Long-Range Distillation: Distilling 10,000 Years of Simulated Climate into Long Timestep AI Weather Models

Martin et al., 2025

ML journal club

Friday, March 06

Quick Overview

- Attempts to achieve skill at S2S timescales by moving away from autoregressive steps—instead, train a probabilistic model to take one large step.

- Model achieves (slight) improvement over climatology & performance similar to ECMWF ensemble at 1 month lead time.

- By training on model output, authors find that model skill can scale with data volume—this is novel.

Background & Motivations

- subseasonal-to-seasonal (S2S) forecasting is difficult…
  - error accumulation in autoregressive rollouts + instability
  - skill of deterministic models saturates to climatology
- Probabilistic forecast can improve performance
  - either through ensemble or probabilistic models (diffusion)
  - modest skill gain compared to NWP ensembles
  - still autoregressive (still accumulate error)

"Fundamental **mismatch of autoregressive training objective** and the **inference task of long-range weather forecasting**"

limited sampling of slow climate variability

Background & Motivations

Train a long, single timestep probabilistic model

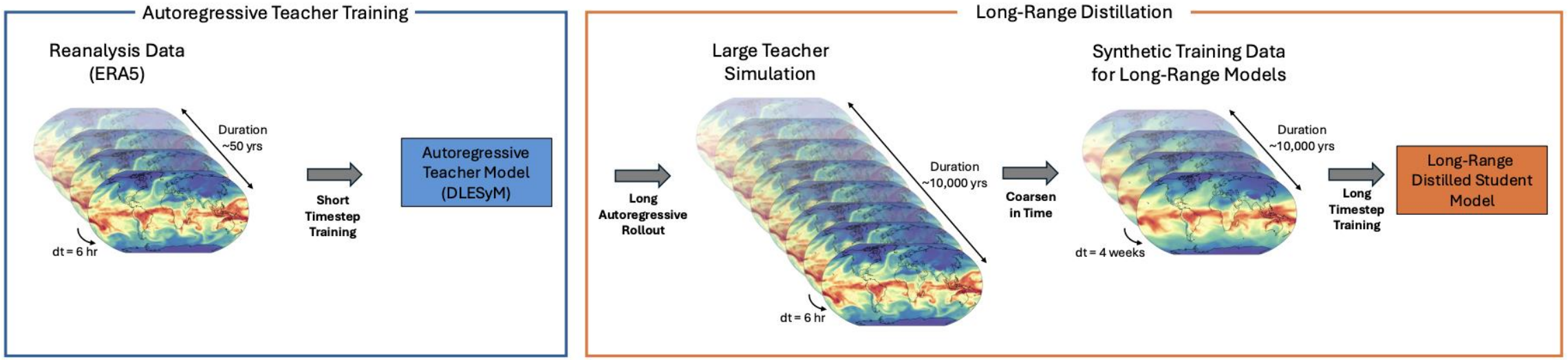
- prevents error accumulation and rollout instability
- easier to calibrate the predictions (control the spread of probabilistic forecasts)

BUT we don’t have enough ERA5 to train on…

Difficult to capture S2S variability without overfitting

Teacher-Student (Long-range Distillation Methodology)

- AR models can be good simulators of atmos variability
  - make synthetic training data for long-step model



- LLMs: performance scales with dataset and model capacity—unlock dataset scaling axis!

Teacher-Student (Long-range Distillation Methodology)

- Target weekly average forecast at 4-week lead time

$$\bar{x}_N = \frac{1}{M}\sum_{-M/2}^{M/2} x_{N+i}$$, N=112, M=28

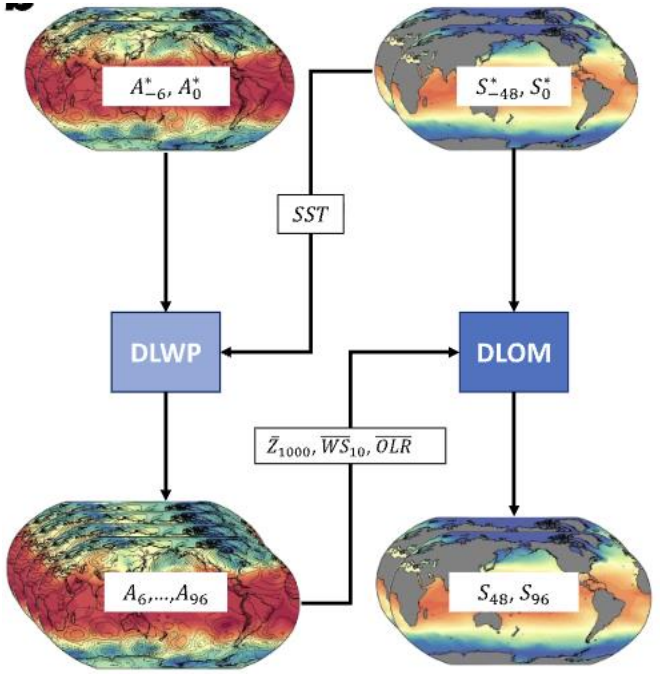
- Student model trained to give the conditional probability of the long-range time

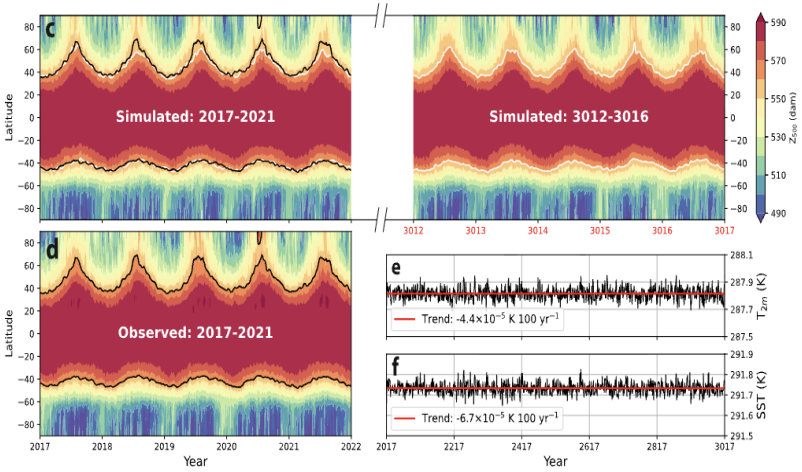
$$p(\bar{x}_N | x_1)^*$$

* **Model gets as input four daily averages.** Even though the DLESyM output is 6-hourly, the training simulations are saved as daily averages. The model gets 4 times as inputs to compensate for the loss of information **(?)**

Teacher: DLESyM

- Indefinitely stable!
- realistic variability (tropical cyclogenesis, midlat blocking events, Indian summer monsoon)
  - trained on reanalysis—should mimic obs. real dynamics





Limitations

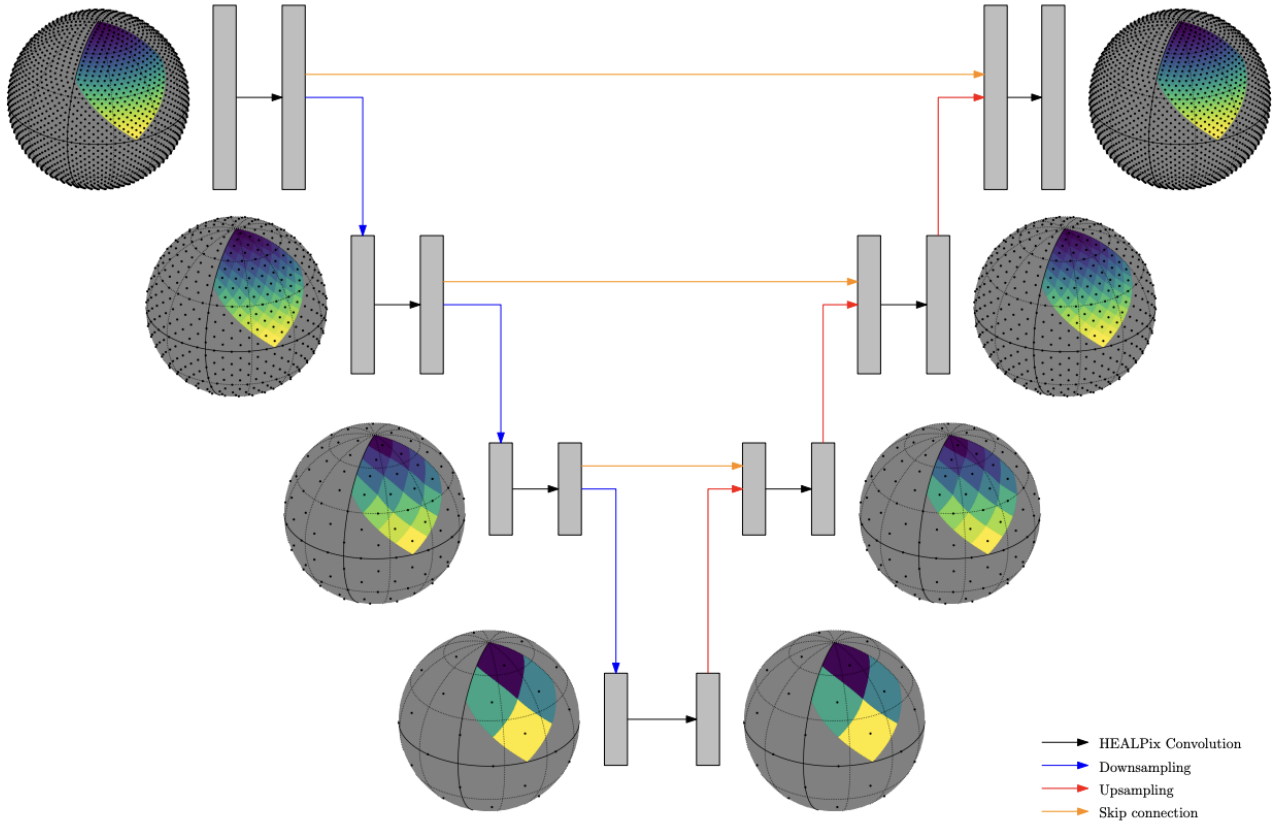
- struggles with ENSO (too weak)
- forecasts few variables
- suppressed error growth for IC pert.

Training simulation:

- 200 simulations of 90 years (18,000 yrs of data)
- some went unstable though… leaves us with 15,000 years of data

Student: DLESyM10K

- Conditional diffusion model
  - add noise, learn to reverse it
  - “conditional” just means it takes in conditioning (in this case the 4 daily averages)
  - model is trained on a range of noise levels (σ) and takes 18 denoising steps to get from the σ_{max} to σ_{min}



Brenowitz et al., 2025

Student: Classifier-Free Guidance

- Need a means to control the ensemble spread of the forecasts
  - typically done by varying the IC pert.
- For the diffusion model this is done via CFG
  - **conditional generation**: model learns to adjust the state x to something consistent with the conditioning (ultimately there is a “right” answer)
  - **unconditional generation**: model just picks a sample form the learned data distribution

$$\nabla_x \log\left(p(x(t)) + w\left[\nabla_x \log(p(x(t)|c)) - \nabla_x \log(p(x(t))\right],\right.$$

w = 1: model ONLY uses conditional score, adhering strongly to conditioning

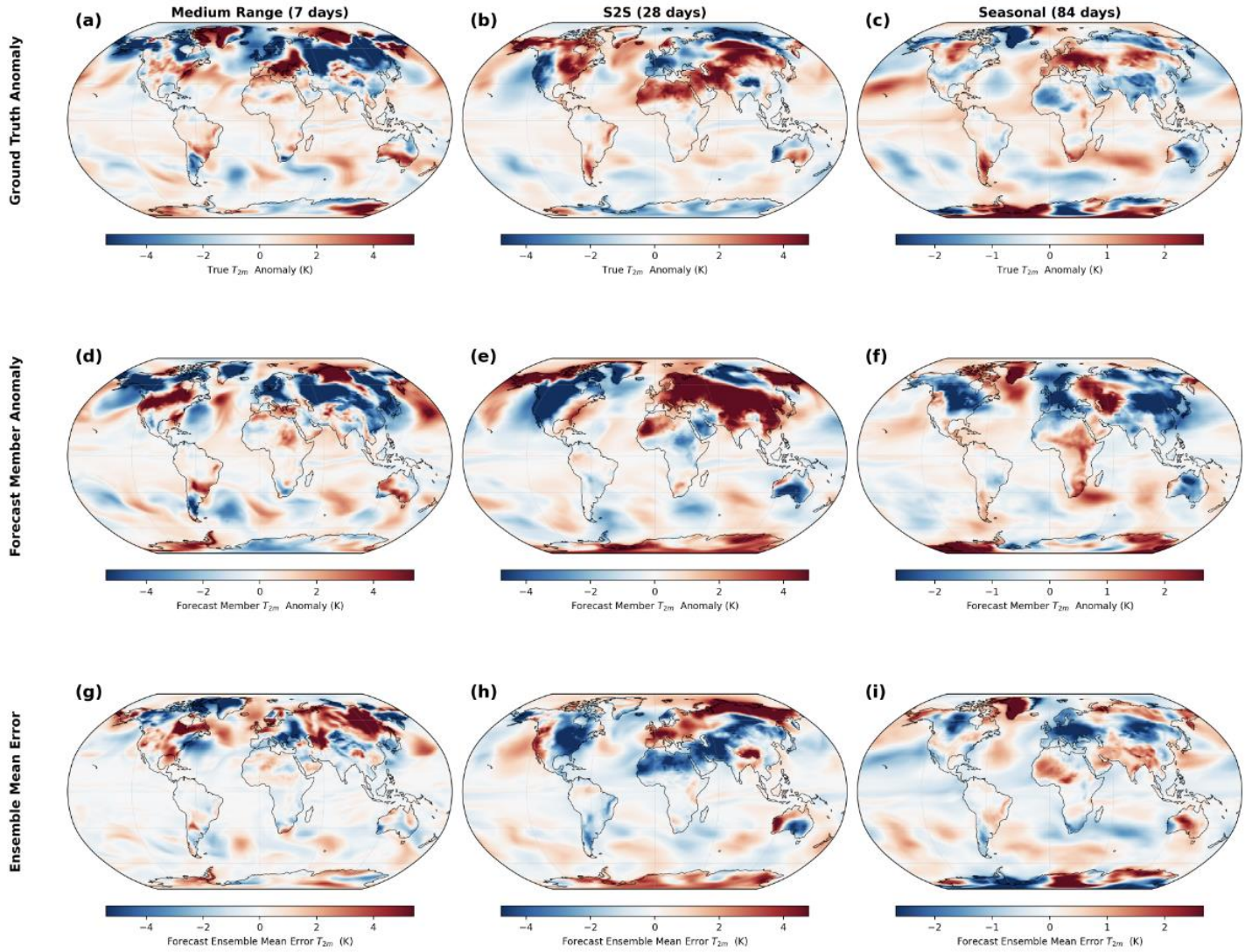
w = 0: model generates samples unconditionally

The balance helps with managing spread (unconditional sampling) while still getting close to the expected output given the conditioning (conditional sampling)

Perfect Model Experiment

- How good are the models (teacher and student) at forecasting an unseen DLESyM simulation?
- “Nature run” ground truth from DLESyM
- select starting point dates and add noise (imperfect ICs)
- run teacher and student ensemble forecasts and eval against NR (medium range, S2S, and seasonal)

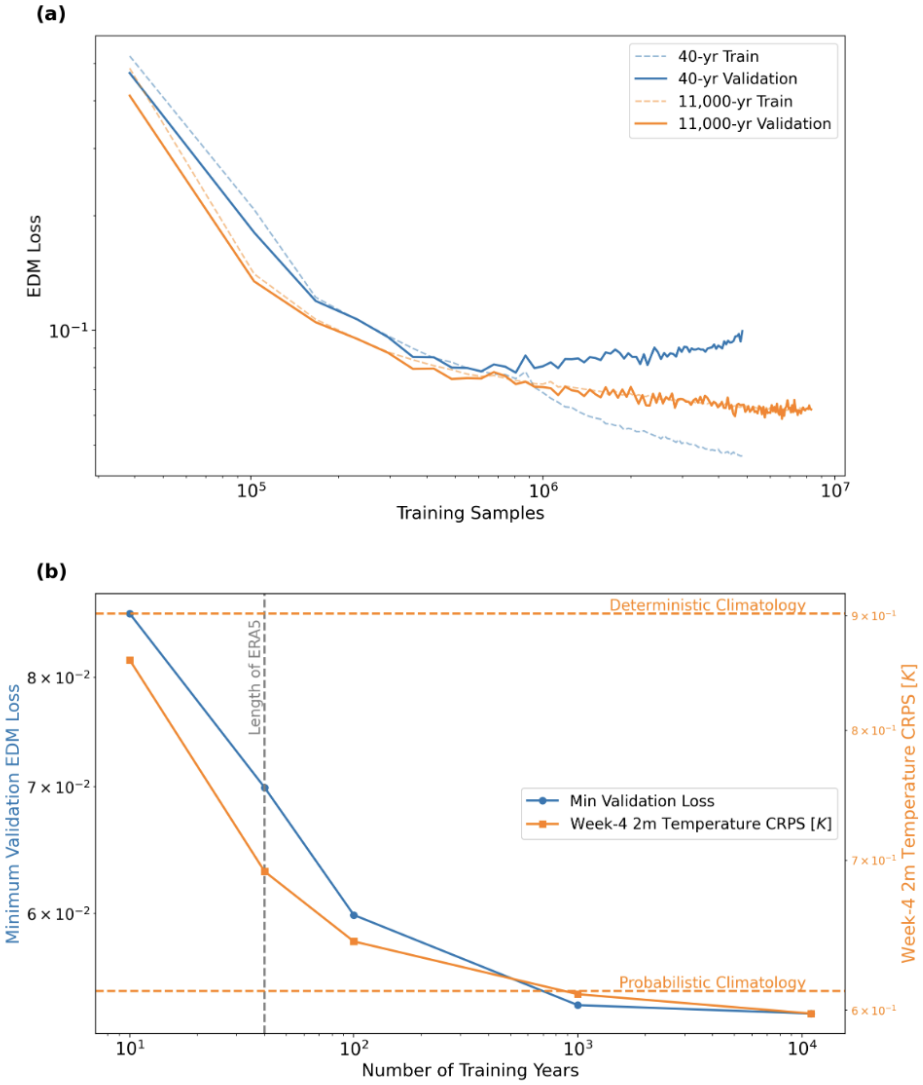
example forecasts from student model



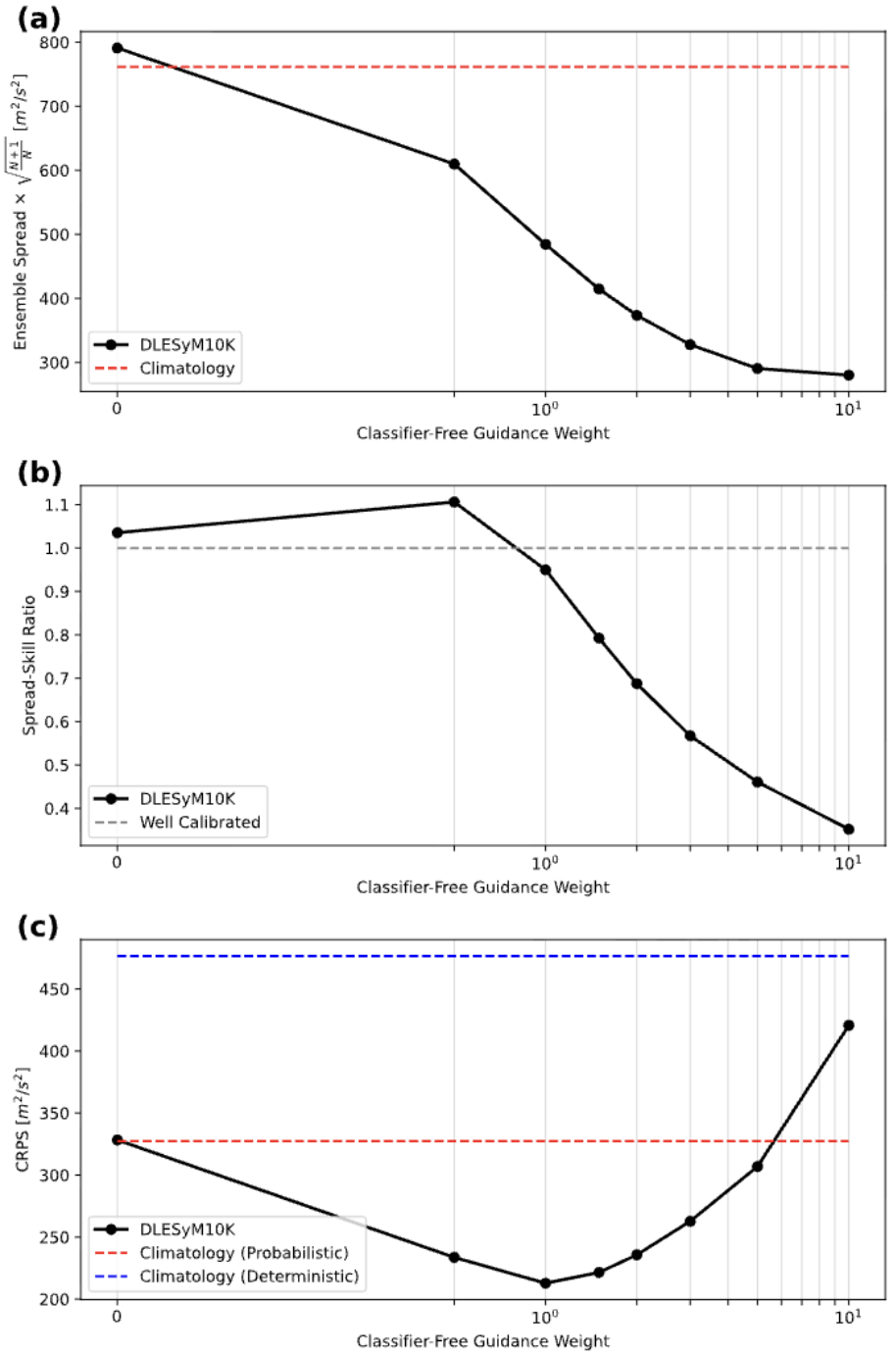
Performance Scales with Training Data

- overfitting for model trained on only 40yrs of data (blue)
- model performance also scales with increasing data!

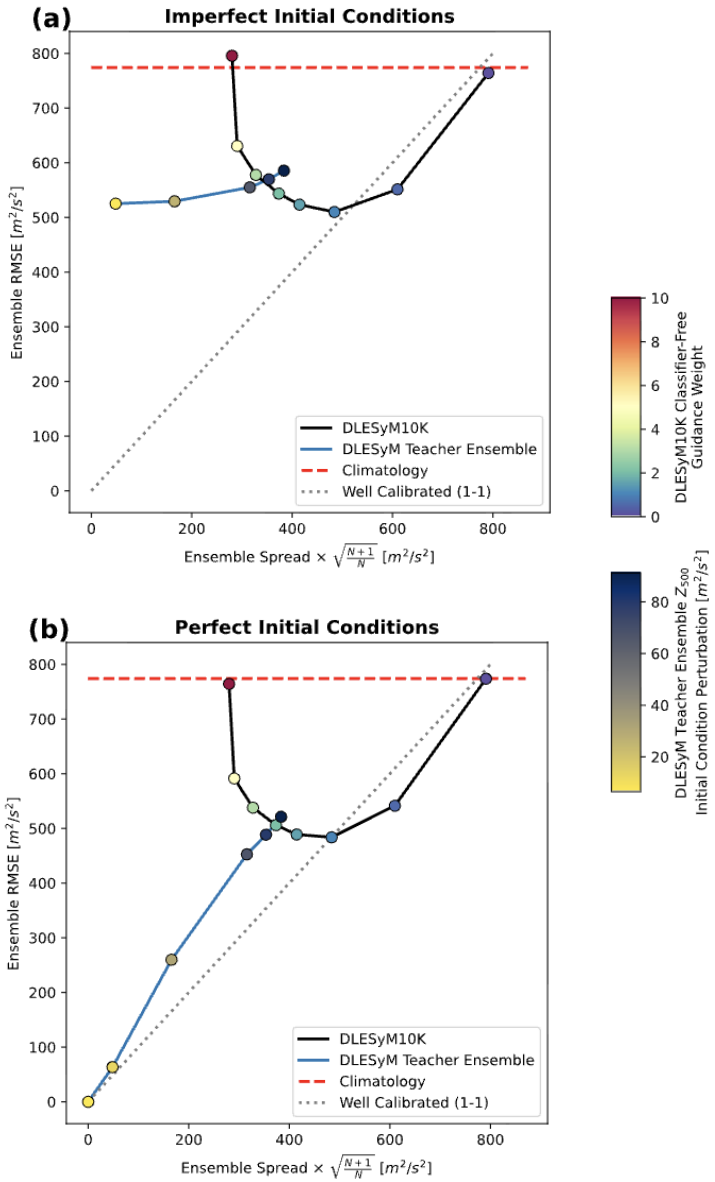
*"generating a **large synthetic corpus** of training data using autoregressive models enables **more skillful long-range models** and is the **first demonstration** of **scaling forecast performance with** an increasing volume of synthetic **training data**"*



More on CFG (I don't quite get it)

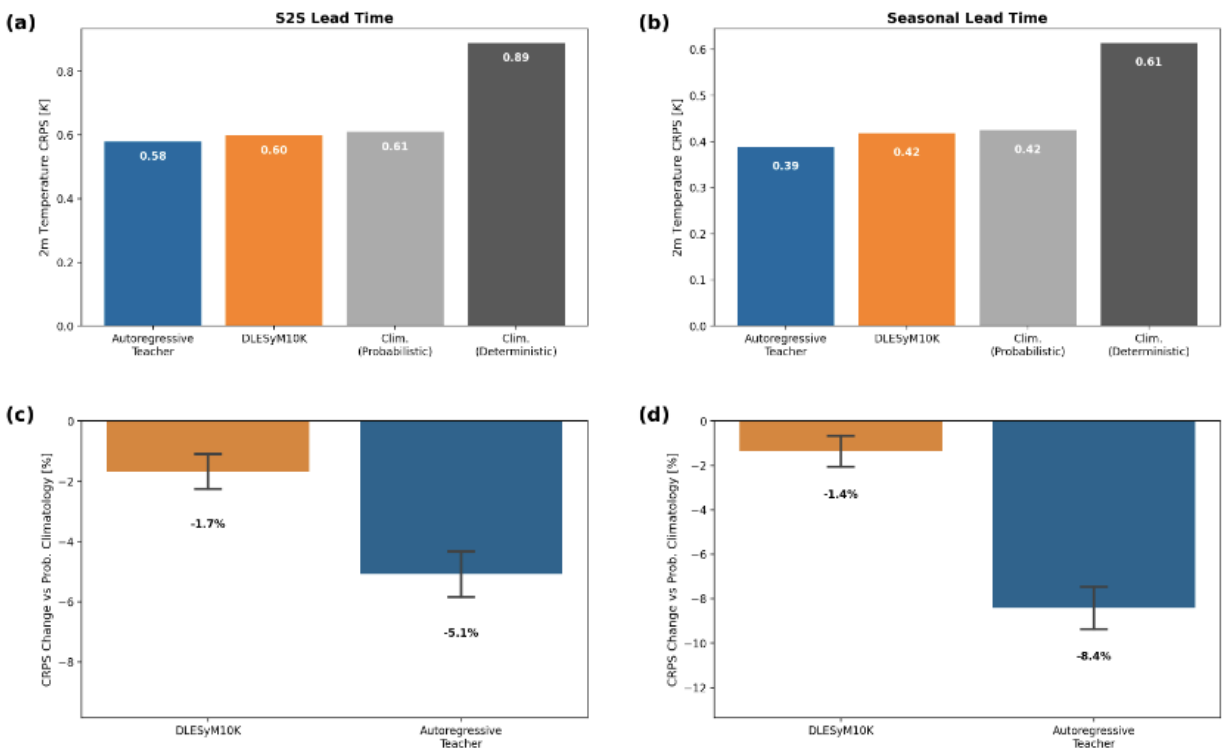


Skill Across Lead Times: medium range

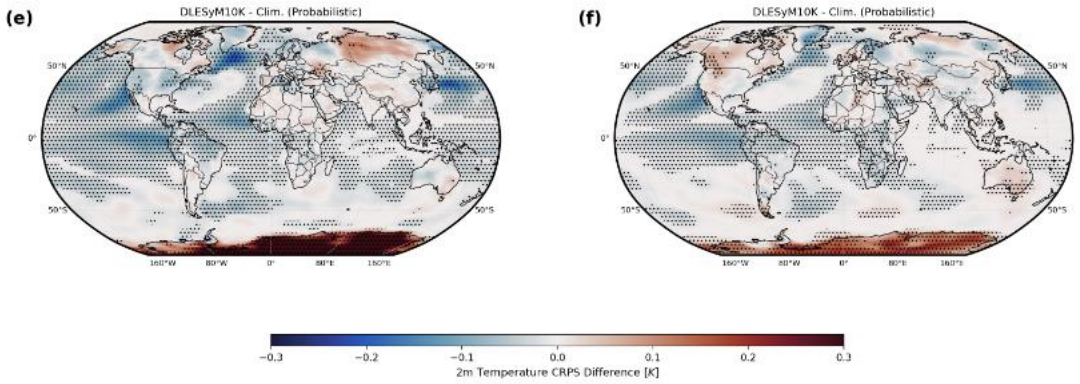


- teacher model is sensitive to IC error (error grows during autoregressive rollout)
- student model is more robust
  - uncertainty mainly comes from learned distribution rather than IC perturbations
- teacher model is consistently under dispersive
- student model achieves good spread-skill ratio with guidance strength = 1

Skill Across Lead Times: medium range



- DLESyM10K better than deterministic climatology and slightly better than prob. climatology
- similar performance to autoregressive ensemble using ICs



Real-World S2S Forecasting

- Need to overcome the domain shift of training on simulations
- two ways of accounting for this:
  - debiasing by adjusting for DLESyM climatological bias
  - finetune on ERA5 data
- DLESyM10K almost as skillful as 50 member ECMWF S2S ensemble