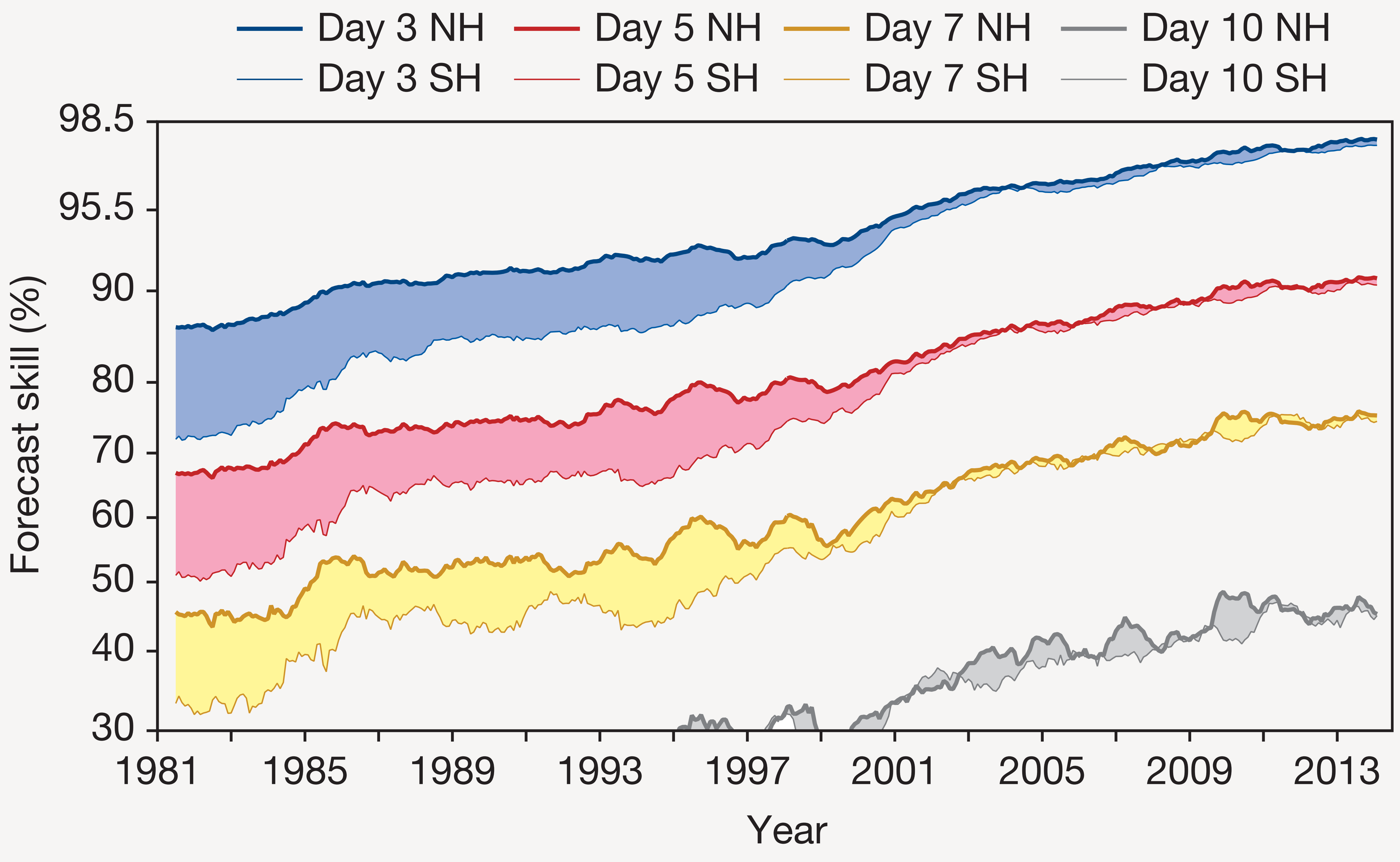
Sub-seasonal Forecasting using Large Ensembles of Data-driven Global Weather Prediction Models

JONATHAN WEYN[1*], DALE DURRAN[1], RICH CARUANA[2]

[1] University of Washington, Seattle, WA
[2] Microsoft Research, Redmond, WA
* Current affiliation: Microsoft

NWP: A "QUIET REVOLUTION"

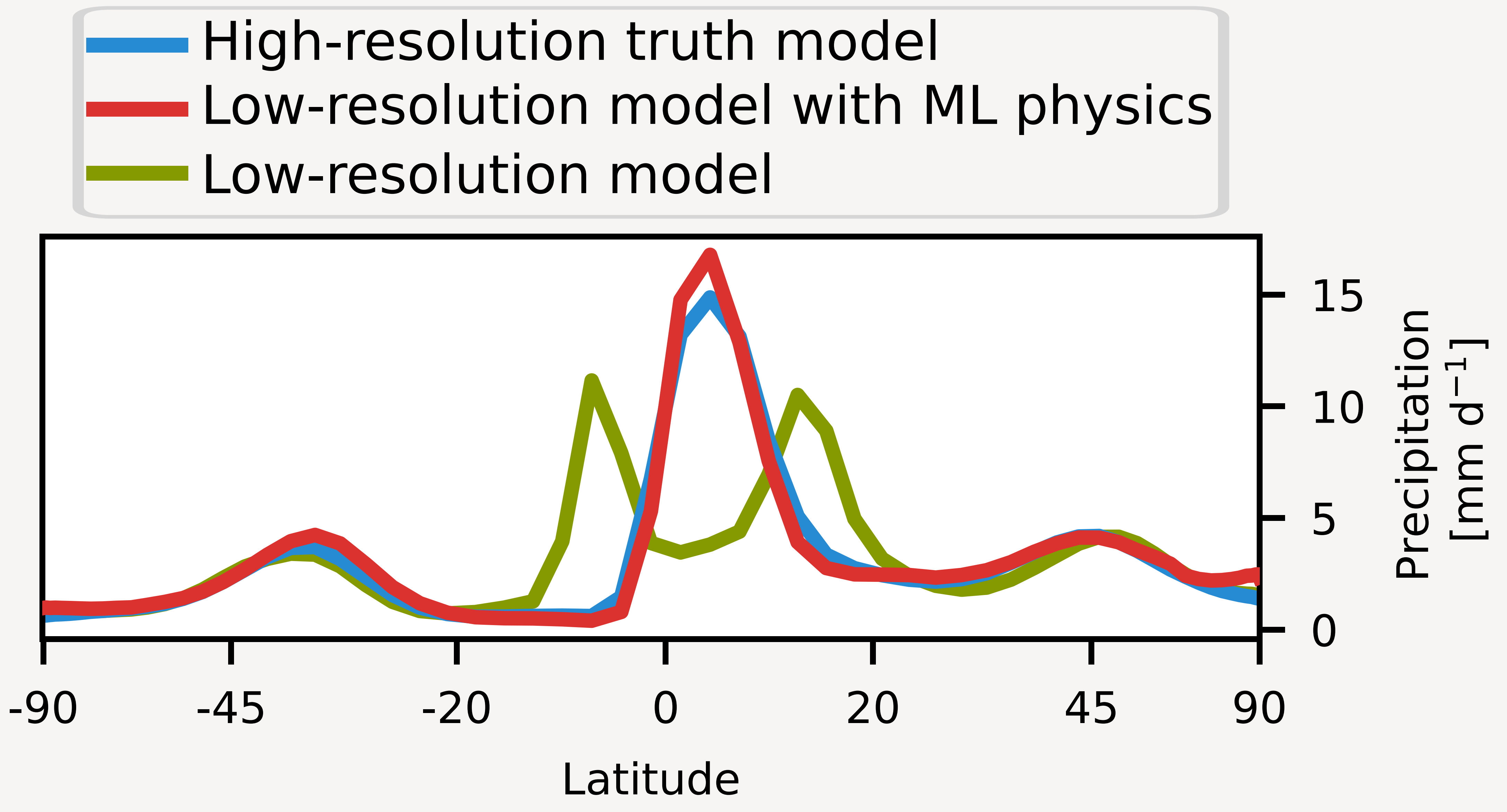
- Weather forecasting has gradually increased in accuracy, due to:
  - Vast advances in numerical representation of atmospheric physics and numerical methods
  - Large network of data collection from satellites
  - Supercomputing power
- *Is it possible to stretch these improvements even further by applying modern machine learning techniques?*



Bauer et al. (2015)

EXPLOSION IN APPLICATION OF ML IN METEOROLOGY

- Post-processing NWP models (Rasp and Lerch 2018)
- Identification and prediction of extreme weather events (Racah et al. 2017, Lagerquist et al. 2019, Herman and Schumacher 2018)
- Improving physical parameterizations in NWP models, and improving their computational efficiency (Rasp et al. 2018, Brenowitz and Bretherton 2018, McGibbon and Bretherton 2019)

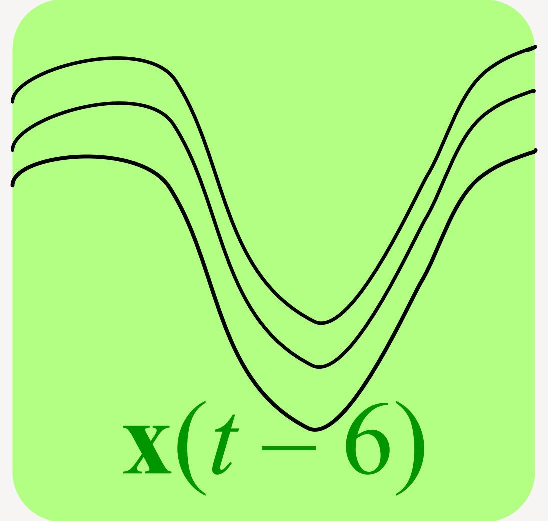


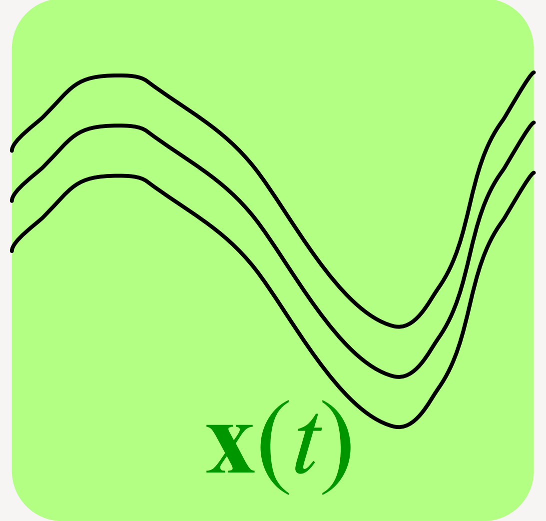
Rasp et al. (2018)

WHAT IF MACHINES COULD PREDICT THE EVOLUTION OF THE ENTIRE ATMOSPHERE?

- We developed our Deep Learning Weather Prediction (DLWP) model
  - Use deep convolutional neural networks (CNNs) on a cubed sphere grid
  - While this is essentially replacing NWP models with deep learning, there are important caveats
    - only a few variables: not a complete forecast
    - subject to using training data dependent on state-of-the-art data assimilation

Indefinitely iterable predictions for every 6 hours





Observed states at two times

$\tilde{\mathbf{x}}(t-6, t) \rightarrow$ Algorithm $\rightarrow$

Predicted states at two times

$\tilde{\mathbf{y}}(t+6, t+12) \rightarrow$ Algorithm $\rightarrow$

Predicted states at two times

$\tilde{\mathbf{y}}(t+18, t+24) \cdots$

$$J_1 = \left\| \mathbf{y}(t+6) - \mathbf{x}(t+6) \right\|_2^2 + \left\| \mathbf{y}(t+12) - \mathbf{x}(t+12) \right\|_2^2$$

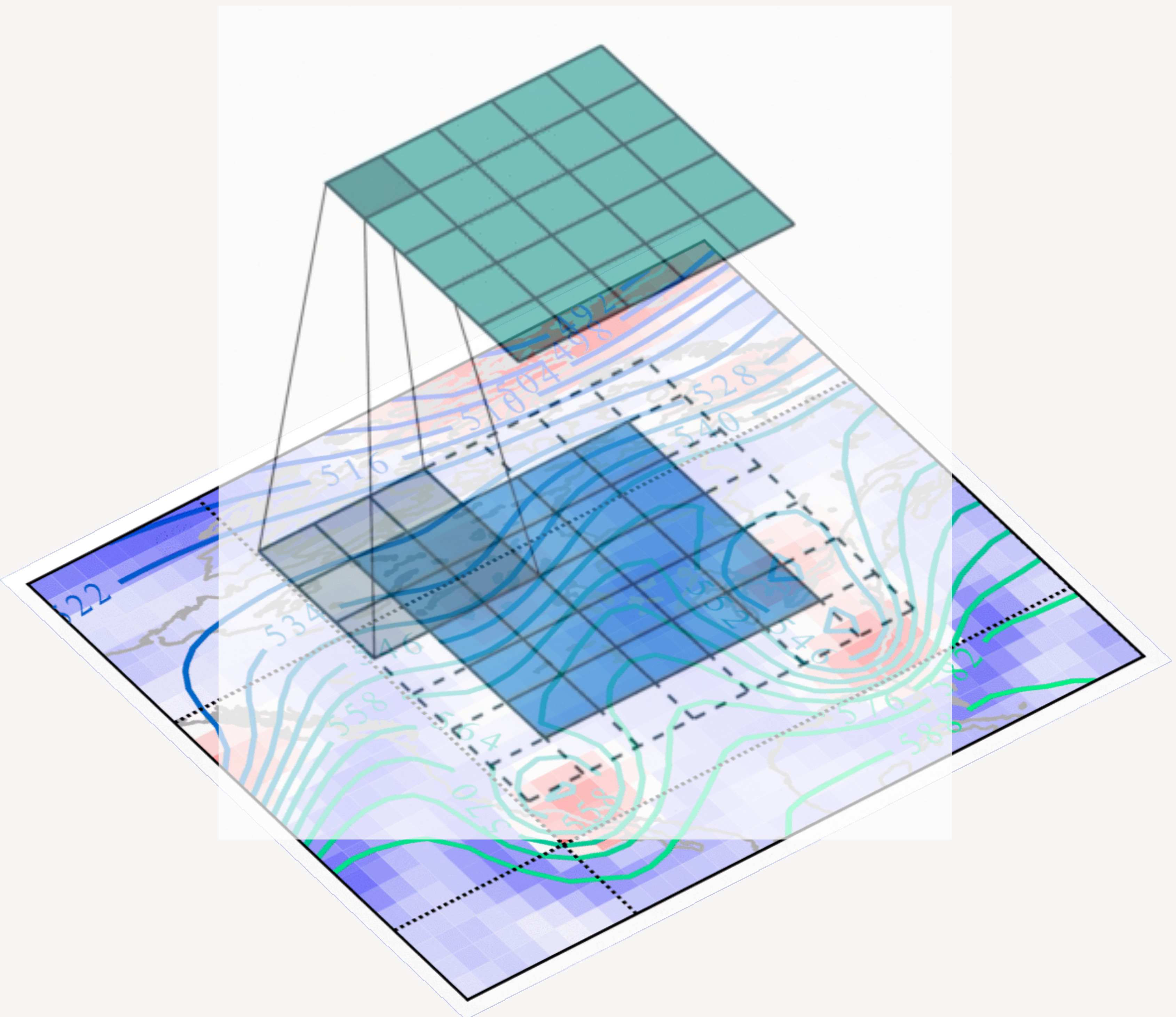
$$J_2 = \left\| \mathbf{y}(t+18) - \mathbf{x}(t+18) \right\|_2^2 + \left\| \mathbf{y}(t+24) - \mathbf{x}(t+24) \right\|_2^2$$

$$J = J_1 + J_2$$

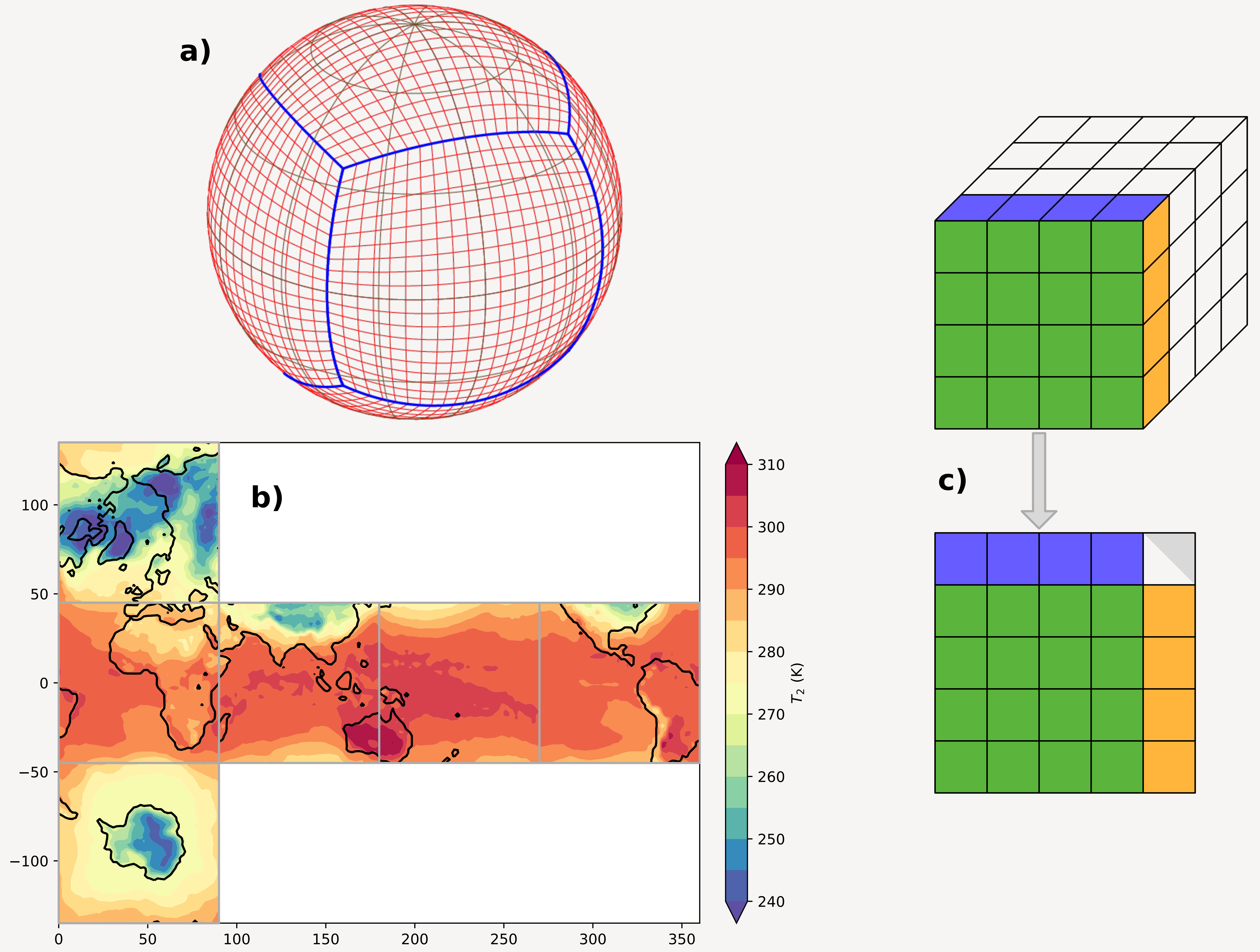
Trained to minimize the error over a 24-hour forecast

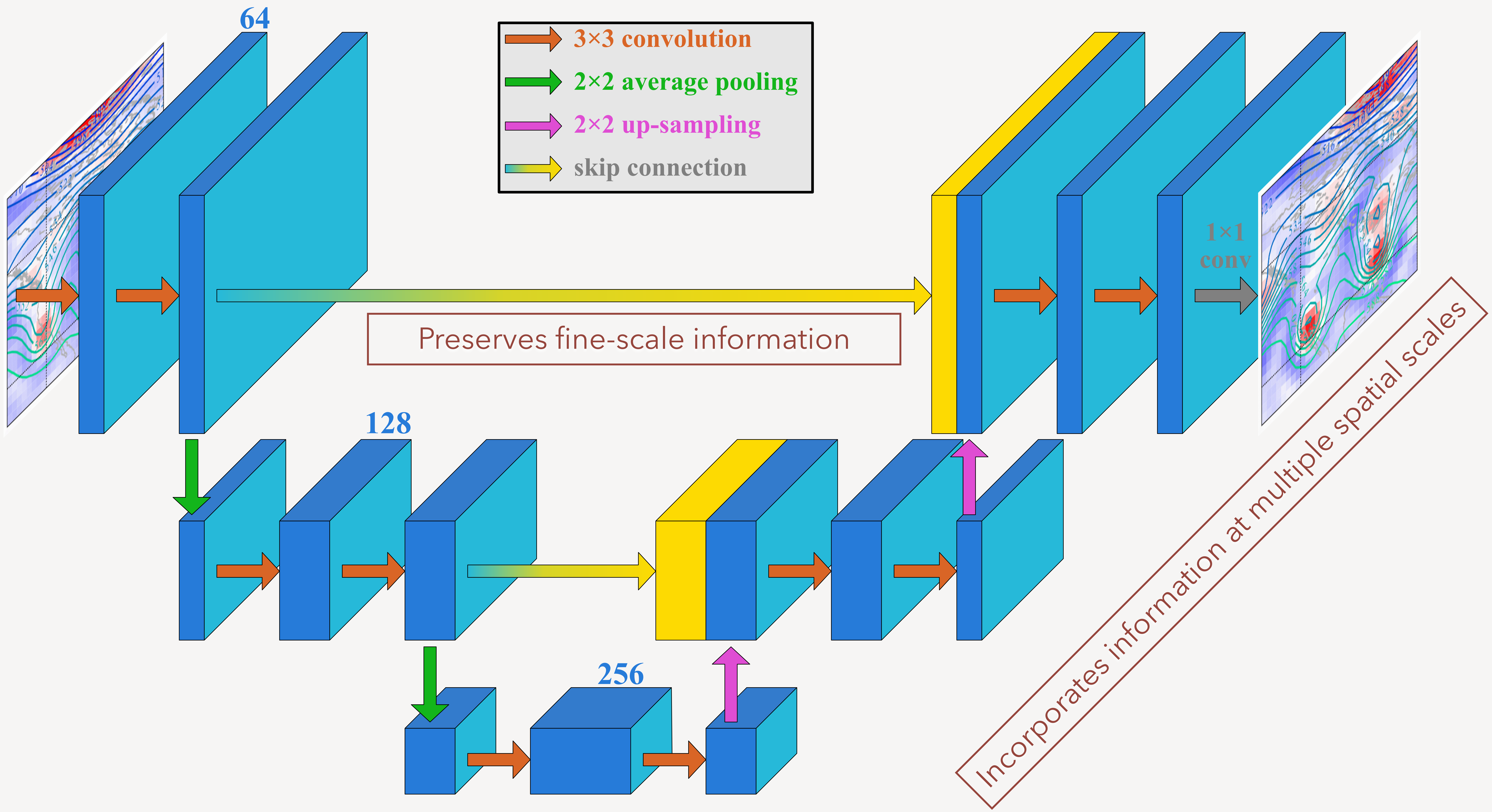
CONVOLUTIONAL NEURAL NETWORKS

- **CNNs**
  - are ideally suited for "images" on a grid
  - account for local spatial correlations
  - identify patterns, edges, shapes
- **However, equirectangular (latitude-longitude) images have heavy distortion near the poles**



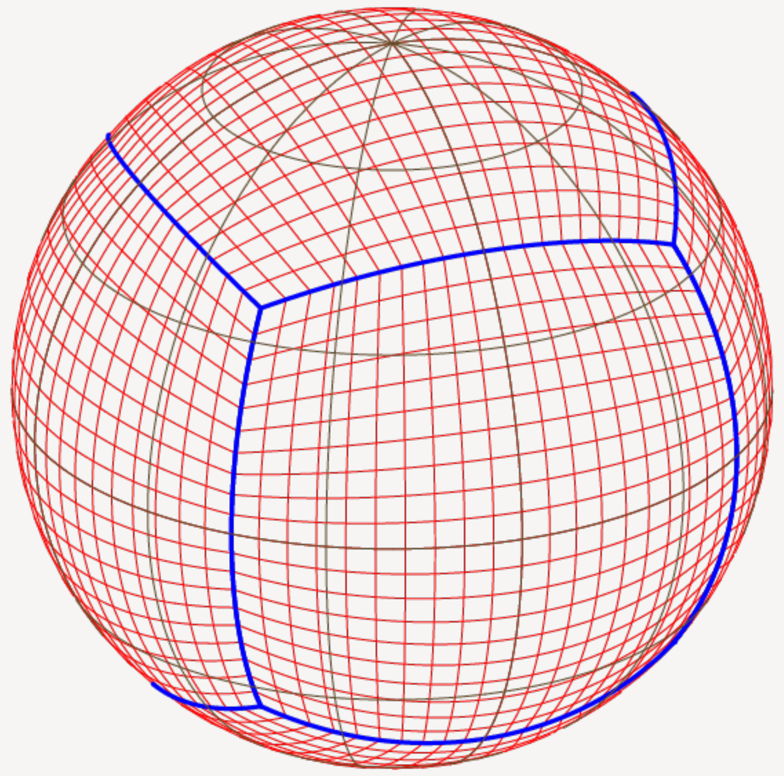
CONVOLUTIONS ON THE CUBED SPHERE




64
3×3 convolution
2×2 average pooling
2×2 up-sampling
skip connection
1×1
conv
Preserves fine-scale information
128
256
Incorporates information at multiple spatial scales

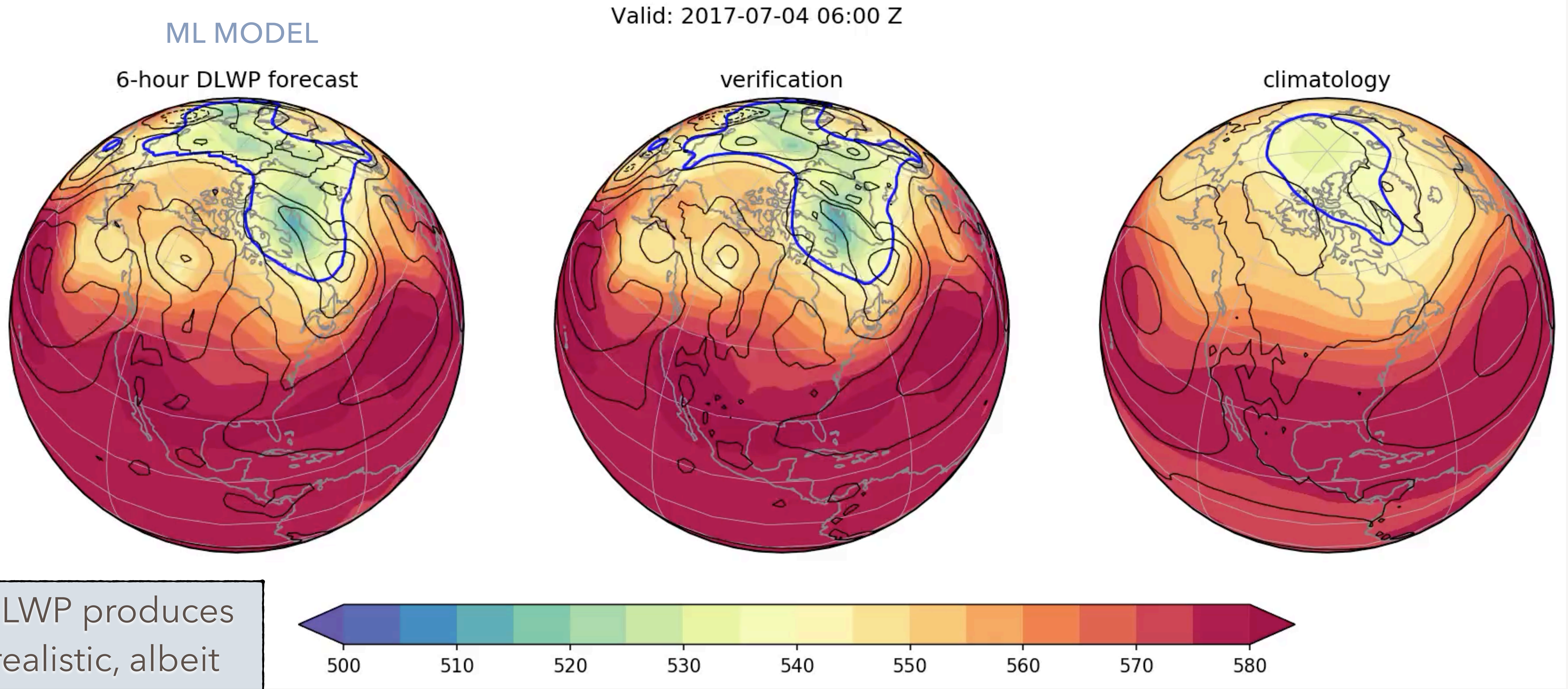
DATA

- **ECMWF ERA5 input fields (6)**
  - ~1.4° resolution; 6-hourly time step
  - Z_{500} , Z_{1000}
  - 300–700 hPa thickness
  - 2-m temperature
  - T_{850}
  - Total column water vapor



- **Prescribed fields (3)**
  - TOA incoming solar radiation
  - land-sea mask
  - topography
- **Data sets**
  - Train: 1979–2012; validate: 2013–16
  - Evaluate: twice weekly in 2017–18

ML MODEL



DLWP produces realistic, albeit smoothed, atmospheric states.

Colors: 500-hPa geopotential height, dkm

Contours: 1000-hPa geopotential height, every 100 m, dashed=negative

https://www.dropbox.com/s/xmi0v81fbcw4t1r/Weyn_ms02.mp4?dl=0

Can our DLWP model form the basis for a good-performing large ensemble prediction system?

- *DLWP is inferior to NWP models. Why bother?*
  - The larger the ensemble, the better!
  - DLWP has a significant computational advantage
    - How long would it take to run a 1000-member ensemble of 1-month forecasts at 1.5° resolution?
      - 10 minutes
      - Single workstation with GPU
    - In comparison, a comparable dynamical model would take about 16 days

Morris MacMatzen / Getty via Vox

OBTAINING CORRECT ENSEMBLE SPREAD

	NWP	DLWP

Observation uncertainty	• Perturb initial conditions
	  • Ensemble 4DVAR
	  • SVD
		• Perturb initial conditions
		  • ERA5 ensemble
		  • *NOT optimal!*

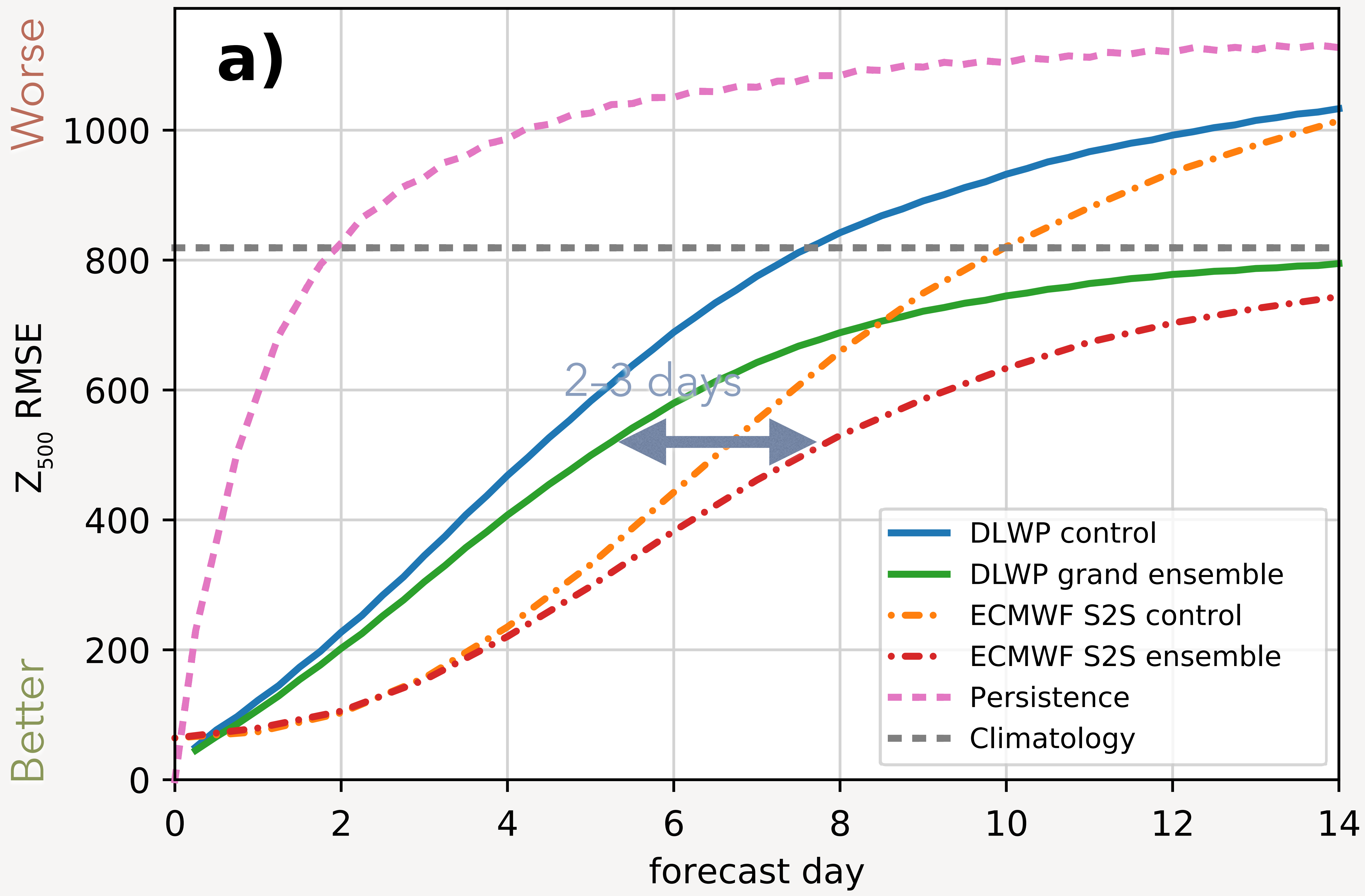
Model uncertainty	• Stochastically perturbed parameterization tendencies (SPPT)
	• Stochastic KE backscatter (SKEB)
		• Random seeds in training
		  • random data sampling
		  • random weight initialization
		  • *"swapping" physics*

DLWP VERSUS THE ECMWF ENSEMBLE

	DLWP	ECMWF
Variables	6 2-D variables	9 prognostic 3-D variables, 91 vertical levels
Resolution	~160 km	~18 km (36 km after day 15)
Other physics	3 prescribed inputs	Many parameterizations
Coupled models	None	Ocean, wave, sea ice
Initial condition perturbations	10 (ERA5)	50 (SVD/4DVAR)
Model perturbations	32 "stochastic" CNNs	stochastic physics perturbations
Ensemble size	320 (+ control)	50 (+ control)

Average forecast error in 500-hPa geopotential

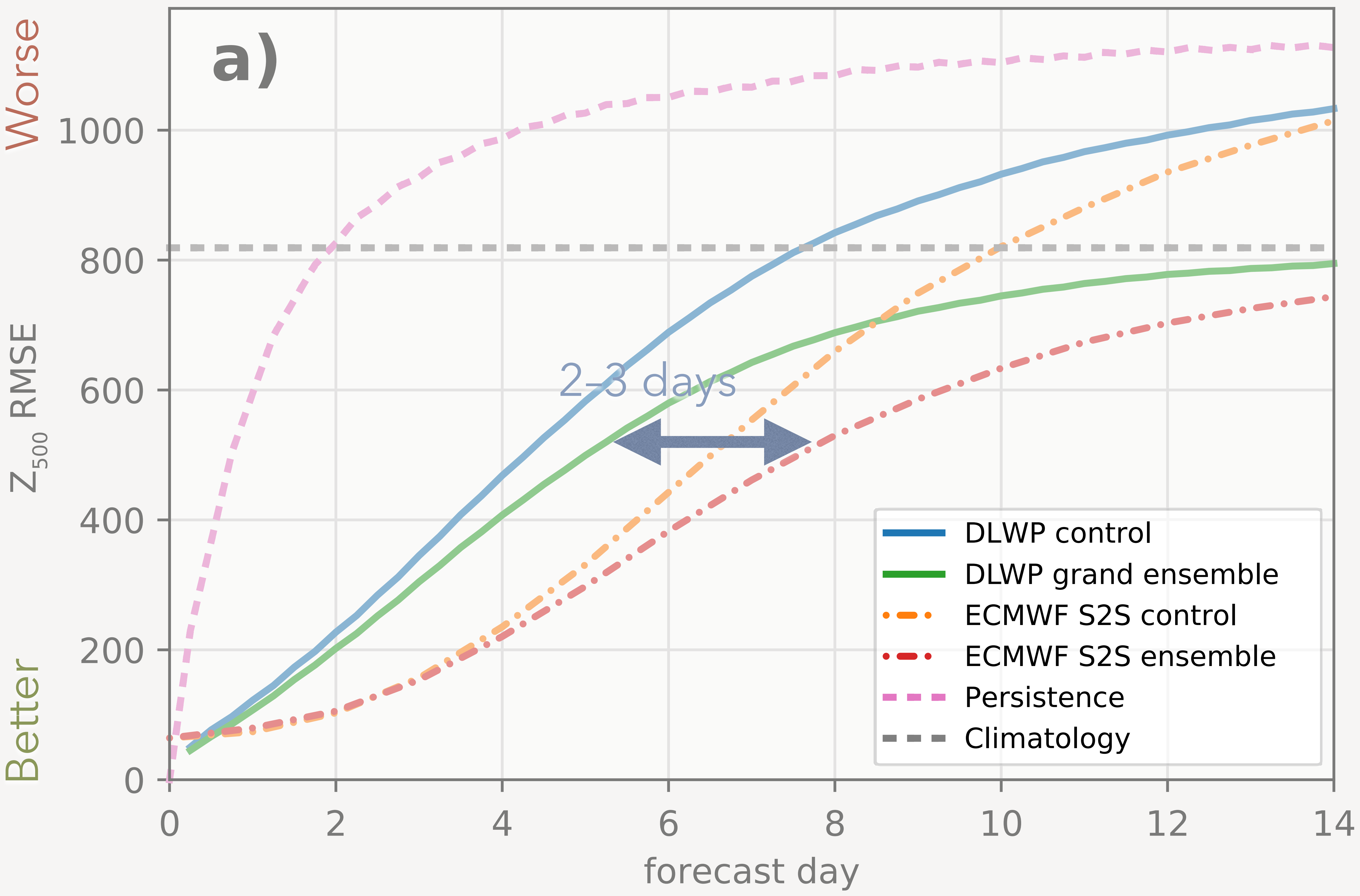
Root-mean-squared error



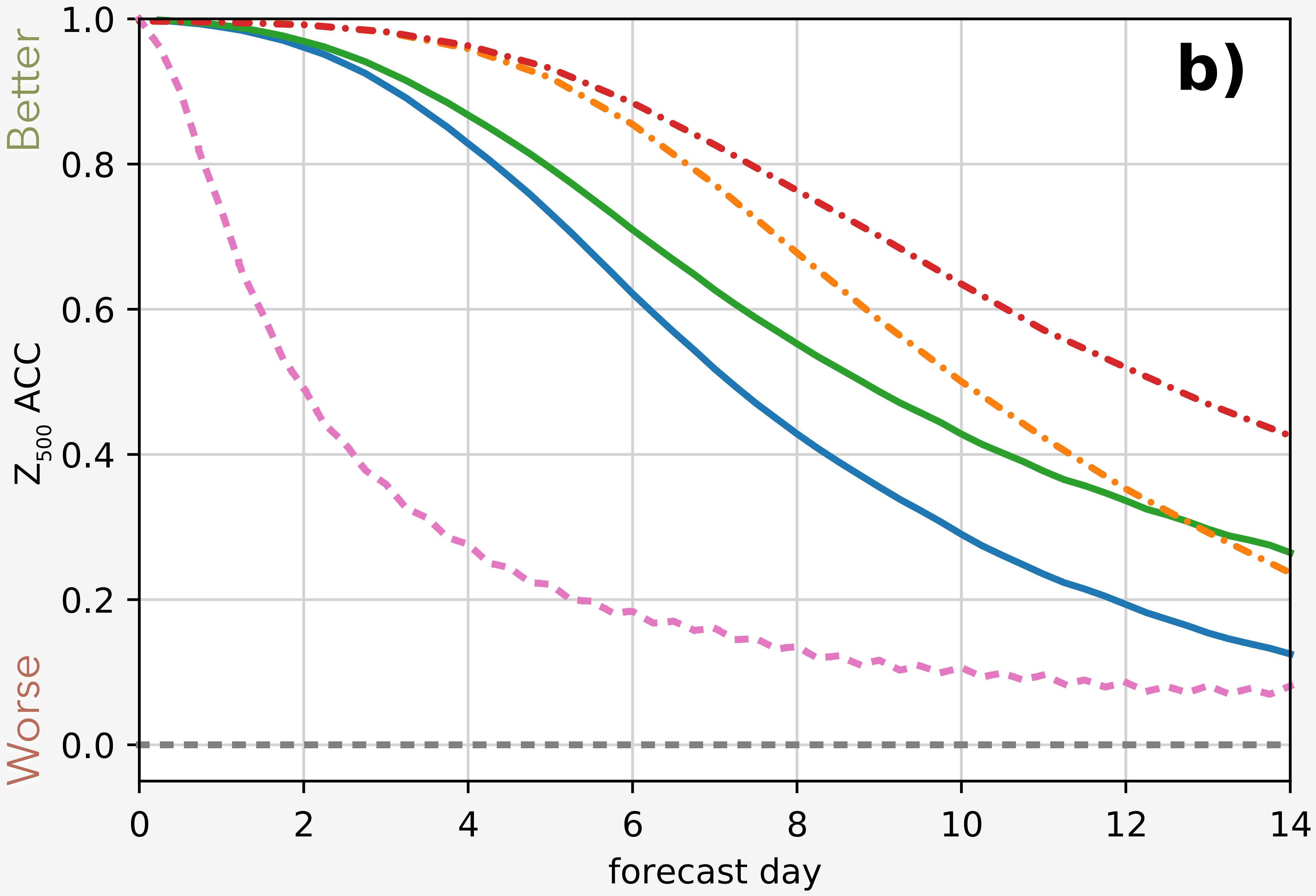
The DLWP ensemble only lags the state-of-the-art ECMWF ensemble by 2-3 days' lead time.

Average forecast error in 500-hPa geopotential

Root-mean-squared error

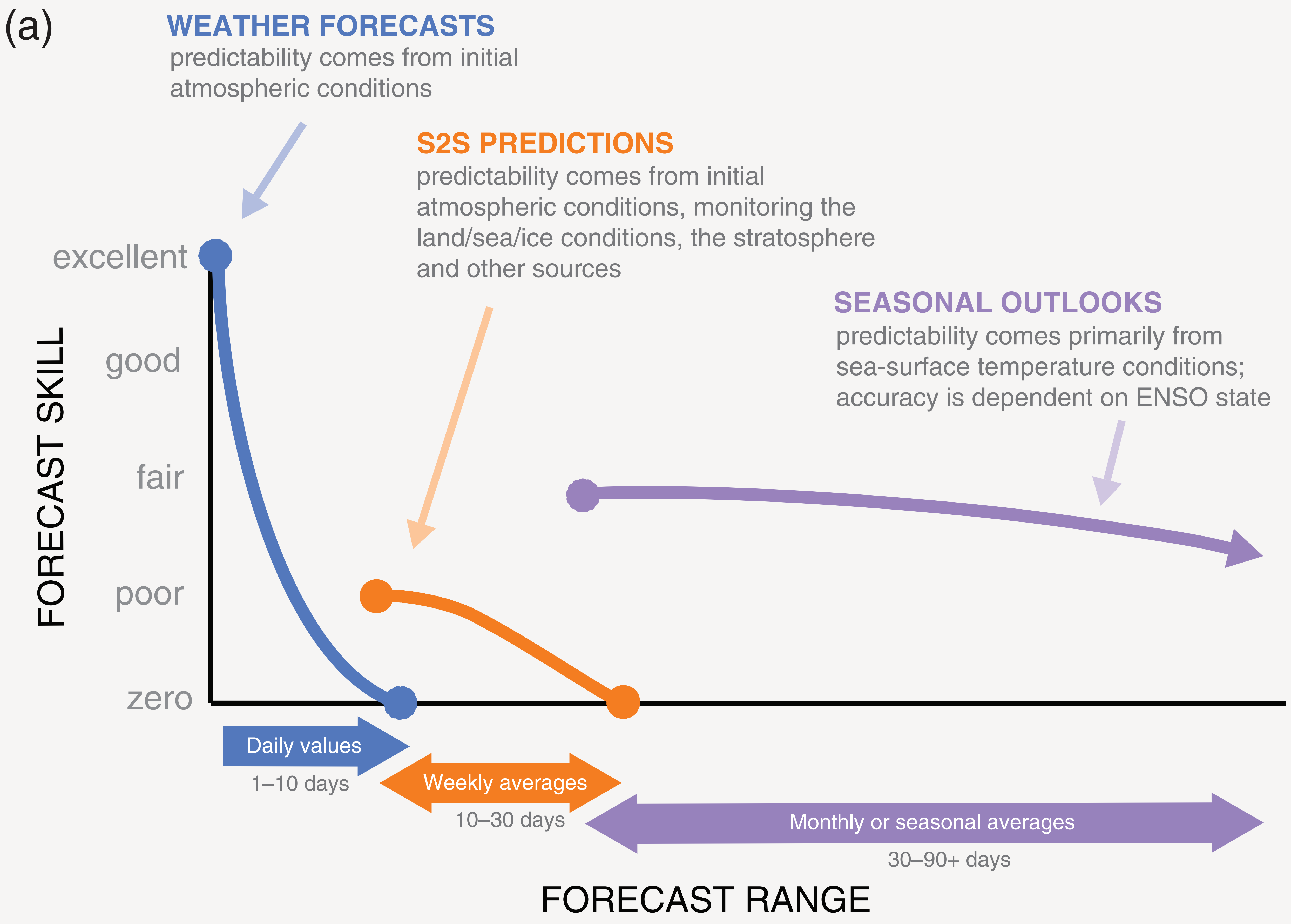


Anomaly correlation coefficient



The DLWP ensemble only lags the state-of-the-art ECMWF ensemble by 2-3 days' lead time.

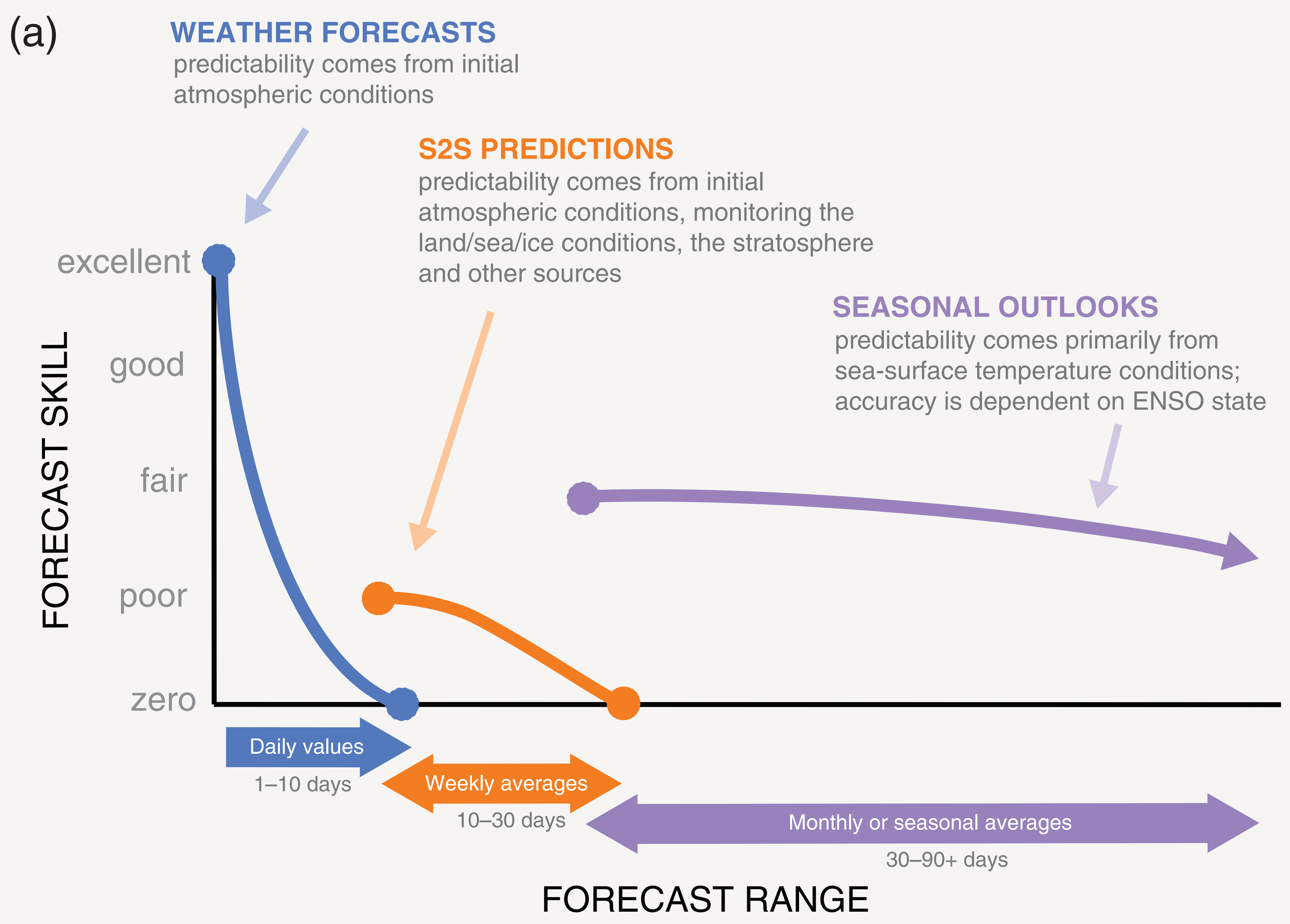
Can our large ensemble produce skillful sub-seasonal-to-seasonal (S2S) forecasts?



White et al. (2017)

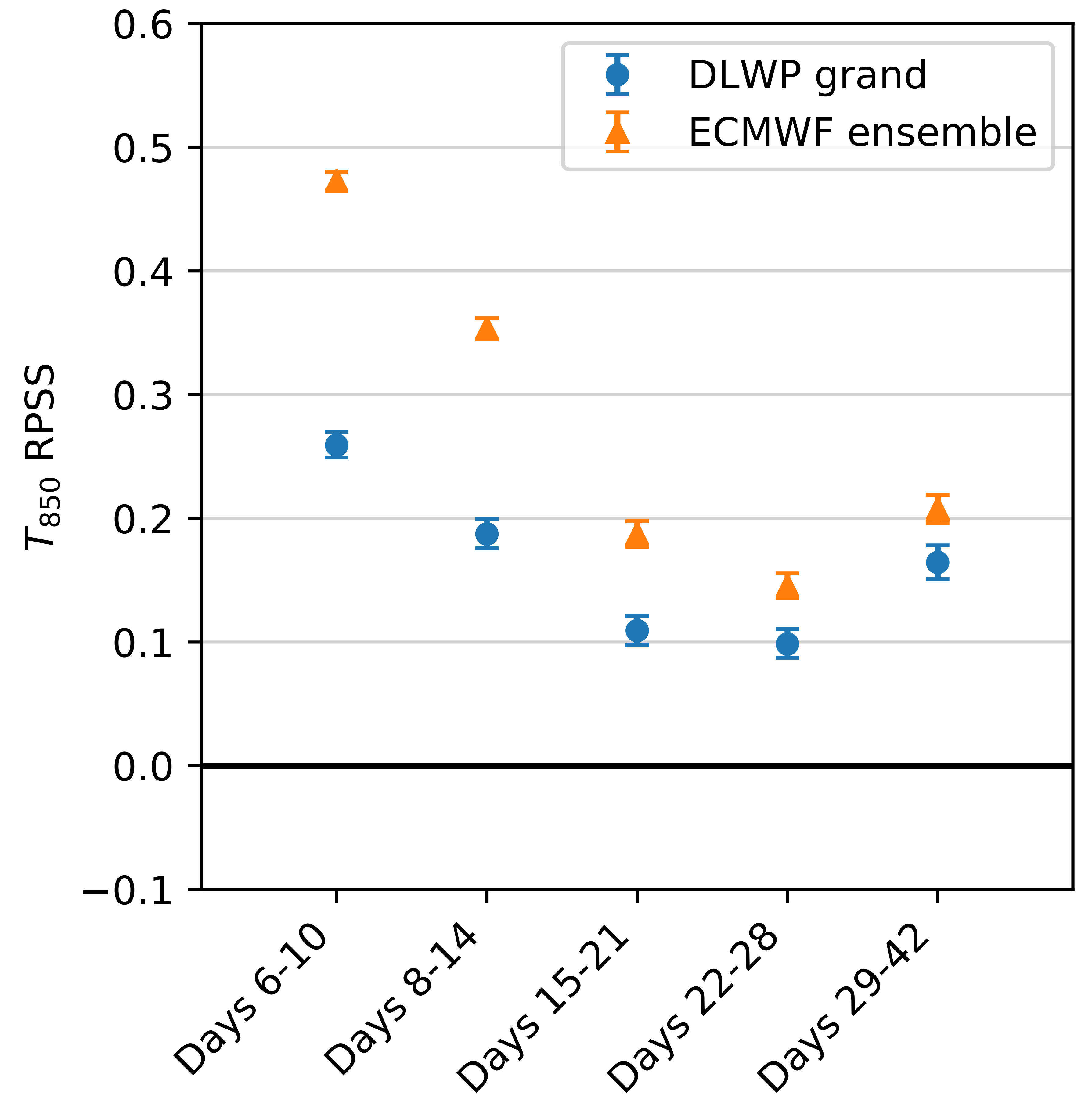
SUB-SEASONAL-TO-SEASONAL FORECASTING

- DLWP cannot be expected to compete with the state-of-the-art
- DLWP is lacking
  - an ocean model
  - land surface parameters e.g. soil moisture
  - sea ice
- We also do not forecast precipitation with the current model iteration
- Cannot represent physics of MJO or ENSO



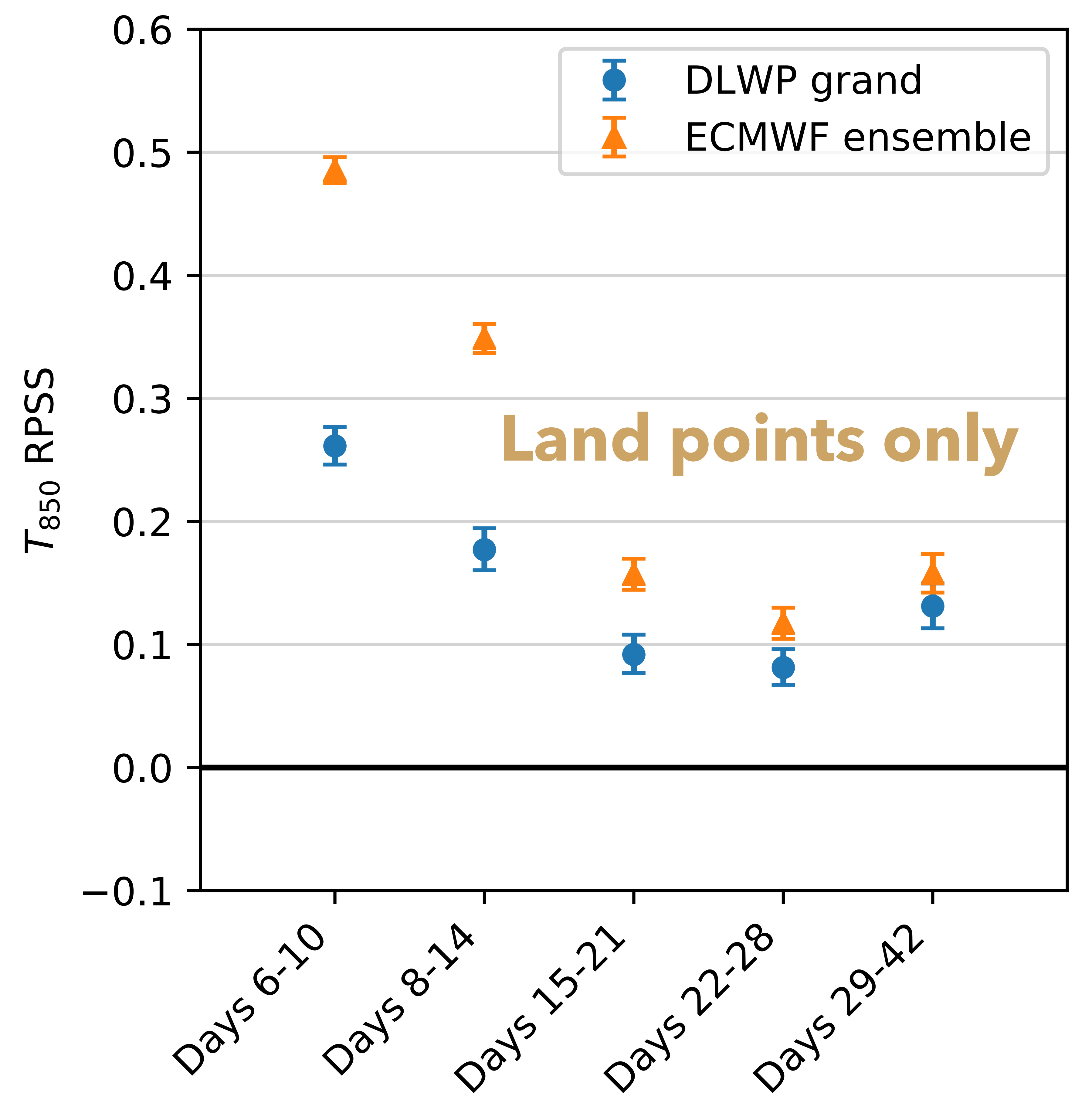
PROBABILISTIC SKILL SCORES OF S2S FORECASTS

- Ranked probability skill score (RPSS)
  - Three equally-probable terciles relative to a 1981–2010 climatology
    - below-, near-, and above-normal
  - Forecast probability is the fraction of ensemble members
  - Evaluate squared error of forecast probability versus observed category
  - Normalize relative to random chance prediction
  - Perfect score is 1, higher is better, negative score is no skill relative to random chance



PROBABILISTIC SKILL SCORES OF S2S FORECASTS

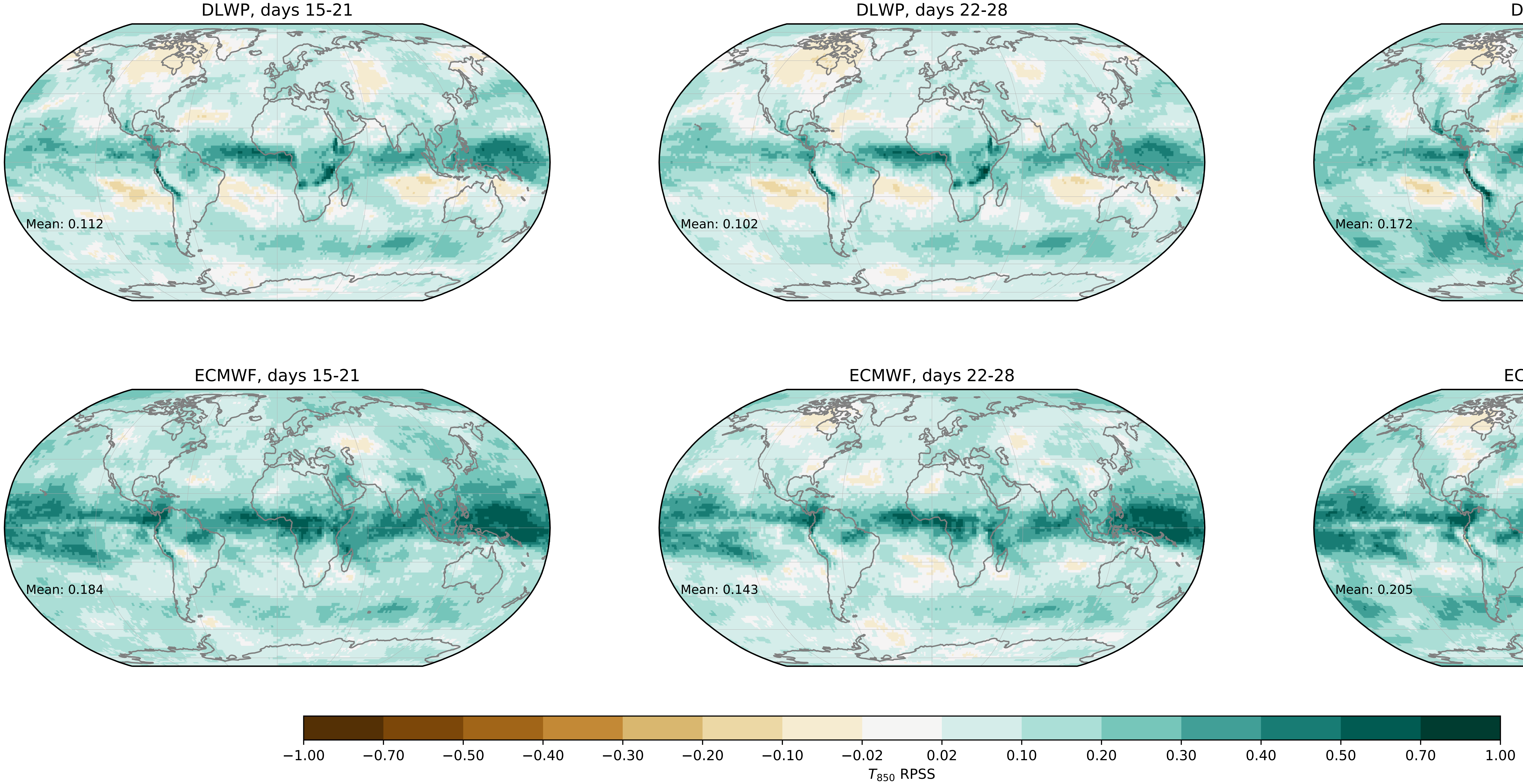
- Ranked probability skill score (RPSS)
  - Three equally-probable terciles relative to a 1981–2010 climatology
    - below-, near-, and above-normal
  - Forecast probability is the fraction of ensemble members
  - Evaluate squared error of forecast probability versus observed category
  - Normalize relative to random chance prediction
  - Perfect score is 1, higher is better, negative score is no skill relative to random chance



Week 3

While most skill comes from the tropics, the DLWP ensemble has positive skill scores over nearly all land masses.

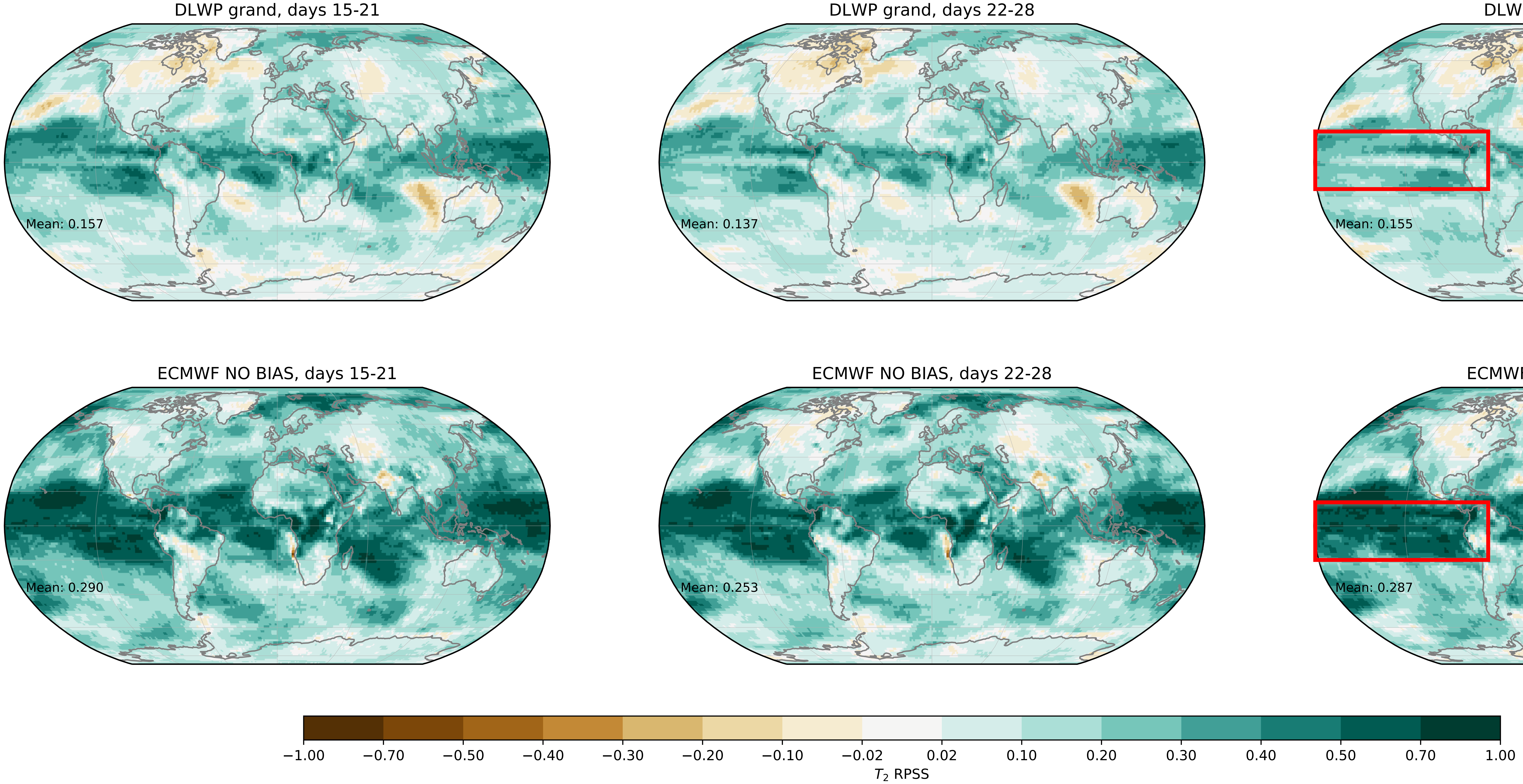
Weeks 5-6



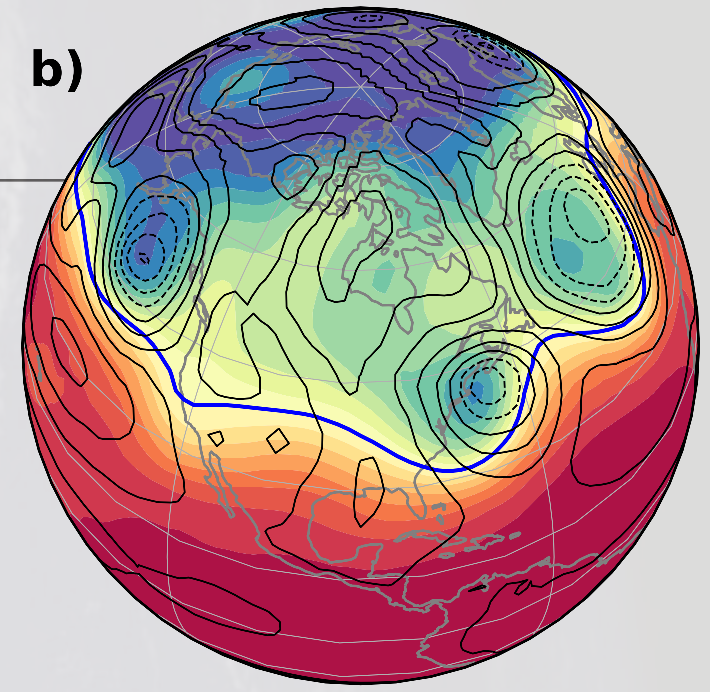
DLWP fails to predict the onset of ENSO patterns in forecasts initialized in boreal autumn.

Week 3

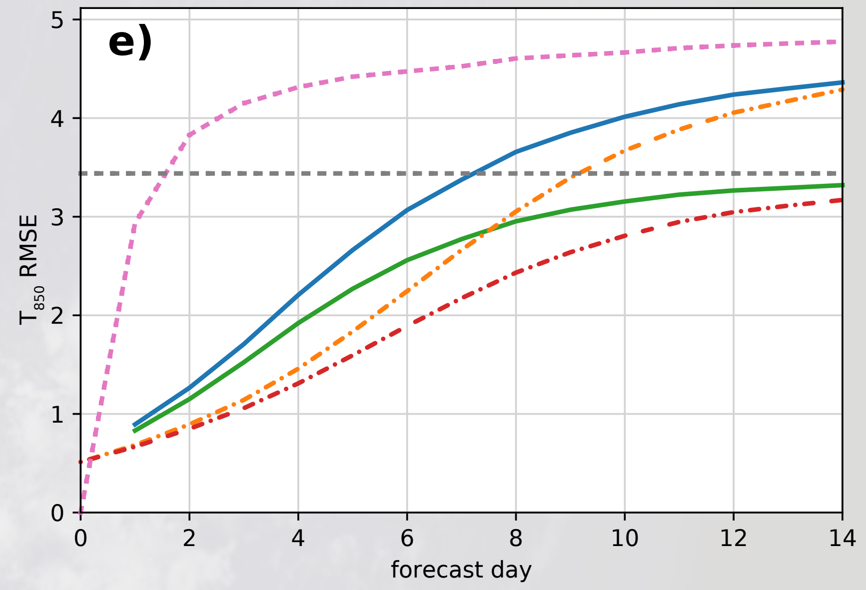
Weeks 5-6



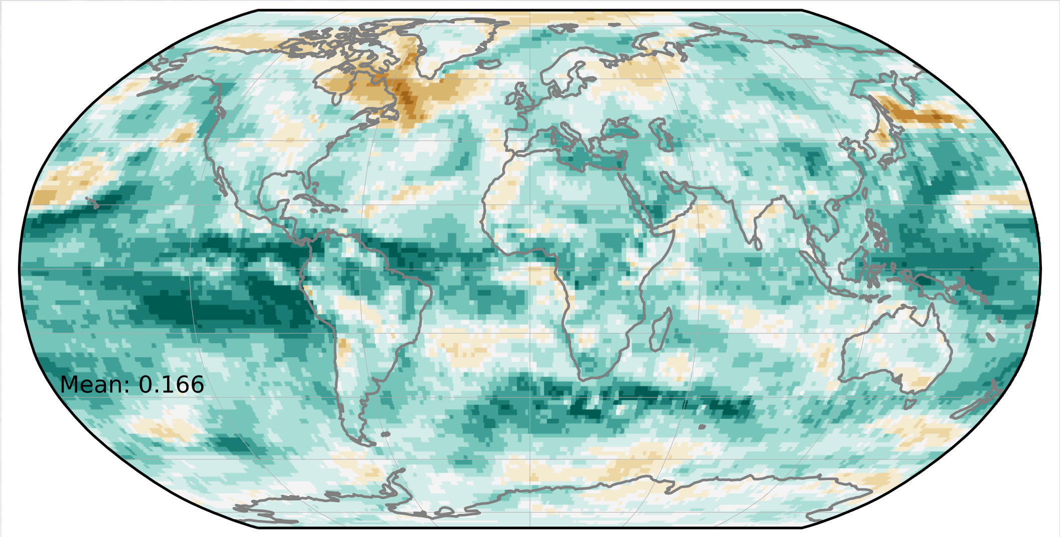
TAKE-AWAYS



- Purely data-driven algorithms (CNNs) can produce realistic, indefinitely-stable global weather forecasts with skill that only lags the state-of-the-art ECMWF model by 2-3 days of lead time.



- By leveraging initial-condition perturbations and the internal randomness of the CNN training process it is possible to produce a high-performing ensemble of data-driven weather models that requires several orders of magnitude less computation power than comparable dynamical models.



- For weekly-averaged forecasts in the S2S forecast range, a notable skill gap for dynamical models, our data-driven ensemble model has useful skill relative to climatology and persistence. It is not far behind the state-of-the-art ECMWF ensemble.