Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

International Max Planck Research School
on Earth System Modelling

Max-Planck-Institut
für Meteorologie

ESA-ECMWF Machine Learning for Earth System Observation
and Prediction Workshop 2021

Self-Attentive Ensemble Transformer:
Representing Ensemble Interactions in Neural Networks for Earth System Models

Tobias Sebastian Finn

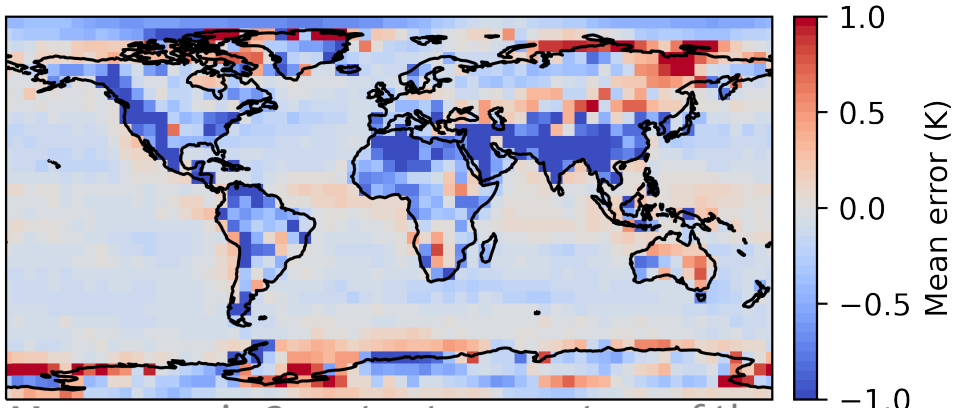
cerea

École des Ponts
ParisTech

Why do we need to post-process ensemble simulations?

Why do we need to post-process ensemble simulations?

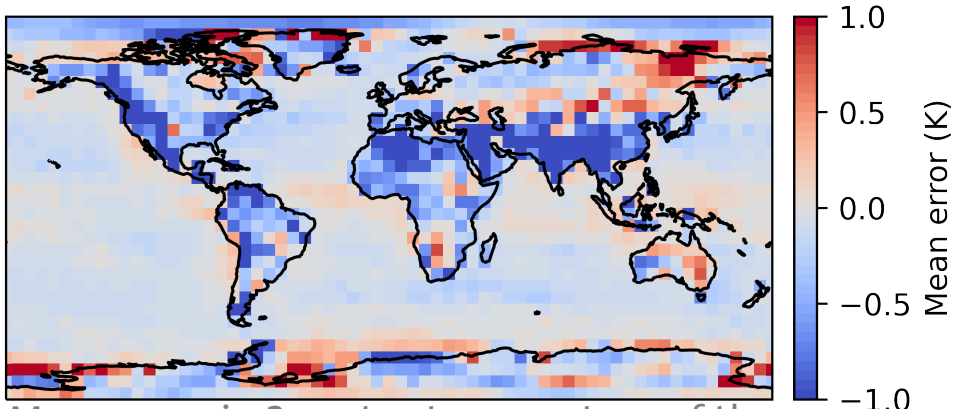
Model bias



Mean error in 2-metre-temperature of the IFS-EPS ensemble mean to the ERA5-reanalysis on a bilinearly regridded grid for 2019

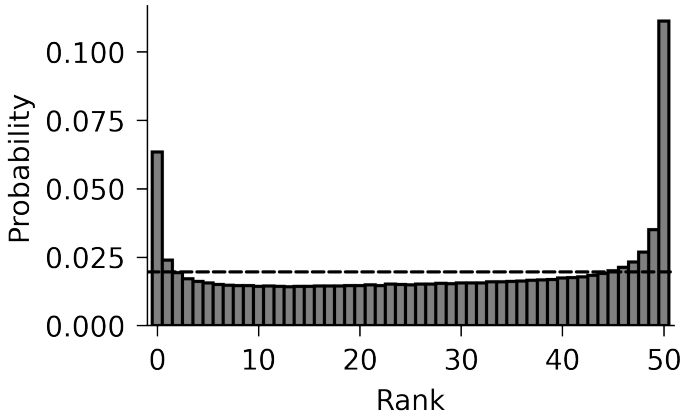
Why do we need to post-process ensemble simulations?

Model bias



Mean error in 2-metre-temperature of the IFS-EPS ensemble mean to the ERA5-reanalysis on a bilinearly regridded grid for 2019

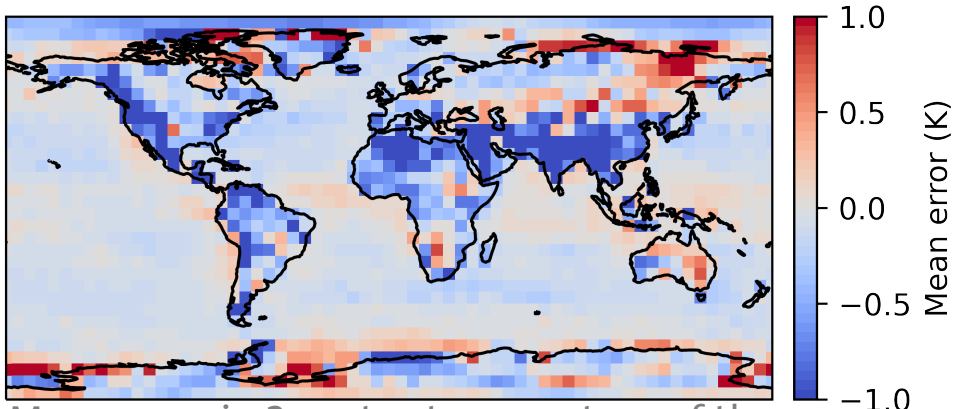
Uncalibrated forecast



Rank histogram in 2-metre-temperature of the IFS-EPS compared to the ERA5-reanalysis for all grid points and the year 2019

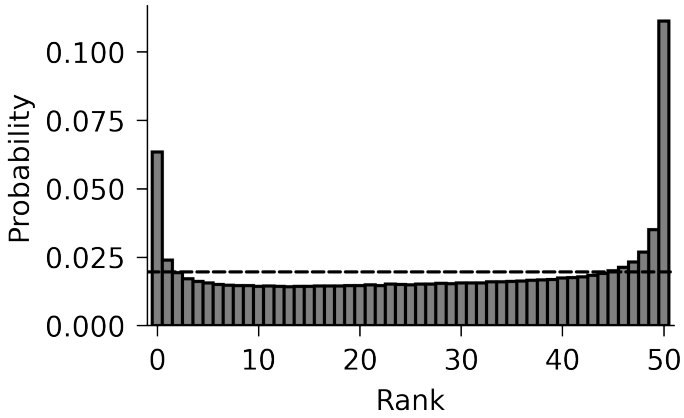
Why do we need to post-process ensemble simulations?

Model bias



Mean error in 2-metre-temperature of the IFS-EPS ensemble mean to the ERA5-reanalysis on a bilinearly regridded grid for 2019

Uncalibrated forecast



Rank histogram in 2-metre-temperature of the IFS-EPS compared to the ERA5-reanalysis for all grid points and the year 2019

+ Forecast non-modelled variables

How to post-process the ensemble data?

How to post-process the ensemble data?

Let’s use neural networks to correct the model bias

How to incorporate ensemble information?

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

	Raw Ensemble
Bias correction	✗
Calibration	✗
Non-parametric	✓
Correlations	✓

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

	Raw Ensemble	PPNN (Rasp & Lerch, 2018)
Bias correction	✗	
Calibration	✗	
Non-parametric	✓	
Correlations	✓	

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

	Raw Ensemble	PPNN (Rasp & Lerch, 2018)
Bias correction	✗	✓
Calibration	✗	✓
Non-parametric	✓	✗
Correlations	✓	✗

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

	Raw Ensemble	PPNN (Rasp & Lerch, 2018)	Bernstein (Bremnes, 2020)
Bias correction	✗	✓	
Calibration	✗	✓	
Non-parametric	✓	✗	
Correlations	✓	✗	

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

	Raw Ensemble	PPNN (Rasp & Lerch, 2018)	Bernstein (Bremnes, 2020)
Bias correction	✗	✓	✓
Calibration	✗	✓	✓
Non-parametric	✓	✗	~
Correlations	✓	✗	?

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

	Raw Ensemble	PPNN (Rasp & Lerch, 2018)	Bernstein (Bremnes, 2020)	Direct
Bias correction	✗	✓	✓	
Calibration	✗	✓	✓	
Non-parametric	✓	✗	~	
Correlations	✓	✗	?	

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

	Raw Ensemble	PPNN (Rasp & Lerch, 2018)	Bernstein (Bremnes, 2020)	Direct
Bias correction	✗	✓	✓	✓
Calibration	✗	✓	✓	✗
Non-parametric	✓	✗	~	✓
Correlations	✓	✗	?	?

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

Use self-attention similar to ensemble data assimilation

→ **Self-attentive ensemble transformer**

	Raw Ensemble	PPNN (Rasp & Lerch, 2018)	Bernstein (Bremnes, 2020)	Direct	**Transfromer** (This talk)
Bias correction	✗	✓	✓	✓	
Calibration	✗	✓	✓	✗	
Non-parametric	✓	✗	~	✓	
Correlations	✓	✗	?	?	

How to post-process the ensemble data?

Let's use neural networks to correct the model bias

How to incorporate ensemble information?

Use self-attention similar to ensemble data assimilation

→ **Self-attentive ensemble transformer**

	Raw Ensemble	PPNN (Rasp & Lerch, 2018)	Bernstein (Bremnes, 2020)	Direct	**Transfromer** (This talk)
Bias correction	✗	✓	✓	✓	✓
Calibration	✗	✓	✓	✗	✓
Non-parametric	✓	✗	~	✓	✓
Correlations	✓	✗	?	?	(✓)

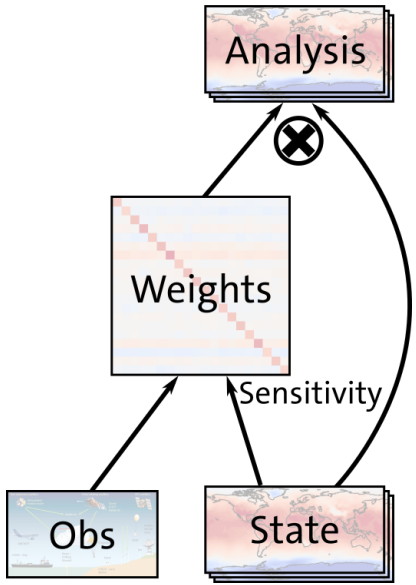
Self-attentive ensemble transformer combines ideas from data assimilation and NLP

Ensemble-based data assimilation

Natural language processing

Self-attentive ensemble transformer combines ideas from data assimilation and NLP

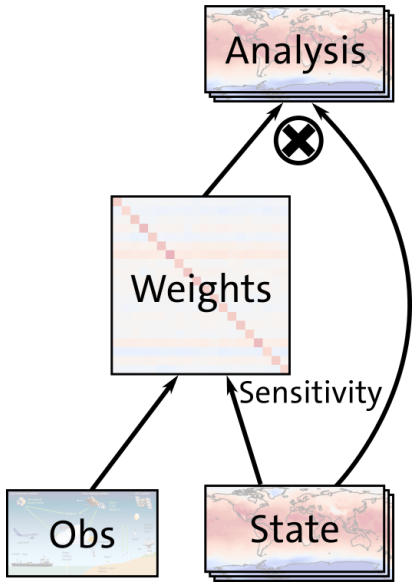
Ensemble-based data assimilation



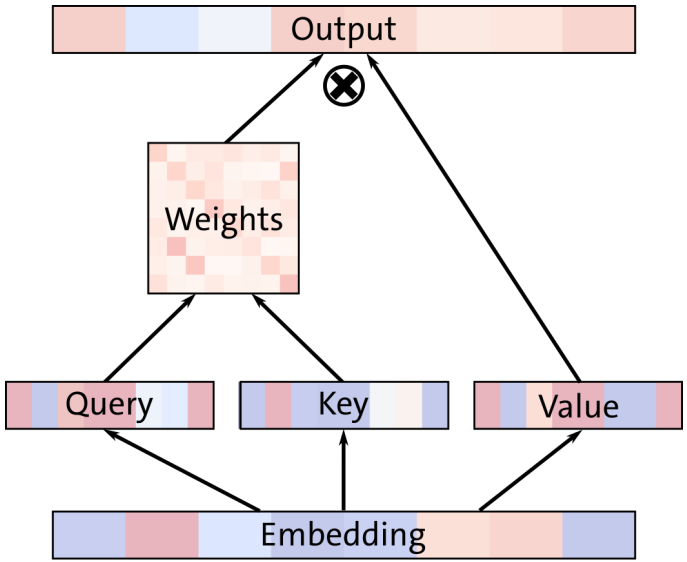
Natural language processing

Self-attentive ensemble transformer combines ideas from data assimilation and NLP

Ensemble-based data assimilation

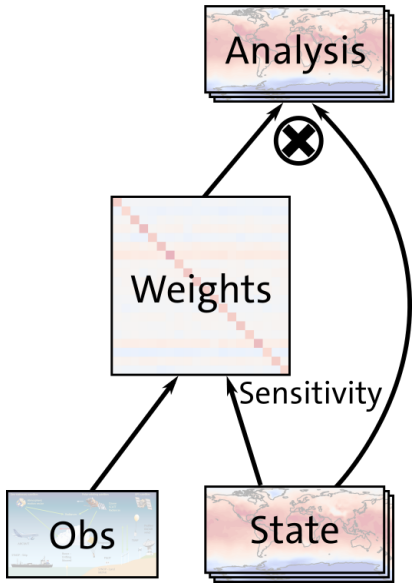


Natural language processing



Self-attentive ensemble transformer combines ideas from data assimilation and NLP

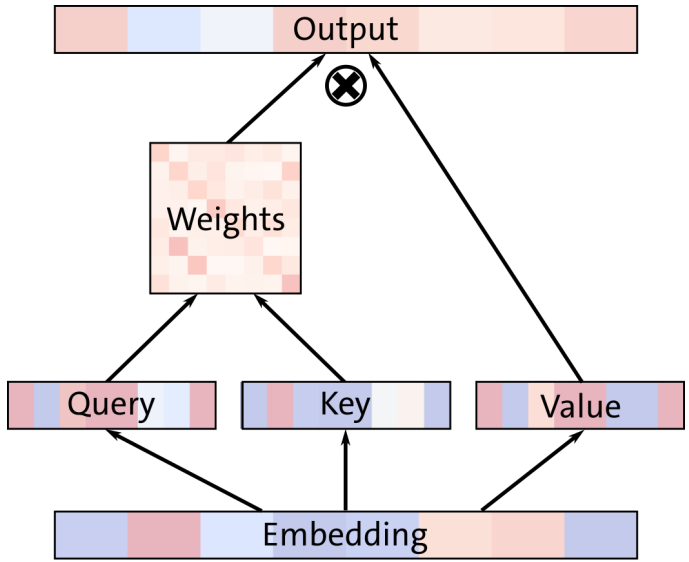
Ensemble-based data assimilation



Global ensemble transformation

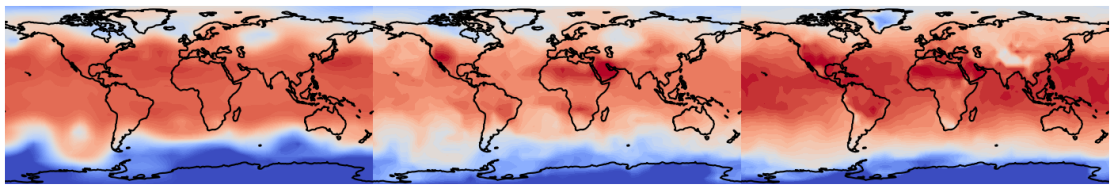
Self-attention

Natural language processing



Global post-processing case

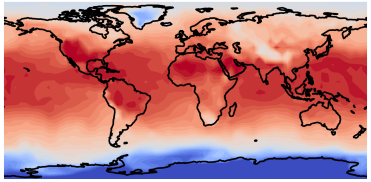
Global post-processing case



Geopotential on 500 hPa | Temperature on 850 hPa | 2-metre-temperature

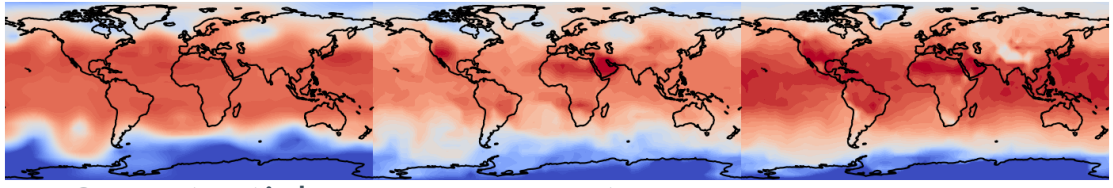
IFS-EPS (50 members, lead time: 48 h)

Global post-processing case



ERA5 reanalysis

2-metre-temperature



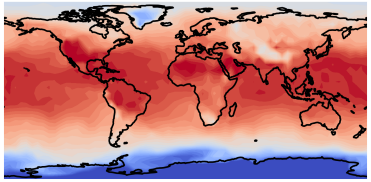
Geopotential on 500 hPa

Temperature on 850 hPa

2-metre-temperature

IFS-EPS (50 members, lead time: 48 h)

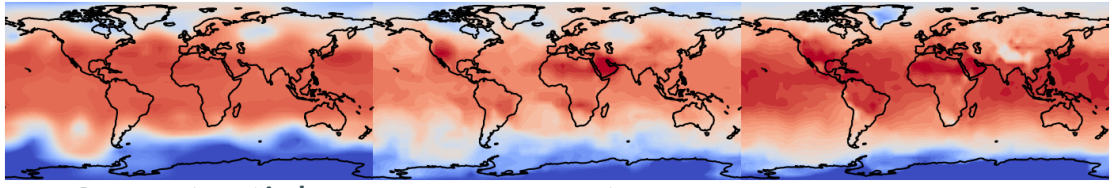
Global post-processing case



ERA5 reanalysis

2-metre-temperature

Global regression with
convolution neural networks



Geopotential
on 500 hPa

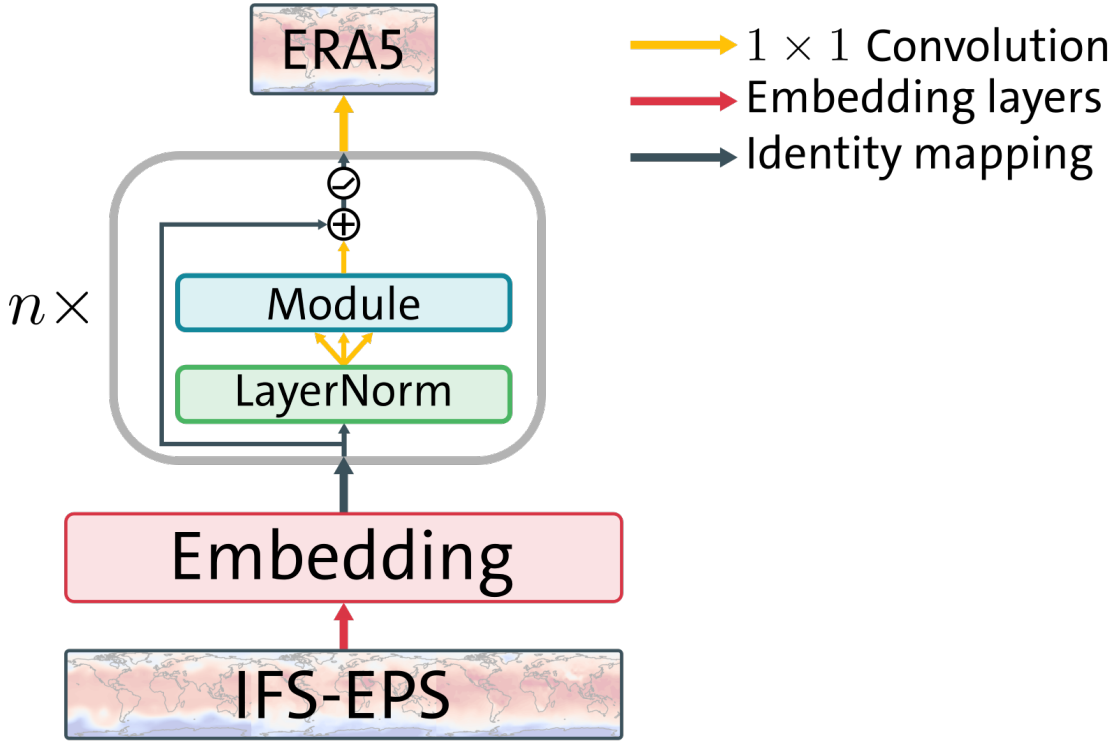
Temperature
on 850 hPa

2-metre-temperature

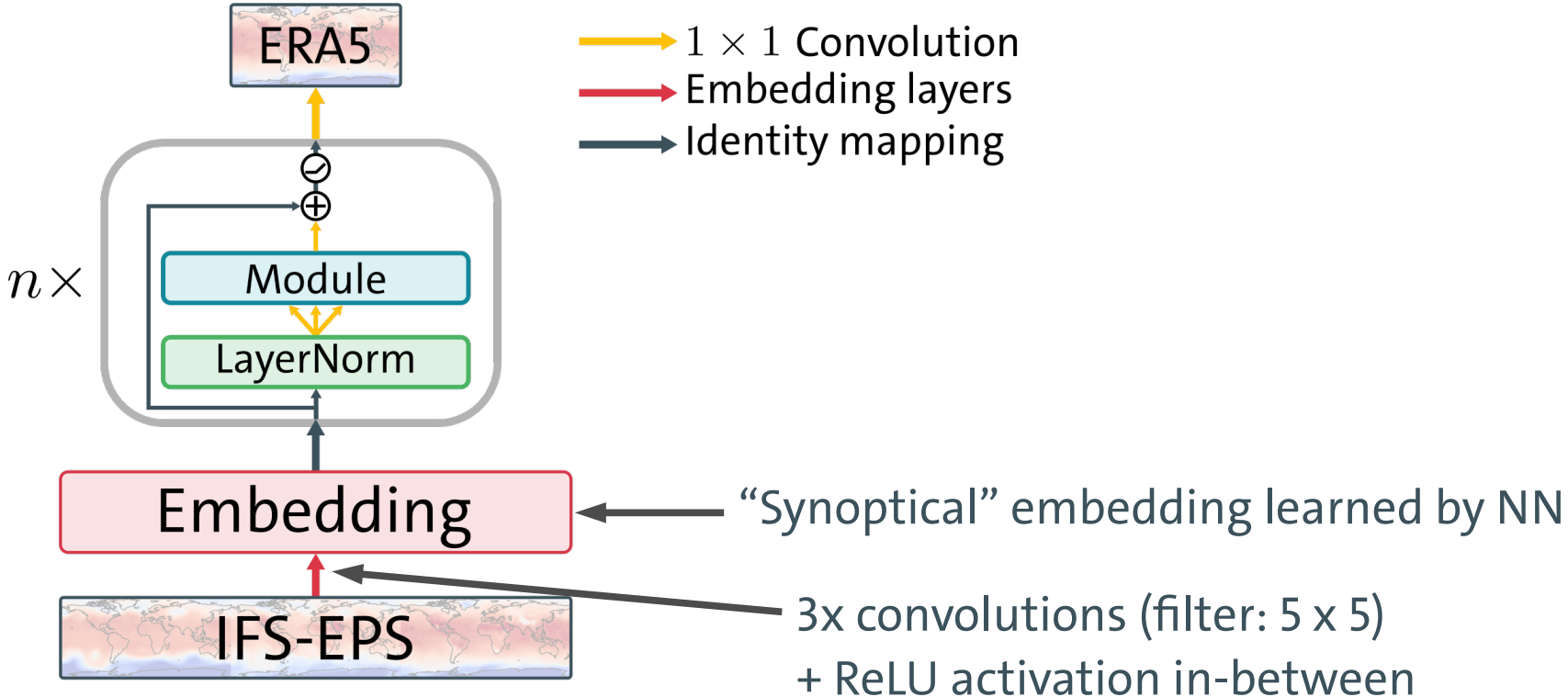
IFS-EPS (50 members, lead time: 48 h)

Ensemble transformer architecture

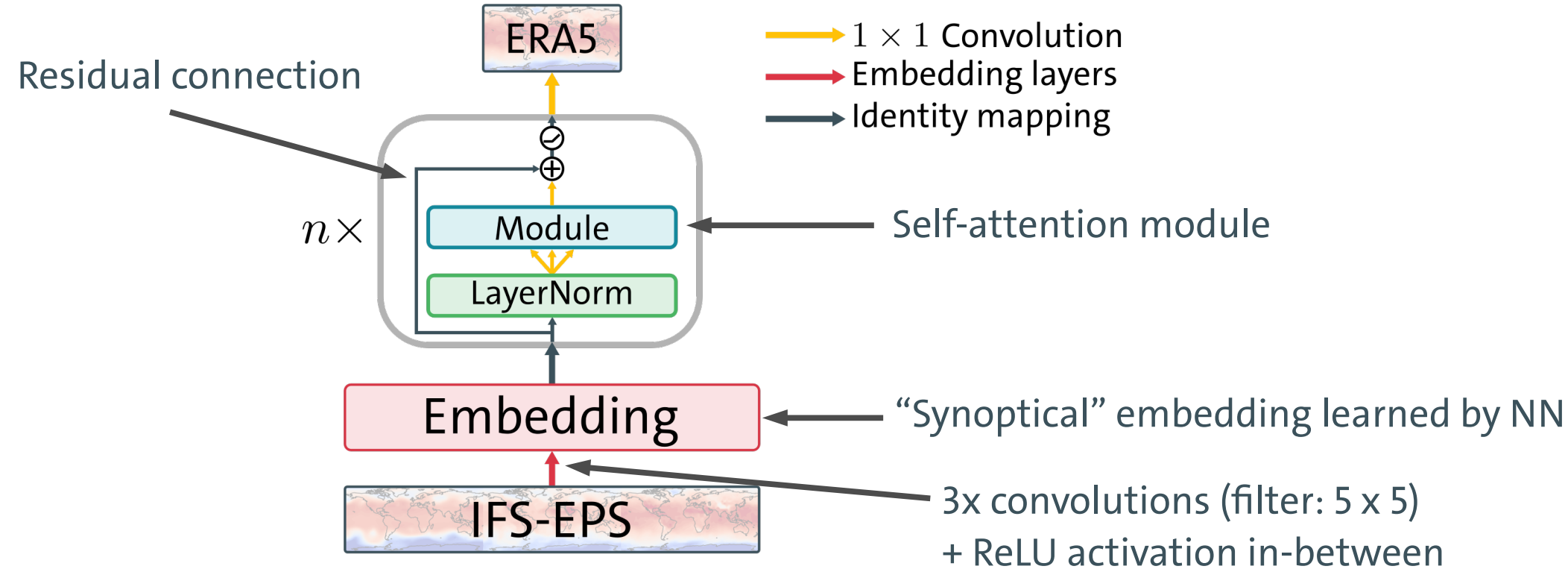
Ensemble transformer architecture



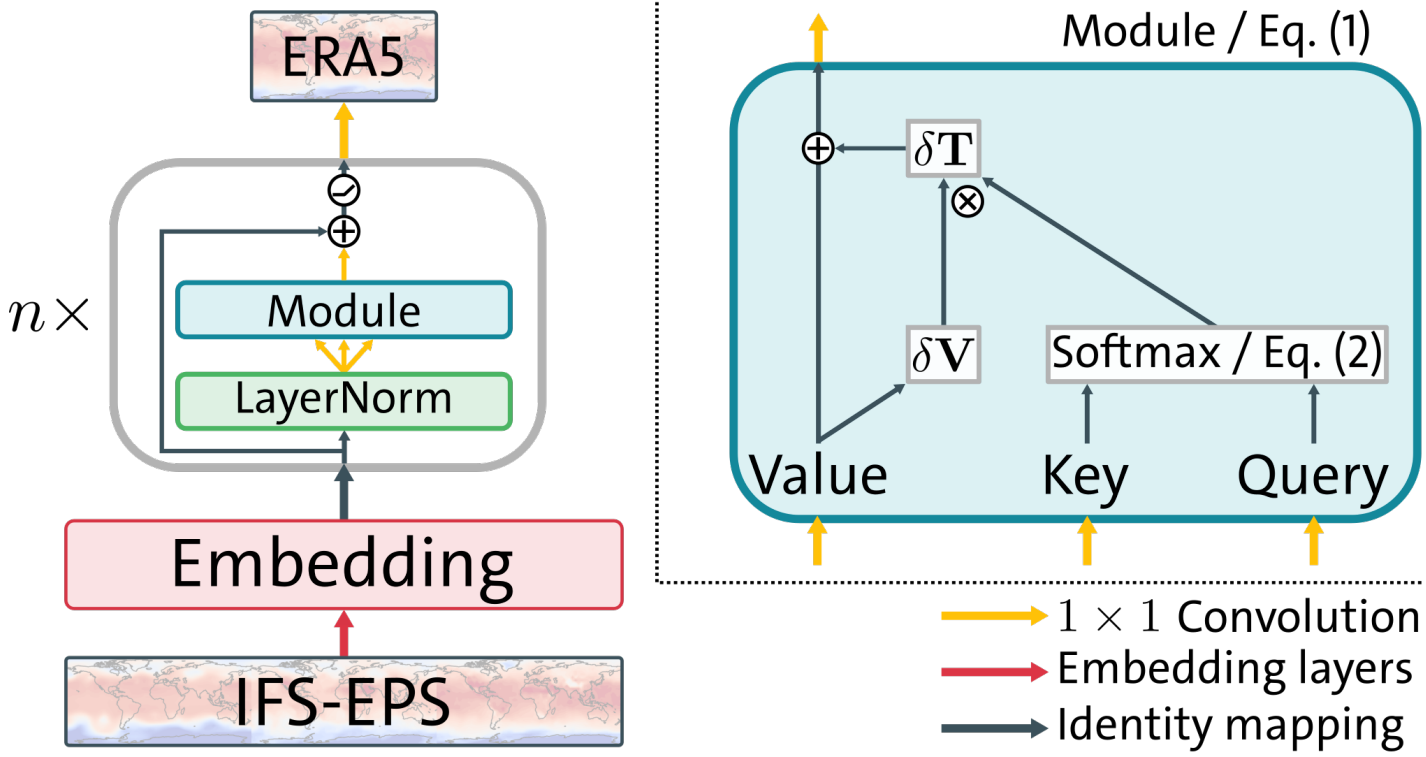
Ensemble transformer architecture



Ensemble transformer architecture



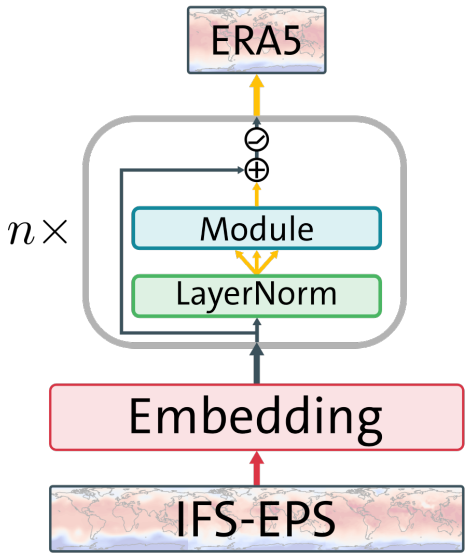
Ensemble transformer architecture



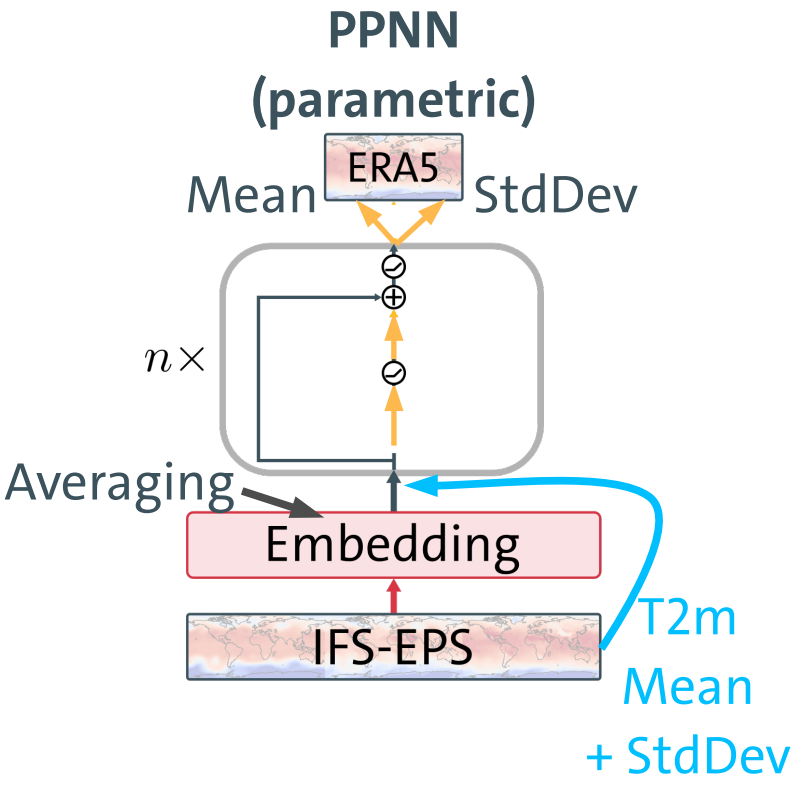
Experiment: Used methods

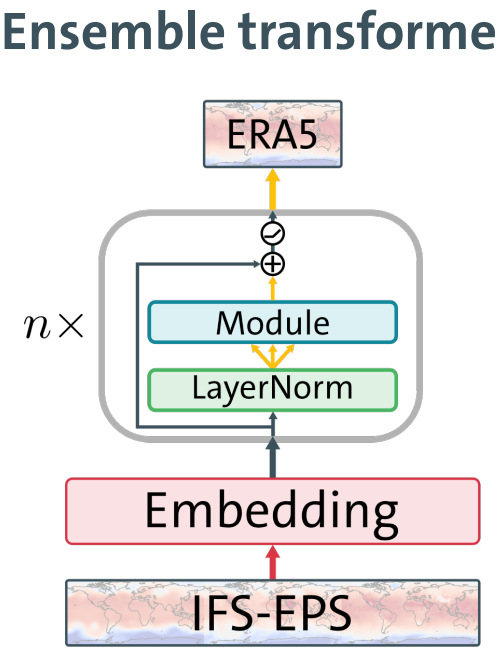
Experiment: Used methods

Ensemble transformer

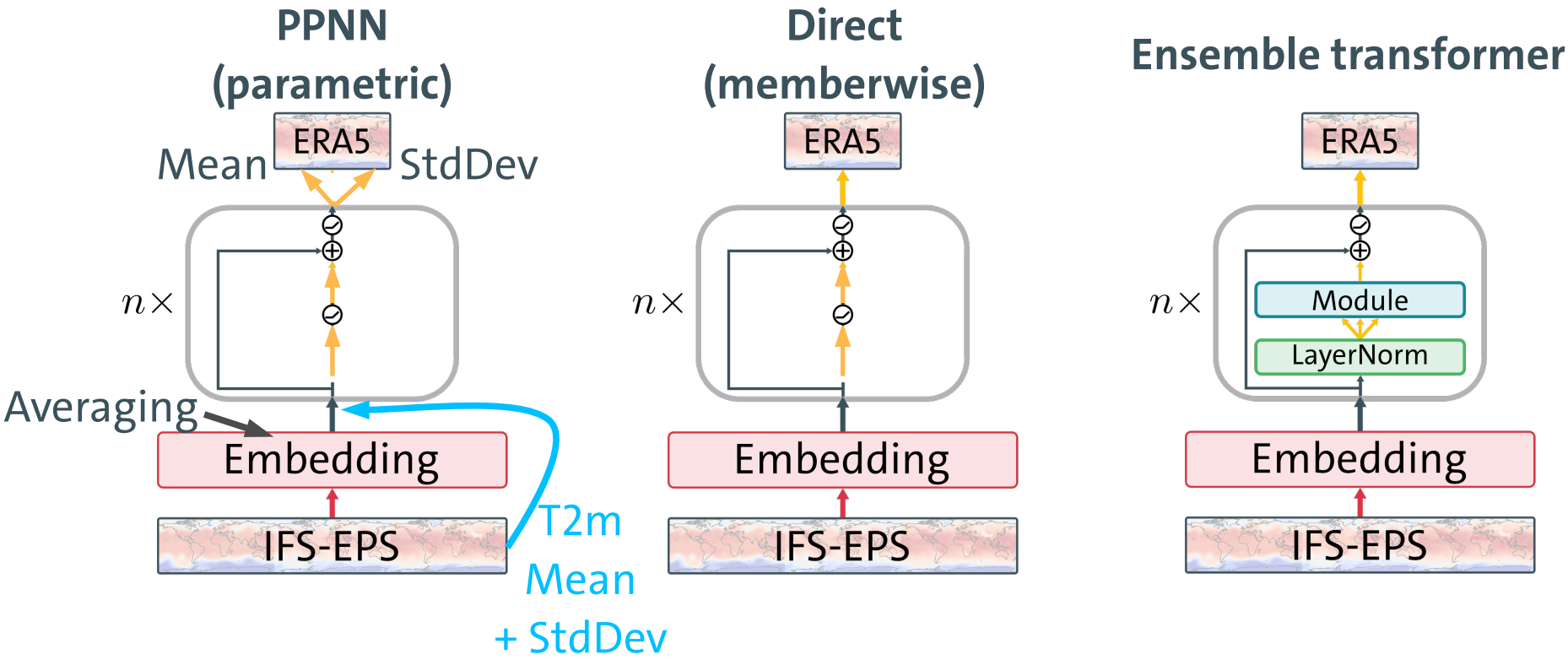


Experiment: Used methods





Experiment: Used methods



Experiment: Training

Experiment: Training

2017 & 2018: Training and validation (10 %)
2019: Testing

Optimizer: Adam
Loss: Latitudinal-weighted continuous ranked probability score (CRPS) for Gaussians

Experiment: Training

2017 & 2018: Training and validation (10 %)
2019: Testing

Optimizer: Adam
Loss: Latitudinal-weighted continuous ranked probability score (CRPS) for Gaussians

Subsampled ensemble members for each training step:

Experiment: Training

2017 & 2018: Training and validation (10 %)
2019: Testing

Optimizer: Adam
Loss: Latitudinal-weighted continuous ranked probability score (CRPS) for Gaussians

Subsampled ensemble members for each training step:

	RMSE (K)	Spread (K)
10 samples	0.91	0.91
20 samples	0.92	0.90
50 samples	0.92	0.89

Experiment: Training

2017 & 2018: Training and validation (10 %)
2019: Testing

Optimizer: Adam
Loss: Latitudinal-weighted continuous ranked probability score (CRPS) for Gaussians

Subsampled ensemble members for each training step:

	RMSE (K)	Spread (K)
10 samples	0.91	0.91
20 samples	**0.92**	**0.90**
50 samples	0.92	0.89

Increased noise during training
+
Increased training speed

Results for the whole year 2019

Results for the whole year 2019

	RMSE (K)	Spread (K)

Results for the whole year 2019

	RMSE (K)	Spread (K)
IFS-EPS	1.12	0.73

PPNN scales slightly with number of layers

Parametric approach (Rasp & Lerch, 2018) →

	RMSE (K)	Spread (K)
IFS-EPS	1.12	0.73
PPNN (1)	0.95	0.87
PPNN (5)	0.93	0.87

Direct approach without self-attention has problems

Parametric approach (Rasp & Lerch, 2018) →

Apply NN to each member independently →

	RMSE (K)	Spread (K)
IFS-EPS	1.12	0.73
PPNN (1)	0.95	0.87
PPNN (5)	0.93	0.87
w/o Self-Attention (1)	0.95	0.70
w/o Self-Attention (5)	0.96	0.70

Transformer has lowest error and best spread-skill ratio

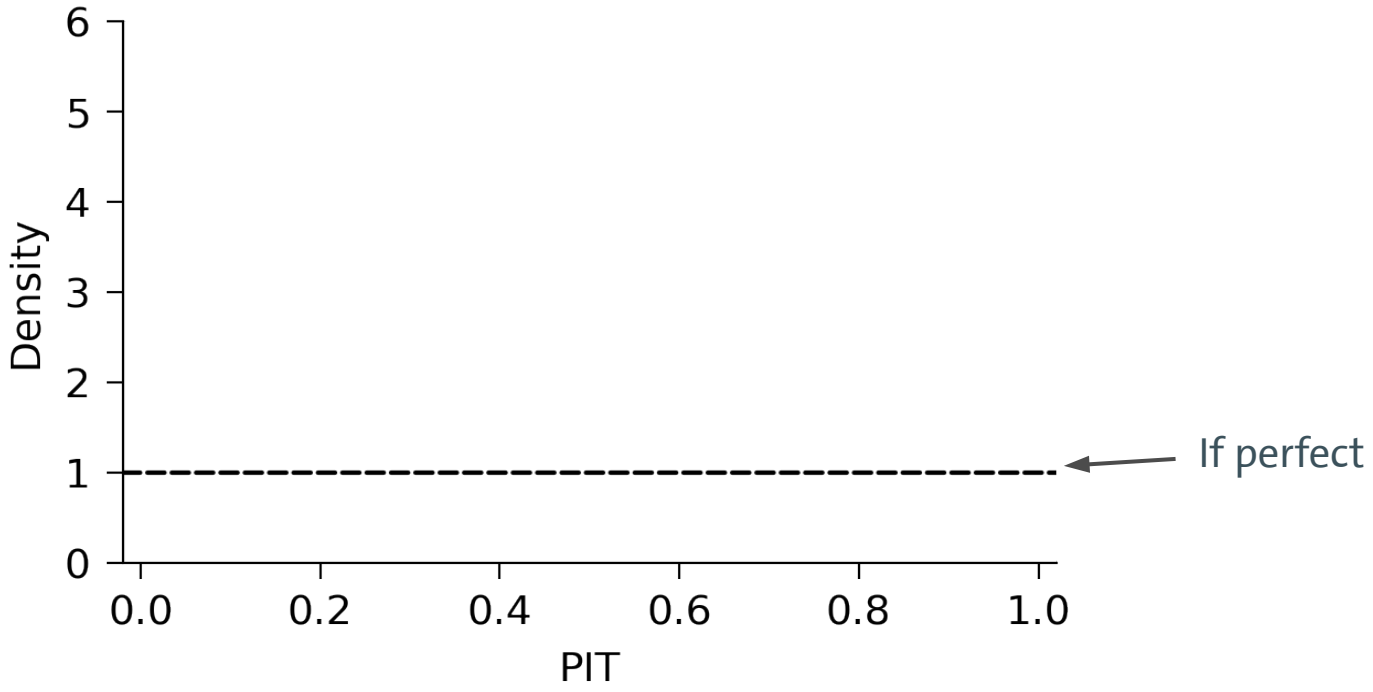
Parametric approach (Rasp & Lerch, 2018) →

Apply NN to each member independently →

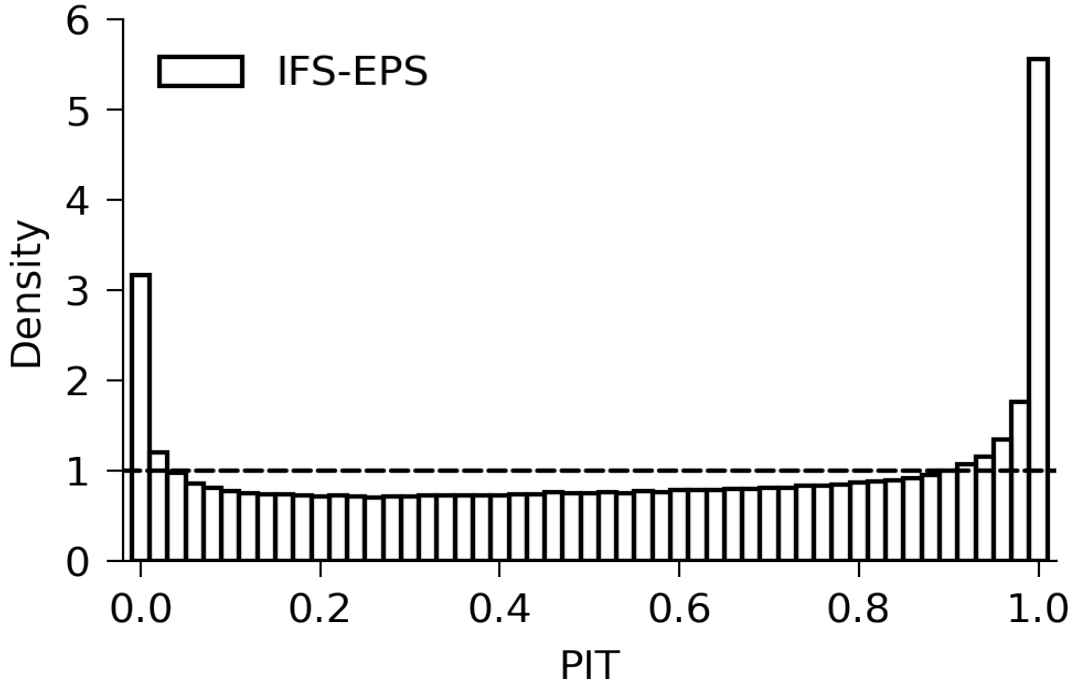
	RMSE (K)	Spread (K)
IFS-EPS	1.12	0.73
PPNN (1)	0.95	0.87
PPNN (5)	0.93	0.87
w/o Self-Attention (1)	0.95	0.70
w/o Self-Attention (5)	0.96	0.70
Transformer (1)	0.91	**0.91**
Transformer (5)	**0.90**	**0.90**

Self-attention can extract additional information from ensemble data
+ helps to calibrate the ensemble

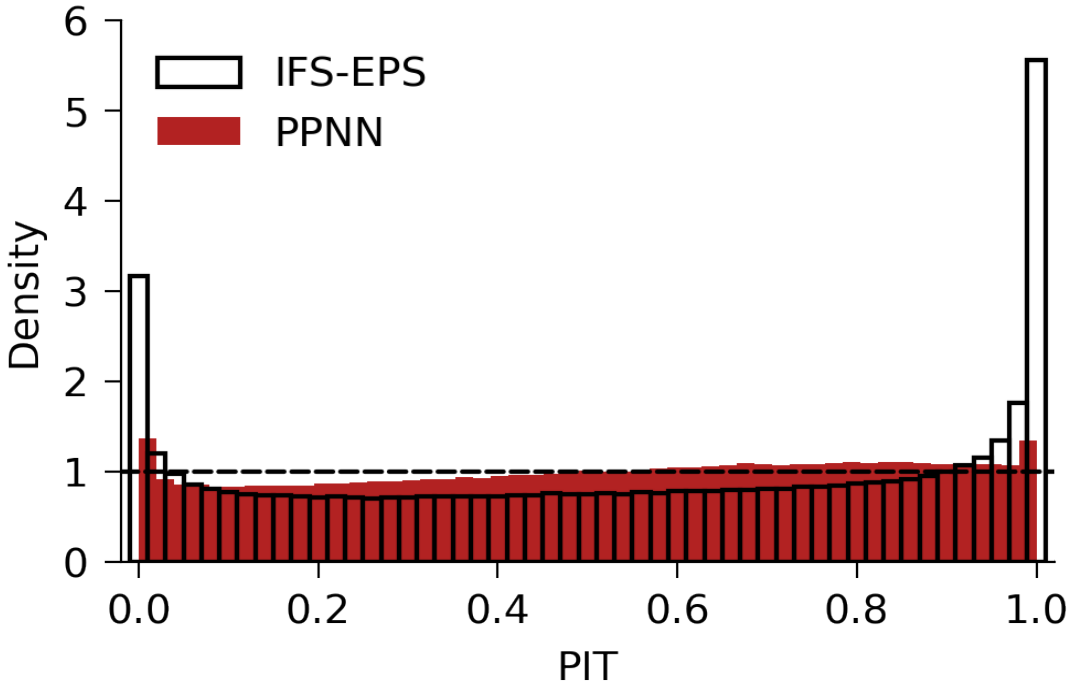
Probability integral transform diagram to check ensemble calibration



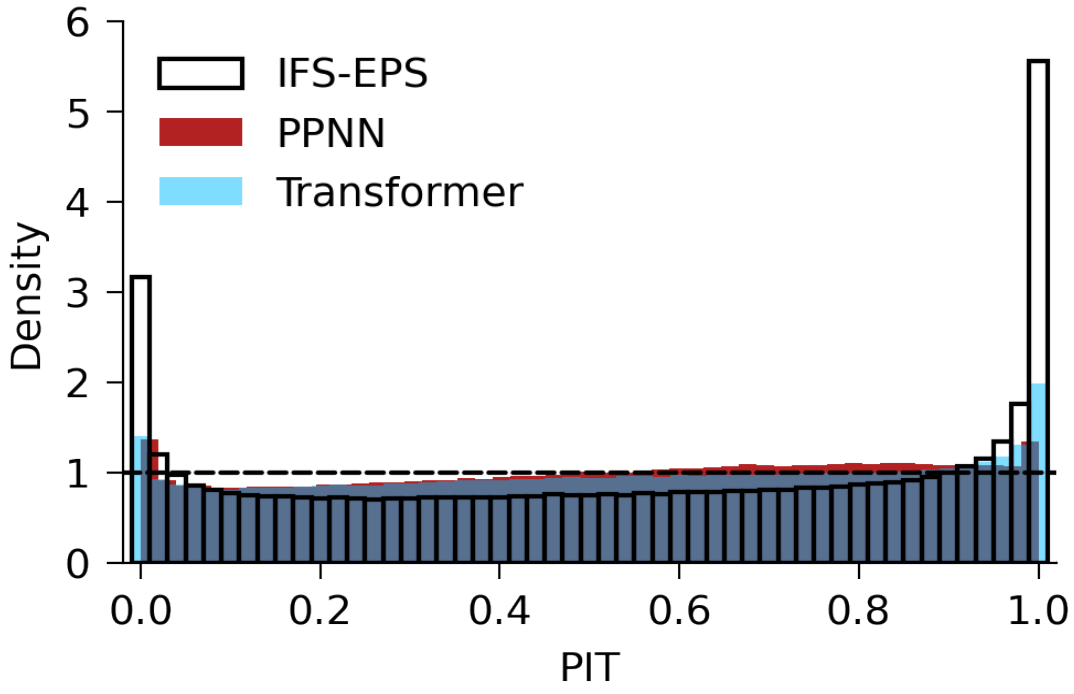
Raw IFS ensemble clearly underdispersive + model bias



PPNN calibrates the ensemble



PPNN and Transformer are similarly good calibrated

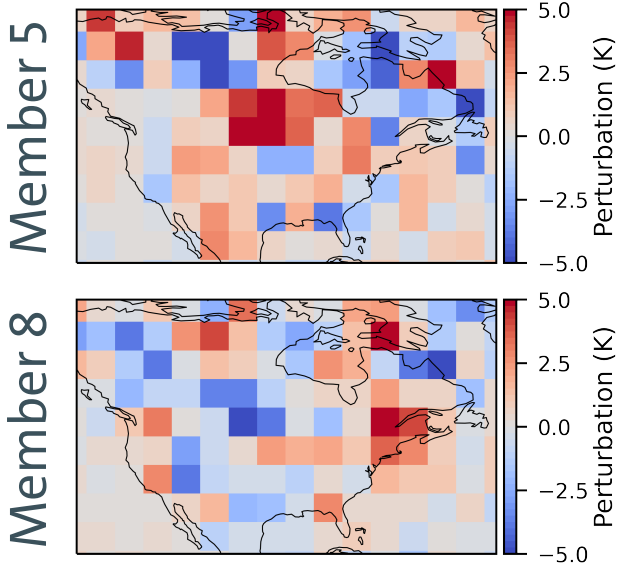


Transformer can represent spatial correlations

Transformer can represent spatial correlations

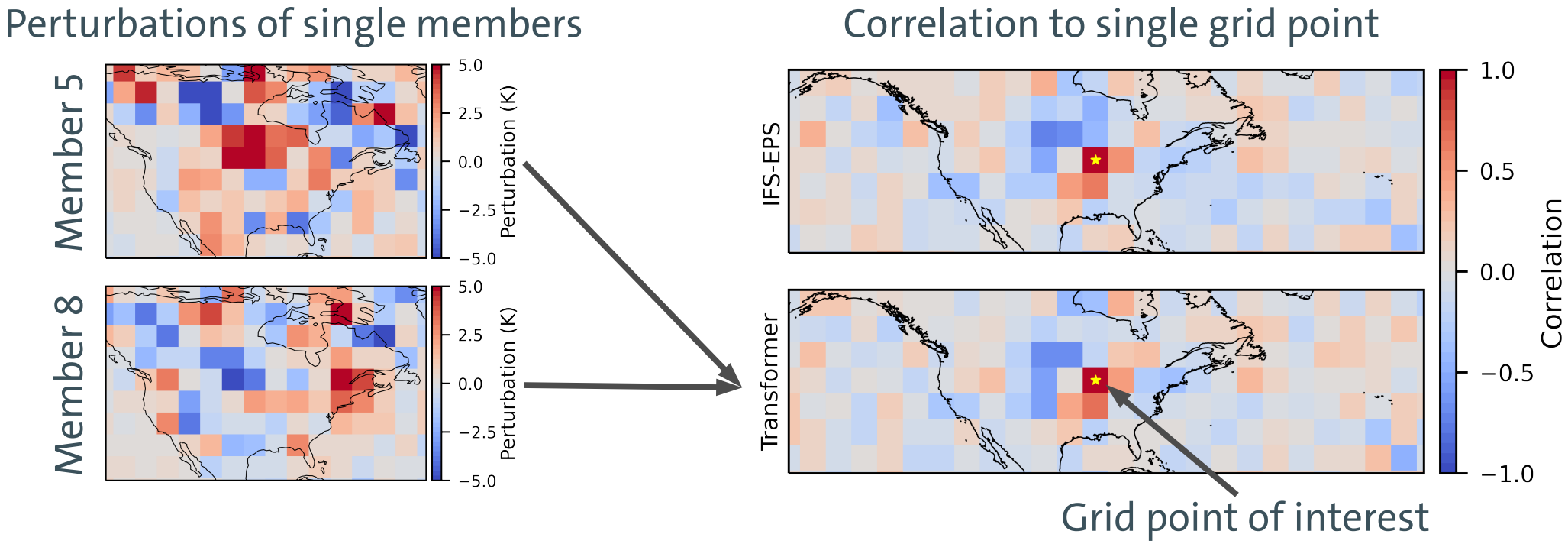
Cold wave in North America – 2019-01-26 12:00 UTC

Perturbations of single members



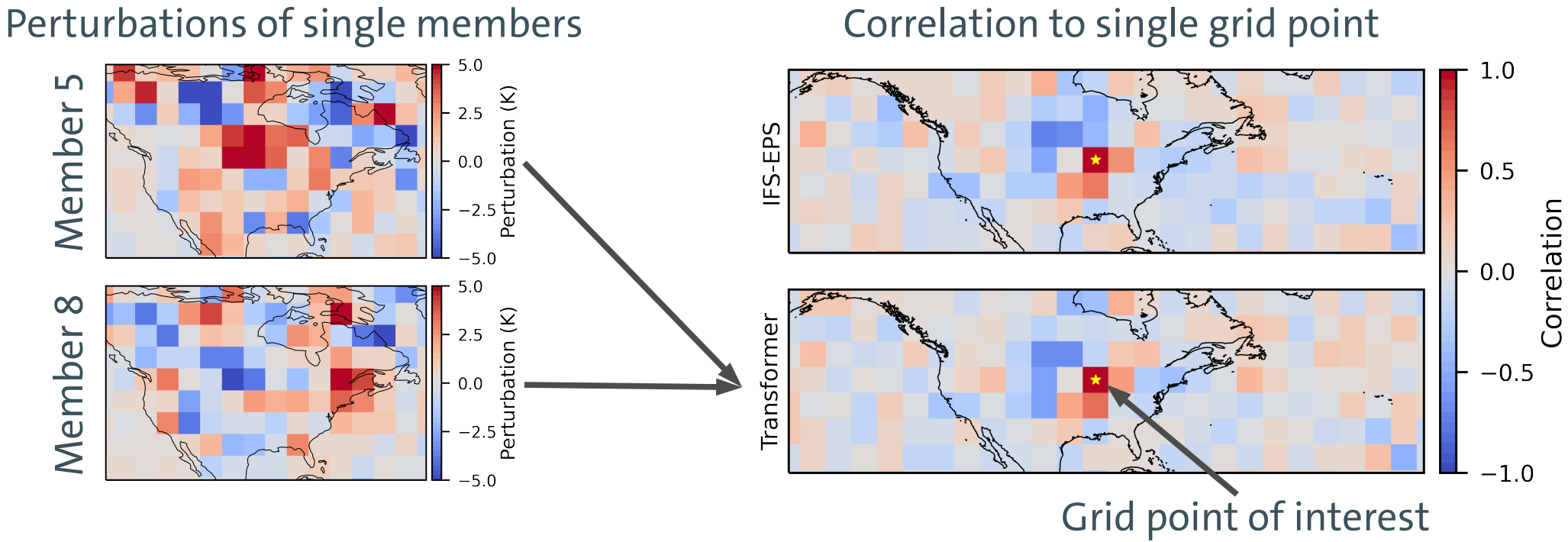
Transformer can represent spatial correlations

Cold wave in North America – 2019-01-26 12:00 UTC



Transformer can represent spatial correlations

Cold wave in North America – 2019-01-26 12:00 UTC



But spurious correlations caused by global self-attention!

Self-attention can explain itself and create informative maps

Attention maps for 2019-09-01 12:00 UTC

Self-attention can explain itself and create informative maps

Attention maps for 2019-09-01 12:00 UTC

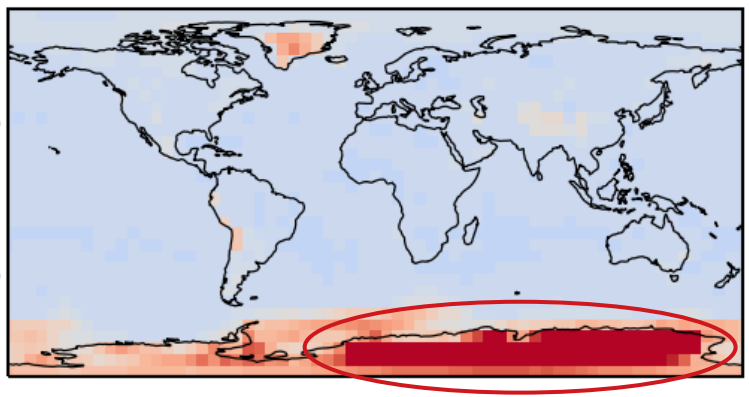
Temperatures below freezing level?

Positive temperature anomalies?

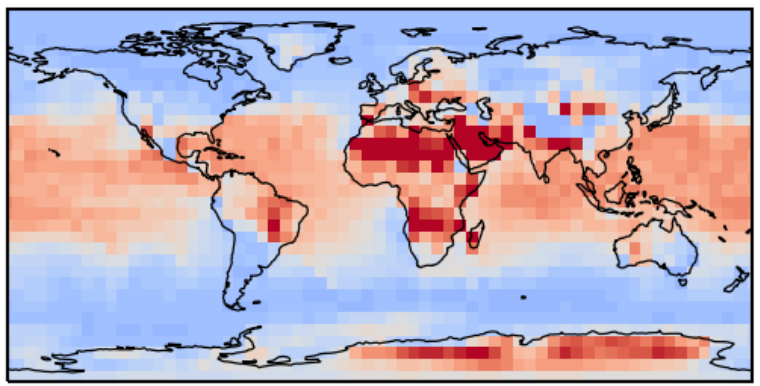
Head →

Layer →

(1,56)



(5,28)



Important areas

Self-attention can explain itself and create informative maps

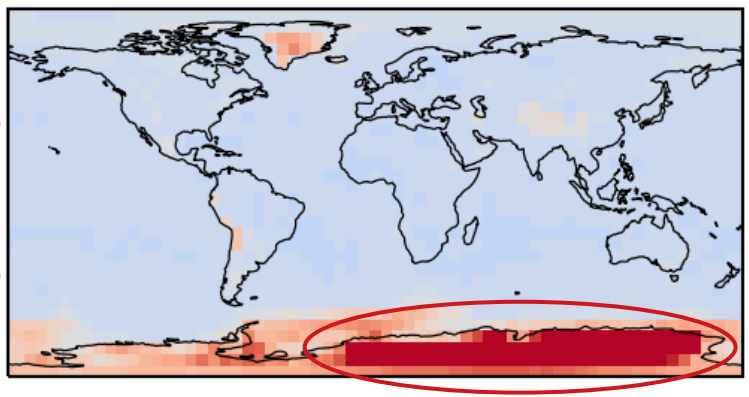
Attention maps for 2019-09-01 12:00 UTC

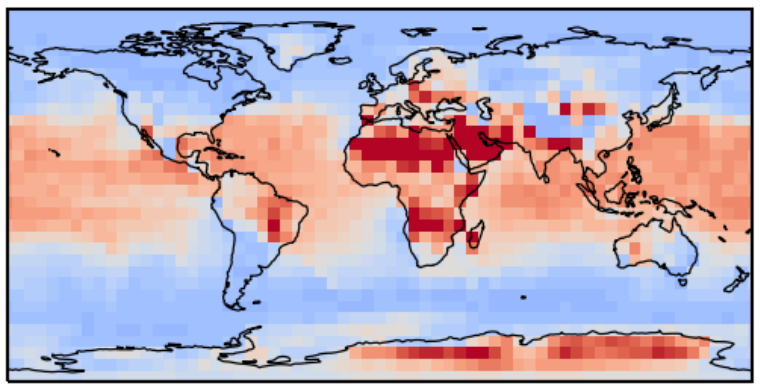
Temperatures below freezing level?

Positive temperature anomalies?

Head → (1,56)

Layer → (5,28)





Important areas

But repetitive heads!

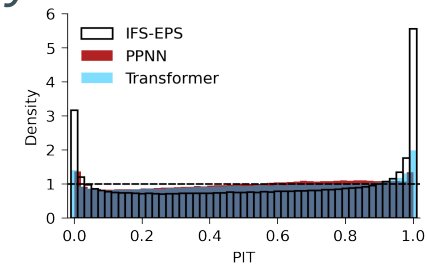
Conclusions of my study

Conclusions of my study

Self-attention can be used to improve post-processing for Earth system models

→ Extraction of additional information → Reduced error

→ Calibration of ensemble for improved uncertainty estimation

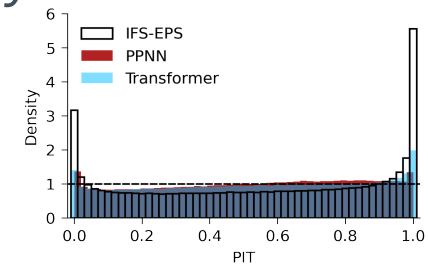


Conclusions of my study

Self-attention can be used to improve post-processing for Earth system models

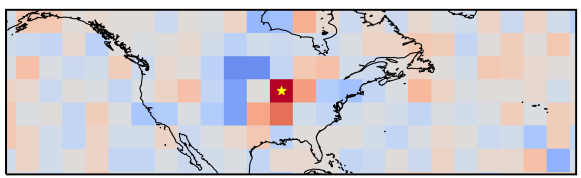
→ Extraction of additional information → Reduced error

→ Calibration of ensemble for improved uncertainty estimation



The ensemble transformer enables member-by-member post-processing with neural networks

→ Non-parametric processing of ensemble members without aggregated statistics



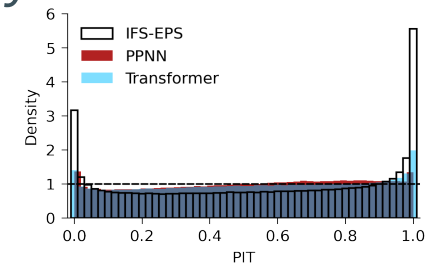
→ Output of spatially correlated forecasts

Conclusions of my study

Self-attention can be used to improve post-processing for Earth system models

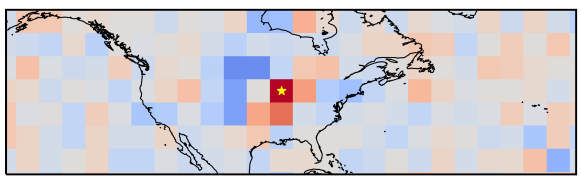
→ Extraction of additional information → Reduced error

→ Calibration of ensemble for improved uncertainty estimation



The ensemble transformer enables member-by-member post-processing with neural networks

→ Non-parametric processing of ensemble members without aggregated statistics



→ Output of spatially correlated forecasts

Do you have questions?

If you have questions

Take a look into the paper (a longer one is in preparation):

Self-Attentive Ensemble Transformer: Representing Ensemble Interactions in Neural Networks for Earth System Models

Tobias Sebastian Finn [1 2]

and/or take a look into the official code:
https://github.com/tobifinn/ensemble_transformer

and/or write me an e-mail:
tobias.finn@enpc.fr

and/or follow me on twitter:
@tobias_finn