More questions than answers:

Linking AI, climate, and statistics to interpret AI weather models

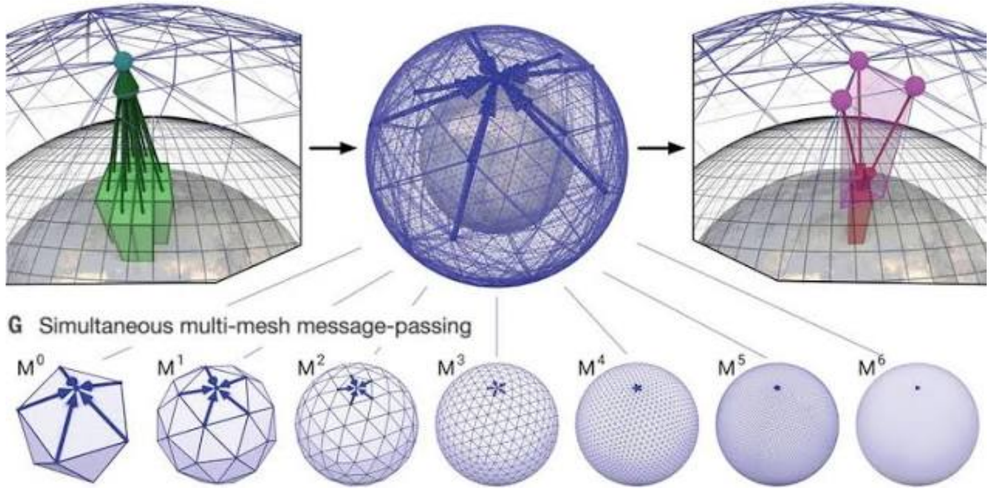
Karen McKinnon, UCLA → JHU

September 28, 2026

IMSI workshop on Unpacking AI Weather Models

Thanks to: Nisha Chandramoorthy and Team Dagstuhl (Claire Monteleoni, Aiden Price, Nico Bianco, and Nicole Ludwig) for many conversations

Data-driven AI weather models are fundamentally different from the true system, and from numerical models



Lam et al (2023), *Science*

?

Can we trust AI weather models?

Can we trust AI weather models?

= are they doing what we expect?

Predictability: do AI weather models give us realistic estimates of the predictability in the climate system?

Causality: do weather events happen for the right reasons?

Probability: do emulators capture the correct distribution of weather events, particularly extremes?

Predictability: do AI weather models give us realistic estimates of the predictability in the climate system?

Do the predictions exhibit appropriate spread, and what is its origin and interpretation?

Predictability: do AI weather models give us realistic estimates of the predictability in the climate system?

Do the predictions exhibit appropriate spread, and what is its origin and interpretation?

Primary focus: generative / diffusion models

$$\Delta X_t \sim p_\theta(\Delta X_t \mid X_t, X_{t-1}, X_{t-2}, \dots)$$

Test system: two-scale Lorenz 96

$$\frac{dX_k}{dt} = X_{k-1}\left(X_{k+1} - X_{k-2}\right) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k},$$

$$\frac{dY_{j,k}}{dt} = cb\, Y_{j+1,k}\left(Y_{j-1,k} - Y_{j-2,k}\right) - cY_{j,k} + \frac{hc}{b}X_k\,.$$

Test system: two-scale Lorenz 96

$$\frac{dX_k}{dt} = \boxed{X_{k-1}}\left(X_{k+1} - X_{k-2}\right) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k},$$

$$\frac{dY_{j,k}}{dt} = cb\, Y_{j+1,k}\left(Y_{j-1,k} - Y_{j-2,k}\right) - cY_{j,k} + \frac{hc}{b}X_k\,.$$

X: slow scale variables

Test system: two-scale Lorenz 96

$$\frac{dX_k}{dt} = X_{k-1}\left(X_{k+1} - X_{k-2}\right) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k},$$

$$\frac{dY_{j,k}}{dt} = cb\, Y_{j+1,k}\left(Y_{j-1,k} - Y_{j-2,k}\right) - cY_{j,k} + \frac{hc}{b}X_k\,.$$

Y: fast scale variables (unresolved)

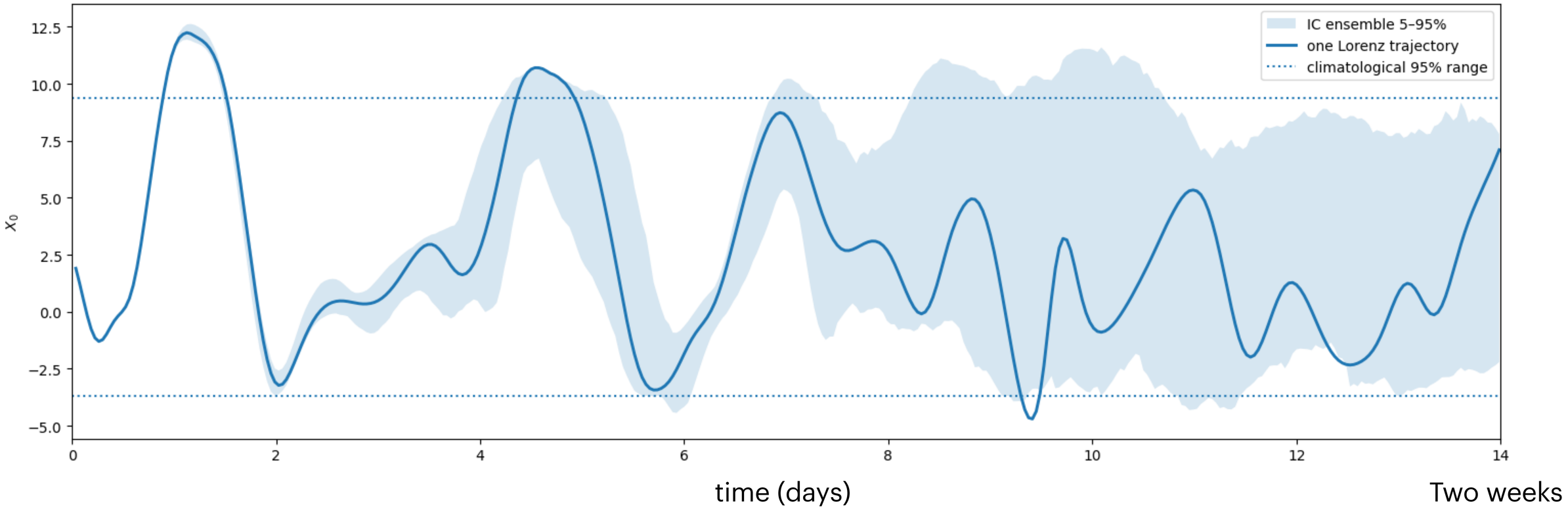
Test system: two-scale Lorenz 96

$$\frac{dX_k}{dt} = X_{k-1}\left(X_{k+1} - X_{k-2}\right) - X_k + F - \boxed{\frac{hc}{b}\sum_{j=1}^{J} Y_{j,k}},$$

$$\frac{dY_{j,k}}{dt} = cb\, Y_{j+1,k}\left(Y_{j-1,k} - Y_{j-2,k}\right) - cY_{j,k} + \frac{hc}{b}X_k\,.$$

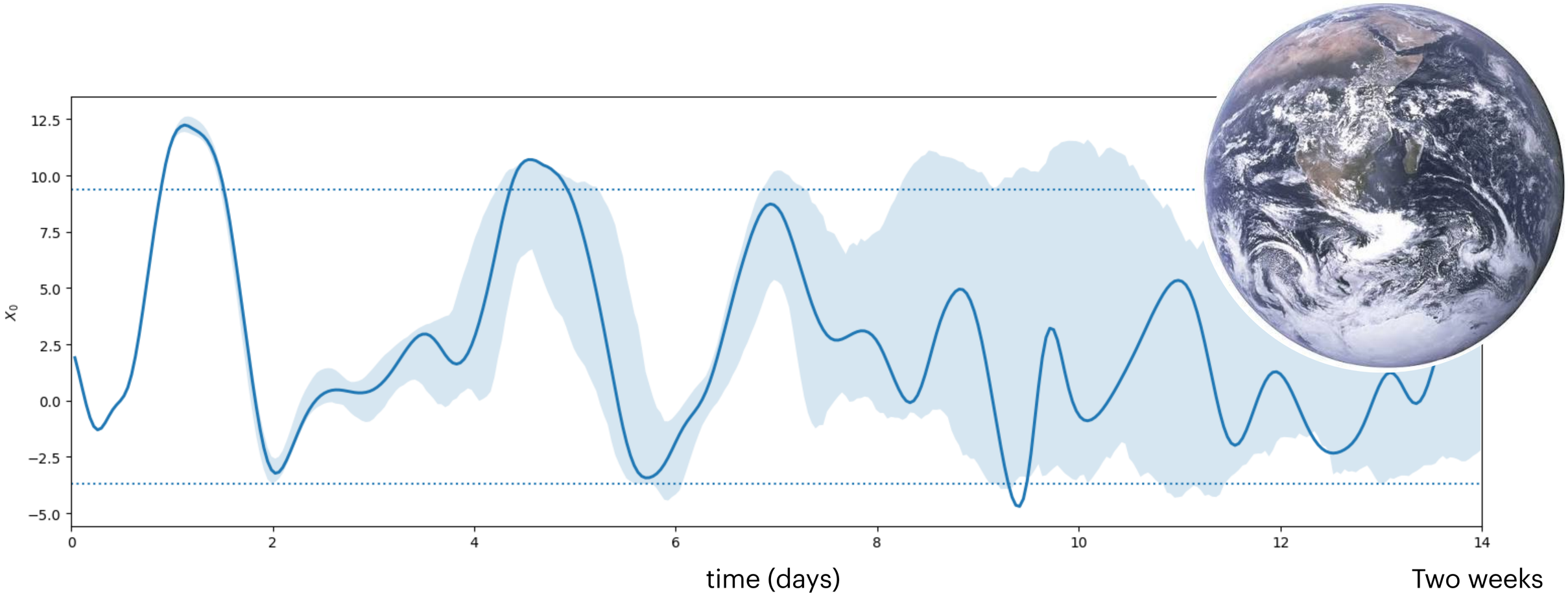
The thing we need to parameterize

In L96, the system is **chaotic, Markovian, and deterministic**



The only source of prediction uncertainty is uncertainty in the initial conditions

In L96, the system is **chaotic, Markovian, and deterministic**



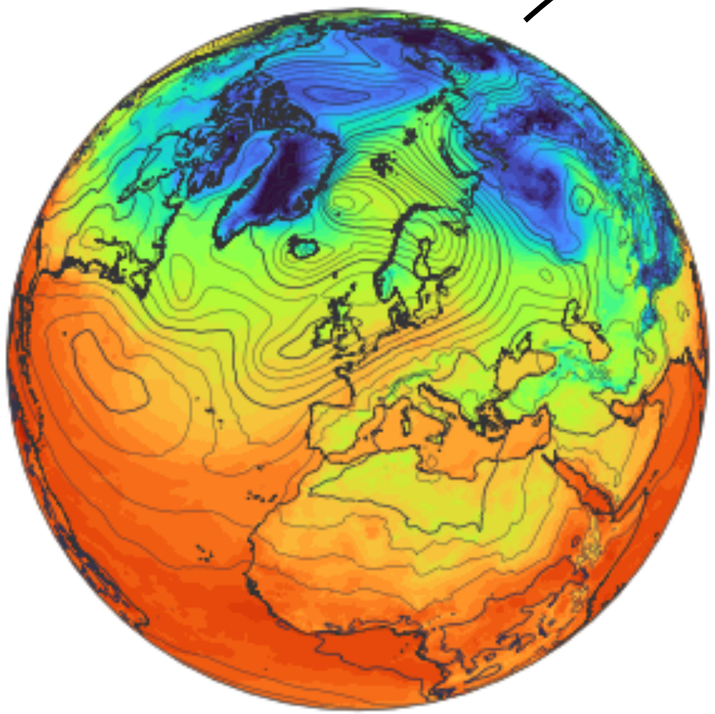
The only source of prediction uncertainty is uncertainty in the initial conditions

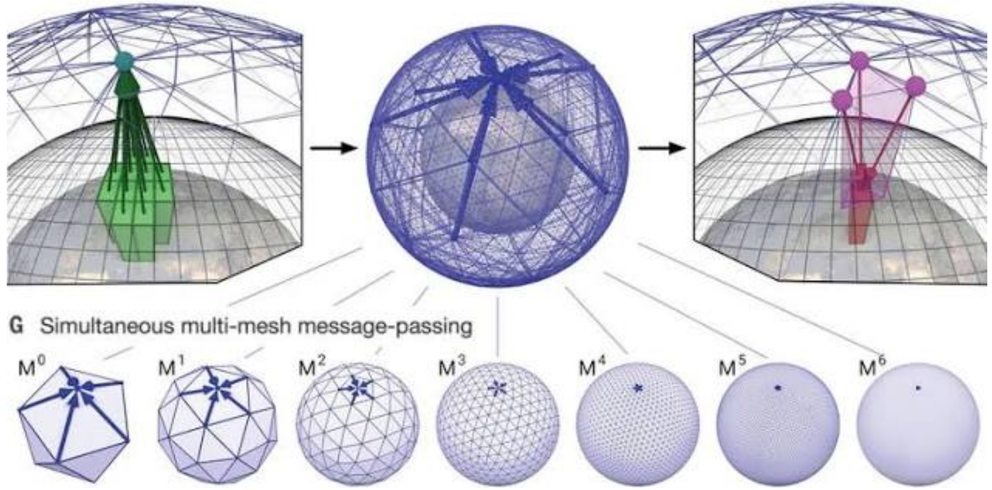
In contrast, generative AI models are inherently **stochastic**, and potentially **non-Markovian**

$$\Delta X_t \sim p_\theta(\Delta X_t \,|\, X_t, X_{t-1}, X_{t-2}, \dots)$$

In the data-driven system, we:

1. Observe a coarsened version of the true system (here, only 36 X's, and no Y's)
2. Do not have perfect representations of the initial conditions in the coarsened variables (true in NWP as well)
3. Train on a finite set of data





In the data-driven system, we:

1. **Observe a coarsened version of the true system (here, only 36 X's, and no Y's)**
2. Do not have perfect representations of the initial conditions in the coarsened variables (true in NWP as well)
3. Train on a finite set of data

Test system: two-scale Lorenz 96 with a parameterization

$$\frac{dX_k}{dt} = X_{k-1}\left(X_{k+1} - X_{k-2}\right) - X_k + F - \boxed{U_k(X)}$$

Mori-Zwanzig formalism: if we train a model only using the coarse dynamics (X), the unobserved dynamics induce **memory** and **stochasticity**

$$\frac{dX(t)}{dt} = F_{\text{det}}(X(t)) + \int_0^t K\Big(X(t-s), s\Big)\, ds + \eta(t),$$

Deterministic Memory Stochasticity

What is the origin of these new terms?

- The memory: due to missing X -> Y -> X dynamics
- The noise: due to unresolved initial state in Y

MZ: Stochasticity in GenAI models emerges naturally due to the unresolved initial state

$$\frac{dX(t)}{dt} = F_{\text{eff}}(X(t)) + \int_0^t K\Big(X(t-s), s\Big)\, ds + \eta(t)$$



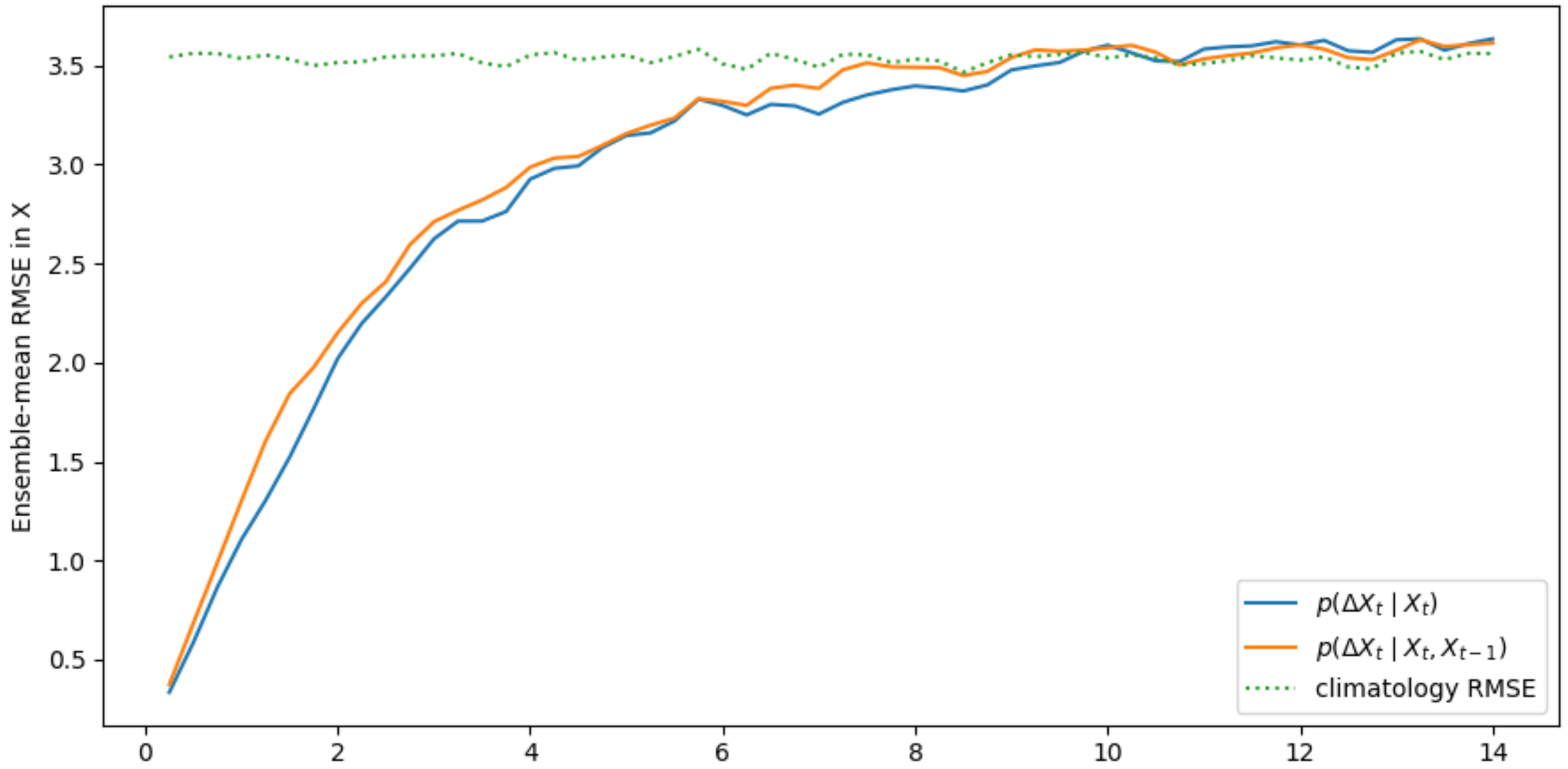
$$\Delta X_t \sim p_\theta(\Delta X_t \mid X_t, X_{t-1}, X_{t-2}, \ldots)$$

What about the memory? Are we leaving potential skill on the table through dropping the non-Markovian term?

$$\frac{dX(t)}{dt} = F_{\text{eff}}(X(t)) + \int_0^t K\Big(X(t-s), s\Big)\, ds + \eta(t)$$

$$\Delta X_t \sim p_\theta(\Delta X_t \mid X_t, X_{t-1}, X_{t-2}, \ldots)$$

At least for a diffusion model trained on L96, the memory does not provide added skill. True for a more complex system? Or when training on different frequencies of data?



In the data-driven system, we:

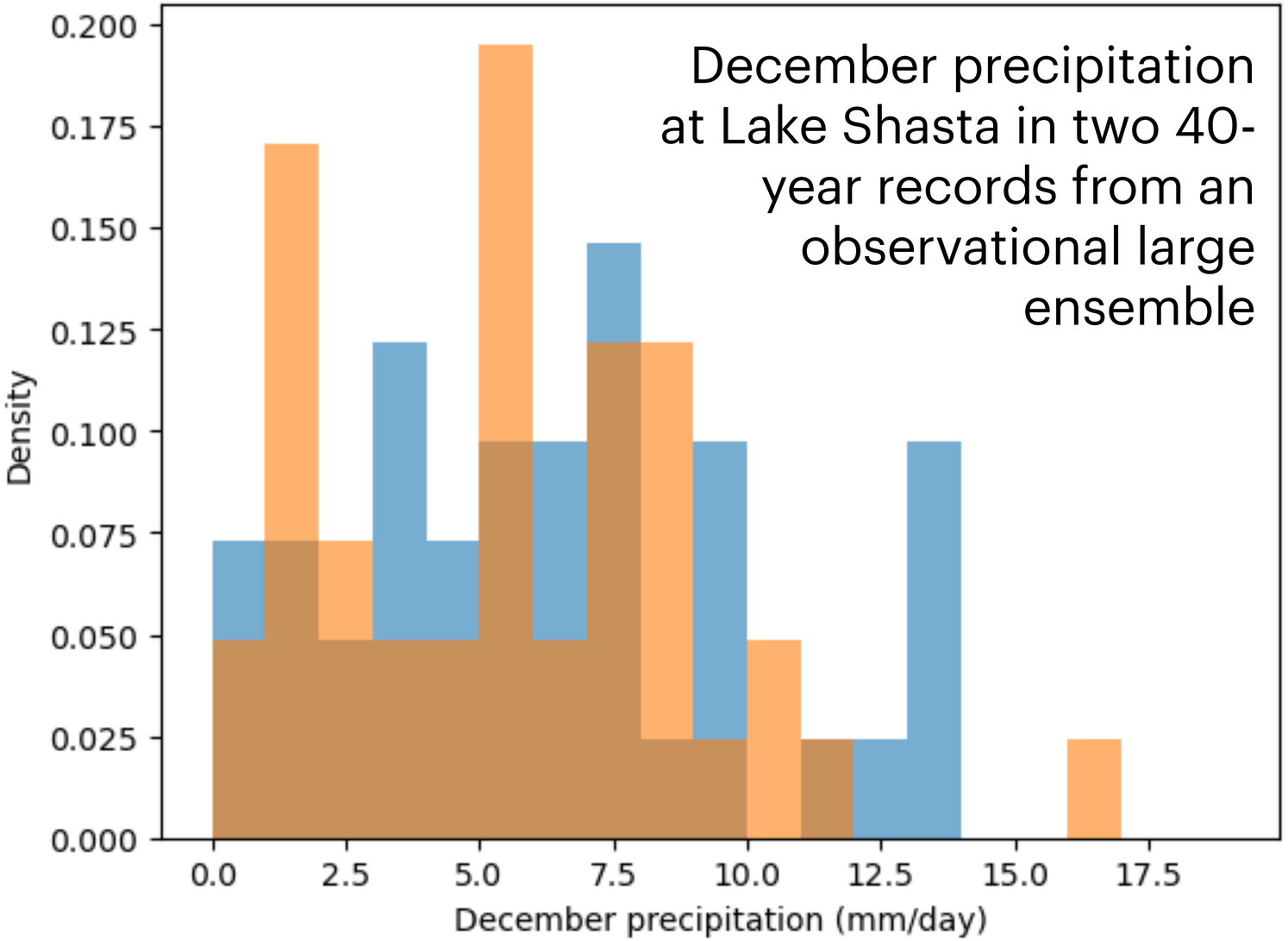
1. Observe a coarsened version of the true system (here, only 36 X's, and no Y's)
2. **Do not have perfect representations of the initial conditions in the coarsened variables (true in NWP as well)**
3. Train on a finite set of data

	Uncertain large-scale ICs	**Uncertain small-scale ICs**
GenAI models	x	x
Deterministic AI models	x	

In the data-driven system, we:

1. Observe a coarsened version of the true system (here, only 36 X’s, and no Y’s)
2. Do not have perfect representations of the initial conditions in the coarsened variables (true in NWP as well)
3. **Train on a finite set of data**

The climate system is not well-sampled with 40 years of data



To what extent can AI models produce events that are fully consistent with the same system, but outside the training data?

If they can, why?

If they can’t, what additional information can we add?

Although imperfect, we have decades of knowledge and associated trust in NWP models (and to some extent climate models).

AI models are fundamentally different, but drawing parallels to our existing physical models may be helpful for trust building.

The Mori-Zwanzig formalism explains why a stochastic, potentially non-Markovian system should emerge from training on the coarse scales alone.

Notably, deterministic AI models do not include this source of uncertainty.

Climate science has showed us that the statistics of weather can differ (even in a stationary climate) across 40-year samples, and we should understand how this affects our AI weather model predictions.

kmckinnon@ucla.edu
karen.mckinnon@jhu.edu

karenamckinnon

website

Although imperfect, we have decades of knowledge and associated trust in NWP models (and to some extent climate models).

AI models are fund r existing physical models may be hel

The Mori-Zwanzig f y non-Markovian system should eme

Notably, determinis certainty.

Climate science has showed us that the statistics of weather can differ (even in a stationary climate) across 40-year samples, and we should understand how this affects our AI weather model predictions.

Come work with me on these questions!

Currently hiring a postdoc and admitting a PhD student (JHU, Env Engineering) to work on related topics.

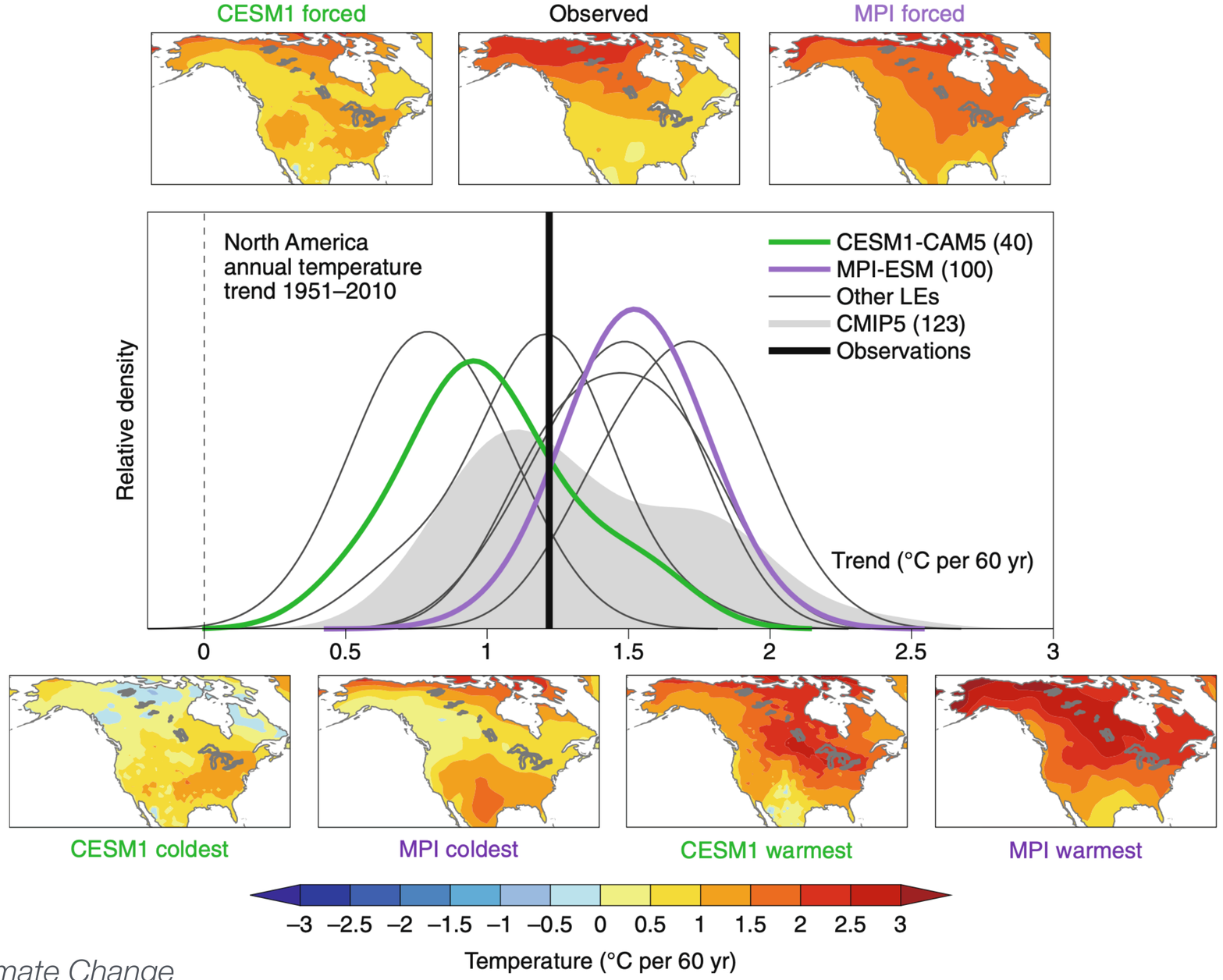
kmckinnon@ucla.edu
karen.mckinnon@jhu.edu

karenamckinnon

website

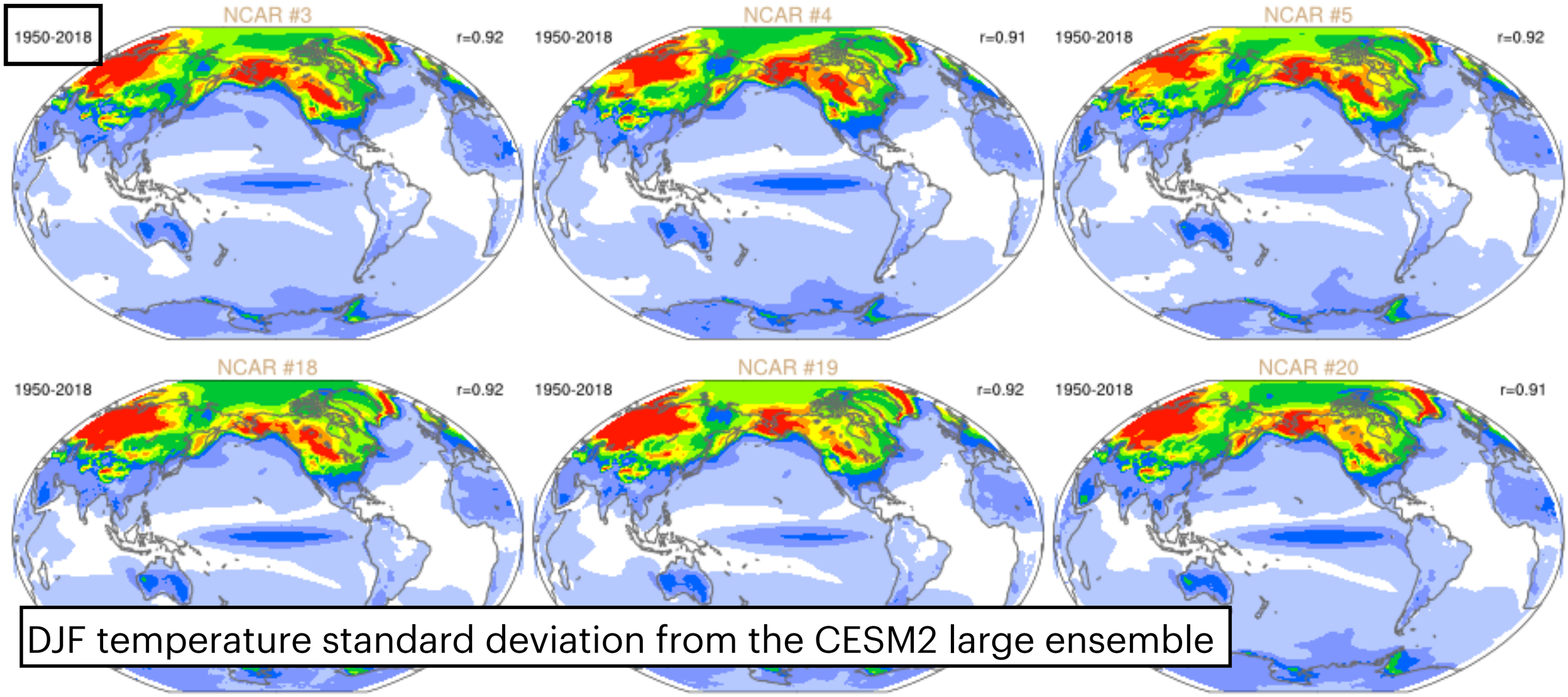
Extras

Sampling of internal variability can lead to very different climate trends over many decades.



Deser et al. (2020) *Nature Climate Change*

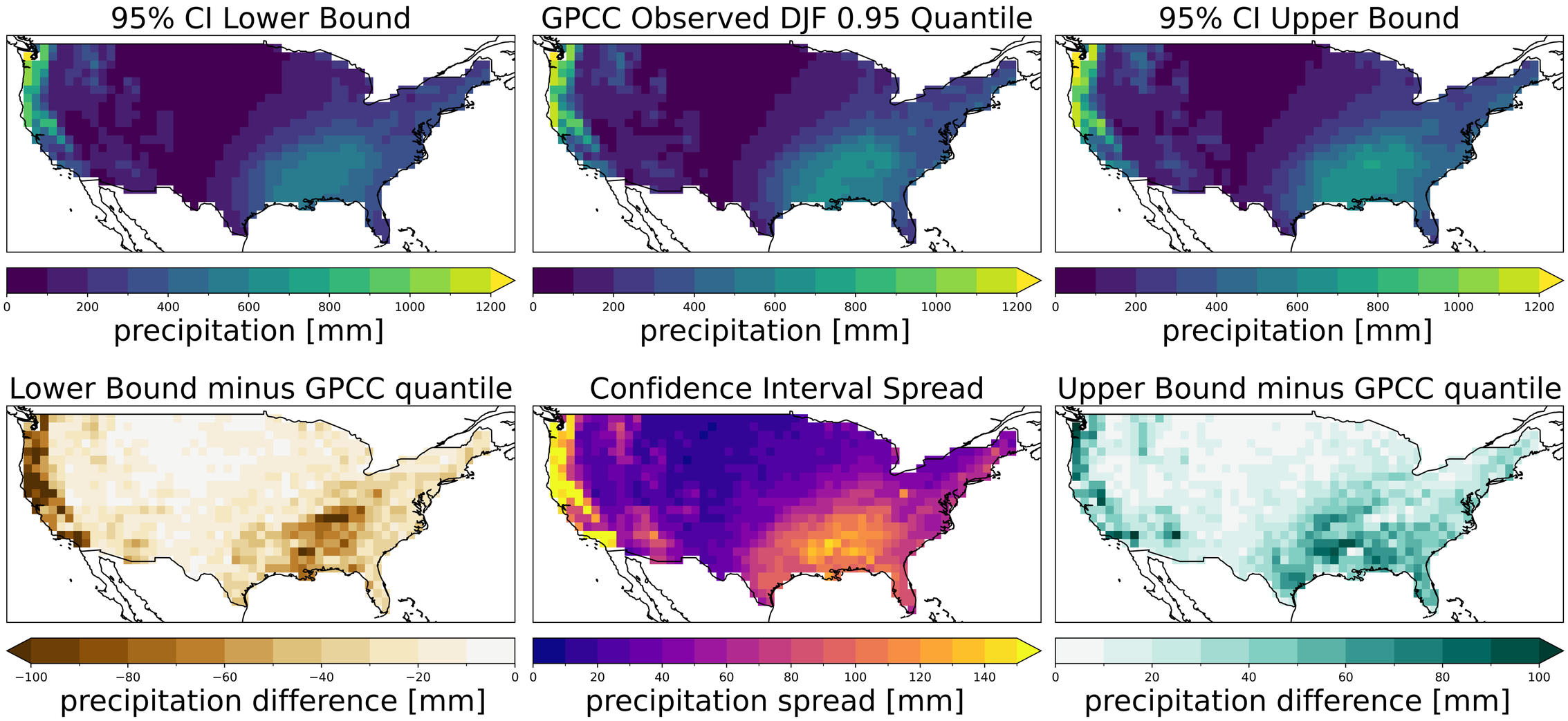
Even 70-year samples can give different estimates of “climate-scale” statistics



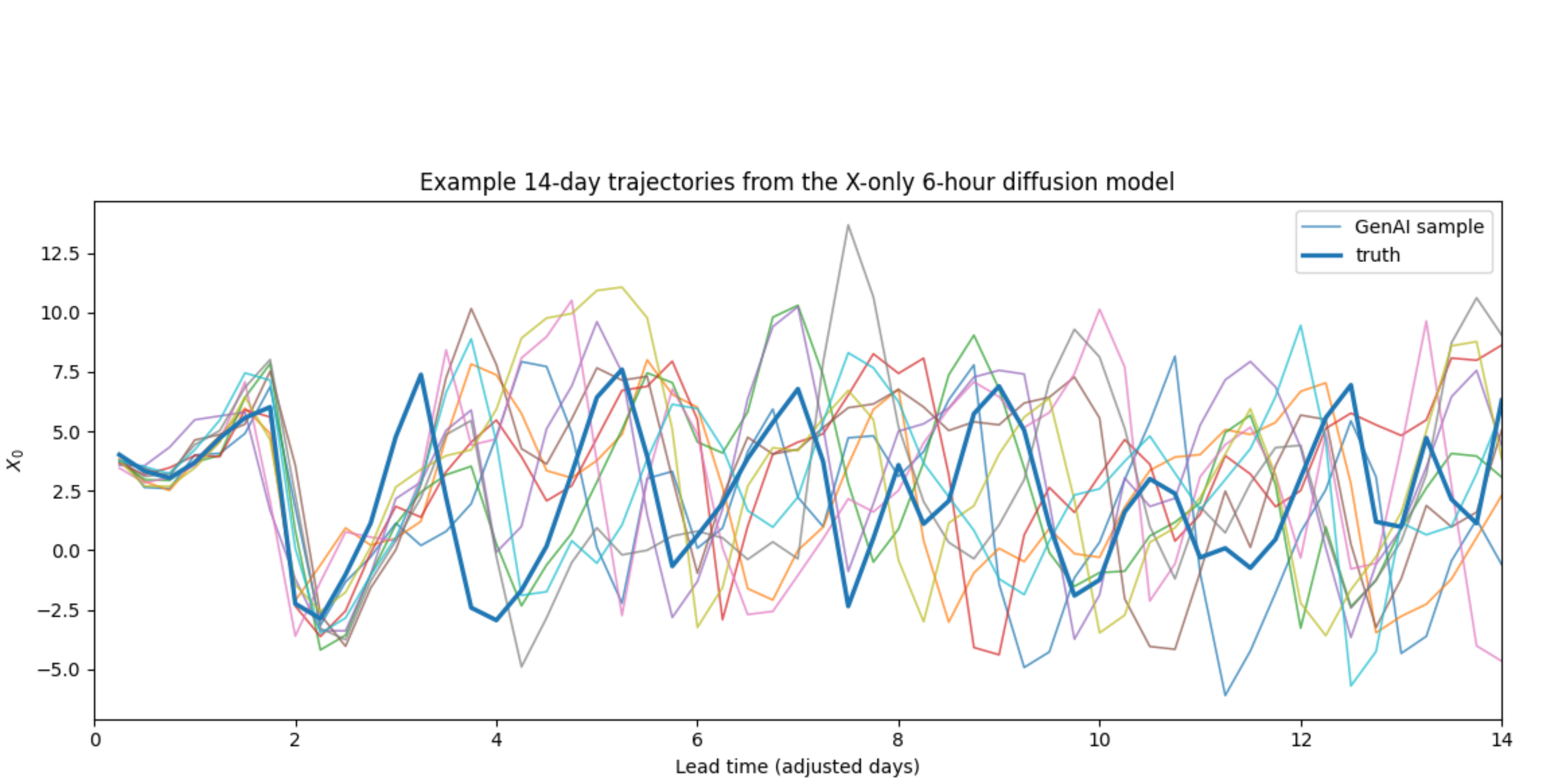
DJF temperature standard deviation from the CESM2 large ensemble

CVDP-LE

...and the role of internal variability increases for extremes and long-tailed distributions



McEvoy, McKinnon, and Poppick, submitted


Example 14-day trajectories from the X-only 6-hour diffusion model
GenAI sample
truth
X_0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
0
2
4
6
8
10
12
14
Lead time (adjusted days)

We believe the true atmosphere is a **deterministic, chaotic system.**

We model the atmosphere using systems of equations that are **deterministic** and **chaotic**, and potentially **stochastic** to account for uncertainty in the model parameterizations.

Data-driven AI models are then (typically) trained on a coarse subset of output from a numerical model (including reanalyses), and represent a different dynamical system. In place of the original spread induced by chaos, ensemble spread can come from (larger) **initial condition spread**, resampling **model parameters**, and explicitly **stochastic predictions**.

What are the sources of uncertainty (forecast spread) in an AI weather ensemble?

Can we match them to a physically-interpretable uncertainty?

	Real system	**NWP**	**Deterministic AI**	**Generative AI**
Initial condition uncertainty	X	X	X	X
Parameter uncertainty		X	X	X
“Coarse-graining” uncertainty		X	X	X
Stochastic sampling				X