Data-Driven Weather Prediction

A Probabilistic Forecasting Framework

Nikola B. Kovachki

Joint work with: Jean Kossaifi, Morteza Mardani, Daniel Leibovici, Suman Ravuri, Ira Shokar, Edoardo Calvello, Mohammad Shoaib Abbas, Peter Harrington, Akshay Subramaniam, Noah Brenowitz, Boris Bonev, Wonmin Byeon, Karsten Kreis, Dale Durran, Arash Vahdat, Mike Pritchard, and Jan Kautz.

NVIDIA Research

Probabilistic Weather Forecasting

Weather Forecasting

- Immediate impact to disaster preparedness and global markets.
- Guide design and application of climate models.

Probabilistic Modeling

- The atmosphere is chaotic and initial conditions are uncertain.
- Deterministic forecasts are impractical for risk management.

Data-Driven Models

- Massive historical datasets.
- Advances in neural architectures and generative modeling.
- Orders-of-magnitude faster inference than traditional solvers.
- Large ensembles enable real-time probabilistic forecasting.

Problem Formulation

Atmospheric State Representation

Consider a stationary, Markovian, discrete-time stochastic process

$$\{x_j\}_{j\in\mathbb{Z}} \subset \mathbb{R}^d; \qquad p(x_j, x_{j+1}) = p(x_k, x_{k+1}),$$

representing atmospheric states.

Learning Objective

Estimate conditional transition density

$$p(x_1 \mid x_0)$$

from identically distributed samples $(x_j^\dagger, x_{j+1}^\dagger) \sim p(x_0, x_1)$.

Forecast Generation

Autoregressive rollout:

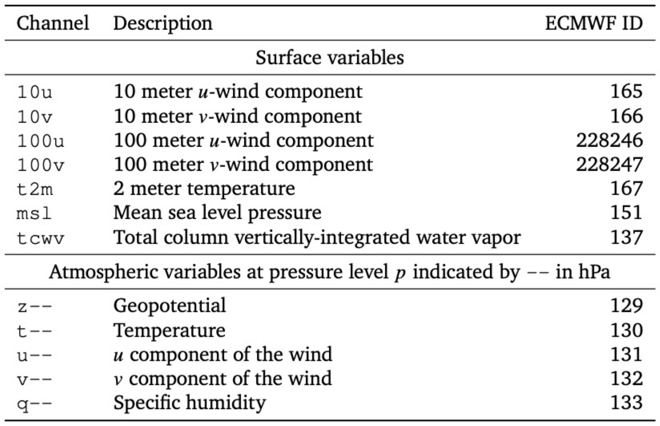
$$\hat{x}_{k+1} \sim p(\cdot \mid \hat{x}_k).$$

Challenges:

(1) high dimensionality $d \gg 1$; (2) distribution shift in rollout.

ERA5 Dataset 1980-2020

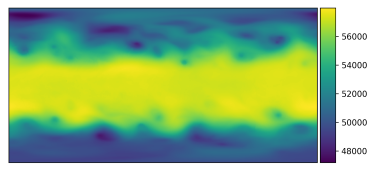
ERA5 Reanalysis Data



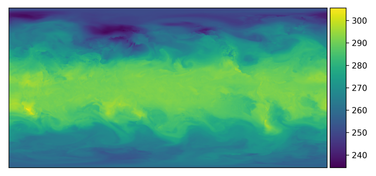
Channel	Description	ECMWF ID
	Surface variables	
`10u`	10 meter *u*-wind component	165
`10v`	10 meter *v*-wind component	166
`100u`	100 meter *u*-wind component	228246
`100v`	100 meter *v*-wind component	228247
`t2m`	2 meter temperature	167
`msl`	Mean sea level pressure	151
`tcwv`	Total column vertically-integrated water vapor	137
	Atmospheric variables at pressure level *p* indicated by `--` in hPa	
`z--`	Geopotential	129
`t--`	Temperature	130
`u--`	*u* component of the wind	131
`v--`	*v* component of the wind	132
`q--`	Specific humidity	133

- Grid: 721 × 1440 equiangular (0.25 deg).
- Sampling interval: 6 hours.
- Structure: 5 atmospheric (13 vertical level) and 7 surface variables.

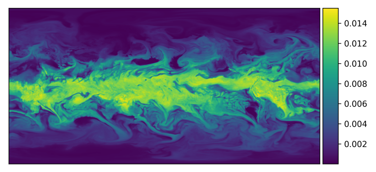
ERA5 Dataset Examples



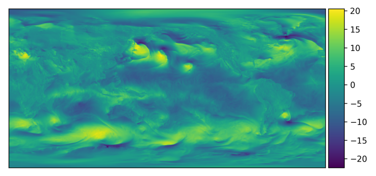
Geopotential (m^2s^{-2}) at 500 hPa



Temperature (K) at 850 hPa



Specific humidity ($kgkg^{-1}$) at 850 hPa



Longitudinal component of wind velocity (ms^{-1}) at 10 m

Properties of Atmospheric Data

Predictability and Length Scales

Atmospheric predictability is scale-dependent:

$$\tau(k) \sim \lambda(k)^{-1},$$

where $\lambda(k)$ is the Lyapunov exponent at wavenumber k.

- Synoptic scales ($\gtrsim$ 1000 km): $\tau \sim$ days.
- Mesoscales ($\lesssim$ 100 km): $\tau \sim$ hours.

At 6 hours, scales below $\approx$ 50–100 km are effectively unpredictable.

Energy Spectrum Considerations

Observed atmospheric kinetic energy spectrum:

$$E(k) \sim \begin{cases} k^{-3} & \text{(synoptic range)} \\ k^{-5/3} & \text{(mesoscale range)} \end{cases}$$

Most forecast-relevant variance lies at low k.

Latent Space Modeling

Latent Representation

Define downsampling operator $B : \mathbb{R}^d \to \mathbb{R}^{d_z}$ and latent spates

$$z_0 = B(x_0), \qquad r_1 = B(x_1 - x_0).$$

Model conditional density:

$$p(r_1 \mid z_0, z_{-1}).$$

- Residuals improves accuracy: map is perturbation of identity.
- Historic state improves stability: B breaks Markovian structure.

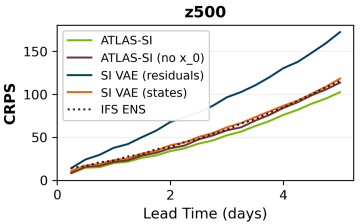
Decoder $D : \mathbb{R}^{d_z} \times \mathbb{R}^d \to \mathbb{R}^d$ approximately solves

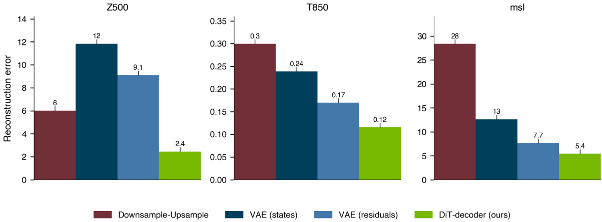
$$x_1 \approx x_0 + D(r_1, x_0).$$

Implementation

- B: bilinear interpolation to 181×360 equiangular grid (1.0 deg).
- D: Diffusion Transformer (DiT) with local attention (Natten).

Comparison to Autoencoders





Emperical Results

- Decoder only model gives more accurate reconstructions.
    - Learned latents contain high-frequency components.
- Rollout in learned latent space accumulates more error.
    - Latents have no temporal consistency.

Probabilistic Modeling Framework

Common Objective

Model latent conditional distribution $p(r_1 \mid z_0, z_{-1})$ via transport map

$$f_\theta : (\xi, z_0, z_{-1}) \to r_1.$$

Three Training Paradigms

Stochastic Interpolants

- Parameterize time-dependent drift term of a forward SDE.
- f_θ is solution operator of SDE.

Diffusion Models

- Parameterize time-dependent score of a data noising process.
- f_θ is solution operator of backward SDE/ODE.

CRPS-Based Generator

- f_θ is parametric transport map trained to minimize MMD variant.

Same transformer backbone supports all probabilistic formulations.

Diffusion Transformer Architecture

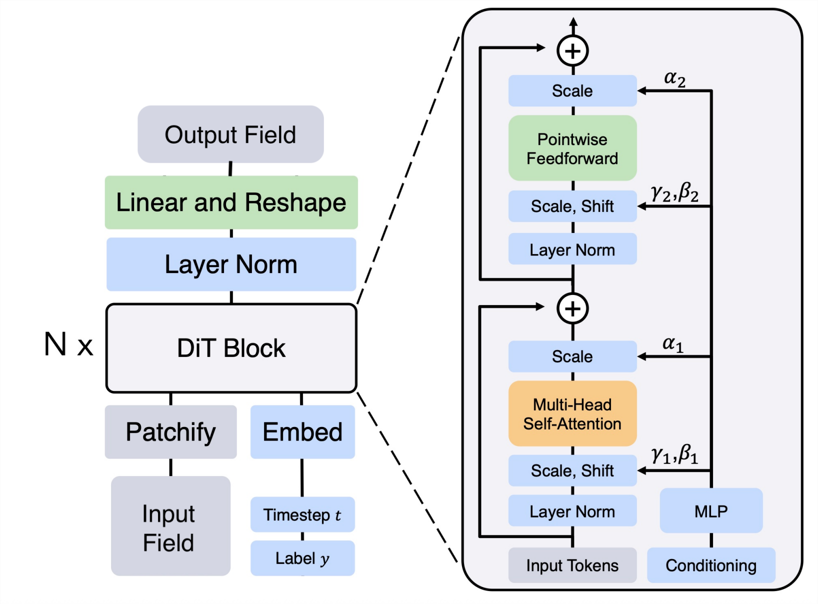


*Image adapted from Peebles, Xie, "Scalable Diffusion Models with Transformers." *International conference on computer vision.* 2023.

Stochastic Interpolants: Formulation

Interpolating Bridge Process

Define stochastic interpolant between z_0 and r_1

$$I_t = \alpha(t) z_0 + \beta(t) r_1 + \sigma(t) W_t, \quad t \in [0,1],$$

where W_t is a Wiener process and

$$\alpha(0) = 1,\ \alpha(1) = 0, \qquad \beta(0) = 0,\ \beta(1) = 1, \qquad \sigma(0) = \sigma(1) = 0.$$

Associated Forward SDE

The conditional law $\rho(r_1 \mid z_0, z_{-1})$ is recovered as terminal law of

$$dX_t = b(X_t, z_0, z_{-1} t)\, dt + \sigma(t)\, dW_t, \qquad X_0 = z_0.$$

Optimal drift minimizes

$$\int_0^1 \mathbb{E} \left| b(I_t, z_0, z_{-1}, t) - \dot{\alpha}(t) z_0 - \dot{\beta}(t) r_1 - \dot{\sigma}(t) W_t \right|^2 dt.$$

Chen et.al. "Probabilistic Forecasting with Stochastic Interpolants and Föllmer Processes." *PMLR.* 2024.

Stochastic Interpolants: Training and Sampling

Training Objective

Given samples $(z^\dagger_{j-1}, z^\dagger_j, r^\dagger_{j+1})$, minimize

$$\mathbb{E}_{t\sim U(0,1),\, \xi\sim\mathcal{N}(0,I)} \left| \hat{b}(I^\dagger_t, z^\dagger_j, z^\dagger_{j-1}, t) - \dot{\alpha}(t) z^\dagger_j - \dot{\beta}(t) r^\dagger_{j+1} - \dot{\sigma}(t)\sqrt{t}\xi \right|^2.$$

- Linear schedules for α, β, σ.
- Reparameterization ensures near-unit variance across t.

Sampling Procedure

Approximate SDE

$$d\hat{X}_t = \hat{b}(\hat{X}_t, z_0, z_{-1}, t)\, dt + \sigma(t)\, dW_t, \quad \hat{X}_0 = z_0.$$

- Discretize with first-order stochastic Runge–Kutta scheme.
- Terminal state $\mathrm{Law}(\hat{X}_1) \approx \mathrm{Law}(r_1 \mid z_0, z_{-1})$.

Diffusion Models: Formulation

Forward Noising Process

Define process

$$dX_t^F = \sqrt{2\sigma(t)\dot{\sigma}(t)}\, dW_t, \qquad X_0^F = r_1.$$

for σ non-negative and increasing. States z_0, z_{-1} remain fixed.

Reverse-Time SDE

Conditional law recovered from

$$dX_t = -2\sigma(t)\dot{\sigma}(t)\nabla_x \log p(X_t, z_0, z_{-1} t)\, dt + \sqrt{2\sigma(t)\dot{\sigma}(t)}\, d\bar{W}_t.$$

$\nabla_x \log p$ is score function of joint density.

- Reverse diffusion approximately transports Gaussian to $p(r_1 \mid z_0, z_{-1})$.
- Objective function for score given by Tweedie's formula

$$\int_0^T \mathbb{E}\big[|\nabla \log p(X_t^{\mathrm{F}}, z_0, z_{-1}, t) - \sigma(t)^{-2}(X_t^{\mathrm{F}} - r_1)|^2\big]\, dt.$$

Diffusion Models: Training and Sampling

Score Matching Objective

Given samples $(z_{j-1}^\dagger, z_j^\dagger, r_{j+1}^\dagger)$, minimize

$$\mathbb{E}_{t,\xi} \left| \hat{s}(r_{j+1}^\dagger + \sigma(t)\xi, z_j^\dagger, z_{j-1}^\dagger, t) - \sigma(t)^{-1}\xi \right|^2 .$$

- Log-normal sampling of noise level σ.
- Reparameterization ensures near-unit variance across t.

Sampling Procedure

Approximate backward SDE

$$d\hat{X}_t = -2\sigma(t)\dot{\sigma}(t)\hat{s}(\hat{X}_t, z_0, z_{-1}t)\,dt + \sqrt{2\sigma(t)\dot{\sigma}(t)}\,d\bar{W}_t.$$

- Second order predictor-corrector method based on Heun.
- Non-uniform time grid with step sizes proportional to σ.

Karras et.al. "Elucidating the Design Space of Diffusion-Based Generative Models." NeuIPS. 2022.

CRPS-Based Generator: Formulation

Direct Conditional Transport Map

Learn deterministic transport map

$$f_\theta : (\xi, z_0, z_{-1}) \mapsto r_1, \qquad \xi \sim \mathcal{N}(0, I_p).$$

such that

$$f_\theta(\cdot, z_0, z_{-1})_\sharp N(0, I_p) \approx p(r_1 \mid z_0, z_{-1}).$$

CRPS Objective

Minimize energy-score

$$\mathbb{E}\left[|f_\theta(\xi, z_0, z_{-1}) - r_1| - \frac{1}{2} |f_\theta(\xi, z_0, z_{-1}) - f_\theta(\xi', z_0, z_{-1})| \right].$$

- First term: forecast accuracy.
- Second term: ensemble diversity.
- Equivalent to MMD with p.s.d. kernel $k(x, y) = -|x - y|$.

Lang et.al. "AIFS – ECMWF's data-driven forecasting system." arXiv:2406.01465. 2024.

CRPS Generator: Training & Spectral Regularization

Training Objective

Given samples $(z^\dagger_{j-1}, z^\dagger_j, r^\dagger_{j+1})$, minimize

$$|f_j(\xi) - r^\dagger_{j+1}| + |f_j(\xi') - r^\dagger_{j+1}| - |f_j(\xi) - f_j(\xi')|$$

where $f_j(\xi) = f_\theta(\xi, z^\dagger_j, z^\dagger_{j-1})$ with 2 ensemble members $\xi, \xi' \sim N(0, I_p)$.

CRPS alone may:

- Under-represent high-frequency modes.
- Be unstable during training due to low ensemble size.

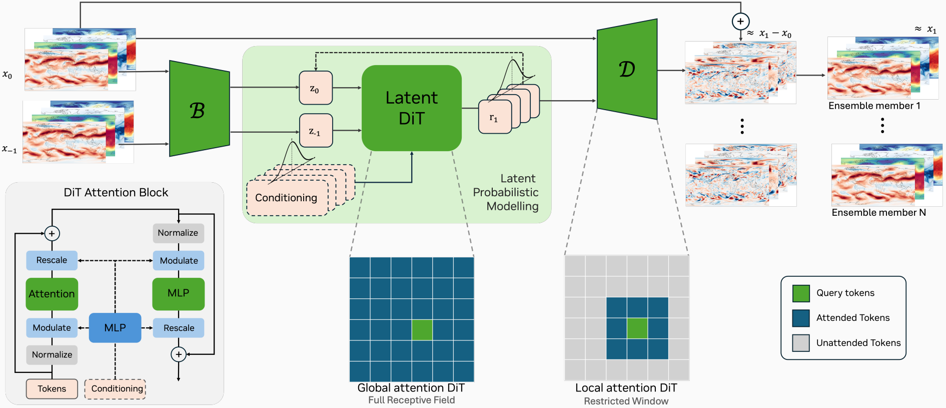
Add spectral term, for $\hat{S}$ spherical harmonic transform,

$$|\hat{S}(f_j(\xi)) - \hat{S}(r^\dagger_{j+1})| + |\hat{S}(f_j(\xi')) - \hat{S}(r^\dagger_{j+1})| - |\hat{S}(f_j(\xi)) - \hat{S}(f_j(\xi'))|$$

Sampling Procedure

Single network evaluation orders of magnitude faster than discrete SDE.

Overview of Methodology



Training, Sampling & Metrics

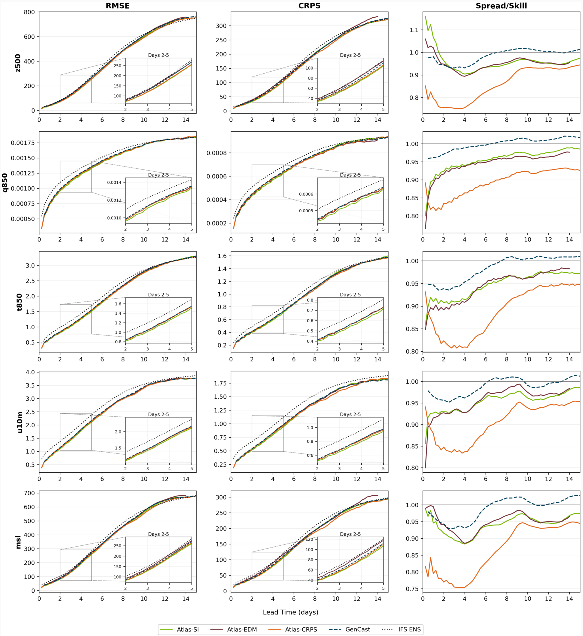
Training

- ERA5 fields Gaussian normalized per-variable (separate states and residuals).
- Additional conditioning fields: cosine-zenith angle, land-sea mask, surface geopotential.
- White noise sampled on spherical harmonic basis (KL expansion).
- Trained with 32 H100 GPUs with hard-restart damped cosine decay.
- 2×3 patch size and embedding dim $e = 3328$ (2.5B parameters).

Sampling and Metrics

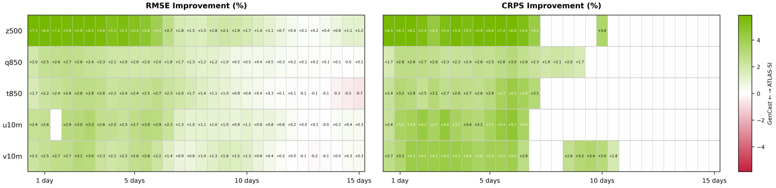
- SDEs discretized with 100 time steps (44s inference time).
- Evaluation based on 56 ensemble members with metrics: RMSE of mean, CRPS, skill-spread ratio.
- Using per date statistics, paired t-test used to indicate statistical significance.

Fifteen Day Rollout Results

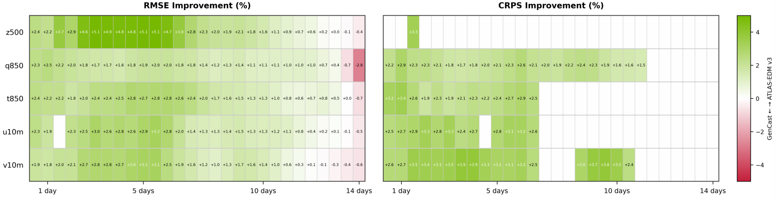


Score card vs IFS

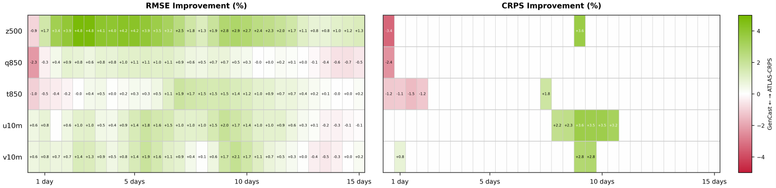
Score card vs GenCast



White cells: not significant (p ≥ 0.05)

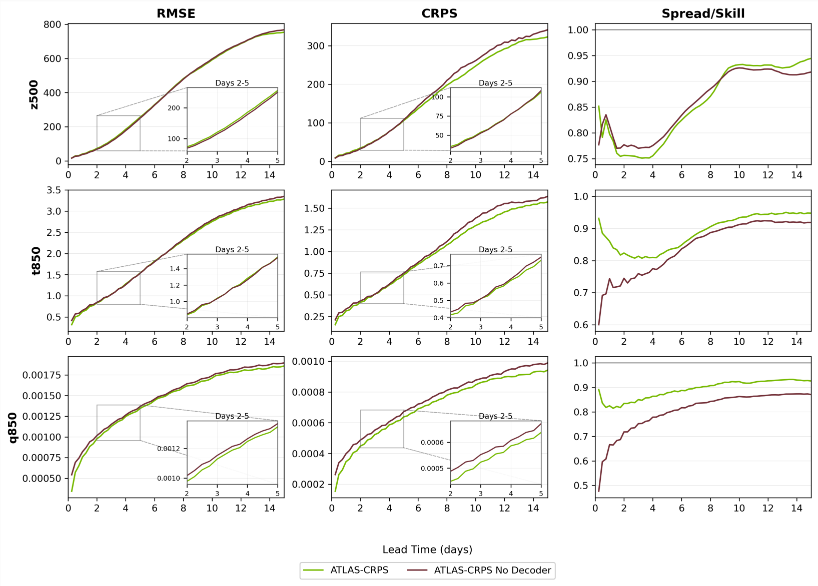


White cells: not significant (p ≥ 0.05)

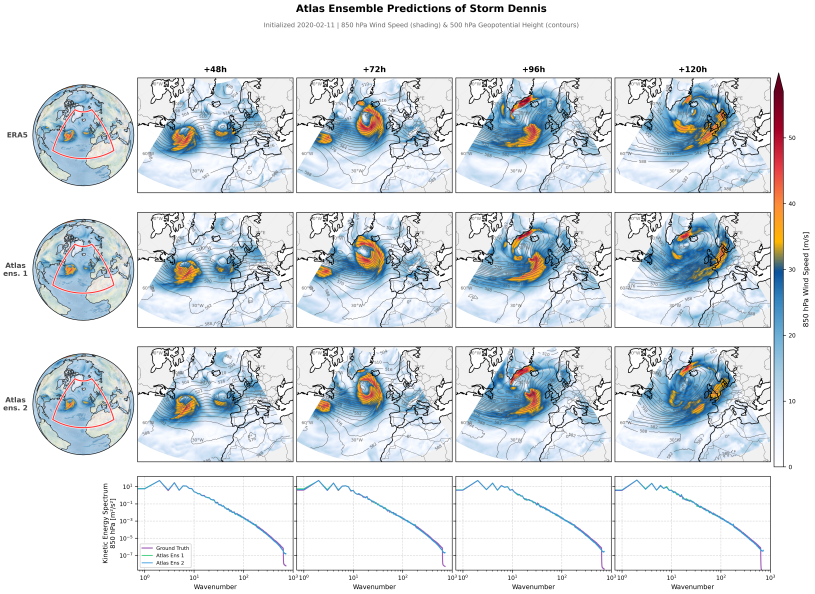


White cells: not significant (p ≥ 0.05)

Role of the Decoder



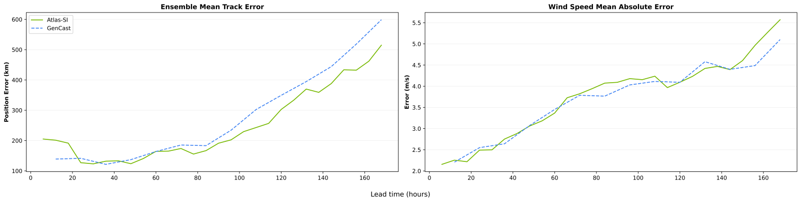
Spectral Analysis: Storm Dennis

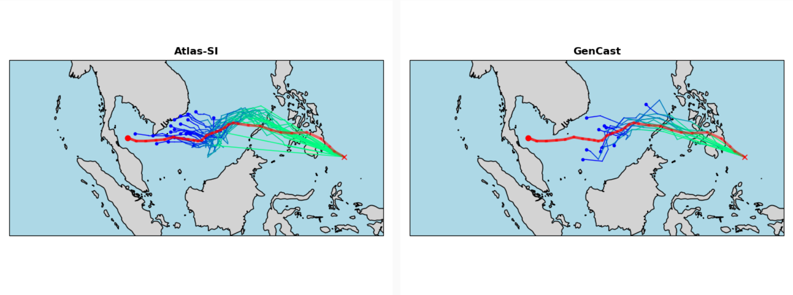


Atlas Ensemble Predictions of Storm Dennis

Initialized 2020-02-11 | 850 hPa Wind Speed (shading) & 500 hPa Geopotential Height (contours)

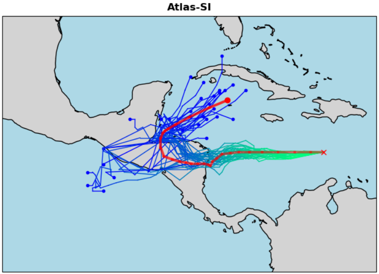
Cyclone Tracking



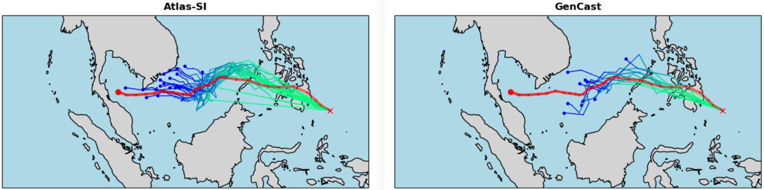


Tropical Storm Krovanh 7 day track, initialized December 17, 2020 UTC 12:00.

Cyclone Tracking Examples



Hurricane Eta 7 day track, initialized October 31, 2020 UTC 18:00.



Tropical Storm Krovanh 7 day track, initialized December 17, 2020 UTC 12:00.

Conclusion

Key Result

Latent transformer framework achieves state-of-the-art probabilistic forecasts.

Technical Insight

Architecture robustness across stochastic interpolants, diffusion, and CRPS training.

Open Questions

Physics constraints, conservation laws, training on ensemble data, long-term stability.