Improving medium-range ensemble forecasts with transformers

JONATHAN WEYN (MICROSOFT, ECMWF)

ZIED BEN BOUALLEGUE, MATTHEW CHANTRY, PETER DEUBEN, JESPER DRAMSCH (ECMWF)

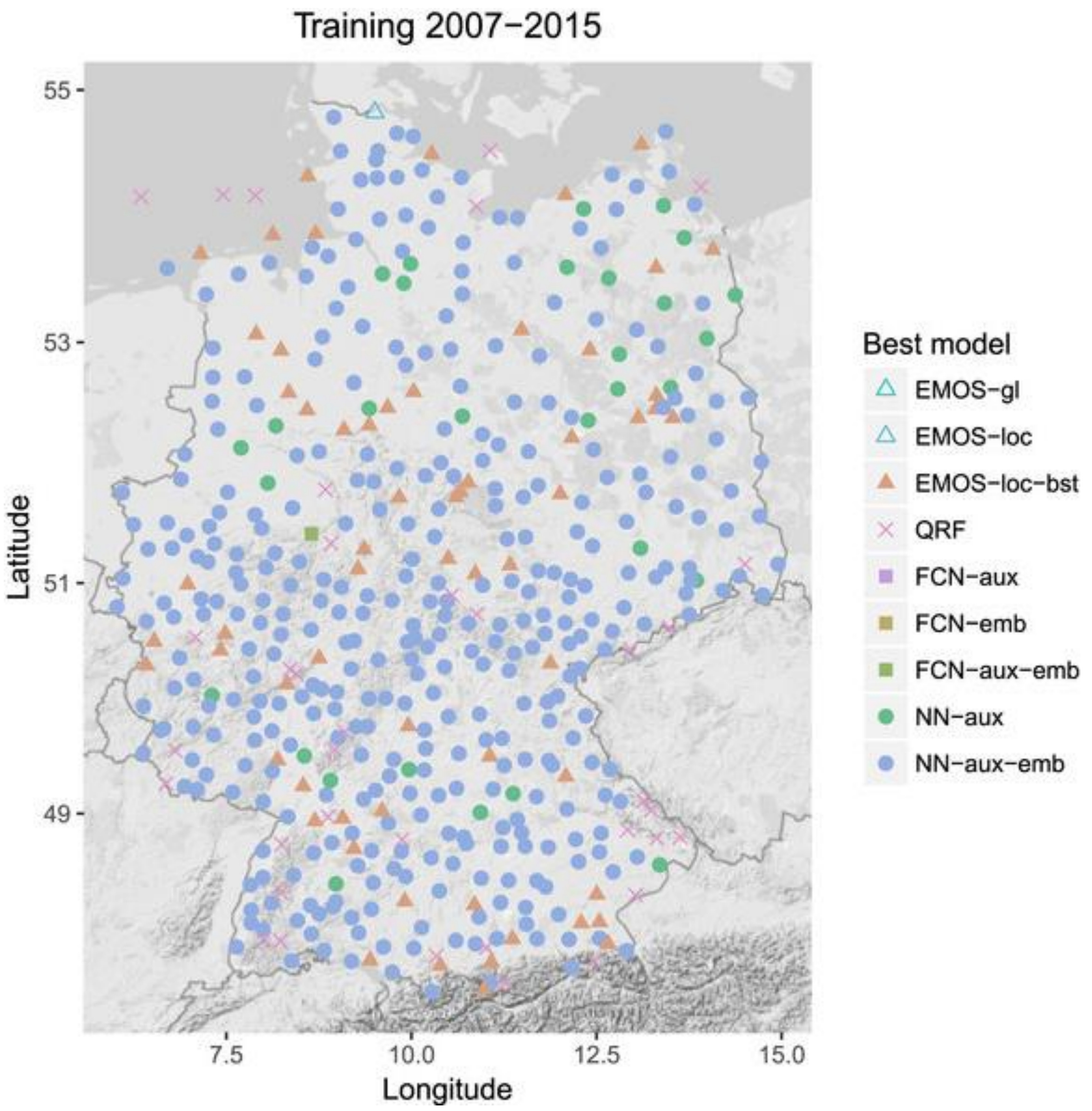
THANKS: TOBIAS FINN (UNIV. OF HAMBURG, MPI)

Can deep learning provide *global, high-resolution, full-ensemble* calibrated medium-range weather forecasts?

- Ensemble numerical weather forecasts still benefit from statistical post-processing
- Methods such as Ensemble Model Output Statistics (EMOS) are rigid, and generally restricted to predicting distribution parameters
- Machine learning methods have successfully been applied to improve predictions of ensemble distribution parameters (usually mean and std)
- Recent introduction of a self-attentive ensemble transformer offers to provide full ensemble distributions of calibrated forecasts

Relevant work

- Rasp and Lerch, “Neural Networks for Postprocessing Ensemble Weather Forecasts” (2018)
    - Neural networks trained to predict mean and std of an ensemble forecast
    - Non-linear NN is the best model at most validation stations across Germany



Relevant work

- Gronqvist et al., “Deep Learning for Post-Processing Ensemble Weather Forecasts”, 2020
    - U-net residual NN architecture applied on global 0.5-degree fields from ECMWF hindcasts
    - Models for bias correction (mean) and uncertainty quantification (std)
    - Able to exceed performance of 10-member hindcast with 5-member input

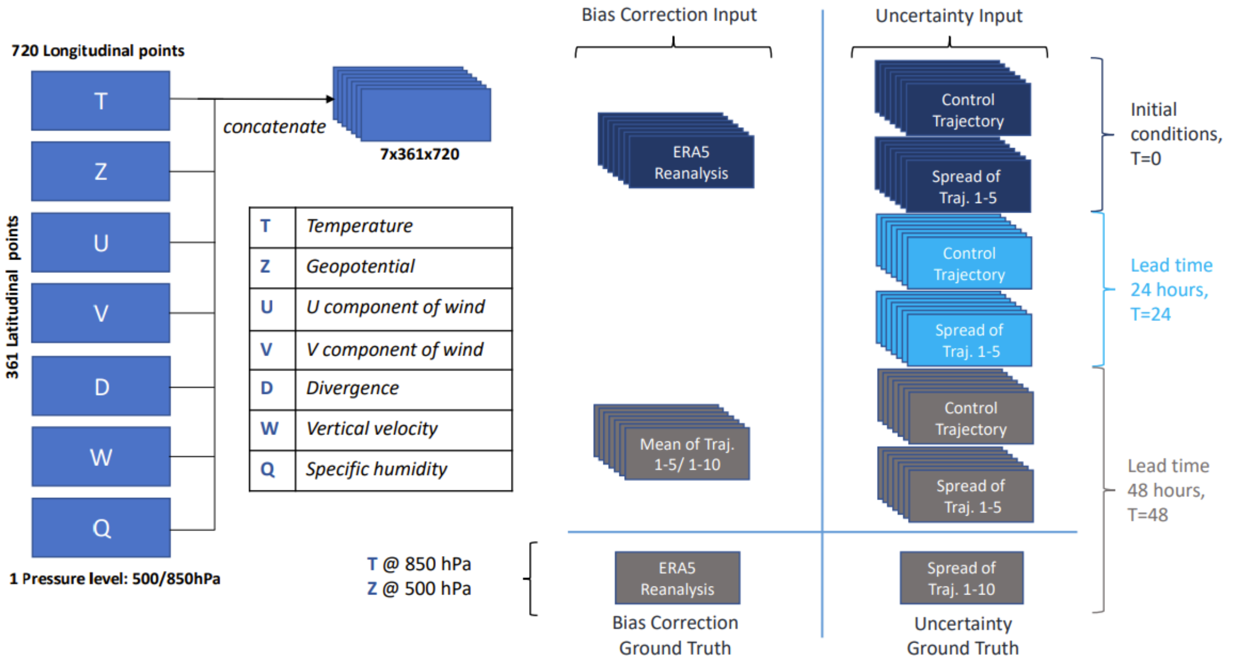
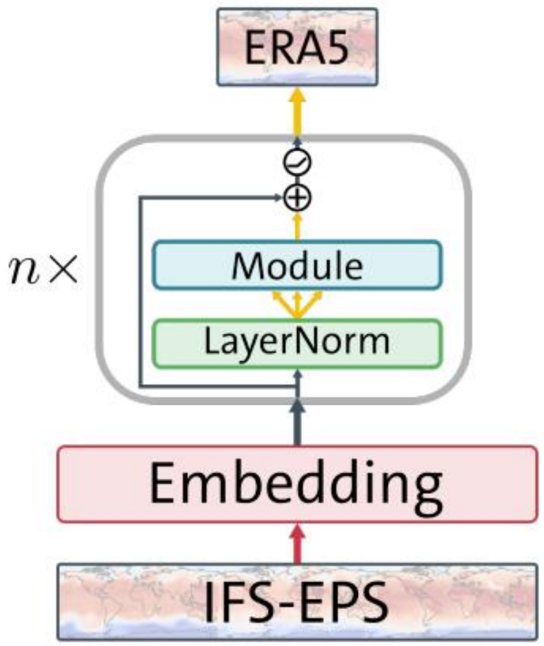


Figure 1: Inputs and ground truths to our neural networks.

The ensemble transformer

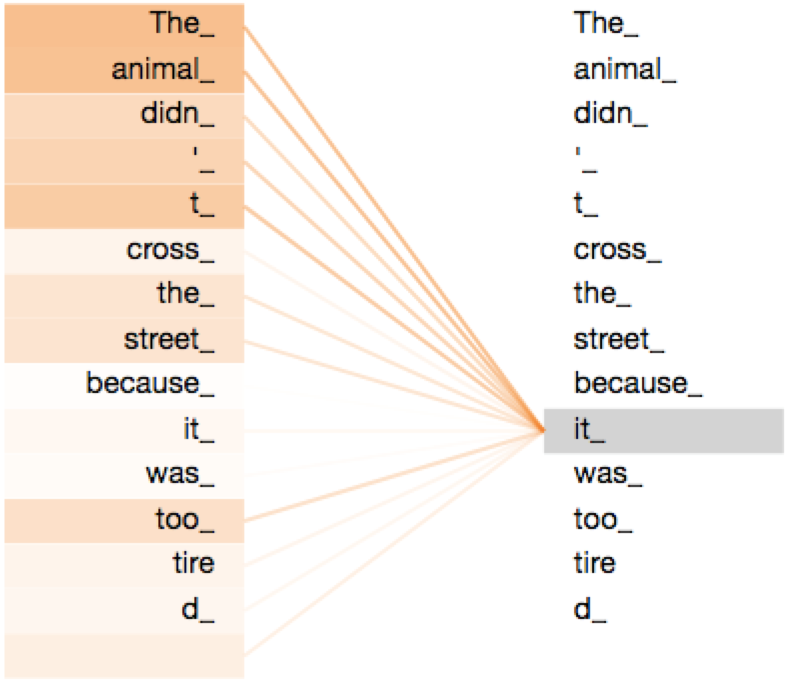
- Finn, “Self-Attentive Ensemble Transformer”, 2021
  - member-by-member approach that applies corrections based on interactions with all other members
  - maintains spatial correlations
  - results for 3-day forecasts of 2-m temperature show 20% improvement in CRPS and spread-skill ratio near 1



Transformer 101

- Developed for natural language processing (NLP) tasks, such as translation
- Unlike recurrent neural networks, computes "attention," or relevance, of each sequence element to every other element

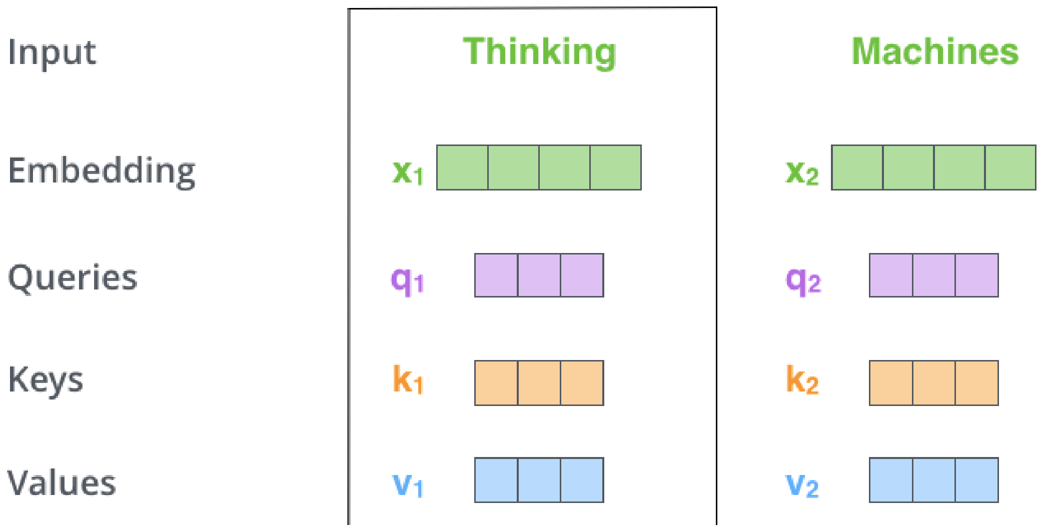
The animal didn't cross the street because it was too tired.



[The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)](https://jalammar.github.io)

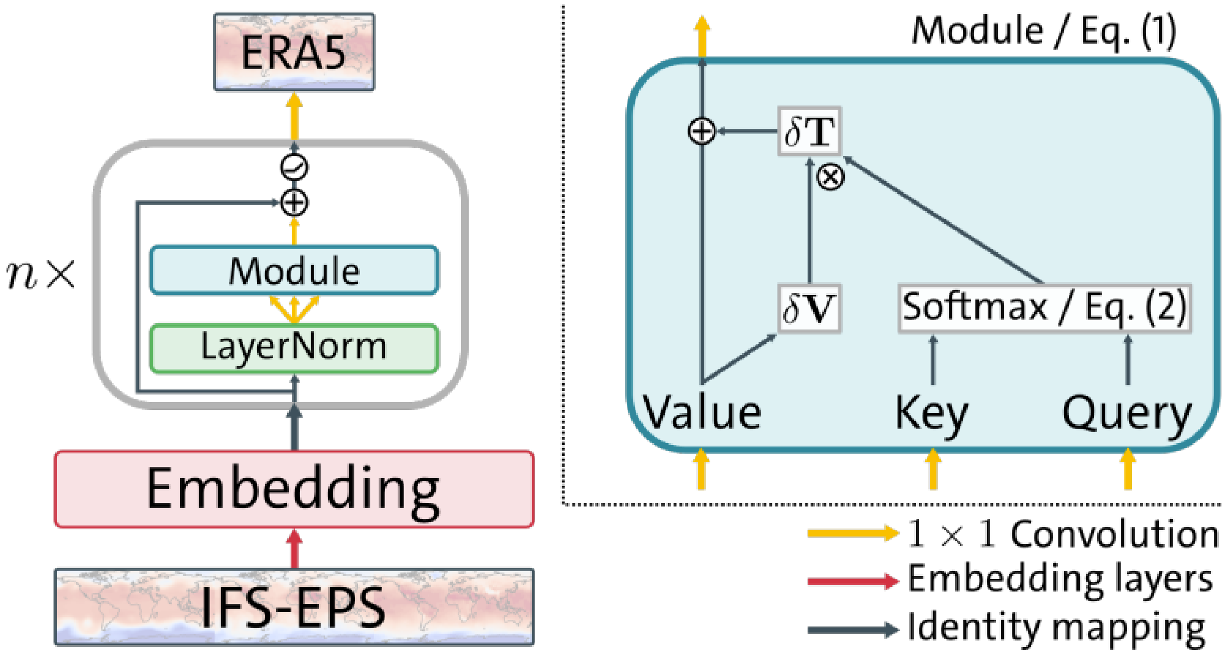
Transformer 101: Self-attention

- Compute a *query*, *key*, and *value* by multiplying each sequential input by arrays of trainable weights
    - key and query initialized as 0 in training
- Compute the similarity (dot-product) between the query of each input to the key of all inputs
- Add to the value of each input the sum of all inputs' values weighted by the normalized similarity

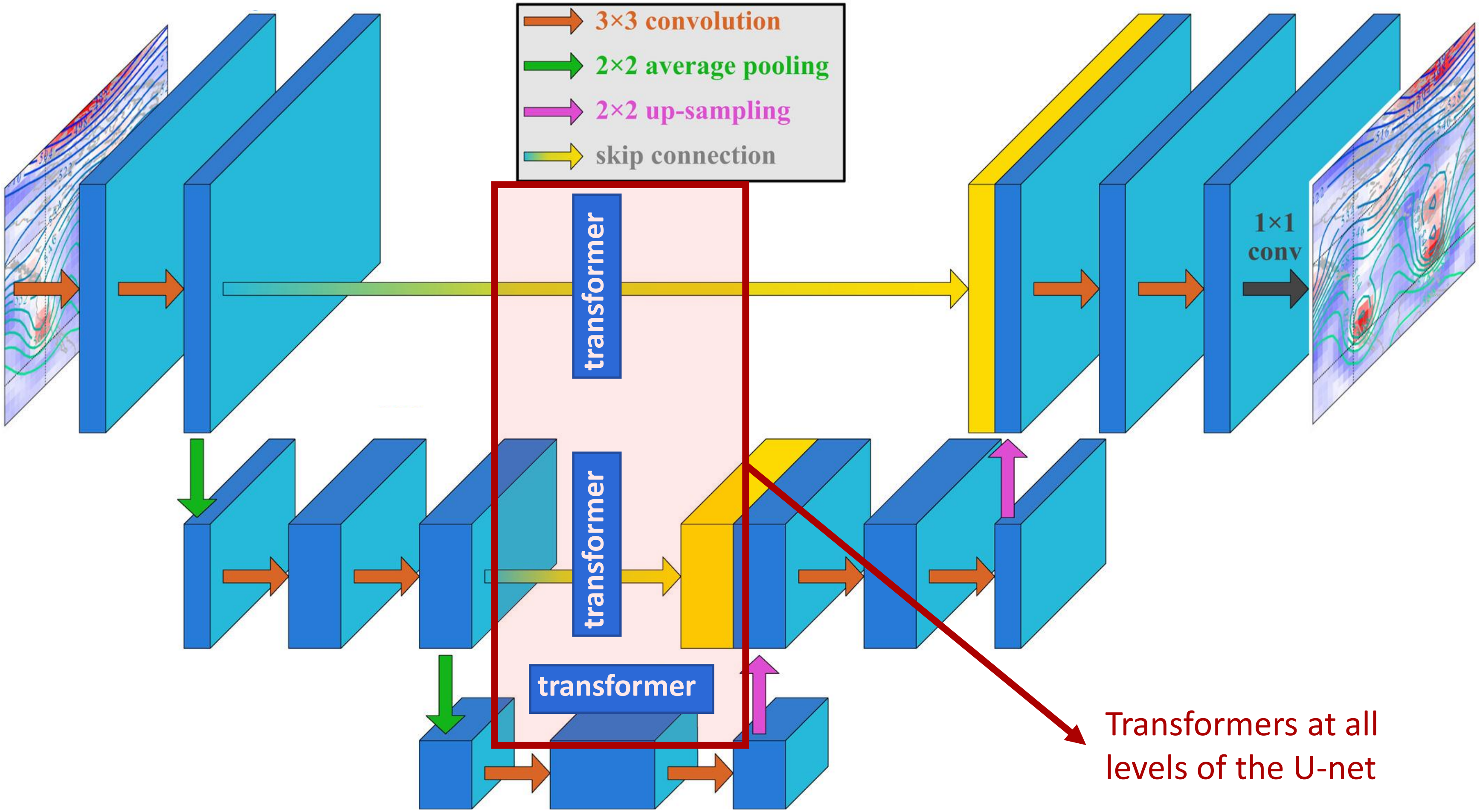


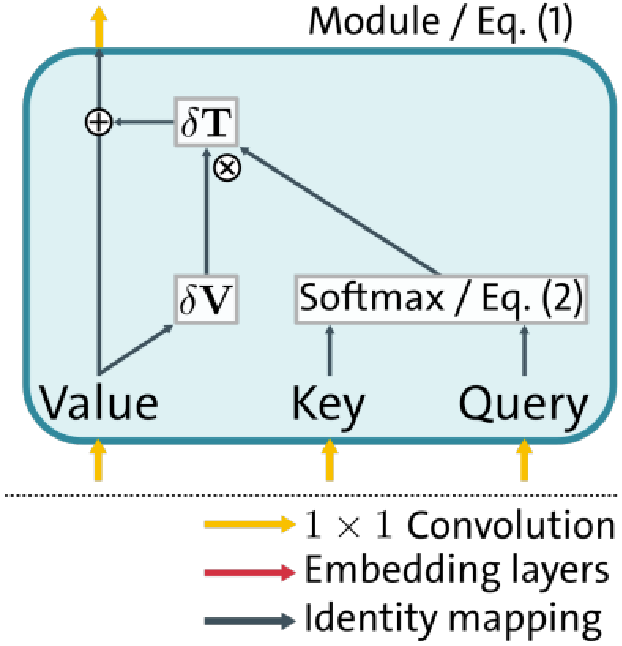
The ensemble transformer

- Apply the transformer along the *ensemble member dimension*
- Values represent an embedding of the original ensemble member
- Keys and queries are used to compute similarity between ensemble members *on a global scale*



PoET: a transformer U-net





Training framework

- Training data consists of 11-member (10 perturbed + 1 control) hindcasts from 2000-2016
- Methodology is agnostic to the number of ensemble members, so testing is done on the full 51-member (50 perturbed + 1 control) operational IFS twice weekly in 2021
- Selection of either 3 or 6 variables, both including prescribed fields
    - 2-m temp, Z500, T850, land-sea mask, topography, TOA insolation, U700, V700, total cloud cover
- Each training example contains the full 11-member hindcast for a single lead time, with lead times every 6 h up to 96 h
- Target analysis data is ERA5 reanalysis 2-m temperature; loss function is CRPS

Benchmark

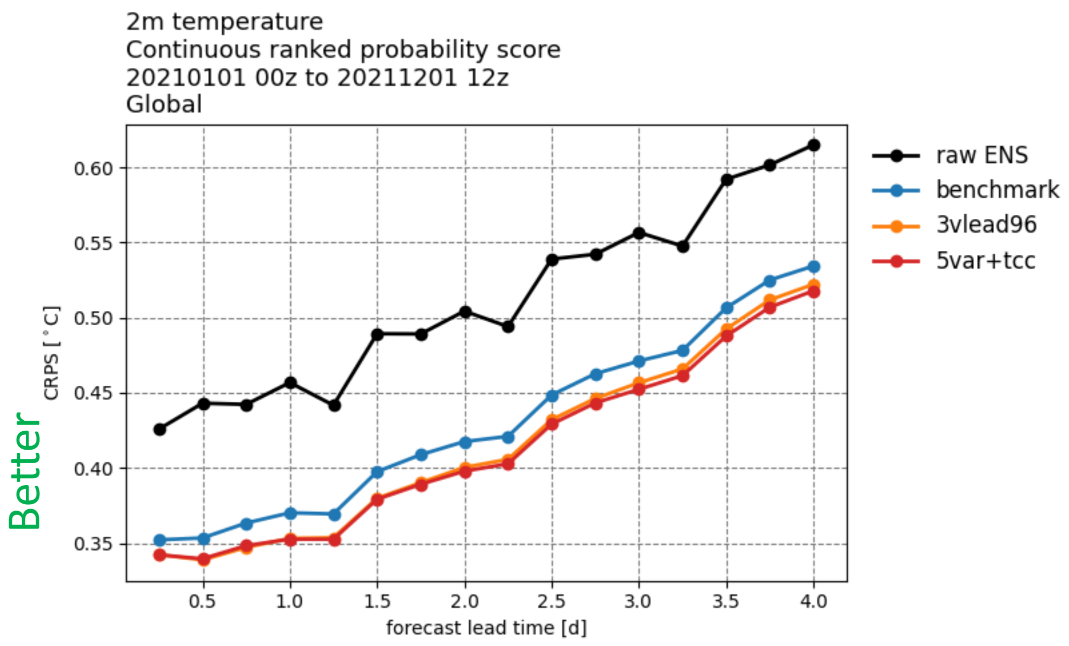
- Member-by-member (MBM) sophisticated linear post-processing approach (Schaeybroeck and Vannitsem, 2015)
- Maximum likelihood estimation of 3 parameters
  - ensemble mean scaling & nudging
  - ensemble spread scaling
- Constrained to preserve weak ensemble reliability (spread = skill) and climatological reliability (forecast variability = observation variability)

PoET vs MBM

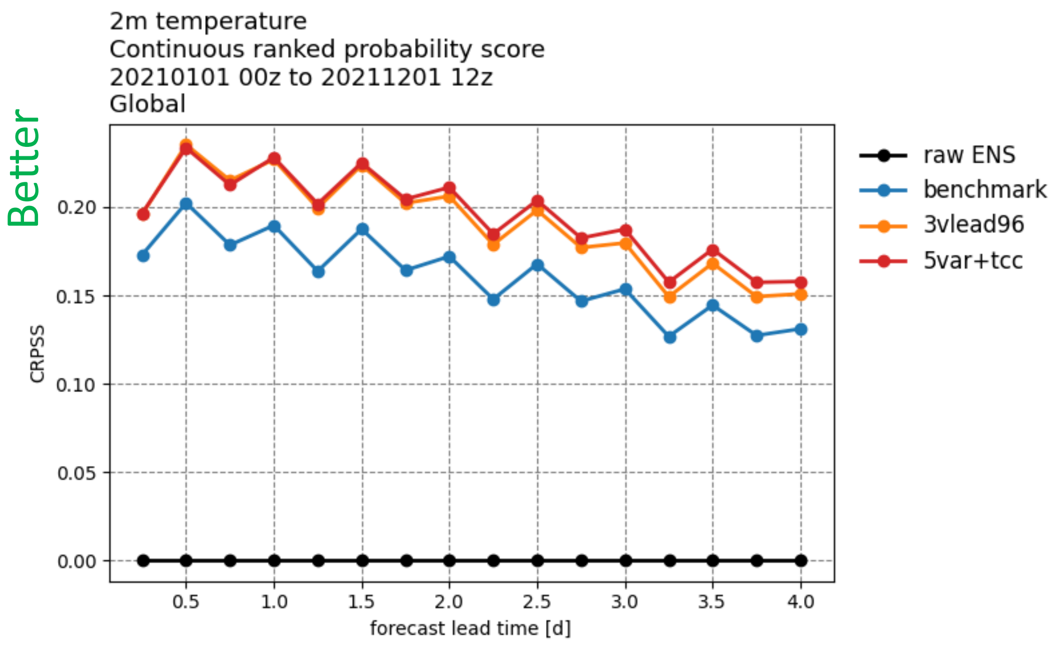
	PoET	MBM benchmark
Inputs	multiple variables	only target variable
Training period	2000-2016	2000-2019
Training lead times	all from 6 to 96	one lead time per model
Training window	all	+/- 15 days from init
Training ensemble members	11	11
Constraints	none	reliability
Spatial context	aware	per-grid-point (rank retained)

Scores

CRPS

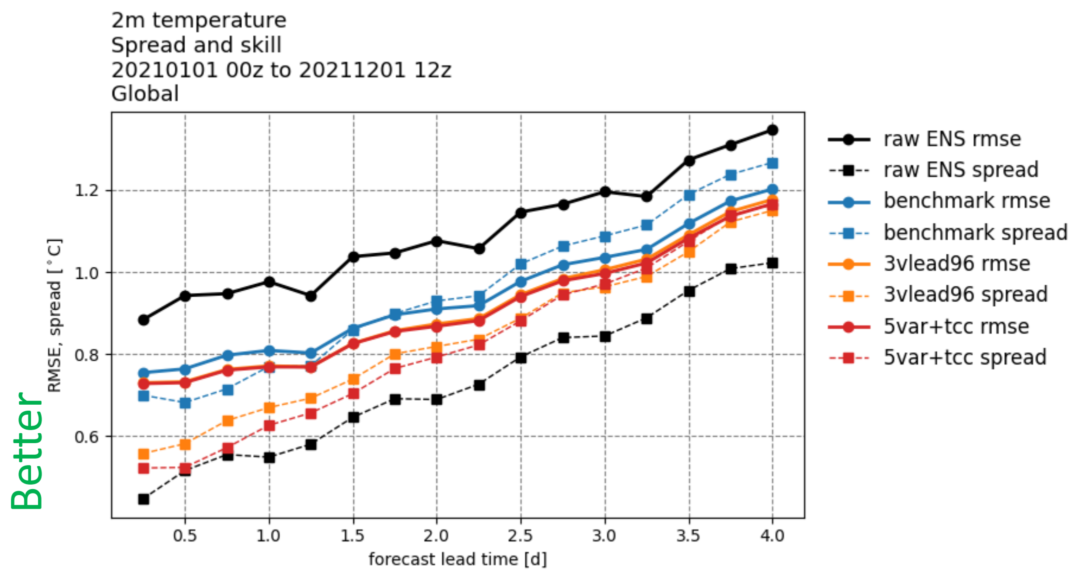


CRPS SCORE

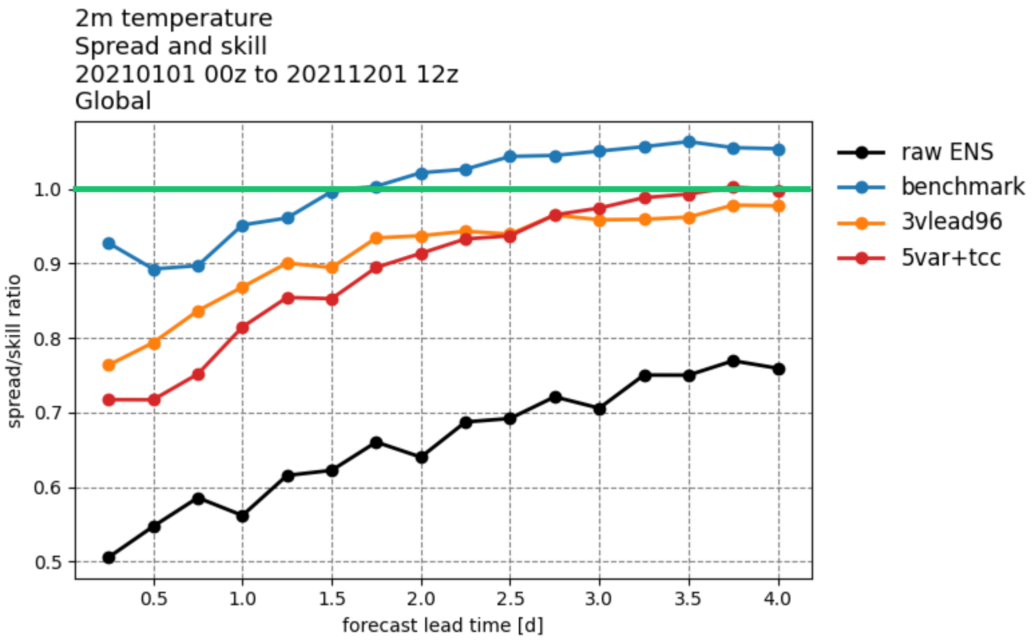


Scores

SPREAD/SKILL



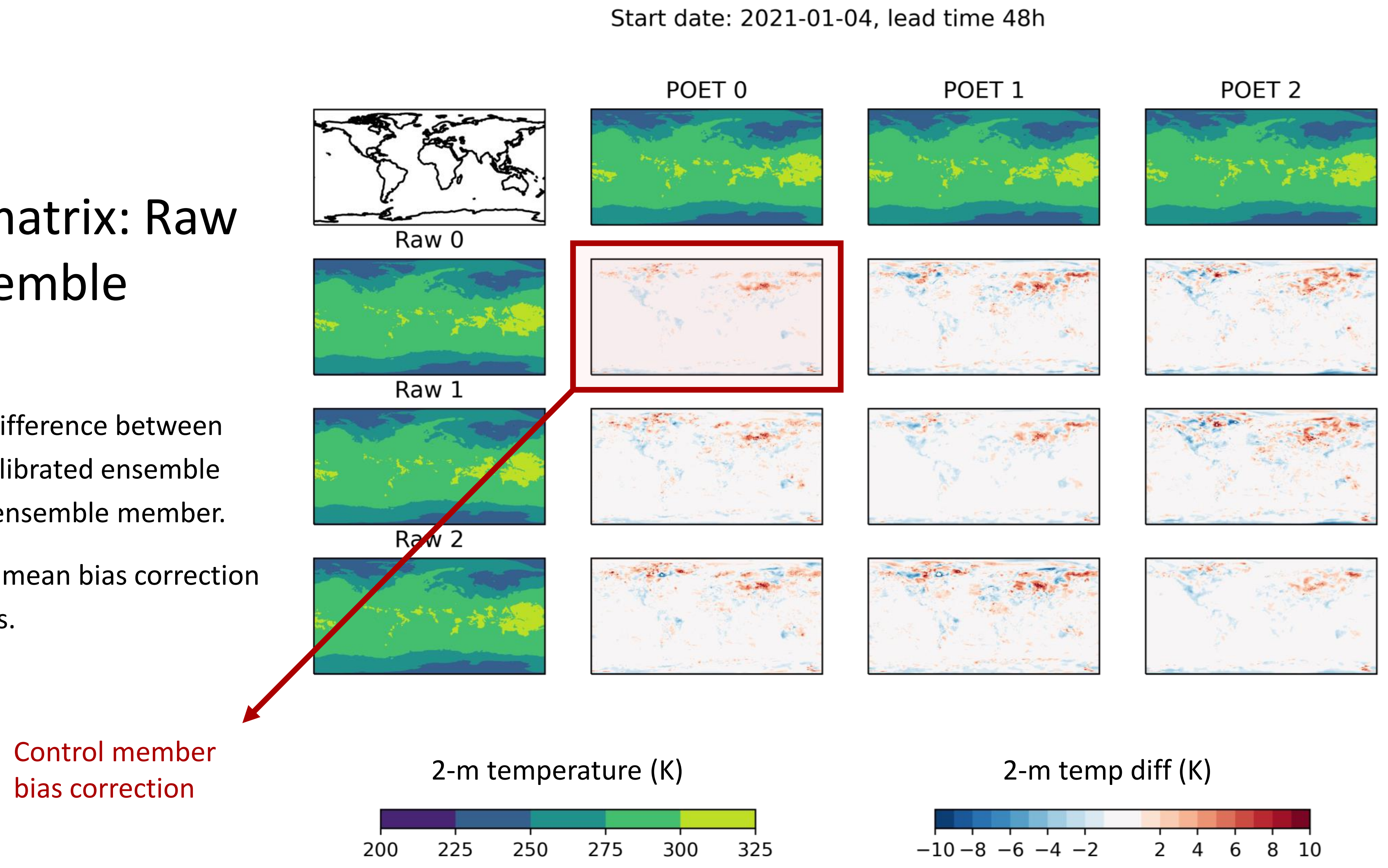
RELIABILITY (SPREAD/SKILL)



Difference matrix: Raw vs PoET ensemble members

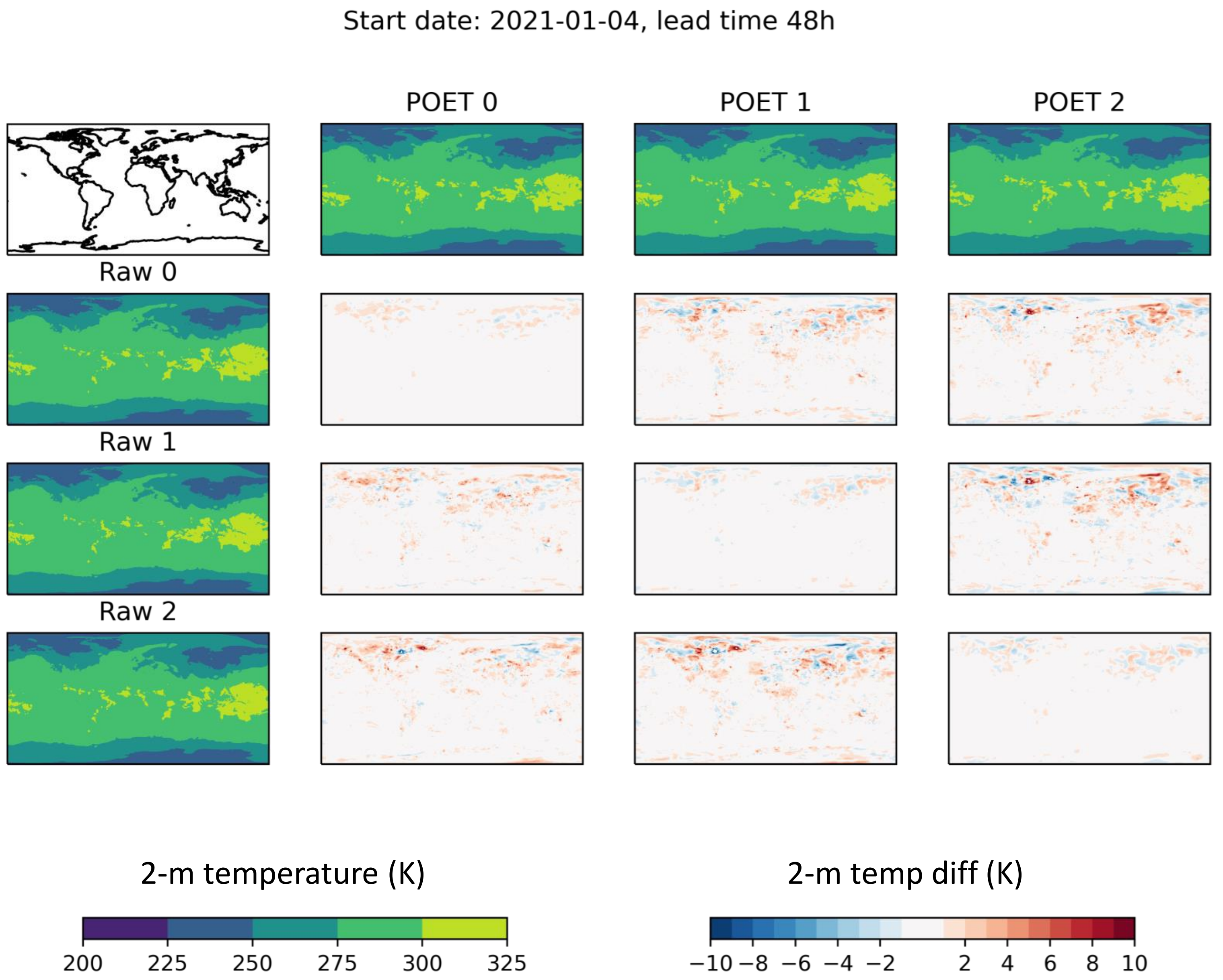
Inner panels are the difference between forecast from PoET-calibrated ensemble member and raw IFS ensemble member.

Dominant feature is a mean bias correction applied in all members.



Difference matrix: Raw vs PoET ensemble members

Inner panels are the difference between forecast *anomalies relative to ensemble mean* from PoET-calibrated ensemble member and raw IFS ensemble member.

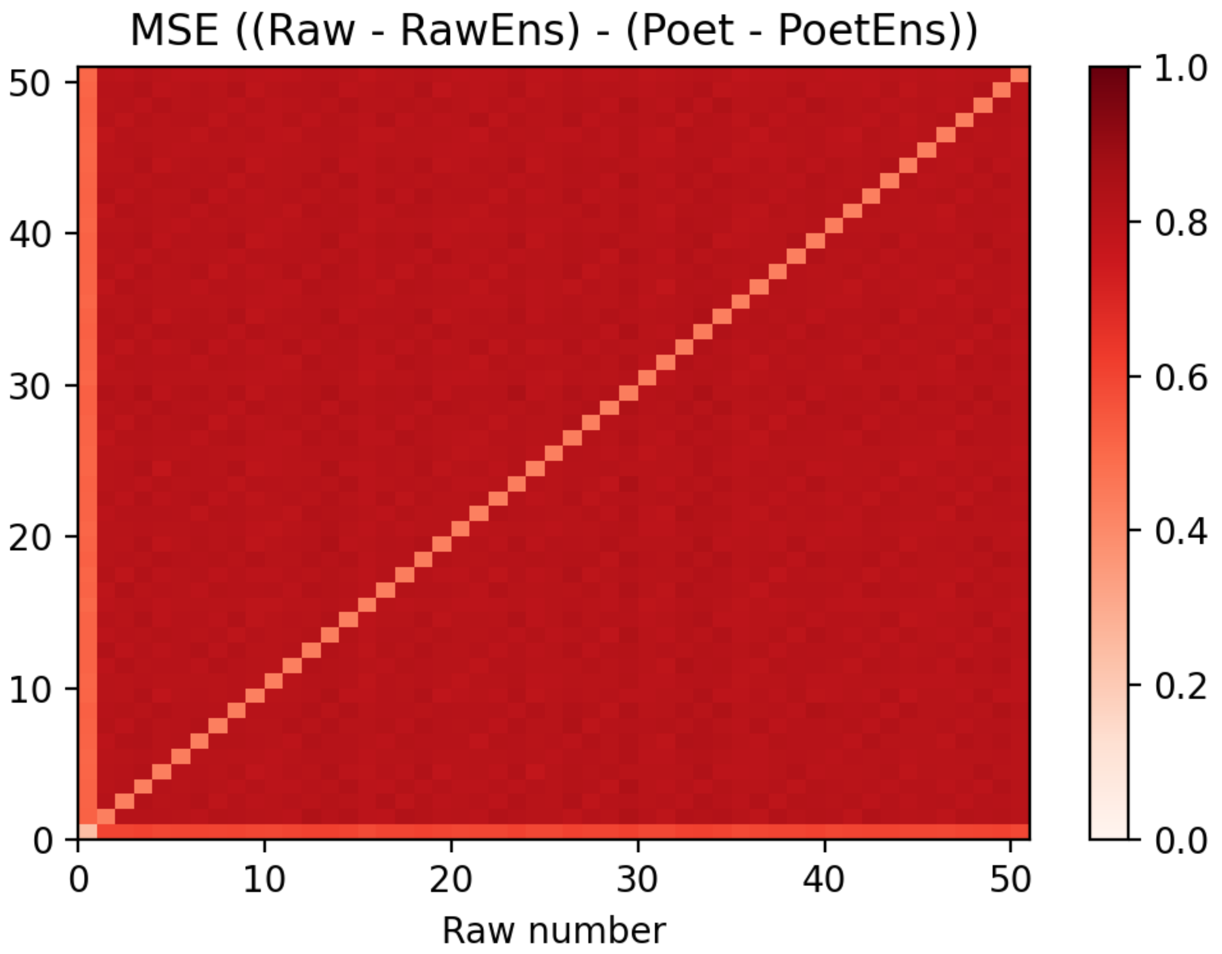


PoET preserves inter-member calibration

This figure computes the MSE between each pair of raw and PoET-calibrated ensemble members.

The diagonal elements are lower, indicating that PoET members still resemble their original member more closely.

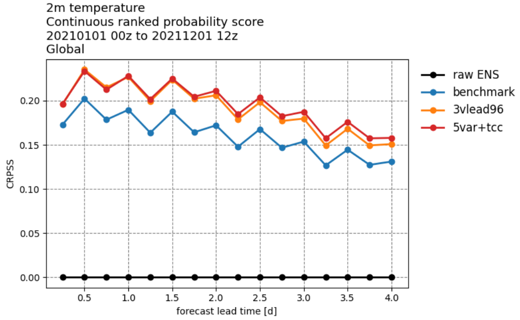
The control member is also more similar to all other members than other members are to each other.

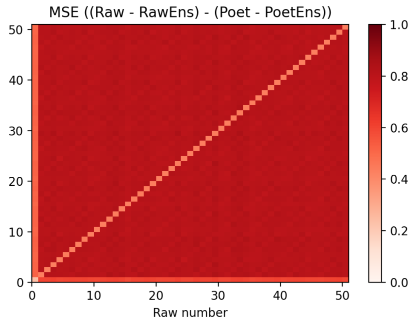


Further work

- Expand input data to include more variables and lead times, and predict more output variables of interest, such as precipitation
  - preliminary results on precipitation show modest improvement with PoET
- Further optimize architecture to allow for inference on full-resolution IFS fields

Take-aways





PoET (Post-processing of Ensembles with Transformers) effectively improves performance of the operational IFS ensemble while improving spread/skill ratio

The ensemble transformer can run on the full operational IFS ensemble while only being trained on an 11-member hindcast, yet still preserving inter-member calibration

The full distribution of the PoET-calibrated ensemble can be used in further ensemble analysis