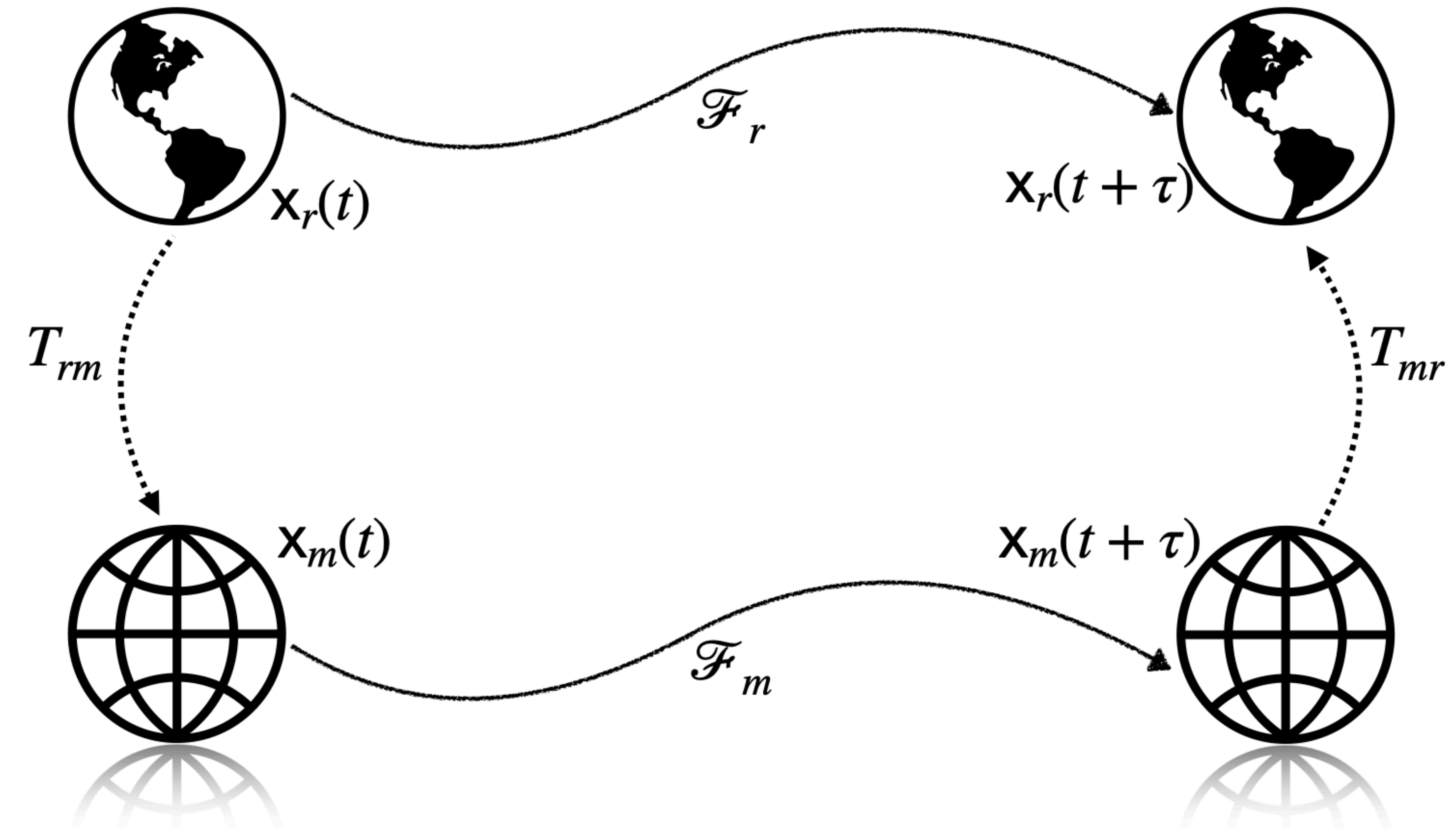
Improving forecasts by learning optimal transforms between dynamical systems and imperfect models

Dan Amrhein
NSF National Center for Atmospheric Research

Niraj Agarwal
NOAA / CIRES

Ian Grooms
University of Colorado



Outline

Forecasting with flawed models: “Fidelity” and “mapping” paradigms

A data-driven approach to optimizing pre- and post-processing operations

Application to Lorenz ’96

Application to CESM SMYLE

Outline

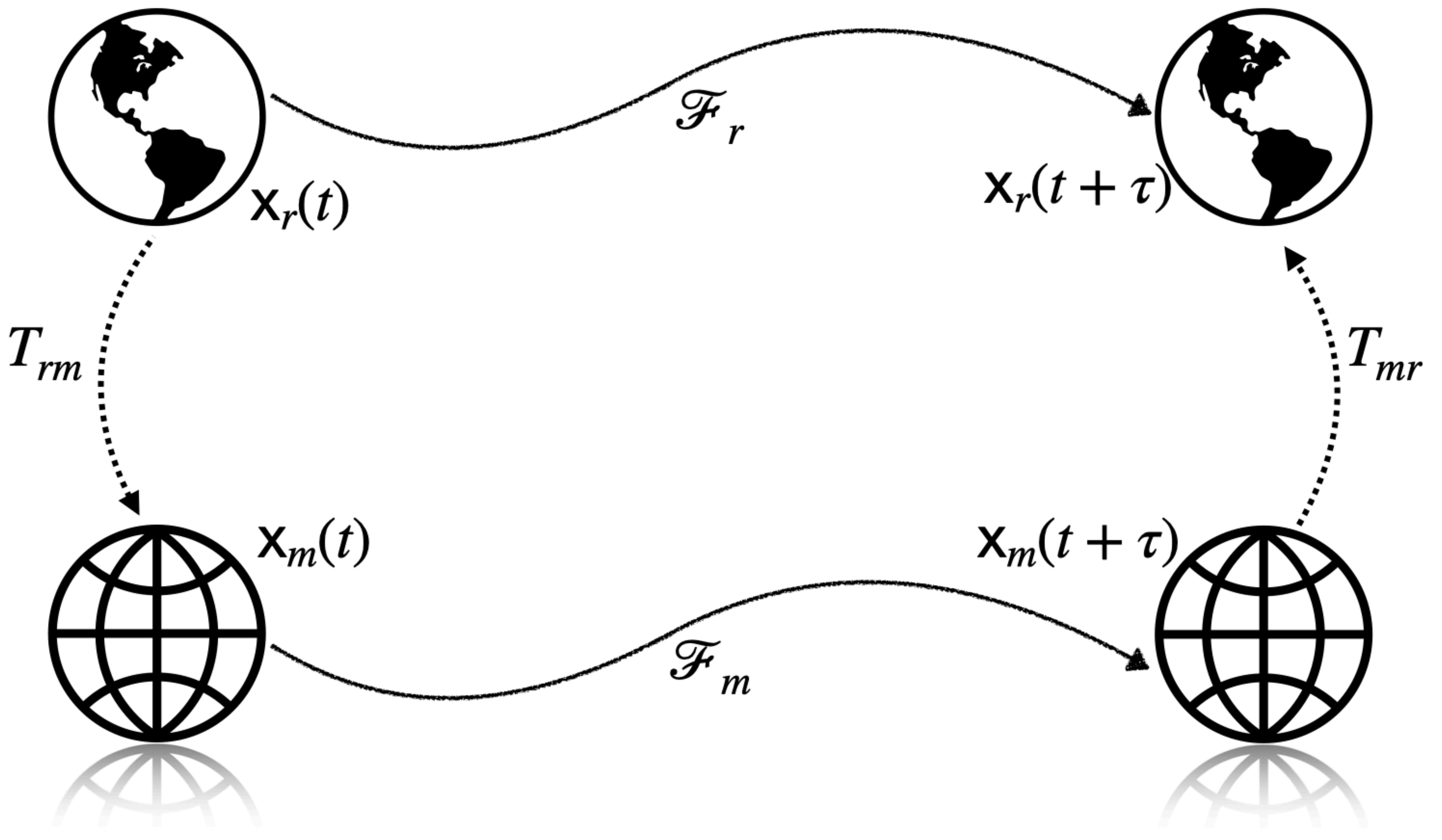
Forecasting with flawed models: “Fidelity” and “mapping” paradigms

A data-driven approach to optimizing pre- and post-processing operations

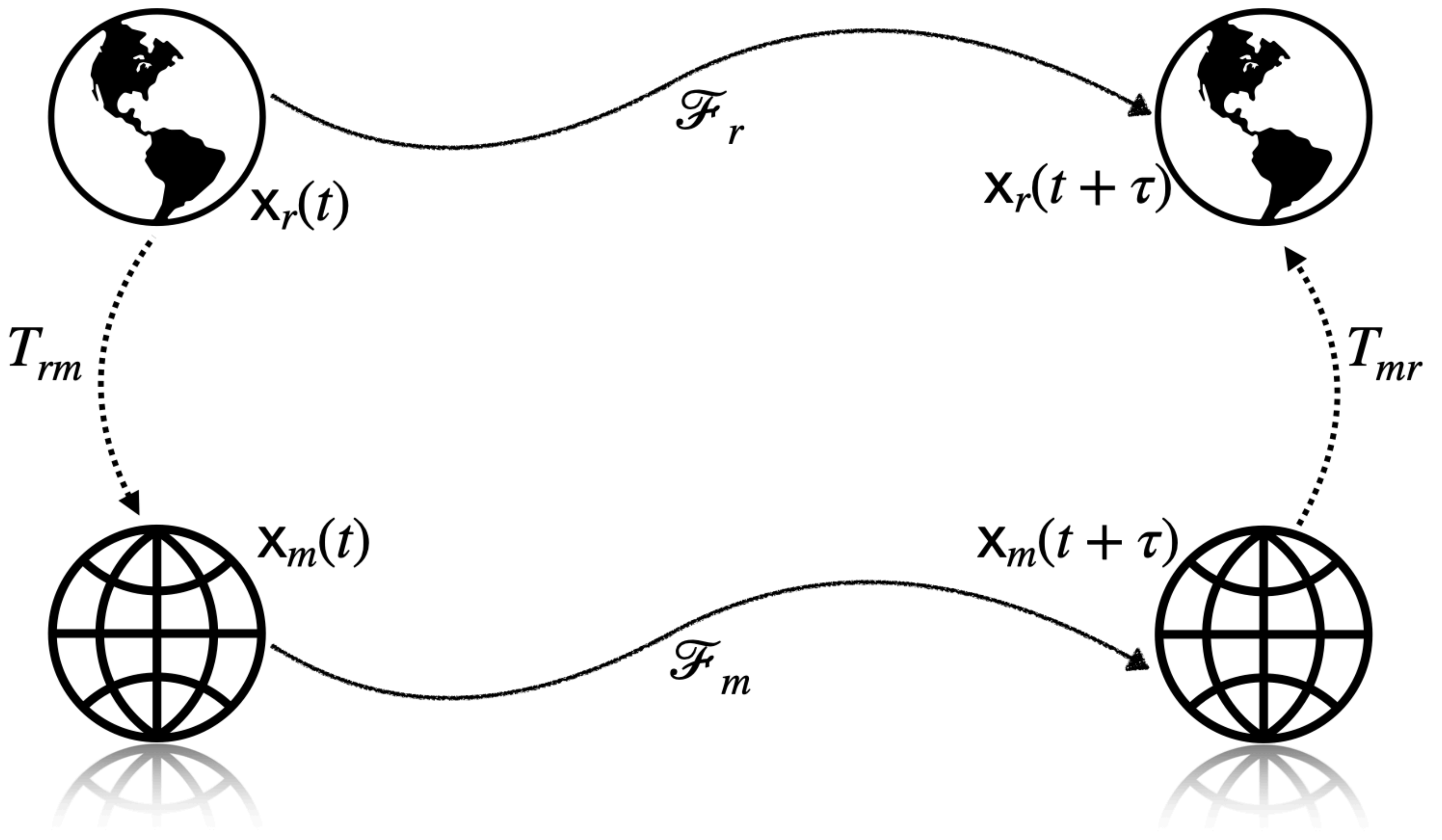
Application to Lorenz ’96

Application to CESM SMYLE

Forecasting with flawed models: "Fidelity" and "mapping" paradigms



Forecasting with flawed models: “Fidelity” and “mapping” paradigms



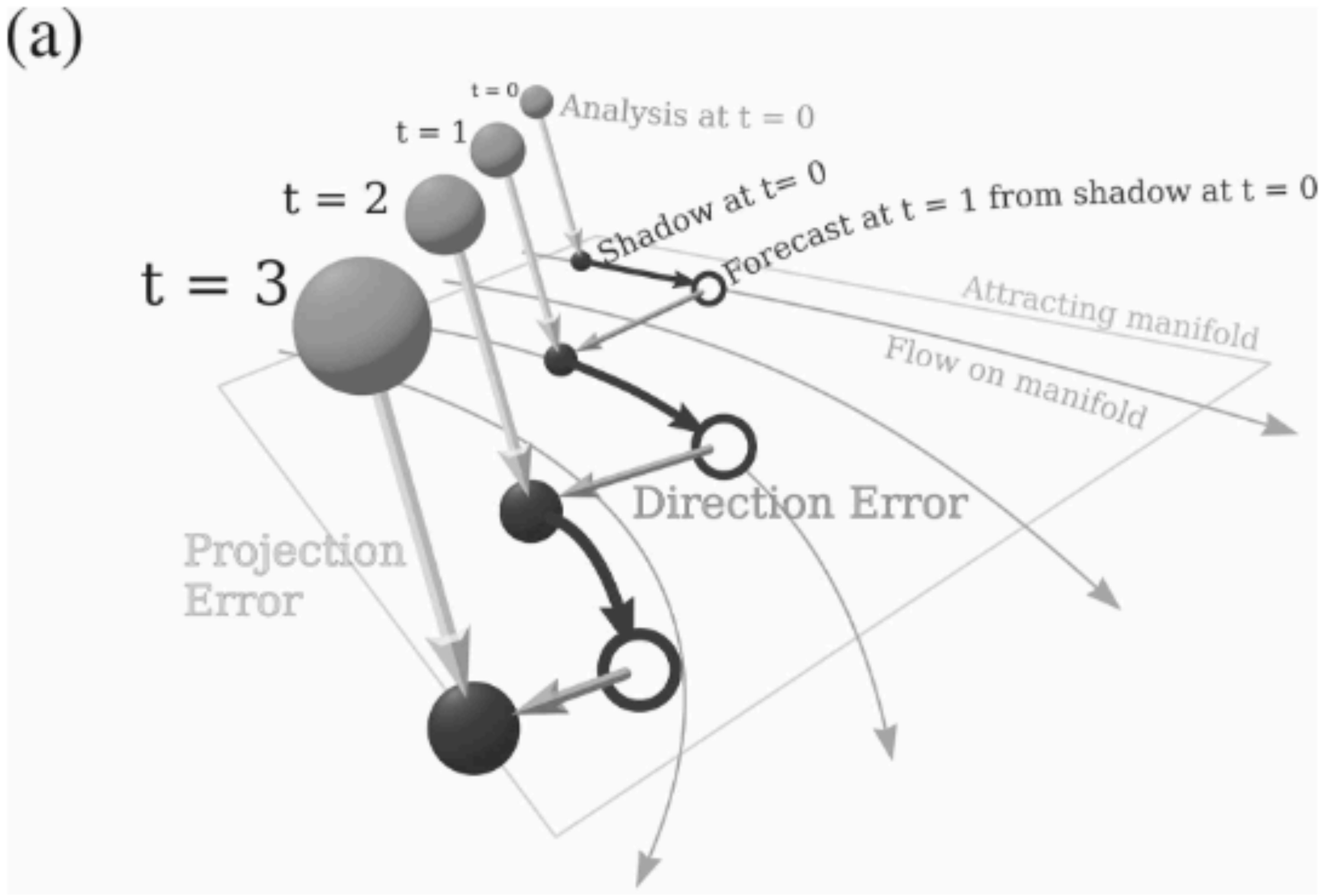
Data assimilation and numerical forecasting with imperfect models: The mapping paradigm

Zoltan Toth*, Malaquias Peña[1]

National Centers for Environmental Prediction (NCEP), Environmental Modeling Center, 5200 Auth Rd., Room 207, Camp Springs, MD 20746, United States

Available online 9 November 2006

The “fidelity” paradigm assumes that nature and the forecast model are the same and, therefore, starts the model with a state as close to the state of nature as possible. This assumption is true only in case of a perfect model (that exists only in simulated numerical experiments). By contrast, the “mapping” paradigm recognizes that numerical models of any system in nature are only imperfect images of reality and searches for a mapping that connects the states observed with the corresponding states in a model. The mapping is then used to find the model state that best represents a state of nature on or near the model attractor for model initialization.



Judd et al. 2008

Outline

Forecasting with flawed models: “Fidelity” and “mapping” paradigms

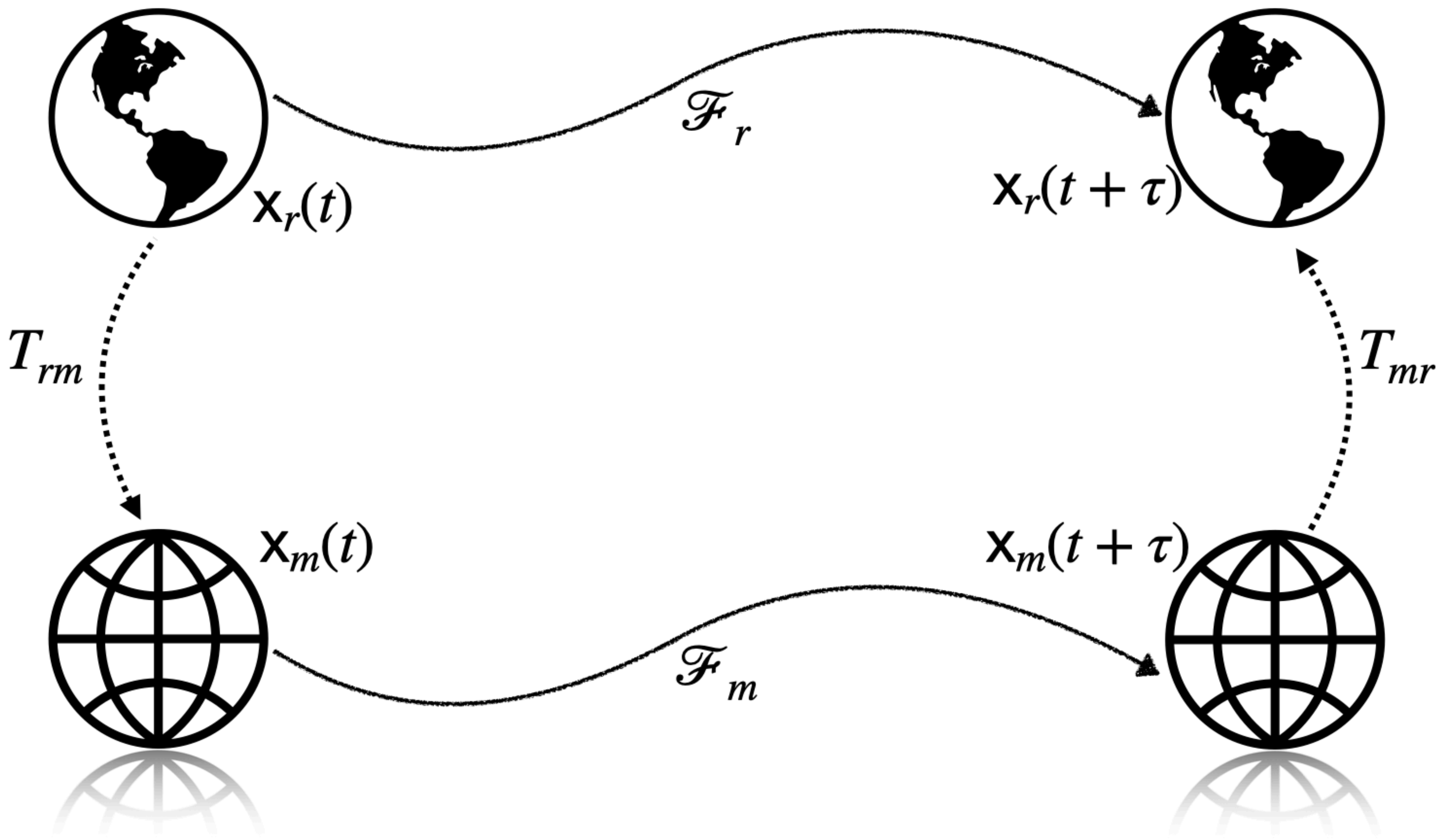
A data-driven approach to optimizing pre- and post-processing operations

Application to Lorenz ’96

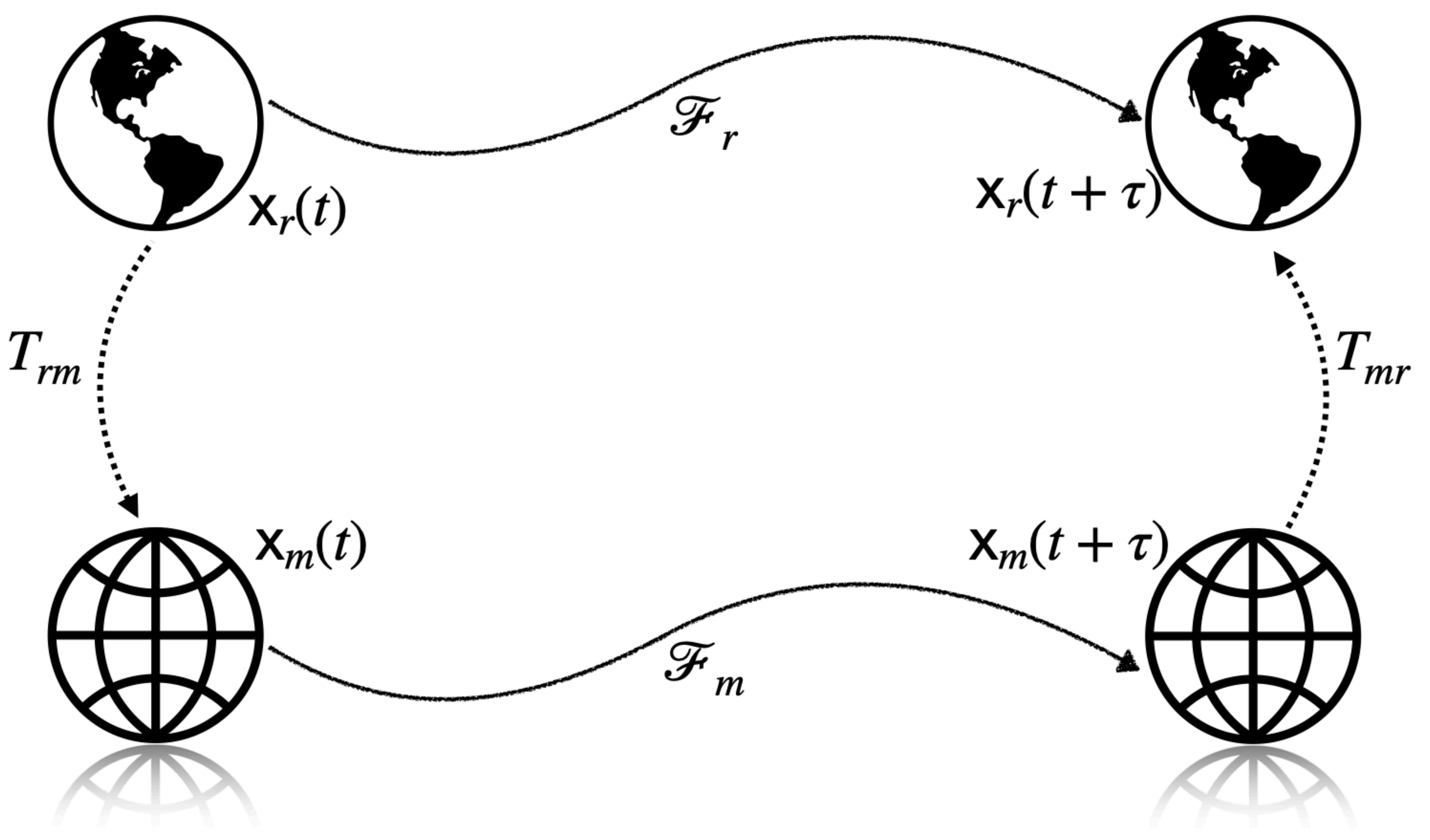
Application to CESM SMYLE

A data-driven approach to optimizing pre- and post-processing operations

$$\left\| \mathbf{x}_r(t+\tau) - [\mathbf{T}_{mr} \circ \mathscr{F}_m \circ \mathbf{T}_{rm}](\mathbf{x}_r(t)) \right\|^2$$



A data-driven approach to optimizing pre- and post-processing operations



$$\left\| \mathbf{x}_r(t+\tau) - [\mathbf{T}_{mr} \circ \mathscr{F}_m \circ \mathbf{T}_{rm}](\mathbf{x}_r(t)) \right\|^2$$

$$J\underline{(\mathbf{T}_{mr}, \mathbf{T}_{rm})} = \frac{1}{N_r} \sum_{i=1}^{N_r} \left\| \mathbf{x}^+_{r,i} - [\mathbf{T}_{mr} \circ \mathscr{F}_m \circ \mathbf{T}_{rm}](\mathbf{x}^-_{r,i}) \right\|^2$$

CATs = “cross-attractor transforms”

A general framework that includes models with reduced state spaces, component submodels, etc.

Outline

Forecasting with flawed models: “Fidelity” and “mapping” paradigms

A data-driven approach to optimizing pre- and post-processing operations

Application to Lorenz ’96: *Can CATS improve skill when forecasting a chaotic, higher-dimensional system with a deterministic, lower-dimensional one?*

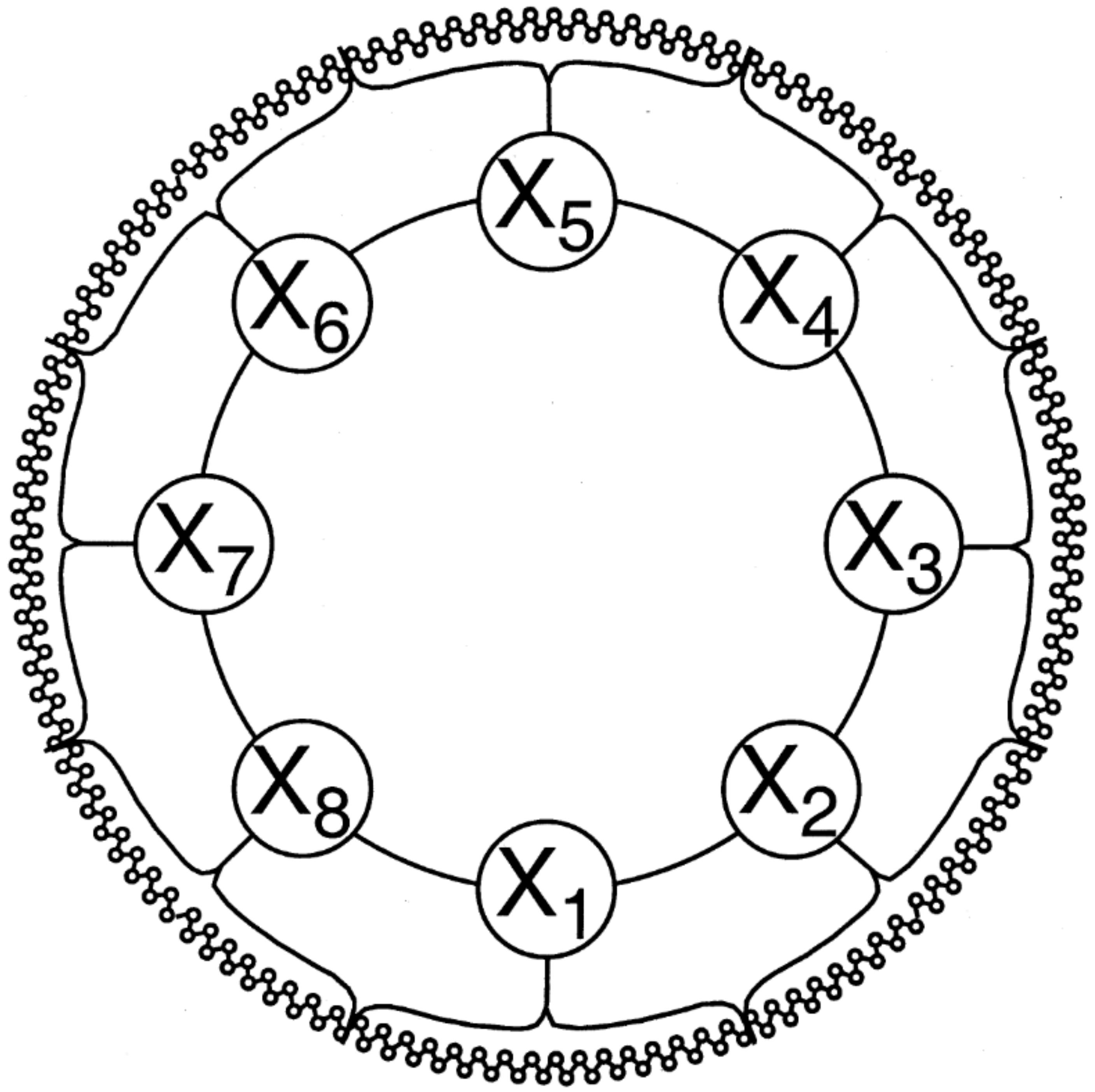
Application to CESM SMYLE

Application to Lorenz ’96

Governing equations

$$\frac{dx_k}{dt} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - hc\bar{y}_k$$

$$\frac{1}{c}\frac{dy_{j,k}}{dt} = -by_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - y_{j,k} + \frac{h}{J}x_k$$



Lorenz 1996; Wilks 2005

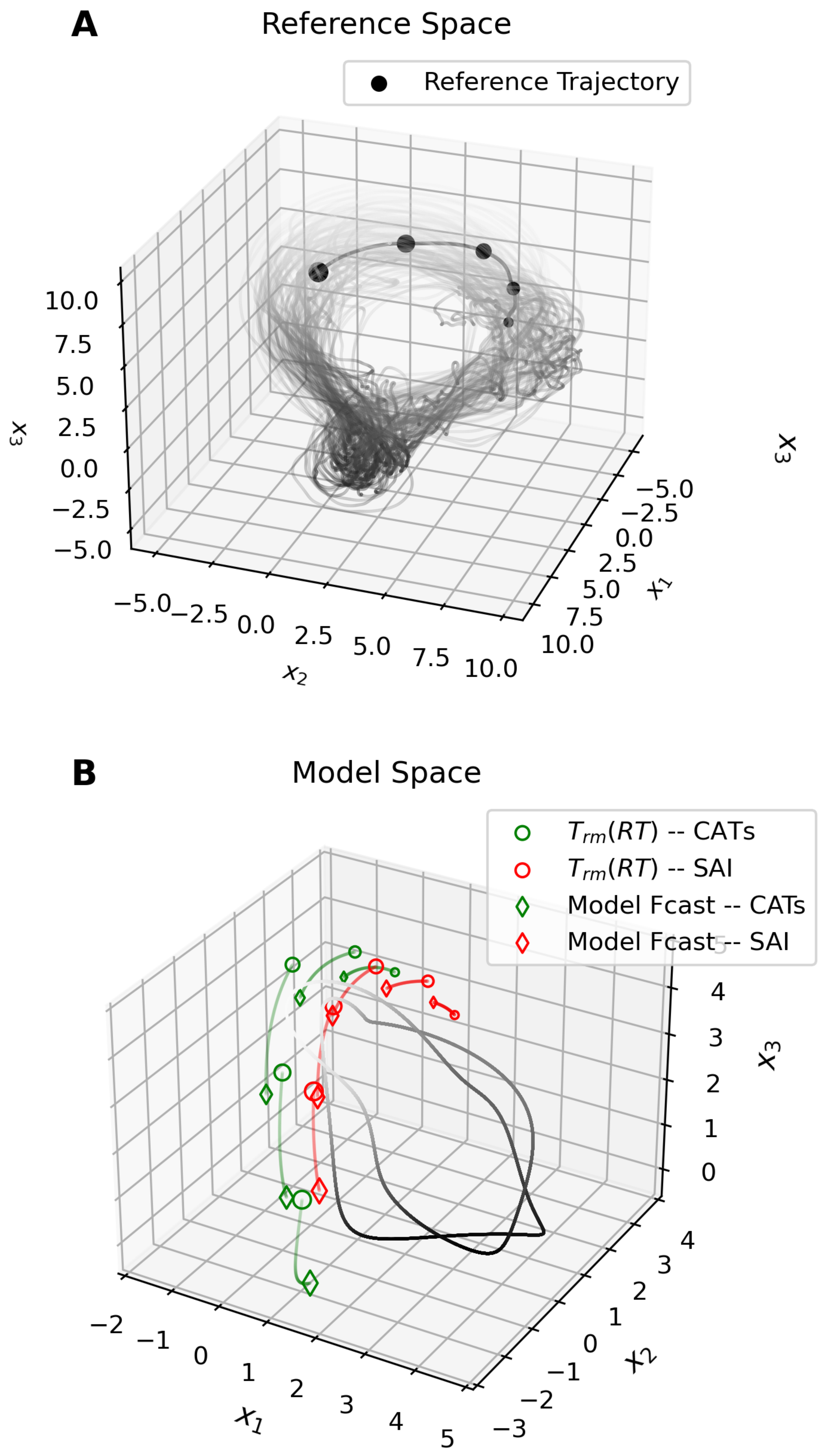
Application to Lorenz '96

Governing equations

$$\frac{dx_k}{dt} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - hc\bar{y}_k$$

$$\frac{1}{c}\frac{dy_{j,k}}{dt} = -by_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - y_{j,k} + \frac{h}{J}x_k$$

Two choices of F generate qualitatively different attractors!

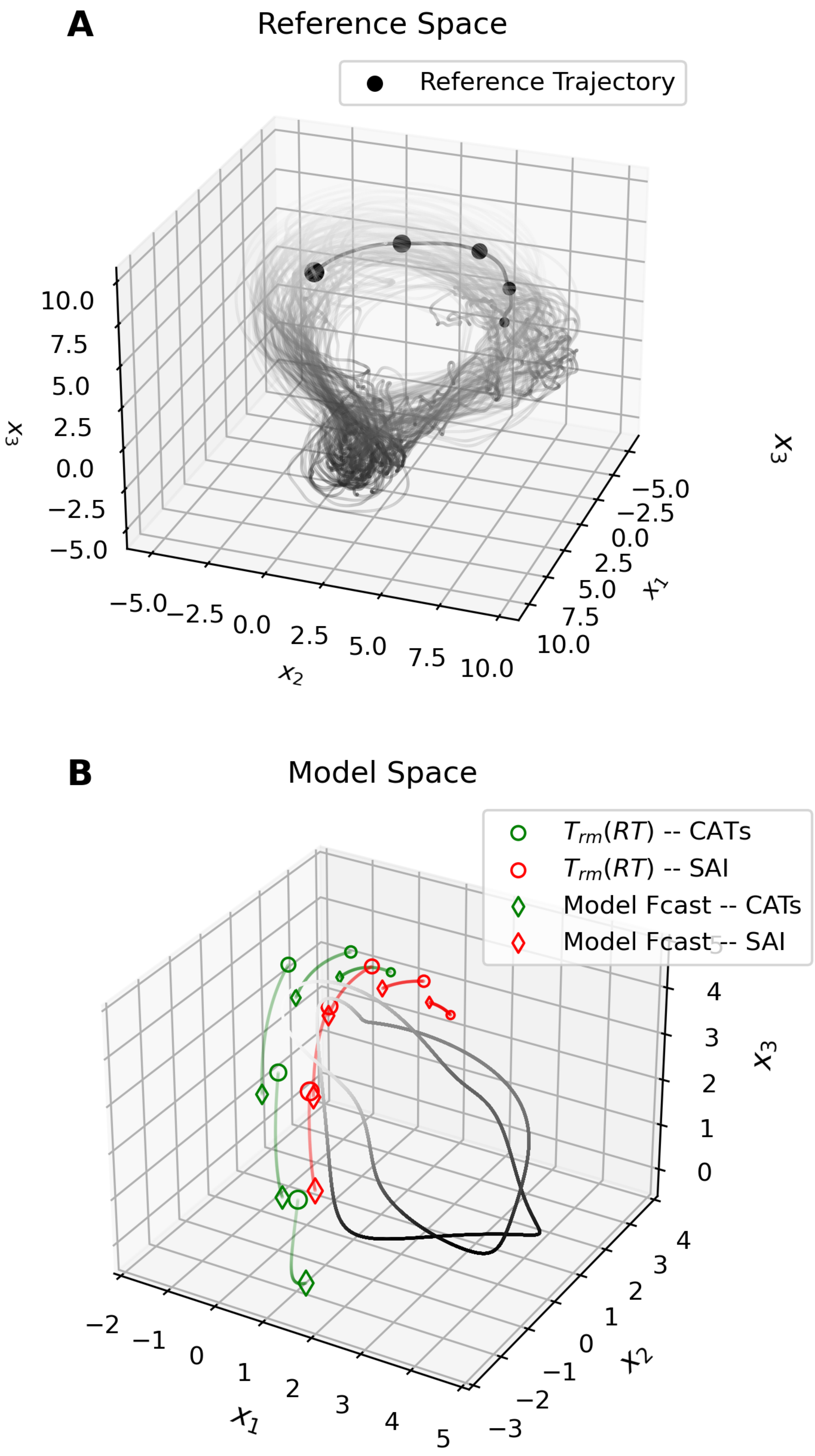


Application to Lorenz ’96

$$J(\mathbf{T}_{mr}, \mathbf{T}_{rm}) = \frac{1}{N_r} \sum_{i=1}^{N_r} \left\| \mathbf{x}^{+}_{r,i} - [\mathbf{T}_{mr} \circ \mathscr{F}_m \circ \mathbf{T}_{rm}](\mathbf{x}^{-}_{r,i}) \right\|^2$$

To compute a Jacobian, we need the gradient of $\mathscr{F}_m$.

We construct an emulator using analogs that is readily differentiable.



Application to Lorenz '96

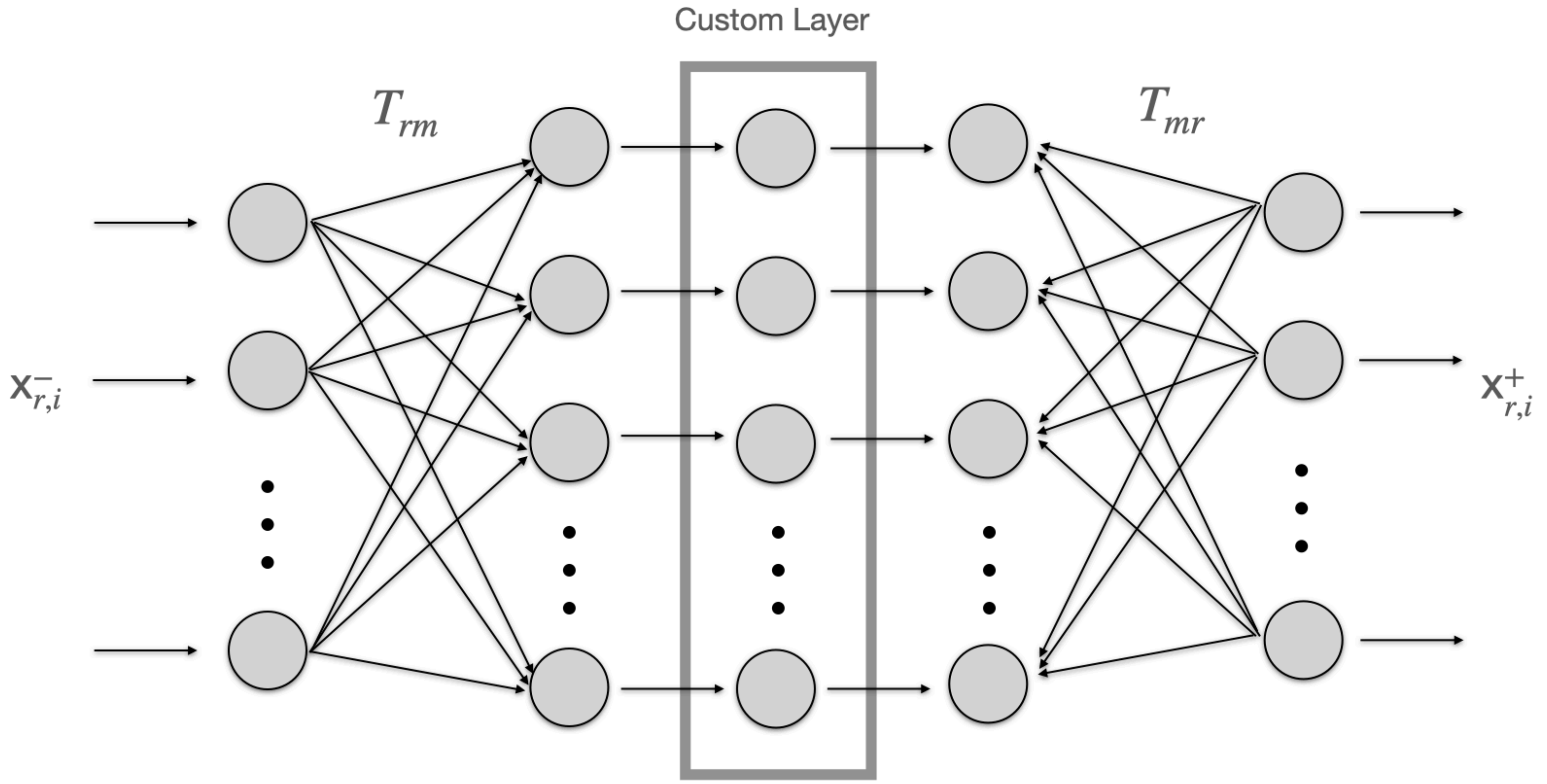
$$J(\mathbf{T}_{mr}, \mathbf{T}_{rm}) = \frac{1}{N_r} \sum_{i=1}^{N_r} \left\| \mathbf{x}^{+}_{r,i} - [\mathbf{T}_{mr} \circ \mathscr{F}_m \circ \mathbf{T}_{rm}](\mathbf{x}^{-}_{r,i}) \right\|^2$$

To compute a Jacobian, we need the gradient of $\mathscr{F}_m$.

We construct an emulator using analogs that is readily differentiable.

A fixed "custom" forecast layer is sandwiched between feed-forward neural networks.

We train on 8000 states from a long L96 simulation and validate on 2000.



(Agarwal, Amrhein, Grooms, in rev.)

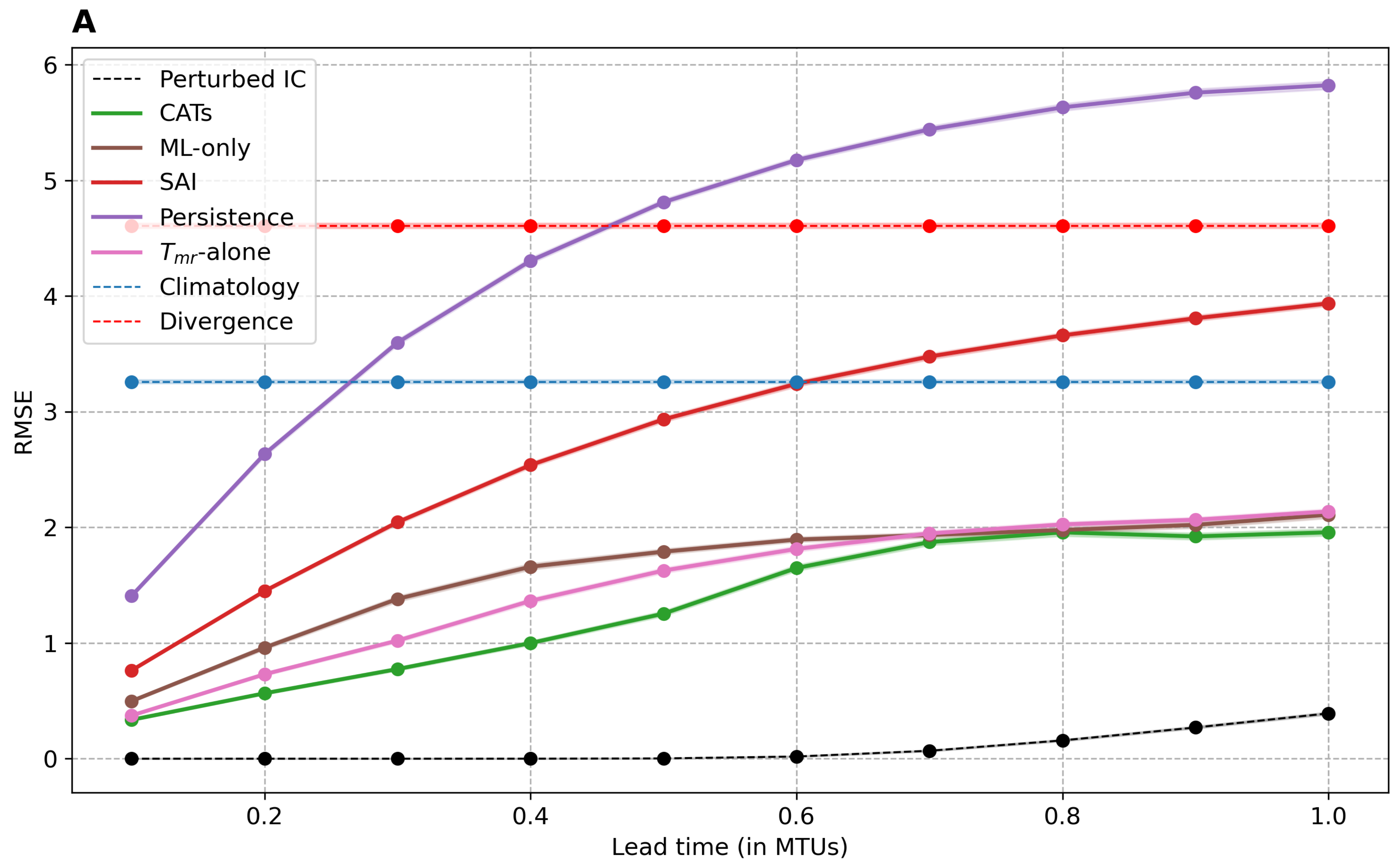
Results: forecast improvement in Lorenz ’96

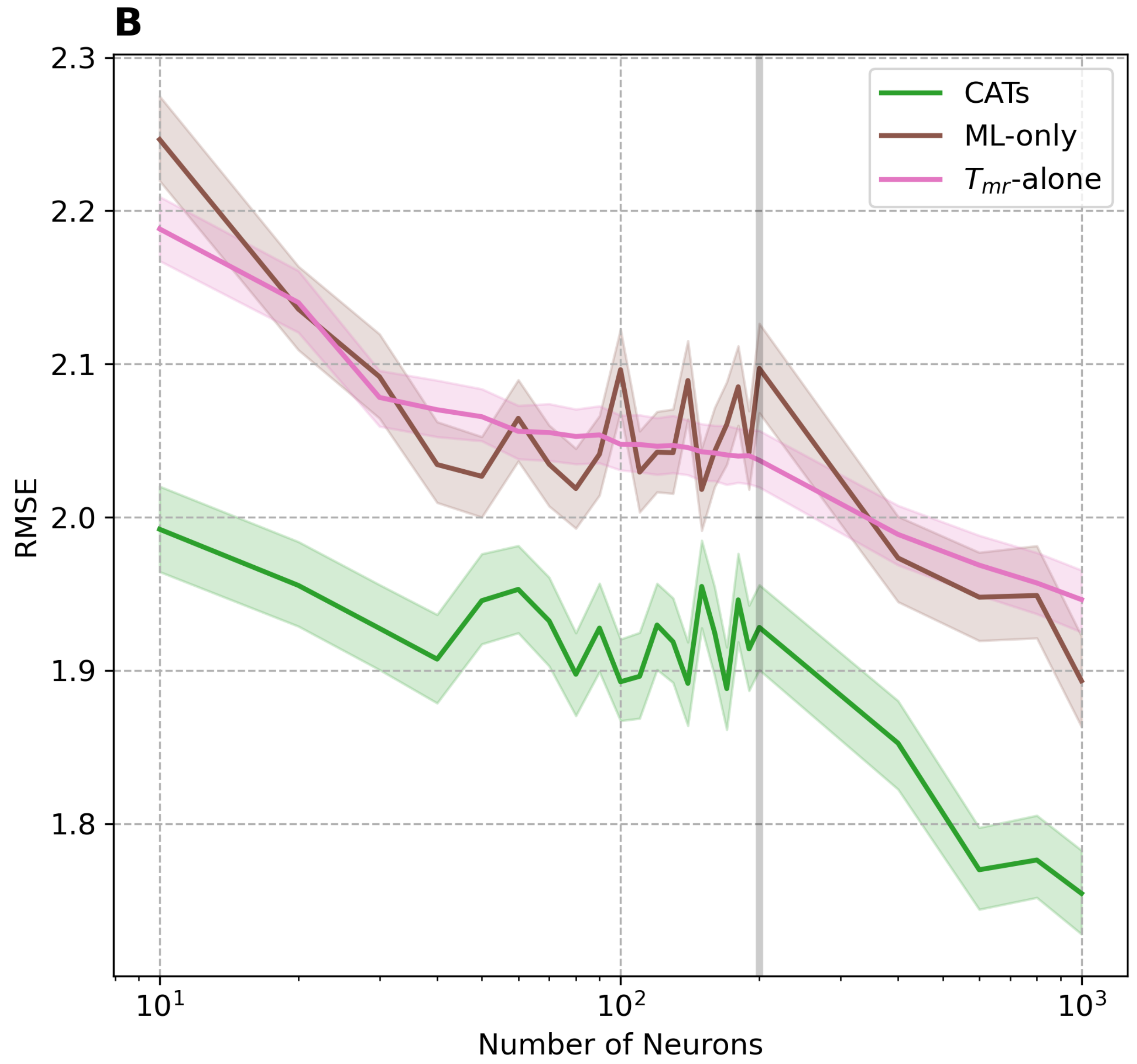
CATs outperforms other approaches (including pure ML emulation) across lead times up to 1 MTU.

A “T_{mr}-only” approach (which does not require an adjoint) also performs well.

Pure emulation typically requires several times more neurons to achieve the same performance.

We hypothesize that we’re in a hybrid Goldilocks regime where we have adequate data to train CATs, but not enough to emulate, an an imperfect model that is adequately skilled to regularize the emulation problem.

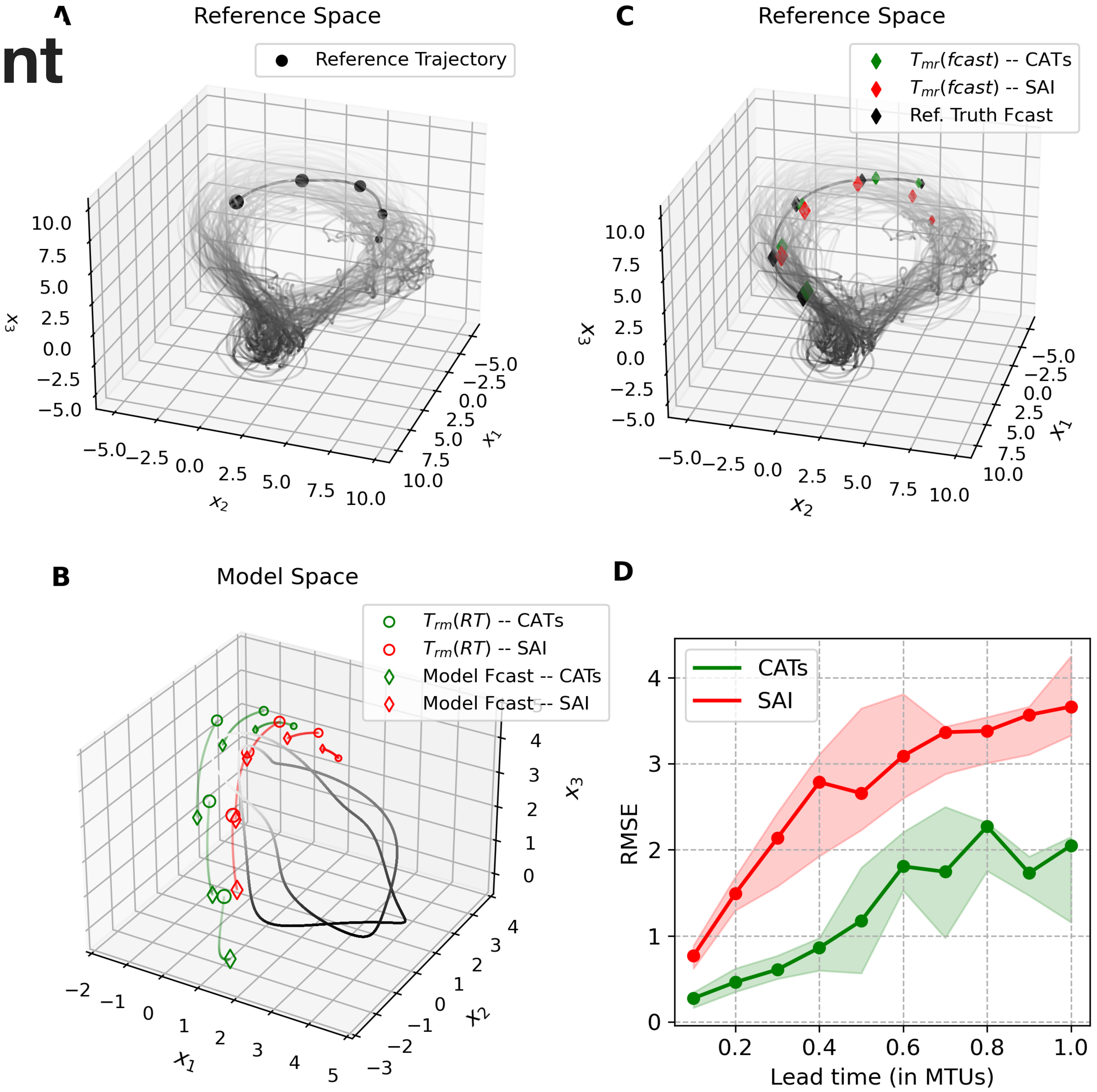




Results: forecast improvement in Lorenz ’96

Reference state trajectories are mapped to (approximate) trajectories in model phase space.

We have not required CATs to initialize on the model attractor. In some cases, off-attractor behavior (shock and drift) may be an avenue for improving skill.



Outline

Forecasting with flawed models: “Fidelity” and “mapping” paradigms

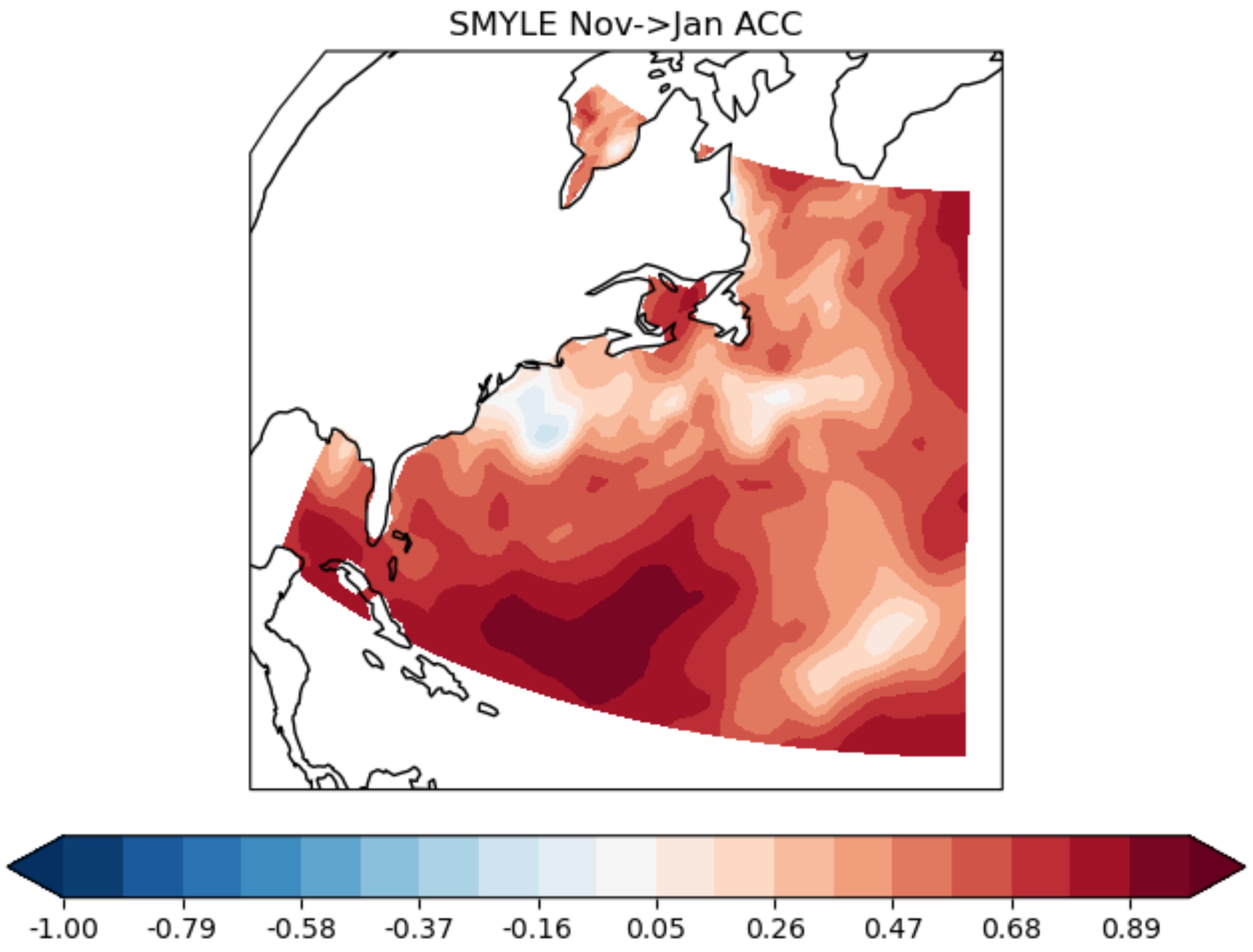
A data-driven approach to optimizing pre- and post-processing operations

Application to Lorenz ’96

Application to CESM SMYLE: *Can CATS improve skill in a high-dimensional, nonstationary Earth System model?*

Application to the CESM
Seasonal to Multi-Year Large Ensemble (SMYLE)

SMYLE (Yeager et al. 2022): 2-year-long hindcast simulations initialized quarterly with 20 ensemble members every year from 1970 to 2019



Can we improve on this?

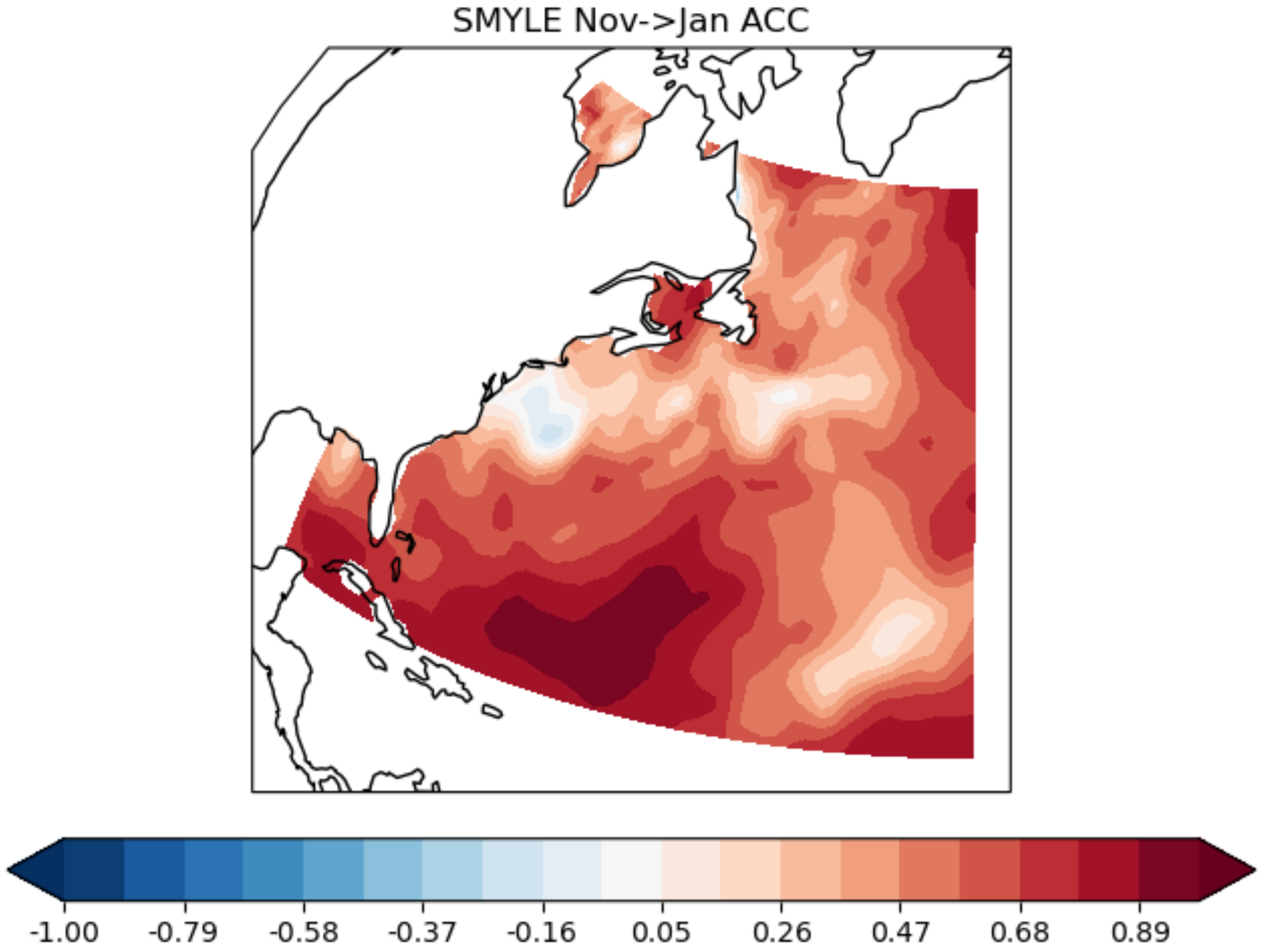
Application to the CESM Seasonal to Multi-Year Large Ensemble (SMYLE)

SMYLE (Yeager et al. 2022): 2-year-long hindcast simulations initialized quarterly with 20 ensemble members every year from 1970 to 2019

Simplify the loss function by assuming fixed T_{rm} and linear (after offset) $\mathbf{T}_{rm}$. We consider only the problem of 2-month predictions initialized in November.

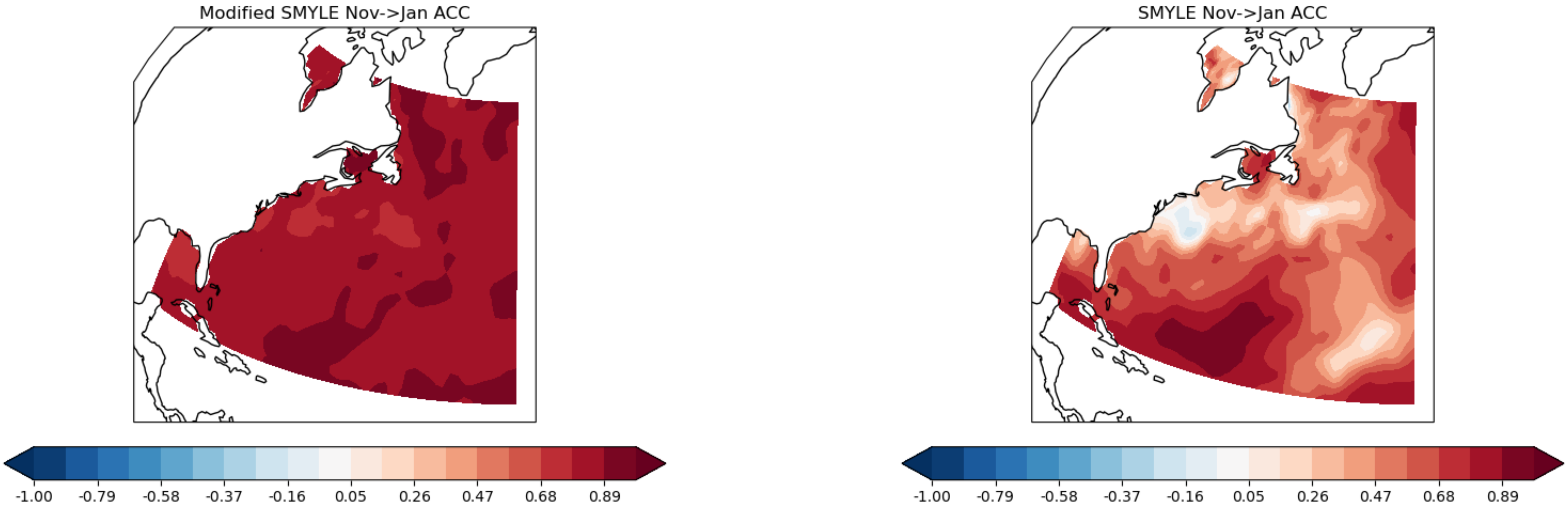
Then we can solve with e.g. ridge regression,
$$\mathbf{T}_{mr} = (\mathbf{X}_r\mathbf{X}_m^T + \lambda\mathbf{I})(\mathbf{X}_m\mathbf{X}_m^T + \lambda I)^{-1}$$

Train on contemporaneous HadSST as “truth”



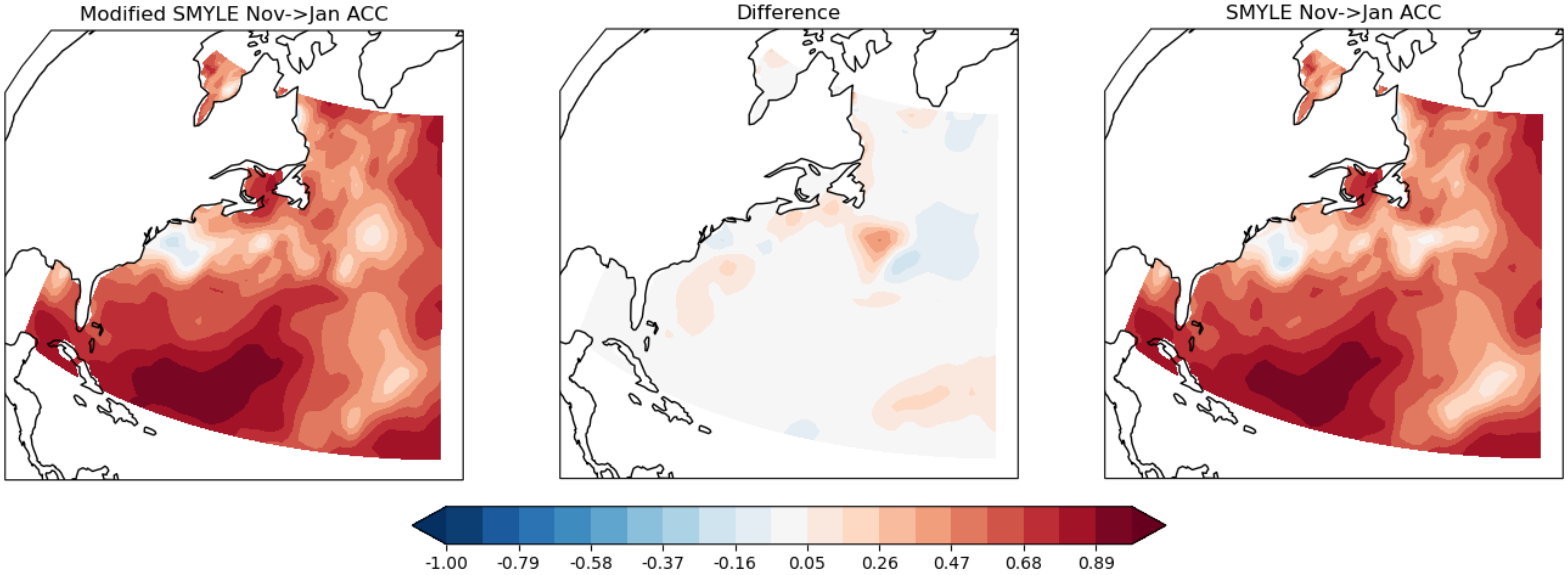
Can we improve on this?

Application to the CESM
Seasonal to Multi-Year Large Ensemble (SMYLE)



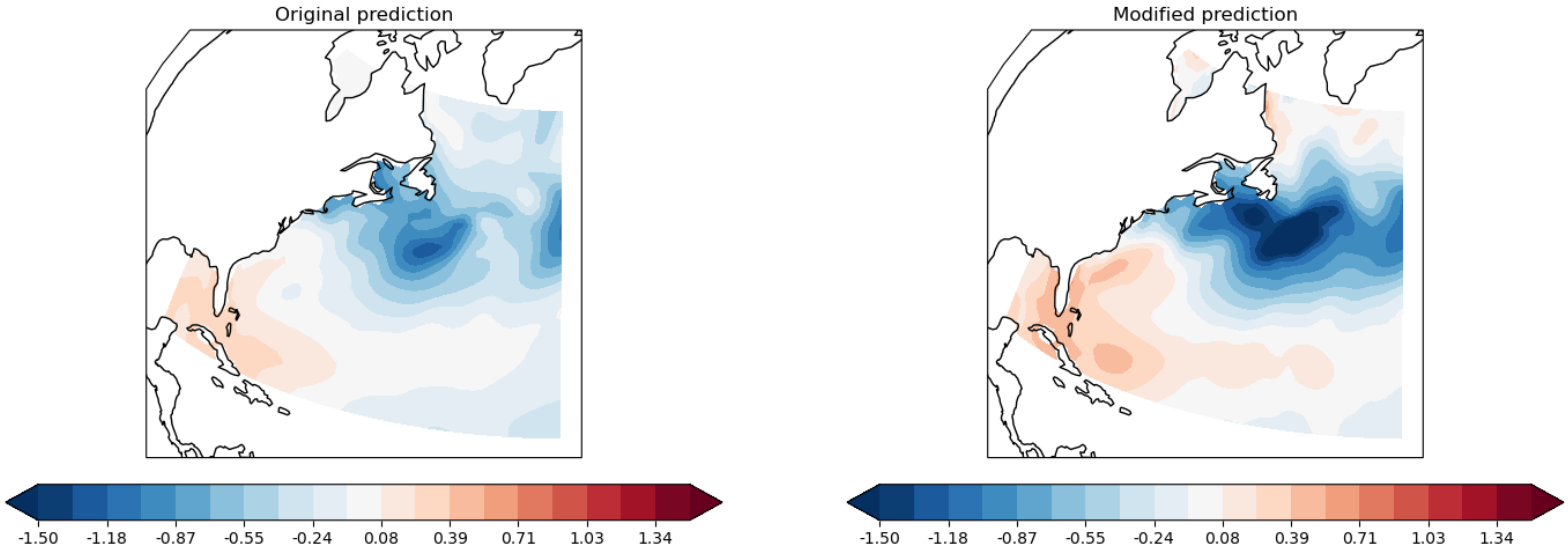
Using calibration = validation = [1970, 2019] shows we can fit the training data.

Application to the CESM
Seasonal to Multi-Year Large Ensemble (SMYLE)



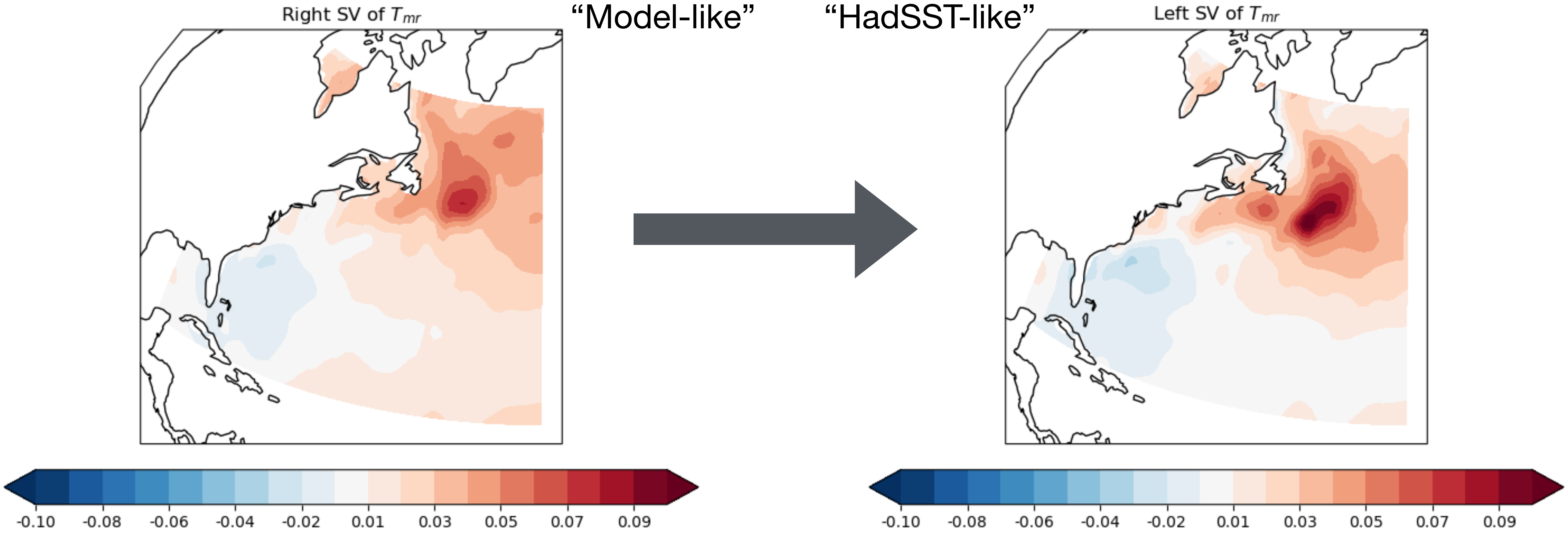
Using calibration = [1972, 2005], validation = [2006, 2019] shows a modest and possibly robust improvement in ACC near the model Northwest Corner.

Application to the CESM
Seasonal to Multi-Year Large Ensemble (SMYLE)



Applying $\mathbf{T}_{mr}$ to individual states reveals how they are mapped back to the reference state from HadSST.

Application to the CESM
Seasonal to Multi-Year Large Ensemble (SMYLE)



Left and right singular vectors of $\mathbf{T}_{mr} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$ reveals specific features in the model that are remapped.

Conclusions

CATs are optimal pre- and post-processing operations trained on previous forecasts. When used with process models, they are a form of hybrid modeling inspired by the “mapping paradigm,” “shadow filtering,” and other past approaches.

They leave model processes intact, and offer avenues (e.g., using singular vector analysis) for explaining how transforms work and possibly diagnosing bias.

We show improved skill in an L96 configuration and suggest inklings of improvement in the North Atlantic in SMYLE SSTs via modification in a dynamically tricky area.

Lots of possible applications! **damrhein@ucar.edu**