Higher-Order Internal Modes of Variability Imprinted in Year-to-Year California Streamflow Changes

A science-centric machine learning study

Shiheng Duan, Giuliana Pallotta, Céline Bonfils
Lawrence Livermore National Laboratory

June 24, 2024

Self Introduction

北京師範大學環境學院
School of Environment, Beijing Normal University

Bachelor of Environmental Engineering from Beijing Normal University. One semester in University of Oklahoma.

UCDAVIS
ATMOSPHERIC SCIENCE

Ph.D. in Atmospheric Science from University of California, Davis.

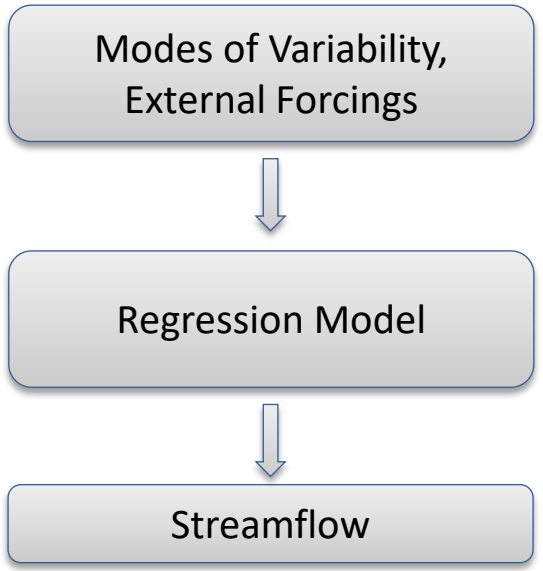
Lawrence Livermore National Laboratory

Postdoc in Climate Section of PLS. Working on Climate Resilience for National Security 22-SI-008.

Motivation and General Framework

- Streamflow is a key component in hydroclimate system.
- It is influenced by anthropogenic climate change and internal variability.
- What is the most important factor that influences streamflow in California?
- Previous studies:
  - Large-scale climate variability (teleconnections).
  - Composite-based analysis.
  - Linear-based correlation.
  - Patterns in local domain.
  - Lack connection between global and local-scale.
  - Independent testing set missing.

- Build a regression model
  - Use internal variability indices and anthropogenic forcings to predict streamflow in California.
  - The explained variance can be used to quantify the predictability.
  - The feature importance can be used to rank the variables.



Inputs and Outputs Datasets

- Available datasets/methods
  - **CMIP6 simulations: augment training samples.**
    - 6 models, each with 10 ensembles.
  - **Modes of variability indices:**
    - PCMDI Metrics Package (PMP): derive modes of variability indices from climate simulations and observations using Empirical Orthogonal Function (EOF) and Common Basis Function (CBF).
  - **Streamflow:**
    - USGS observations, independent testing set.
    - CMIP-forced simulations.
      - Inter-Sectoral Impact Model Intercomparison Project (ISIMIP): bias-corrected and downscaled temperature and precipitation to force regional hydrology models. **Quarter degree, do not tune for all basins.**
      - Long-Short Term Memory (LSTM) model: takes daily time series of basin-mean precipitation, temperature. Trained towards USGS observations. Forced by BCSD-CMIP6 for historical projections.

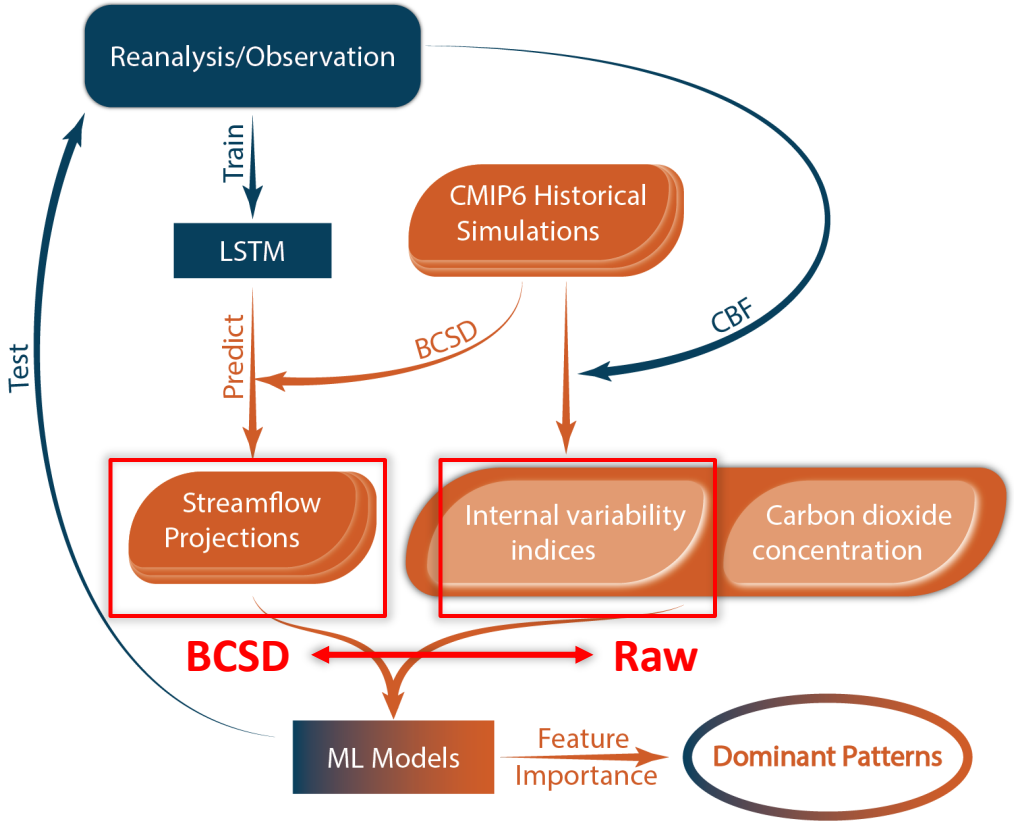


Figure. Schematic work flow. Blue for reanalysis and red for climate simulations.

Validation of EOF/CBF and BCSD

- Validate whether BCSD would change the variability
  - IPSL-CM6A-LR (GCM)
  - BCSD (with NOAA OISST as reference)
- CBF (right) shows BCSD would not change variability
- EOF (left) shows swapping of EOFs is needed for higher-order EOFs
  - PC-6 from BCSD matches well with PC-5 from GCM
  - Further discussion in Lee et al., 2021

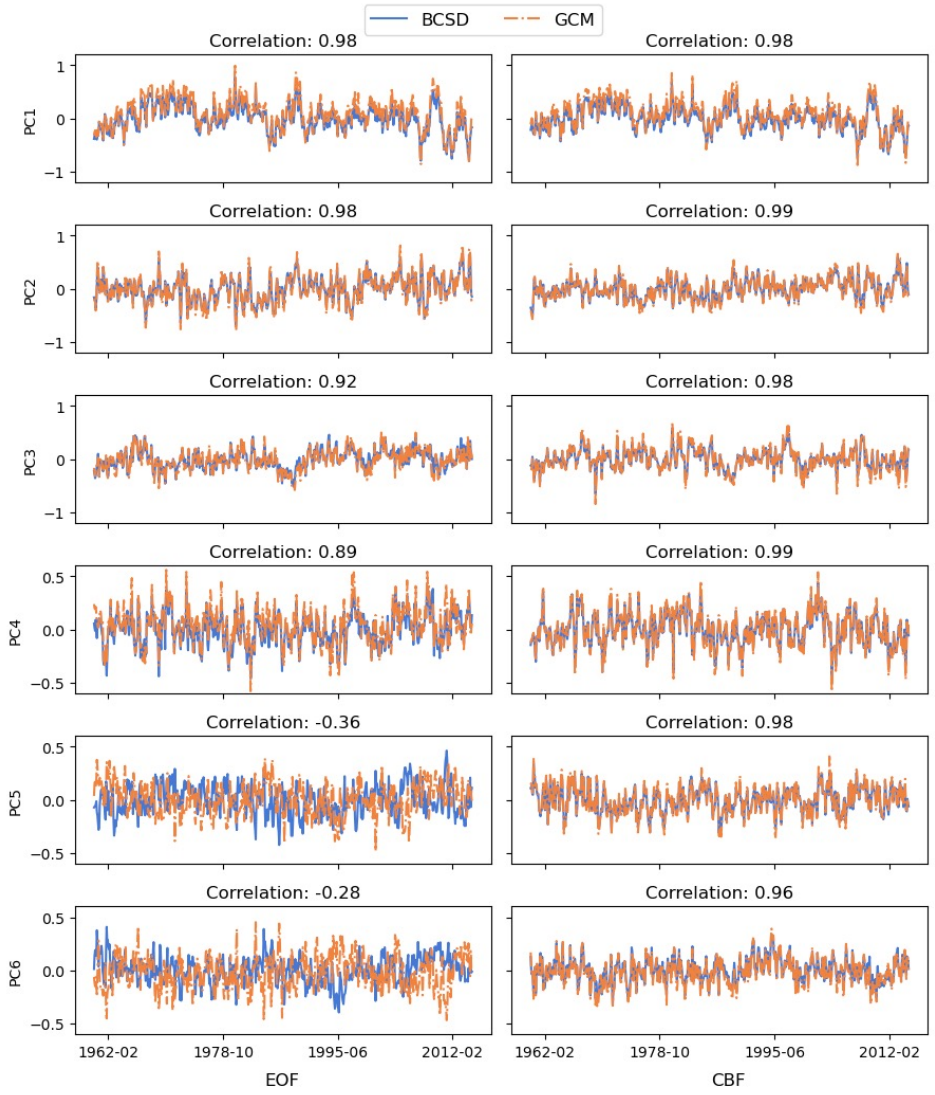


Figure. PC time series of the PDO domain from BCSD data (blue lines) and raw IPSL-CM6A-LR simulation.

ML Model Performance

- ML models:
  - Linear Regression (LR)
  - Lasso
  - Ridge
  - Linear Orthogonal Decomposition (LOD)
  - Partial Least Squares Regression (PLS)
  - Automated Machine Learning (AutoGluon)
- Output:
  - *Seasonal peak streamflow: peak month, month prior and after*
  - *Standardized anomalies*
- Input:
  - *Concurrent or previous season variability indices*

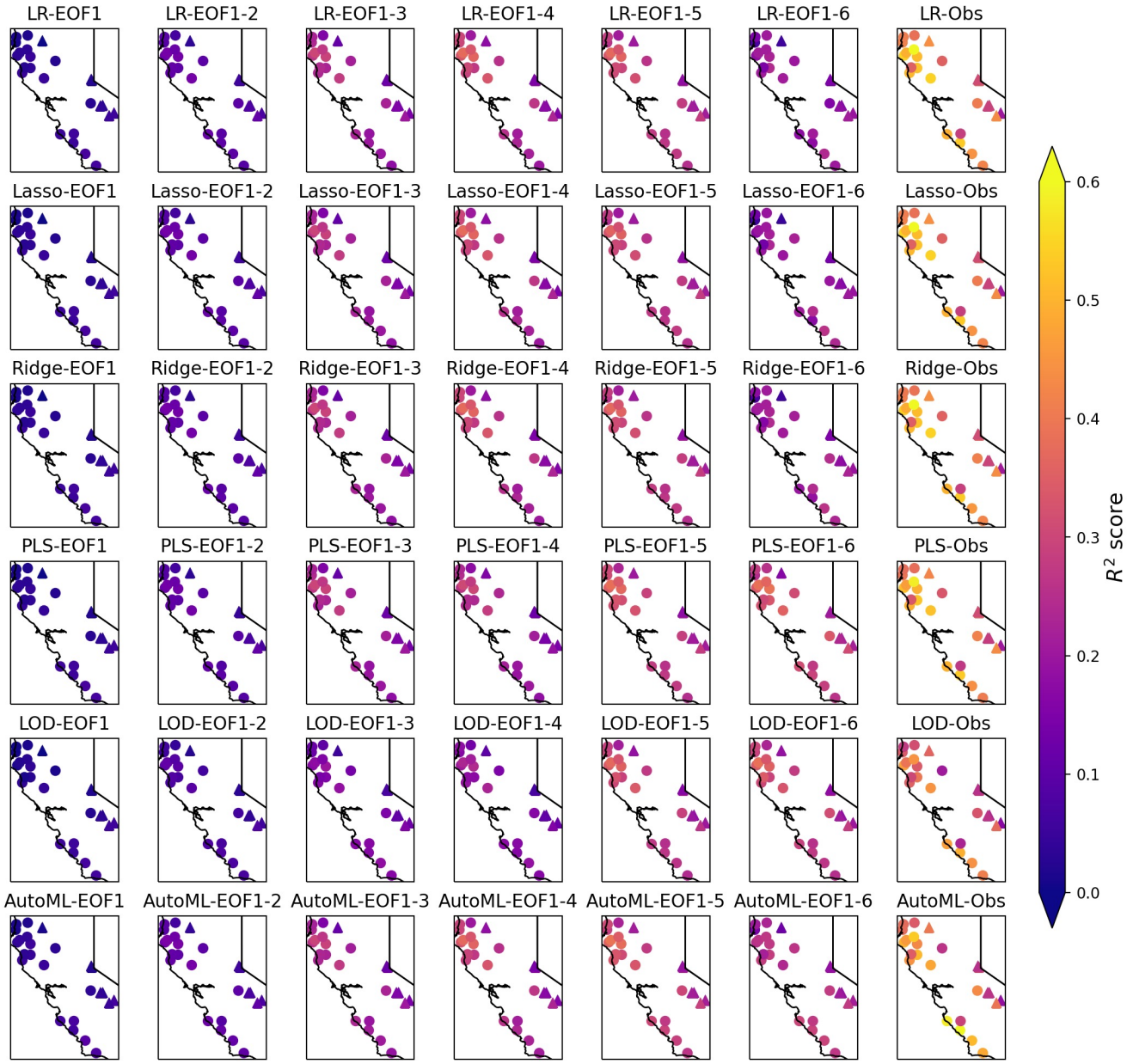


Figure. Cross-validation results from ML models (left 6 cols) and performance on observations (rightmost col).

Triangles for basins with peaks in summer time and squares for basins with winter time peaks.

ML Model Performance

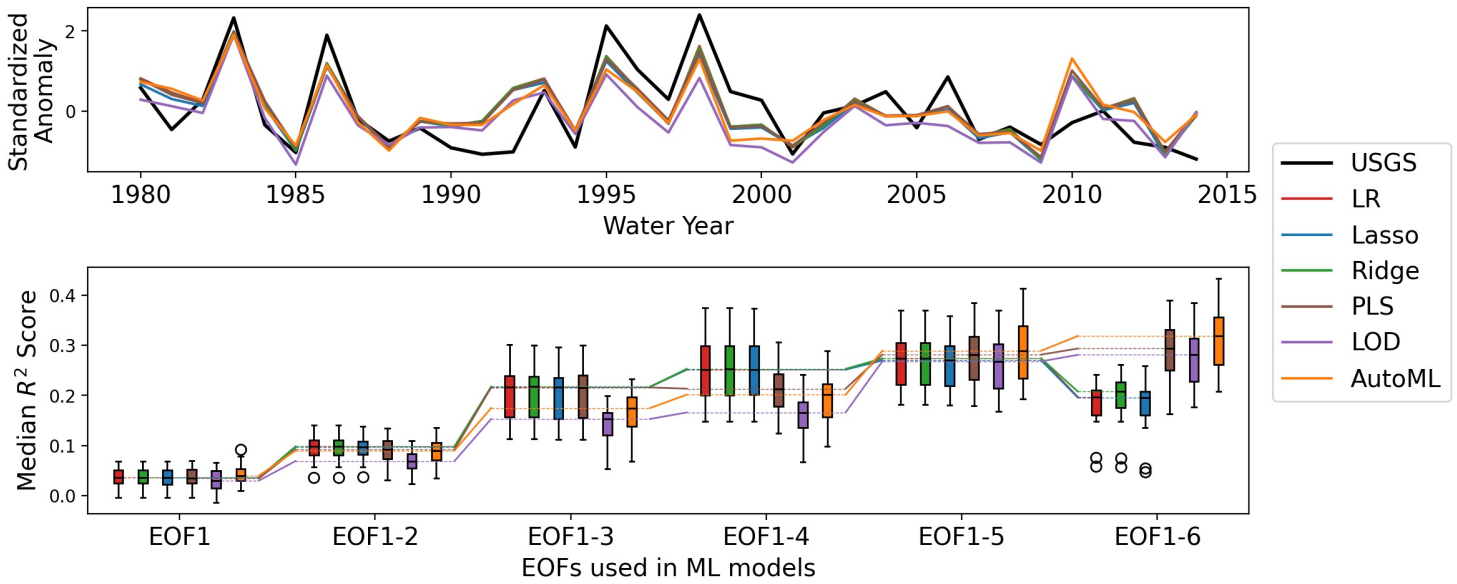


Figure. ML predictions of standardized peak flow anomalies vs observation for a selected basin (top); Cross-validation median performance with respect to the number of EOFs (bottom).

- ML models can generalize to reanalysis/observation
- Prediction skill increases with the number of EOFs
- AutoML, PLS and LOD have better regularization skills

Feature Importance

- Diagnose the feature importance by adding variables into ML models (isolate the predictability)
  - Compare the predictability from each variable separately
  - Cumulatively add variables to account the correlations among variables ("forward pass")
  - Feature importance for California in general (median R2)
- Retrain all the models with the cross-validation setting

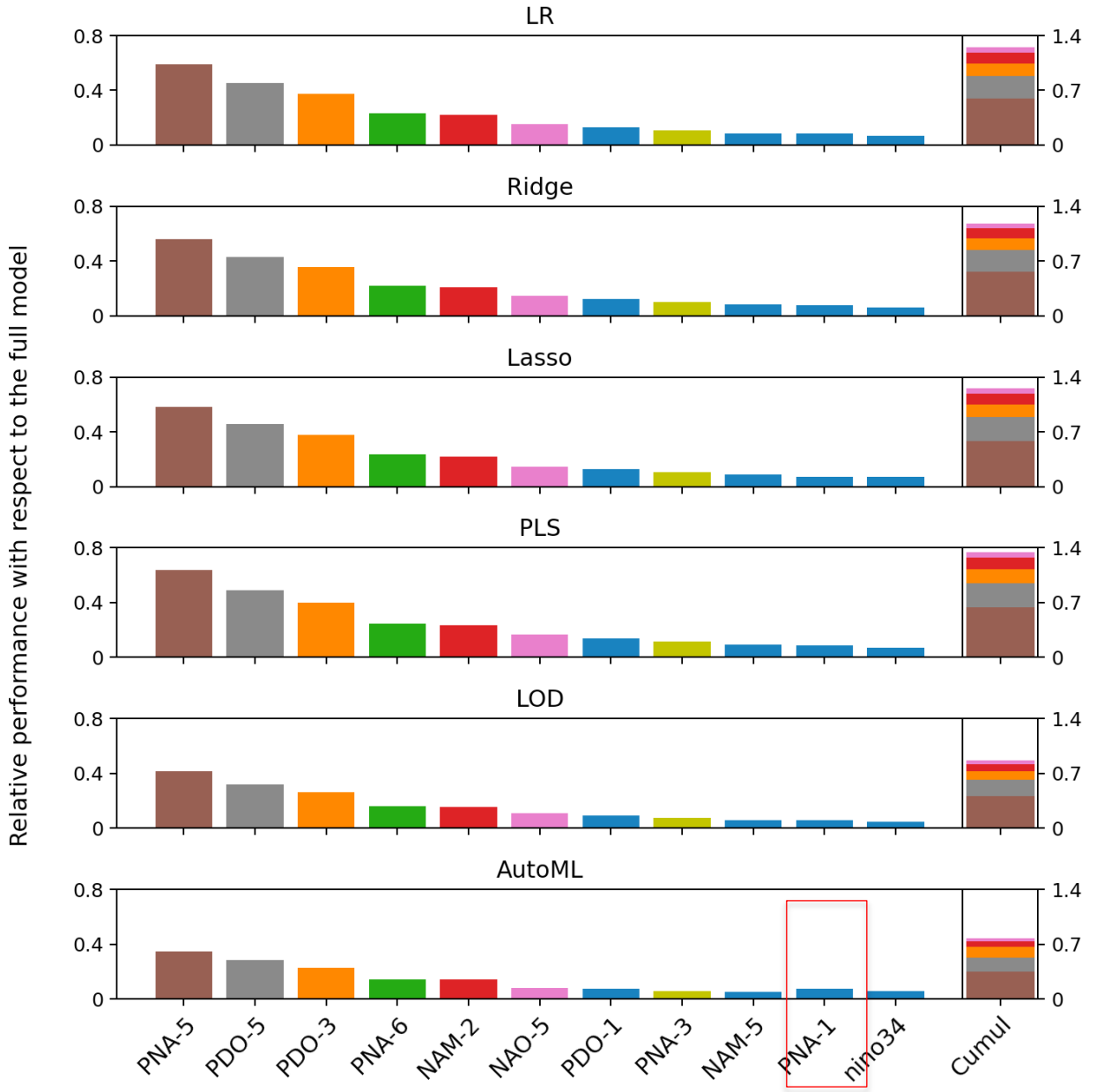


Figure. Feature importance ranked by predictability. Left columns for models with individual predictor. Rightmost column for models with variables added cumulatively.

Feature Importance

Focus on the common agreements from different ML models: **PNA-5** and **PDO-5**

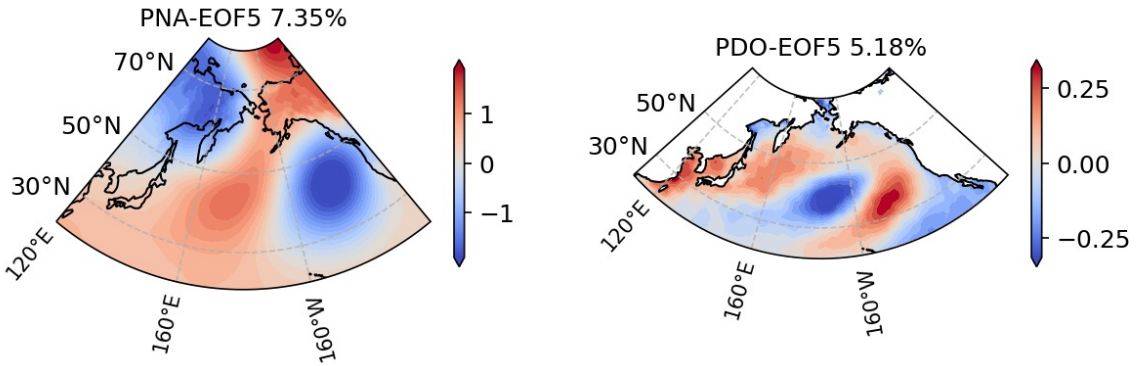


Figure. PNA-5 SLP pattern and PDO-5 SST pattern.

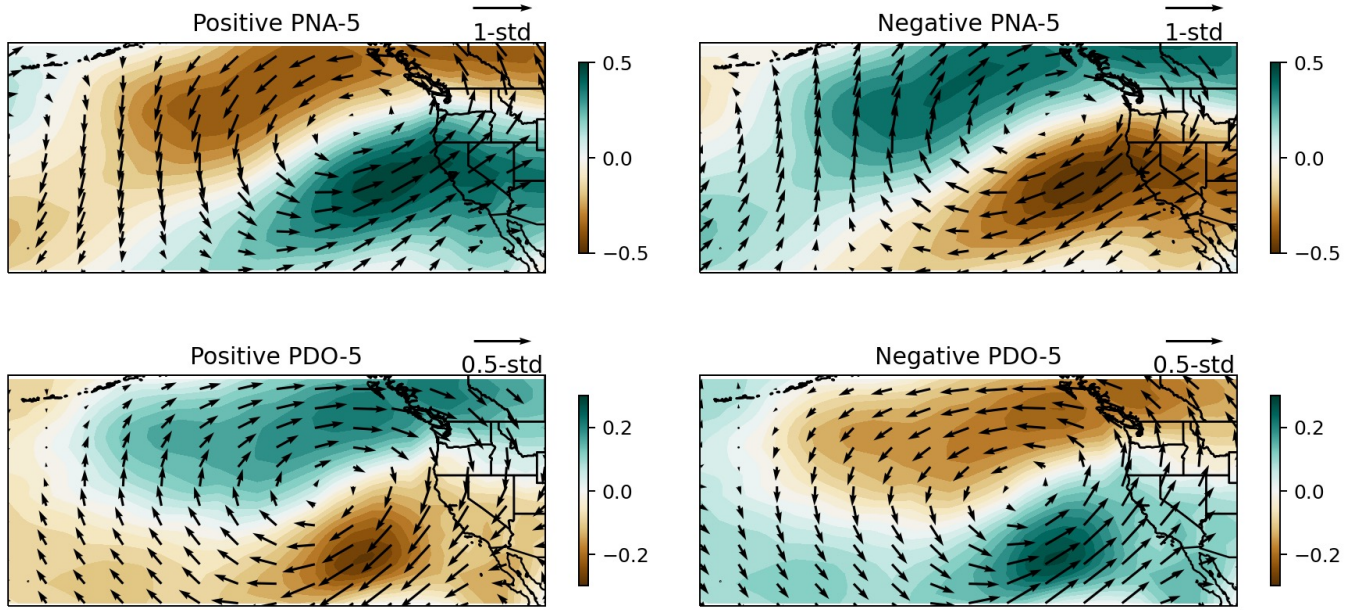


Figure. Composites of standardized anomaly of integrated water vapor transport from IPSL-CM6A-LR with respect to PNA-5 (top) and PDO-5 (bottom).

Feature Sensitivity

- How does PNA-5/PDO-5 affect streamflow? Is it linear or nonlinear?
  - Use **AutoML** as a testbed since it considers potential nonlinear relationship and has the highest overall skill score.
  - Use **accumulated local effect (ALE)** to account for correlations among features.
- The response shows “quasi-”linear.

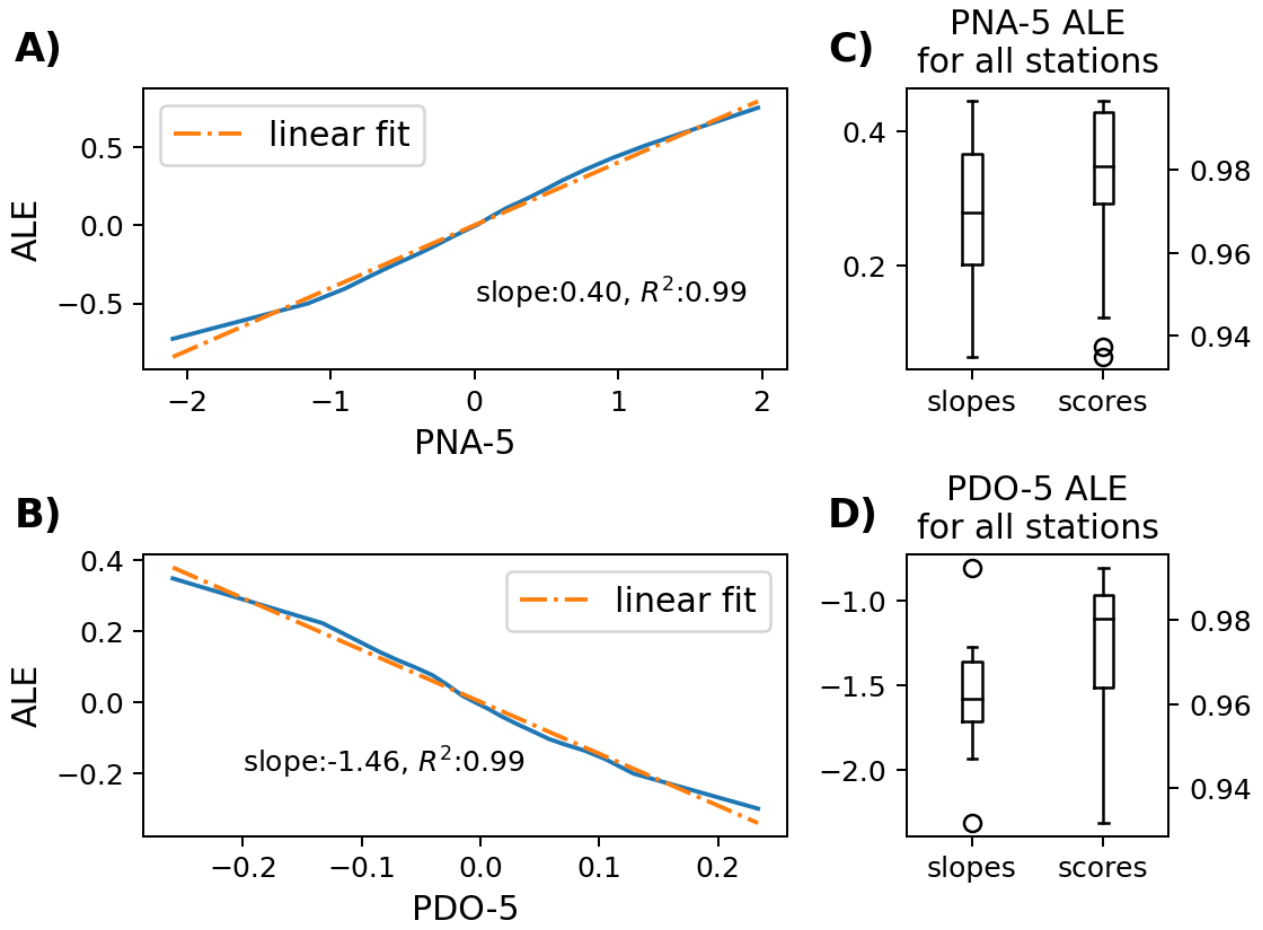


Figure. Accumulated Local Effect (ALE) for PNA-5 (A) and PDO-5 (B) at a selected basin. Distribution of ALE slope and linear fit score over California for PNA-5 (C) and PDO-5 (D)

Conclusions

- Higher-order modes of variability show better predictability for basin-scale streamflow changes than well-know variability indices;
- By taking the higher-order of EOFs from well-known variability domains, we can connect between local features and larger-scale patterns.
- ML models agree on the feature importance, and PNA-5 and PDO-5 are the two most dominant features.
- The relationship is quasi-linear.
- BCSD preserves variability. Useful for variability-related studies.

Future Work

- Western US snowpack detection: since BCSD can preserve internal variability, similar framework can be used to generate snowpack projections (CMIP6->BCSD->DL/process-based models) and applications.

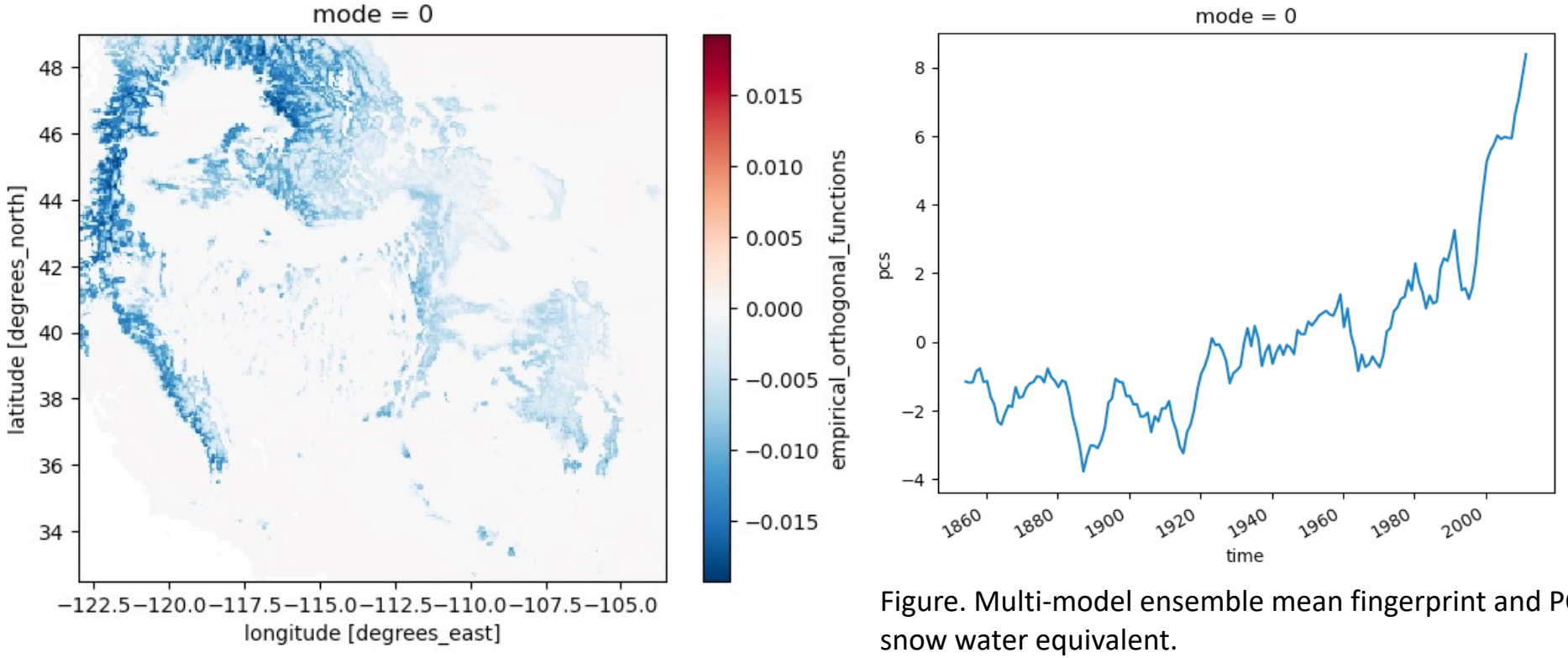


Figure. Multi-model ensemble mean fingerprint and PC of normalized snow water equivalent.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Internal Variability Domains

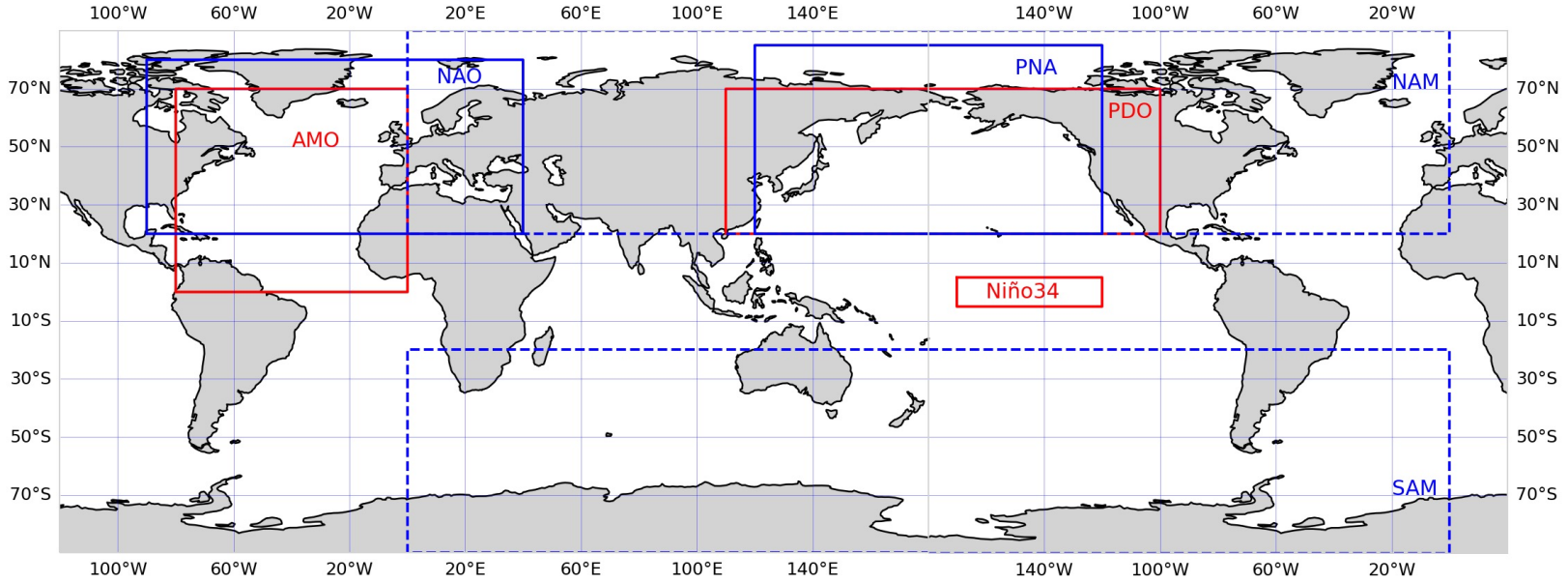
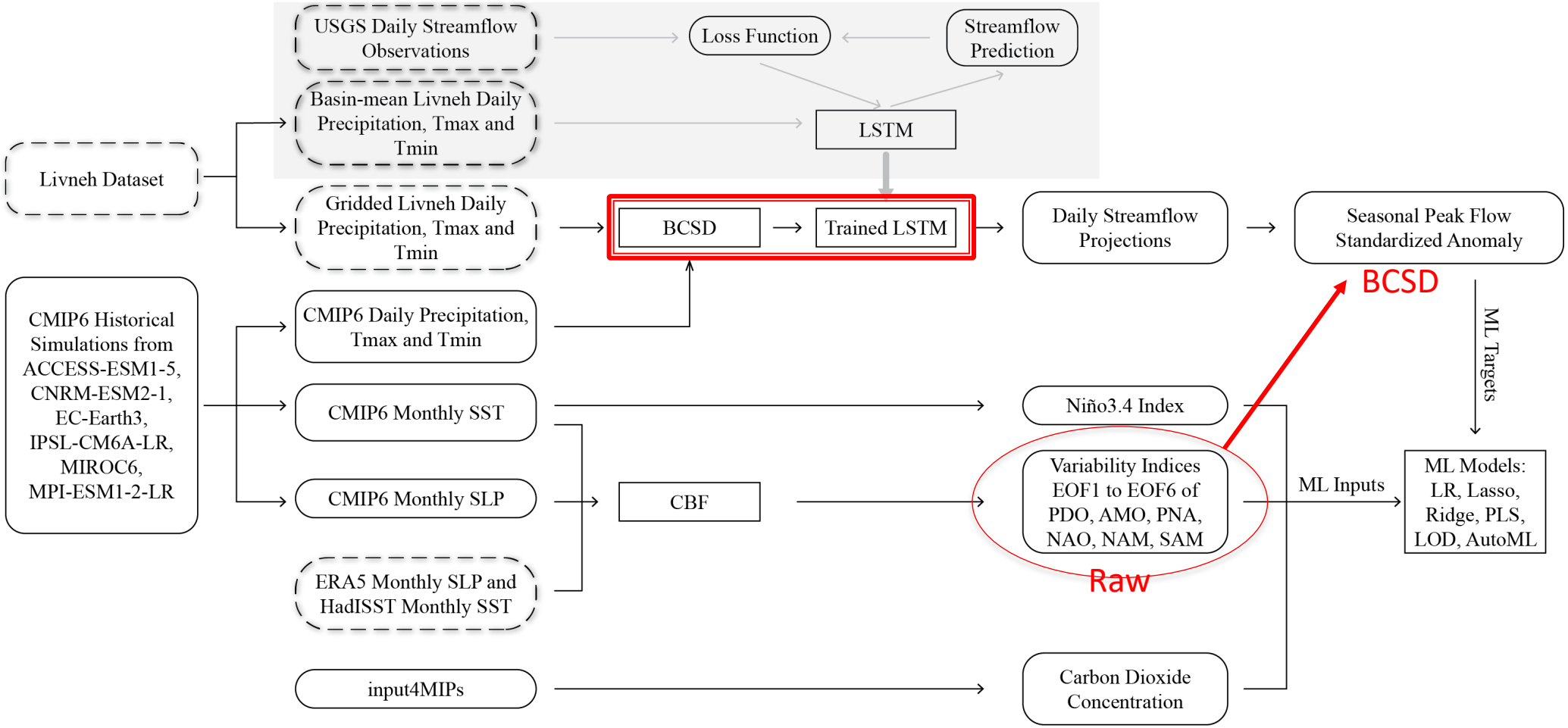


Figure. Internal variability domains. Dashed lines for NAM and SAM.

Technical Workflow



LSTM Performance

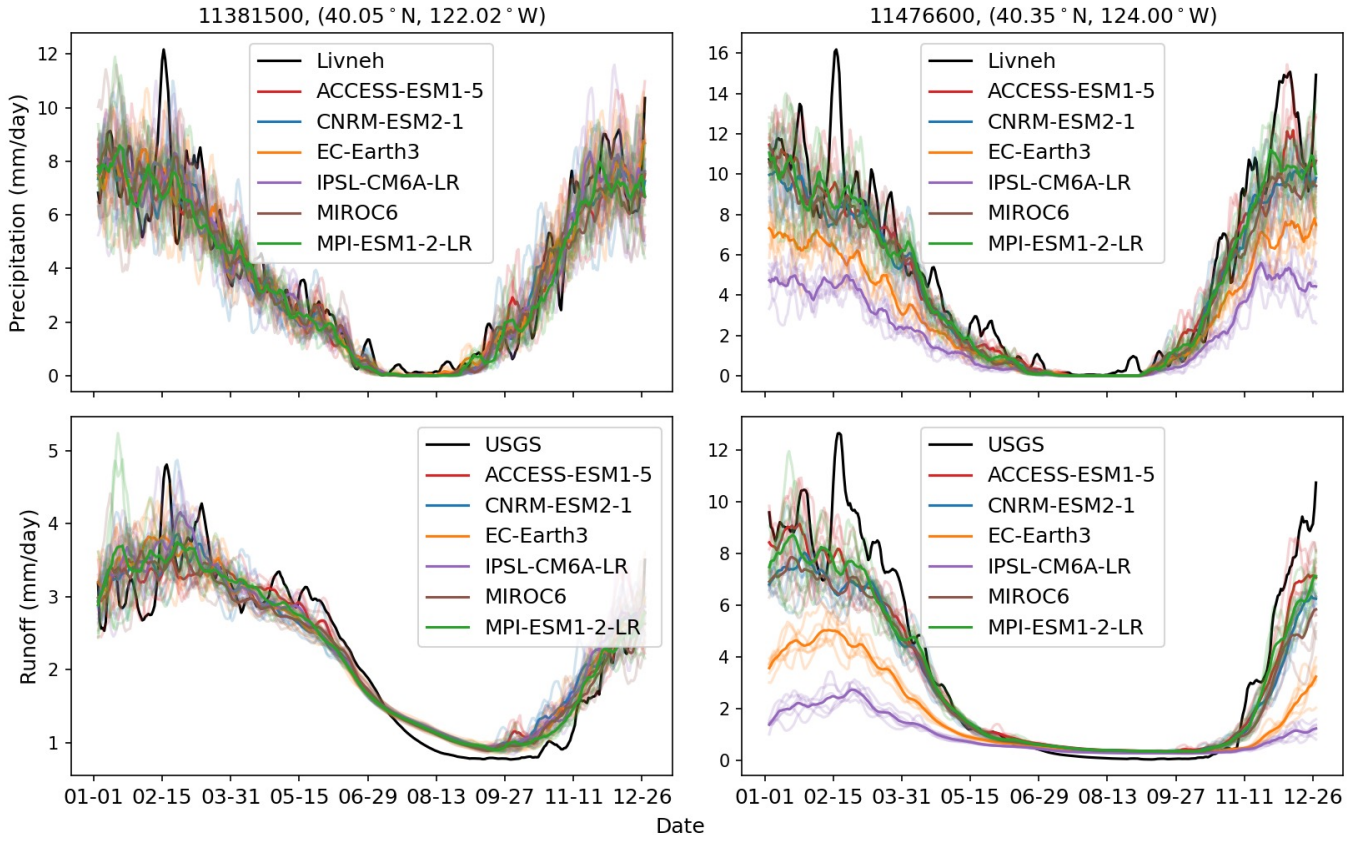
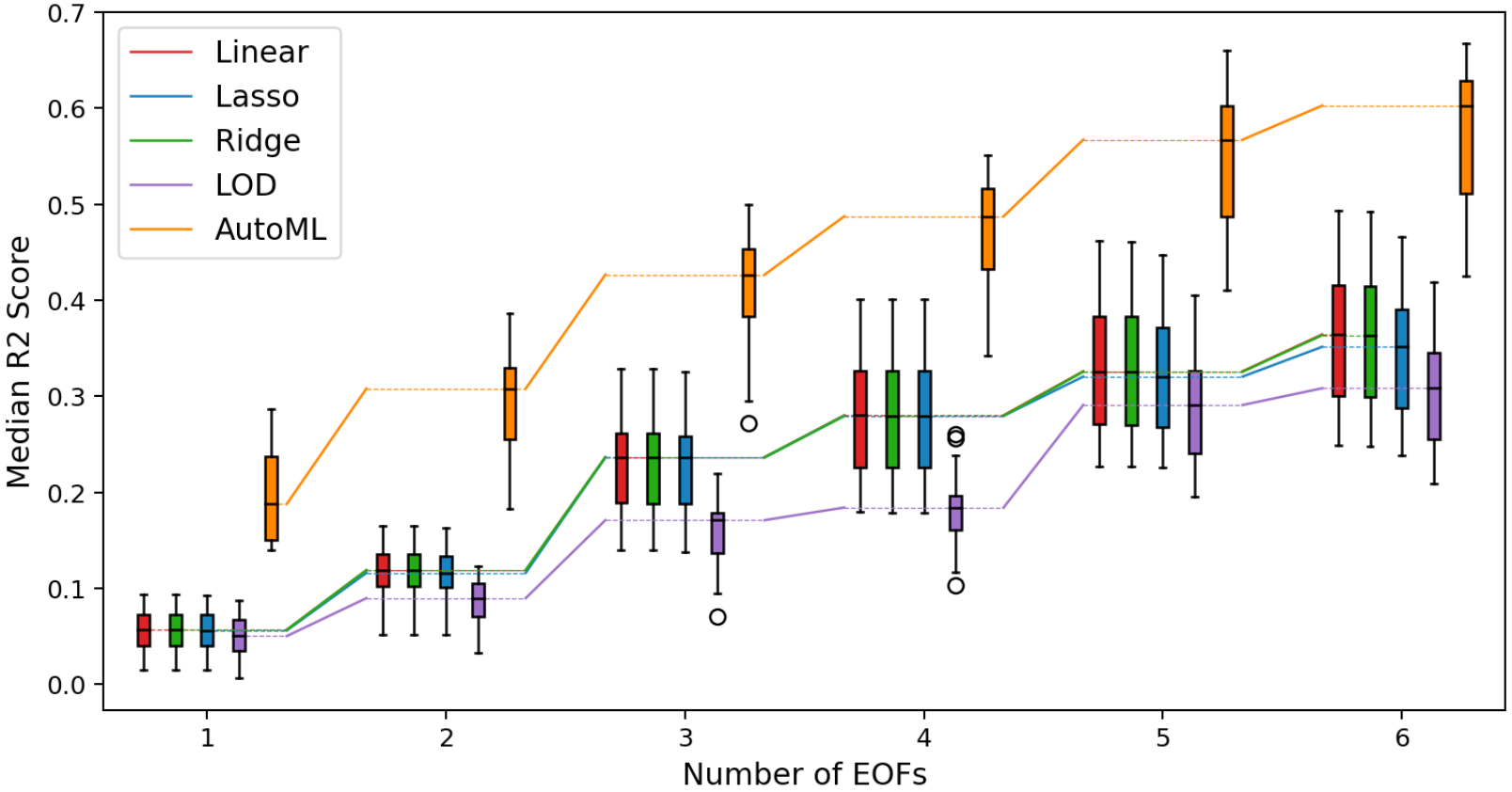


Figure. Daily precipitation climatology (top) and streamflow climatology (bottom) from two selected basins (left and right columns).

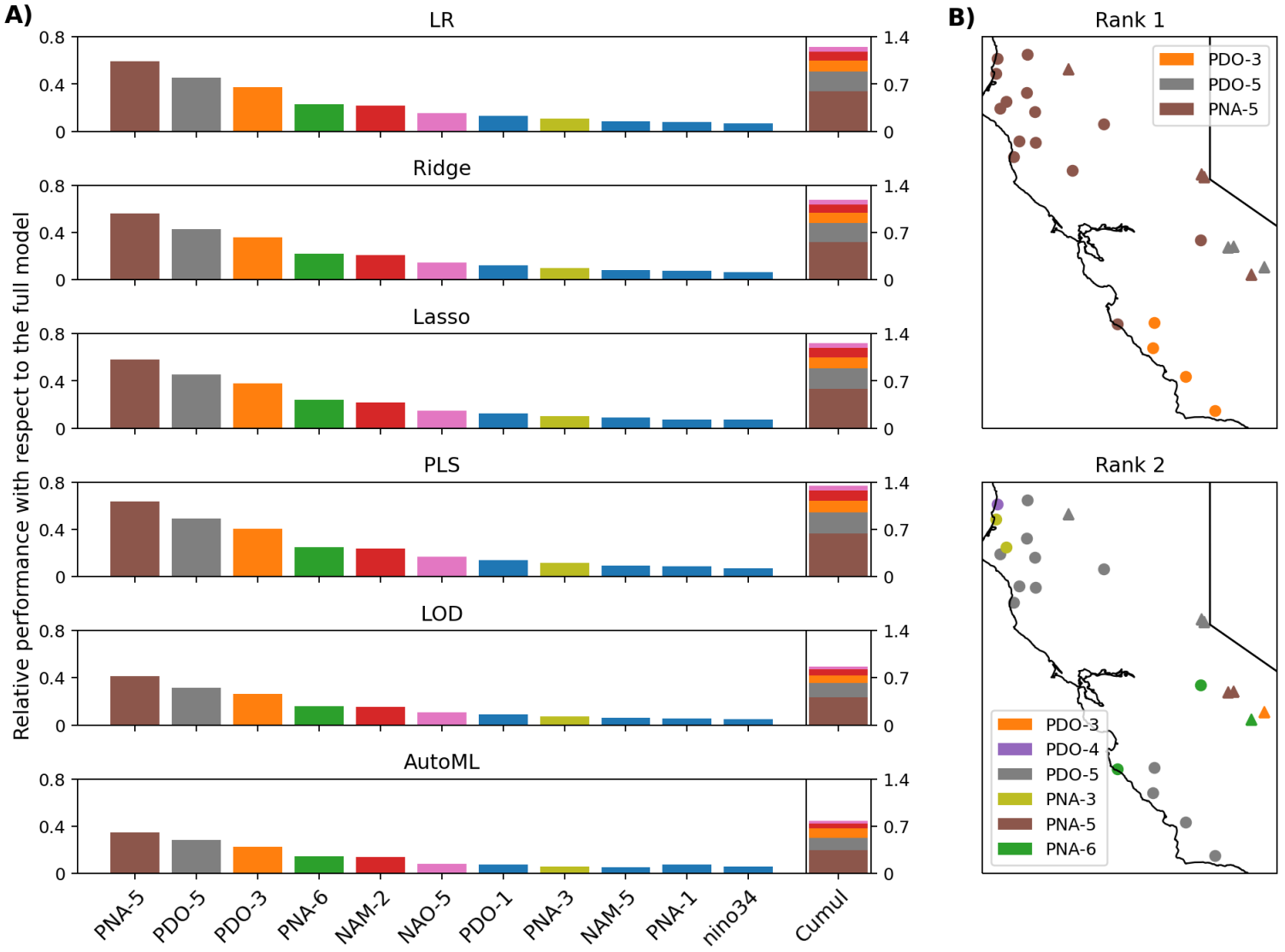
- LSTM accuracy of daily streamflow prediction has been proved
- For historical projection, the bottleneck is the biases in precipitation
- To mitigate the biases, use standardized streamflow anomalies
- Output/Target: LSTM-projected seasonal peak streamflow
  - *Peak month, month prior and after*
  - *Standardized anomalies*
- Input: seasonal variability indices
  - *Concurrent season or previous season*

Training cross-validation R2



Training R2 with respect to number of EOFs.

Feature Importance



Cross validation