GEOSPATIAL DIFFUSION-BASED EVOLUTION SYNTHESIS (GEODES) FOR AUGMENTATION OF LOCATION-SPARSE PHENOMENA

Anonymous authors
Paper under double-blind review

ABSTRACT

Geospatial synthesis models are designed to generate realistic environmental conditions, yet they often struggle to capture the detailed dynamics of Location-Sparse Phenomena (LSP): complex events like storms, wildfires and oil spills that occur sparsely and irregularly across the globe. Existing approaches frequently rely on structural simplifications, such as restrictive static regional boundaries or computationally expensive, overly smoothed global grids. Instead, we introduce *Geospatial Diffusion-based Evolution Synthesis (GeoDES)*, a phenomenon-centered spatiotemporal diffusion model. GeoDES' correlated noise schedule, non-autoregressive design and temporal inflative training paradigm serve to increase fidelity and decrease compute requirements for LSP. GeoDES efficiently synthesizes high-quality events to augment hazard datasets and stress-test downstream models. On extratropical cyclone generation tasks, GeoDES outperforms baselines by achieving 21% lower Peak Vorticity Error and 26% higher Synoptic Anomaly Correlation Coefficient. Additionally, evaluations on the xBD multi-disaster dataset demonstrate high fidelity, with 41.27 lower Fréchet Inception Distance than the next-best method. GeoDES brings realistic synthesis to the broad class of LSP and offers a step towards robust geospatial data augmentation.

1 INTRODUCTION

The most hazardous geospatial events—including volcanic eruptions, cyclone storms, oil spills and conflict zones—are notoriously difficult to model. Because these events dictate major humanitarian and ecological outcomes, modeling their lifecycles is crucial. To study this broad class of events under a unified framework, we coin the term **"Location-Sparse Phenomena" (LSP)**.

Modeling LSP is challenging because their sparse occurrence severely limits observational data. Generative models typically require massive datasets, causing them to struggle with isolating complex, fine-grained event dynamics from overwhelming quantities of non-event background data. Consequently, models fail to learn underlying physical rules governing these rare occurrences.

We introduce the LSP term to establish a unified concept crossing the *image/video, stationary/mobile and RGB/sensor data boundaries*. To the best of our knowledge, there is no existing terminology that unifies these tasks. Geospatial disciplines have historically tackled LSP in silos, relying on simplifications that limit utility. For instance, the area of weather forecasting remains largely constrained to either global or regional models (Shi et al., 2025b). Both fail for tracking detailed, far-moving weather LSP like hurricanes: global models require immense compute and yield overly smoothed predictions, while regional models are restricted to static boundaries that cannot capture the full lifecycle of these mobile phenomena. Conversely, the area of remote sensing typically frames change modeling (Gupta et al., 2019) as a static RGB image task, limiting cross-pollination of methodologies and hindering the adoption of novel models that could improve structural consistency.

We address these challenges with **Geospatial Diffusion-based Evolution Synthesis (GeoDES)**, a spatiotemporal diffusion model specifically designed for LSP. GeoDES is visualized in Figure 1. We adopt a *phenomenon-centered* approach (e.g., Booth et al., 2018), treating lifecycles as a video anchored to a moving center (Figure 2). By dynamically tracking the LSP, GeoDES captures fine-scale dynamics at high resolutions (e.g., $0.45° \times 0.45°$). Stripping away non-event training data drastically reduces memory usage, enabling computationally accessible training on standard GPUs.

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.

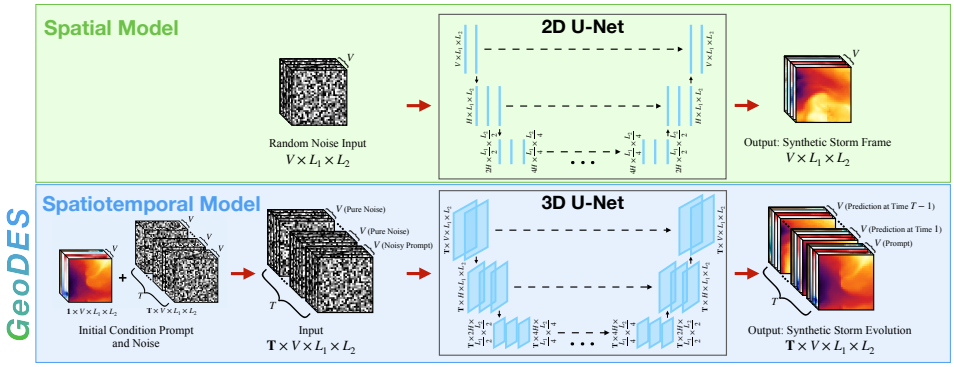


Figure 1: To increase performance and decrease compute requirements (Section 3), we use a three-phase training pipeline, consisting of (1) two-dimensional spatial pretraining, (2) temporal inflation and (3) three-dimensional spatiotemporal fine-tuning. Phase 1 is depicted at top, green background, and phase 3 (the GeoDES model) is at bottom, blue background. We use a dataset of phenomenon-centered bounding boxes with height L_1, width L_2 and timestep dimension T, for V variables (e.g., RGB colors or wind/temperature/humidity; V is analogous to channels C in standard image models). Model hidden dimension is H. GeoDES uses a prompt of initial event conditions to predict the following $T - 1$ timesteps of evolution.

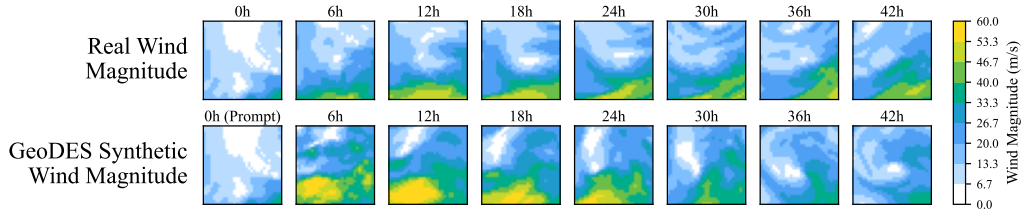


Figure 2: Prompted with 0-hour cyclone storm wind magnitude data, GeoDES generates an alternative storm lifecycle (bottom) that maintains the complex details of the real storm (top) over 42 hours.

To model complex temporal dynamics from limited data, GeoDES treats time as an independent dimension to enable *non-autoregressive synthesis*. We first train a 2D-image spatial model to learn spatial and cross-channel relationships, before using temporal inflation (Singer et al., 2022) to create a 3D-video spatiotemporal model. Synthesizing the entire lifecycle simultaneously mitigates the compounding errors of autoregressive methods, yielding outputs directly compatible with domain-specific analyses (Booth et al., 2018; Clancy et al., 2022; Naud et al., 2025).

To rigorously evaluate GeoDES, we apply our approach to two tasks: **(i)** a meteorological application to extratropical cyclones and **(ii)** multi-disaster damage prediction to assess cross-domain generalization capabilities. For the first task, evaluated on North Atlantic extratropical cyclones from 2016-2024 in Section 3.1, GeoDES outperforms existing methods across a variety of metrics, achieving 21% lower Peak Vorticity Error and 26% higher Synoptic Anomaly Correlation Coefficient than the next strongest baseline. For the second task, evaluated on the xBD dataset (Gupta et al., 2019) for remote-sensing disaster evolution (e.g., tornadoes, hurricanes, and fire damage), GeoDES successfully synthesizes the post-disaster state, with 41.27 lower Fréchet Inception Distance than the next best method. In summary, our core contributions are:

- *Task Formalization:* We formalize the data augmentation task of conditional LSP synthesis, a framework for modeling locally detailed, globally sparse events given an initial prompt condition.
- *Spatiotemporal Diffusion for LSP:* We present GeoDES, a spatiotemporal diffusion model that uses phenomenon-centered framing and temporal inflation to enable detailed synthesis of rare events with limited training observations.
- *State-of-the-Art Generative Realism:* We demonstrate that GeoDES significantly outperforms existing baselines in preserving phenomenon structures and energy distributions, for the specific class of cyclone storms and for a broad dataset of diverse geospatial disasters.

We formally define GeoDES in Section 2. Evaluations in Section 3 demonstrate GeoDES' generative fidelity on cyclone storms and disaster impact prediction. We then discuss prior work and conclusions in Sections 4 and 5.

Algorithm 1: GeoDES Three-Stage Training Procedure

Input: Spatiotemporal dataset D_{3D}; rand. initialized U-Nets $\Theta_{2D}^{(0)}, \Theta_{3D}^{(0)}$; frame count T; learning rate η; noise processes $\mathcal{N}_{\text{iid}}$ (i.i.d. Gaussian) and $\mathcal{N}_{\text{corr}}$ (correlated, Eq. 2)
Output: Trained 3D spatiotemporal model Θ_{3D}^{*}

Procedure TRAIN(Θ, D, $\mathcal{N}$)**:**
  while *not converged* **do**
    $\mathbf{x}_0 \sim \text{Norm}(D)$, $s \sim \mathcal{U}\{1, \ldots, S\}$ ▷ get normalized batch & diffusion steps
    $\mathbf{x}_s \leftarrow \mathcal{N}(\mathbf{x}_0, s)$ ▷ form noisy batch
    $\mathcal{L}(\Theta) \leftarrow \mathbb{E}\big[\, w_{\text{SNR}}(s)\, \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\Theta(\mathbf{x}_s, s)\|_2^2 \,\big]$ ▷ SNR-weighted loss, Eq. 1
    $\Theta \leftarrow \Theta - \eta\, \nabla_\Theta \mathcal{L}(\Theta)$ ▷ update weights
  return Θ

▷ Phase 1: 2D spatial pretraining

$D_{2D} \leftarrow \{\mathbf{x}_\tau^{(n)} : \mathbf{x}^{(n)} \in D_{3D},\, \tau = 1, \ldots, T\}$ ▷ flatten $(N, T, *) \to (NT, *)$

$\Theta_{2D}^{*} \leftarrow \text{TRAIN}(\Theta_{2D}^{(0)},\, D_{2D},\, \mathcal{N}_{\text{iid}})$

▷ Phase 2: temporal inflation

foreach *parameter* $w_{2D} \in \Theta_{2D}^{*}$ *with matching* $w_{3D} \in \Theta_{3D}^{(0)}$ **do**
  if $\dim(w_{2D}) = \dim(w_{3D})$ **then**
    $w_{3D} \leftarrow w_{2D}$ ▷ copy
  else
    $w_{3D} \leftarrow \frac{1}{T}\,\text{Repeat}(w_{2D}, T)$ ▷ inflate $(*) \to (T, *)$

▷ Phase 3: 3D spatiotemporal fine-tuning

$\Theta_{3D}^{*} \leftarrow \text{TRAIN}(\Theta_{3D}^{(0)},\, D_{3D},\, \mathcal{N}_{\text{corr}})$ ▷ $\boldsymbol{\epsilon}(t)$ conditioned on $\mathbf{x}_0(0)$ (Eq. 2)

return Θ_{3D}^{*}

2 THE GEOSPATIAL DIFFUSION-BASED SYNTHESIS (GEODES) METHOD

We design GeoDES, a custom image-to-video diffusion model, to resolve the issues of prior ML models for LSP synthesis. Rather than attempting to learn a complex spatiotemporal manifold from scratch—which leads to physical hallucinations or dynamic range collapse as demonstrated by our Section 3.4 ablations—we decouple spatial structure from temporal evolution. We use a three-phase pipeline, consisting of (1) two-dimensional spatial pretraining, (2) temporal inflation and (3) three-dimensional spatiotemporal fine-tuning. This end-to-end procedure, formalized in Algorithm 1 and visualized in Figure 1, serves as the foundation for the architectural details described below.

GeoDES is based on Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020), which learns to generate synthetic data by reversing a Markovian process that progressively adds Gaussian noise to a true data sample x_0 over N steps. Further background on diffusion models is provided in Appendix B and a glossary of terms is available in Appendix A.

2.1 PHASE 1: THE SPATIAL MODEL ($f_{\Theta_{2D}}$)

The goal of the first phase is to learn a spatial model (the top half of Figure 1) , denoted as $f_{\Theta_{2D}}$ for weight matrix Θ_{2D}. LSP are governed by strict spatial dependencies (e.g., the physical relationship between pressure gradients and wind fields in storms, or the movement patterns of oil distributed across the ocean). By training a two-dimensional image model first, $f_{\Theta_{2D}}$ learns to represent these multivariate spatial relationships without the compounding complexity of temporal dynamics.

Suppose each 2D training datapoint is a 3D tensor $\mathbf{x}_0 \in D_{2D} \subset \mathbb{R}^{V \times L_1 \times L_2}$, where V denotes the number of variables (analogous to channel dimension C in standard vision models, but encompassing traditional RGB channels *and* domain-specific variables, like temperature and sea level pressure) and L_1 and L_2 represent the spatial grid dimensions. To ensure $f_{\Theta_{2D}}$ captures the full variability of the data without suffering dynamic range collapse that causes model outputs to become overly uniform , we first apply a variable-specific hybrid normalizer. We use log-scaling for long-tailed distributions (e.g., wind magnitude) and linear scaling for centered distributions (e.g., temperature). Each variable is scaled according to the 1st and 99th percentiles present in the training distribution.

The network $f_{\Theta_{2D}}$ is an unconditional pixel-space U-Net (Ronneberger et al., 2015). We begin with randomly-initialized weights $\Theta_{2D}^{(0)}$ and optimize using a variance-preserving schedule (Song et al.,

2020) and Signal-to-Noise Ratio (SNR)-weighted objective (Hang et al., 2023), i.e. $\mathcal{L}_{\Theta_{2D}}$ for

$$\mathcal{L}_{\Theta} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},s}\left[\lambda(s) \cdot \|\boldsymbol{\epsilon} - f_{\Theta}(\mathbf{x}_s, s)\|_2^2\right], \tag{1}$$

where $s \in [S]$ is the diffusion step and $\lambda(s)$ is the SNR weight. This weighting forces the model to dedicate representational capacity to intermediate noise scales—which correspond to meaningful, large-scale event structures—rather than wasting capacity on imperceptible, high-frequency noise.

2.2 Phase 2: Temporal Inflation

Once the spatial priors Θ_{2D} are learned, we transition to the spatiotemporal network, denoted as $f_{\Theta_{3D}}$ for weight matrix Θ_{3D}. To introduce time without destroying the multivariate and spatial representations learned by $f_{\Theta_{2D}}$, we perform temporal inflation (Phase 2 of Algorithm 1), adapting the strategy of Singer et al. (2022). Let T denote the number of event evolution timesteps (as opposed to diffusion noise steps S). We initialize $\Theta_{3D}^{(0)}$ by copying all spatially-equivalent blocks from Θ_{2D} to $\Theta_{3D}^{(0)}$. For any spatial block w_{2D} that requires an expanded temporal dimension in the 3D model, we initialize the corresponding weights w_{3D} at each time t as $w_{3D}(t) = w_{2D}/T$.

2.3 Phase 3: The Spatiotemporal Network ($f_{\Theta_{3D}}$)

Beginning with the inflated architecture weights $\Theta_{3D}^{(0)}$, Phase 3 fine-tunes $f_{\Theta_{3D}}$ on temporal dynamics. Suppose a 3D datapoint is a 4D tensor $\mathbf{x}_0 \in \mathbb{R}^{T\times V\times L_1\times L_2}$.

Training Phase: We use a standard forward diffusion process for training. Because geospatial state is highly autocorrelated across time, we find using standard independent Gaussian noise for each frame disrupts this inherent consistency, allowing for physically impossible state jumps. While diffusion methods usually rely solely on network convolutions learning temporal smoothness, we explicitly inject this prior into the forward process by correlating the noise at frame t with the preceding frame $t-1$ via a tunable parameter ρ. Given uncorrelated noise $\boldsymbol{\epsilon}_u \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we set $\boldsymbol{\epsilon}(0) = \boldsymbol{\epsilon}_u(0)$ and subsequent frames $t \in [1, T)$ as

$$\boldsymbol{\epsilon}(t) = \rho \cdot \boldsymbol{\epsilon}(t-1) + (\sqrt{1-\rho^2}) \cdot \boldsymbol{\epsilon}_u(t). \tag{2}$$

In our experiments, we set $\rho = 0.95$. Real geospatial variables exhibit complex, variable-specific covariance structures; thus, developing dynamic, physics-informed noise schedules remains a valuable direction for future work. The reverse process then trains $f_{\Theta_{3D}}$ to denoise $\boldsymbol{\epsilon}$ using the same SNR-weighted objective of Equation 1, i.e. $\mathcal{L}_{\Theta_{3D}}$.

Sampling Phase: At sample time, the goal is to predict future evolution conditioned strictly on an initial ($t = 0$) prompt frame, denoted $\mathbf{x}_0(0)$. To prevent the model from seeing the remainder of the non-synthetic data (the frames $t \in [1, T)$), the forward process forms the noisy sample $\mathbf{x}_n$ by setting all non-initial frames to correlated noise ($\mathbf{x}_n(1, T-1) = \boldsymbol{\epsilon}(1, T-1)$), while the initial frame receives standard pure diffusion noise, $\mathbf{x}_n(0) = \sqrt{\bar{\alpha}_n} \cdot \mathbf{x}_0(0) + \sqrt{1-\bar{\alpha}_n} \cdot \boldsymbol{\epsilon}(0)$.

During the reverse denoising process, we anchor the temporal generation by re-injecting the unnoised prompt $\mathbf{x}_0(0)$ at every step of the sampling sequence (Lugmayr et al., 2022; Ho et al., 2022). This process ensures that the generated event remains tied to the initial prompt.

3 Evaluations

We design our experiments to validate the following four claims:

1. **GeoDES outperforms prior works in storm realism.** Because the utility of synthetic weather data depends entirely on its physical plausibility, we verify that our architecture captures atmospheric dynamics more accurately than existing state-of-the-art baselines.
2. **GeoDES is effective for change modeling tasks.** *Change modeling* involves predicting an "after" scene given a "before" scene; we specifically demonstrate high fidelity synthesis for the task of disaster aftermath prediction.
3. **GeoDES maintains reasonable compute requirements during training and inference.** A key limitation of global models is their immense computational cost. We demonstrate that our localized, hazard-centered approach is accessible on standard hardware.
4. **Ablations: Individual components are critical to GeoDES' success.** We validate the importance of SNR weighting, stochastic sampling, logarithmic normalization and temporal inflation introduced in Section 2.

We address these claims sequentially in Sections 3.1 through 3.4. Definitions, while provided throughout, are also listed in Appendix A.

Table 1: Comparison of generative storm realism for North Atlantic extratropical cyclones. GeoDES successfully maintains physical constraints and extreme energy distributions, avoiding the spatial smoothing of global Weather Foundation Models. Our method, GeoDES, is shown in blue background, other ML models are shown in white background and the naive climatology baseline of simply averaging all training data together is shown in gray background. In the far left column, (↓) and (↑) respectively indicate lower or higher scores are better, while (·1) indicates that scores closer to 1 are better. Best results are bolded.

	GeoDES	SVD	CoDiCast	CEF	ClimaX FT	ClimaX	Aurora	*Climatology*
Peak Vorticity Error (↓)	$\mathbf{5.04 \pm 0.47}$	6.41 ± 0.22	10.52 ± 1.72	11.04 ± 0.77	10.68 ± 1.01	12.32	13.80	*7.86*
HF Spectral Ratio (·1)	$\mathbf{0.95 \pm 0.17}$	0.65 ± 0.01	3.65 ± 1.09	0.59 ± 0.02	0.41 ± 0.02	0.02	0.00	*0.29*
Frequency Bias Index (·1)	$\mathbf{1.02 \pm 0.17}$	0.91 ± 0.00	2.26 ± 0.09	0.60 ± 0.05	1.86 ± 0.03	1.04	0.00	*0.75*
Fractions Skill Score (↑)	$\mathbf{0.76 \pm 0.04}$	$\mathbf{0.76 \pm 0.00}$	0.65 ± 0.01	0.61 ± 0.01	0.72 ± 0.01	0.00	0.00	*0.71*
Synoptic ACC (↑)	$\mathbf{0.39 \pm 0.03}$	0.31 ± 0.00	0.06 ± 0.01	0.19 ± 0.02	0.28 ± 0.00	0.09	0.10	*0.64*
Synoptic ACC T-Final (↑)	$\mathbf{0.44 \pm 0.03}$	0.24 ± 0.00	0.06 ± 0.01	0.18 ± 0.02	0.24 ± 0.00	0.05	0.11	*0.64*

3.1 Cyclone Storm Synthesis

We begin by validating our first claim.

Claim 1: GeoDES outperforms prior modeling approaches in storm realism.

Setup: We extract North Atlantic extratropical cyclones from ERA5 reanalysis (Hersbach et al., 2020) across five atmospheric variables (Sea level pressure, U/V wind components, temperature, humidity). To prevent temporal leakage, we split by time (1940–2015 train, N_{train}=8589; 2016–2024 test, N_{test}=893). While global baselines forecast over the full Earth grid before storm extraction, GeoDES and localized baselines operate directly on 1600×1600 km bounding boxes centered on minimum sea level pressure (App. C.1). Evaluations average 3 runs (hyperparameters in App. C.4).

Baselines: We benchmark GeoDES against: (1) *Stable Video Diffusion (SVD)* (Blattmann et al., 2023) as a localized, storm-centered baseline; (2) the global diffusion models of *Continuous Ensemble Forecasting (CEF)* (Andrae et al., 2025) and *CoDiCast* (Shi et al., 2025a); (3) the global foundation models of *ClimaX* (zero-shot and fine-tuned) (Nguyen et al., 2023) and zero-shot *Aurora* (Bodnar et al., 2024) and (4) a naive *climatology* baseline (averaging all training datapoints) (Rasp et al., 2020). A structural comparison of all baselines is provided in Appendix Table 5 and more details in App. C.3. To ensure fair resource comparison, we benchmark against models of comparable computational scale (e.g., SVD has 1.5B parameters, Aurora 1.25B, GeoDES 1.8B). Baseline details are in App. C.3; inference efficiency is analyzed in Sec. 3.3.

Metrics: We evaluate structural and physical realism via: (1) *Peak Relative Vorticity Error* (↓, measures core cyclonic circulation) (Holton & Hakim, 2012); (2) *High-Frequency (HF) Spectral Ratio* (·1, ideal is 1.0; assesses fine-scale kinetic energy) (Skamarock, 2004); (3) *Frequency Bias Index (FBI)* (·1, ideal is 1.0; evaluates extreme wind pixel recovery) (Wilks, 2011); (4) *Fractions Skill Score (FSS)* (↑, spatial overlap of extreme winds) (Roberts & Lean, 2008); and (5) *Synoptic ACC* (↑, spatially filtered pattern correlation) (Wilks, 2011). Full formulations and supplemental metrics (e.g., RMSE) are detailed in Appendix D.3.

Results: Table 1 includes results for our primary metrics across all evaluated methods on the North Atlantic basin test set ($n = 893$) and Figure 3 visualizes an example datapoint. ***GeoDES outperforms probabilistic methods across evaluation metrics.*** In particular, GeoDES is the only model capable of recovering the kinetic energy distribution of ground-truth storms, achieving a near-perfect Frequency Bias Index of 1.02 and a High-Frequency (HF) Spectral Ratio of 0.95.

Baseline weather models generally exhibit one of two contrasting failure modes. While CoDiCast collapses into overly noisy, high-frequency predictions that fail to capture proper cyclone structures, SVD and the global models (CEF, ClimaX FT, ClimaX, and Aurora) produce overly blurred, low-energy forecasts. Meanwhile, the statistical climatology baseline manages to achieve deceptively high spatial correlation scores by predicting a heavily smoothed average of training storms; however, its poor kinetic energy and frequency metrics reveal that it completely lacks the physical realism required for actual storm synthesis.

Successes and failure modes of each method are also demonstrated by their outputs in Figure 3. While GeoDES synthesizes realistic, dynamic cyclones, fine-tuned global models (CoDiCast, CEF, ClimaX FT) exhibit unphysical distortions such as the North-South stripes in middle frames of CEF and the lack of calm/white area in CoDiCast. Meanwhile, zero-shot models (ClimaX, Aurora) generate overly blurry forecasts. Furthermore, encoder-based models (CoDiCast, Aurora) fail to even reproduce initial storm conditions.

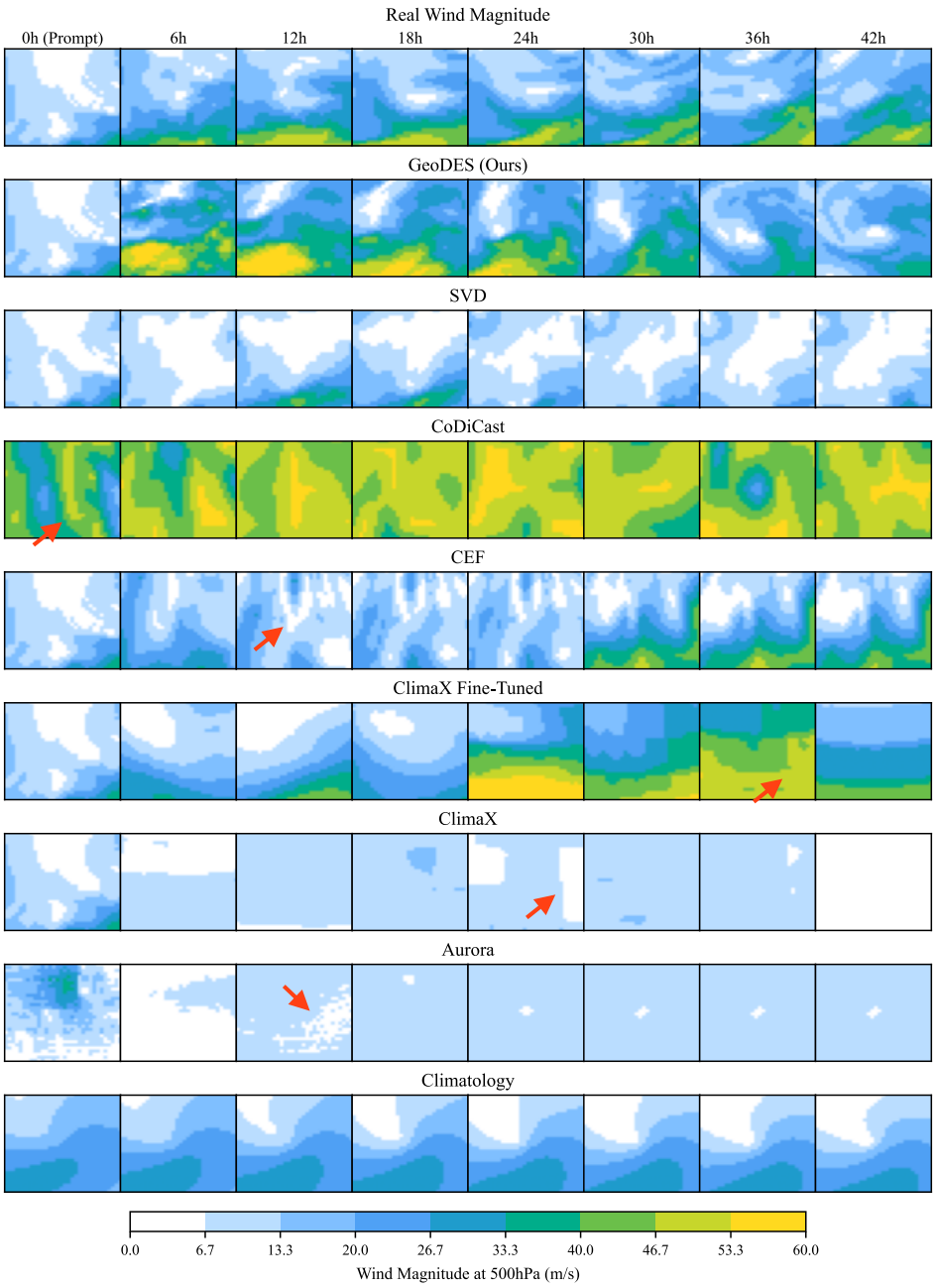


Figure 3: Wind magnitude of a real, non-synthetic extratropical cyclone in the North Atlantic (57.60° N, 26.87° W at 0h) on October 3-5, 2018, and output from each synthesis method. GeoDES creates a realistic, novel storm, while other methods generate blurry (ClimaX, Aurora) or noisy (CoDiCast) outputs. For instance, Aurora generates nearly perfectly uniform predictions with only a small dot in the center of the storm, while the North-South bands in the CEF prediction do not look realistic. Red arrows indicate just some of the many unphysical distortions.

Table 2: Post-disaster image synthesis (Section 3.2). Averaged across 2 runs; Best results bolded.

	GeoDES	Pix2Pix	Palette
Frechét Inception Distance (↓)	$\mathbf{195.63 \pm 68.71}$	265.77 ± 0.68	236.90 ± 22.93
Structural Similarity (↑)	$\mathbf{0.41 \pm 0.06}$	0.29 ± 0.00	0.31 ± 0.04
Peak Signal-to-Noise Ratio (↑)	15.08 ± 0.93	$\mathbf{16.27 \pm 0.21}$	14.21 ± 0.13

3.2 Disaster Impact Modeling

To prove the versatility of GeoDES, we use the xBD disaster impacts dataset as an example of a change modeling task, to validate our second claim:
Claim 2: GeoDES is effective for change modeling tasks.

Setup: We evaluate GeoDES on xBD (Gupta et al., 2019), a dataset of satellite imagery covering multiple categories of environmental disasters such as hurricanes, floods and wildfires. Because xBD is formatted for pre- and post-disaster damage assessment, we provide the pre-disaster image as the $t = 0$ prompt and task models with synthesizing the post-disaster $t = 1$. Details in Appendix E.

We compare GeoDES to two popular disaster modeling approaches, the Pix2Pix GAN (Isola et al., 2017) and Palette diffusion model (Saharia et al., 2022), using three generative metrics: *Fréchet Inception Distance* (↓) (Heusel et al., 2017) to measure alignment with the true post-disaster distribution, *Structural Similarity* (↑) (Wang et al., 2004) to observe the preservation of invariant pre-disaster structures and *Peak Signal-to-Noise Ratio* (↑) (Wang et al., 2021) to gauge pixel-level reconstruction.

Results: Table 2 includes results and Figure 4 visualizes example outputs. GeoDES leads on both Fréchet Inception Distance (195.63) and Structural Similarity (0.41), while remaining competitive with Pix2Pix and Palette on Peak Signal-to-Noise Ratio (15.08). Pix2Pix achieves the highest Peak Signal-to-Noise Ratio (16.27), which is expected given its explicit L_1 pixel-wise optimization objective, though this comes at the cost of perceptual degradation (Fréchet Inception Distance of 265.77). Overall, ***GeoDES provides a highly effective balance of geospatial fidelity and generative realism, confirming its robust generalization to standard RGB domains and multi-LSP datasets.***

3.3 Computational Requirements

Due to the significant computational requirements associated with atmospheric modeling, we now return our focus to meteorology as a proving ground for GeoDES' computational footprint:
Claim 3: GeoDES maintains reasonable compute requirements during training and inference.

Setup: We evaluate the computational footprint of GeoDES for tracking far-moving storms across time, versus the meteorological baselines introduced in Section 3.1. Table 3 reports the model parameter count, sampling paradigm (per-storm versus continuously over the 2016–2024 test span) and the resulting number of inference steps, as well as the FLOPs, latency, VRAM and energy required to produce one 42-hour sample. Due to the prohibitive RAM of tracing Aurora's execution graphs, exact FLOP counts for Aurora could not be profiled natively. We instead estimate FLOPs for Aurora according to transformer scaling laws (Kaplan et al., 2020). Refer to Appendix F for additional details.

To assess train-time compute, we also measure the 2D spatial model used to train GeoDES. To sample the same number of frames as the 3D model (i.e., $T = 8$), the 2D model uses 9.4 TFLOPs as opposed to 123.7 TFLOPs for the 3D model shown in Table 3.

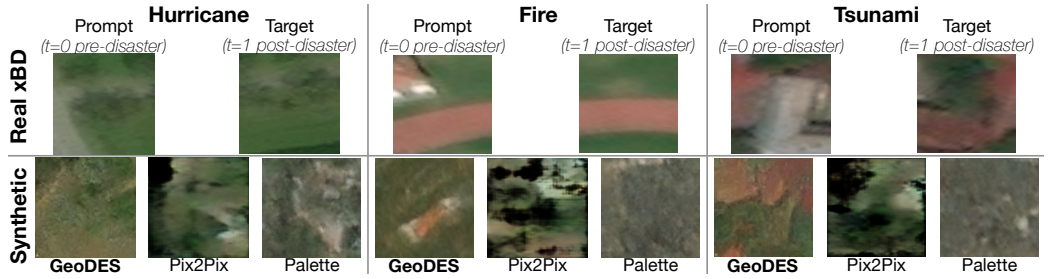


Figure 4: Real and synthetic data for xBD disaster experiment in Sec. 3.2. Left to right: US Hurricane Florence (2018), California Woolsey Fire (2018), Palu Tsunami (2018). GeoDES exhibits higher rates of invariant structure preservation, as demonstrated by higher Structural Similarity (Table 2) than baselines, maintaining footprints of roads and buildings in post-disaster images.

Table 3: Test-time computational requirements for each evaluated model, including model size, sampling paradigm, inference steps, FLOPs, latency, VRAM and energy to produce one 42-hour sample. While GeoDES contains more parameters than other models, all other metrics such as latency and energy usage are competitive with other models. *: FLOPs could not be empirically measured for Aurora due to system RAM limitations; refer Appendix F for estimated derivation.

	GeoDES	SVD	CoDiCast	CEF	ClimaX	Aurora
Params (M)	1802	1525	63	56	111	1256
Paradigm	Storm	Storm	Continuous	Continuous	Continuous	Continuous
Inference Steps	893	893	13145	13145	13145	13145
FLOPs (T)	123.7	57.6	93.5	34.6	2.2	$\sim$ 4557*
Latency (s)	17.3	7.7	1029.2	4.1	0.9	47.6
VRAM (GB)	7.1	6.0	1.1	0.3	1.6	13.4
Energy (Wh)	1.44	0.64	85.77	0.34	0.07	3.97

Results: *At test time:* GeoDES demonstrates practical inference requirements. While GeoDES has the most parameters (1.8B), the storm-centered approach restricts its computation strictly to the spatial bounding box of the storm. This design choice bypasses the compute explosion characteristic of high-resolution global Weather Foundation Models (WFMs). For instance, while Aurora contains fewer parameters (1.25B) than GeoDES, processing its high-resolution global grid requires over 36 times as many FLOPs and nearly double the VRAM (13.4 GB).

GeoDES also avoids computational traps of autoregressive models. Despite CoDiCast's smaller parameter count (63M), its autoregressive sequence generation results in severe per-sample latency (1029.2s) and energy consumption (85.77 Wh); GeoDES uses a fraction of the time (17.3s) and energy (1.44 Wh). Continuous-running models (CoDiCast, CEF, ClimaX, Aurora) compound this cost over the 2016–2024 period, performing over $13,000$ inference steps, compared to GeoDES, which is only prompted 893 times to sample the entire testset. Consequently, heavier architectures like CoDiCast cumulatively exceed 1 MWh for the entire testset, heavily outpacing the storm-centric paradigm of GeoDES.

Ultimately, GeoDES maintains highly accessible compute requirements that enable its use on a single commercial GPU. By constraining our modeling directly to the phenomenon's bounding box, ***GeoDES achieves the state-of-the-art physical realism demonstrated in Table 1 using compute resources that are rigorously comparable to, or significantly lower than, the evaluated baselines.***

At train time: GeoDES' temporal inflation scheme results in significant time and computational resource savings compared to training the 3D spatiotemporal model from scratch. Because half of the GeoDES training epochs are allocated to the 2D stage, we reduce the computational bottleneck associated with spatiotemporal diffusion. In particular, because the 2D spatial generation phase requires only 9.4 TFLOPs per sample, compared to the 123.7 TFLOPs at 3D temporal phase, GeoDES achieves its final performance with only 53.8% of the total theoretical compute that would have been required to train the 3D architecture from scratch. Derivation details in Appendix F.

3.4 Ablations

Claim 4: Ablation analysis demonstrates the effectiveness of GeoDES as a whole.

Setup: We test four ablation variants of GeoDES:

- *No SNR:* Signal-to-noise weighting is disabled.
- *Non-Stoch.:* The model is sampled deterministically.
- *Linear Norm.:* All atmospheric channels are normalized linearly.
- *3D W/Out 2D:* A randomly initialized spatiotemporal GeoDES model is trained entirely from scratch for the total number of training epochs.

Primary evaluation metrics are in Table 4; supplementary metrics are in Appendix D.3 and Table 9.

Results: The ablation study confirms that all components of the GeoDES pipeline are necessary to achieve high-quality storm synthesis.

Temporal Inflation (3D W/Out 2D): Training the 3D spatiotemporal model from scratch yields significant performance degradation, particularly in spatial overlap and navigational accuracy. Without the spatial representation learned during 2D pretraining, Fractions Skill Score drops sharply from

Table 4: Ablation study of the GeoDES architecture evaluated on the North Atlantic storm test set. Results demonstrate that the entire GeoDES pipeline contributes to physical realism, with the 2D-to-3D temporal inflation phase proving especially critical for model stability. Best results bolded.

	GeoDES	No SNR	Non-Stoch.	Linear Norm.	3D W/Out 2D
Peak Vorticity Error (↓)	5.04 ± 0.47	5.21 ± 0.36	5.43 ± 0.28	$\mathbf{4.96 \pm 0.05}$	5.07 ± 0.04
HF Spectral Ratio (·1)	$\mathbf{0.95 \pm 0.17}$	0.71 ± 0.19	0.76 ± 0.08	0.79 ± 0.11	1.09 ± 0.35
Frequency Bias Index (·1)	$\mathbf{1.02 \pm 0.17}$	0.93 ± 0.08	0.95 ± 0.04	1.06 ± 0.12	1.10 ± 0.19
Fractions Skill Score (↑)	$\mathbf{0.76 \pm 0.04}$	0.73 ± 0.05	0.72 ± 0.02	$\mathbf{0.76 \pm 0.00}$	0.62 ± 0.08
Synoptic ACC (↑)	$\mathbf{0.39 \pm 0.03}$	0.36 ± 0.04	0.37 ± 0.00	0.37 ± 0.01	0.33 ± 0.01
Synoptic ACC T-Final (↑)	$\mathbf{0.44 \pm 0.03}$	0.41 ± 0.04	0.39 ± 0.01	0.42 ± 0.00	0.38 ± 0.02

0.76 to 0.62 and Synoptic ACC falls to 0.33. This confirms that two-stage training is not merely a computational optimization (as established in Section 3.3), but a critical requirement for stability.

Hybrid Normalization (Linear Norm.): While relying purely on linear normalization yields a competitive Peak Vorticity Error (4.96), the model fails to accurately preserve the energy distribution of extreme wind, demonstrated by its degraded HF Spectral Ratio (0.79 down from 0.95). This demonstrates that applying log-scaling to long-tailed variables helps to prevent dynamic range collapse.

Noise and Sampling Mechanisms (No SNR & Non-Stoch.): Removing signal-to-noise ratio weighting or using deterministic sampling impairs the model's ability to represent fine-grained, chaotic dynamics of weather systems. Both ablations result in higher Peak Vorticity Errors (> 5.20) and drops in high-frequency energy recovery (HF Spectral Ratios falling from 0.95 to 0.71 and 0.76, respectively). This proves that properly weighted stochasticity is required to maintain realistic atmospheric turbulence.

4 Related Work

GeoDES is most closely related to the areas of synoptic storms research, weather modeling and diffusion architectures. We now provide more background on each of these areas.

Geospatial ML traditionally operates on fixed spatial grids, a paradigm that has rapidly advanced weather modeling through regional networks (Andrychowicz et al., 2023) and global Weather Foundation Models (WFMs) (Nguyen et al., 2023; Lam et al., 2023). However, this fixed-coordinate approach struggles with "location-sparse phenomena"—dynamic events like extratropical cyclones, tornadoes, or volcanic plumes that feature complex, fine-grained details but occur sparsely across the globe. Static regional models cannot track these moving phenomena across vast oceans, while deterministic global models waste immense computational capacity on calm areas and tend to blur fine-grained features. While probabilistic global frameworks (Price et al., 2023) mitigate this blurring, they impose prohibitive costs when processing the entire Earth's surface. Consequently, despite meteorologists frequently using localized, moving bounding boxes to analyze severe weather (Booth et al., 2018; Holton & Hakim, 2012), the efficient ML modeling of the complete spatiotemporal evolution of location-sparse systems remains relatively under-explored.

Diffusion modeling (Ho et al., 2020) is a popular generative approach for image and video synthesis (Ho et al., 2022; Blattmann et al., 2023). In geospatial applications, diffusion has primarily been applied to fixed-grid global forecasting (e.g., CoDiCast (Shi et al., 2025a) and CEF (Andrae et al., 2025)), static regional domains (Mardani et al., 2024; Gao et al., 2023), or static image-to-image translation tasks (e.g., Palette (Saharia et al., 2022)). Applying diffusion to location-sparse phenomena requires untethering the architecture from fixed coordinate systems. While diffusion shows promise for scalable ensemble generation (Li et al., 2023), adapting these models to dynamically track and evolve isolated, high-frequency events over time—without simulating the entire global geospatial grid—remains a critical missing step.

5 Conclusion

We present Geospatial Diffusion-based Evolution Synthesis (GeoDES), a model for the synthesis of location-sparse phenomena. We demonstrate that GeoDES' phenomenon-centered approach and 2D-to-3D temporal inflation strategy enable realistic synthesis, useful to data augmentation and validation of downstream models.

AI USE STATEMENT

In this work, we used generative AI tools to receive feedback on the paper, copyedit the paper for readability, to create starter code for generating plots and to find related works. We did not use generative AI tools for the required disclosure tasks of generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis or interpreting results. We have reviewed all AI-assisted work. For example, we proofread all LLM outputs for accuracy, verified and tested all LLM-generated code, performed a manual literature search and reviewed all cited papers ourselves. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several efforts towards reproducibility. Beyond describing our method in detail in Section 2, we have explained our dataset creation process in-depth for the extratropical cyclone task in Appendix C.1 and the disaster change prediction task in Appendix E and the hyperparameter processes for these experiments are described in Appendices C.4 and E, respectively. We have taken steps to share compute requirements to run GeoDES and baselines through the experiment in Section 3.3. Anonymized source code is included with our submission and open source code will be released upon the publication of the paper.

REFERENCES

Martin Andrae, Tomas Landelius, Joel Oskarsson, and Fredrik Lindsten. Continuous ensemble weather forecasting with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 26392–26416, 2025.

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*, 2023.

Roman Attinger, Elisa Spreitzer, Maxi Boettcher, Heini Wernli, and Hanna Joos. Systematic assessment of the diabatic processes that modify low-level potential vorticity in extratropical cyclones. *Weather and Climate Dynamics*, 2:1073–1091, November 2021. doi: 10.5194/wcd-2-1073-2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan Weyn, Haiyu Dong, Anna Vaughan, et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 1(8), 2024.

James F. Booth, Catherine M. Naud, and Jeyavinoth Jeyaratnam. Extratropical Cyclone Precipitation Life Cycles: A Satellite-Based Analysis. *Geophysical Research Letters*, 45(16):8647–8654, 2018. ISSN 1944-8007. doi: 10.1029/2018GL078977. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2018GL078977. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL078977.

J. L. Catto. Extratropical cyclone classification and its use in climate studies. *Reviews of Geophysics*, 54(2):486–520, 2016. doi: https://doi.org/10.1002/2016RG000519. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016RG000519.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James

Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Robin Clancy, Cecilia M Bitz, Edward Blanchard-Wrigglesworth, Marie C McGraw, and Steven M Cavallo. A cyclone-centered perspective on the drivers of asymmetric patterns in the atmosphere and sea ice during arctic cyclones. *Journal of Climate*, 35(1):73–89, 2022.

Elizabeth E Ebert. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, 15(1):51–64, 2008.

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.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery, 2019. URL https://arxiv.org/abs/1911.09296.

Johannes Hagemann, Samuel Weinbach, Konstantin Dobler, Maximilian Schall, and Gerard de Melo. Efficient parallelization layouts for large-scale distributed model training, 2024. URL https://arxiv.org/abs/2311.05610.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7441–7451, 2023.

M. K. Hawcroft, L. C. Shaffrey, K. I. Hodges, and H. F. Dacre. How much Northern Hemisphere precipitation is associated with extratropical cyclones? *Geophysical Research Letters*, 39(24), 2012. ISSN 1944-8007. doi: 10.1029/2012GL053866. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2012GL053866. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2012GL053866.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

James R Holton and Gregory J Hakim. *An introduction to dynamic meteorology*, volume 88. Academic press, 2012.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

Ian T Jolliffe and David B Stephenson. *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons, 2011.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

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.

Lizao Li, Rob Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Seeds: Emulation of weather forecast ensembles with diffusion models. *arXiv preprint arXiv:2306.14066*, 2023.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471, 2022.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual diffusion modeling for km-scale atmospheric downscaling. 2024.

Jonathan E Martin. *Mid-latitude atmospheric dynamics: a first course*. John Wiley & Sons, 2013.

Catherine M. Naud, Jonathan E. Martin, Poushali Ghosh, Gregory S. Elsaesser, James F. Booth, and Derek J. Posselt. Lifecycle-type matters for extratropical cyclone precipitation production. *Geophysical Research Letters*, 52(8):e2025GL115153, 2025. doi: https://doi.org/10.1029/2025GL115153. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025GL115153. e2025GL115153 2025GL115153.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Simon Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Nigel M Roberts and HW Lean. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97, 2008.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022. URL https://arxiv.org/abs/2111.05826.

Jimeng Shi, Bowen Jin, Jiawei Han, Sundararaman Gopalakrishnan, and Giri Narasimhan. Codicast: Conditional diffusion model for global weather forecasting with uncertainty quantification. International Joint Conferences on Artificial Intelligence Organization, 2025a.

Jimeng Shi, Azam Shirali, Bowen Jin, Sizhe Zhou, Wei Hu, Rahuul Rangaraj, Shaowen Wang, Jiawei Han, Zhaonan Wang, Upmanu Lall, Yanzhao Wu, Leonardo Bobadilla, and Giri Narasimhan. Deep learning and foundation models for weather prediction: A survey, 2025b. URL https://arxiv.org/abs/2501.06907.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

William C Skamarock. Evaluating mesoscale nwp models using kinetic energy spectra. *Monthly weather review*, 132(12):3019–3032, 2004.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Joshua C Steiner. A regional comparison of bomb cyclones in the central plains and western atlantic. Master's thesis, The Ohio State University, 2021.

Yuanhao Wang et al. Spatial-temporal super-resolution of satellite imagery via conditional pixel synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27333–27345, 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Daniel S Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

Ian R. Young and Agustinus Ribal. Multiplatform evaluation of global trends in wind speed and wave height. *Science*, 364(6440):548–552, 2019. doi: 10.1126/science.aav9527. URL https://www.science.org/doi/abs/10.1126/science.aav9527.

A Glossary of Terms and Abbreviations

We include definitions for common abbreviations in this section, as well as important technical terms.

- **ACC (Anomaly Correlation Coefficient):** A standard meteorological metric that measures the spatial pattern correlation of predicted fields to evaluate spatial and navigational accuracy (Wilks, 2011).
- **Advection:** The horizontal transport of atmospheric properties, such as heat (thermal advection) or moisture, driven by the wind field.
- **Autoregressive Synthesis:** A generation approach where the model predicts each subsequent timestep based on the previous one(s).
- **CEF (Continuous Ensemble Forecasting):** A diffusion-based global forecasting ensemble that operates via a continuous deterministic ODE solver (Andrae et al., 2025).
- **Change Modeling:** A predictive task that involves generating a subsequent "after" scene given an initial "before" scene; utilized in this work to evaluate disaster impact and aftermath synthesis.
- **Climatology (Baseline):** In weather forecasting, a naive statistical baseline that predicts the long-term historical average of atmospheric variables, rather than running a dynamic, physics-based simulation.
- **DDPM (Denoising Diffusion Probabilistic Model):** A generative machine learning framework that synthesizes data by learning to reverse a Markovian forward diffusion process that progressively adds Gaussian noise to a sample, see Appendix B (Ho et al., 2020).
- **ERA5:** The fifth generation ECMWF atmospheric reanalysis of the global climate, used as the primary data source for training and evaluating the models in this study (Hersbach et al., 2020).

- **Extratropical Cyclone:** An evolving, mid-latitude weather system driven by atmospheric temperature gradients, associated with strong wind, heavy precipitation and large-scale energy transfer (Martin, 2013).
- **FBI (Frequency Bias Index):** A metric evaluating extreme intensity, calculated as the ratio of predicted extreme wind pixels (> 20 m/s) to actual extreme wind pixels (Wilks, 2011).
- **FID (Fréchet Inception Distance):** A computer vision metric used to measure alignment with the true post-disaster distribution by evaluating generative image quality (Heusel et al., 2017).
- **Fine-tuning:** A machine learning procedure where a pre-trained model is further trained on a task-specific dataset (such as climate fields) to adapt its generalized capabilities to a specialized domain.
- **Front:** A sharp boundary or transition zone between two distinct air masses (e.g., cold and warm air). Fronts drive mid-latitude weather and are characterized by tight gradients in temperature, pressure, and humidity.
- **FSS (Fractions Skill Score):** A neighborhood-based spatial verification metric that measures the spatial overlap accuracy specifically for extreme (> 20 m/s) wind distributions (Roberts & Lean, 2008).
- **GeoDES (Geospatial Diffusion-based Evolution Synthesis):** The primary architecture proposed in this work; a storm-centered, spatiotemporal diffusion model designed for synthesizing large-scale storm structure evolution.
- **Geopotential:** The gravitational potential energy of a unit mass at a certain altitude. In weather modeling, it is commonly used in place of geometric height to track the boundaries of pressure levels.
- **Hallucination:** An undesirable effect of generative model outputs appearing structurally coherent, but physically incorrect or purely fabricated, often failing to respect underlying scientific laws.
- **hPa (Hectopascal):** A metric unit of pressure widely used in meteorology to quantify atmospheric pressure levels (e.g., 500hPa, 925hPa). One hectopascal is equivalent to one millibar.
- **Hypsometric Equation:** A meteorological equation that relates the atmospheric pressure ratio between two levels to the mean virtual temperature of that layer; used in this work to derive sea-level pressure from geopotential and temperature data.
- **Latent Space:** A compressed, low-dimensional mathematical representation of data where a model encodes key features, allowing for efficient processing and synthesis of complex spatial structures.
- **Location-Sparse Phenomena (LSP):** A term introduced in this work to describe complex, globally distributed events (e.g., storms, wildfires, oil spills) that occur irregularly and cover only a fraction of the Earth's surface at any given time.
- **MCMS (Modeling, Analysis and Prediction Climatology of Mid-latitude Storminess):** Codebase and methodology used to identify and track the coordinates of extratropical cyclones within the global ERA5 dataset, see Appendix C.1.
- **MSE (Mean Squared Error):** Standard loss function used during training, measuring the pixel-wise difference between ground-truth and predicted atmospheric states (Ho et al., 2022).
- **Non-Autoregressive Synthesis:** A generation approach where the model synthesizes an entire sequence of data (like a time-series of storm frames) simultaneously rather than step-by-step. This contrasts with autoregressive synthesis.
- **Palette:** An image-to-image diffusion model evaluated as a baseline architecture for disaster impact synthesis (Saharia et al., 2022).
- **Pix2Pix:** A conditional Generative Adversarial Network (GAN) utilized as an image-to-image baseline for disaster damage prediction (Isola et al., 2017).
- **Pressure Level:** A vertical coordinate system used in atmospheric sciences where elevation is defined by atmospheric pressure (e.g., 500hPa, 925hPa) rather than geometric altitude.
- **PSD (Power Spectral Density):** A measure of a signal's energy distribution across different spatial frequencies (wavenumbers). In this work, it visualizes a model's ability to resolve fine-grained, high-energy atmospheric turbulence.
- **PSNR (Peak Signal-to-Noise Ratio):** A computer vision metric used to gauge the pixel-level reconstruction quality of synthesized images against ground-truth targets (Wang et al., 2021).

- **Reanalysis:** A retrospective meteorological dataset created by fusing historical observations with numerical weather prediction models via data assimilation, providing a globally complete historical record (e.g., ERA5).
- **Relative Vorticity:** A macroscopic measure of the local rotation or spin of the fluid flow (wind) in the atmosphere; used to quantify the organized rotational intensity of a cyclone's core.
- **RMSE (Root Mean Square Error):** A standard pixel-wise evaluation metric calculated across all un-normalized atmospheric variables (Roberts & Lean, 2008).
- **SACC (Synoptic Anomaly Correlation Coefficient):** An adaptation of the ACC metric that applies a spatial Gaussian filter prior to correlation, mitigating the double-penalty effect for structurally sound but slightly displaced storm predictions (Wilks, 2011).
- **SLP (Sea-Level Pressure):** One of the five primary atmospheric variables analyzed in this study; the minimum SLP is tracked to determine the center coordinates of the cyclonic storms.
- **SNR (Signal-to-Noise Ratio):** A weighting mechanism applied to the reverse denoising training objective to prioritize physical accuracy and stabilize intermediate noise scales (Hang et al., 2023).
- **Specific Humidity:** One of the five primary atmospheric variables analyzed in this study, measured at the 500hPa pressure level to track moisture distribution and frontal boundaries.
- **SSIM (Structural Similarity):** A computer vision metric used to observe the preservation of invariant pre-disaster spatial structures (e.g., roads and footprints) in generated images (Wang et al., 2004).
- **SVD (Stable Video Diffusion):** An off-the-shelf image-to-video diffusion baseline model evaluated for its zero-shot physical viability in storm synthesis (Blattmann et al., 2023).
- **Synoptic Scale:** A horizontal length scale in meteorology referring to large-scale weather systems spanning roughly 1,000 to 2,500 kilometers across, such as extratropical cyclones.
- **Temperature:** One of the five primary atmospheric variables analyzed in this study, measured at the 925hPa pressure level to capture near-surface thermodynamic gradients.
- **Transformer:** A prominent neural network architecture that relies on attention mechanisms to capture long-range dependencies in data; widely used as the backbone for modern foundation models.
- **U-component of Wind:** One of the five primary atmospheric variables analyzed in this study, representing the East-West horizontal velocity of wind at the 500hPa pressure level.
- **U-Net:** A convolutional neural network architecture commonly used in diffusion models, featuring a symmetric encoder-decoder structure that preserves spatial resolution for tasks like storm synthesis.
- **V-component of Wind:** One of the five primary atmospheric variables analyzed in this study, representing the North-South horizontal velocity of wind at the 500hPa pressure level.
- **WFM (Weather Foundation Model):** Large-scale machine learning architectures (such as ClimaX and Aurora) pretrained on vast amounts of atmospheric data to perform generalized weather tasks.
- **xBD:** A multi-disaster dataset of pre- and post-disaster satellite imagery, used in this work to assess building damage and evaluate cross-domain change modeling capabilities (Gupta et al., 2019).
- **Zero-shot Evaluation:** An evaluation paradigm where a machine learning model is tested on a task or dataset it was not explicitly trained for, demonstrating its ability to generalize to unseen scenarios without parameter updates.

B Background on Diffusion Modeling

A Diffusion Model (DM) is capable of generating synthetic datapoints by learning a mapping from pure Gaussian noise to the desired data distribution (Ho et al., 2020). Denote a training dataset by $D \subseteq \mathbb{R}^d$. The *forward diffusion process* is responsible for transforming a real datapoint $\mathbf{x}_0 \in D$ with distribution $q(\mathbf{x}_0)$ into a tensor of Gaussian noise $\mathbf{x}_S$ over S diffusion steps. This process is represented by a Markov chain and acts according to a variance schedule $\{\beta_1, \beta_2, \ldots, \beta_s\}$, to produce

$$q(\mathbf{x}_s|\mathbf{x}_{s-1}) = \mathcal{N}(\mathbf{x}_s; \sqrt{1-\beta_s}\mathbf{x}_{s-1}, \beta_s\mathbf{I}) \text{ and} \tag{3}$$

$$q(\mathbf{x}_{1:S}|\mathbf{x}_0) = \prod_{s=1}^{S} q(\mathbf{x}_s|\mathbf{x}_{s-1}), s \in \{1, \ldots, S\}. \tag{4}$$

The forward process can also be described in the closed form across steps as

$$q(\mathbf{x}_s|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_s; \sqrt{\bar{\alpha}_s}\mathbf{x}_0, (1-\bar{\alpha}_s)\mathbf{I}),$$

where $\alpha_s = 1 - \beta_s$ and $\bar{\alpha}_s = \prod_{i=1}^{s} \alpha_i$.

In the *reverse denoising process*, the denoiser network uses $\mathbf{x}_S$ to predict a denoised version of $\mathbf{x}_0$ in steps. Given that $p_\Theta(\mathbf{x}_s)$ is the distribution at noise step s parameterized by the denoiser Θ, this process is performed as

$$p_\Theta(\mathbf{x}_{0:S}) = p(\mathbf{x}_S) \prod_{s=1}^{S} p_\Theta(\mathbf{x}_{s-1}|\mathbf{x}_s). \tag{5}$$

For each noise step $s \in \{1, \ldots, S\}$, the training process aims to minimize the Kullback-Leibler (KL) Divergence (Kullback & Leibler, 1951) between the predicted and ground-truth sample distribution at noise step $s-1$,

$$\mathcal{L}_s = D_{KL}(q(\mathbf{x}_{s-1}|\mathbf{x}_s)||p_\Theta(\mathbf{x}_{s-1}|\mathbf{x}_s)).$$

The term $q(\mathbf{x}_{s-1}|\mathbf{x}_s)$ is computed from $q(\mathbf{x}_s|\mathbf{x}_{s-1})$ using Bayes theorem and the open-form equation for the forward process. Note that in practice, most diffusion implementations (including GeoDES) use mean squared error (MSE) loss in place of KL Divergence, for computational efficiency and to encourage sharper, more detailed outputs (Ho et al., 2022).

C Weather Modeling Experimental Setup

In this appendix, we provide further information on the setups for the extratropical cyclone experiments performed in Section 3. Appendix C.1 includes information on how non-synthetic data, as well as model outputs, were processed, while Appendix C.3 provides further elaboration on the baselines included and Appendix C.4 describes model hyperparameters. Appendix F provides details on how computational requirements were determined as presented in Section 3.3.

C.1 Data Processing Details

Dataset and task design: We focus on extratropical cyclones, which are mid-latitude low-pressure systems that are primary drivers of global atmospheric energy transfer (Martin, 2013). Forming along surface temperature gradients (fronts) (Catto, 2016), they are associated with severe winds and heavy precipitation (Hawcroft et al., 2012; Attinger et al., 2021), making them at times highly destructive weather events (Steiner, 2021).

Global atmospheric data is sourced from the ERA5 reanalysis dataset (Hersbach et al., 2020), with training data spanning from January 1940 to August 2015 and test data from January 2016 through December 2024. This temporal partition prevents co-occurrence leakage in the global model baselines, which may otherwise observe a test set storm during training if a separate train set storm happens to occur simultaneously. As a byproduct of shifting global climate, this split induces a slight distributional shift towards intense storms in the test set (Figure 5).

As our target use-case is the study and evaluation of LSP by climate scientists, we focus on a lightweight selection of five climate variables useful to analyzing an extratropical cyclone: sea-level pressure (SLP), U- and V-components of wind at the 500hPa pressure-level, temperature at 925hPa and humidity at 500hPa. While global models are trained on the full training time span and permitted to view the full global data, cyclone-centered baselines are trained solely on cyclone-centered bounding boxes (such as Figure 3). At test-time, the outputs of cyclone-centered models are used directly, while cyclone-centered bounding boxes must first be extracted from global models' outputs for a fair comparison. Detailed processing and extraction steps are outlined below.

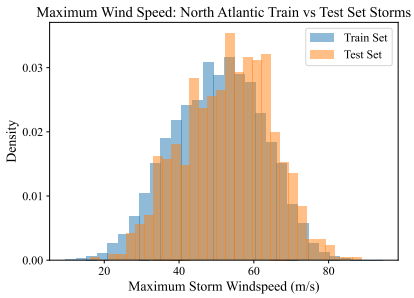


Figure 5: Normalized distribution of maximum wind speeds for train vs test set. The test set exhibits a slightly higher intensity bias, reflecting recent climatological trends.

Storm selection: Extratropical cyclones are identified within the ERA5 Sea Level Pressure 1940-2024 6-hourly global data on a $0.25° \times 0.25°$ grid using the NASA Modeling, Analysis and Prediction Climatology of Mid-latitude Storminess (MCMS) codebase[1]. For each extratropical cyclone, the codebase returns a list of timeframes in which the cyclone occurs, as well as the latitude and longitude of the storm's low-pressure center at each timeframe. We filter to only include storms that begin equatorward of $70°$ and that persist for at least 8 timeframes. For storms longer than 8 timepoints, we only include the first 8 timeframes. Next, we use a region mask to categorize and retain storms that occur in the North/South Atlantic and Pacific basins.

Global models:
Train-time procedure: Global model codebases are designed to expect continuous 6-hourly global data over a given period. Accordingly, models are trained on 6-hourly data over our entire training period (January 1940-August 2015).

Test-time procedure: We sample each model using a 7-step lead time. For each storm in our test set (January 2016-December 2024), we select the model output that is prompted from the timestep that the storm begins. Beginning at the ground-truth initial storm location, we then track the storm through the synthetic global forecast in subsequent timesteps by locating the minimum Sea Level Pressure (SLP) within a $10° \times 10°$ bounding box centered on the previous frame's location. At each timestep, we then extract a 1600×1600km equal-area bounding box, centered at the storm's minimum SLP location. We perform interpolation as necessary to reach a $32 \times 32 \times V$ tensor, where V is the number of climate variables.

ClimaX SLP Derivation: Because pretrained ClimaX does not offer SLP, we instead condition on and predict geopotential at 925 hPa and 2-meter temperature. The ClimaX-predicted SLP is then derived via the hypsometric equation,

$$P_{slp} = \ell e^{\Phi/(R_d T_{2m})},$$

where ℓ is the reference pressure level (925 hPa), Φ is the predicted geopotential, R_d is the specific gas constant for dry air and T_{2m} is 2-meter temperature. Consequently, storm tracking for ClimaX relies on geopotential minima in place of SLP.

Patch-based models (GeoDES and SVD):
Train-time procedure: Using the ground-truth storm location data created by MCMS for train-time storms in a given basin of interest, we extract bounding boxes around each storm at each timestep of its lifecycle, centered on its minimum SLP location. Interpolating as needed, each storm is represented as a $32 \times 32 \times V$ tensor, where the first two dimensions represent grid area and V is the number of climate variables ($V = 5$ in our experiments). GeoDES and SVD are trained on these samples directly.

Test-time procedure: Forming ground-truth test set datapoints in an analogous fashion to the GeoDES and SVD train set datapoints above, both models are then prompted with the 0-th timeframe of each storm in the test set.

[1] https://github.com/jfbooth/MIKE_BAUERS_MCMSV4

Table 5: Comparison of the methods evaluated in Section 3. The first row specifies the underlying architecture of each model, the second row lists whether the model operates directly on pixel space or uses an additional encoder model, while the third row states whether the model is a pretrained WFM.

	GeoDES	SVD	CEF	CoDiCast	ClimaX	Aurora
Architecture	Diffusion	Diffusion	Diffusion	Diffusion	Transformer	Transformer
Encoder vs Pixel Space	Pixel	Pixel	Pixel	Encoder	Pixel	Encoder
Pretrained Foundation Model	No	No	No	No	Yes	Yes

Evaluations:
Given the storm-centered bounding boxes created by or extracted from each method, evaluations are performed against the ground-truth test set storm-centered bounding boxes created during the GeoDES/SVD test-time procedure.

C.2 Metrics Details

Metrics: We evaluate our architectures using a hybrid suite of metrics drawn from both deep learning and meteorology, broadly categorized into three areas:

Physical and Structural Coherence:

- *Peak Relative Vorticity Error:* The maximum rotational intensity of the cyclone via spatial gradients ($\zeta = \frac{dv}{dx} - \frac{du}{dy}$), identifying models that fail to generate tight, organized cyclones (Holton & Hakim, 2012).
- *High-Frequency (HF) Spectral Ratio:* The ratio of integrated energy in the high-frequency domain of the Power Spectral Density between the prediction and ground truth (Skamarock, 2004; Ravuri et al., 2021). This metric exposes models that output overly-smoothed, safe predictions, or inflate intensity scores by outputting unconstrained, high-frequency noise. An ideal ratio is 1.0.

Extreme Intensity and Distribution:

- *Frequency Bias Index (FBI):* The ratio of predicted extreme wind pixels (> 20 m/s) to actual extreme wind pixels (Wilks, 2011). An FBI of 1.0 indicates perfect extreme energy recovery, while > 1.0 indicates generative hallucination (over-forecasting) and < 1.0 indicates unphysical smoothing.
- *Fractions Skill Score (FSS):* A neighborhood-based spatial verification metric (using a 3×3 grid cell window) that measures the spatial overlap accuracy specifically for the extreme (> 20 m/s) wind distribution (Roberts & Lean, 2008). FSS is formulated as 1 minus the ratio of the fractional Mean Squared Error (MSE) to a worst-case reference MSE, yielding a score strictly bounded between 0.0 (no overlap) and 1.0 (perfect overlap).

Spatial and Navigational Accuracy:

- *Synoptic Anomaly Correlation Coefficient (ACC):* Measures the spatial pattern correlation of the predicted fields (Wilks, 2011; Jolliffe & Stephenson, 2011). To avoid penalizing structurally sound storms that are offset from ground-truth examples, we apply a spatial Gaussian filter ($\sigma = 1.5$) prior to correlation, allowing for a fair comparison against heavily smoothed autoregressive baselines.

We also provide average and per-variable root mean square error (RMSE), synoptic and non-synoptic ACC scores in Appendix D.3. Our primary metrics listed above were selected to give a fair evaluation of weather synthesis, rather than forecasting performance: Pixel-wise metrics (e.g., RMSE) can be misleading for weather synthesis tasks, penalizing realistic features that are slightly displaced compared to ground truth, through a phenomenon known as the "double penalty" problem (Roberts & Lean, 2008; Ebert, 2008).

C.3 Methods Details

Methods: We evaluate GeoDES against a diverse suite of baselines (Table 5), including statistical climatology, prior diffusion-based architectures and pretrained global Weather Foundation Models (WFMs). To ensure relevance for computationally-accessible, specialized storm analysis, these baselines were selected according to three criteria: (1) feasibility of execution on a single GPU; (2)

the capacity to be trained and sampled using a sparse, five-variable atmospheric state rather than a full deep-column ERA5 initialization; and (3) the availability of a public codebase and, for WFMs, pretrained weights. We describe each baseline below.

- *Climatology:* As a naive baseline, we average all training datapoints together, forming a single composite storm climatology (Rasp et al., 2020). We predict this average for all test datapoints.
- *Stable Video Diffusion (SVD):* Blattmann et al. (2023) represents an off-the-shelf diffusion baseline. We train a randomly initialized SVD model in pixel-space on linearly normalized data. Similar to GeoDES, time is represented as a native dimension within the model, enabling non-autoregressive sampling.

The remainder of our evaluation suite is composed of global forecasting models, selected due to the lack of modern patch-based baselines. To adapt these methods to our setting, we execute their training and inference scripts over global ERA5 data. For each storm, we extract the initial frame ($t = 0$) from the spatial coordinates of the ground truth storm, then track the storm through the synthetic global forecast via minimum Sea Level Pressure (SLP). More details are in Appendix C.1.

- *Continuous Ensemble Forecasting (CEF):* Andrae et al. (2025) is a diffusion-based global forecasting ensemble operating on a 5.625° grid. Conditioned on an initial global state, CEF maps the forecast into a latent space and uses a continuous deterministic ODE solver to generate ensemble trajectories in parallel.
- *CoDiCast:* Shi et al. (2025a) is a diffusion-based global forecasting model designed to operate on a 5.625° grid. Conditioned on an initial state encoded via a pretrained deterministic autoencoder, CoDiCast autoregressively predicts each subsequent atmospheric timestep.
- *ClimaX WFM (Public and Fine-tuned Variants):* Nguyen et al. (2023) is a transformer-based global WFM. Because pretrained ClimaX does not offer SLP, we instead derive SLP from ClimaX geopotential and 2-meter temperature predictions—see Appendix C.1. We use the highest-resolution checkpoint (1.40625°). ClimaX's pretraining set spans January 1940-December 2015, which aligns with our train/test split. We include two variants of ClimaX in our evaluations: *ClimaX FT*, which we fine-tune on our five-variable subset of ERA5, and the publicly-available *ClimaX*, without additional fine-tuning.
- *Aurora WFM:* Bodnar et al. (2024) is a transformer-based global WFM with an additional encoder. We use the 0.25° checkpoint. We note that publicly available Aurora checkpoints are fine-tuned on 2016–2021 data. Consequently, evaluation on our 2016–2024 test set contains temporal data leakage. Even with this artificial advantage, we demonstrate in Sections 3.1 and D.1 that off-the-shelf global WFMs fail to resolve the fine-grained rotational dynamics of extratropical cyclones.

C.4 Hyperparameters

GeoDES: GeoDES was implemented in Python 3.11.11 with Diffusers 0.35.0 and PyTorch 2.6.0. U-Net block out channels are set to $[512, 1024, 2048]$ and we leave the layers per block and cross-attention dimension to the diffusers library defaults of 2 and 768, respectively.

We train GeoDES for a maximum total of 40 epochs (20 at 2D phase and 20 at 3D phase). The 2D and 3D phases are both performed with a patience of 5 epochs, with patience determined according to the training loss. The learning rate is $1e{-5}$ at the start of both 2D and 3D phases, selected from a search over $\{1e{-3}, 1e{-4}, 1e{-5}, 1e{-6}, 1e{-7}, 1e{-8}\}$. We perform the learning rate search at the 2D phase, select the two learning rates with lowest losses ($1e{-4}$ and $1e{-5}$), then perform the 3D loss search using both of these 2D settings to select the 3D model with lowest overall loss. Both 2D and 3D phases use the diffusers cosine schedule learning rate decay implementations with no warmup. We use MSE loss function and SNR of 5.

The training phase uses an AdamW optimizer and a DDPM scheduler with 1000 timesteps. Our batch size is 1, with 8 gradient accumulation steps and no weight decay. Model weights are updated directly without the use of an Exponential Moving Average (EMA).

At the 3D phase, we train and sample with BFloat16 (bf16) mixed precision and use temporal noise correlation $\rho = 0.95$ (see Section 2.3 for details). At sample time, we use the DDIM sampler

with 50 timesteps and $\eta = 1.0$ (we find that $\eta = 0.0$ also produces viable, but more conservative, predictions).

Baselines: For the prior work baselines (CoDiCast, CEF, ClimaX, ClimaX FT, Aurora), we use the highest-resolution model available, with the hyperparameters selected in their respective codebases. We use BFloat 16 mixed precision fine-tuning for ClimaX FT. For SVD, we use the same hyperparameter search protocol as the GeoDES model to reach a learning rate of $1e{-}4$.

D Additional Storm Modeling Realism Results

We now present additional analyses of our main paper experiments presented in Section 3. Appendix D.1 revisits the Section 3.1 experiment on extratropical cyclones, specifically exploring GeoDES' ability to realistically synthesize *extreme* storms. Appendix D.2 provides visualizations to accompany our main results on generative storm realism in Section 3.1, while Appendix D.3 contains additional performance metrics (average and per-variable RMSE, synoptic and non-synoptic ACC) for the experiments presented in Sections 3.1, D.1 and 3.4.

D.1 Severe Storm Representation

Claim 1A: GeoDES' realistic storm synthesis capabilities are especially apparent for *extreme* storms.

Setup: We follow a similar experimental setup to Section 3.1, evaluating strictly on the 90th percentile of most extreme storms within the North Atlantic basin test set (n=89), defined by storm-lifetime maximum wind magnitude (Young & Ribal, 2019). Given the relative rarity of extreme training examples, compounded by the out-of-distribution severity bias of the test set (see Figure 5), this subset represents a challenging synthesis task. Results are presented in Table 6 with supplemental metrics in Appendix D.3 and Table 8.

Results: Under the extreme storm setting, the performance gap between GeoDES and the baseline models widens regarding physical realism and the representation of extreme-energy behavior. While ClimaX FT achieves a highly competitive Fractions Skill Score (0.81 compared to GeoDES's 0.82) and the climatology baseline continues to display misleading ACC scores (refer to Section 3.1), most methods struggle to accurately represent the kinetic energy and structural integrity of severe storms. For example, the global WFMs (ClimaX and Aurora) experience catastrophic spatial smoothing. Both models produce FBI scores of 0.00 and HF Spectral Ratios near 0.00, indicating that their predictions are overly regularized by the climatological mean, preventing them from synthesizing extreme weather.

In contrast, **GeoDES generates realistic extreme storms.** GeoDES matches or slightly exceeds its full-dataset performance on key structural metrics. Despite the severity of this data subset, GeoDES correctly preserves the rotational core (Peak Vorticity Error of 4.47 ± 1.04) and safely recovers the vast majority of the extreme wind distribution (Frequency Bias Index of 0.83 ± 0.13). This performance demonstrates that the GeoDES architecture correctly parameterizes severe weather without reverting to the mean or hallucinating unphysical noise.

Table 6: Comparison of generative storm realism for the subset of most extreme North Atlantic extratropical cyclones (defined as top 10% of storm-lifetime wind magnitude). GeoDES maintains its strong synthesis performance demonstrated in Table 1 in this more challenging setting. In the far left column, (↓) and (↑) respectively indicate lower or higher scores are better, while $(\cdot 1)$ indicates that scores closer to 1 are better. Best results are bolded.

	GeoDES	SVD	CoDiCast	CEF	ClimaX FT	ClimaX	Aurora	*Climatology*
Peak Vorticity Error (↓)	$\mathbf{4.47 \pm 1.04}$	5.13 ± 0.02	6.90 ± 2.28	11.80 ± 0.95	11.14 ± 1.18	12.96	14.37	*8.28*
HF Spectral Ratio $(\cdot 1)$	$\mathbf{0.71 \pm 0.14}$	0.61 ± 0.01	2.02 ± 0.94	0.36 ± 0.02	0.27 ± 0.01	0.01	0.00	*0.19*
Frequency Bias Index $(\cdot 1)$	0.83 ± 0.13	$\mathbf{0.88 \pm 0.00}$	1.79 ± 0.08	0.42 ± 0.06	1.55 ± 0.08	0.00	0.00	*0.59*
Fractions Skill Score (↑)	$\mathbf{0.82 \pm 0.07}$	0.78 ± 0.00	0.75 ± 0.01	0.53 ± 0.04	0.81 ± 0.01	0.00	0.00	*0.79*
Synoptic ACC (↑)	$\mathbf{0.45 \pm 0.05}$	0.33 ± 0.01	0.06 ± 0.01	0.18 ± 0.02	0.34 ± 0.00	0.12	0.13	*0.70*
Synoptic ACC T-Final (↑)	$\mathbf{0.50 \pm 0.06}$	0.27 ± 0.00	0.04 ± 0.02	0.18 ± 0.01	0.28 ± 0.01	0.05	0.14	*0.70*

D.2 Supplemental Plots and Visualizations

In this section, we provide additional figures to visualize the performance of each method from Section 3.1. For explanations of each method, please refer to Appendix C.

Example storm outputs: We visualize the sea-level pressure, temperature and humidity outputs of each method compared to a real storm in Figures 6, 7, 8, respectively. The wind magnitude outputs for the same storm across all methods are presented in Figure 3.

Together, these figures support the narrative told by the performance metrics in Table 1. Global Weather Foundation Models (ClimaX, Aurora) and other global autoregressive models (CEF) suffer from catastrophic smoothing. They fail to resolve the sharp thermodynamic gradients of moving weather fronts (visible in humidity and temperature) and wash out the high-frequency energy bands of the wind field. CoDiCast presents a high-energy failure mode, with overly energetic wind field and steep spatial gradients in its SLP, temperature and humidity predictions. GeoDES is the only architecture capable of preserving both the sharp frontal boundaries and the organized, high-frequency turbulence of the cyclone, balancing structural realism with intensity.

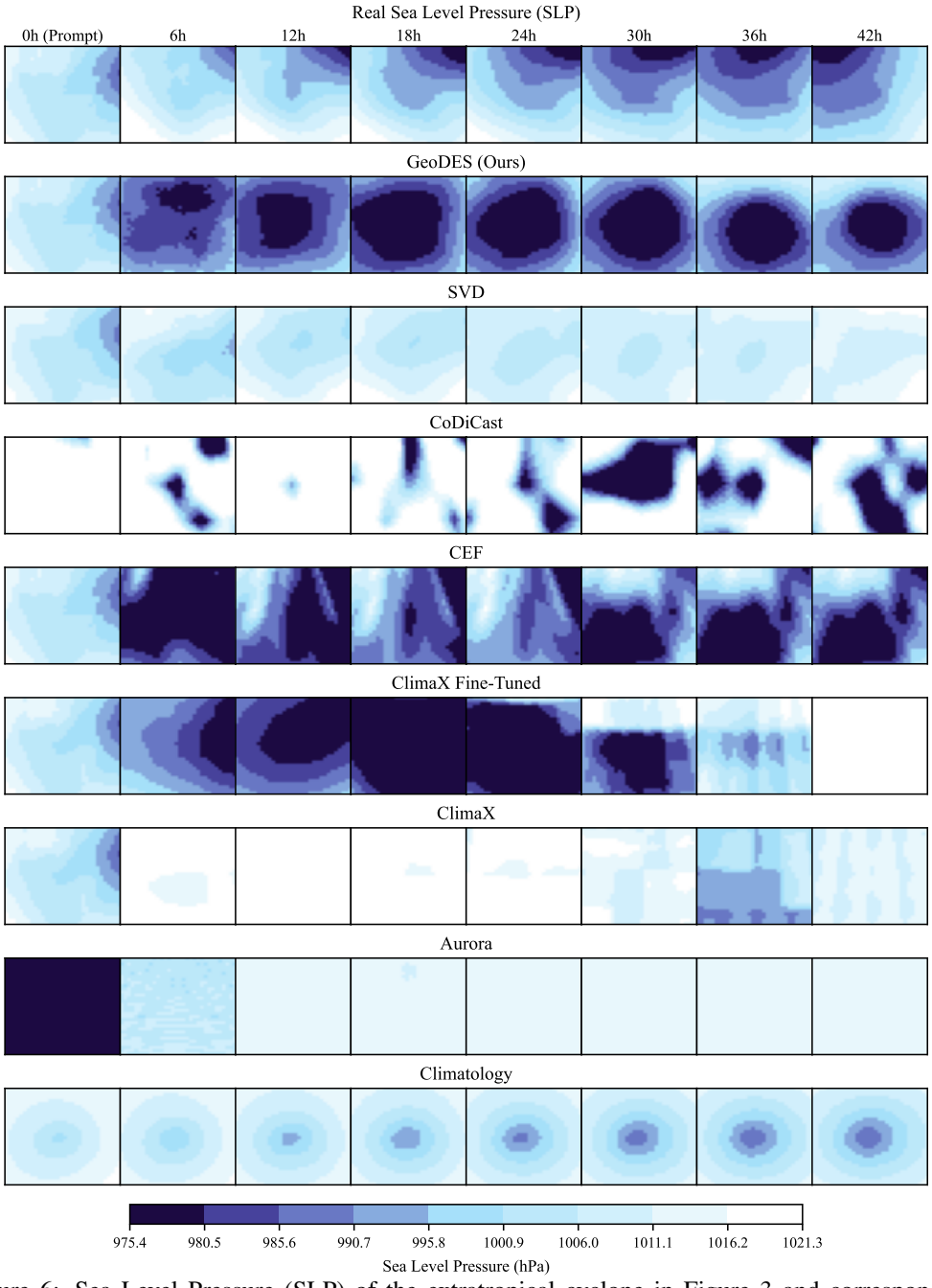


Figure 6: Sea Level Pressure (SLP) of the extratropical cyclone in Figure 3 and corresponding predictions from each synthesis method. While global WFMs correctly identify the macroscopic background pressure gradient, they fail to resolve the deep, localized minimum of the storm core. Conversely, GeoDES perfectly captures the tight pressure gradients and depth of the localized cyclone core.

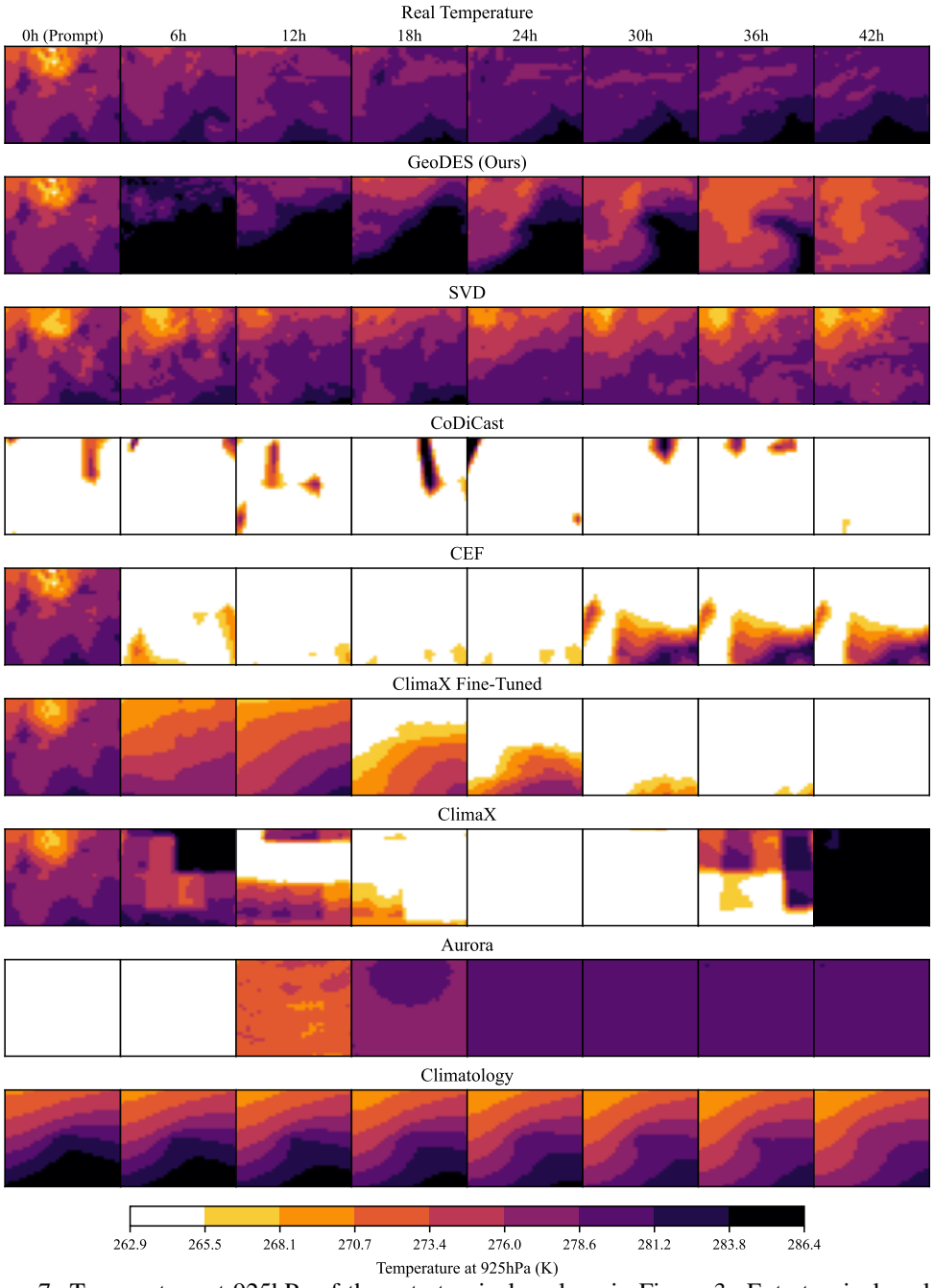


Figure 7: Temperature at 925hPa of the extratropical cyclone in Figure 3. Extratropical cyclones are fundamentally driven by the interaction of contrasting air masses. While baseline models act as spatial low-pass filters that blur these boundaries, GeoDES successfully preserves the sharp thermodynamic gradients of the moving weather fronts and the swirling thermal advection of the storm over the 42-hour rollout.

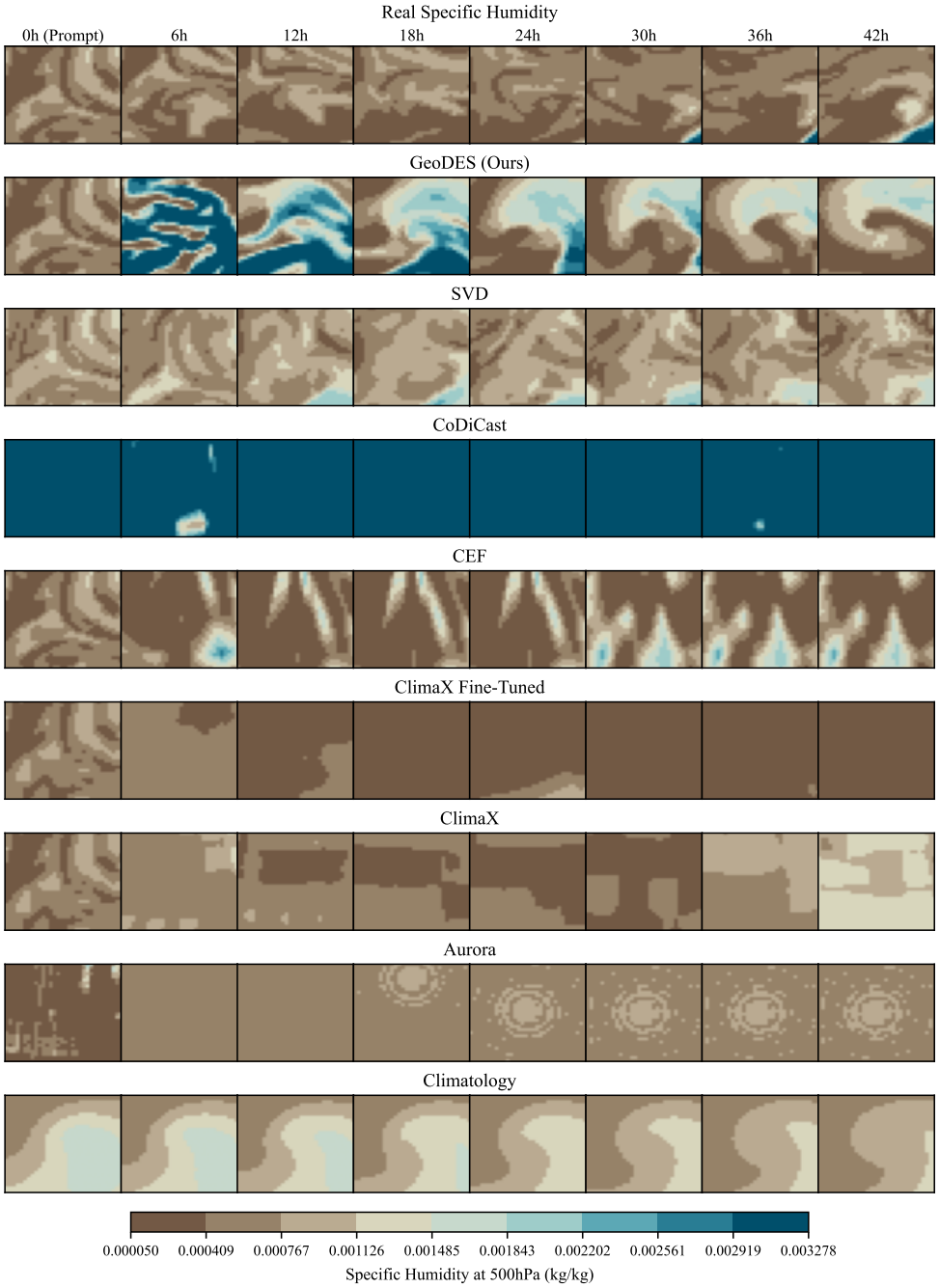


Figure 8: Specific Humidity at 500hPa of the extratropical cyclone in Figure 3. Similar to the temperature fields, GeoDES accurately maintains the structural integrity and sharp gradients of the moisture fronts, avoiding the catastrophic spatial smoothing exhibited by the global baselines and the unconstrained noise injection of CoDiCast.

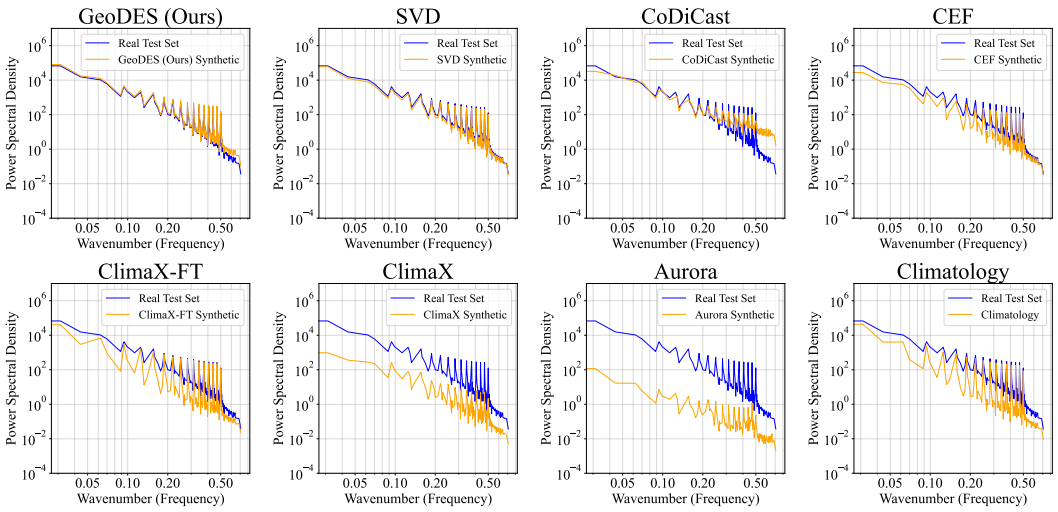


Figure 9: Power Spectral Density (PSD) of Wind Magnitude, visualizing how kinetic energy is distributed across spatial scales from macro-scale (left) to fine-grained, localized structures (right). Global models (ClimaX, Aurora) fall below the ground-truth test set (blue), indicating spatial smoothing. CoDiCast diverges above the ground truth at high frequencies, indicating noise hallucination. GeoDES uniquely preserves the kinetic energy distribution across spatial scales.

Power Spectral Density (PSD) Graphs: Figure 9 contains one PSD graph per method, with the synthesis method represented by the orange line and the non-synthetic test set represented by the blue line. A PSD plot visualizes how kinetic energy is distributed across different spatial scales within a storm. The x-axis represents spatial frequency (wavenumber), transitioning from large, macro-scale weather patterns on the left to fine-grained, highly localized storm structures on the right. A physically realistic model will produce a synthetic energy spectrum that tightly tracks the ground-truth observational spectrum.

Most models (SVD, CEF, ClimaX FT, ClimaX and Aurora) have PSD lines that fall consistently beneath the test set PSD line, resembling the climatology baseline PSD. This behavior is indicative of overly smooth predictions and a failure to generate fine grained storm structures, which agrees with the general failure mode of these models as described in Section 3.1 according to Table 1. Meanwhile, CoDiCast's PSD line falls above the test set line in the right half of their respective plots, which represents an over-prediction of high-energy features. While GeoDES presents a slight over-estimation bias in the highest-frequency features, across the full range of spatial frequencies, *GeoDES is the model that tracks the closest to the non-synthetic test set.*

Single-variable Storm-Lifetime Distributions: We use violin plots to compare methods' storm-lifetime distributions of maximum wind speed in Figure 10, minimum sea-level pressure in Figure 11 and maximum humidity in Figure 12.

Most methods fare acceptably with reproducing the general range and distribution of these single variables, except for the bias of global models towards calmer weather conditions (e.g. lower wind magnitudes, higher SLP and lower humidity). The two most notable exceptions are Aurora, with a strong bias towards high minimum SLP, and CoDiCast, with a bias towards low minimum SLP and apparent outlier values in its wind magnitude and humidity distributions. Despite both methods' use of batch-based normalization, these issues are perhaps artifacts of both methods' autoregressive sampling approaches. The core $(25\% - 75\%)$ percentiles of GeoDES' distributions track similarly to the test set distribution, although the violin plots represent a drawback of its normalization strategy: clipping between the 1st and 99th percentiles as described in Section 2 prevents the model from generating the outlier values that naturally occur in the non-synthetic test set. Future work in this area could explore more advanced normalization strategies that enable the generation of reasonable outlier values without allowing the predictions to explode into the overly energetic failure mode demonstrated by CoDiCast.

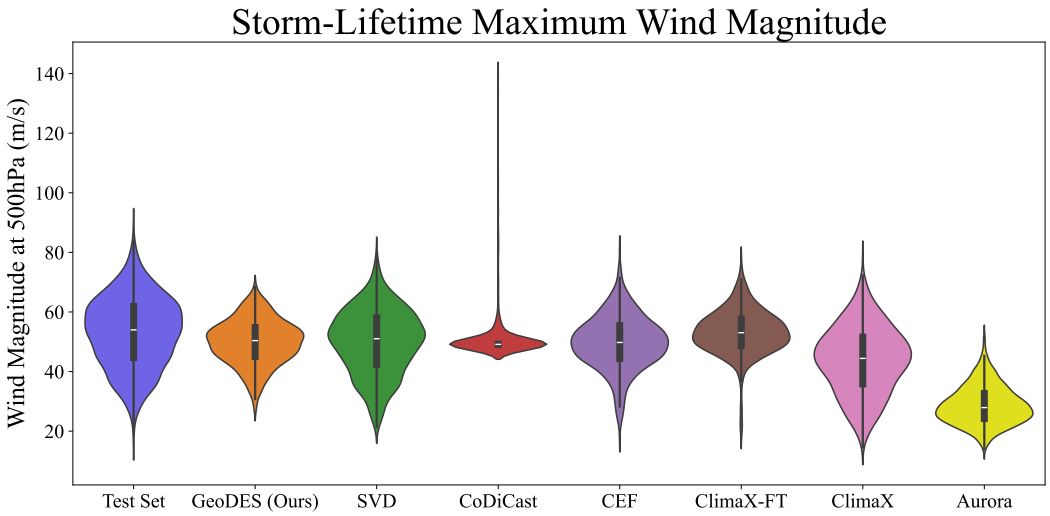


Figure 10: Violin plots detailing the distributions of storm-lifetime maximum wind speeds across the North Atlantic test set. While GeoDES successfully recovers the core distribution without succumbing to the unphysical runaway hallucinations of autoregressive baselines (e.g., CoDiCast), its maximum tail is bounded at roughly 75 m/s due to the 99th-percentile clipping enforced by the hybrid normalizer. This clipping prevents the undesirable outlier effects demonstrated by CoDiCast.

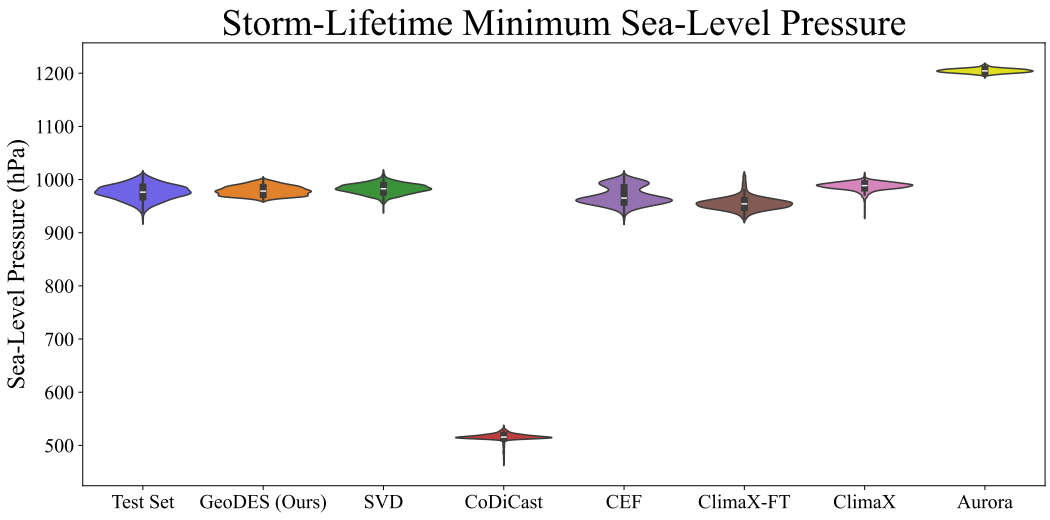


Figure 11: Distributions of storm-lifetime minimum Sea-Level Pressure (SLP) across the North Atlantic test set. Global foundation models exhibit a strong bias toward higher pressures (calmer weather), with Aurora suffering from severe pressure overestimation. CoDiCast hallucinates unphysically deep pressure minimums. GeoDES closely tracks the ground-truth distribution.

Wind-Pressure Relationship: Figure 13 displays the relationships between storm-lifetime maximum wind speed (y-axis) and minimum sea-level pressure (x-axis) for each method, in orange, compared to the real test set, in blue. Each marker represents one storm. These plots are useful for discerning methods' multivariate relationship modeling capabilities.

GeoDES' wind-pressure relationship is most similar to that of the non-synthetic test set. Similar to the patterns presented in previous metrics and plots, ClimaX and Aurora both underestimate wind magnitude and overestimate sea-level pressure in most storms. Oddly, Aurora's storms fall into three clear groups along the sea-level pressure axis (we perform one run, using the publicly available Aurora checkpoint and deterministic sampling; thus, these groups are not the result of

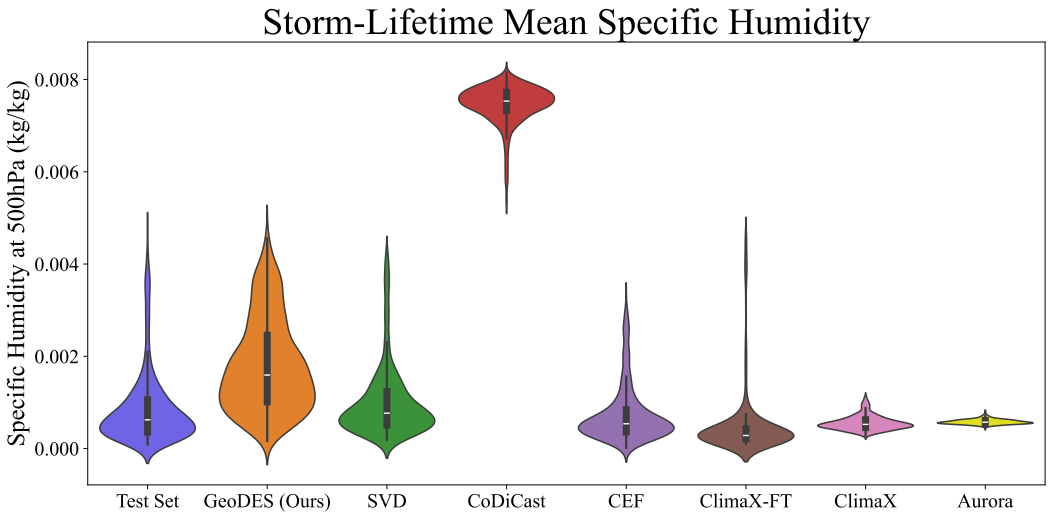


Figure 12: Distributions of storm-lifetime mean specific humidity at 500hPa across the North Atlantic test set. GeoDES captures the range of the true distribution, while most baseline models compress toward the calm-weather mean or generating unphysical outliers (CoDiCast).

differences across runs). ClimaX predictions fall into two general groups as well, with one group falling within the area occupied by the ground-truth test set storms and the other group falling into an area indicative of calm weather. ClimaX FT largely improves on these issues and resembles the test set more closely, although still with a bias towards calmer weather indicated by the large number of synthetic datapoints that fall below the area of test set points. A similar pattern is visible for the CEF model, while CoDiCast continues to output overly high-energy predictions.

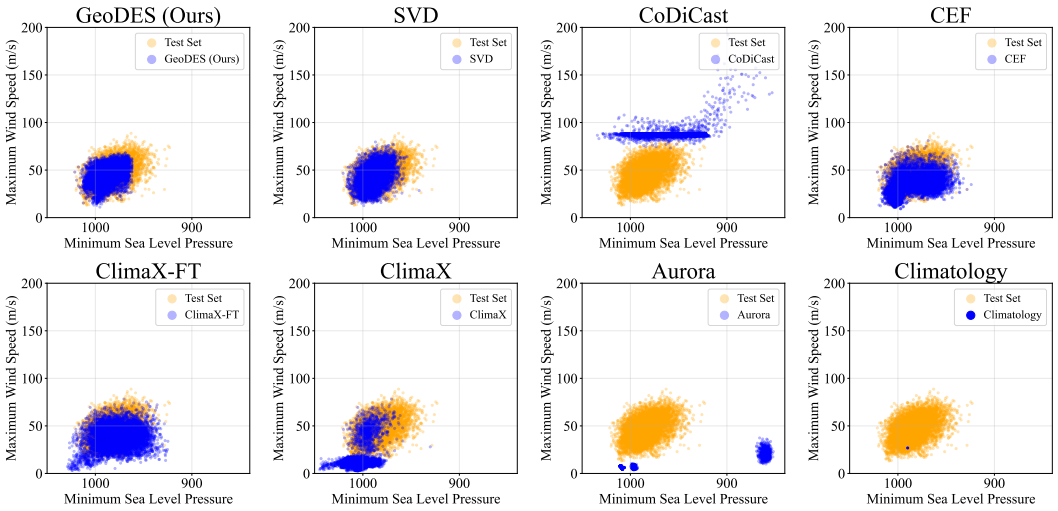


Figure 13: The multivariate physical relationship between storm-lifetime maximum wind speed and minimum sea-level pressure. Each marker represents a single storm system. GeoDES (orange) best matches the ground-truth test set distribution (blue), demonstrating its ability to jointly model the low pressure storm core with the corresponding kinetic increase in wind speed. Global models systematically underestimate wind magnitude while overestimating pressure.

D.3 SUPPLEMENTAL METRICS

We provide the following supplementary metrics for the synthesis experiments presented in Sections 3.1, D.1 and 3.4:

- *Overall Root Mean Square Error (RMSE):* The standard pixel-wise RMSE across all unnormalized variables (Wilks, 2011).
- *Per-variable RMSE:* The standard pixel-wise RMSE calculated individually for each of the five atmospheric variables (sea-level pressure, U- and V-wind components, temperature and specific humidity) to identify variable-specific generative performance (Wilks, 2011).
- *Skill vs Persistence:* A comparative skill score evaluating the model's predictive accuracy relative to a naive persistence forecast (where the initial conditions at $t = 0$ are held constant). It is calculated as $1 - (\text{MSE}_{\text{model}}/\text{MSE}_{\text{persistence}})$, where positive values indicate forecasting skill superior to persistence (Wilks, 2011).
- *Per-Variable Synoptic Anomaly Correlation Coefficient (SACC):* The spatial pattern correlation calculated individually for each atmospheric variable at the final forecast timestep (T=Final), following the application of a spatial Gaussian filter ($\sigma = 1.5$) to mitigate the double-penalty effect for structurally sound but slightly displaced features (Jolliffe & Stephenson, 2011).
- *Anomaly Correlation Coefficient (ACC):* The standard spatial pattern correlation between the predicted and ground-truth anomaly fields averaged across all timesteps, without any spatial smoothing applied, providing a strict evaluation of localized structural alignment (Jolliffe & Stephenson, 2011).
- *Per-Variable Anomaly Correlation Coefficient (ACC):* The standard, un-smoothed ACC calculated individually for each atmospheric variable at the final forecast timestep (T=Final) to isolate spatial correlation capabilities across disparate thermodynamic and kinematic fields (Jolliffe & Stephenson, 2011).

Results for the main experiment in Section 3.1 are in Table 7, Section D.1 on extreme storms in Table 8 and Section 3.4 ablations in Table 9. For each metric, best-performing models are bolded—as a naive baseline, we present climatology but do not bold its results. We analyze these results together, grouped by metric:

RMSE and Skill: As stated in Appendix C.2, RMSE suffers from the "double penalty" problem where synthesis methods that generate realistic novel storms are penalized twice (for the presence of the storm features that did not occur in the ground truth storm, as well as the lack of storm features that did occur in the ground truth storm), whereas overly smoothed methods are only penalized for the lack of storm features occurring in the ground truth (Roberts & Lean, 2008; Ebert, 2008). We include RMSE for completeness, due to the ubiquity of this metric.

GeoDES has the best overall synthesis model RMSE and skill vs persistence scores for both overall storm generation (Table 7) and extreme storm generation (Table 8). Still, Climatology achieves the best scores by a wide margin—for example, GeoDES' general storms RMSE is 13.95 compared to the Climatology's 9.79 in Table 7. This discrepancy highlights the known issues associated with this metric for weather synthesis assessment. Accordingly, it is difficult to assess whether a method scores low RMSE because it forecasts mostly accurate features and mostly in the same locations compared to the test set, or because it only predicts smooth predictions for all inputs to minimize its error. This pattern continues in the ablations, where the Linear Normalization method scores lowest RMSE in Table 9. Because outlier values skew the linear normalization scheme and prevent the model from making effective use of its dynamic range, it is reasonable that the model learns to generate smooth outputs that are more similar to the climatology or global baselines such as ClimaX and Aurora.

Table 7: Supplemental metrics to the North Atlantic storm synthesis experiment in Section 3.1. We report overall and per-variable Root Mean Square Error (RMSE), Skill vs. Persistence, time-averaged overall Synoptic Anomaly Correlation Coefficient (SACC), T=Final per-variable SACC, standard time-averaged overall Anomaly Correlation Coefficient (ACC) and T=Final per-variable ACC. GeoDES achieves the best overall generative performance among the evaluated dynamic models. Best results are bolded (excluding the statistical Climatology baseline).

	GeoDES	SVD	CoDiCast	CEF	ClimaX FT	ClimaX	Aurora	*Climatology*
Overall RMSE (↓)	$\mathbf{13.95 \pm 0.24}$	15.38 ± 0.00	36.21 ± 4.87	16.10 ± 0.24	18.19 ± 1.38	16.50	14.47	*9.79*
SLP RMSE (↓)	15.27 ± 1.46	15.73 ± 0.00	35.74 ± 5.07	19.63 ± 0.46	22.83 ± 1.02	20.71	**14.02**	*10.58*
Wind U RMSE (↓)	18.22 ± 0.11	21.65 ± 0.00	46.91 ± 6.63	17.84 ± 0.45	19.53 ± 1.32	**17.15**	18.22	*12.98*
Wind V RMSE (↓)	15.57 ± 0.29	16.62 ± 0.00	45.90 ± 12.15	20.10 ± 1.04	17.39 ± 0.10	**14.32**	15.16	*10.56*
Temperature RMSE (↓)	$\mathbf{12.75 \pm 1.62}$	13.80 ± 0.00	29.69 ± 5.66	13.71 ± 0.15	21.09 ± 3.83	20.80	17.00	*9.32*
Humidity RMSE (↓)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	0.01 ± 0.00	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	**0.00**	**0.00**	*0.00*
Skill vs Persistence (↑)	$\mathbf{-0.66 \pm 0.03}$	-0.83 ± 0.00	-3.30 ± 0.58	-0.91 ± 0.03	-1.16 ± 0.16	-0.96	-0.72	*−0.16*
Avg. SACC (↑)	$\mathbf{0.39 \pm 0.03}$	0.31 ± 0.00	0.06 ± 0.01	0.19 ± 0.02	0.28 ± 0.00	0.09	0.10	*0.64*
SLP SACC (↑)	$\mathbf{0.54 \pm 0.05}$	0.29 ± 0.00	0.28 ± 0.02	0.38 ± 0.07	0.33 ± 0.04	0.15	0.44	*0.73*
Wind U SACC (↑)	0.61 ± 0.04	0.31 ± 0.01	-0.02 ± 0.06	0.35 ± 0.09	$\mathbf{0.64 \pm 0.02}$	0.14	0.08	*0.77*
Wind V SACC (↑)	$\mathbf{0.48 \pm 0.02}$	0.21 ± 0.01	0.00 ± 0.02	-0.31 ± 0.15	-0.32 ± 0.03	0.02	0.08	*0.71*
Temperature SACC (↑)	0.45 ± 0.08	0.29 ± 0.01	0.04 ± 0.04	0.44 ± 0.09	$\mathbf{0.51 \pm 0.02}$	-0.04	-0.11	*0.68*
Humidity SACC (↑)	$\mathbf{0.10 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	-0.01 ± 0.00	0.07 ± 0.04	0.02 ± 0.05	0.00	0.04	*0.27*
Avg. ACC (↑)	$\mathbf{0.37 \pm 0.04}$	0.32 ± 0.01	0.06 ± 0.01	0.18 ± 0.03	0.27 ± 0.00	0.08	0.08	*0.63*
SLP ACC (↑)	$\mathbf{0.54 \pm 0.04}$	0.29 ± 0.00	0.28 ± 0.02	0.38 ± 0.07	0.33 ± 0.04	0.13	0.42	*0.74*
Wind U ACC (↑)	0.59 ± 0.04	0.30 ± 0.01	-0.01 ± 0.05	0.33 ± 0.09	$\mathbf{0.61 \pm 0.02}$	0.12	0.05	*0.76*
Wind V ACC (↑)	$\mathbf{0.45 \pm 0.02}$	0.20 ± 0.01	0.00 ± 0.01	-0.29 ± 0.15	-0.30 ± 0.03	0.01	0.07	*0.69*
Temperature ACC (↑)	0.43 ± 0.08	0.27 ± 0.01	0.03 ± 0.04	0.42 ± 0.09	$\mathbf{0.49 \pm 0.02}$	-0.04	-0.10	*0.66*
Humidity ACC (↑)	$\mathbf{0.09 \pm 0.01}$	0.08 ± 0.01	-0.01 ± 0.00	0.06 ± 0.03	0.02 ± 0.04	0.00	0.01	*0.25*

Table 8: Supplemental metrics to extreme result experiment in Section D.1, where we evaluate on the top 10% most severe storms in the North Atlantic test set by maximum wind magnitude. We report overall and per-variable Root Mean Square Error (RMSE), Skill vs. Persistence, time-averaged overall Synoptic Anomaly Correlation Coefficient (SACC), T=Final per-variable SACC, standard time-averaged overall Anomaly Correlation Coefficient (ACC) and T=Final per-variable ACC. GeoDES maintains superior spatial pattern correlation (SACC and ACC) and the lowest overall RMSE among dynamic models even under extreme, out-of-distribution severity conditions.

	GeoDES	SVD	CoDiCast	CEF	ClimaX FT	ClimaX	Aurora	*Climatology*
Overall RMSE (↓)	$\mathbf{14.79 \pm 1.31}$	17.50 ± 0.00	37.43 ± 5.93	18.26 ± 0.16	18.11 ± 1.13	19.88	18.16	*11.69*
SLP RMSE (↓)	$\mathbf{16.35 \pm 0.99}$	19.27 ± 0.01	35.82 ± 3.34	18.77 ± 0.60	21.01 ± 2.11	27.13	19.97	*14.20*
Wind U RMSE (↓)	$\mathbf{20.88 \pm 2.42}$	25.94 ± 0.00	52.01 ± 11.78	23.98 ± 0.29	21.06 ± 0.49	24.17	25.90	*16.99*
Wind V RMSE (↓)	$\mathbf{16.15 \pm 0.32}$	18.30 ± 0.00	46.35 ± 11.17	22.30 ± 1.13	19.82 ± 0.18	17.20	18.22	*12.47*
Temperature RMSE (↓)	$\mathbf{11.20 \pm 2.66}$	12.35 ± 0.00	28.11 ± 4.44	15.54 ± 0.37	18.90 ± 2.81	18.97	15.72	*6.13*
Humidity RMSE (↓)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	0.01 ± 0.00	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	**0.00**	**0.00**	*0.00*
Skill vs Persistence (↑)	$\mathbf{-0.32 \pm 0.12}$	-0.56 ± 0.00	-2.33 ± 0.53	-0.63 ± 0.01	-0.61 ± 0.10	-0.77	-0.62	*−0.04*
Avg. SACC (↑)	$\mathbf{0.45 \pm 0.05}$	0.33 ± 0.01	0.06 ± 0.01	0.18 ± 0.02	0.34 ± 0.00	0.12	0.13	*0.70*
SLP SACC (↑)	$\mathbf{0.58 \pm 0.10}$	0.33 ± 0.01	0.29 ± 0.03	0.41 ± 0.04	0.32 ± 0.04	0.13	0.52	*0.79*
Wind U SACC (↑)	0.74 ± 0.07	0.38 ± 0.01	-0.04 ± 0.08	0.29 ± 0.07	$\mathbf{0.78 \pm 0.02}$	0.21	0.27	*0.90*
Wind V SACC (↑)	$\mathbf{0.51 \pm 0.06}$	0.21 ± 0.00	-0.03 ± 0.04	-0.30 ± 0.19	-0.27 ± 0.04	-0.01	0.04	*0.75*
Temperature SACC (↑)	0.55 ± 0.12	0.32 ± 0.01	0.01 ± 0.08	0.44 ± 0.12	$\mathbf{0.57 \pm 0.03}$	-0.08	-0.20	*0.79*
Humidity SACC (↑)	$\mathbf{0.13 \pm 0.02}$	0.11 ± 0.00	-0.02 ± 0.03	0.06 ± 0.04	0.00 ± 0.05	-0.01	0.06	*0.26*
Avg. ACC (↑)	$\mathbf{0.43 \pm 0.06}$	0.32 ± 0.00	0.06 ± 0.02	0.18 ± 0.03	0.33 ± 0.00	0.10	0.10	*0.69*
SLP ACC (↑)	$\mathbf{0.58 \pm 0.09}$	0.33 ± 0.00	0.29 ± 0.02	0.41 ± 0.04	0.32 ± 0.04	0.12	0.49	*0.79*
Wind U ACC (↑)	0.71 ± 0.07	0.37 ± 0.00	-0.04 ± 0.08	0.28 ± 0.07	$\mathbf{0.75 \pm 0.02}$	0.18	0.18	*0.88*
Wind V ACC (↑)	$\mathbf{0.49 \pm 0.06}$	0.20 ± 0.00	-0.03 ± 0.04	-0.29 ± 0.19	-0.26 ± 0.04	-0.01	0.03	*0.73*
Temperature ACC (↑)	0.52 ± 0.12	0.31 ± 0.00	0.01 ± 0.07	0.42 ± 0.13	$\mathbf{0.55 \pm 0.03}$	-0.08	-0.18	*0.77*
Humidity ACC (↑)	$\mathbf{0.11 \pm 0.02}$	0.09 ± 0.00	-0.02 ± 0.03	0.06 ± 0.03	0.00 ± 0.04	0.00	0.02	*0.24*

SACC and ACC: Synoptic and non-synoptic ACC induce similar relative orderings of methods. Synoptic scores are globally higher than non-synoptic scores, indicating that all methods benefit from the use of the synoptic spatial filter. *GeoDES scores the highest average ACC and SACC in overall storm generation, extreme storm generation and the ablations (Tables 7, 8 and 9).* Most global baselines perform poorly on these metrics, while Climatology is surprisingly competitive, with best overall (S)ACC in extreme storm generation (Table 8). We highlight that *SVD's lower performance highlights the importance of GeoDES' approach to achieving high performance as a storm-centered method, without the global context afforded to the other baselines.*

Table 9: Supplemental metrics to ablations in Section 3.4. We report overall and per-variable Root Mean Square Error (RMSE), Skill vs. Persistence, time-averaged overall Synoptic Anomaly Correlation Coefficient (SACC), T=Final per-variable SACC, standard time-averaged overall Anomaly Correlation Coefficient (ACC) and T=Final per-variable ACC. The table highlights the performance degradation associated with removing Signal-to-Noise Ratio (SNR) weighting, deterministic sampling (Non-Stoch.), using standard linear scaling (Linear Norm.) and removing the 2D pretraining phase (3D W/Out 2D). The Linear Normalization ablation achieves the lowest RMSE, mathematically demonstrating that lower RMSE often correlates with unphysical spatial smoothing and dynamic range collapse rather than structural realism.

	GeoDES	No SNR	Non-Stoch.	Linear Norm.	3D W/Out 2D
Overall RMSE (↓)	13.95 ± 0.24	14.61 ± 0.21	14.19 ± 0.07	$\mathbf{13.87 \pm 0.86}$	18.93 ± 4.88
SLP RMSE (↓)	15.27 ± 1.46	17.56 ± 1.42	$\mathbf{14.77 \pm 0.34}$	14.89 ± 1.13	26.02 ± 9.33
Wind U RMSE (↓)	18.22 ± 0.11	17.72 ± 0.82	19.19 ± 0.24	$\mathbf{17.51 \pm 0.79}$	24.62 ± 8.26
Wind V RMSE (↓)	15.57 ± 0.29	$\mathbf{14.72 \pm 0.16}$	15.13 ± 0.03	16.31 ± 2.05	15.59 ± 0.06
Temperature RMSE (↓)	$\mathbf{12.75 \pm 1.62}$	15.02 ± 1.54	13.84 ± 0.30	12.89 ± 0.32	15.73 ± 0.96
Humidity RMSE (↓)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$
Skill vs Persistence (↑)	-0.66 ± 0.03	-0.74 ± 0.02	-0.69 ± 0.01	$\mathbf{-0.65 \pm 0.10}$	-1.25 ± 0.58
Avg. SACC (↑)	$\mathbf{0.39 \pm 0.03}$	0.36 ± 0.04	0.37 ± 0.00	0.37 ± 0.01	0.33 ± 0.01
SLP SACC (↑)	$\mathbf{0.54 \pm 0.05}$	0.53 ± 0.03	0.47 ± 0.02	0.51 ± 0.01	0.34 ± 0.18
Wind U SACC (↑)	$\mathbf{0.61 \pm 0.04}$	0.52 ± 0.15	0.50 ± 0.00	0.56 ± 0.04	0.59 ± 0.08
Wind V SACC (↑)	0.48 ± 0.02	0.49 ± 0.03	0.44 ± 0.02	0.44 ± 0.03	$\mathbf{0.54 \pm 0.10}$
Temperature SACC (↑)	0.45 ± 0.08	0.41 ± 0.01	0.43 ± 0.02	$\mathbf{0.48 \pm 0.02}$	0.37 ± 0.13
Humidity SACC (↑)	0.10 ± 0.01	0.09 ± 0.06	0.10 ± 0.01	$\mathbf{0.12 \pm 0.03}$	0.07 ± 0.03
Avg. ACC (↑)	$\mathbf{0.37 \pm 0.04}$	0.35 ± 0.05	0.35 ± 0.00	0.35 ± 0.01	0.30 ± 0.03
SLP ACC (↑)	$\mathbf{0.54 \pm 0.04}$	$\mathbf{0.54 \pm 0.03}$	0.48 ± 0.02	0.52 ± 0.01	0.33 ± 0.20
Wind U ACC (↑)	$\mathbf{0.59 \pm 0.04}$	0.50 ± 0.14	0.48 ± 0.00	0.54 ± 0.04	0.56 ± 0.07
Wind V ACC (↑)	0.45 ± 0.02	0.46 ± 0.03	0.41 ± 0.02	0.42 ± 0.03	$\mathbf{0.51 \pm 0.09}$
Temperature ACC (↑)	0.43 ± 0.08	0.39 ± 0.01	0.41 ± 0.02	$\mathbf{0.45 \pm 0.02}$	0.34 ± 0.13
Humidity ACC (↑)	0.09 ± 0.01	0.08 ± 0.05	0.08 ± 0.01	$\mathbf{0.10 \pm 0.02}$	0.06 ± 0.03

Table 10: GeoDES results for each basin after 2D-phase pretraining on all basins, according to the setup in Appendix D.4. Model performance on fidelity metrics scales with the observational fidelity of the underlying basin data, rather than raw data volume, achieving highest ACC/SACC in the densely-observed Northern Hemisphere (North Pacific/Atlantic) and fuzzier predictions in the satellite-dependent Southern Hemisphere, despite greater overall data quantity and accordingly better energy-matching metrics (Frequency Bias Index, Peak Vorticity Error) in the Southern Hemisphere.

	N. Atlantic	S. Atlantic	N. Pacific	S. Pacific
# Train	8589	11262	5782	16358
# Test	893	1280	614	1917
Overall RMSE (↓)	14.60	15.05	12.96	14.63
Peak Vorticity Error (↓)	5.10	1.91	4.92	2.72
HF Spectral Ratio (·1)	0.71	1.30	1.04	0.91
Frequency Bias Index (·1)	1.16	1.04	0.79	1.17
Fractions Skill Score (↑)	0.77	0.78	0.66	0.83
Synoptic ACC (↑)	0.33	0.36	0.37	0.34
Synoptic ACC T-Final (↑)	0.38	0.40	0.40	0.41
ACC (↑)	0.32	0.34	0.35	0.33
ACC T-Final (↑)	0.36	0.38	0.38	0.39

D.4 Biases Introduced Via Whole-World Pretraining and Performance Across Basins

Setup: To evaluate the generalization capabilities of our diffusion architecture, we experimented with training on all storms from any of the four ocean basins at the 2D pretraining phase, followed by fine-tuning on a specific basin at the 3D fine-tuning phase. Notably, the distribution of storms in the pretraining dataset is highly asymmetric, following the order: North Pacific $<$ North Atlantic $<$ South Atlantic $<$ South Pacific.

Results: Table 10 displays an at first unusual relationship between data volume and model performance. GeoDES achieved its highest spatial accuracy and optimal variance in the North Pacific (Overall RMSE 12.96, Spectral Ratio 1.03), despite this basin representing the smallest fraction of the pretraining data.

Rather than a failure of the architecture, this hemispheric split highlights a well-documented disparity in global meteorological observational fidelity. Ground-truth reanalysis in the Northern Hemisphere is aided by dense on-site observations and aircraft reconnaissance, providing more structurally precise, high-fidelity targets. In contrast, Southern Hemisphere reanalysis relies predominantly on smoother, satellite-derived estimates characterized by higher spatial uncertainty.

The model's performance directly reflects this information gap. The sharp, high-quality gradients in the Northern Hemisphere data allowed GeoDES to learn precise spatial and multivariate relationships during the 2D phase, requiring minimal correction during 3D specialization. The high volume of South Atlantic data—a basin characterized by high wind shear and unorganized convection—taught the model a chaotic structural prior, contributing to the prediction of over-energized, high-frequency fields (Spectral Ratio of 1.30) generated when fine-tuning for that domain. Ultimately, GeoDES scales not with raw data volume, but with the physical fidelity of the underlying observations.

E Details on Disaster Modeling (xBD) Experiment

Dataset: We extract 64×64 bounding boxes from the raw satellite images: three randomly selected non-overlapping boxes per satellite image in the xBD-defined train split and one box per satellite image in the xBD validation and test splits. Accordingly, using variable names defined in Figure 1, each datapoint is $T \times V \times L_1 \times L_2 = 2 \times 3 \times 64 \times 64$. The train split has 8397 datapoints, while the validation and test splits each contain 933 datapoints.

Implementations: We use the publicly available PyTorch implementation of Pix2Pix[2]. We implement the Palette model as described in its paper (Saharia et al., 2022) using PyTorch and Diffusers libraries.

Hyperparameters: We use default hyperparameters for GeoDES, including a learning rate of $1e-5$ at 2D and 3D training stages. We run Pix2Pix with batch size 256, learning rate $1e-4$ and otherwise default hyperparameters, after performing a hyperparameter search over learning rate in $\{1e-4, 2e-4, 3e-4, 4e-4\}$. Palette is trained for 40 epochs, batch size 128 and learning rate $\{5e-5\}$, after performing a learning rate hyperparameter search over $\{1e-5, 5e-5, 1e-4, 5e-4\}$. For both Pix2Pix and Palette, the best hyperparameter setting was selected according to structural similarity on the validation set.

F Details on Computational Requirements Estimation

Model parameter count and FLOPs are both calculated using the Python `fvcore` library implementations for all models except the TensorFlow-implemented CoDiCast, for which we calculate these metrics from scratch. All measurements are performed on the same machine with an NVIDIA RTX A6000 GPU.

Estimation of Aurora FLOPs: Because Aurora operates on global $0.25°$ inputs, empirical FLOP calculation requires prohibitive amounts of system RAM and we instead estimate the FLOPs for this model.

Aurora is a 1.25B-parameter transformer-based model, with inputs of dimensions 721×1440 per variable. A Perceiver encoder model performs 3D convolutions to extract 4×4 patches, compressing the grid to $180 \times 360 = 64800$ tokens per level. Vertically, the encoder cross-attends the inputs to a latent depth of $Z = 4$ levels (one surface and three atmospheric), regardless of the number of input variables or atmospheric levels[3]. Accordingly, the encoded input contains $4 * 64800 = 259200$ tokens.

According to Kaplan et al. (2020), the cost of a single transformer inference-time forward pass may be estimated as $2PN$ for a P-parameter model and N input tokens. Accordingly, each Aurora forward pass is approximately $2 * (1.256 \times 10^9) * (2.59 \times 10^5) \approx 651$ TFLOPs. Multiplying by 7 inference timesteps to achieve the 48-hour forecasts within Section 3.1 yields a total approximation of 4557 TFLOPs.

We validate this FLOP estimate by calculating the corresponding Model FLOPs Utilization (MFU, Chowdhery et al. (2022)). As stated in Table 3, we observe a 47.6-second wall-clock latency for Aurora, which indicates a rate of 95.74 TFLOP/second. These evaluations were performed on an NVIDIA RTX A6000, with a dense FP16/BF16 Tensor Core peak throughput of 154.8 TFLOP/second. As such, the MFU is calculated as $\frac{\text{Model TFLOP/second}}{\text{Theoretical Peak Hardware TFLOP/second}} \approx 61.8\%$. Because Aurora performs a dense volumetric forward pass over a 259200 token sequence at each autoregressive step, the computation maintains high arithmetic intensity. As such, the observed MFU aligns well with the established real-world vs theoretical upper bounds of dense PyTorch Transformer utilization (Hagemann et al., 2024; Pope et al., 2023).

Estimation of Saved Compute Using 2D vs 3D GeoDES Training: The 3D model uses $123,662$ GFLOPs to produce all 8 frames of a single datapoint. The 2D model uses $1,182$ GFLOPs to produce a single (non-temporal) frame, or $9,455$ GFLOPs to produce 8 frames. As such, we estimate the saved compute:

$$\frac{1}{2} + \frac{1}{2} \cdot \frac{9455}{123662} = 0.538.$$

[2]Pix2Pix Implementation: `https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix`

[3]According to the official Microsoft Aurora implementation, `https://github.com/microsoft/aurora`