UCDAVIS
ATMOSPHERIC SCIENCE

Using Machine Learning Models to Reveal Teleconnections -- A case study on the North American Monsoon

Shiheng Duan, David John Gagne, Paul Ullrich

UC Davis, NCAR

Nov 15, 2021

Previous work on teleconnections

- Teleconnections are mostly defined as the mode of climate variables, such as sea surface temperature and sea level pressure. The popular indices include Nino34, Nino3, SOI, PAO.
- Saha, M., Mitra, P., & Nanjundiah, R. S. (2016) used an autoencoder to identify predictors of Indian monsoon.
- Tang, Y., & Duan, A. (2021) used a CNN model to predict the East Asian summer monsoon.

- Idea: Use ML models for the North America Monsoon (NAM) precipitation prediction.
- Can we reveal teleconnection areas from the interpretation of ML models?
- Can we show the physical causation from ML models?

Research tasks

Identify the North American Monsoon Area.

Build ML models for daily precipitation prediction and analyze the possible teleconnection areas.

Build ML models for monthly scale and analyze the corresponding teleconnection effects.

Identify the NAM domain

- Global monsoon domain identification:
  - Wind reversal
  - Precipitation pattern: local summer-minus-winter precipitation exceeds 300 mm, and the local summer precipitation exceeds 55% of the annual total.
  - NAM: from subtropical America, expanding to the southwestern of US.

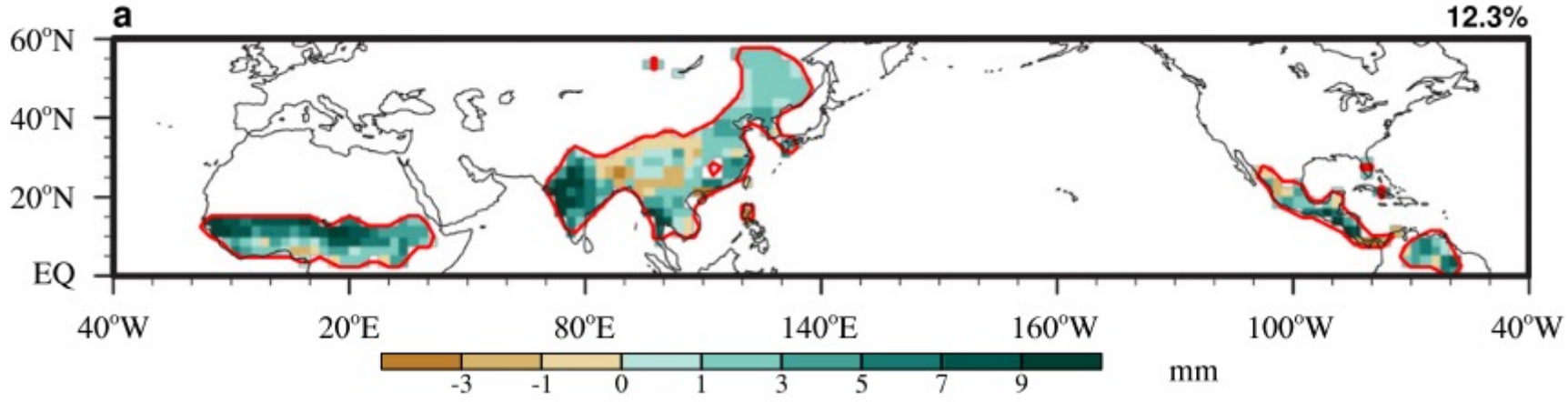


Figure: Spatial pattern of the first EOF mode of the decadal variation of the summer monsoon precipitation over the NH land monsoon regions. (Wang et al., 2018)

Identify the NAM domain

- Localized NAM domain: determined by the ensemble results from Self-organizing maps (SOMs).
  - CPC-global precipitation dataset is used. The cubic root of LTDM precipitation is first normalized to [0, 1], and then used for SOMs clustering. (Swenson and Grotjahn, 2019)
  - The number of nodes ranges from 10 to 20.

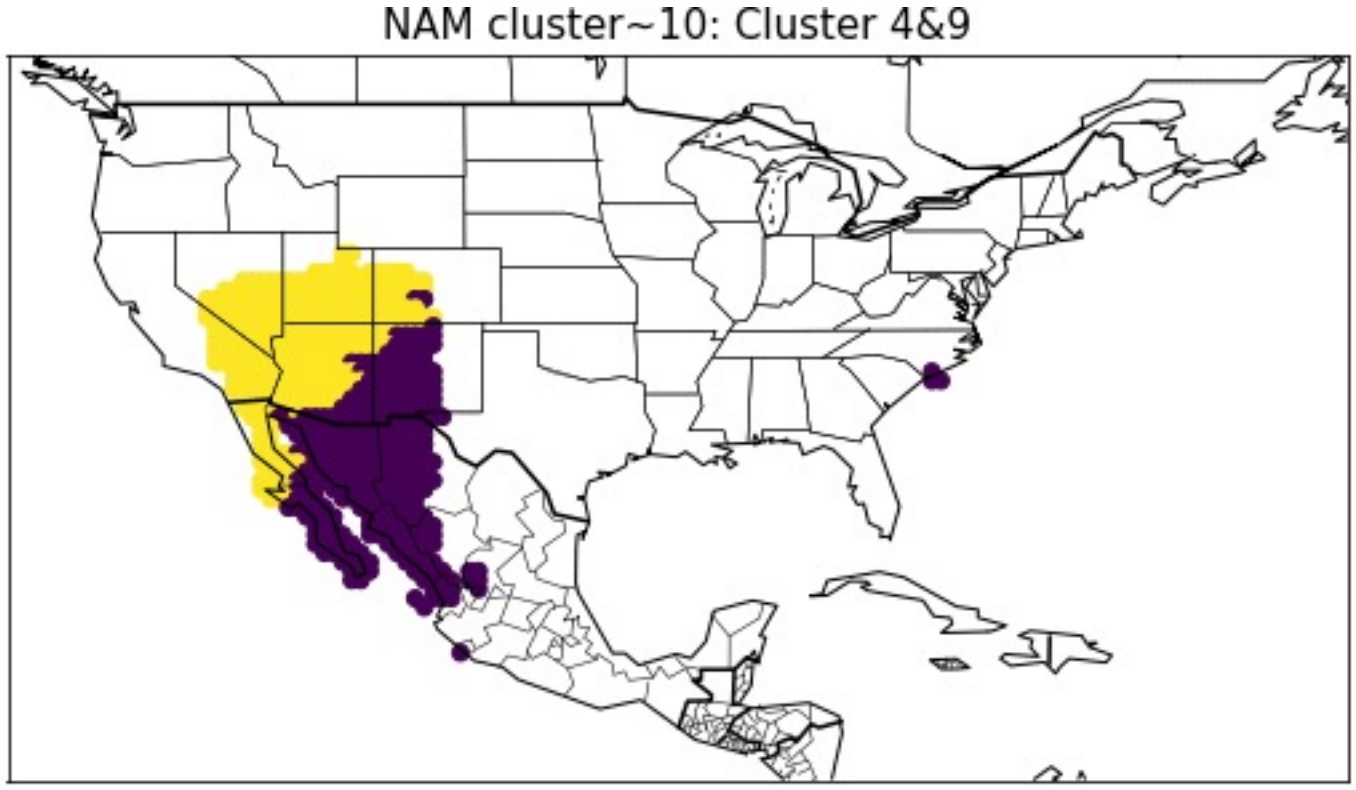


Figure: Different results from SOMs ensembles.

Identify the NAM domain

The NAM domain is the intersection of 11 ensembles from SOMs, excluding Baja California (Englehart and Douglas, 2001) and any other singular points.

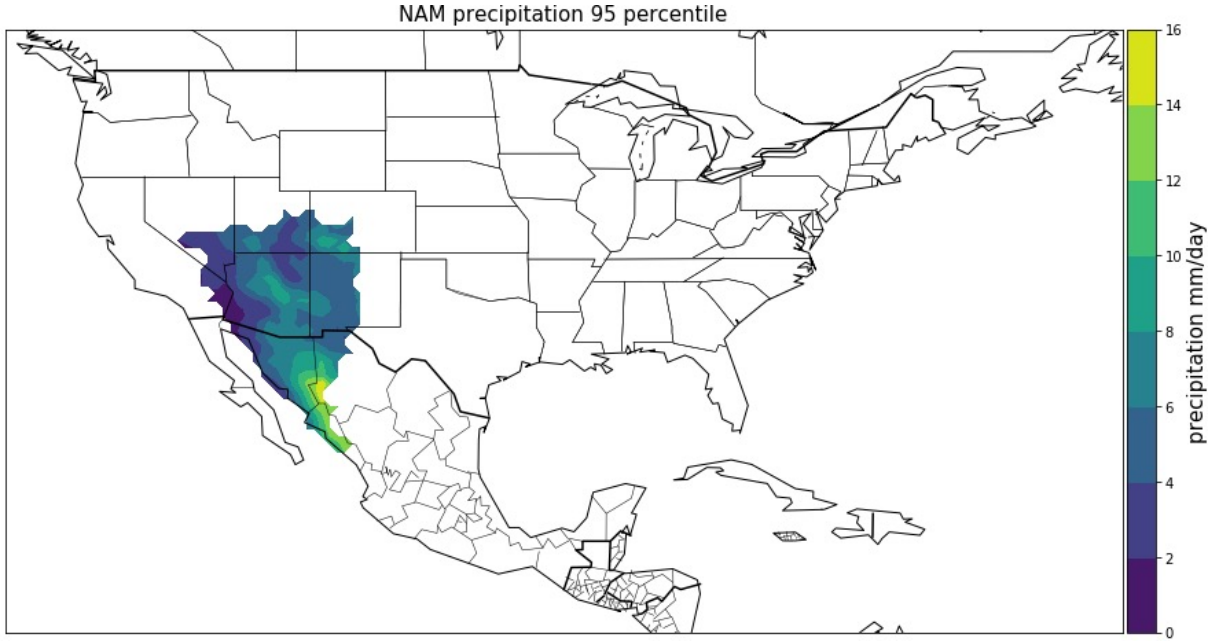


Figure: NAM region with 95th percentile precipitation rate.

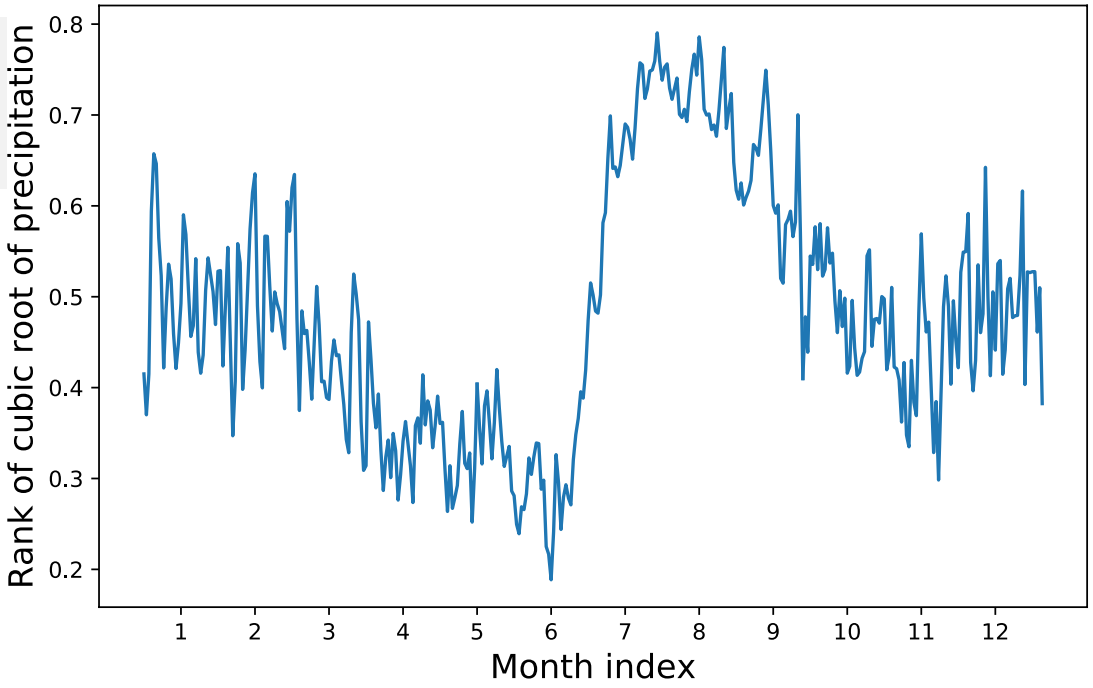


Figure: Normalized cubic root of precipitation of the NAM region.

The rapid increase in precipitation signal after June shows the monsoon impact.

The identified domain is then used as the NAM mask and can be applied to any other dataset.

NAM precipitation in climate datasets

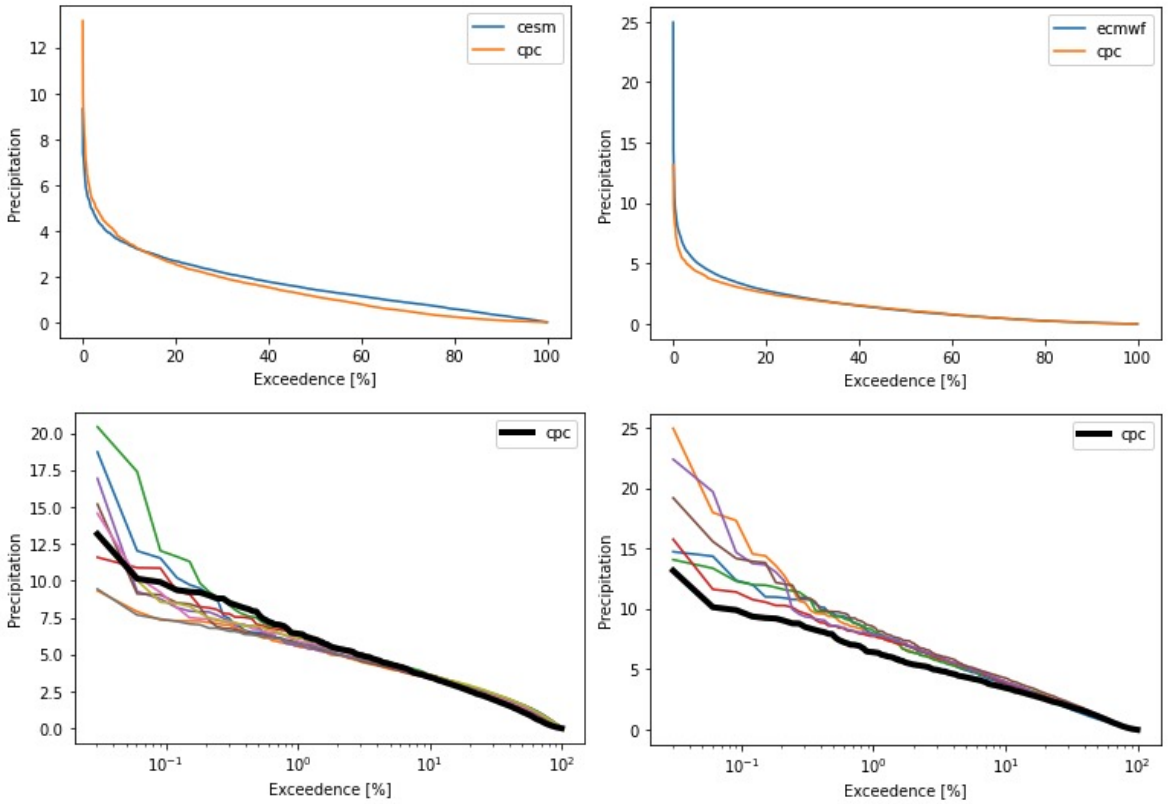


Figure: Comparing NAM precipitation from CESM LENs and ECMWF (HighRes) against CPC.

Dataset	95th percentile	95th percentile for each ensemble
CPC	4.343	
CESM1	4.210	4.082
		4.477
		4.322
		4.163
		4.001
		4.059
		4.188
		4.237
		4.353
ECMWF	5.110	4.913
		5.072
		5.028
		5.181
		5.153
		5.305

Research tasks

Identify the North American Monsoon Area.

Build ML models for daily precipitation prediction and analyze the possible teleconnection areas.

Build ML models for monthly scale and analyze the corresponding teleconnection effects.

Daily precipitation prediction model

Hypothesis:
The disturbance from a remote area will affect the meteorology conditions over the NAM area, and by ML models, we can trace back to the origin of the predictability.

ML models:
PCA+LSTM
Use PCA to decompose high-dimensional meteorology fields and LSTM for time dependency

CNN+LSTM
Use CNN to extract spatial patterns and LSTM for time dependency

Inputs: PSL, Z500, Q850 anomaly
Outputs: mean precipitation in the NAM region.

Daily PCA model

PCA model with no time dependency:
Linear regression with different PCs.

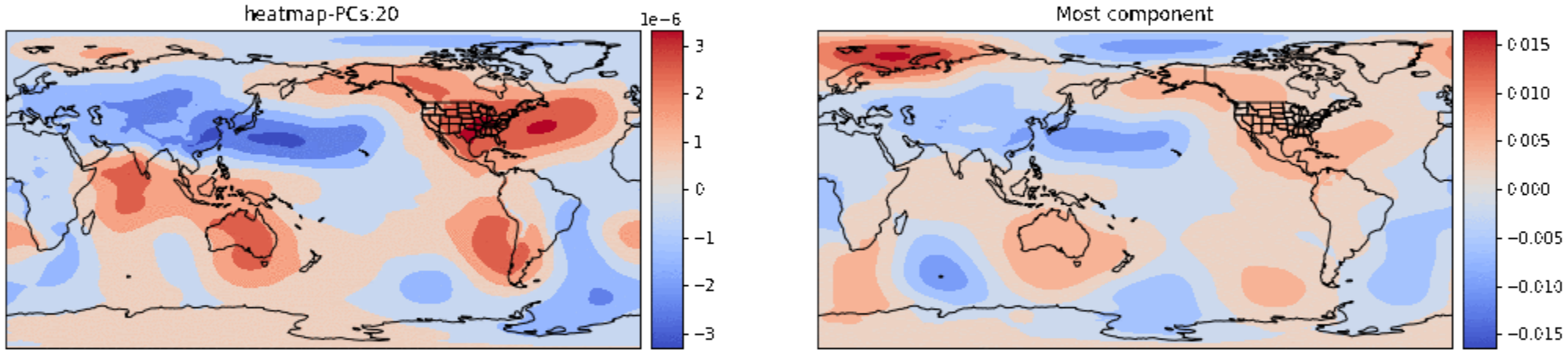


Figure: Heatmap from the linear model. Heatmap = (Linear weight * component)/# of PCs.

Daily PCA model

PCA+LSTM
Time lag: 14 days.
day 0 to day 13 as inputs and day 14 as output.

Model architecture:

Layer	Output Shape	
Input	Batch, 14, PCs	Day 0 – Day 13 PCA variables
LSTM	Batch, 14, Hidden	LSTM outputs
Dense	Batch, 14, 150	
Slice	Batch, 1, PCs	Take the last frame, PC prediction
Dense	Batch, 1, 1	Day 14 precipitation prediction

Loss function: 5*MSE(PC_true, PC_pred) + MSE(Precip_true, Precip_pred)

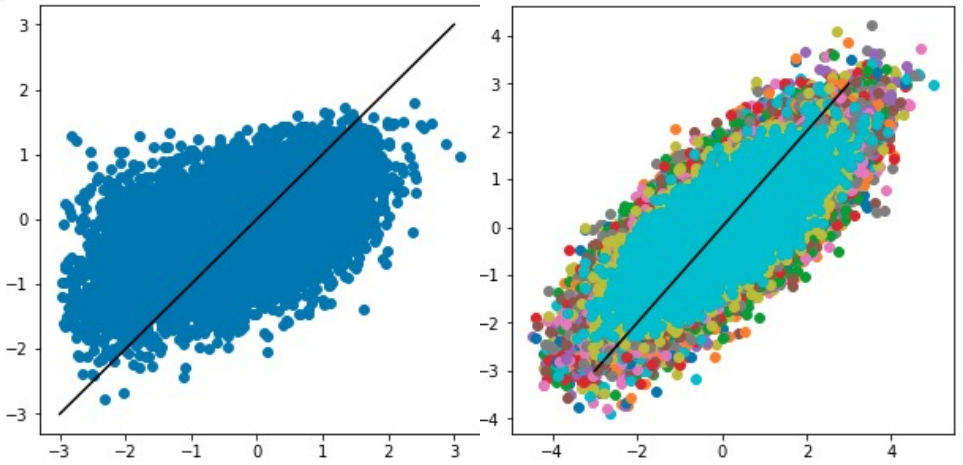
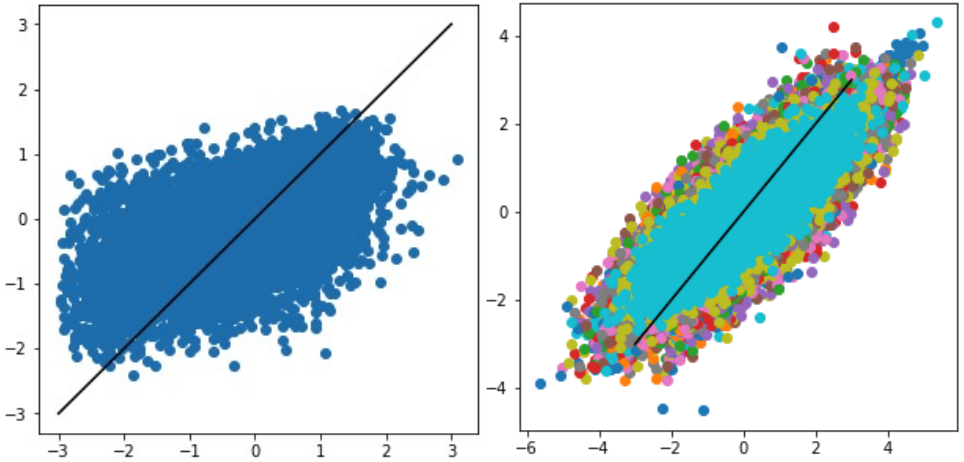


Figure: Prediction result with PSL(up) and Z500(bottom). Precipitation on the left (R2=0.304 with PSL and 0.335 for Z500), PCs on the right.



Daily PCA model

Integrated gradients of PC prediction

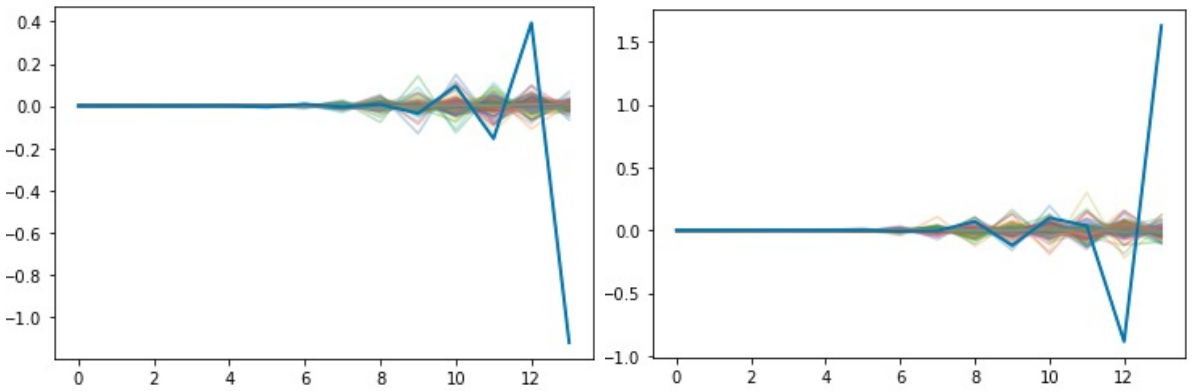


Figure: IG fields of PC prediction. Left for PC0 and right for PC10.

Integrated gradients of precipitation prediction

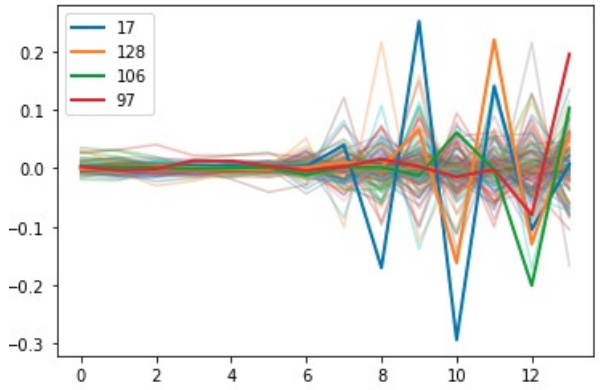


Figure: IG fields of Precipitation prediction.

Heatmap = PC x attr x sign
sign = transformed / abs(transformed)

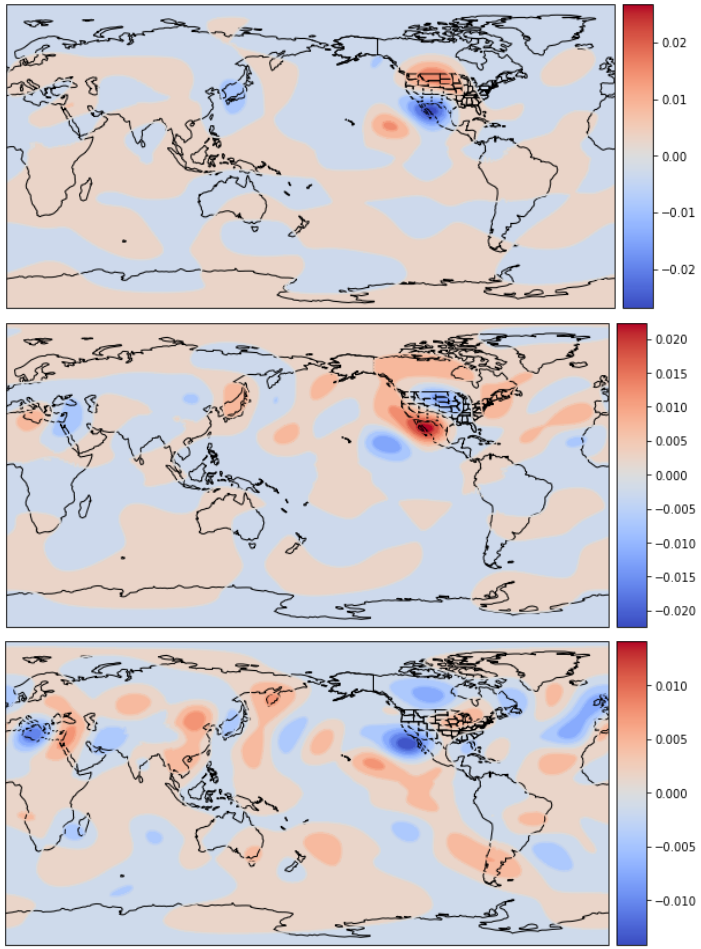


Figure: Heatmap from principal components.

Daily CNN model

CNN+LSTM
Time lag: 14 days.
day 0 to day 13 as inputs and day 14 as output.

Model architecture:

Layer	Output Shape	
Input	Batch, 14, (192, 288), 1	Day 0 – Day 13 PSL or Z500
Conv-Pool	Batch, 14, (16, 10), 16	(16, 10) spatial
Dense	Batch, 14, 256	
LSTM	Batch, 14, 512	
Dense	Batch, 1, 1	Day 14 precipitation prediction

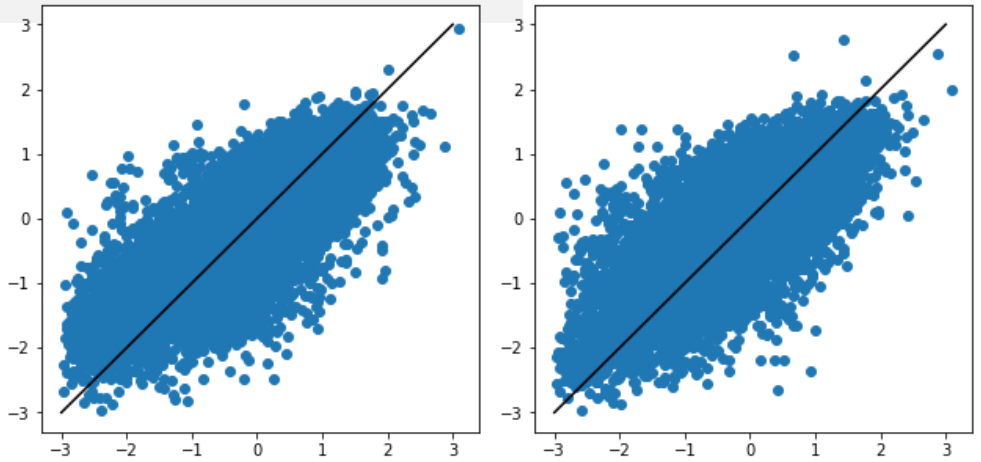


Figure: Prediction result with PSL(left) and Z500(right).
R2=0.582 with PSL and 0.564 for Z500

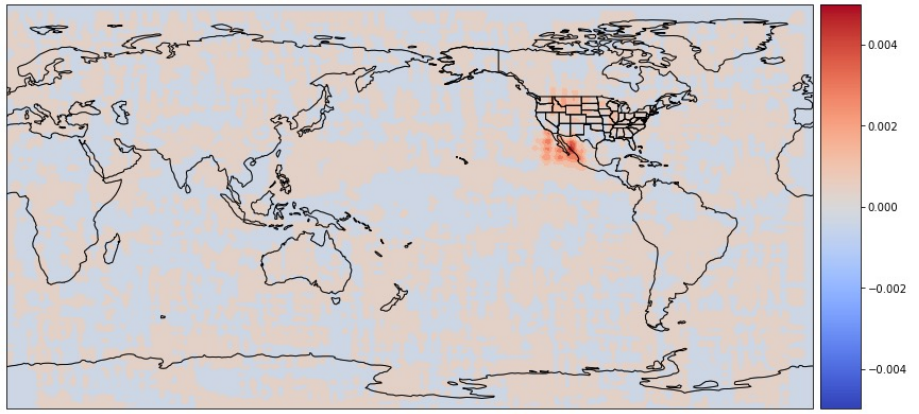


Figure: mean of IGs from precipitation over 95th percentile.

Research tasks

Identify the North American Monsoon Area.

Build ML models for daily precipitation prediction and analyze the possible teleconnection areas.

Build ML models for monthly scale and analyze the corresponding teleconnection effects.

Monthly precipitation prediction model

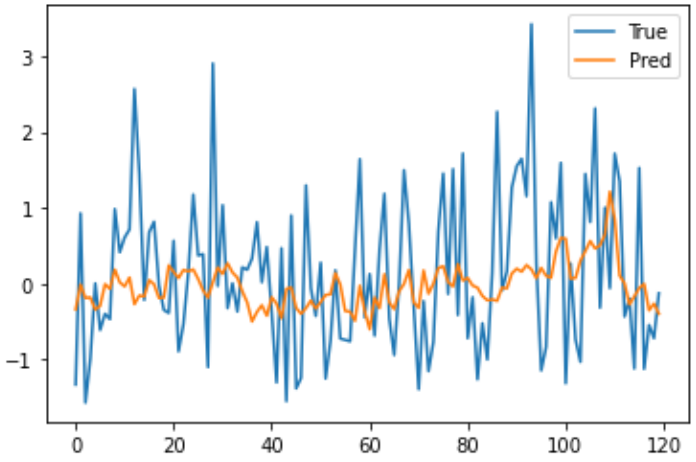
The teleconnection impact is weak in daily scale. Switch to monthly field to focus on the longer time scale.

CNN Model
Time lag: 1 month.
Month 0 as input and Month 1 as output.

Model architecture:

Layer	Output Shape	
Input	192, 288	Spatial dimension for CESM
Conv-pool	9, 15	Spatial size for feature map
Dense	256	
Dense	1	Precipitation anomaly prediction

Input features: 'PSL', 'Z500', 'TMQ', 'Q850', 'TREFHT'
R2_score: 0.095



	Positive	Negative
True	1699	2025
False	1373	1089

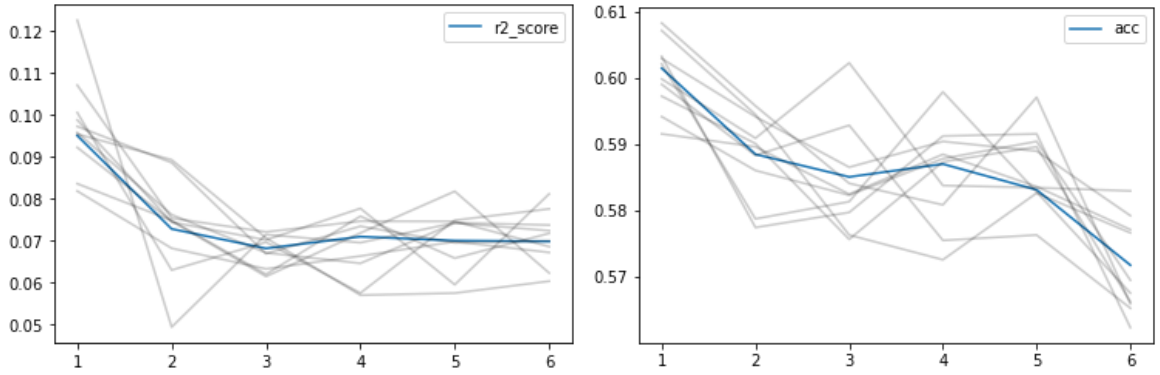


Figure: Prediction skill with different month lags.

Monthly precipitation prediction model

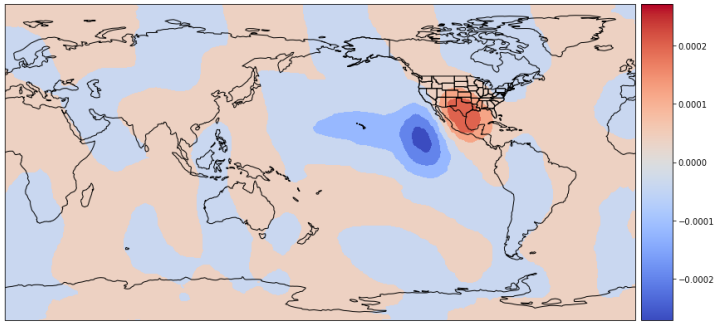


Figure: Mean of gradient (up) and mean of absolute gradient (bottom).

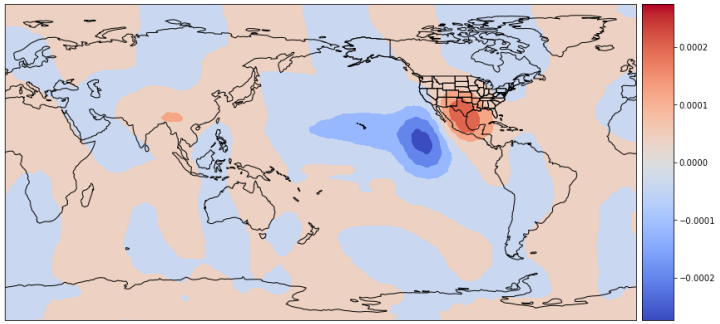


Figure: Mean of gradient in DJF (up) and mean of gradient in JJA (bottom).

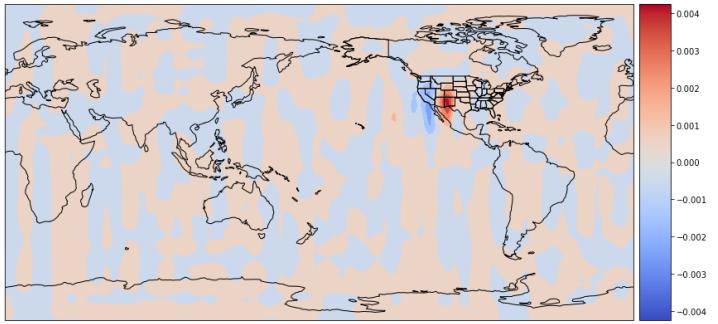
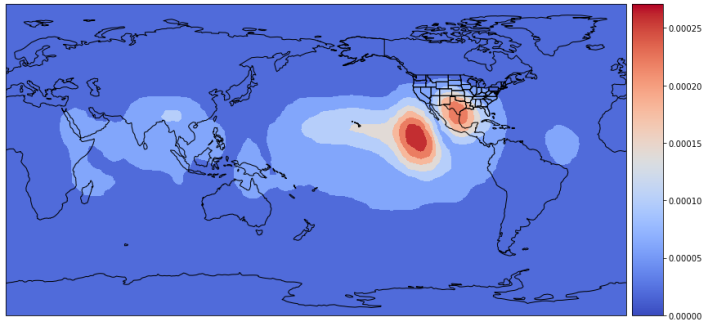
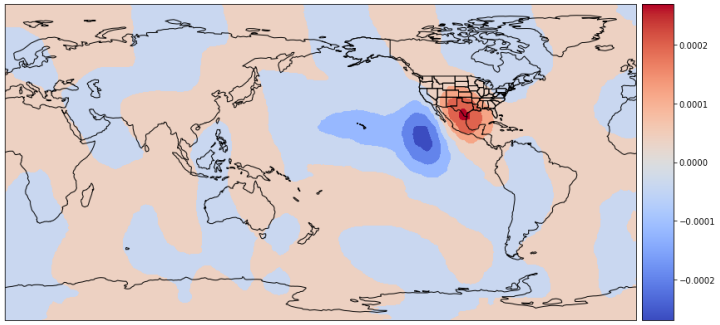
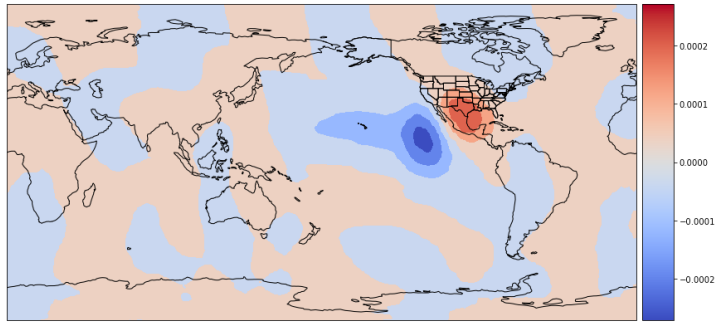


Figure: Mean of gradient with concurrent month (up) and mean of gradient with one month lag (bottom).







Monthly precipitation predictability

To show the predictability from these areas, we permute the samples and compare the drop in R2 score.

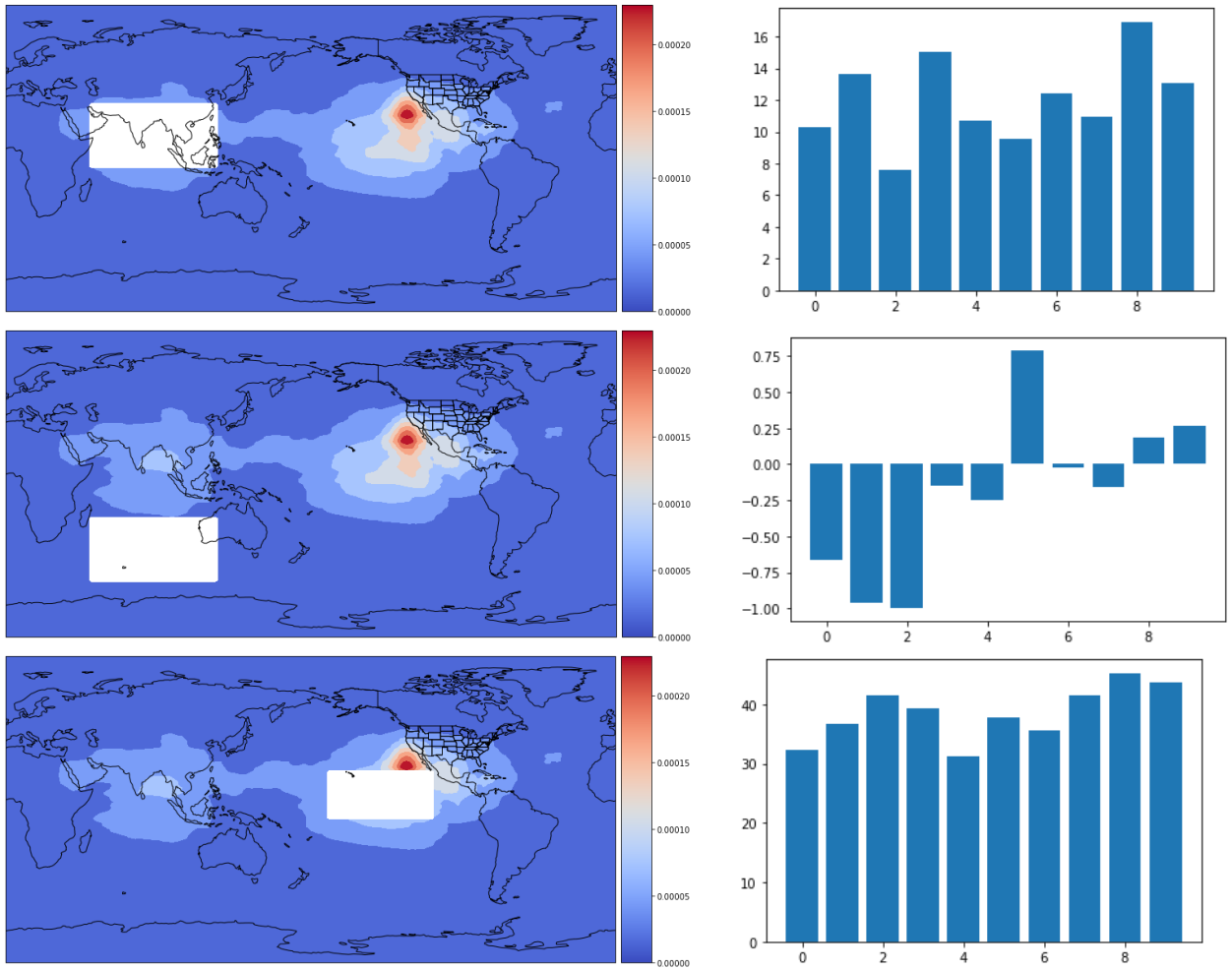


Figure: Permutation area (left) and r2 drop ratio (right).

Compare with teleconnection indices, train and test with the same ensembles as CNN model:

Nino34: 0.053
IOD: 0.009

	R2_drop (%)
PSL	2.27
Z500	2.39
TMQ	-0.07
Q850	0.83
TREFHT	1.78

Monthly precipitation causation

Partial dependency plot: permute the PSL in the South Asia and Pacific area and see the response of the predicted precipitation.

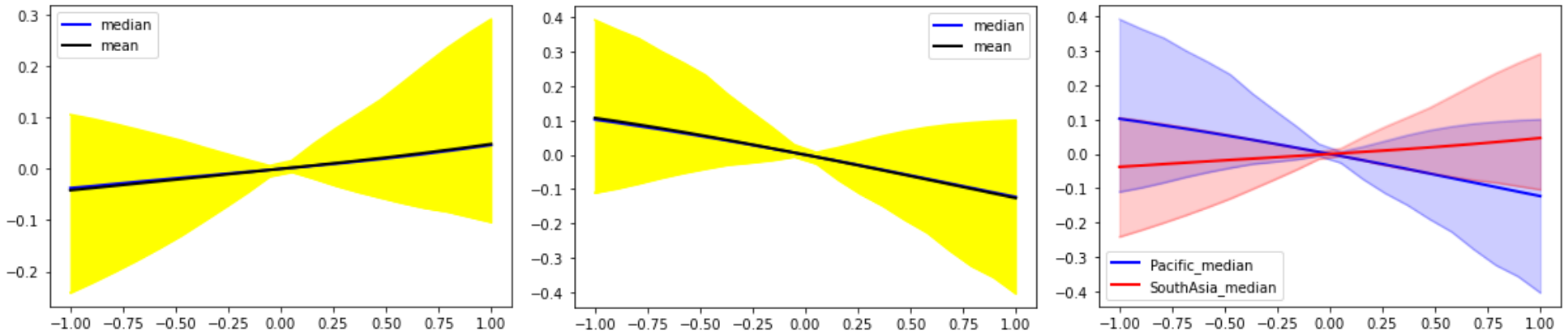
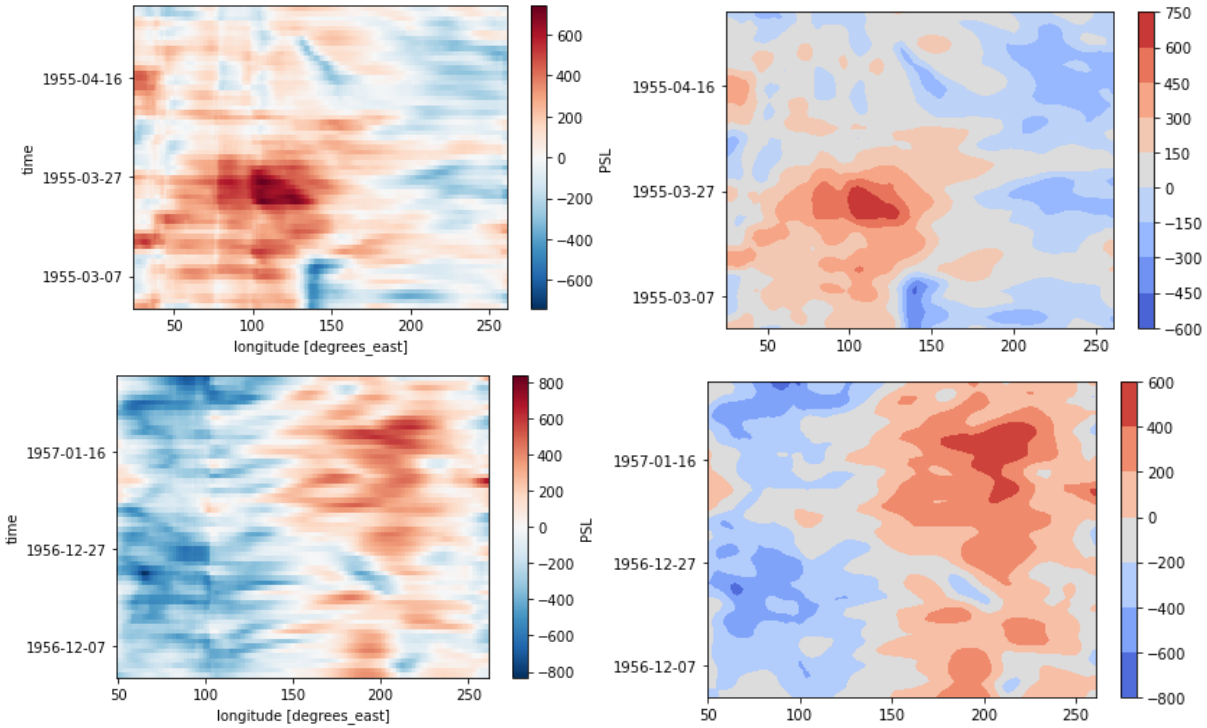


Figure: Response of prediction when perturb PSL in South Asia (left) and Pacific (middle).

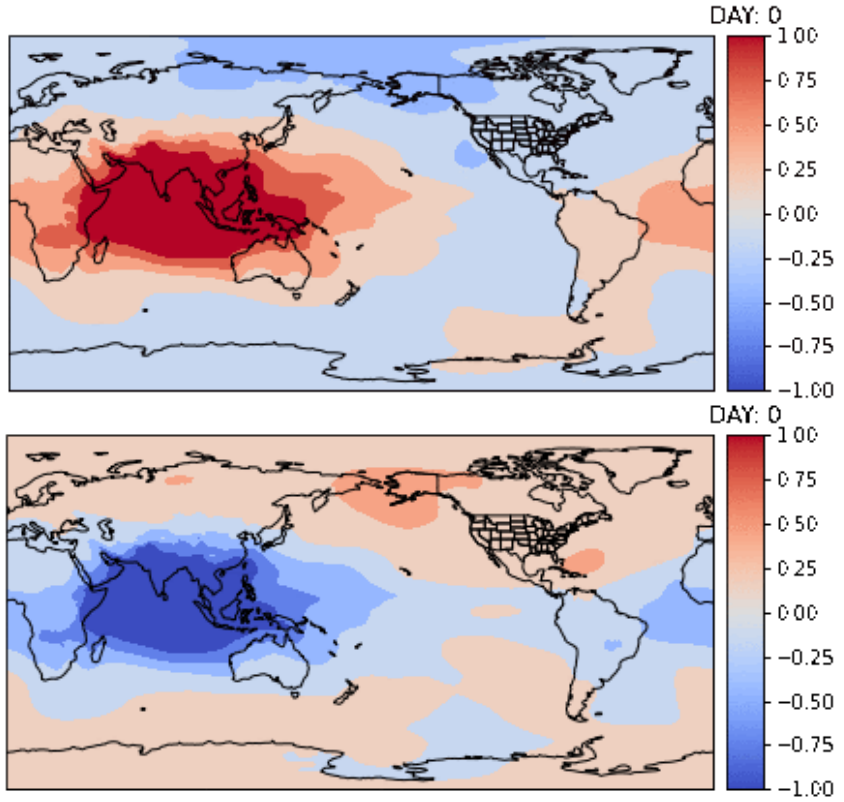
Decrease the PSL over South Asia (Pacific) will decrease (increase) the precipitation in the NAM region.

Monthly precipitation causation

Use Hovmoller diagram to check if there is any wave train from South Asia to the NAM area.



Composite daily fields



Summary

- Teleconnection is hard to be detected with daily precipitation.
- With global features as input, the CNN model highlights the Pacific and South Asia area as the potential source of predictability.
- Nino34 can provide a better prediction compared with other indices, but with limited domain, it is not comparable to the CNN model.
- The interpretation can not reveal the causation of monthly precipitation prediction. We still need physical-based models to test our hypothesis.

Thanks