Improving sub-seasonal monsoon forecasts over India with Machine Learning

Andrew W. Robertson, **Emile Esmail**, Azhar Ehsan, Bohar Singh, Aaron Kaplan, Avijit Dey, Susmitha Joseph

Columbia Climate School Center for Climate Systems Research @ NASA GISS
Indian Institute of Tropical Meteorology, Pune

awr2001@columbia.edu

IWM-8, Pune, March 17, 2025

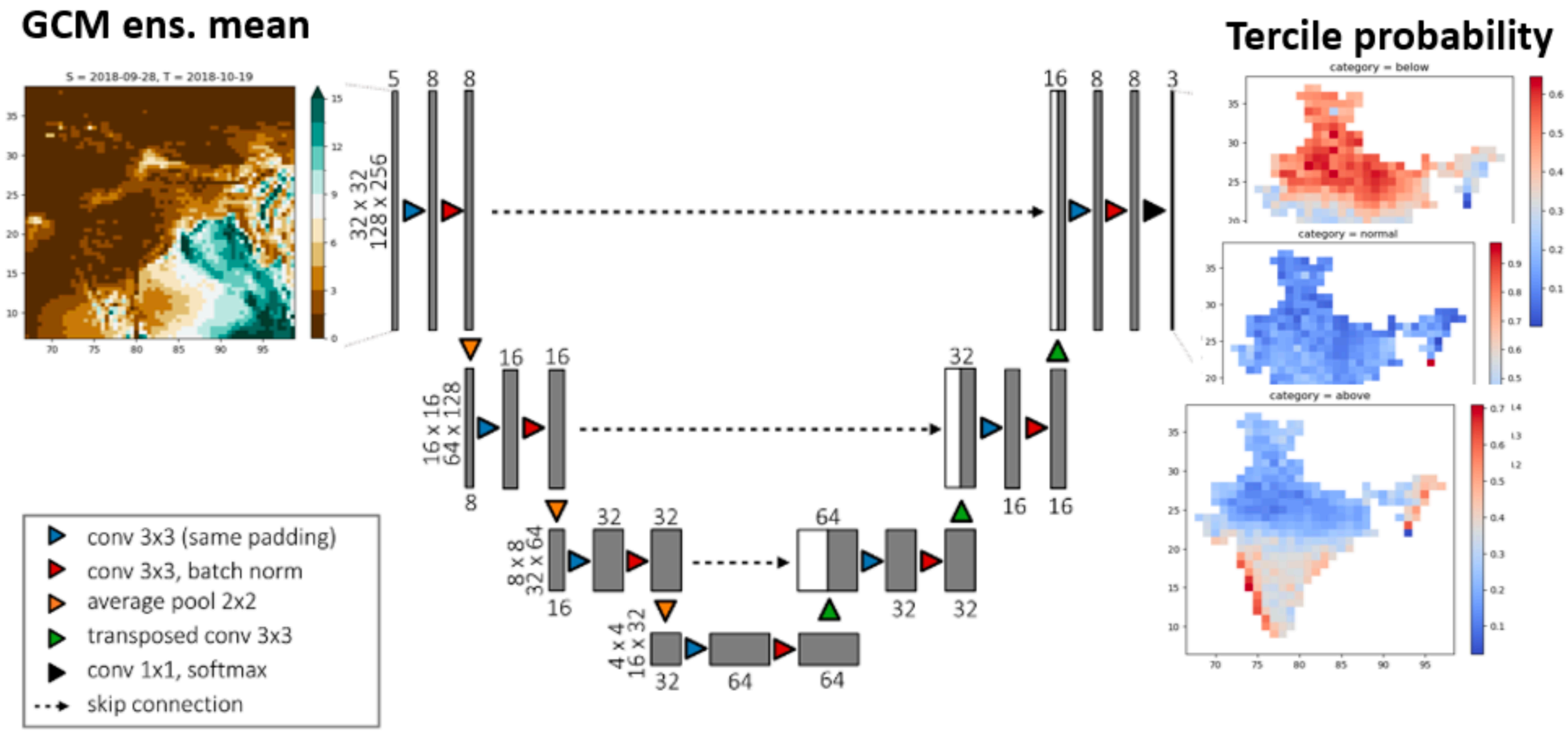
Introduction

- Goal is to improve sub-seasonal to seasonal (S2S) forecasts of Indian summer monsoon rainfall using machine learning
- We use convolutional neural networks (CNNs) to post-process forecasts of precipitation from three sub-seasonal ensemble prediction systems (NCEP GEFSv12, ECMWF, IITM ERPv2)
- Based on U-Net architecture used globally by Horat & Lerch (2023)
- Benchmark skill against logistic regression

Datasets

Summary of Data and Model Characteristics			
Model	**GEFSv12**	**ECMWF**	**IITM ERPv2**
Resolution	$1° \times 1°$	$1.5° \times 1.5°$	$0.5° \times 0.5°$
Hindcast Period	1989-2018	2003-2018	2003-2018
# Members	11	10	18
Initialization Dates	Every Wednesday	Every Monday and Thursday	Once in a week but for fixed dates, e.g., 4th May, 11th May, 18th May, 25th May, 1st June, 8th June, 15th June, and so on.
Observation: Indian Meteorological Department (IMD) rainfall regridded to GCM spatial resolution			
Training Season: June–September			
Lead Times: 0–7 days (Week 1), 7–14 days (Week 2), 14–28 days (Weeks 3&4)			

U-Net Architecture



Adapted from
Horat & Lerch (2023)

- A convolutional neural network (CNN) with an encoder-decoder structure
- Input is ensemble mean GCM precipitation
- Final convolution layer produces tercile probabilities (softmax function)
- Tercile edges computed on a weekly basis (include seasonality)

- We compare the Unet to a **baseline Extended Logistic Regression** (Wilks 2009), a linear probabilistic forecast calibration method. Extended Logistic Regression (ELR) uses the ensemble mean $\overline{x}_{\text{ens}}$, and an additional explanatory variable $g(q)$ can be used to produce the probability p of non-exceedance of the quantile q :

$$\ln\left(\frac{p}{1-p}\right) = f\left(\overline{x}_{\text{ens}}\right) + g(q),$$

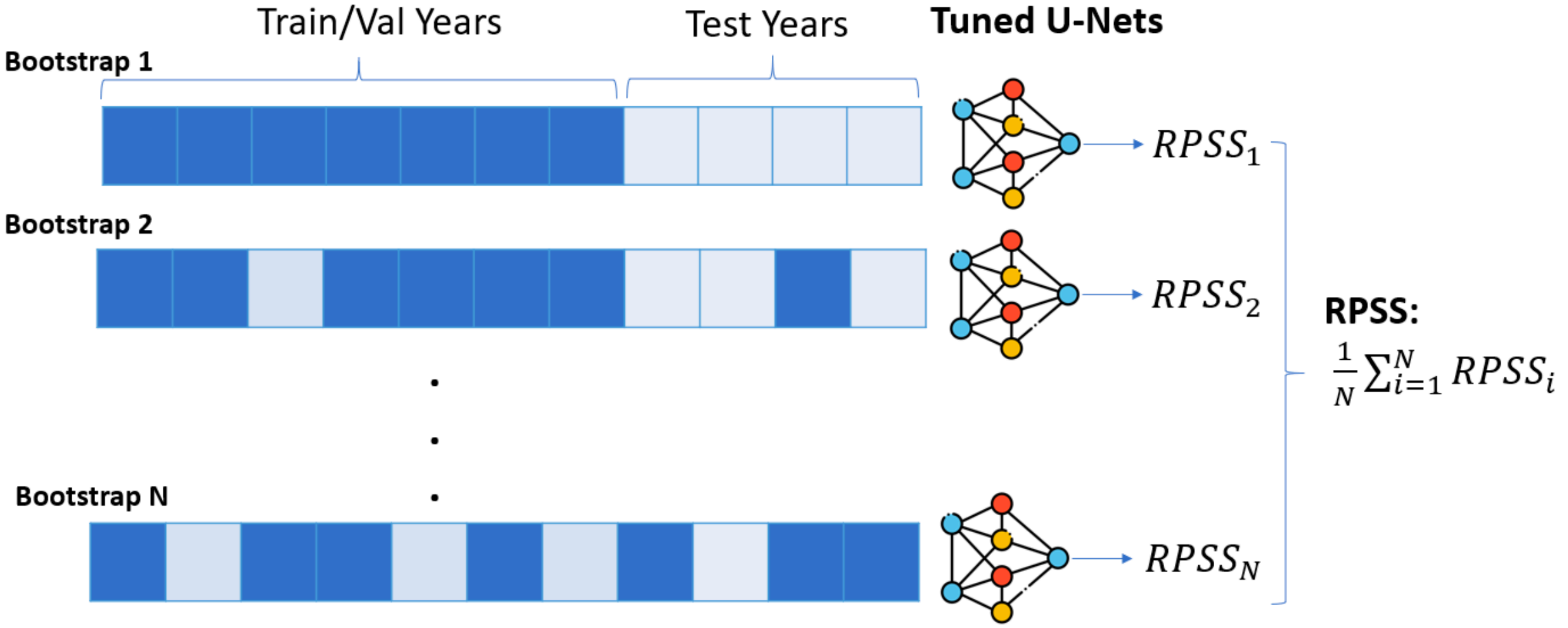
where

$$f = b_0 + b_1\overline{x}_{\text{ens}} \quad \text{and} \quad g = b_2 q.$$

Training/Validation/Test

70%/20%/10% proportion of hindcasts

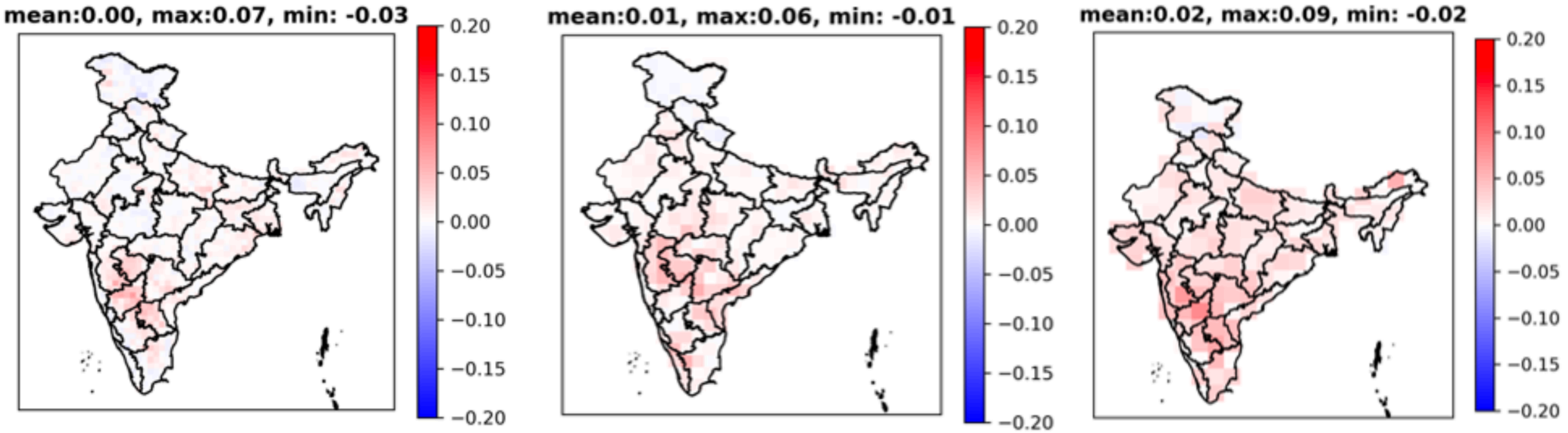
To assess performance, we employ a bootstrap approach, or **Monte-Carlo Cross Validation** (MCCV) with $N = 10$ bootstraps, to compute individual skill scores on fine-tuned U-Nets before averaging them.



Weeks 3-4 Individual model RPSS Skill - full hindcast periods

Baseline ELR

IITM | GEFSv12 | ECMWF

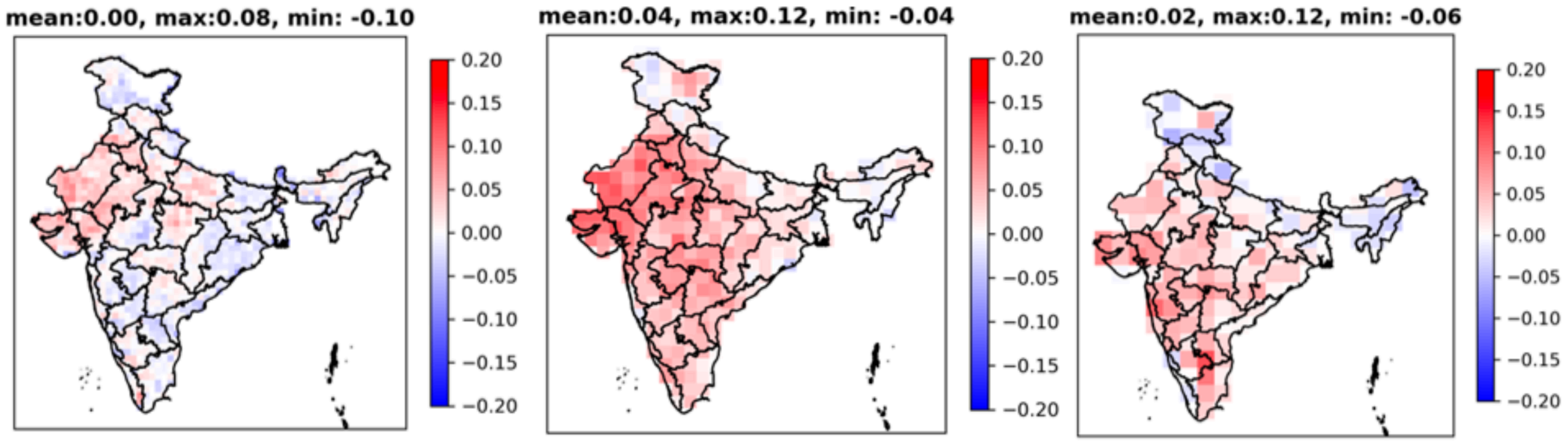


(a) ELR

U-Net beats baseline RPSS

U-Net

IITM | GEFSv12 | ECMWF



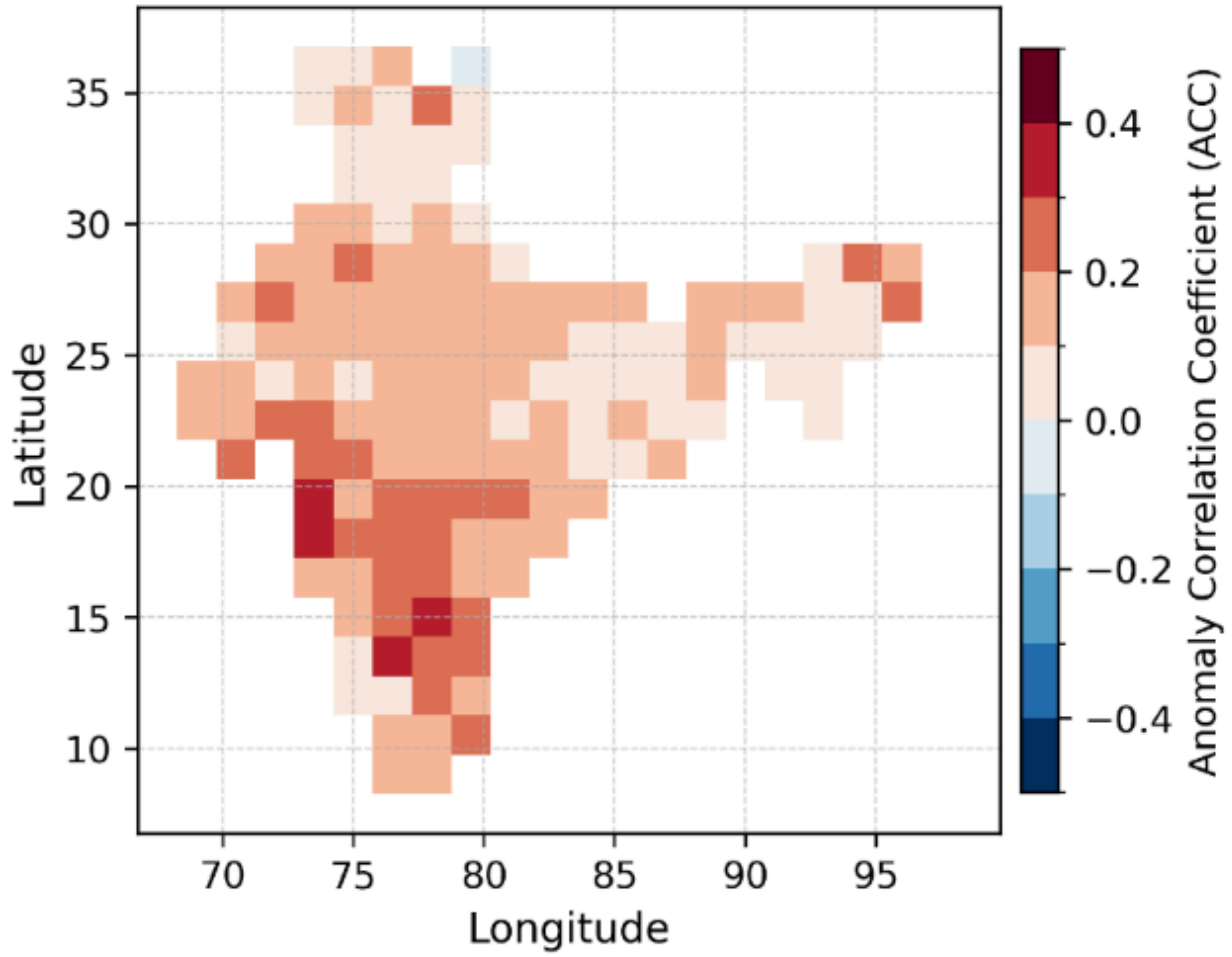
(b) UNET

2003-2018 | 1989-2018 | 2003-2022

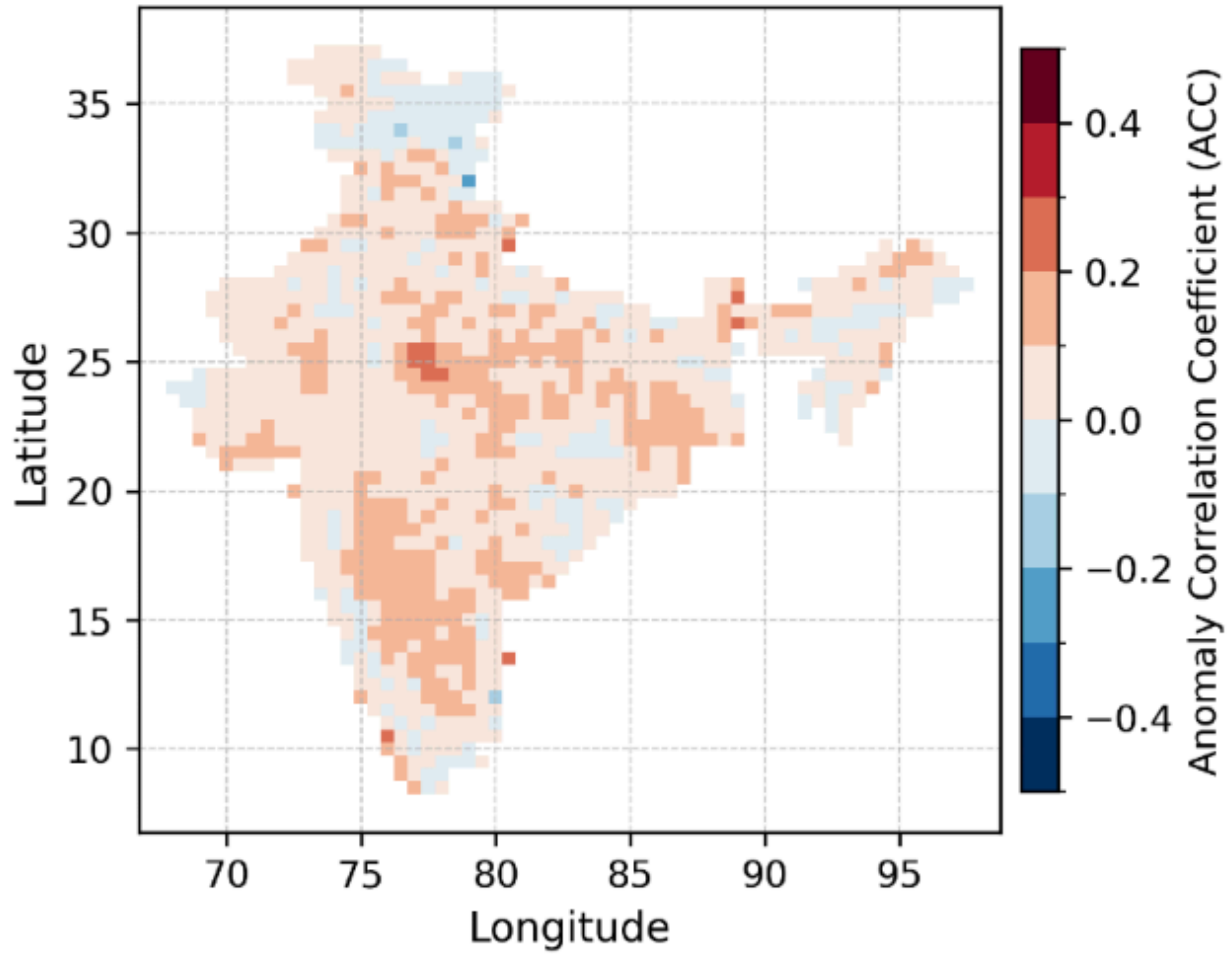
Models with longest hindcasts have highest RPSS

Weeks 3-4 Anomaly Correlation Skill

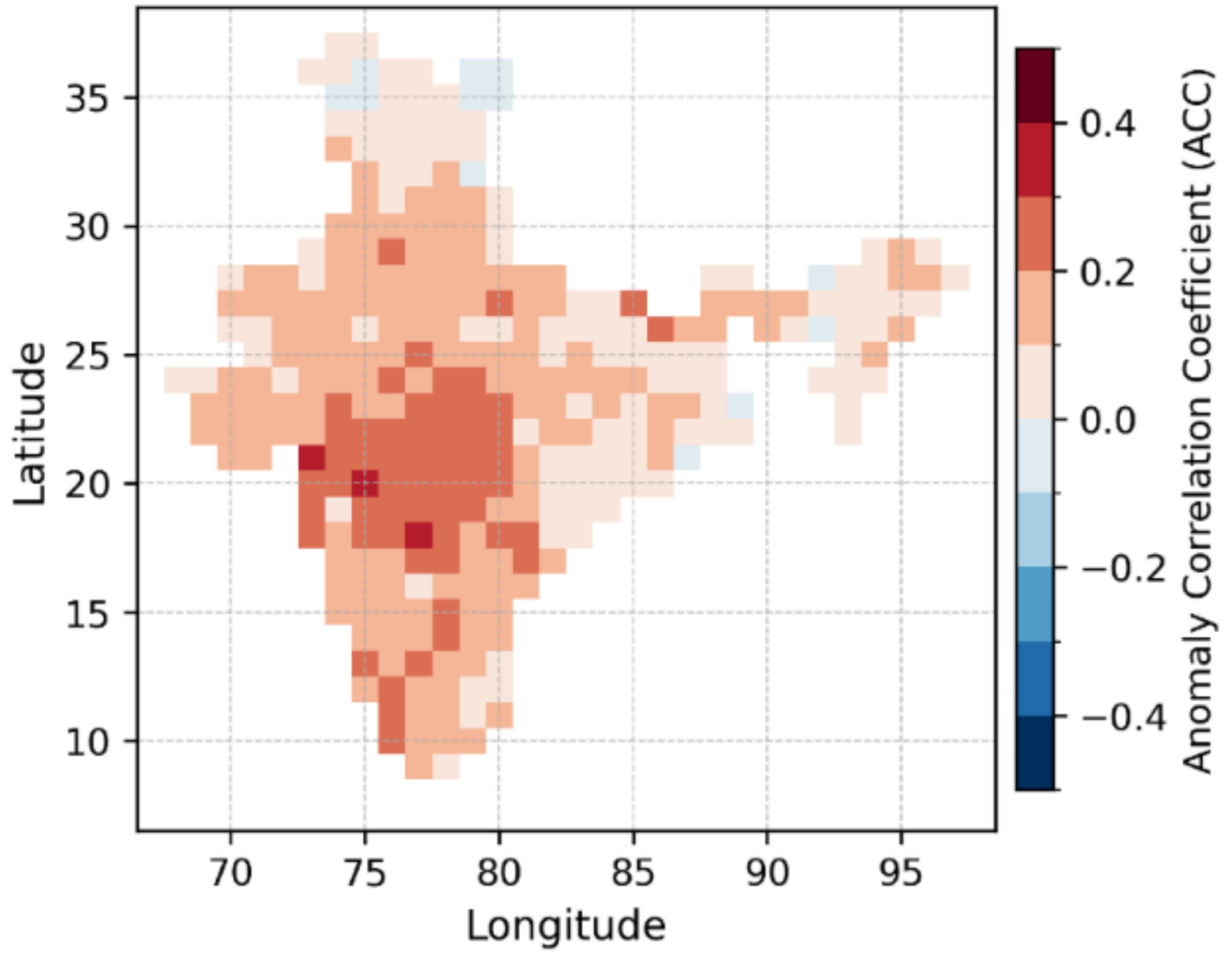
Based on weekly anomalies



(a) ECWMF (2003-2022



(b) IITM (2003-2018)

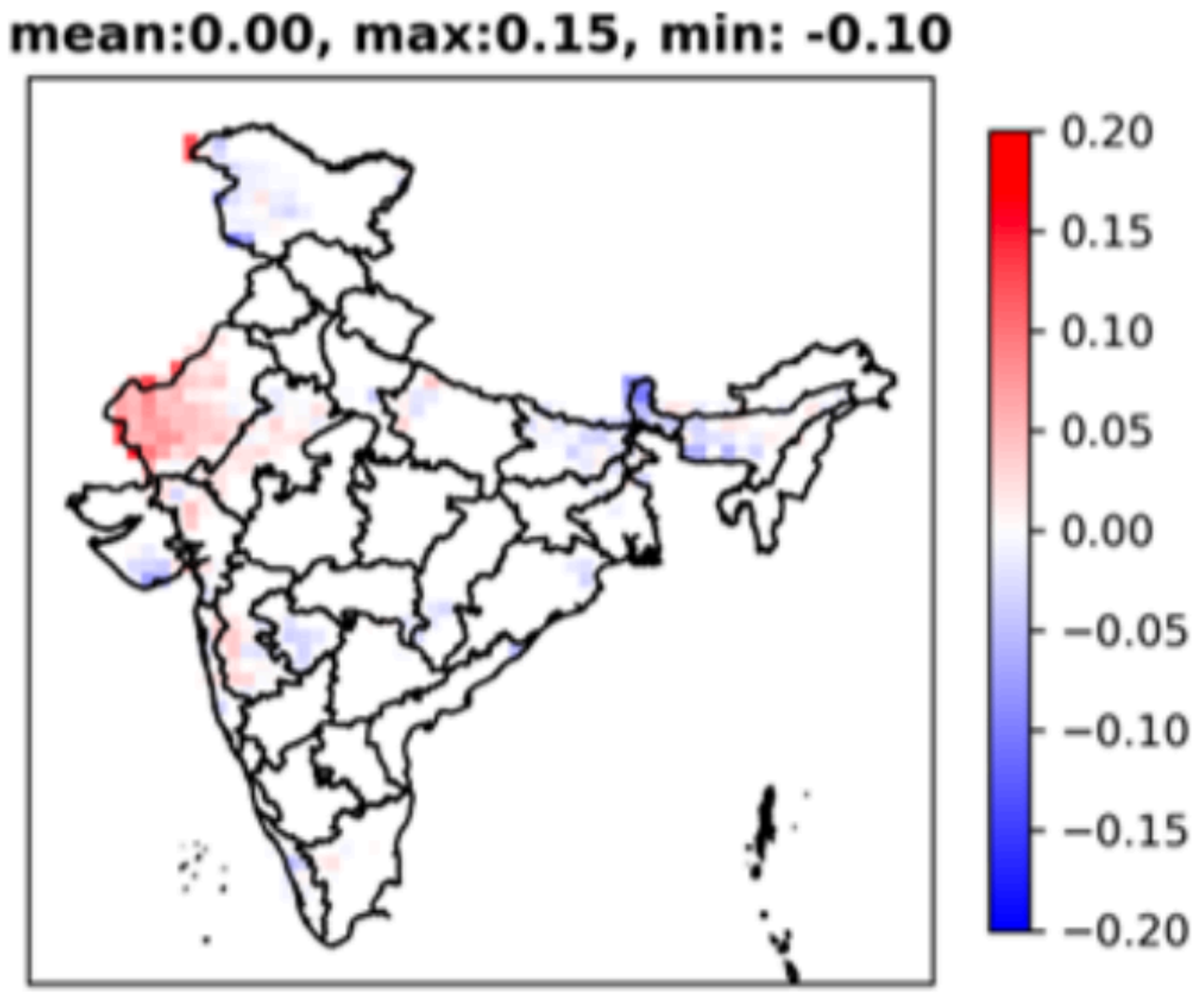


(c) GEFS (1989-2018)

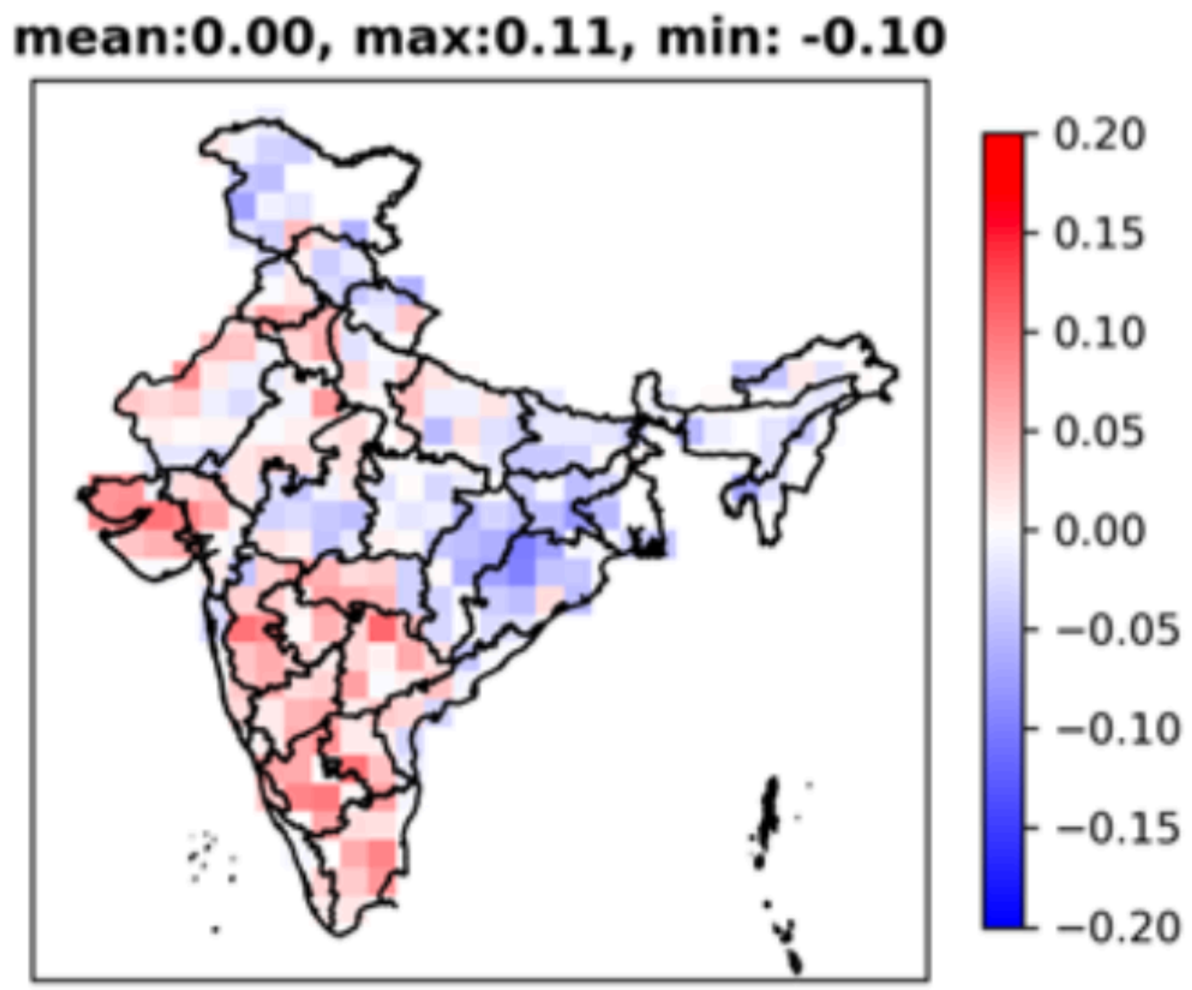
Relative RPSS performance of GCM+UNet is consistent with the simple anomaly correlation.

U-Net Multi-model ensemble Weeks 3-4 RPSS Skill Common 2003-2018 Period

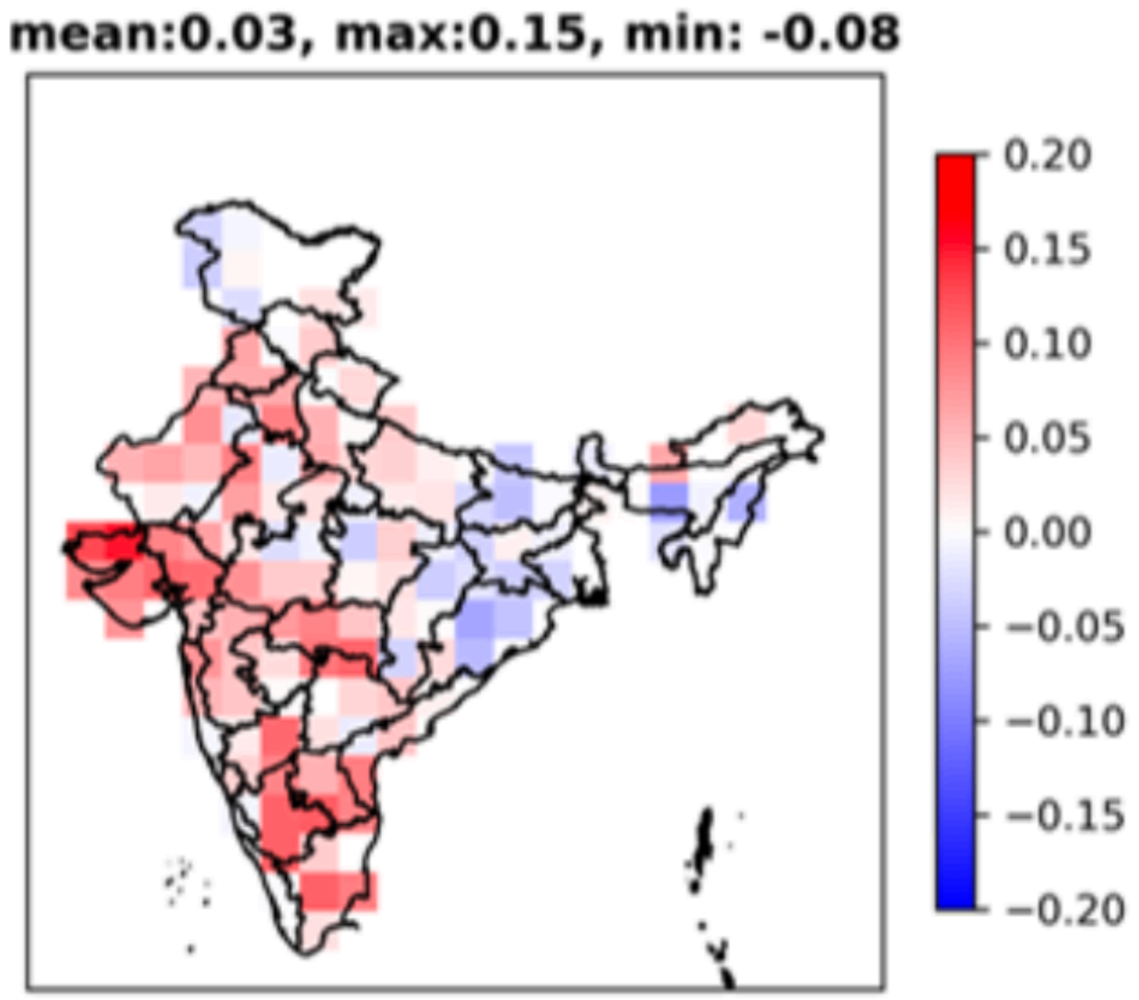
IITM



GEFSv12

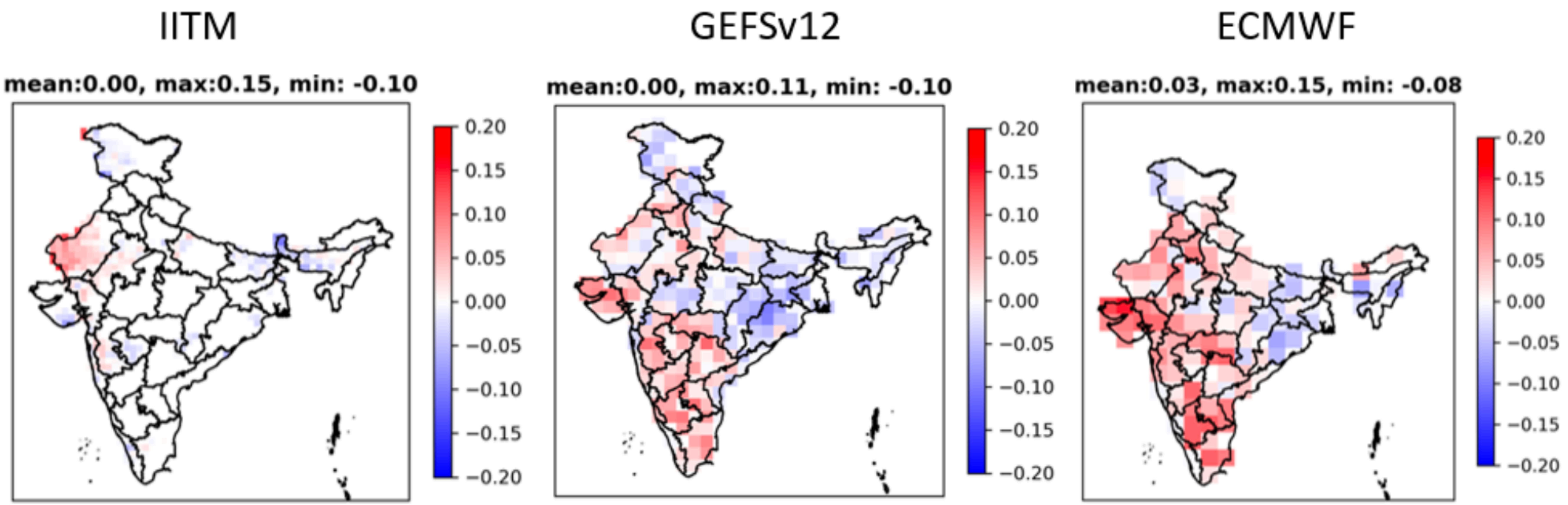


ECMWF



Common set of hindcasts (2003-2018) and a common 1 × 1 spatial resolution

U-Net Multi-model ensemble Weeks 3-4 RPSS Skill Common 2003-2018 Period



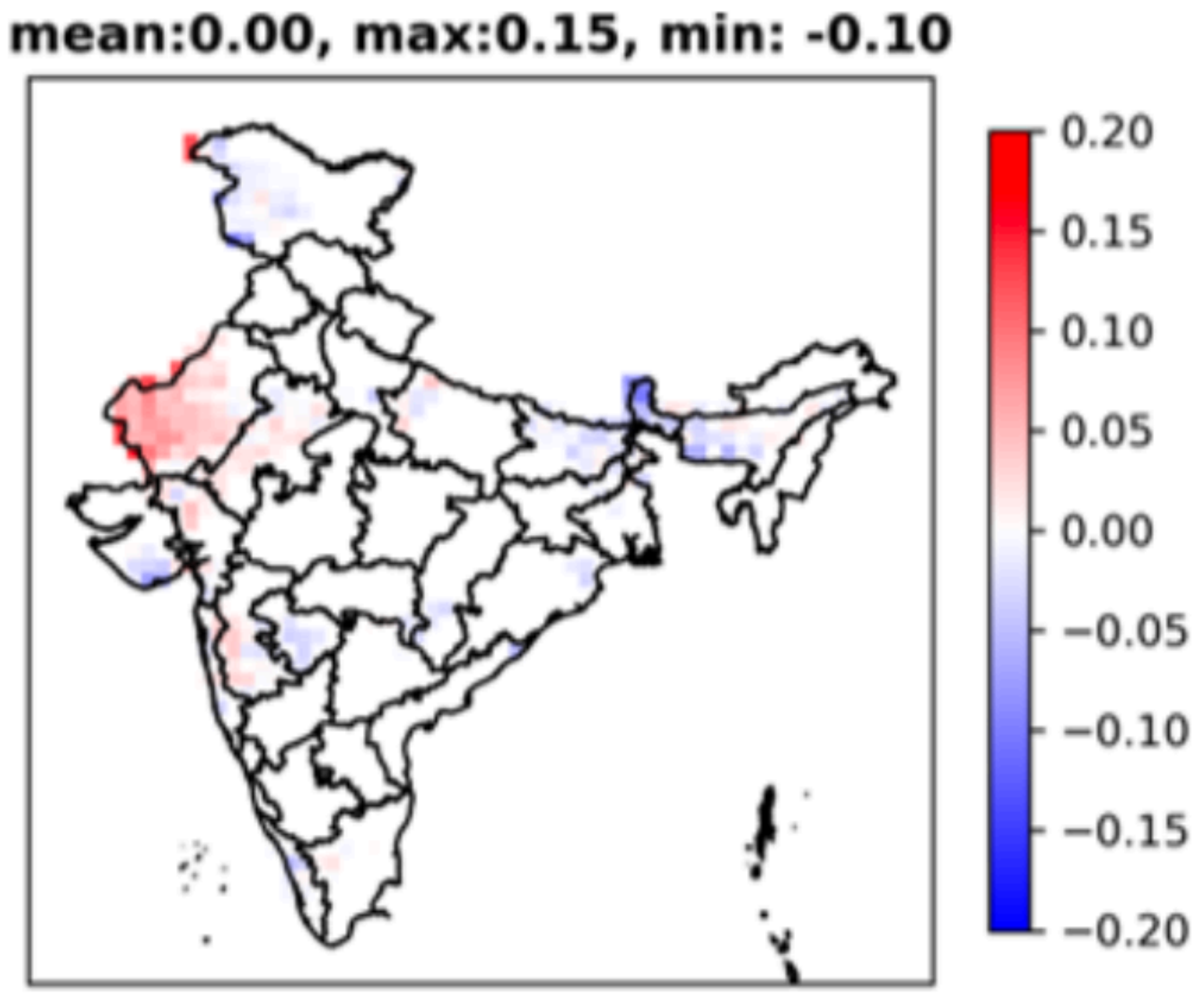
Common set of hindcasts (2003-2018) and a common 1 × 1 spatial resolution

We aggregate each individual model’s forecasts by averaging their output probabilities and normalizing

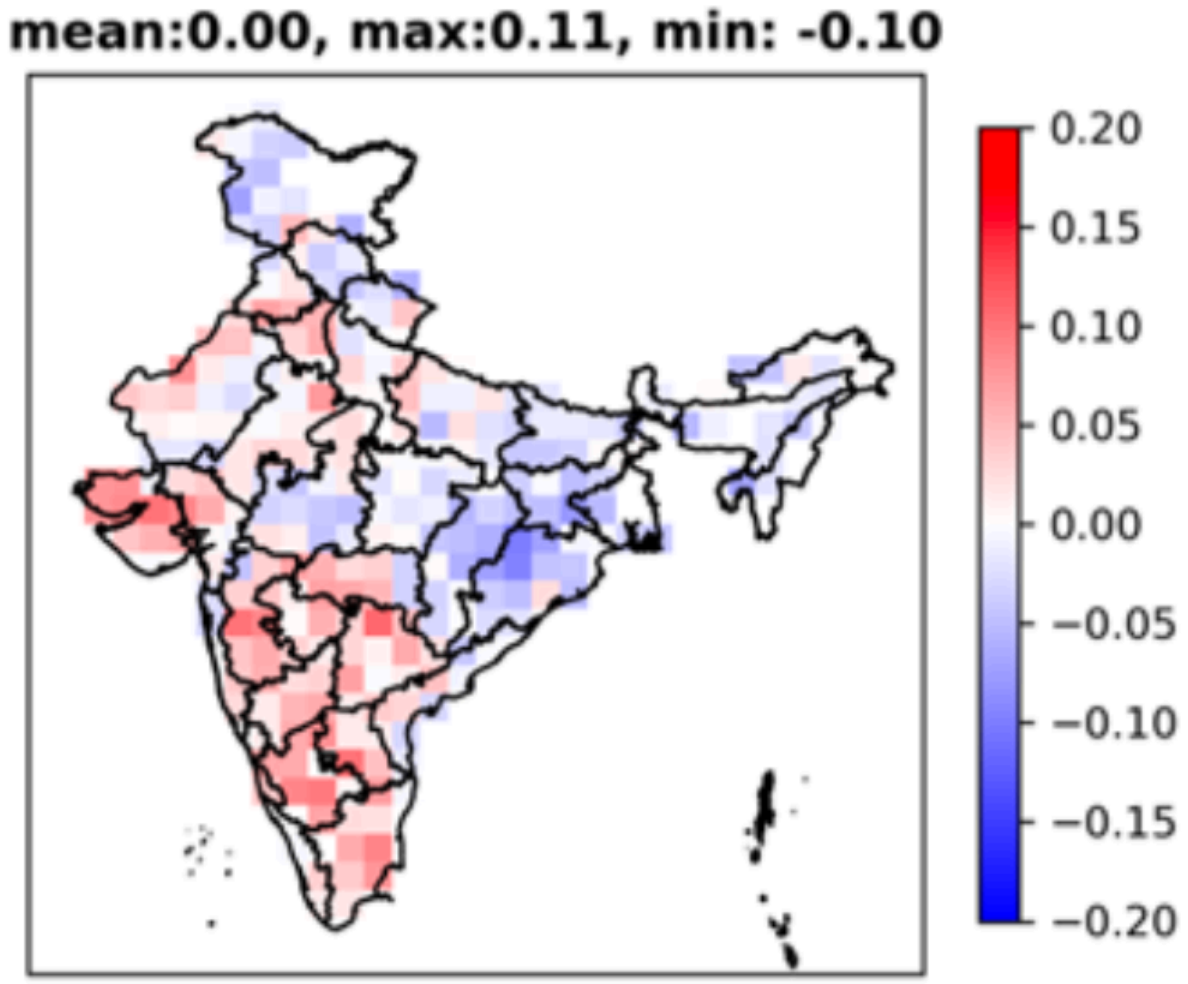
U-Net Multi-model ensemble Weeks 3-4 RPSS Skill Common 2003-2018 Period

The MME performs on par with the best individual model for every lead time and architecture. (UNet or ELR). Also gives more positive RPSS values across the domain.

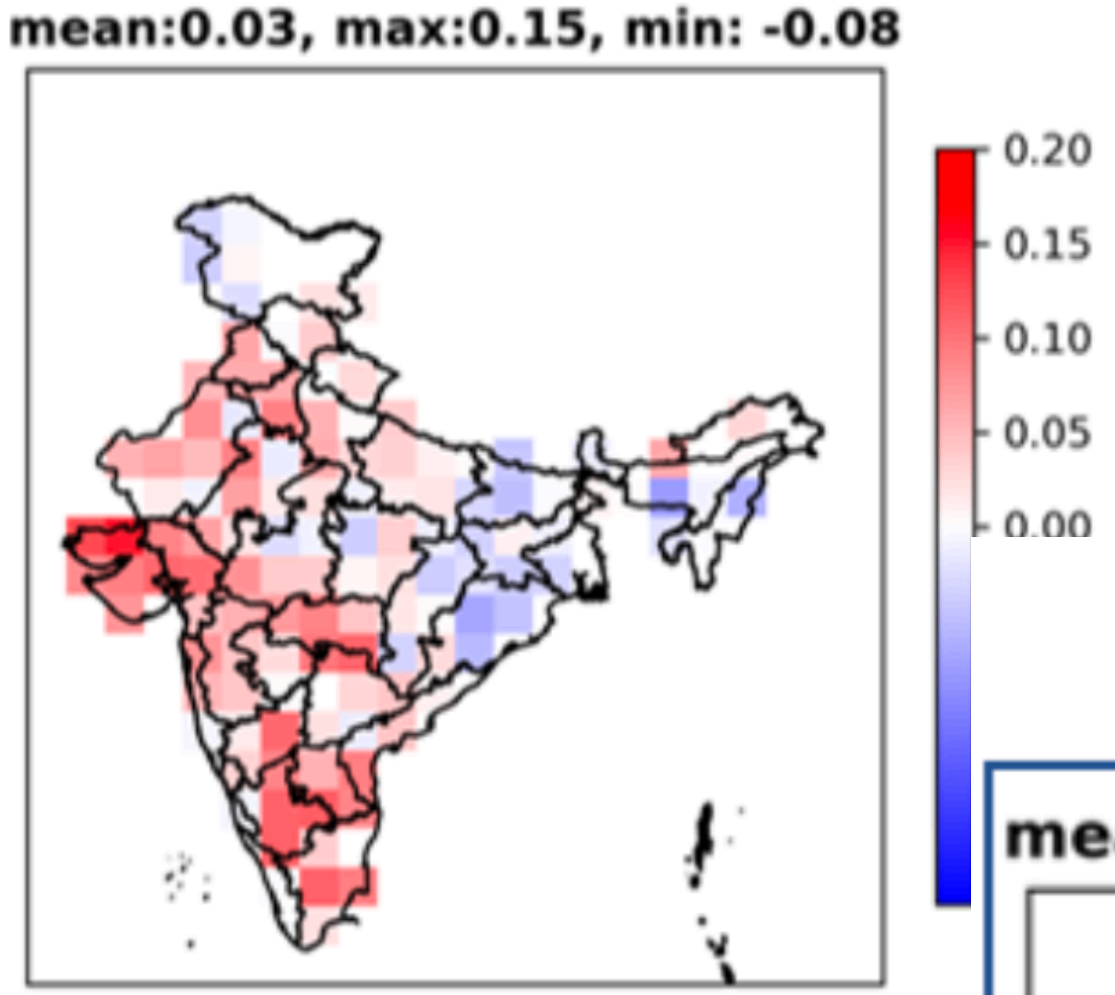
IITM



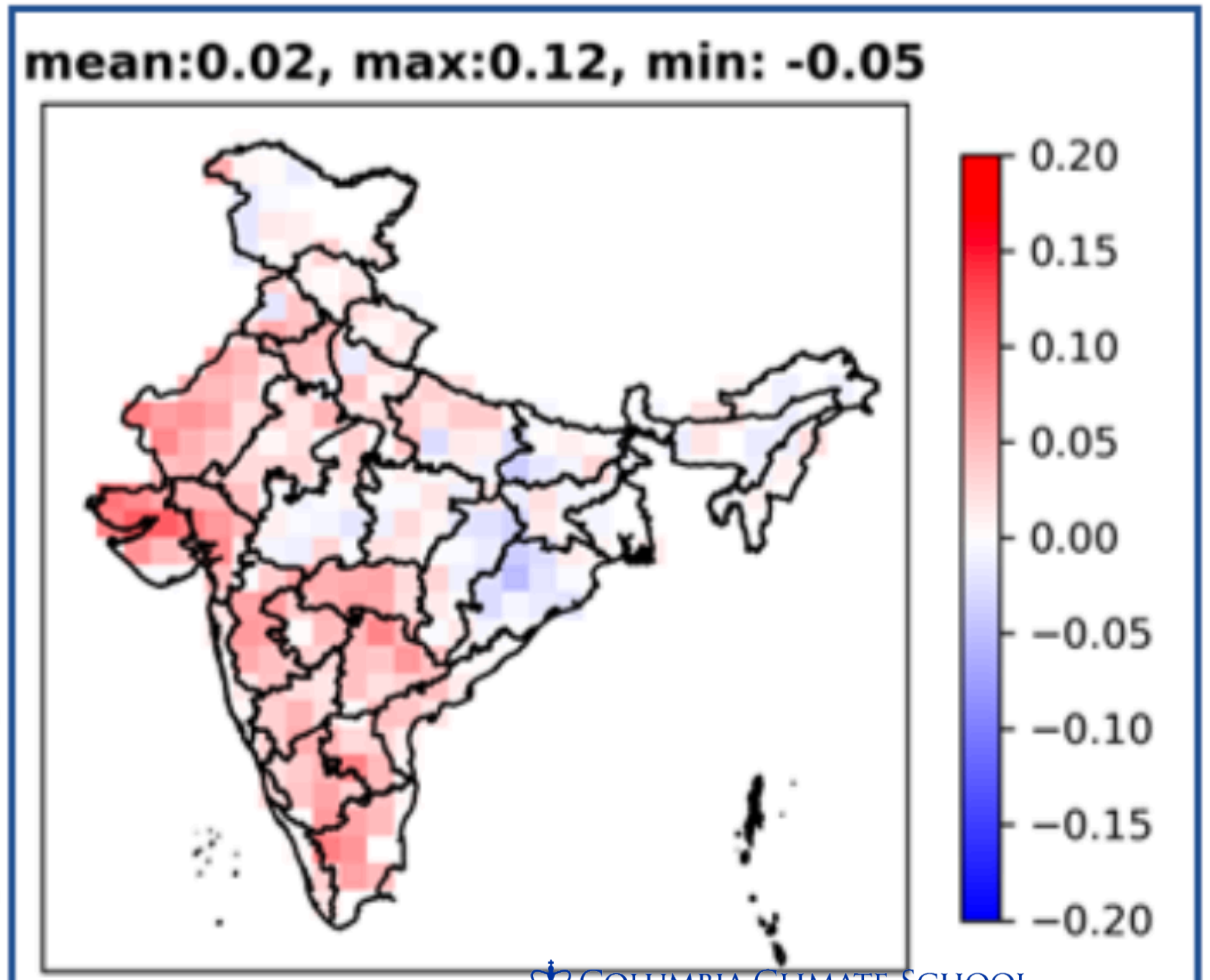
GEFSv12



ECMWF



Multi-Model

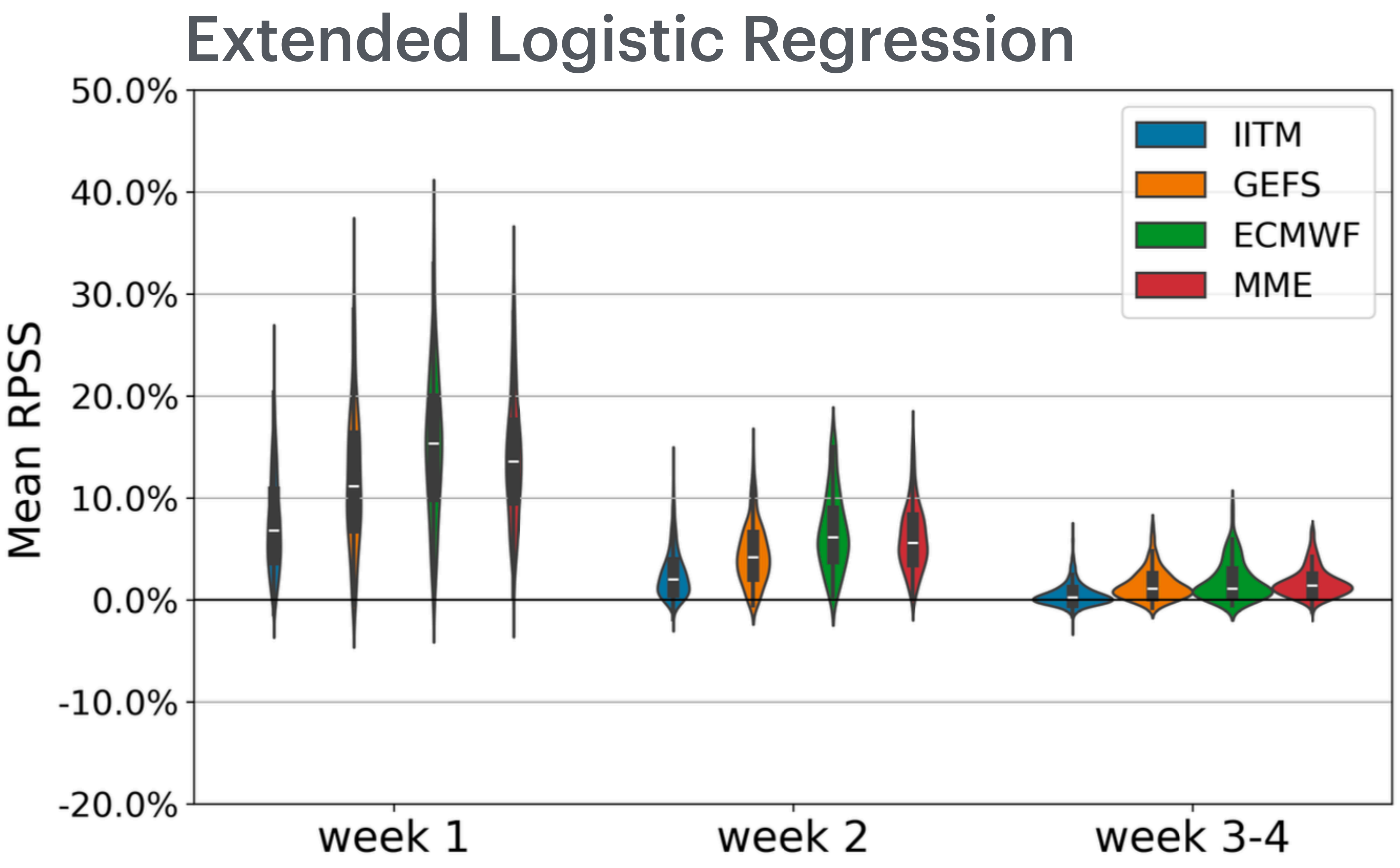


Common set of hindcasts (2003-2018) and a common 1 × 1 spatial resolution

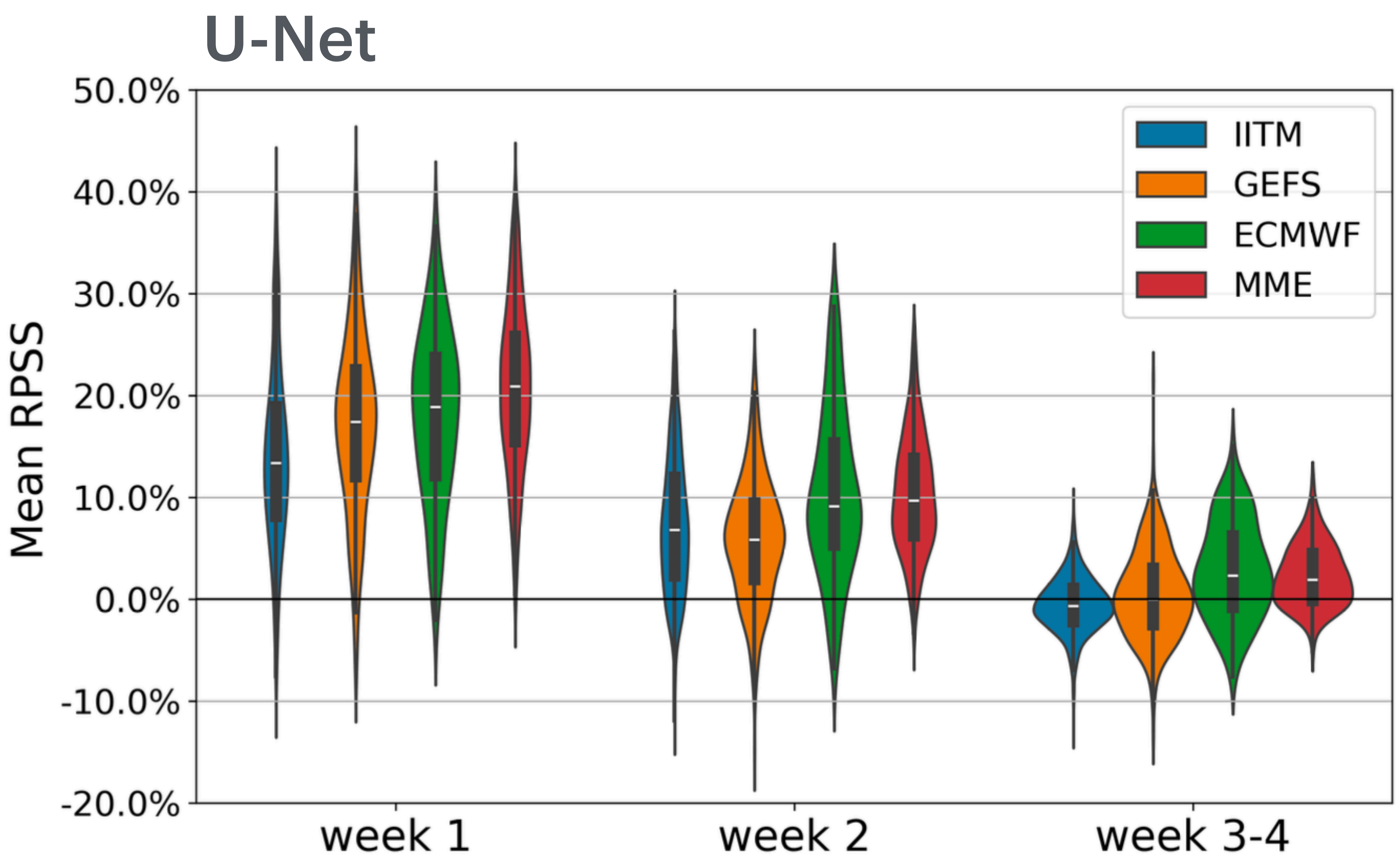
We aggregate each individual model's forecasts by averaging their output probabilities and normalizing

RPSS Gridpoint Distributions

Averages over 10 bootstrap samples



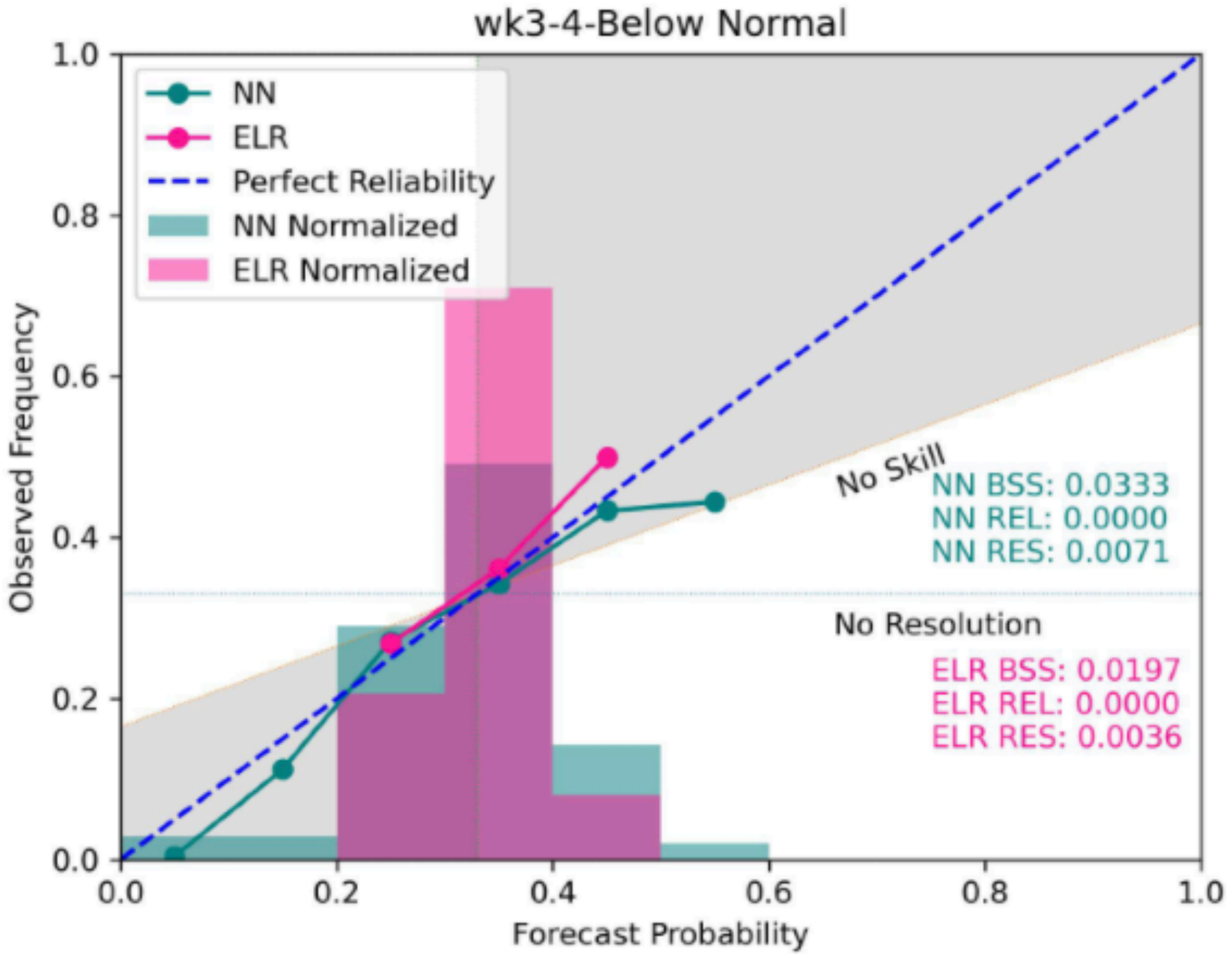
(a) ELR



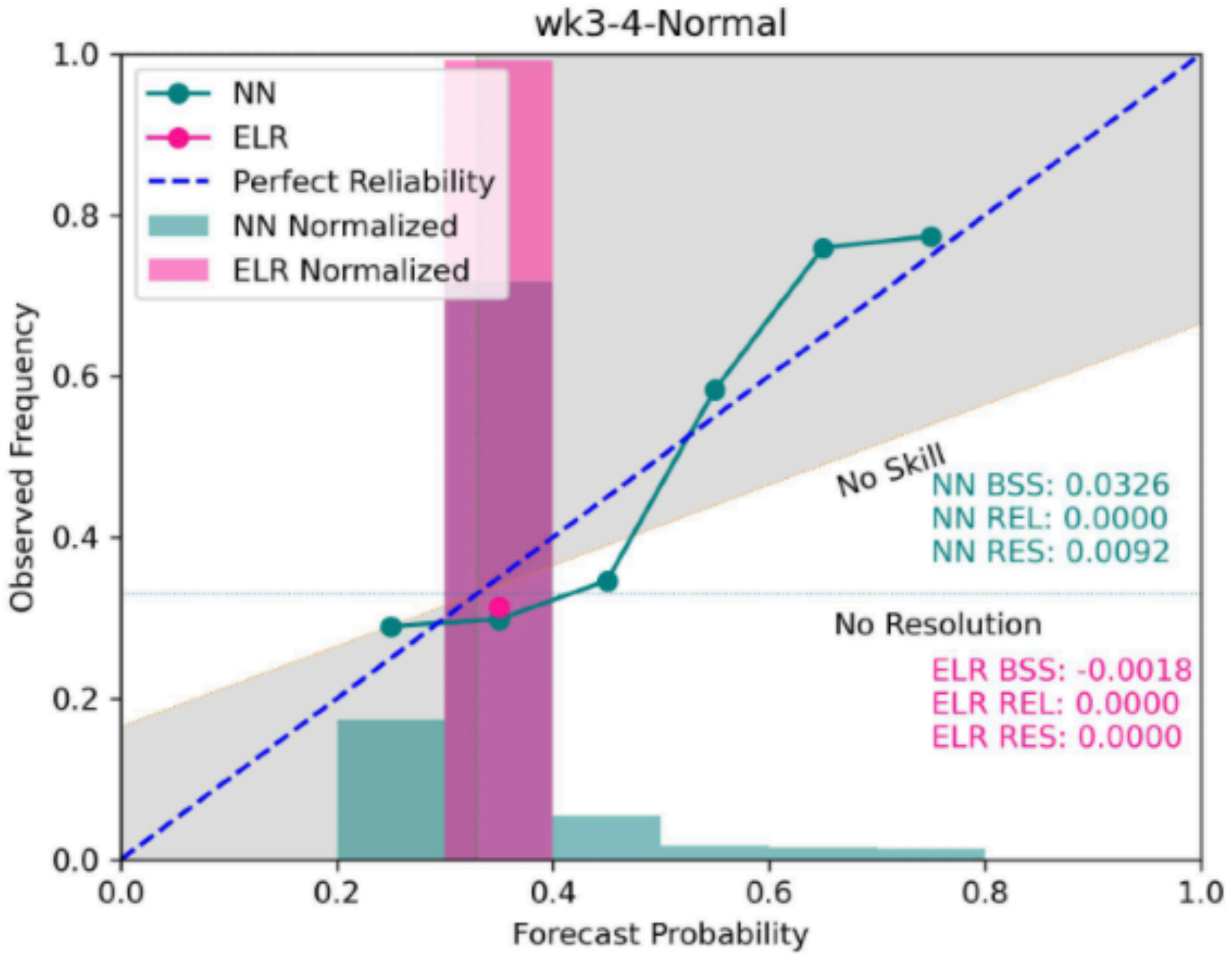
(b) UNet

MME skill exceeds (or equals) best individual model at all lead times.
The U-Net increases the skill at many points, but decreases at others.

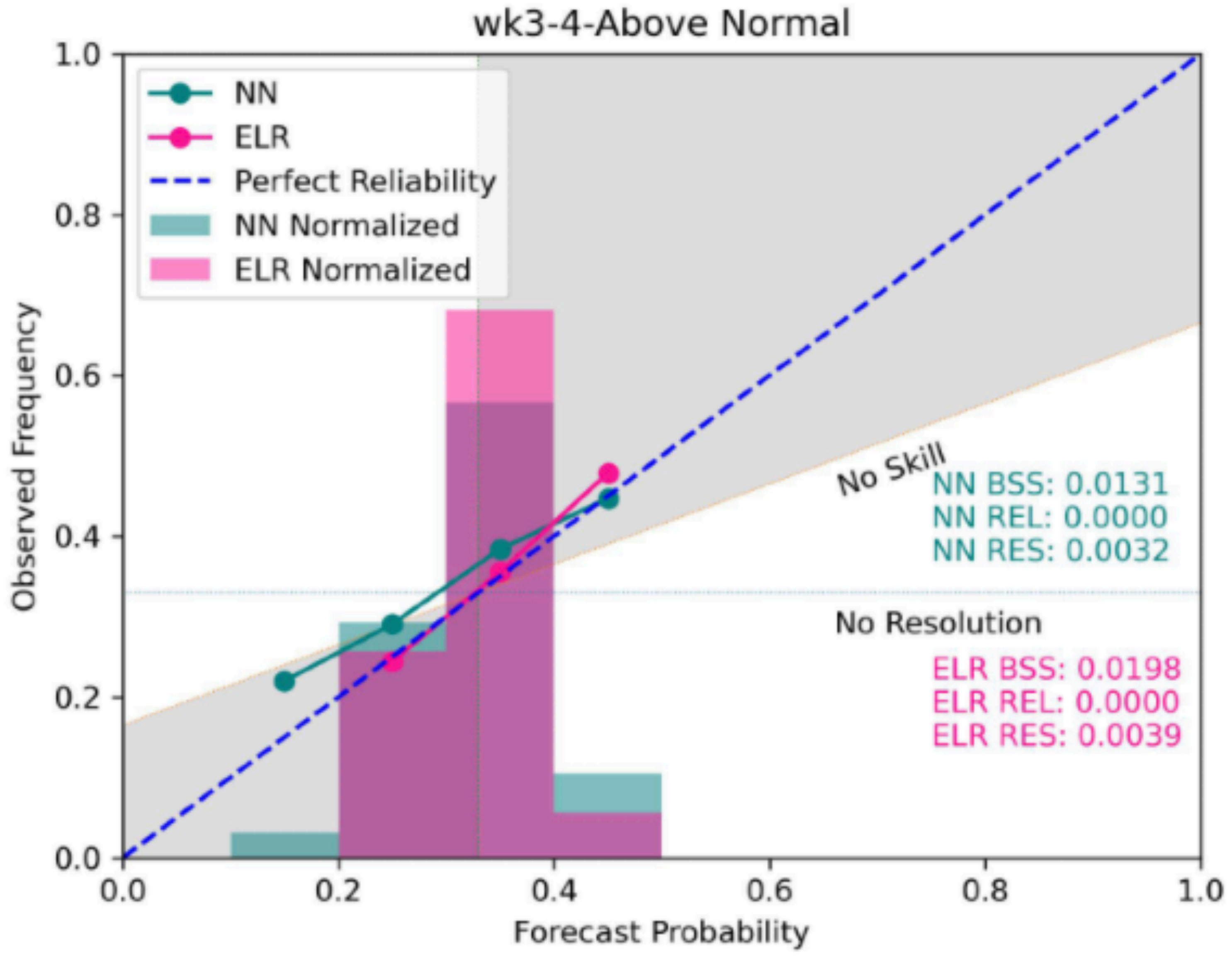
Weeks 3-4 Hindcast Reliability



(g) Week 3–4: Below Normal



(h) Week 3–4: Normal



(i) Week 3–4: Above Normal

U-Net forecasts are sharper than the baseline ELR
Both U-Net and ELR have good and comparable reliability.

Example MME U-Net Weeks 3-4 Forecast

Issued 14 Jul 2023 - Target Period 28 Jul-11 Aug

Extended Logistic Regression **U-Net** **Obs**

Every model tuned on its full hindcast period using a train/validation split of 80%/20%

Conclusions

- U-Net outperforms Extended Logistic Regression (ELR) in terms of calibration skill, despite challenges related to the relatively short hindcast data records
- The multi-model ensemble (MME) further enhances the forecast skill, achieving performance comparable to the best individual model and showing consistent improvements across the domain.
- Real-time forecasts for the summer of 2023 revealed both successful predictions and areas for improvement