Diagnosing Biases in Soil Moisture–Evaporation Coupling Metrics in Earth System Models using Covariance Discriminant Analysis

Nazanin Tavakoli[1](ntavakol@gmu.edu), Paul A. Dirmeyer[1,2]

[1]George Mason University, [2]Center for Ocean-Land-Atmosphere Studies (COLA)

Grants: 80NSSC20K1803
80NSSC21K1801

How does land affect the atmosphere?

- Land-Atmosphere (L-A) interactions involve a series of feedback mechanisms within the Earth system.
- The land-surface states serve as initial elements in the feedback mechanisms, and soil moisture (SM) is among its most influential components.
- Soil moisture plays a critical role in shaping **climate and weather conditions** *(Hirschi et al., 2011; Seneviratne et al., 2006)* by **partitioning** the available energy at the land surface between latent (LE) and sensible (H) heat fluxes, which in turn affect near-surface atmospheric conditions, boundary-layer properties, and ultimately cloud formation and precipitation.
- These L-A feedbacks can be quantified with **coupling metrics** that combine statistical techniques with an understanding of physical processes.
- Many studies have been made to quantify L-A coupling metrics using **Earth system models**, **reanalysis**, **observations**.

Adapted from Santanello et al. (2018)

Energy cycle
Water cycle
Bridge the energy and water cycles

L-A coupling metrics

Large discrepancy in L–A coupling strength and type across Earth system models.[1-8]

Models and observations remain inconsistent in capturing L-A interactions. [9-11]

Reanalysis products and observational data also disagree in capturing L-A interactions. [12-13]

Tavakoli & Dirmeyer *(2026; under review at Scientific Data)* have developed a **global dataset of L-A coupling metrics** based on observations, while accounting for the **random observational errors** in **SM satellite** measurements using the **Markovian framework**.

Selected refs:

[1] Koster & Suarez (1999); [2] Koster et al. (2004); [3] Koster et al. (2006); [4] Hohenegger et al. (2009);
[5] Seneviratne & Koster (2012); [6] Flato et al. (2013); [7] Sippel et al. (2017); [8] Moghisi et al. (2024);
[9] Dirmeyer (2013); [10] McColl et al. (2019); [11] Phillips et al. (2017);
[12] Hsu & Dirmeyer (2022); [13] Dirmeyer et al. (2021).

Research Gap, Questions & Significance

Observational-Adjusted Metrics (SM, HF: LE, H, EF):

Adjusted Pearson's Correlation

$$R_{Adjusted} = \frac{\mathrm{cov}[SM_O, HF]}{\sqrt{(v\mathrm{ar}[SM_O] - \delta^2) \cdot \mathrm{var}[HF]}}$$

δ^2: Random Error Variance

Adjusted Terrestrial Coupling Index *(Dirmeyer et al.,2011)*

$$I_{Adjusted}(SM, HF) = R_{Adjusted} * \sigma_{HF}$$

Model-Derived Metrics

Pearson's Correlation

$$R_{model}(SM, HF) \frac{\mathrm{cov}[SM, HF]}{\sqrt{\mathrm{var}[\mathrm{SM}] \cdot \mathrm{var}[HF]}}$$

Terrestrial Coupling Index

$$I_{model}(SM, HF) = R_{model} * \sigma_{HF}$$

Objective: This project focuses on applying **Covariance Discriminant Analysis (CDA)** to covariance matrices (both model and observation) to identify the linear combinations of SM and HF where the **greatest separation occurs between the two distributions**.

Since covariance matrices encode var(SM), var(HF), and cov(SM, HF), and since R and TCI are derived from these quantities, this study highlights the **directions in which the model variance structure differs most from observations**. This tells us about **the structure of disagreement:**

- Where is the model's SM–HF structure significantly different from observations? Is there a systematic seasonal pattern in this structure?
- Is the disagreement dominated by SM variance, HF variance, or their covariance? And how much are their contributions?

Why It Matters: Improve understanding of ecosystem dynamics, climate variability, and clarify implications for global earth system model simulations. Guide model development by targeting improvements in their parameterizations.

Data description

	Products	***Spatial Resolution***	***Periods***
Observations	***Soil moisture***		
	CCI (v09.1)	*0.25˚ × 0.25˚*	*1979-2023*
	Surface heat flux (LE,H,EF)		
	GLEAM (v4.1a)	*0.1˚ × 0.1˚*	*2000-2019*
Model	*Community Land Model (CLM5)*	*~0.94° lat × 1.25° lon*	*2001-2026*

2001-2023 (Daily data)
Soil moisture (SM), Evaporation (E)

Four seasons: December–January–February (DJF)
March–April–May (MAM)
June–July–August (JJA)
September–October–November (SON)

Spatial resolution: 0.94°×1.25° lat–lon

How to compute CDA eigenvalues? (Step 1 & 2)

Step 1: Extract 2×2 covariance matrices for observations and model using paired time series of (SM, E):

Let's consider soil moisture (SM) & evaporation (E):

$$x_{obs}(t) = [SM_{obs}(t), E_{obs}(t)]^T, x_{mod}(t) = [SM_{mod}(t), E_{mod}(t)]^T$$

Then the covariance matrices are:

$$\Sigma_{obs} = \begin{bmatrix} var(SM_{obs}) & cov(SM_{obs}, E_{obs}) \\ cov(SM_{obs}, E_{obs}) & var(E_{obs}) \end{bmatrix}$$

$$\Sigma_{mod} = \begin{bmatrix} var(SM_{mod}) & cov(SM_{mod}, E_{mod}) \\ cov(SM_{mod}, E_{mod}) & var(E_{mod}) \end{bmatrix}$$

Step 2: Compute difference matrix:

The difference in covariance structure between observations and model is:

$$\Delta\Sigma = \Sigma_{obs} - \Sigma_{mod} = \begin{bmatrix} d_{var,SM} & d_{cov} \\ d_{cov} & d_{var,E} \end{bmatrix}$$

$$d_{var,SM} = var(SM_{obs}) - var(SM_{mod})$$
$$d_{var,E} = var(E_{obs}) - var(E_{mod})$$
$$d_{cov} = cov(SM_{obs}, E_{obs}) - cov(SM_{mod}, E_{mod})$$

How to compute CDA eigenvalues? (Step 3)

Step 3: Solve the generalized eigenvalue problem:

CDA is approximated $\boldsymbol{\lambda}$ by solving a generalized eigenvalue problem of:

$$\Delta\boldsymbol{\Sigma}.\boldsymbol{v} = \boldsymbol{\lambda}.\Sigma_{obs}.\boldsymbol{v}$$

$\boldsymbol{\lambda}$ = the corresponding **eigenvalue** for that eigenvector, which tells us how strong the disagreement is between Σ_{obs} and Σ_{mod} along their respective eigenvectors.

$\boldsymbol{v}$ = eigenvector (a 2-component vector of weights for SM and E)

Consider (SM,E) as a 2D space, $\boldsymbol{v}$ is a direction in the SM–E space. It defines a linear combination for the mode-1:

$$z_{obs}^{(1)}(t) = a * SM_{obs}(t) + b * E_{obs}(t) \qquad v_1 = \begin{pmatrix} a \\ b \end{pmatrix}$$

$$z_{mod}^{(1)}(t) = a * SM_{mod}(t) + b * E_{mod}(t)$$

$$\lambda_i = 1 - \frac{Var\left(z_{mod}^{(i)}\right)}{Var\left(z_{obs}^{(i)}\right)}$$

How to compute CDA eigenvalues? (Step 4)

Step 4: Contributions to disagreement & fractions of disagreement (absolute contributions):

How much of the total CDA variance disagreement along eigenmode 1 is due to the model–obs difference in the variance of SM, variance of E, and their covariance?

For a given CDA eigenvector $v = \begin{pmatrix} a \\ b \end{pmatrix}$, the signed disagreement along that mode is:

$$D = v^T \Delta\Sigma v = a^2 d_{var,SM} + b^2 d_{var,E} + 2ab\, d_{cov} = C_{SM} + C_E + C_{cov}$$

$$C_{SM} = a^2 d_{var,SM},\ C_E = b^2 d_{var,E},\ C_{cov} = 2ab\, d_{cov}$$

C_{SM}: Contribution of the difference in SM variance to the total variance disagreement along that mode.
C_E: Contribution from difference in E variance.
C_{cov}: contribution from difference in SM–E covariance.

If we take the absolute values and normalize, then the fractions of disagreement ($f_{SM} + f_E + f_{cov} = 1$) are:

$$f_{SM} = \frac{|C_{SM}|}{|C_{SM}| + |C_E| + |C_{cov}|}$$

$$f_E = \frac{|C_E|}{|C_{SM}| + |C_E| + |C_{cov}|}$$

$$f_{cov} = \frac{|C_{cov}|}{|C_{SM}| + |C_E| + |C_{cov}|}$$

f_{SM}: Fraction of the magnitude of total CDA disagreement along eigenmode 1 that comes from the SM variance difference between model and observation.

Permutation test (statistical significance)

The permutation test assesses whether the observed eigenvalues could arise by chance if model and observational samples actually came from the same underlying joint distribution of (SM, E).

Null hypothesis is that both observation and model come from the same population, so we can mix them.
H_0: Σ_{obs} and Σ_{mod} are not distinguishable; model and observation are exchangeable samples.

Step 1: Form paired arrays of (SM, E) from observations and model: $obs_{pairs} = [SM_{obs}, E_{obs}]$, $mod_{pairs} = [SM_{mod}, E_{mod}]$

Step 2: Randomly permute the time indices of obs_{pairs} and $model_{pairs}$ separately, and pool all shuffled pairs together.

Step 3: Randomly split pooled pairs into two groups of equal size: Group 1 → synthetic obs, Group 2 → synthetic model

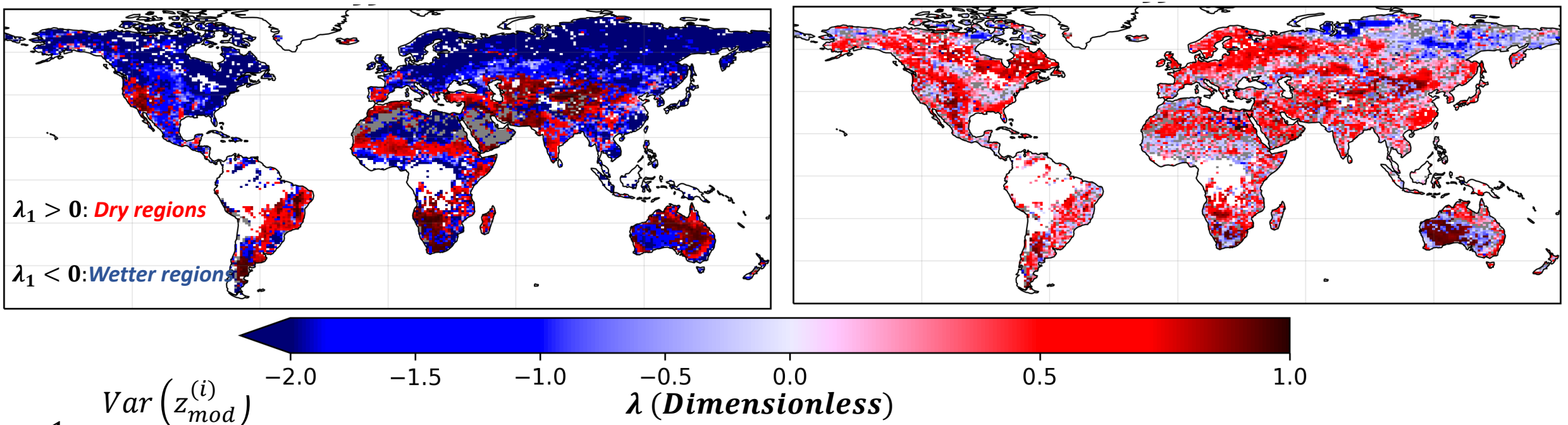
Step 4: Compute 2×2 covariance matrices for both groups: Σ_1, Σ_2 ,
Compute $\Delta\Sigma^* = \Sigma_1 - \Sigma_2$, and solve the same generalized eigenproblem: $\Delta\Sigma^* . \boldsymbol{v}^* = \boldsymbol{\lambda}^* . \Sigma_1 . \boldsymbol{v}^*$
Extract the first two eigenvalues for this permutation.

Step 5: P-Value Calculation
After 1000 iterations, I obtain empirical null distributions for λ_1 and λ_2 using a two-sided test in terms of magnitude $|\lambda|$:
$p - value = \frac{1}{B}\sum_{b=1}^{B}(|\lambda_b^*| \geq |\lambda_{obs}|)$, λ_b^* is the eigenvalues from the B permutations (here B=1000)

Results: *Where and how strongly does CLM deviate from obs in the first & second SM–E modes during JJA?*

JJA

λ_1 λ_2

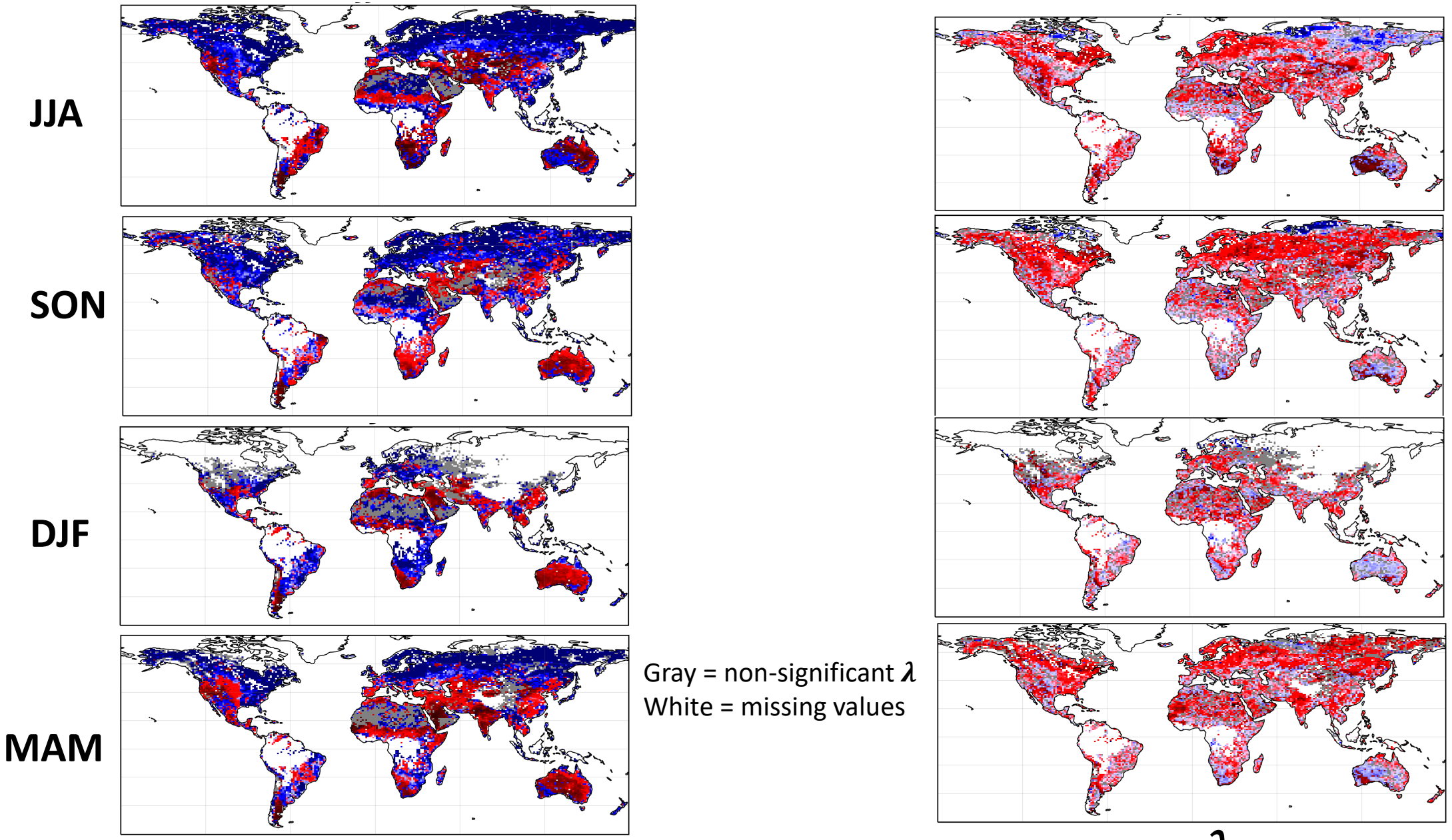


$$\lambda_i = 1 - \frac{Var\left(z_{mod}^{(i)}\right)}{Var\left(z_{obs}^{(i)}\right)}$$

Gray = non-significant λ
White = missing values

- $|\lambda|$: how big the relative variance disagreement is.
- Sign of λ tells us whether the model variance is too small or too large compared to obs along that mode.
- $\lambda \approx 0$: Model variance along that SM–E mode is close to observed variance.
- $\lambda > 0 \rightarrow (Var\left(z_{mod}^{(i)}\right) < Var\left(z_{obs}^{(i)}\right))$: Model variance along that SM–E mode is smaller than observed variance (underestimates variance)
- $\lambda < 0 \rightarrow (Var\left(z_{mod}^{(i)}\right) > Var\left(z_{obs}^{(i)}\right))$: Model variance along that SM–E mode is larger than observed variance (overestimates variance)

Results: *Where and how strongly does CLM deviate from obs in the first & second SM–E modes during seasons?*



Results: *Variance explained by each eigenmode*

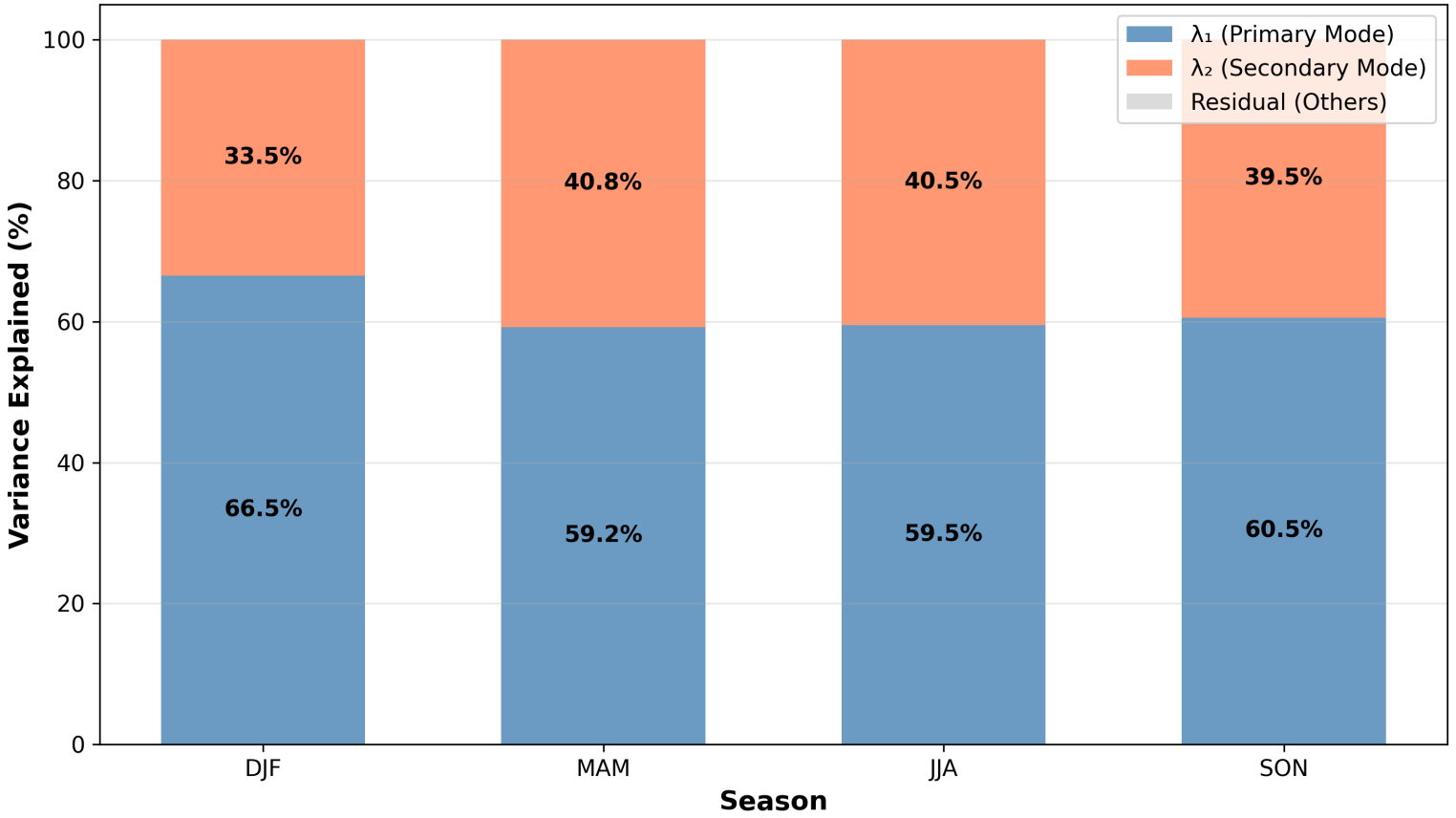
$$Variance\ explained\ by\ \lambda_i = \frac{\sum_{j=1}^{N} \lambda_i^2(j)}{\sum_{k=1}^{2}\sum_{j=1}^{N} \lambda_i^2(j)} * 100\%$$

$j = grid\ cell\ index\ (1\ to\ N)$

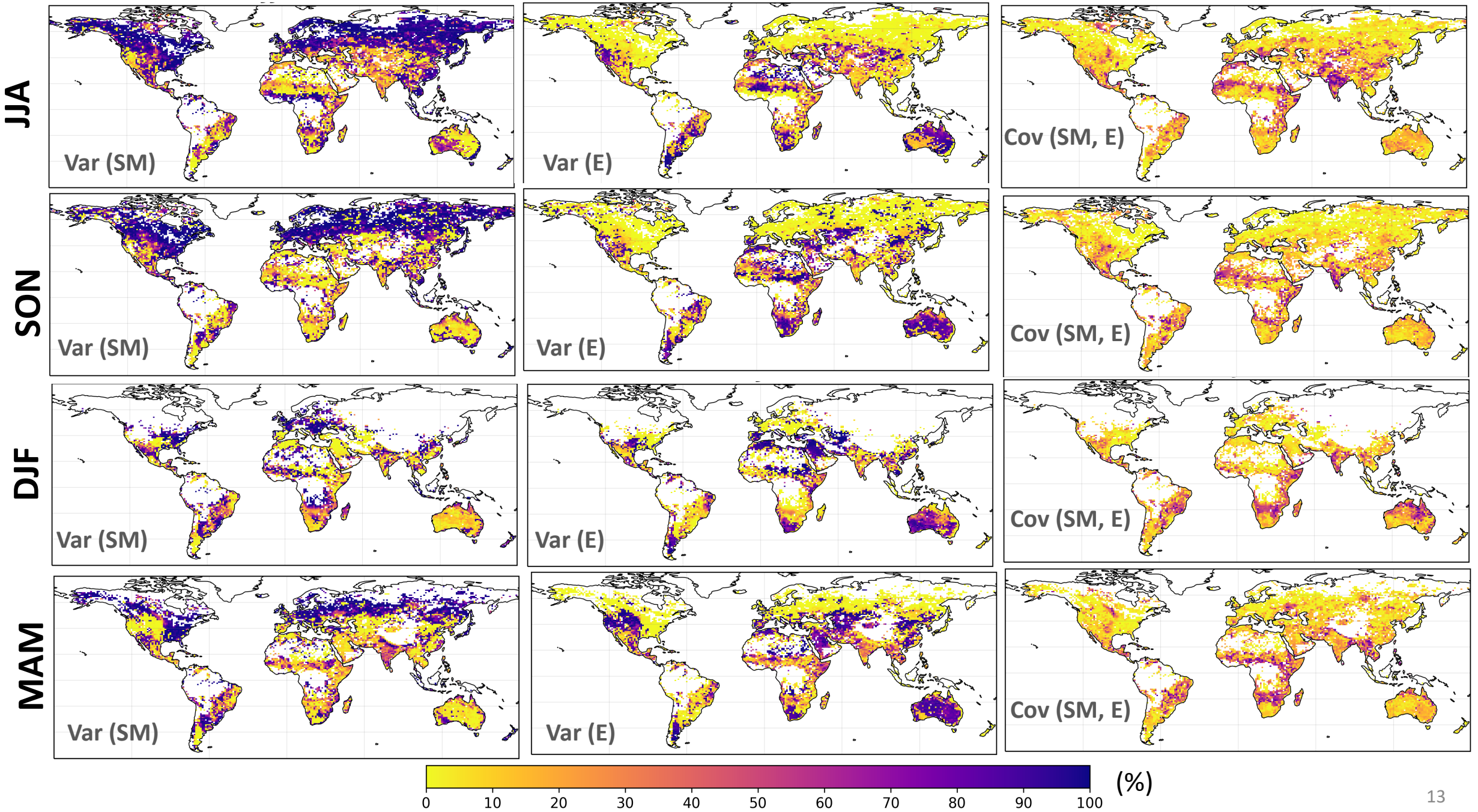
$N = number\ of\ significant\ grid\ cells\ (where\ p < \alpha\ \&\ |\lambda| \leq \lambda_{max})$

$$\sum_{j=1}^{N} \lambda_i^2(j) = sum\ of\ squared\ eigenvalues\ for\ mode\ i\ across\ all\ significant\ locations$$

$$\sum_{k=1}^{2}\sum_{j=1}^{N} \lambda_i^2(j) = total\ varinace\ from\ both\ modes$$



Results: *Fraction of mode-1 disagreement attributable to SM variance, E variance, and SM-E covariance differences*



Takeaways & Future Work

We applied Covariance Discriminant Analysis on covariance matrices of soil moisture and evaporation to identify directions and magnitude where CLM model structure differs most from observations (CCI-GLEAM).

- The CLM model underestimates variance along the first SM–E mode compared to the observed variance, mostly over semi-arid regions during JJA.
- The CLM model overestimates variance along the first SM–E mode mode compared to the observed variance, mostly over humid and wetter regions during JJA.
- There is a systematic seasonal pattern in the global over- and underestimation of the first and second modes.
- Is the disagreement dominated by SM variance, E variance, or their covariance?
  - Soil moisture has a higher contribution in high latitudes.
  - Evaporation contribution is highest in drier regions.
  - Covariance contribution is mostly highest in transitional zones and monsoon regions.

Future steps: Apply this approach to

Sensible heat flux (H) and evaporative fraction (EF) of CLM

CESM2 model, MPAS-NoahMP

Thank you!

Contact Information: ntavakol@gmu.edu