Swift: An Autoregressive Consistency Model for Efficient Weather Forecasting

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

Introduction

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

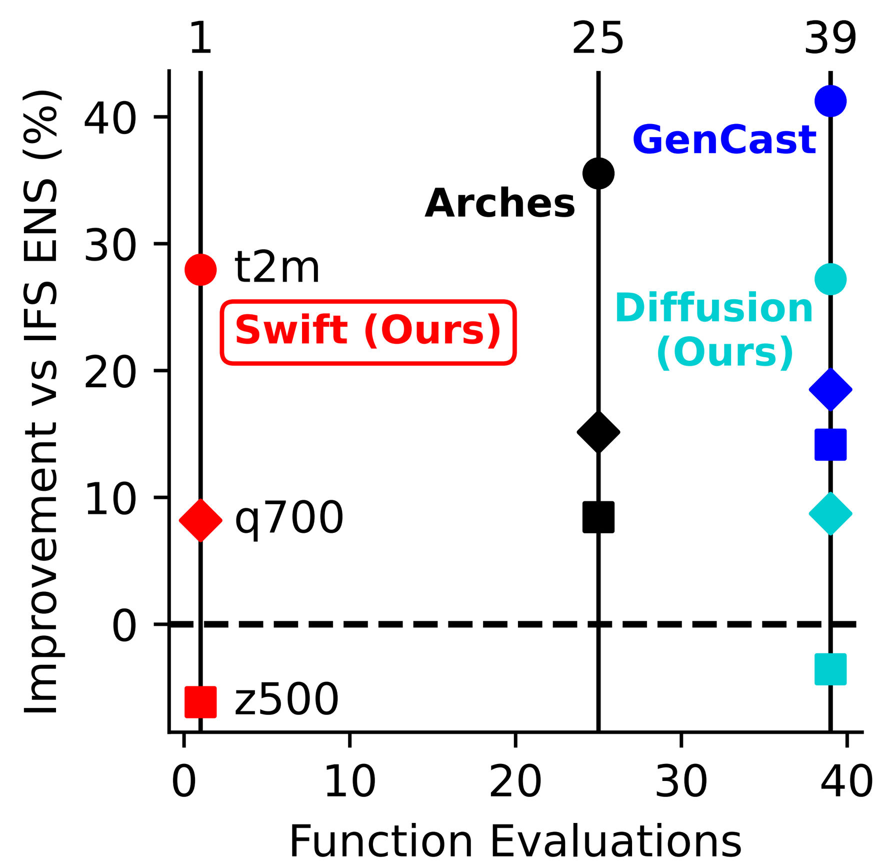


Figure 1: 24h forecast error vs compute

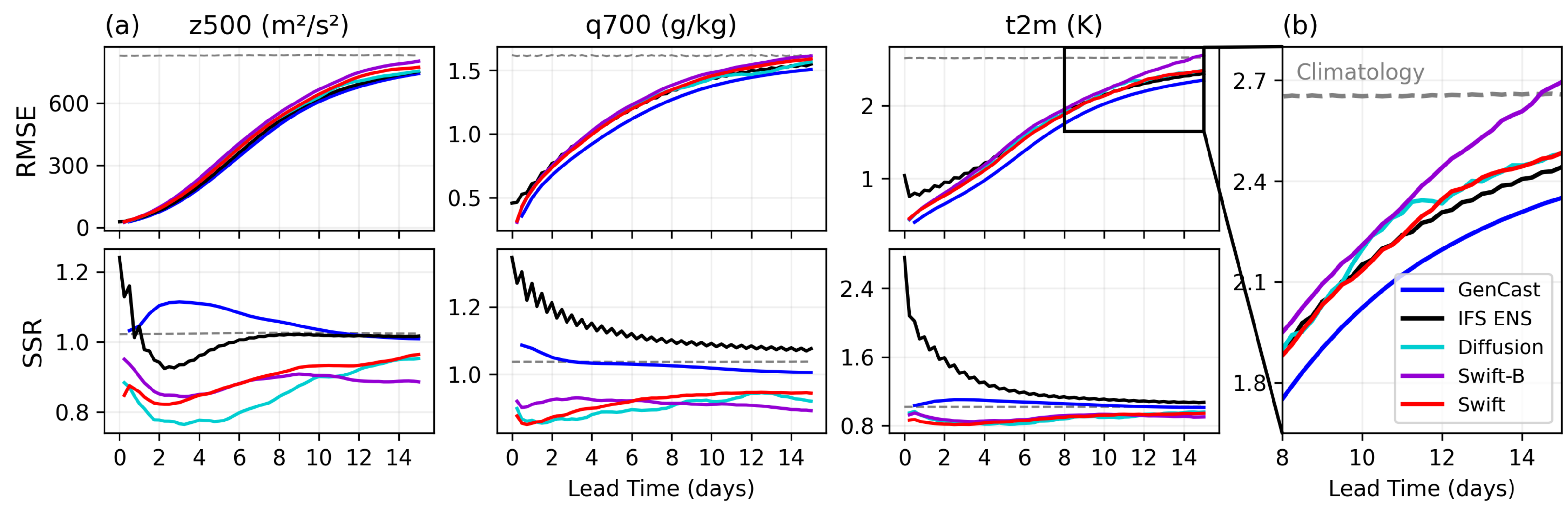


Figure 2: global forecast skill: latitude-weighted rmse and spread/skill compared to baselines

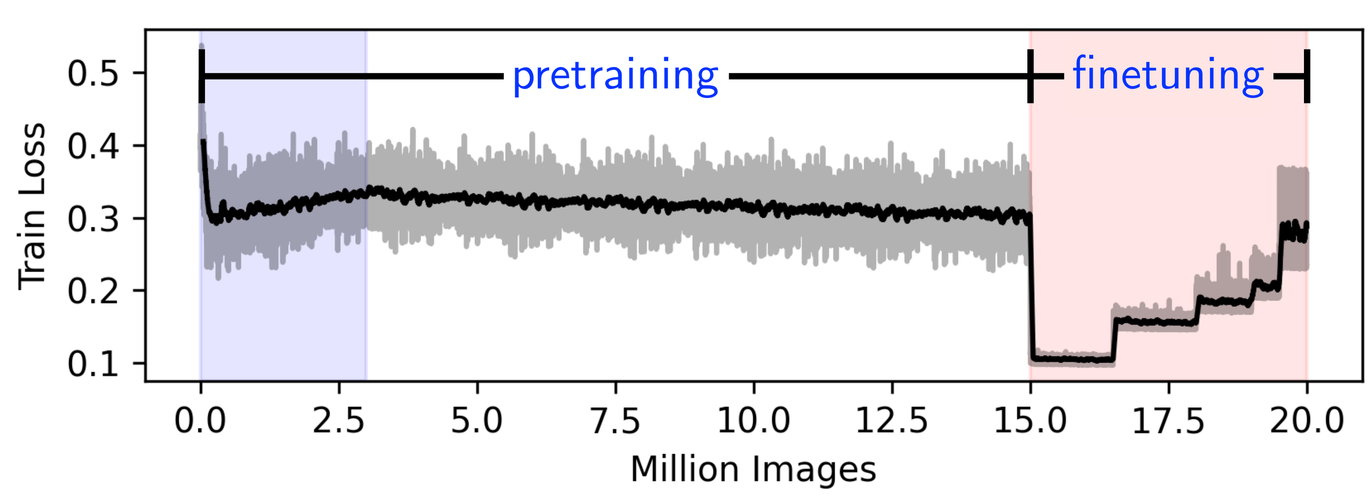
Methodology

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

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)

Architecture: 225M parameter conditional (initial/forcing + noise level and data time delta) non-hierarchical swin transformer with 2×2 local patch size

Two-Stage Probability Flow Training



① **Model Pretraining**: latitude $\kappa(s)$ and pressure $\alpha(v)$ weighted v-prediction loss for both the diffusion baseline and base consistency model

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

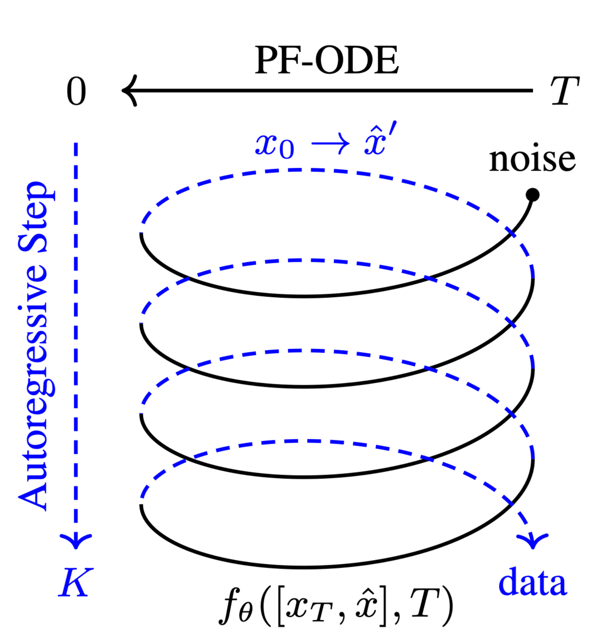
Diffusion (TrigFlow): $\mathbb{E}_{\mathbf{x}_0,\mathbf{z},t}\left[\left\| \sigma_d \mathbf{F}_\theta\left(\frac{\mathbf{x}_t}{\sigma_d}, t\right) - \left(\cos(t)\,\mathbf{z} - \sin(t)\,\mathbf{x}_0\right) \right\|_2^2 \right]$

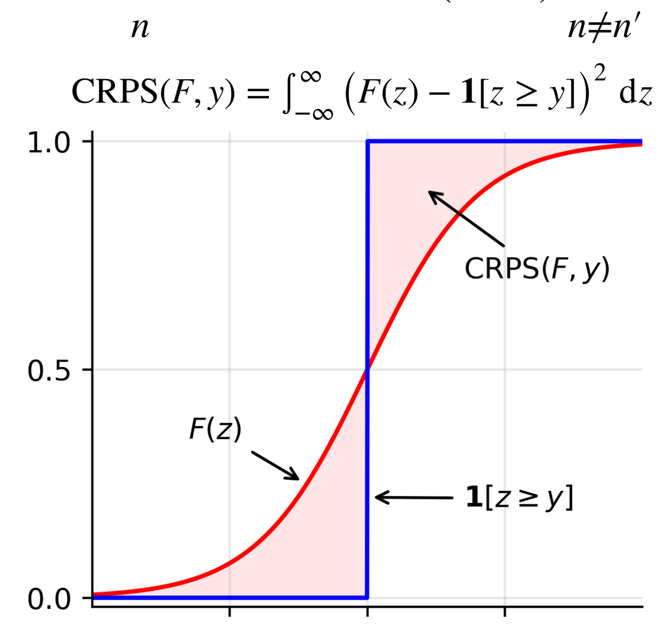
Consistency (sCM): $\mathbb{E}_{\mathbf{x}_t,t}\left[\left\| \mathbf{F}_\theta\left(\frac{\mathbf{x}_t}{\sigma_d}, t\right) - \mathbf{F}_{\theta^-}\left(\frac{\mathbf{x}_t}{\sigma_d}, t\right) - \cos(t)\frac{\mathrm{d}\mathbf{f}_{\theta^-}(\mathbf{x}_t, t)}{\mathrm{d}t} \right\|_2^2 \right]$

② **Multi-Step Finetuning**: single sampling step with $K = 1$–8 autoregressive steps and $N = 2$ ensembles, backpropagating through time on last step loss

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

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





* jason.stock@anl.gov

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

Experimental Results

Sampling Efficiency (Figure 1): 39× fewer forward passes for our consistency model (Swift) with comparable errors to diffusion baselines across variables

Forecast Skill (Figure 2): finetuned Swift (1-step sampler) is competitive to the IFS ENS and GenCast (20-step 2nd-order sampler), spread/skill tradeoffs

Long-Term Stability (Figure 3): consistency models reproduce realistic wave modes with coherent physical structure and realistic power spectra (not shown)

Case Studies (Figure 4): 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

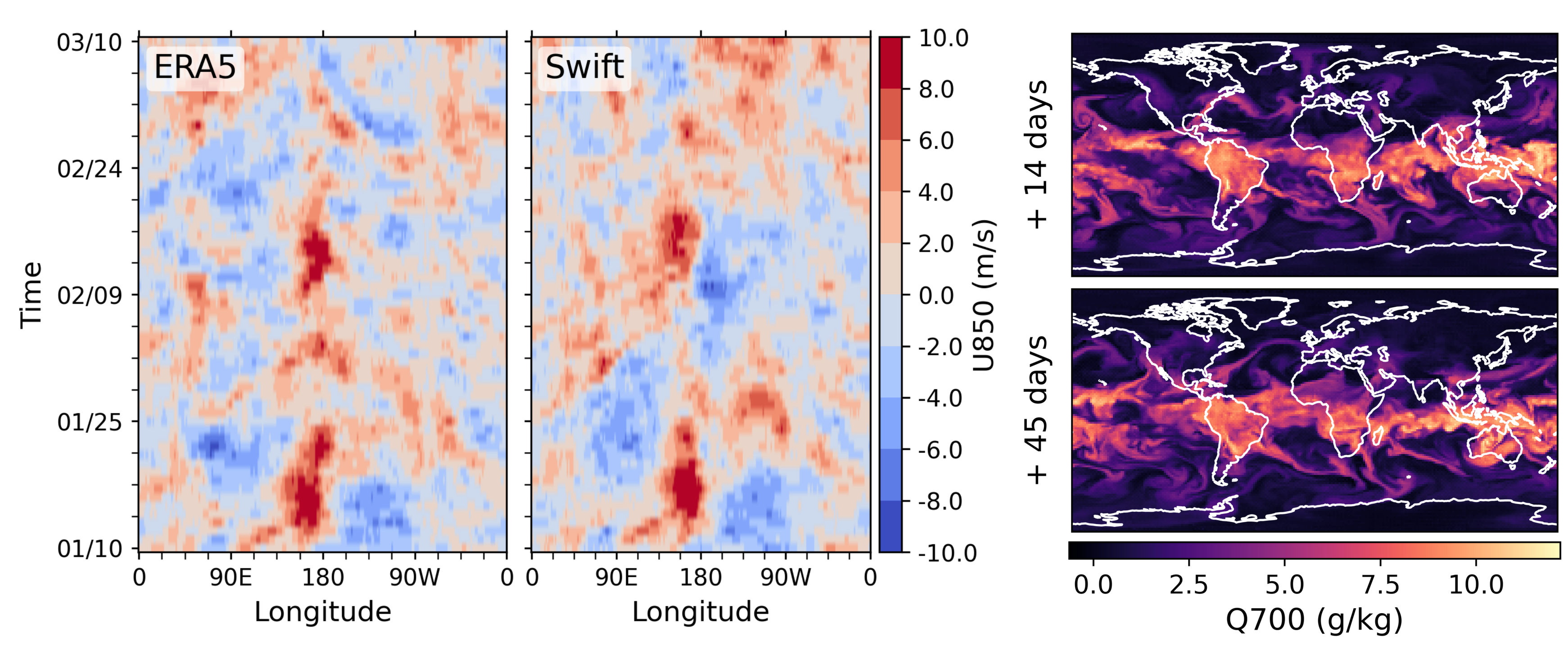


Figure 3: long-term stability: Hovmöller diagram and forecast fields

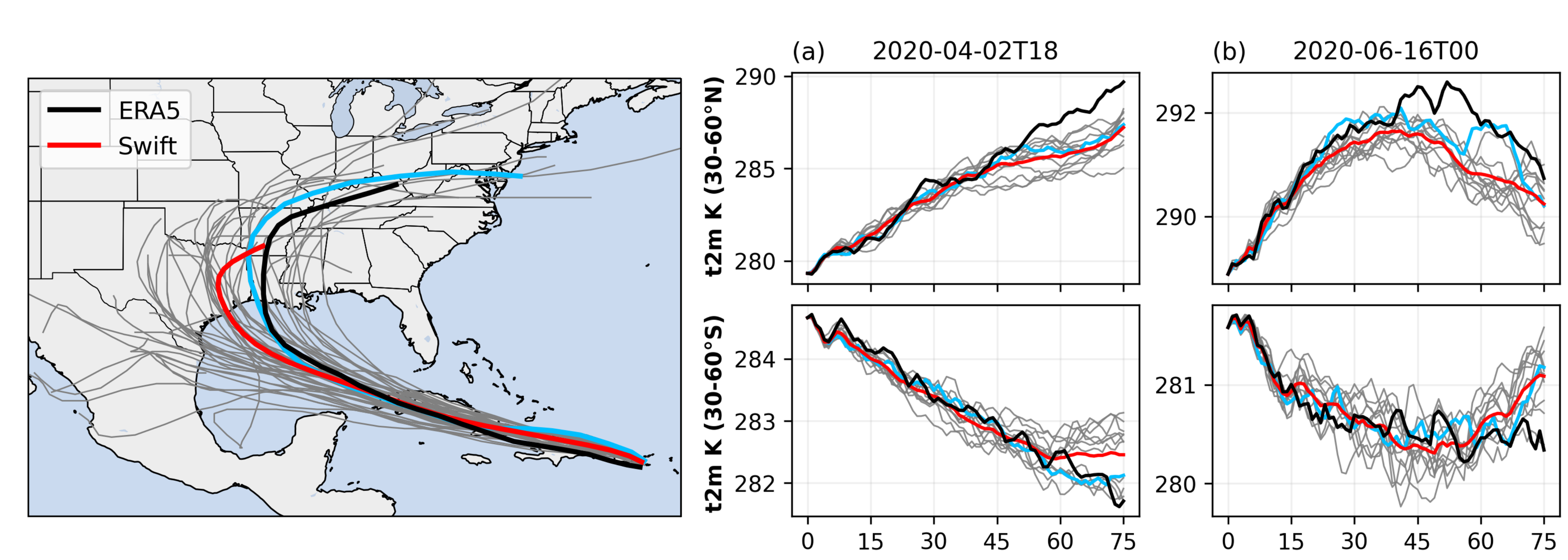


Figure 4: case studies: Hurricane Laura tracks and seasonal cycles

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
- **Classifier Free Guidance**: incurs 2× NFEs but could further improve skill