, \mathcal{Y} \sim \mathcal{N}(\boldsymbol{\mu}_\tau, \sigma_\tau^2 I), \quad \boldsymbol{\mu}_\tau = (1-\tau)\mathcal{X} + \tau\mathcal{Y}, \quad \sigma_\tau = (1-\tau)\sigma. \tag{9}$$

Velocity loss. The SI velocity field is defined as the conditional expectation of the pathwise time derivative,

$$v(\tau, x) \;:=\; \mathbb{E}\Big[\partial_\tau \hat{\mathcal{Y}}_\tau \mid \hat{\mathcal{Y}}_\tau = x\Big]. \tag{10}$$

Differentiating (7) pathwise gives

$$\partial_\tau \hat{\mathcal{Y}}_\tau \;=\; \mathcal{Y} - \mathcal{X} - \sigma\boldsymbol{\epsilon}. \tag{11}$$

We learn v via L^2 regression of a parameterized field v_θ against this pathwise target:

$$\mathcal{L}_v[v_\theta] \;=\; \mathbb{E}\left\|v_\theta(\hat{\mathcal{Y}}_\tau, \mathcal{X}, \mathcal{Z}, \tau) - (\mathcal{Y} - \mathcal{X} - \sigma\boldsymbol{\epsilon})\right\|^2. \tag{12}$$

The unique minimizer of (12) over measurable functions of $\hat{\mathcal{Y}}_\tau$ is the conditional expectation (10); this is the standard L^2-projection property (Albergo et al., 2025b, Theorem 7).

Eq. (12) simplifies via the linearity of the target in $\boldsymbol{\epsilon}$. Expanding the squared norm yields a constant $\sigma^2\mathbb{E}\|\boldsymbol{\epsilon}\|^2$ term (independent of θ) and a cross-term $2\sigma\,\mathbb{E}[\langle v_\theta - (\mathcal{Y} - \mathcal{X}),\, \boldsymbol{\epsilon}\rangle]$ that vanishes under antithetic noise pairing $(\boldsymbol{\epsilon}, -\boldsymbol{\epsilon})$. The training-time loss therefore reduces to

$$\mathcal{L}_v[v_\theta] \;\equiv_\theta\; \mathbb{E}\left\|v_\theta(\hat{\mathcal{Y}}_\tau, \mathcal{X}, \mathcal{Z}, \tau) - (\mathcal{Y} - \mathcal{X})\right\|^2, \tag{13}$$

the velocity term in Eq. (3).

B.2 Score via Tweedie's Identity, Denoiser Loss, and Joint Objective

Tweedie's identity (Robbins, 1956; Efron, 2011). Let $X = \boldsymbol{\mu} + \kappa\boldsymbol{\epsilon}$ with $\boldsymbol{\mu} \sim p(\boldsymbol{\mu})$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \perp \boldsymbol{\mu}$, and $\kappa > 0$. Then

$$\mathbb{E}[\boldsymbol{\epsilon} \mid X = x] = -\kappa\, \nabla_x \log p_X(x). \tag{14}$$

Proof. Differentiating $p_X(x) = \int p(\boldsymbol{\mu})\, \mathcal{N}(x; \boldsymbol{\mu}, \kappa^2\mathbf{I})\, d\boldsymbol{\mu}$ under the integral and dividing by $p_X(x)$:

$$\nabla_x \log p_X(x) = \frac{\mathbb{E}[\boldsymbol{\mu} \mid x] - x}{\kappa^2}.$$

Since $\boldsymbol{\epsilon} = (X - \boldsymbol{\mu})/\kappa$, $\mathbb{E}[\boldsymbol{\epsilon} \mid x] = -\kappa\, \nabla_x \log p_X(x)$.

Application to BridgeCast. The conditional Gaussian (9) matches the Tweedie hypothesis with $\kappa = \sigma_\tau$. Applying (14) to the marginal ρ_τ yields

$$\nabla_x \log \rho_\tau(x) = -\frac{1}{\sigma_\tau}\, \mathbb{E}[\boldsymbol{\epsilon} \mid \hat{\mathcal{Y}}_\tau = x], \quad \tau \in [0, 1), \tag{15}$$

which matches (Albergo et al., 2025b, Eq. 2.18) under $\gamma(\tau) \leftrightarrow \sigma_\tau$.

Denoiser loss. Following noise-prediction diffusion (Ho et al., 2020), we train η_θ to regress $\boldsymbol{\epsilon}$:

$$\mathcal{L}_\eta[\eta_\theta] = \mathbb{E}\left\|\eta_\theta(\hat{\mathcal{Y}}_\tau, \mathcal{X}, \mathcal{Z}, \tau) - \boldsymbol{\epsilon}\right\|^2. \tag{16}$$

The L^2 minimizer is $\eta_\theta^\star = \mathbb{E}[\boldsymbol{\epsilon} \mid \hat{\mathcal{Y}}_\tau]$. Combined with (15), the score is recovered at inference as

$$\nabla_x \log \rho_\tau(x) = -\frac{\eta_\theta(\hat{\mathcal{Y}}_\tau, \mathcal{X}, \mathcal{Z}, \tau)}{\sigma_\tau}. \tag{17}$$

This is denoising score matching at noise level σ_τ (Vincent, 2011). The denoiser parameterization is preferred over direct score regression since the latter's target diverges as σ_τ^{-1} as $\tau \to 1$ (Albergo et al., 2025b, Remark 9), whereas $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is uniformly bounded.

Joint objective. Combining (13) (with antithetic sampling) and (16):

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\Big[\left\|v_\theta - (\mathcal{Y} - \mathcal{X})\right\|^2 + \left\|\eta_\theta - \boldsymbol{\epsilon}\right\|^2\Big], \tag{18}$$

where v_θ, η_θ share inputs $(\hat{\mathcal{Y}}_\tau, \mathcal{X}, \mathcal{Z}, \tau)$ and the expectation is over $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) \sim \mathcal{D}_{\text{tr}}$, $\tau \sim \mathcal{U}([0, 1])$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This is Eq. (3).

B.3 Inference SDE

By (Albergo et al., 2025b, Corollary 10) (cf. Song et al., 2021b for the score-based diffusion analog), for any $\varepsilon \geq 0$, the SDE become,

$$d\hat{\mathcal{Y}}_\tau = [v + \varepsilon\, s]\, d\tau + \sqrt{2\varepsilon}\, dW_\tau \tag{19}$$

preserves the marginals $\{\rho_\tau\}$ shared with the probability-flow ODE $d\hat{\mathcal{Y}}_\tau = v\, d\tau$. Substituting the score recovery (17):

$$d\hat{\mathcal{Y}}_\tau = \left[v_\theta - \frac{\varepsilon_{\text{SDE}}}{\sigma_\tau}\,\eta_\theta\right] d\tau + \sqrt{2\varepsilon_{\text{SDE}}}\, dW_\tau. \tag{20}$$

Euler–Maruyama discretization with $\Delta\tau = 1/S$, step index $i \in \{0, \dots, S-1\}$, $\tau_i = i\Delta\tau$:

$$\hat{\mathcal{Y}}_{\tau_{i+1}} = \hat{\mathcal{Y}}_{\tau_i} + \left[v_\theta - \frac{\varepsilon_{\text{SDE}}}{\sigma_{\tau_i}}\eta_\theta\right]\Delta\tau + \sqrt{2\varepsilon_{\text{SDE}}\Delta\tau}\,\boldsymbol{\xi}_i, \tag{21}$$

with $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ drawn fresh per step. Per member initial conditions $\hat{\mathcal{Y}}_0 = \mathcal{X} + \sigma\boldsymbol{\epsilon}^{(m)}$ provide ensemble spread; $\varepsilon_{\text{SDE}} = 0$ recovers the probability-flow ODE (η_θ is unused in this limit). A numerical floor $\sigma_\tau \geq 10^{-3}$ prevents division by zero as $\tau \to 1$ (Albergo et al., 2025b, Algorithm 5). This yields Algorithm 2.

C Experiments

C.1 Evaluation Metrics

Root Mean Square Error (RMSE). Computed on the ensemble mean $\bar{\mathcal{Y}}$ for stochastic methods and on the single prediction for deterministic methods:

$$\text{RMSE} = \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega}\left(\bar{\mathcal{Y}}_i - \mathcal{Y}_i\right)^2}, \qquad \bar{\mathcal{Y}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\mathcal{Y}}^{(m)}, \tag{22}$$

where Ω denotes the set of ocean pixels. Lower is better.

Continuous Ranked Probability Score (CRPS). A strictly proper scoring rule that jointly assesses calibration and sharpness (Gneiting & Raftery, 2007). For an ensemble of size M, we use the unbiased fair estimator (Matheson & Winkler, 1976):

$$\text{CRPS} = \frac{1}{M}\sum_{m=1}^{M}\left\|\hat{\mathcal{Y}}^{(m)} - \mathcal{Y}\right\|_1 - \frac{1}{2M(M-1)}\sum_{m=1}^{M}\sum_{m'=1}^{M}\left\|\hat{\mathcal{Y}}^{(m)} - \hat{\mathcal{Y}}^{(m')}\right\|_1. \tag{23}$$

For deterministic baselines (M=1), CRPS reduces to the mean absolute error $|\bar{\mathcal{Y}} - \mathcal{Y}|$.

Threshold-Weighted CRPS (twCRPS). A weighted variant of CRPS that emphasizes performance on extreme events (Fricker et al., 2013). Let q_{95} denote the train-set 95th percentile of SWH. The twCRPS uses a chaining function $v(x) = \max(x, q_{95})$ to penalize errors only above the threshold:

$$\text{twCRPS} = \mathbb{E}_{\hat{\mathcal{Y}}}\left|v(\hat{\mathcal{Y}}) - v(\mathcal{Y})\right| - \tfrac{1}{2}\,\mathbb{E}_{\hat{\mathcal{Y}},\hat{\mathcal{Y}}'}\left|v(\hat{\mathcal{Y}}) - v(\hat{\mathcal{Y}}')\right|, \tag{24}$$

estimated via the same fair-ensemble formula as (23) applied to the chained outputs $v(\hat{\mathcal{Y}}^{(m)})$. Reported in meters.

Critical Success Index (CSI). A categorical skill score for binary event detection (Schaefer, 1990). We threshold the ensemble mean and ground truth at the train-set 95th percentile to obtain binary masks, then compute the contingency table (true positives TP, false positives FP, false negatives FN):

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{25}$$

CSI ranges in $[0, 1]$, rewards correct event hits, and penalizes both missed and false alarms. It is widely used in nowcasting evaluation (Ribeiro & Pucer, 2026).

Heidke Skill Score (HSS). A categorical score that corrects CSI for chance agreement (Heidke, 1926). Using the same contingency table:

$$\text{HSS} = \frac{2(\text{TP}\cdot\text{TN} - \text{FP}\cdot\text{FN})}{(\text{TP}+\text{FN})(\text{FN}+\text{TN}) + (\text{TP}+\text{FP})(\text{FP}+\text{TN})}. \tag{26}$$

HSS is bounded above by 1; values above 0 indicate skill exceeding a random forecast.

Mean Wave Period and Direction CRPS. For the auxiliary wave variables we use channel-specific CRPS variants. **CRPS-MWP** (s) applies (23) to mean wave period. **CRPS-MWD** (°) is the mean angular distance between the ensemble-mean direction $\bar{\theta}$ (recovered from the $(\sin, \cos)$ ensemble means) and the target θ^*:

$$\text{CRPS-MWD} = \min\big(|\bar{\theta} - \theta^*|,\ 360^\circ - |\bar{\theta} - \theta^*|\big). \tag{27}$$

Both are reported per channel.

C.2 Baselines

Baselines. We compare BridgeCast against eight representative baselines spanning deterministic spatiotemporal forecasters and probabilistic generative models, all trained on the same paired $(\mathcal{X}, \mathcal{Z}, \mathcal{Y})$ tuples and evaluated under the same protocol.

The deterministic baselines cover four major architectural families relevant to gridded geoscience forecasting. **U-Net** (Ronneberger et al., 2015) is the canonical encoder–decoder convolutional baseline, configured with base channel width 64 (GFS) or 32 (IFS) to match training-set size. **SimVP** (Gao et al., 2022a) uses depthwise-separable spatial–temporal convolutional blocks for parameter-efficient video prediction. **AFNO** (Guibas et al., 2022) replaces self-attention with adaptive Fourier neural operators and is the standard spectral-mixing baseline used in modern weather emulators (e.g. FourCastNet). **EarthFormer** (Gao et al., 2022b) introduces cuboid attention along factorized spatial and temporal axes; we use the $[4, 4, 4]$ cuboid configuration tuned for regional grids. **EarthFarseer** (Wu et al., 2024) augments a SimVP-style backbone with a Fourier-Transformer (FoTF) module to capture long-range dependencies; it is our most parameter-heavy deterministic baseline ($\sim$25.6 M). **TimeXer** (Wang et al., 2024) is a transformer that explicitly separates exogenous (atmospheric) and endogenous (wave) channels through cross-attention with a residual prediction head — the closest deterministic analog to our two-head design.

The probabilistic baselines fall into two families. **DiffCast** (Yu et al., 2024) is a denoising diffusion model trained from Gaussian noise to the target distribution with 1000 training steps and 250-step DDIM sampling at inference; we use its standard U-Net backbone with channel multipliers $(1, 2, 4)$. **FlowCast** is a Conditional Flow Matching (Lipman et al., 2023) variant we construct on top of an EarthFormer backbone, sampled with a 10-step Euler ODE solver; it shares the bridge-style training objective with BridgeCast but uses a Gaussian source rather than a forecast-anchored source.

BridgeCast is comparable in parameter count to FlowCast and lighter than DiffCast and EarthFarseer. This places BridgeCast's gains squarely on the modeling side rather than on raw scale.

C.3 Data Details

C.3.1 Forecast Archives

GFS-Wave (primary). We use the operational NCEP GFS-Wave (WAVEWATCH III) forecast (Tolman, 2009), taking all four daily initialization cycles (00, 06, 12, 18 UTC) and extracting four wave channels: significant wave height (SWH), mean wave period (MWP), and mean wave direction encoded as $(\sin\theta_d, \cos\theta_d)$ to avoid the $0/2\pi$ discontinuity. GFS-Wave is distributed at 0.5° on the regional crops we use, matching the ERA5 wave grid which is produced only at 0.5° (Hersbach et al., 2020). The regional GFS-Wave archive begins in March 2021 and bounds the GFS training-set size.

IFS-WAM (cross-system). We additionally use ECMWF IFS-WAM forecasts (ECMWF, 2023) from the public ECMWF open-data archive at 00 and 12 UTC, lead times $0, 3, \ldots, 240$ h. Native IFS-WAM resolution is 0.25°; we stride-2 subsample to 0.5° on the same regional bounding boxes as GFS-Wave so the input grid is bit-identical to the GFS training grid. We cap IFS at $\ell_{\max} = 168$ h so the evaluation contract is identical to GFS-Wave. The IFS open-data archive begins 2024-03-01, which limits the IFS training budget to roughly one seventh of the GFS budget — a deliberate stress test of the corrector in the small-data regime.

Atmospheric exogenous $\mathcal{Z}$. For each forecast system, atmospheric fields are taken from the same forecast cycle and lead time as the wave forecast and regridded to the wave grid before training, so $\mathcal{Z}$ and $\hat{\mathcal{Y}}^{\text{p}}$ share an identical $H \times W$ spatial extent and no spatial resampling occurs inside the

network. $\mathcal{Z}$ contains $C' = 11$ channels: raw 10 m wind components (u_{10}, v_{10}) and mean sea-level pressure (msl), plus eight deterministic derivations chosen to make wind-forcing physics visible to the network — 10 m wind speed, wind direction $(\sin, \cos)$, u_{10}^2, the per-frame finite-difference time derivatives of u_{10}, v_{10}, and msl, and the wind-forcing magnitude $u_{10}^2 + v_{10}^2$. The derived channels are mathematically redundant given the raw fields but accelerate convergence on the small training budgets used here, following the inductive-bias argument made by CorrDiff (Mardani et al., 2025) for its derived input fields. $\mathcal{Z}$ enters BRIDGECAST as conditioning only and is never part of the supervised target.

Target $\mathcal{Y}$. The supervised target is the corresponding ERA5 wave field (Hersbach et al., 2020) at the same valid times and the same $0.5°$ grid, with the same four channels. ERA5 wave variables are assimilated against altimeter and buoy observations and are the standard reference for wave-model verification.

C.3.2 REGIONS AND SAMPLE CONSTRUCTION

Study regions. We evaluate on three coastal regions spanning different wave climates: the **Gulf of Mexico** (15.0–31.0° N, 98.0–75.0° W; grid 33×47; ≈ 0.85 ocean), a fetch-limited basin dominated by tropical cyclones; the **North Sea** (48.0–63.5° N, 4.0° W–19.5° E; grid 32×48; ≈ 0.78 ocean), an extratropical-storm regime with the highest baseline GFS-Wave errors of the three; and the **California coast** (30.0–47.0° N, 135.0–115.0° W; grid 35×41; ≈ 0.81 ocean), a long-fetch swell regime in which the raw forecast is already comparatively strong and residual-correction gains are hardest to obtain. Land cells are excluded from every loss and every metric, for BRIDGECAST and for every baseline.

Sample geometry. Each sample is one trajectory of $S = 24$ frames at $\Delta t = 3$ h, covering a 72-hour window starting at lead ℓ. Frame index $s \in \{0, \dots, S-1\}$ within a sample issued at lead ℓ has valid lead time $\ell + s\Delta t$ hours. Fixing $S = 24$ for every sample keeps the temporal token count identical across the entire 3–168 h forecast horizon while allowing arbitrary positions inside the 7-day window via the lead offset ℓ.

Lead-time conditioning: one model, all leads. BRIDGECAST trains a single model jointly across all forecast lead times, conditioning on a per-frame valid-lead encoding $\tilde{\ell}_s = (\ell + s\Delta t)/\ell_{\max}$ with $\ell_{\max} = 168$ h, supplied to AdaLN-Zero alongside the interpolant time τ.

Lead-time sampling rationale. Operational use of marine forecasts concentrates at short leads (next-day decisions for shipping, port operations, offshore work, and fisheries dominate query traffic). A uniform-in-ℓ training distribution would over-represent long leads relative to where forecast skill matters in deployment. We therefore sample $\ell \sim \log \mathcal{U}[3, 168]$ h on the 3 h grid, which (i) concentrates training mass in proportion to operational utility and (ii) acts as importance-sampling reweighting of the joint loss across leads — the same role that non-uniform timestep sampling plays in modern diffusion training, where it has been shown to reduce gradient variance and accelerate convergence (Ho et al., 2020; Song et al., 2021b).

GFS-Wave (NOMADS, $0.5°$, four daily cycles since March 2021) materializes the ℓ-sampler at export time: $\ell \sim \log \mathcal{U}[3, 168]$ h is drawn independently per cycle, yielding a variable count of trajectories per cycle (mean ≈ 2.3, range 1–10). The variability reflects the per-cycle stochastic draw and the constraint that cycles late in the archive period drop trajectories at long ℓ where the ERA5 target window extends past the cache end. Validation and test use a 72 h non-overlapping initialization stride for independent samples.

IFS-WAM (ECMWF Open Data via AWS public replica bucket, two daily cycles since 2024-03-01) is much smaller: the official ECMWF Open Data Portal retains only the most recent ∼12 forecast runs, and the full MARS historical archive requires a commercial Service Agreement we do not hold. This caps the IFS training budget at roughly $1/7$ the GFS budget. Given the small archive, we adopt deterministic per-cycle export with exactly two ℓ draws per (00/12 UTC) initialization from the 3 h grid $\{3, 6, \dots, 168\}$ h. Validation and test retain every available initialization; trajectories therefore overlap in valid time (median 3 h spacing), and test-set metrics aggregate over unique initialization times to avoid double-counting overlapping windows.

Table 4: Performance comparison on IFS's dataset. ↑ (↓) indicates a larger (smaller) value is better. **Red** (Blue) values indicate the best (second-best) values per forecasting term.

		Gulf of Mexico					North Sea					California Coast				
	Model	RMSE↓	CRPS↓	CSI↑	HSS↑	twCRPS↓	RMSE↓	CRPS↓	CSI↑	HSS↑	twCRPS↓	RMSE↓	CRPS↓	CSI↑	HSS↑	twCRPS↓
Short-Term	IFS-Wave	0.2110	0.1397	0.6829	0.8015	0.0133	0.4328	0.2771	0.6630	0.7855	0.0202	0.2342	0.1676	0.6385	0.7659	0.0181
	U-Net	0.1564	0.0936	0.6889	0.8076	0.0123	0.2687	0.1944	0.5982	0.7229	0.0178	0.1908	0.1302	0.6596	0.7836	0.0130
	EarthFormer	0.1738	0.1067	0.6049	0.7380	0.0152	0.2940	0.2111	0.5712	0.6992	0.0213	0.2030	0.1412	0.6088	0.7357	0.0146
	EarthFarseer	0.1372	0.0874	0.6808	0.8014	0.0118	0.2559	0.1794	0.6006	0.7274	0.0176	0.1820	0.1274	0.6578	0.7799	0.0124
	SimVP	0.1758	0.1117	0.6398	0.7698	0.0136	0.3256	0.2387	0.6331	0.7615	0.0171	0.2122	0.1560	0.6696	0.7897	0.0130
	AFNO	0.1540	0.0980	0.7150	0.8266	0.0113	0.2928	0.2119	0.6274	0.7533	0.0164	0.1821	0.1206	0.6883	0.8036	0.0116
	TimeXer	0.1590	0.1002	0.7651	0.8611	0.0083	0.3529	0.2067	0.7373	0.8407	0.0115	0.1733	0.1115	**0.7746**	**0.8672**	0.0080
	DiffCast	0.1936	0.1127	0.7329	0.8385	0.0069	0.4789	0.2790	0.6678	0.7908	0.0115	0.2326	0.1448	0.7141	0.8251	0.0079
	FlowCast	0.1770	0.0922	0.6870	0.8058	0.0096	0.3907	0.2170	0.5323	0.6723	0.0206	0.1382	0.0724	0.7487	0.8476	0.0065
	BridgeCast-o	**0.1072**	**0.0527**	**0.7808**	**0.8719**	0.0061	**0.1829**	**0.0957**	**0.7619**	**0.8578**	0.0079	**0.1296**	0.0671	0.7738	0.8660	0.0063
	BridgeCast-s	0.1107	0.0619	0.7785	0.8704	**0.0057**	0.1834	0.0961	0.7594	0.8560	**0.0078**	0.1298	**0.0665**	0.7744	0.8664	**0.0062**
Mid-Term	IFS-Wave	0.2298	0.1515	0.6431	0.7727	0.0143	0.4705	0.3054	0.6164	0.7474	0.0288	0.2501	0.1708	0.6090	0.7428	0.0172
	U-Net	0.1721	0.1049	0.6271	0.7602	0.0129	0.3136	0.2174	0.5900	0.7192	0.0223	0.2054	0.1359	0.5862	0.7209	0.0137
	EarthFormer	0.1932	0.1166	0.5036	0.6380	0.0165	0.3405	0.2349	0.5531	0.6855	0.0252	0.2267	0.1502	0.5103	0.6456	0.0155
	EarthFarseer	0.1794	0.1100	0.5840	0.7240	0.0140	0.3286	0.2215	0.5439	0.6780	0.0246	0.2200	0.1461	0.5241	0.6610	0.0142
	SimVP	0.1972	0.1255	0.5375	0.6816	0.0152	0.3620	0.2589	0.6127	0.7444	0.0225	0.2254	0.1609	0.5869	0.7208	0.0137
	AFNO	0.1751	0.1109	0.6340	0.7657	0.0125	0.3296	0.2328	0.6298	0.7585	0.0202	0.2036	0.1335	0.5918	0.7232	0.0126
	TimeXer	0.1880	0.1195	0.6734	0.7970	0.0112	0.4064	0.2533	0.6623	0.7851	0.0191	0.2062	0.1319	0.6381	0.7656	0.0112
	DiffCast	0.2491	0.1393	0.6569	0.7848	0.0092	0.4627	0.2668	0.6335	0.7625	0.0167	0.2460	0.1487	0.6316	0.7621	0.0096
	FlowCast	0.1951	0.1010	0.6428	0.7729	0.0098	0.4177	0.2256	0.5370	0.6771	0.0227	0.1688	0.0827	0.6479	0.7695	0.0092
	BridgeCast-o	**0.1465**	**0.0699**	**0.6875**	**0.8071**	0.0092	**0.2627**	**0.1331**	**0.6956**	**0.8102**	0.0141	**0.1627**	0.0825	**0.6635**	**0.7844**	0.0091
	BridgeCast-s	0.1487	0.0740	0.6862	0.8061	**0.0085**	0.2649	**0.1331**	0.6949	0.8097	**0.0135**	0.1638	**0.0809**	0.6623	0.7834	**0.0088**
Long-Term	IFS-Wave	0.3200	0.2059	0.5205	0.6718	0.0185	0.7497	0.4870	0.4374	0.5838	0.0461	0.4004	0.2548	0.4431	0.5860	0.0242
	U-Net	0.2703	0.1573	0.4640	0.6191	0.0172	0.6267	0.3940	0.4155	0.5596	0.0352	0.3532	0.2147	0.4424	0.5881	0.0190
	EarthFormer	0.2581	0.1537	0.3730	0.5033	0.0195	0.5891	0.3869	0.3661	0.4956	0.0369	0.3197	0.2057	0.3989	0.5367	0.0191
	EarthFarseer	0.2604	0.1563	0.4227	0.5738	0.0184	0.6132	0.3974	0.3687	0.5071	0.0383	0.3217	0.2062	0.3833	0.5133	0.0192
	SimVP	0.2636	0.1645	0.4095	0.5575	0.0188	0.5905	0.4026	0.4316	0.5789	0.0361	0.3201	0.2153	0.4485	0.5919	0.0190
	AFNO	0.2770	0.1650	0.4861	0.6407	0.0166	0.6472	0.4082	0.4319	0.5799	0.0341	0.3623	0.2168	0.4302	0.5763	0.0181
	TimeXer	0.2605	0.1621	0.5138	0.6664	0.0165	0.6315	0.4152	0.4427	0.5905	0.0348	0.3134	0.1976	0.4655	0.6085	0.0180
	DiffCast	0.3126	0.1644	0.5092	0.6637	0.0141	0.7034	0.3765	0.4362	0.5852	0.0317	0.3552	0.1899	0.4511	0.5980	0.0160
	FlowCast	0.2720	0.1310	0.5024	0.6556	0.0140	0.6636	0.3325	0.3924	0.5314	0.0331	**0.2849**	**0.1377**	0.4759	0.6177	**0.0153**
	BridgeCast-o	**0.2296**	0.1097	**0.5285**	**0.6786**	0.0144	**0.5619**	0.2945	0.4626	0.6098	0.0315	0.2954	0.1524	**0.4786**	**0.6211**	0.0175
	BridgeCast-s	0.2307	**0.1061**	0.5278	0.6779	**0.0135**	0.5689	**0.2816**	**0.4629**	**0.6099**	**0.0296**	0.3003	0.1463	0.4774	0.6200	0.0159
Overall	IFS-Wave	0.2639	0.1698	0.6055	0.7433	0.0157	0.5839	0.3681	0.5526	0.6934	0.0342	0.3108	0.2007	0.5537	0.6942	0.0199
	U-Net	0.2113	0.1225	0.5774	0.7203	0.0144	0.4493	0.2788	0.5237	0.6620	0.0266	0.2665	0.1636	0.5456	0.6866	0.0155
	EarthFormer	0.2169	0.1289	0.4705	0.6082	0.0174	0.4434	0.2875	0.4846	0.6212	0.0291	0.2620	0.1694	0.4836	0.6220	0.0167
	EarthFarseer	0.2088	0.1243	0.5384	0.6850	0.0153	0.4478	0.2808	0.4799	0.6201	0.0289	0.2580	0.1658	0.4923	0.6312	0.0158
	SimVP	0.2215	0.1380	0.5038	0.6514	0.0163	0.4555	0.3090	0.5442	0.6862	0.0270	0.2625	0.1802	0.5487	0.6881	0.0155
	AFNO	0.2155	0.1289	0.5915	0.7322	0.0138	0.4670	0.2940	0.5536	0.6945	0.0249	0.2696	0.1621	0.5472	0.6865	0.0145
	TimeXer	0.2143	0.1327	0.6274	0.7620	0.0128	0.4954	0.3070	0.5838	0.7216	0.0240	0.2473	0.1533	0.5934	0.7289	0.0133
	DiffCast	0.2678	0.1449	0.6146	0.7523	0.0107	0.5641	0.3084	0.5597	0.7014	0.0216	0.2885	0.1629	0.5790	0.7183	0.0116
	FlowCast	0.2241	0.1108	0.5994	0.7389	0.0112	0.5185	0.2636	0.4833	0.6261	0.0264	**0.2156**	**0.1016**	0.5986	0.7301	**0.0111**
	BridgeCast-o	**0.1777**	**0.0823**	**0.6417**	**0.7728**	0.0107	**0.3933**	0.1876	**0.6071**	**0.7408**	0.0198	0.2175	0.1061	**0.6066**	**0.7385**	0.0118
	BridgeCast-s	0.1794	0.0842	0.6404	0.7718	**0.0100**	0.3955	**0.1851**	**0.6071**	**0.7408**	**0.0190**	0.2196	0.1035	0.6050	0.7371	0.0112

Lead-time bands. We report results in three operationally meaningful bands chosen *a priori*: **Short-Term** (3–24 h), **Mid-Term** (27–72 h), and **Long-Term** (75–168 h).

C.3.3 Splits and Sample Counts

Splits are strictly chronological. For GFS-Wave: training 2021-03–2023-12, validation 2024, test 2025, yielding ∼8,080 training trajectories per region (Gulf 8,082; NS 8,090; California 8,010) and 121 non-overlapping validation and 121 test trajectories per region. For IFS-WAM: training 2024-03–2024-12, validation 2025-01–2025-06, test 2025-07–2025-12, yielding 1,230 training, 724 validation, and 693 test trajectories per region (716 raw, with 23 exact $(init, \ell)$ collisions deduplicated before metric computation). Normalization statistics are computed from the training split only and frozen for validation and test.

C.4 Additional Results

C.4.1 IFS-Wave to ERA5 Benchmark

Table 4 demonstrates that BridgeCast generalizes beyond GFS: when adapted to the IFS-Wave forecast system, the model systematically corrects forecast errors across all three regions and three lead bands and the overall evaluation. Both BridgeCast-o and BridgeCast-s remain top 2 across nearly every (region, lead band) cell.

Table 5: Mean-wave-period (MWP) and mean-wave-direction (MWD) CRPS on GFS's dataset. **Red** (Blue) values indicate the best (second-best) per column. MWP CRPS in seconds (Gaussian CRPS for ensembles, $|\text{mean}-y|$ MAE for $M=1$); MWD CRPS in degrees (mean angular distance, regardless of M). Lead bands as in Table 1.

		Gulf of Mexico		North Sea		California Coast	
	Model	CRPS-MWP (s)	CRPS-MWD (°)	CRPS-MWP (s)	CRPS-MWD (°)	CRPS-MWP (s)	CRPS-MWD (°)
Short-Term	GFS-Wave	1.1789	18.1443	2.1466	23.6186	3.1095	24.5094
	U-Net	0.3038	9.6514	0.5430	15.6090	0.7313	14.2569
	EarthFormer	0.3223	10.1632	0.4402	11.7093	0.5345	9.9825
	EarthFarseer	0.3196	10.7905	0.4456	11.5879	0.5360	10.3003
	SimVP	0.3373	12.6680	0.4635	12.7184	0.5787	11.2427
	AFNO	0.3150	10.2438	0.4403	11.6142	0.5559	**9.6755**
	TimeXer	0.4500	16.1298	0.7732	20.5555	0.8732	14.0816
	DiffCast	0.4403	11.5434	0.8880	13.8405	2.0067	17.2921
	FlowCast	0.3789	15.2576	0.3947	16.7287	**0.4388**	12.3990
	BridgeCast-o (ours)	0.2000	**8.7965**	0.3033	**9.9402**	0.7044	10.7848
	BridgeCast-s (ours)	**0.1981**	8.8533	**0.3018**	9.9519	0.6726	10.6976
Mid-Term	GFS-Wave	1.2186	19.5821	2.0705	25.4573	3.0214	26.0714
	U-Net	0.3154	11.2001	0.5487	17.4141	0.7224	14.9428
	EarthFormer	0.3237	11.2796	0.4722	13.4342	0.5392	10.5984
	EarthFarseer	0.3254	12.2747	0.4737	13.7686	0.5240	**10.1840**
	SimVP	0.3475	13.9634	0.4986	14.1441	0.5520	11.3051
	AFNO	0.3303	11.7190	0.4681	13.9231	0.5441	10.5546
	TimeXer	0.4559	17.3956	0.7616	22.1130	0.8191	14.7116
	DiffCast	0.4584	13.1540	0.8543	16.2941	1.9572	19.0256
	FlowCast	0.3706	16.4341	0.4014	17.1562	**0.4162**	12.5427
	BridgeCast-o (ours)	0.2269	**10.4695**	0.3167	12.5428	0.7172	12.5543
	BridgeCast-s (ours)	**0.2205**	10.5735	**0.3164**	**12.5278**	0.6610	12.5271
Long-Term	GFS-Wave	1.2491	24.9397	2.0496	33.7038	3.1956	30.2076
	U-Net	0.3946	17.5566	0.7026	27.8330	0.8941	17.6861
	EarthFormer	0.3983	17.6914	0.6252	**24.2660**	0.6941	13.9438
	EarthFarseer	0.3941	18.4182	0.6323	24.8971	0.7004	**13.9104**
	SimVP	0.4305	19.9169	0.6431	25.4706	0.7044	14.3971
	AFNO	0.4035	18.0726	0.6339	25.0677	0.7070	14.2969
	TimeXer	0.5172	22.8283	0.8762	31.0028	0.9850	18.1178
	DiffCast	0.5265	19.2589	0.8998	26.7177	2.2066	23.0010
	FlowCast	0.4074	21.5450	0.5289	27.6177	**0.5695**	15.9678
	BridgeCast-o (ours)	0.3106	**17.3444**	0.4756	24.3245	0.8758	15.8839
	BridgeCast-s (ours)	**0.2936**	17.4844	**0.4560**	24.2744	0.8050	15.9087
Overall	GFS-Wave	1.2260	21.5010	2.0677	29.2188	3.1021	27.4883
	U-Net	0.3445	13.4756	0.6217	22.2269	0.7895	15.9149
	EarthFormer	0.3522	13.6378	0.5423	18.4381	0.5978	11.8002
	EarthFarseer	0.3510	14.4893	0.5474	18.8612	0.5940	**11.6156**
	SimVP	0.3782	16.1239	0.5645	19.4154	0.6140	12.4824
	AFNO	0.3567	14.0138	0.5447	19.0343	0.6087	11.8850
	TimeXer	0.4790	19.3684	0.8173	26.2254	0.8899	15.9263
	DiffCast	0.4825	15.3532	0.8787	21.0368	2.0611	20.3509
	FlowCast	0.3859	18.2891	0.4616	22.0955	**0.4779**	13.8369
	BridgeCast-o (ours)	0.2560	**12.9524**	0.3913	**17.9221**	0.7777	13.6248
	BridgeCast-s (ours)	**0.2474**	13.1022	**0.3829**	17.9533	0.7227	13.5947

C.4.2 Mean Wave Period and Direction (GFS)

Table 5 reports CRPS for the auxiliary wave variables (mean wave period, mean wave direction) on GFS. BridgeCast substantially improves over GFS-Wave for both channels — e.g., overall CRPS-MWP drops from 1.23 s to 0.25 s on the Gulf of Mexico. The California region is more competitive: FlowCast edges BridgeCast on long-term MWP, consistent with the long-fetch swell regime where GFS-Wave is already comparatively strong.

C.4.3 Mean Wave Period and Direction (IFS)

We observed that IFS-Wave is systematically better than GFS-Wave on MWP and MWD, narrowing the gap to the post-processed output. BridgeCast still wins on most cells; the exception is California long-term MWP, where FlowCast edges ahead. We suspect this is due to the Pacific long-fetch swell regime, in which much of the wave signal arrives from outside the cropped domain, limiting the contribution of BridgeCast's in-domain wind–wave coupling.

C.4.4 SDE Noise Sweep

We observed three patterns. On GFS, the curves are flat over a wide ε_{SDE} range — Gulf-of-Mexico All-band CRPS varies by less than 4% across $[10^{-3}, 3\times10^{-2}]$, and North Sea is nearly invariant

Table 6: Mean-wave-period (MWP) and mean-wave-direction (MWD) CRPS on IFS's dataset. **Red** (Blue) values indicate the best (second-best) per column. MWP CRPS in seconds (Gaussian CRPS for ensembles, $|\text{mean}-y|$ MAE for $M=1$); MWD CRPS in degrees (mean angular distance, regardless of M). Lead bands as in Table 4.

		Gulf of Mexico		North Sea		California Coast	
	Model	CRPS-MWP (s)	CRPS-MWD (°)	CRPS-MWP (s)	CRPS-MWD (°)	CRPS-MWP (s)	CRPS-MWD (°)
Short-Term	IFS-Wave	0.4825	10.3872	0.5160	10.7447	0.8127	6.9564
	U-Net	0.2638	8.9300	0.3395	8.5967	0.3518	7.4276
	EarthFormer	0.2891	9.7657	0.3694	9.7133	0.3569	7.8750
	EarthFarseer	0.2586	8.8383	0.3528	8.6533	0.3606	7.4498
	SimVP	0.3209	10.5803	0.4212	12.4824	0.3845	9.2524
	AFNO	0.2754	9.5769	0.3469	8.9027	0.3434	8.0315
	TimeXer	0.3187	10.5029	0.4130	11.2852	0.4038	8.6414
	DiffCast	0.3269	11.2017	0.4425	11.7581	0.3239	7.9811
	FlowCast	0.2456	11.9952	0.3593	13.5704	**0.2214**	6.2442
	BridgeCast-o (ours)	**0.1643**	**7.7254**	0.1973	**7.1847**	0.2457	**6.0555**
	BridgeCast-s (ours)	0.1720	7.9537	**0.1959**	7.1848	0.2407	6.0717
Mid-Term	IFS-Wave	0.4997	12.1454	0.5381	12.6323	0.8108	7.8890
	U-Net	0.2804	10.3052	0.3699	10.1466	0.3709	7.7827
	EarthFormer	0.3019	10.7854	0.3984	11.0200	0.3873	7.9654
	EarthFarseer	0.2877	10.9556	0.3907	10.5443	0.4077	8.3241
	SimVP	0.3388	11.8804	0.4472	13.6954	0.3966	9.4051
	AFNO	0.2913	10.8475	0.3804	10.5866	0.3717	8.5253
	TimeXer	0.3438	11.9428	0.4574	12.8496	0.4320	9.2887
	DiffCast	0.3491	12.7037	0.4195	13.0103	0.3935	8.6919
	FlowCast	0.2536	12.7814	0.3702	14.4779	**0.2417**	**6.6746**
	BridgeCast-o (ours)	0.1880	**9.3481**	0.2355	**8.9935**	0.2769	6.8038
	BridgeCast-s (ours)	**0.1879**	9.5560	**0.2288**	9.0132	0.2640	6.8207
Long-Term	IFS-Wave	0.6153	18.7502	0.7651	23.6216	0.8552	12.1642
	U-Net	0.3785	16.8379	0.5937	21.3694	0.5165	11.8668
	EarthFormer	0.3631	16.2652	0.5705	20.7036	0.5105	11.1356
	EarthFarseer	0.3652	16.7016	0.5728	21.3109	0.5271	11.6492
	SimVP	0.4010	17.2658	0.6121	21.7407	0.5029	12.0682
	AFNO	0.3994	17.2320	0.6262	21.8795	0.5197	12.3993
	TimeXer	0.4186	17.2684	0.6342	21.2561	0.5408	12.1015
	DiffCast	0.4023	18.7721	0.5674	22.8323	0.4572	12.4695
	FlowCast	0.3037	18.2279	0.4972	23.3455	**0.3472**	**10.0059**
	BridgeCast-o (ours)	0.2634	**15.0289**	0.4262	**18.8640**	0.4099	10.2840
	BridgeCast-s (ours)	**0.2485**	15.1735	**0.3967**	19.0302	0.3650	10.3958
Overall	IFS-Wave	0.5398	14.3401	0.6175	16.3982	0.8268	9.3148
	U-Net	0.3141	12.5204	0.4475	14.0368	0.4210	9.2094
	EarthFormer	0.3227	12.6672	0.4579	14.3893	0.4284	9.1025
	EarthFarseer	0.3129	12.8115	0.4525	14.2397	0.4454	9.4235
	SimVP	0.3597	13.7016	0.5046	16.4859	0.4339	10.3425
	AFNO	0.3287	13.0210	0.4656	14.4858	0.4215	9.8577
	TimeXer	0.3683	13.7203	0.5166	15.7329	0.4682	10.2228
	DiffCast	0.3654	14.7463	0.4762	16.4192	0.4069	9.9698
	FlowCast	0.2711	14.6825	0.4152	17.5874	**0.2777**	**7.8263**
	BridgeCast-o (ours)	0.2128	**11.2305**	0.3006	**12.3769**	0.3212	7.9698
	BridgeCast-s (ours)	**0.2082**	11.4180	**0.2894**	12.4223	0.2969	8.0056

across the entire range. The per-band optima migrate rightward as lead time grows (Short → Mid → Long), consistent with longer leads benefitting from larger path noise. IFS curves are monotone-increasing, with the smallest swept value winning in nearly every band, suggesting the IFS-trained model is already well-calibrated at the prior-draw level. The per-band validation optima $\varepsilon^\star$ are shown in Fig. 6, though the flat GFS curves indicate $\varepsilon_{\text{SDE}} \approx 2\times10^{-2}$ would yield comparable performance.

C.5 Implementation Details

Architecture. BridgeCast uses a two-head DiT (Peebles & Xie, 2023): hidden $D=256$, depth 6, 8 attention heads, head dim 32, MLP ratio 4, no dropout. Wave grid is tokenized per-pixel; ATM tokens are 2×2 down-projected before cross-attention. Inputs are $C_{\text{in}}=5$ (4 wave channels + ocean mask), outputs $C_{\text{out}}=4$ (SWH, MWP, MWD-sin, MWD-cos), ATM channels $C_{\text{atm}}=11$, trajectory length $T=24$ (3-h cadence, $\tau_{\max}=168$ h). τ is injected via sinusoidal embedding concatenated with the t embedding through a 2-layer SiLU MLP and AdaLN-Zero modulation. The two heads $(\hat{b}, \hat{\eta})$ share the trunk and split via separate AdaLN-Zero final projections (zero-init modulators, Xavier-init output projections); a 1×1 residual $\mathcal{X}\to\hat{b}$ skip is added on the velocity head only. The ocean mask is applied multiplicatively to both heads' outputs. Total parameters: ≈13.69 M.

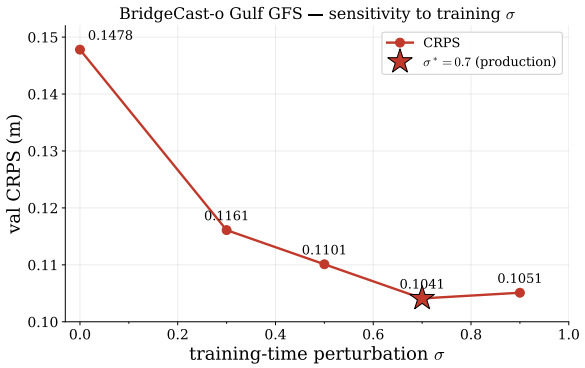


Figure 5: SDE noise sweep

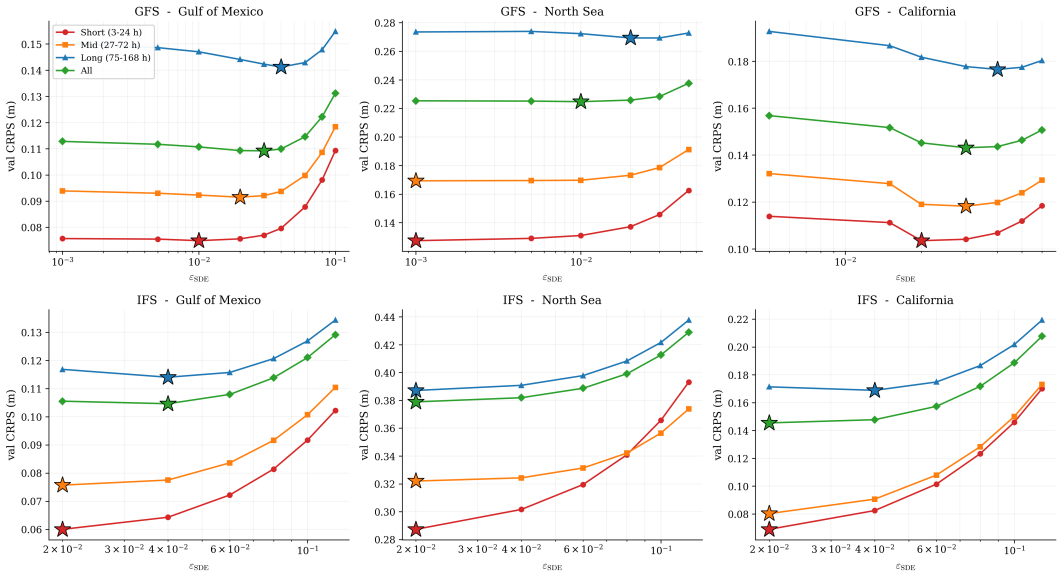


Figure 6: SDE-noise sweep on the validation set. Each panel shows val-set CRPS as a function of ε_{SDE} for one lead band; rows are forcing system (`GFS`, `IFS`), columns are region (Gulf of Mexico, North Sea, California). Stars mark the per-band winners $\varepsilon^\star$ used by BRIDGECAST-s in main results.

Training. AdamW (PyTorch defaults for $(\beta_1, \beta_2, \epsilon)$), weight decay 10^{-4}, peak LR 3×10^{-4} with 5-epoch linear warm-up followed by cosine decay to 10^{-6}. Up to 100 epochs with patience 20 on validation SWH-RMSE. Per-GPU batch size 8, gradient accumulation 4 (effective batch 32), gradient clipping at norm 1.0. EMA on all parameters with decay 0.999. Mixed precision (`torch.cuda.amp`). Single GPU per run (no DDP).

Loss. Channel weights $w = [5, 4, 1, 1]$ on (SWH, MWP, MWD-sin, MWD-cos); ocean mask multiplied inside the squared error and used as the denominator for normalization. Source noise scale $\sigma = 0.7$, score-loss weight $\lambda_\eta = 1$. SI time $t \sim \mathcal{U}(0, 1)$ per minibatch element; forecast lead τ comes from the data sample directly (no resampling). Antithetic noise pairing $(\boldsymbol{\epsilon}, -\boldsymbol{\epsilon})$ is used at every training step.

Inference. Euler–Maruyama with $S = 10$ steps from $\tau = 0$ to $\tau = 1$, evaluating the model at $\tau_i = i/S$ for $i = 0, \ldots, S - 1$. The stochastic increment is omitted in the final step. The score denominator $(1 - \tau_i)\sigma$ is floor-clamped at 10^{-3}. The initial state is $\hat{\mathcal{Y}}_0^{(k)} = \mathcal{X} + \sigma\boldsymbol{\epsilon}^{(k)}$, with $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ drawn fresh per ensemble member. Ensemble size $K = 20$. $\varepsilon_{\text{SDE}} = 0$ recovers

the probability-flow ODE (BRIDGECAST-o); $\varepsilon_{\text{SDE}} > 0$ gives the SDE sampler (BRIDGECAST-s). Both variants generate ensembles through random initial perturbations. For BRIDGECAST-s, per-region noise levels are selected on the validation split (Gulf/NS/Cali = $0.030/0.010/0.030$ for GFS, $0.040/0.020/0.020$ for IFS).

Data. Six (forcing, region) pairs: GFS-Wave from 2021-03 onward ($\approx$8 000 train trajectories per region, 121 val, 121 test) and IFS-WAM from 2024-03 onward (1 230 train, 724 val, 716 test). Wave grids: Gulf 33×47, North Sea 32×48, California 35×41. Each trajectory is 24 frames at 3-h cadence covering forecast leads $\tau \in [3, 168]$ h. Targets are ERA5 reanalysis; inputs are physical-model wave forecasts ($\mathcal{X}$) plus 11 ATM forcing channels ($\mathcal{Z}$). All channels are z-scored using training-set statistics; MWD is sin/cos encoded. Total test frames: 60 264 across the six pairs.

Compute. Training: Single NVIDIA A100 GPU per run on an internal server. Inference benchmark: NVIDIA RTX 6000 Ada, FP32, $M{=}1$, $S{=}10$; BRIDGECAST-s at $\approx$665–980 ms per 24-frame trajectory depending on region and forcing source.