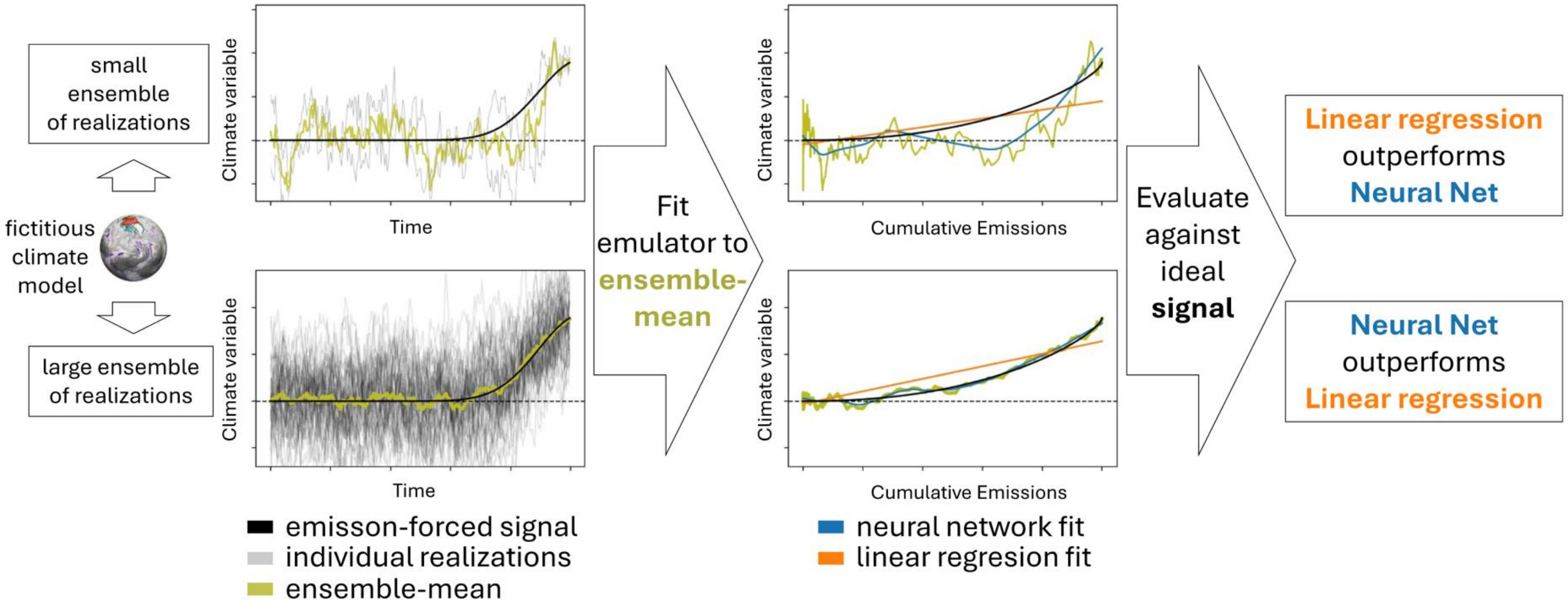
MLJC overview

- Eric, Will, and Andy
- Paper discussions — interesting developments for ML in atmos/climate science
  - “Science”-focused papers (e.g., this one)
  - Methodology-focused papers
    - If leading a methodology-focused paper, present like nobody has read the paper
- Watch parties
- Talks (your research or someone else’s)
- Workshops
  - Summer workshops
  - Open-to-all workshop in winter quarter (?)

Volunteers for future meetings

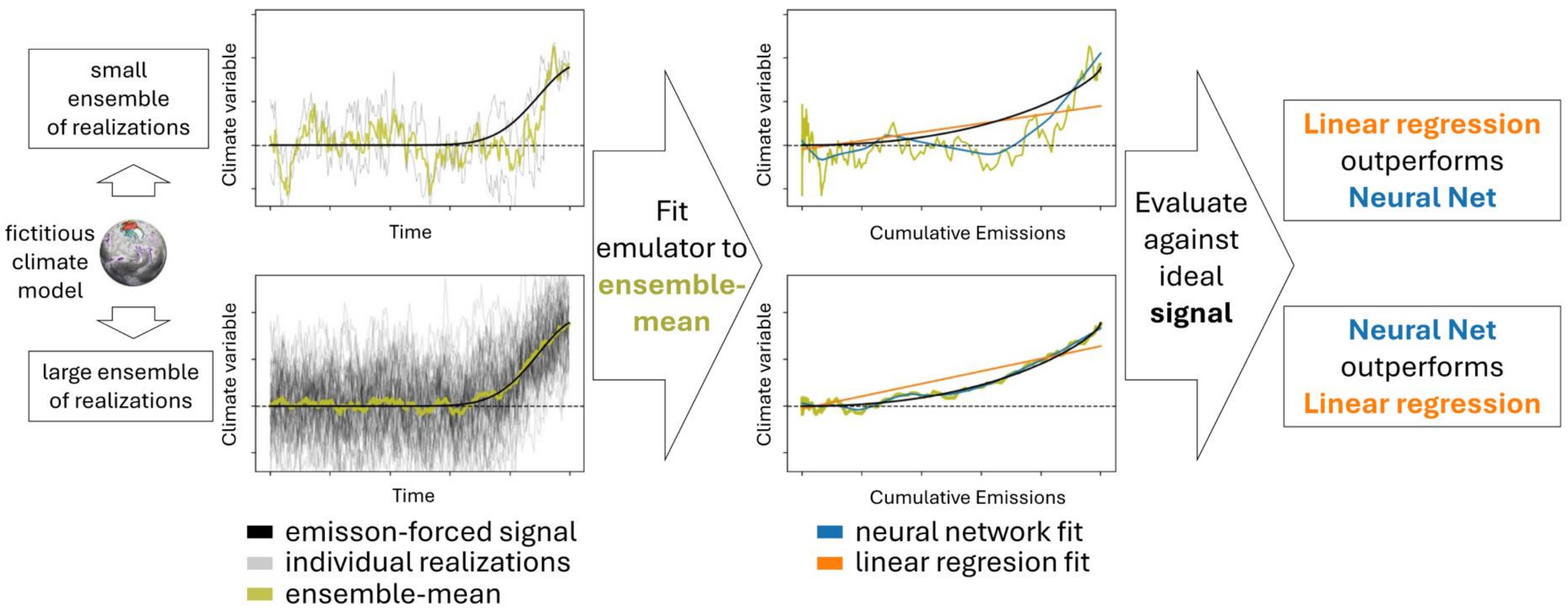
- Oct 31
  - Leader: Will
  - Pizza: Eliot
- Nov 14
  - Leader: Andy!
  - Pizza: Eric
- Dec 5
  - Leader:
  - Pizza:

Chat with your neighbor ~3 mins



- What was the main point of the paper?
- Did you like the paper? Why?
  - Was it easy to understand?
- What additional info do you wish you had?

Main point



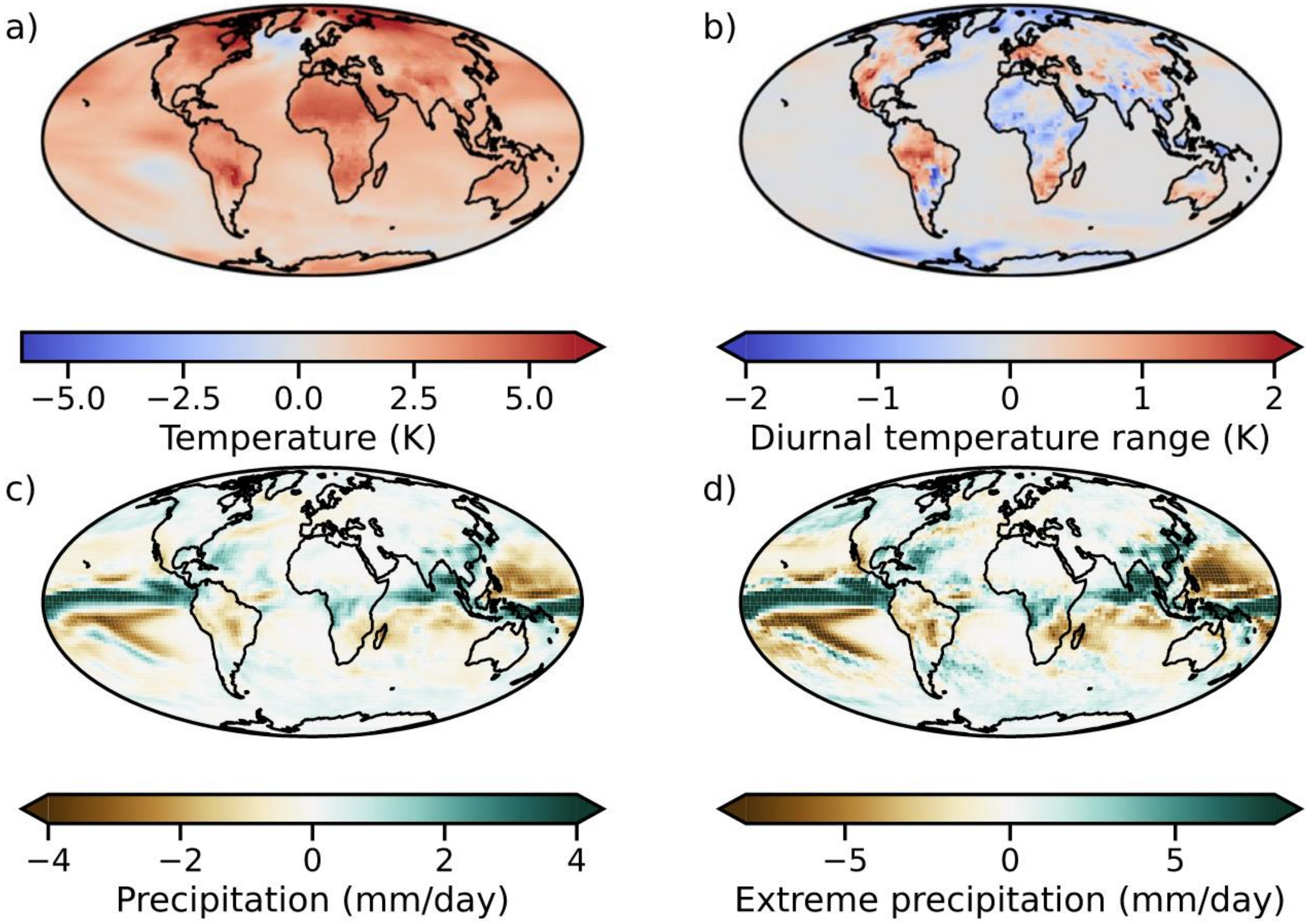
- Few ensemble members in ClimateBench leads to overfitting
    - Linear pattern scaling does better for many (nonlinear variables)
- Need a lot of ensemble members for benchmarking!
- (unstated) Evaluation against null hypothesis model necessary

<5 min. overview of necessary methods

- ClimateBench
- Linear pattern scaling
- CNN-LSTM
- Bias-variance tradeoff

ClimateBench

- Hist- and ssp- runs
- Inputs of CO_2, methane, SO_2, and black carbon
- Training + predicting on ensemble mean
  - ssp2-4.5 held out
  - **3 members**
- Compare to MPI-ESM1.2-LR (this study)
  - **50 members**



Watson-parris et al. (2022)

Linear pattern scaling (LPS)

Global avg. temperature scales w/ cumulative CO_2

Cum. CO_2

$$\widehat{\overline{T}}^{\text{surf}}_{t,e} = w^{\text{global}} x_{t,e} + b^{\text{global}}$$

Regional patterns scale w/ global avg. temperature

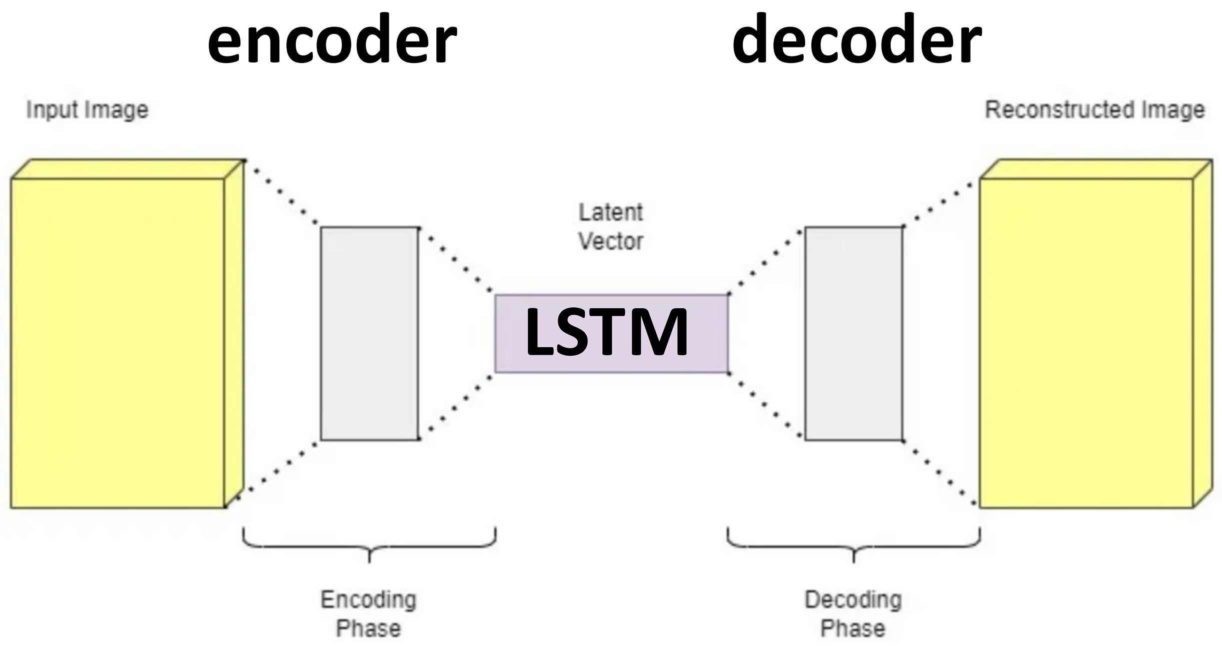
$$\hat{y}_{i,j,t,e} = w^{\text{local}}_{i,j} \widehat{\overline{T}}^{\text{surf}}_{t,e} + b^{\text{local}}_{i,j}$$

Fixed pattern

- Params = 2 * (# grid cells)
- Not sure why Watson-Parris et al. (2022) didn't use this model?

CCN-LSTM (encoder-LSTM-decoder)

- Includes all inputs
  - 20 filters of shape (3, 3)
  - Pooling removes spatial information
  - Latent LSTM state is 20
- 10-year memory
- No hyperparameter optimization



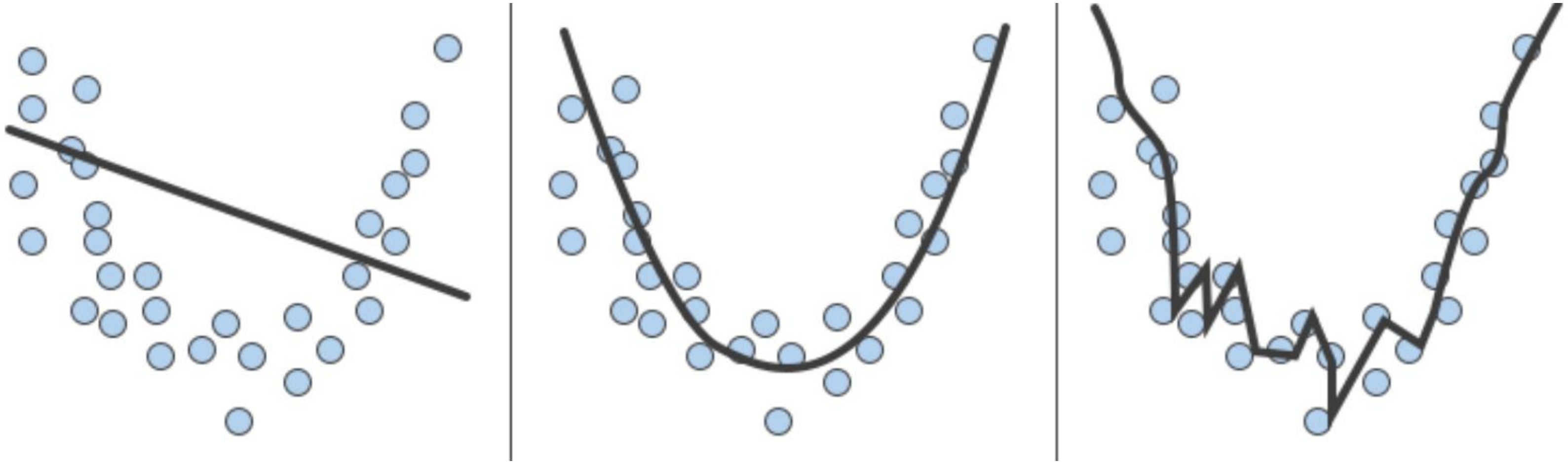
Bias-variance tradeoff: when do I use what model?

$$\mathrm{Bias}(\hat{y}) = \mathbb{E}[\hat{y}] - y$$

$$\mathrm{Var}(\hat{y}) = \mathbb{E}\left[(\hat{y} - y)^2\right]$$

$$\mathrm{MSE}(\hat{y}) = (\mathrm{Bias}(\hat{y}))^2 + \mathrm{Var}(\hat{y})$$

High bias



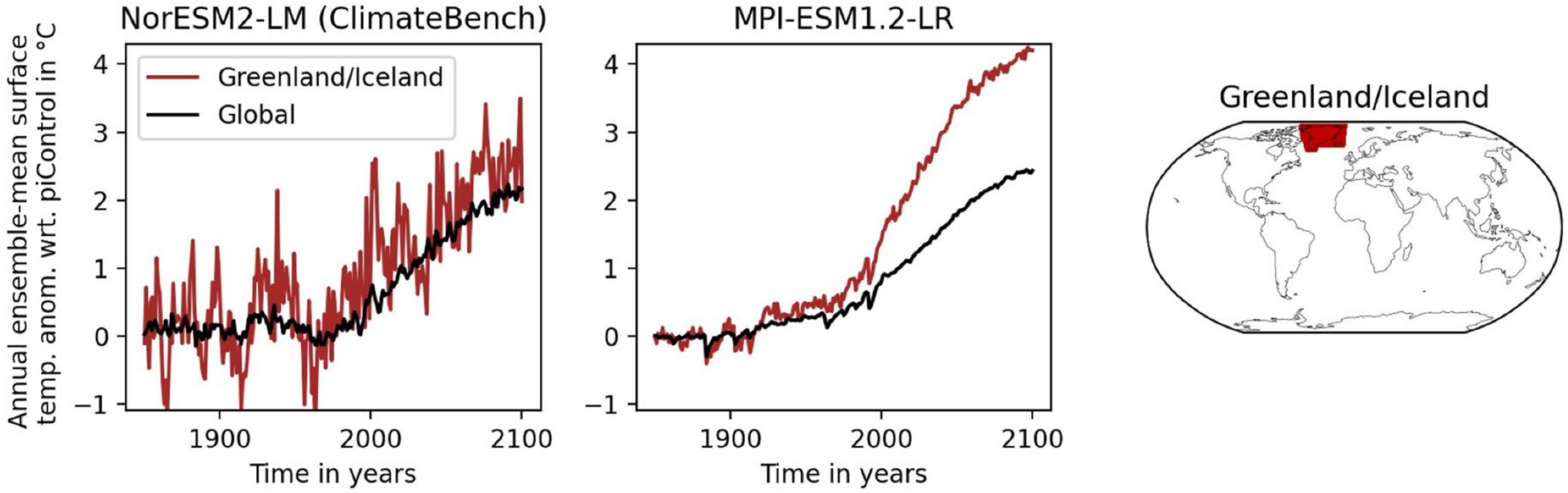
High variance

Results

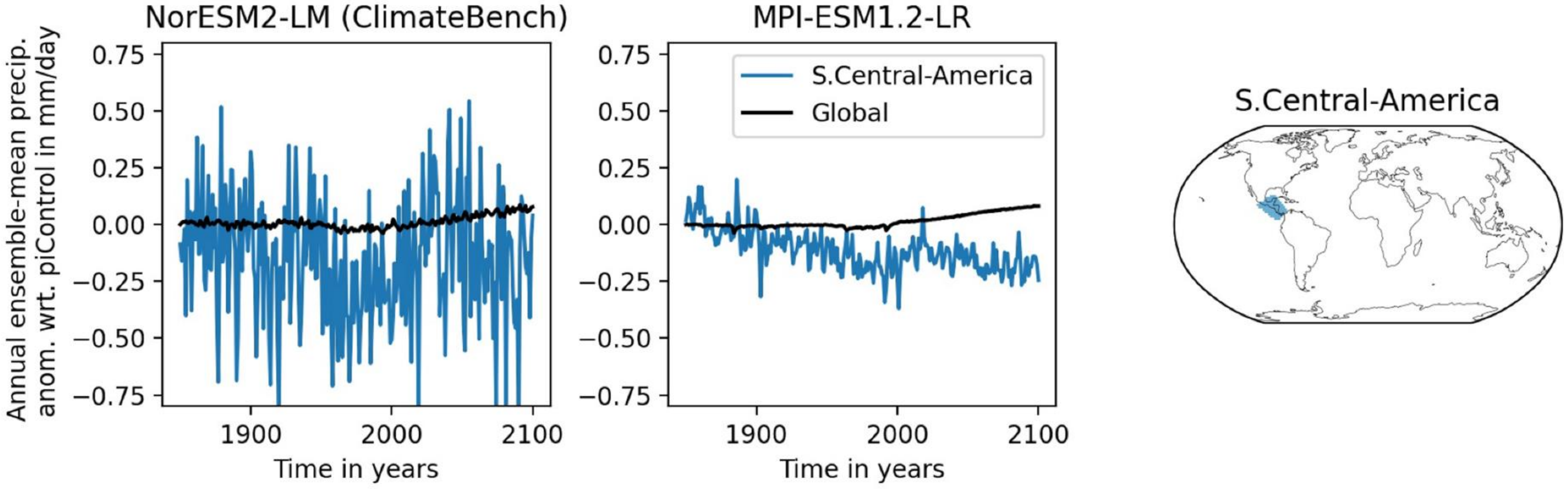
- Why use a large ensemble over a small ensemble?
- LPS vs. heavy-hitters
- Small vs. large ensemble theory

Why use a large ensemble? (Fig. 2)

small **large**



(a) Surface temperature anomalies



(b) Precipitation anomalies

LPS performance validation (Fig. 4)

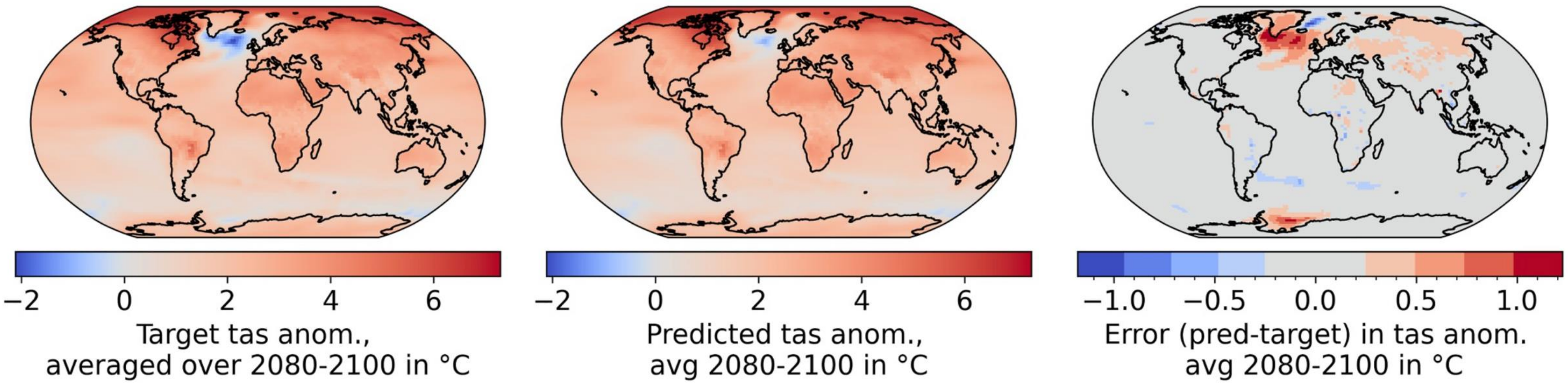


Figure 4. Linear pattern scaling error map. The left plot shows the target surface temperature anomalies (tas) from the ssp245 ClimateBench test set, which are averages over 3 realizations and 21 years (2080–2100). The middle plot shows the linear pattern scaling predictions and the right plot the error of predictions minus the target. The other variables are plotted in Figures C1–C3.

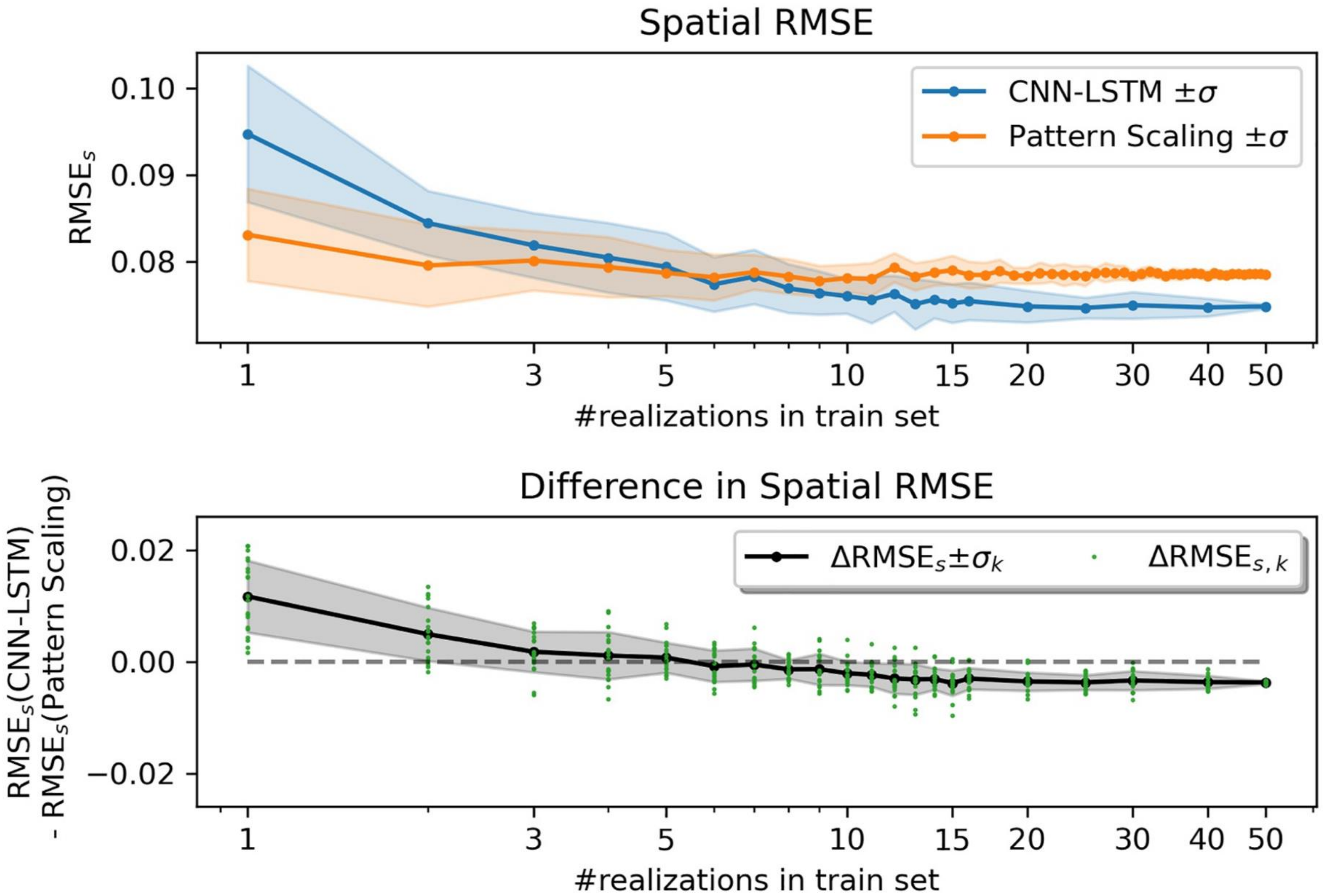
LPS on ClimateBench

Table 1
ClimateBench Results Table Including Linear Pattern Scaling

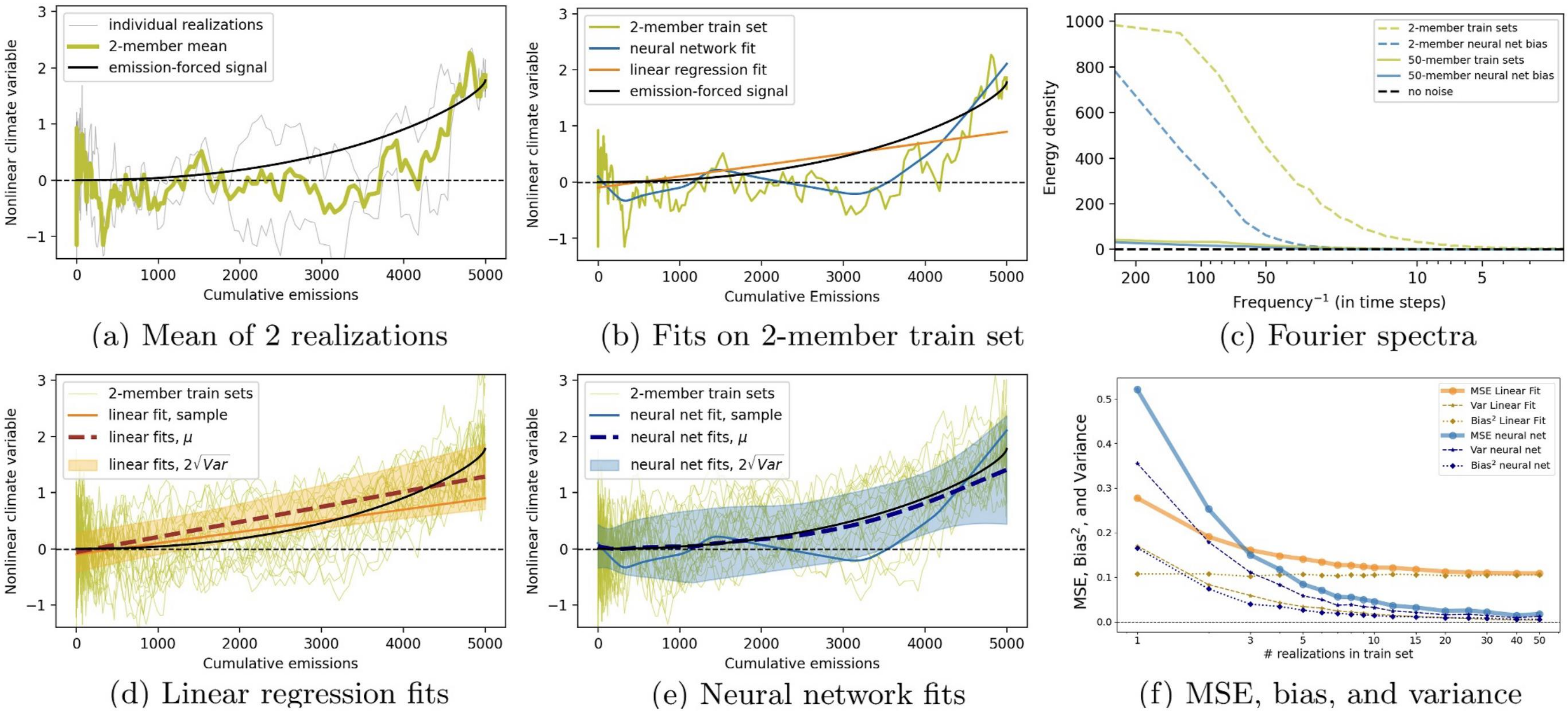
	Ref.	#param	Surface temperature			Diurnal temperature range			Precipitation			90th percentile precipitation		
			Spatial	Global	Total	Spatial	Global	Total	Spatial	Global	Total	Spatial	Global	Total
Gaussian process	WP22	n/a	0.109	0.074′	0.478	9.21	2.68	22.6	2.34	0.341	4.05	2.56	0.429	4.70
CNN-LSTM	WP22	365 K	0.107	0.044′	0.327	9.92	1.38	16.8	2.13	0.209	3.18	2.61	0.346	4.34
CNN-LSTM (reproduced)	N23	365 K	0.123	0.080′	0.524	7.47	1.23	13.6	2.35	**0.151**	**3.10**	3.11	**0.282**	4.52
Random forest	WP22	47.5 K	0.108	0.058′	0.4′	9.20	2.65	22.5	2.52	0.502	5.04	2.68	0.543	5.40
Cli-ViT	N23	unavail.	0.086′	0.044′	0.305	7.00	1.76	15.8	2.22	0.241	3.43	2.80	0.329	4.45
ClimaX	N23	108M	0.085′	0.043′	0.297	**6.69**	**0.810**	**10.7**	2.19	0.183	3.11	2.68	0.342	4.39
Linear pattern scaling	ours	27.7 K	**0.0786**	**0.0410**	**0.284**	8.02	2.15	18.8	**1.87**	0.268	3.20	**2.25**	0.357	**4.03**
Target std. dev.	WP22	–	0.052′	0.072′	0.414	2.51	1.49	9.97	1.35	0.268	2.69	1.76	0.457	4.04

- Why does LPS perform better?
- Is this a fair comparison?

LPS vs. CNN-LSTM on large ensemble (Fig. 5, precip)



Theory (Fig. 7)



(a) Mean of 2 realizations

(b) Fits on 2-member train set

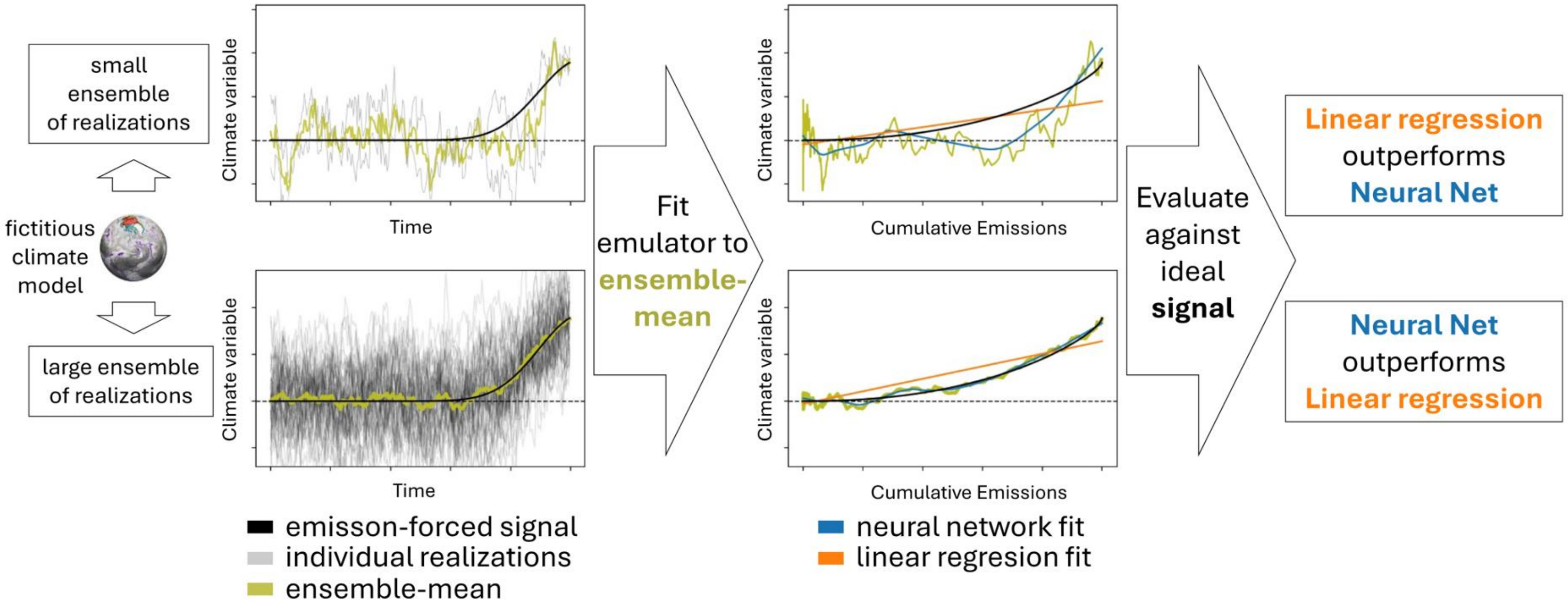
(c) Fourier spectra

(d) Linear regression fits

(e) Neural network fits

(f) MSE, bias, and variance

Chat with your neighbor ~3 mins



- What was the main point of the paper?
- **Did you like the paper? Why?**
  - **Was it easy to understand?**
- **What additional info do you wish you had?**
  - Out-of-sample forcings (less correlated input space)