Swift: An Autoregressive Consistency Model for Efficient Weather Forecasting

Jason Stock[1], Troy Arcomano[1,2], Rao Kotamarthi[1]

[1] Argonne National Laboratory
[2] Allen Institute for AI

Motivation

Problem: diffusion models are effective for probabilistic weather forecasting, but their slow iterative solvers during inference make them impractical for S2S applications, where long-lead times and domain calibration is essential

Limitations with prior work:

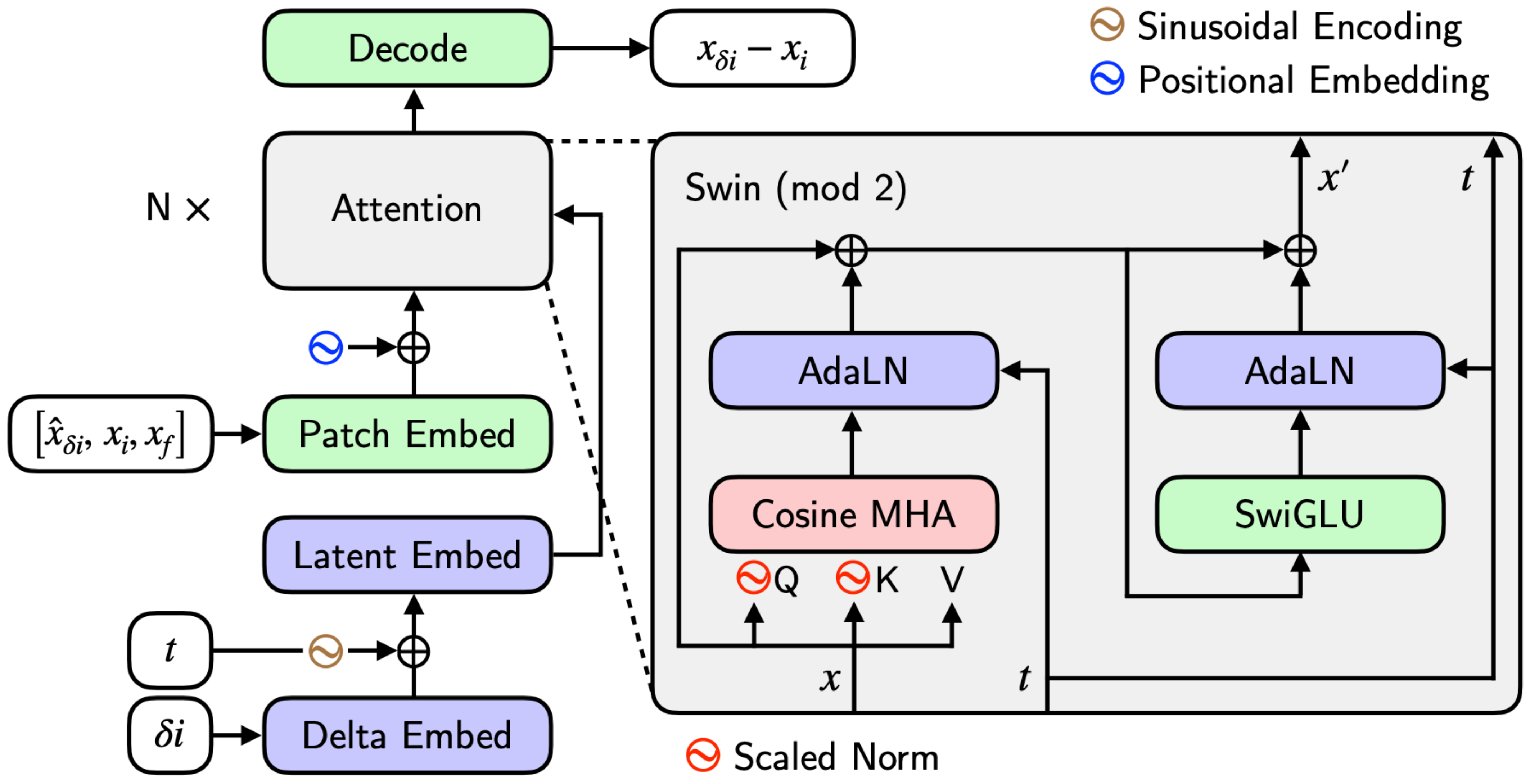
- GenCast takes 39 NFEs and ArchesWeatherGen takes 25 NFEs per step — multi-step finetuning is way too costly
- Lower the temporal resolution form 6h to 12 or 24h so there are simply make fewer function evals
- Do without diffusion and train multiple models (FGN) with sensitive perturbations

What do we want? The speed of a single deterministic model + probabilisitic formulation + physical consistency

Methodology

Task: model the global evolution of the atmosphere, learning $p\left(x_{i+1} \mid x_i\right)$ by estimating the temporal residual $x_{\delta i} - x_i : \delta i \sim U\{6,12,24\}$ with a neural net (helps to regularize training and is more flexible)

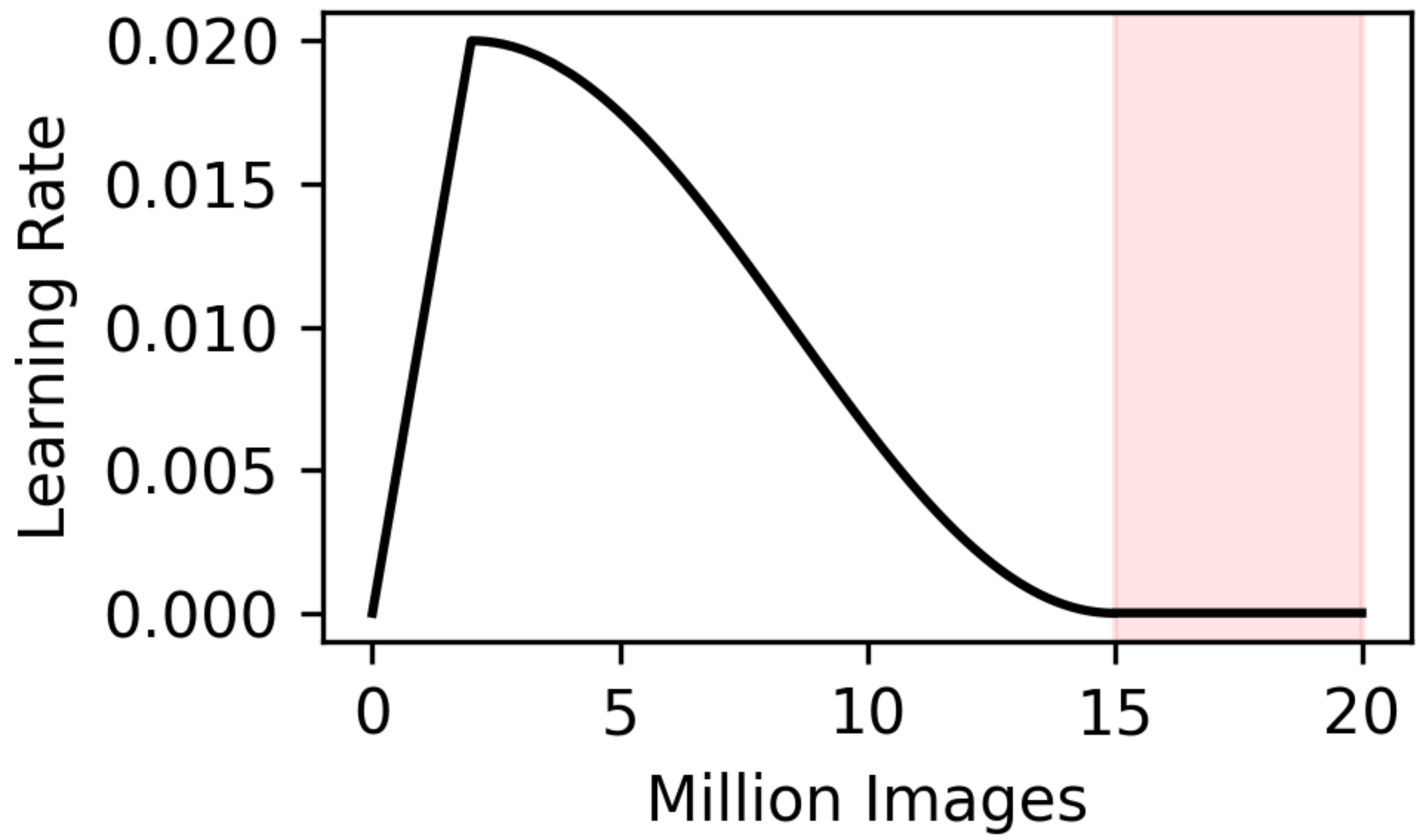
ERA5 Dataset: 6-hourly 1.4° WB2 data with 4 surface, 5 atmospheric (at 13 vertical levels), and 3 forcing variables; train: 1979–2018, val: 2019, test: 2020 (compute budget constraints, but working on full-resolution 0.25° ERA5)

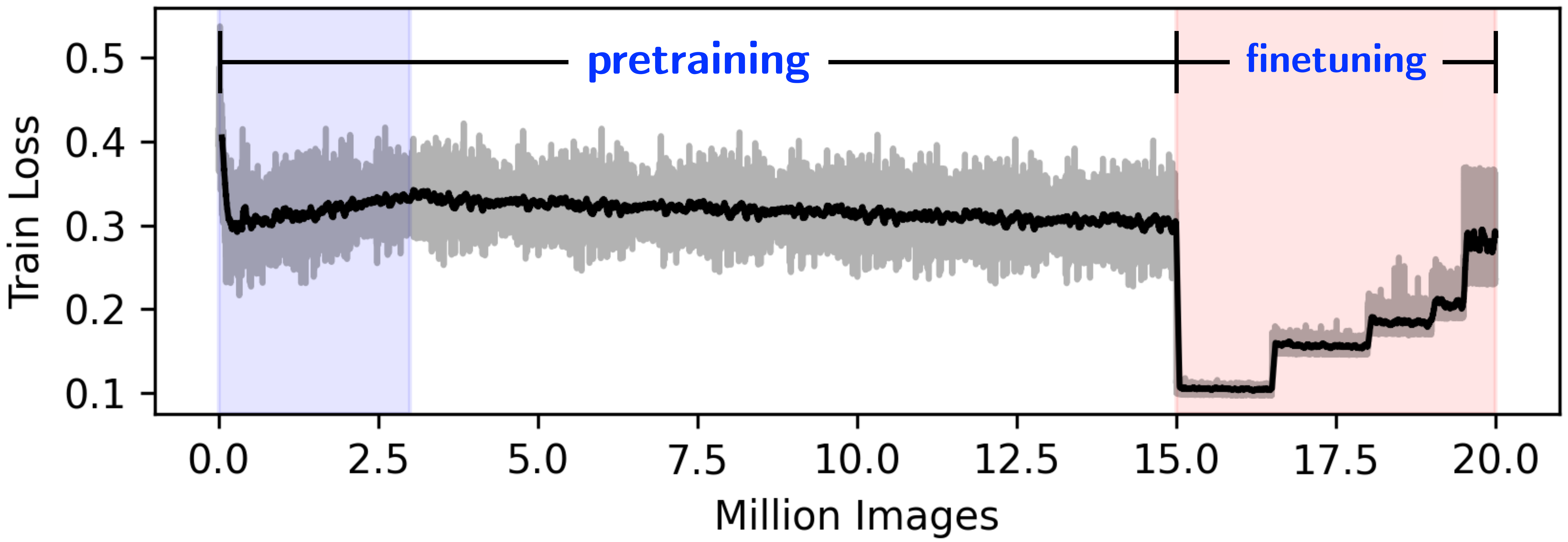


Type	Variable	Description
surface	`t2m`	2 meter temperature
surface	`10u`	10 meter u-wind component
surface	`10v`	10 meter v-wind component
surface	`mslp`	Mean sea level pressure
atmos.	`z---`	Geopotential
atmos.	`t---`	Temperature
atmos.	`u---`	u-wind component
atmos.	`v---`	v-wind component
atmos.	`q---`	Specific humidity
static	`z`	Geopotential at surface
static	`lsm`	Land-sea mask
clock	`n/a`	TOA incident solar radiation

Two-Stage Probability Flow Training

Novelty: first consistency and finetuned probability flow model for global weather and subseasonal-to-seasonal (S2S) forecasting!





① Model Pretraining

Latitude $\kappa(s)$ and pressure $\alpha(v)$ weighted v-prediction loss

$$L(\theta) = \frac{1}{|S|} \sum_{s\in S} \sum_{v\in V} \kappa(v)\,\alpha(s)\,\ell_{v,s}(\theta)$$

where $\ell(\theta)$ is the objective for the baseline diffusion and base consistency models

Diffusion (TrigFlow): $\mathbb{E}_{\mathbf{x}_0,\mathbf{z},t}\left[\left\| \sigma_d \mathbf{F}_\theta\left(\frac{\mathbf{x}_t}{\sigma_d}, t\right) - \left(\cos(t)\,\mathbf{z} - \sin(t)\,\mathbf{x}_0\right) \right\|_2^2\right]$

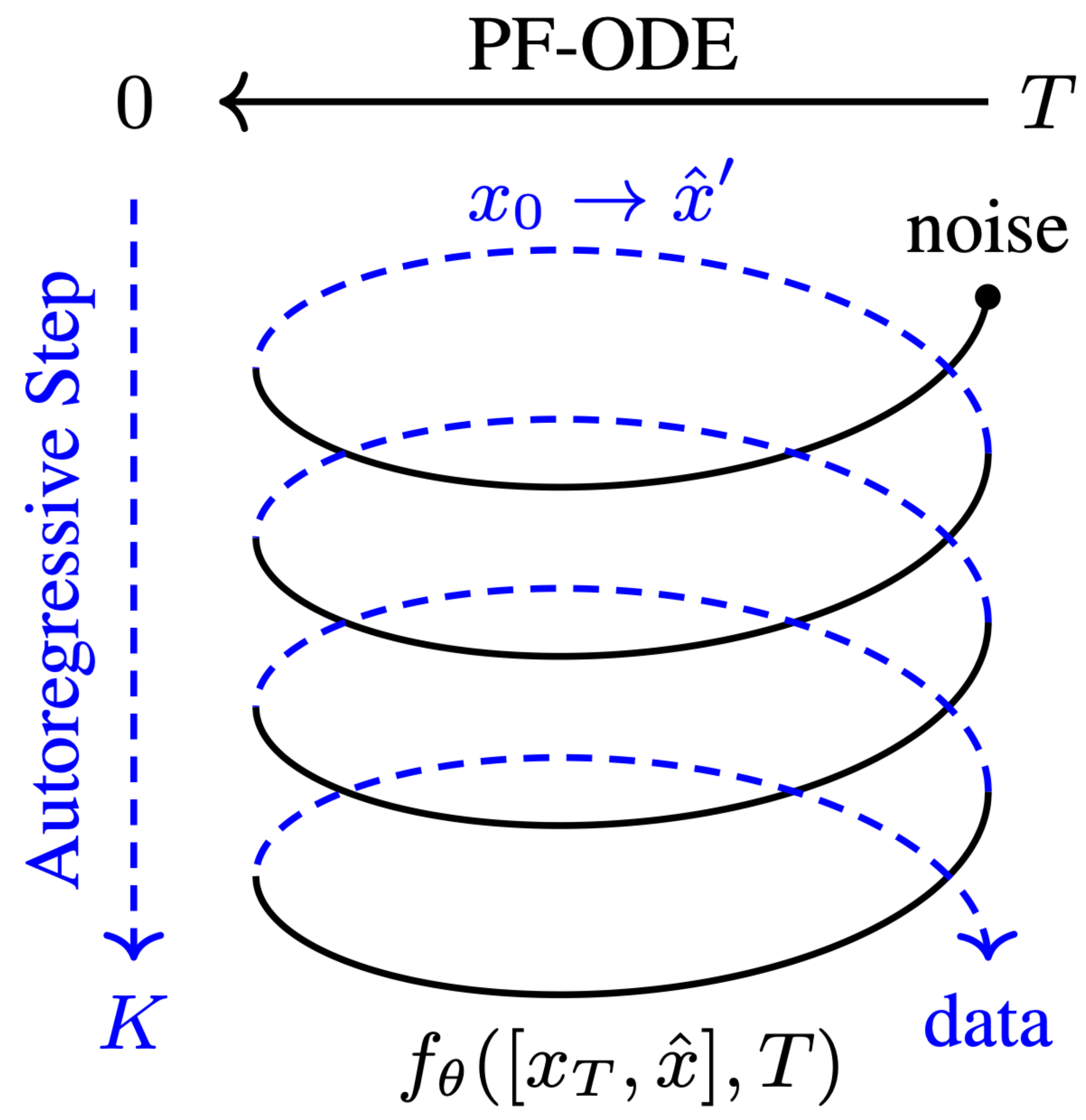
Consistency (sCM): $\mathbb{E}_{\mathbf{x}_t,t}\left[\left\| \mathbf{F}_\theta\left(\frac{\mathbf{x}_t}{\sigma_d}, t\right) - \mathbf{F}_{\theta^-}\left(\frac{\mathbf{x}_t}{\sigma_d}, t\right) - \cos(t)\frac{\mathrm{d}\mathbf{f}_{\theta^-}(\mathbf{x}_t, t)}{\mathrm{d}t} \right\|_2^2\right]$

② Multi-Step Finetuning

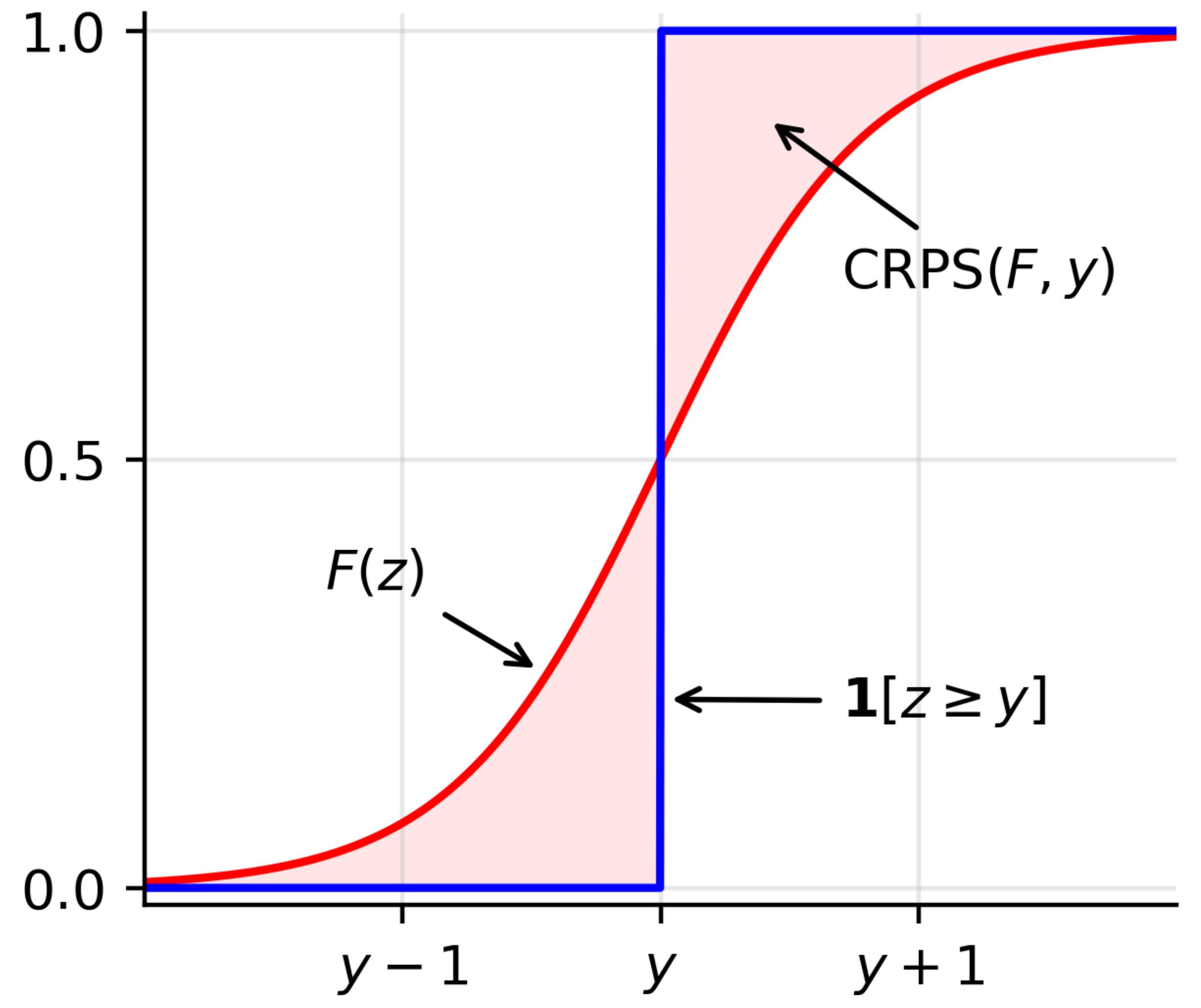
With the **single step consistency model**, we take $K = 1$–8 autoregressive steps and $N = 2$ ensembles, backpropagating through time on last step loss

$$L_{\text{CRPS}} = \frac{1}{|S|} \sum_{s \in S} \sum_{v \in V} \kappa(v)\, \alpha(s)\, \text{CRPS}\left(\hat{\mathbf{y}}^{1:2}_{v,s}, \mathbf{y}_{v,s}\right)$$

Continuous Ranked Probability Score: $\frac{1}{N}\sum_{n} |\hat{y}^n - y| - \frac{1}{2N(N-1)} \sum_{n \neq n'} |\hat{y}^n - y^{n'}|$



$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbf{1}[z \geq y]\right)^2 \,\mathrm{d}z$$



Medium-Range Forecast Skill

Swift (1-step consistency sampler) is competitive to the IFS ENS and GenCast (20-step 2nd-order sampler), though finetuning shows a spread/skill tradeoffs in early lead times

Swift is 39× faster!

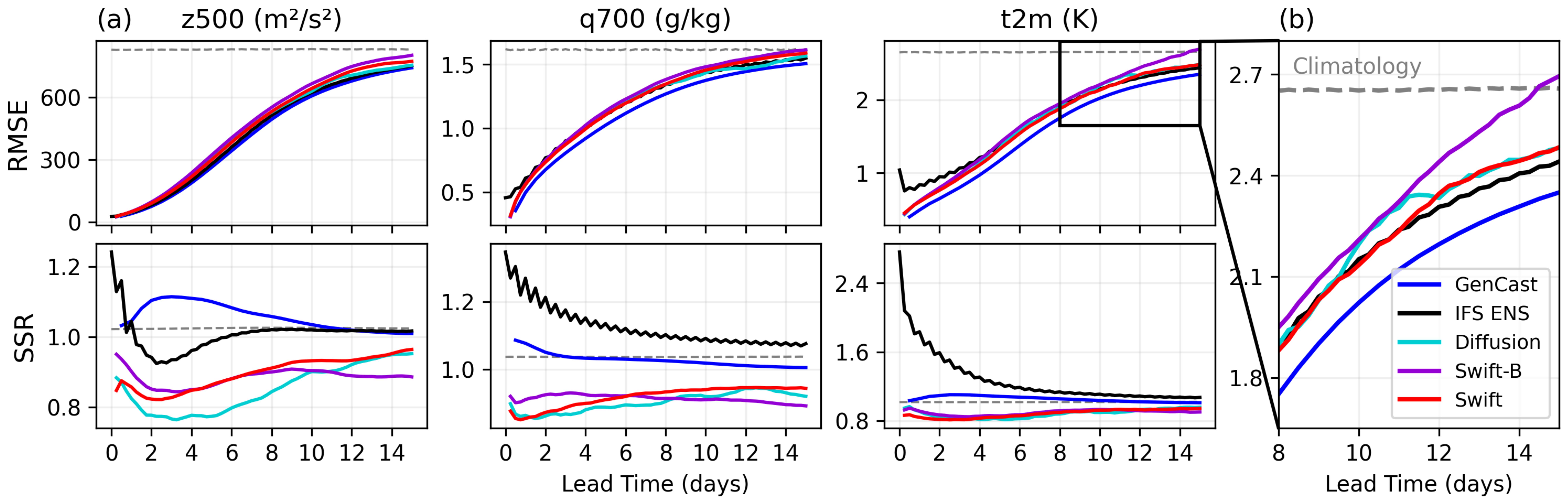


Figure 1: subset of latitude-weighted rmse and spread/skill metrics compared to baselines

Long-Term Stability

Swift reproduce realistic wave modes with coherent physical structure and realistic power spectra

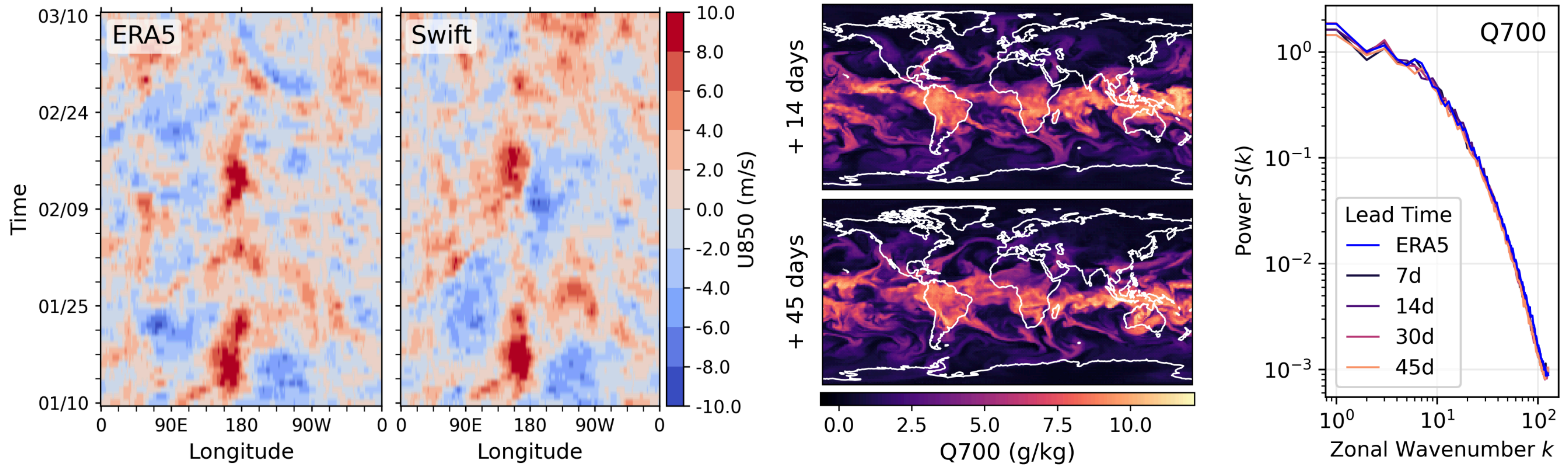
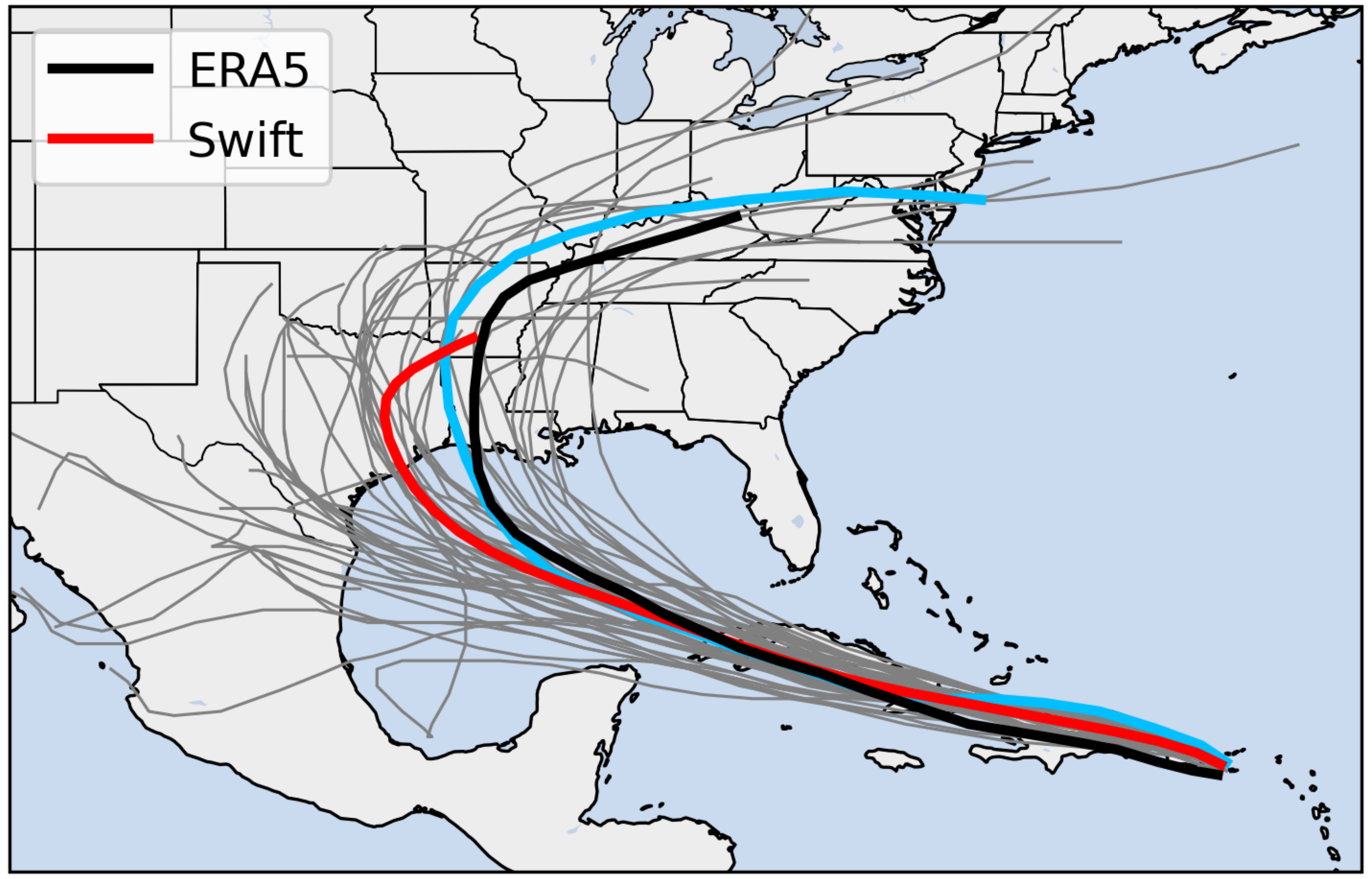


Figure 2: Hovmöller diagram, qualitative forecast fields, and Q700 power spectra

Case Studies

Swift correctly (a) captures the northern trend for Hurricane Laura at a lead time of 5 days and (b) reproduces the expected response to change in forcing conditions on seasonal scales



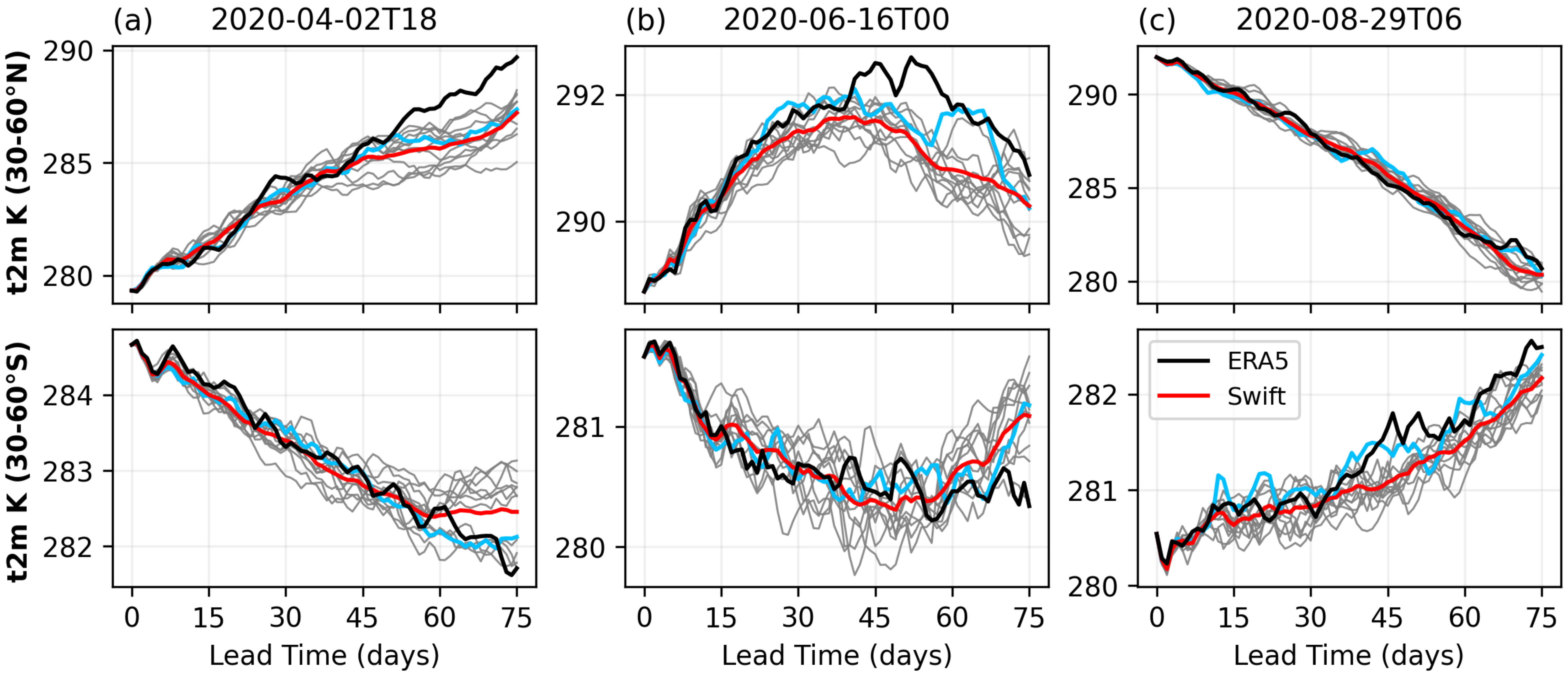
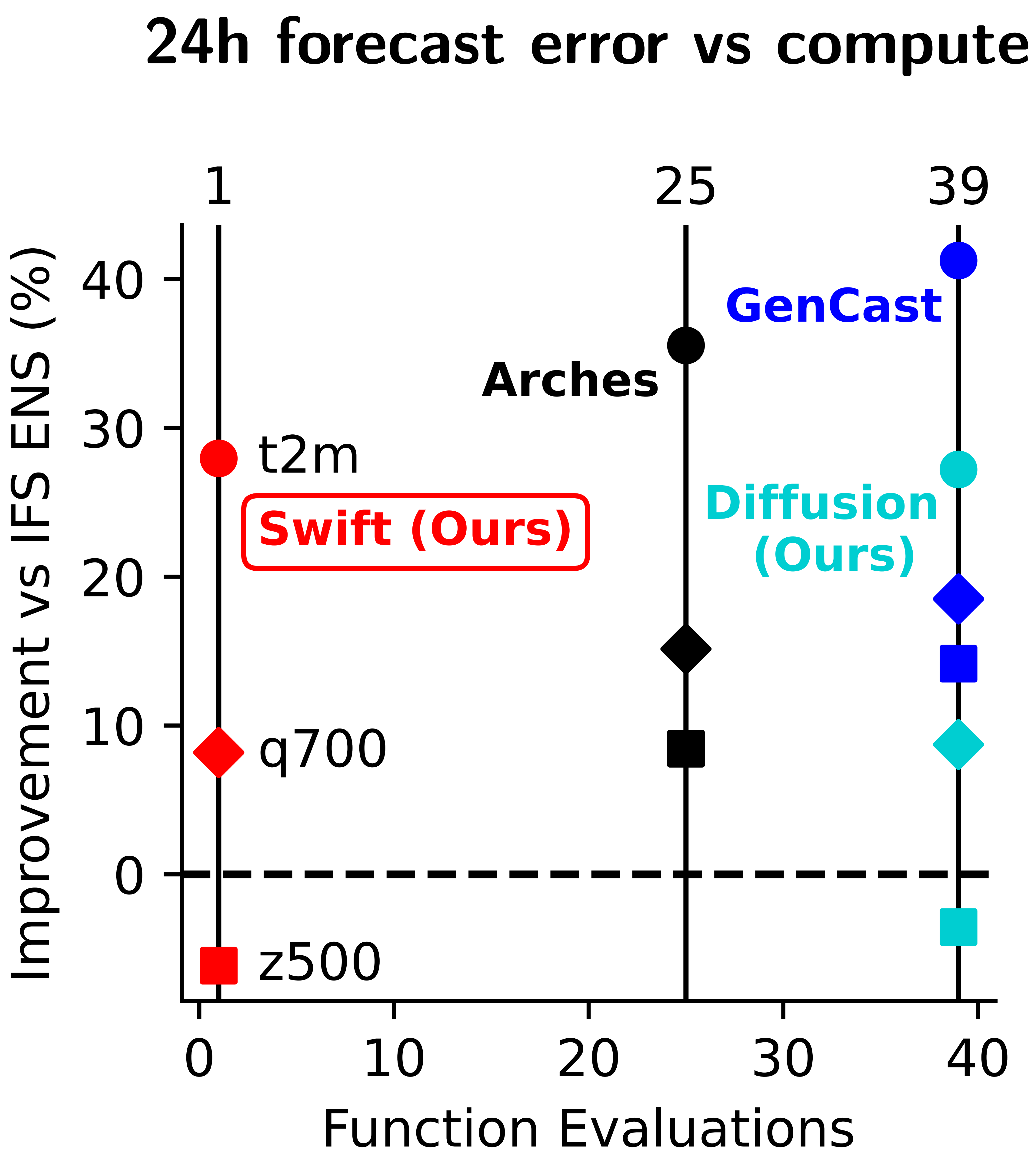


Figure 3: (left) Hurricane Laura tracks and (right) seasonal cycle forecasts



Conclusion

Takeaway: finetuning consistency models with CRPS deliver fast, few-step ensemble forecasts with comparable skill to diffusion at longer time scales

Challenges: consistency models are hard to train (instability and divergence) and remain expensive to train, only solving part of the compute problem

- **Model Distillation**: knowledge transfer from larger TrigFlow models to a smaller consistency model
- **Improved Performance**: Classifier Free Guidance + Architecture Improvement (GNN/FNO) + 0.25° Data