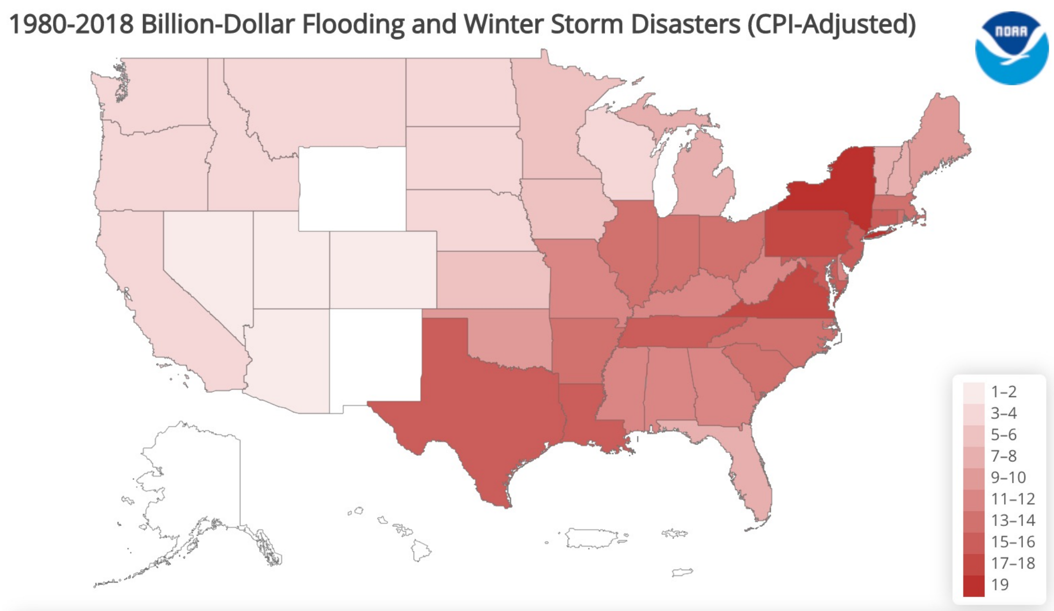
Examining the Utility of Random Forests to Forecast Subseasonal Extreme Precipitation in the Contiguous United States

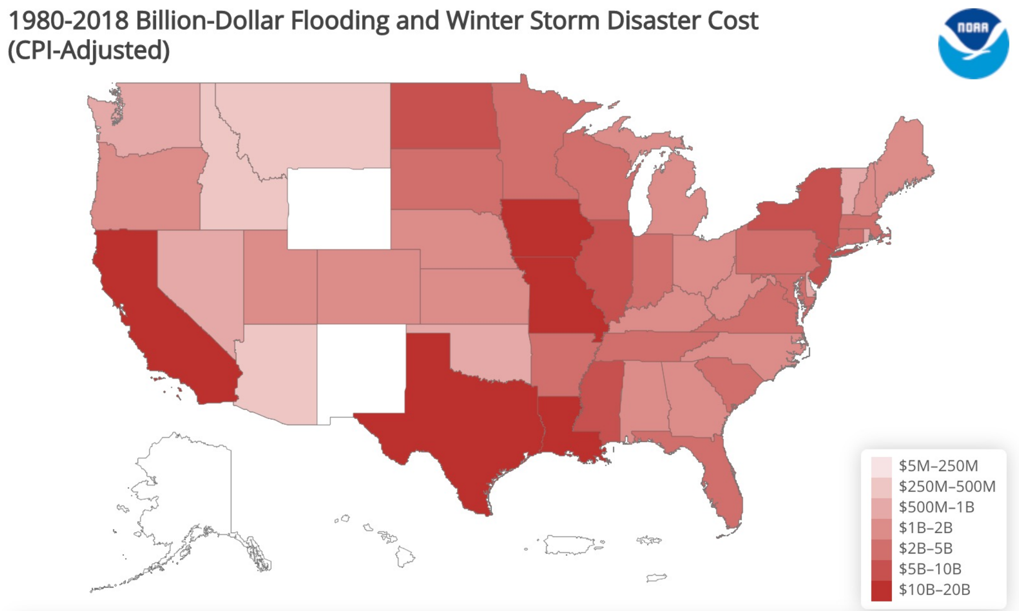
Ty A. Dickinson, Jason C. Furtado, Michael B. Richman

46th Annual Climate Diagnostics and Prediction Workshop
October 26, 2021

Motivation

- Extreme precipitation affects nearly the entire CONUS.
- Stakeholders in water resources, agriculture, energy, and emergency managers and tribal leaders all are impacted.





Credit: National Center for Environmental Information

Motivation

- Forecasting extreme precipitation is exceptionally difficult!
    - Statistical methods can help improve forecasts (e.g., Herman and Schumacher 2018).
- Barlow et al. (2019) review paper details extreme events with duration up to 1 week.
- <u>**What about extreme precipitation events on longer timescales?**</u>
    - ➢Subseasonal forecasts currently lack skill.
    - ➢Can we make improvements to subseasonal extreme precipitation forecasting?

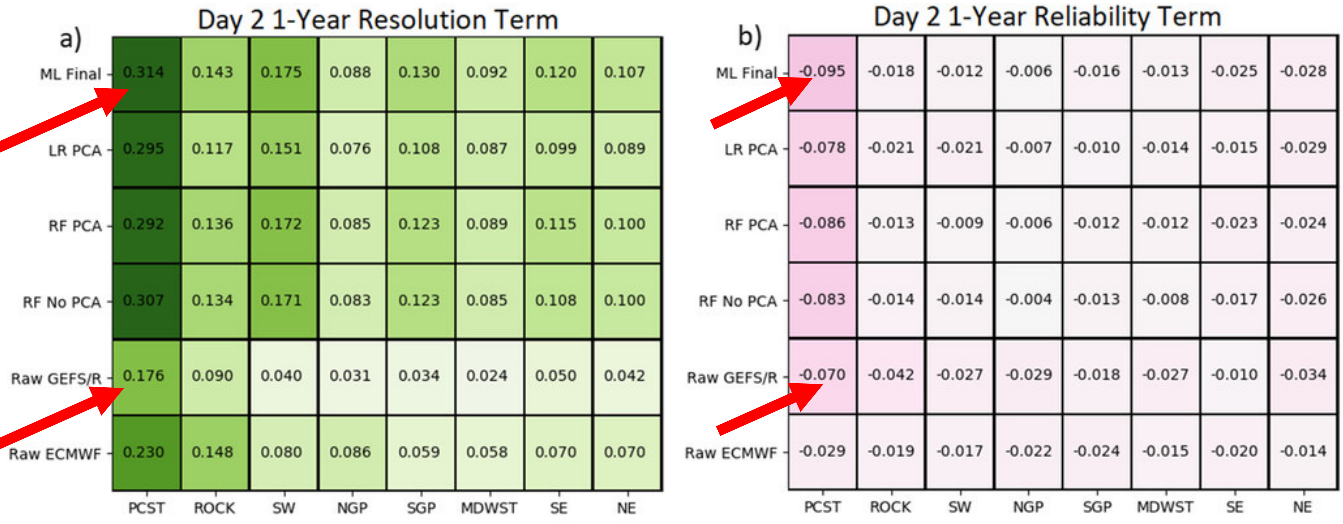
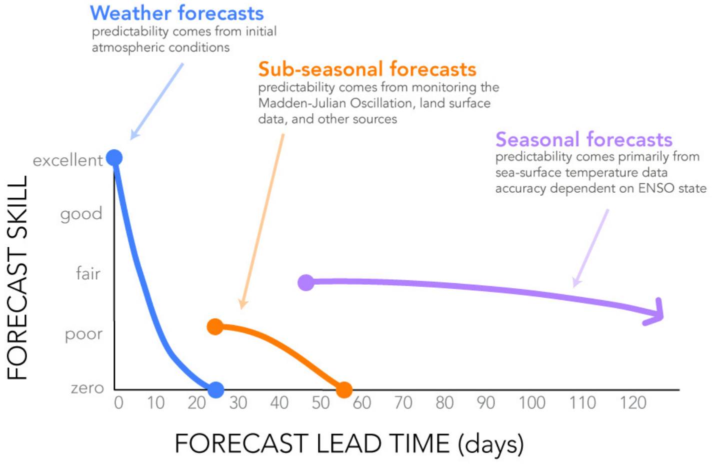


Figure 10 from Herman and Schumacher (2018)



Adapted from the International Research Institute for Climate and Society

The PRES^{2}iP Team

- Prediction of Rainfall Extremes at Subseasonal to Seasonal Periods
- Goals of the project:
  1. Define databases of S2S extreme events.
  2. **Quantify statistical and dynamical links between S2S extreme events and synoptic-scale and global scale precursors.**
  3. **Improve capability to predict S2S extreme events.**
  4. Increase communication between research scientists and stakeholder communities.

Research Questions

1. How well can a random forest (RF) classify days as being extreme or not extreme within the Central Plains and Ohio River Valley?
2. Which atmospheric variables are most important in classifying extreme versus non-extreme days and where is their importance maximized?

Overview of the Database

- Events in the database (Dickinson et al. 2021) are **large-scale, longer-duration extreme** events.
  - ➢ Exceed 99th percentile, more than 7 days of above normal daily precipitation, have areal extent ≥ 200,000 km^2.
  - ➢ *k*-means clustering on events to get "regions"
- Timeframe: 1950 – 2018
  - ➢ Training data: 15 Jan 1950 – 28 Dec 2000
  - ➢ Testing data: 15 Jan 2001 – 28 Dec 2018

❖ **Event independence** ensured in database generation.

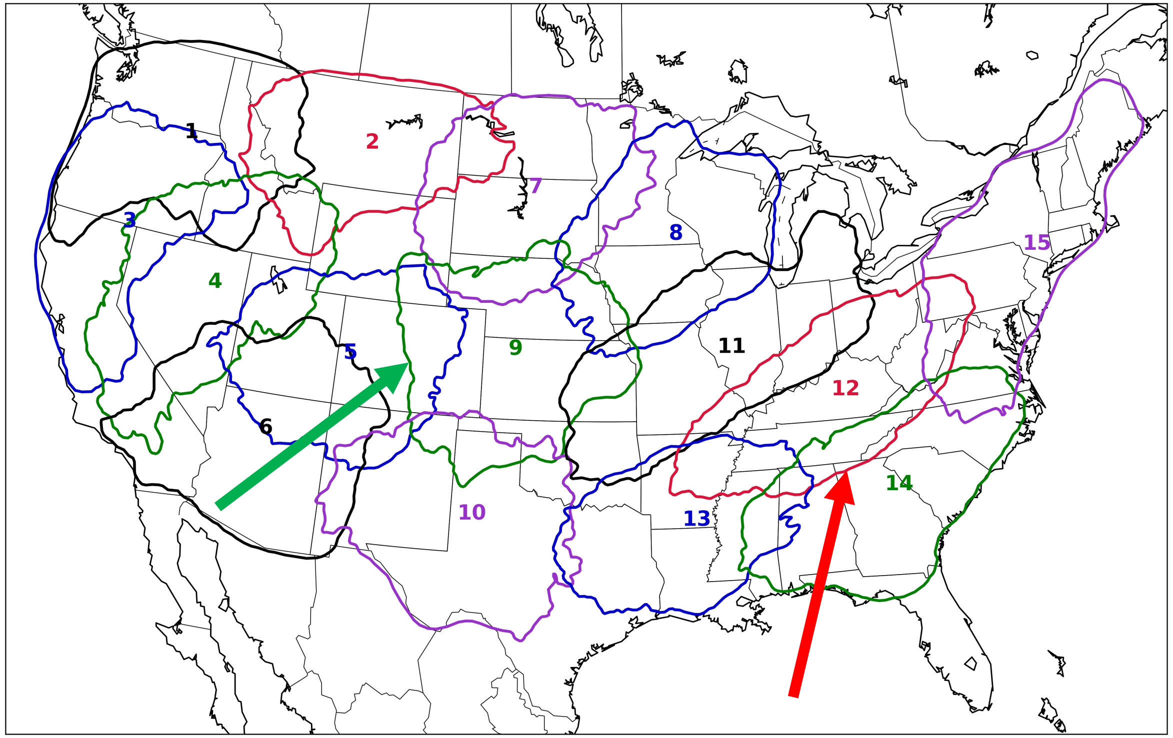


Figure 9 from Dickinson et al. (2021)

Online table of events can be found at http://pres2ip.com/extreme-event-tables

Downloadable database available at https://github.com/tydickinson29/PRES2iPpy/tree/master/pres2ippy/databases

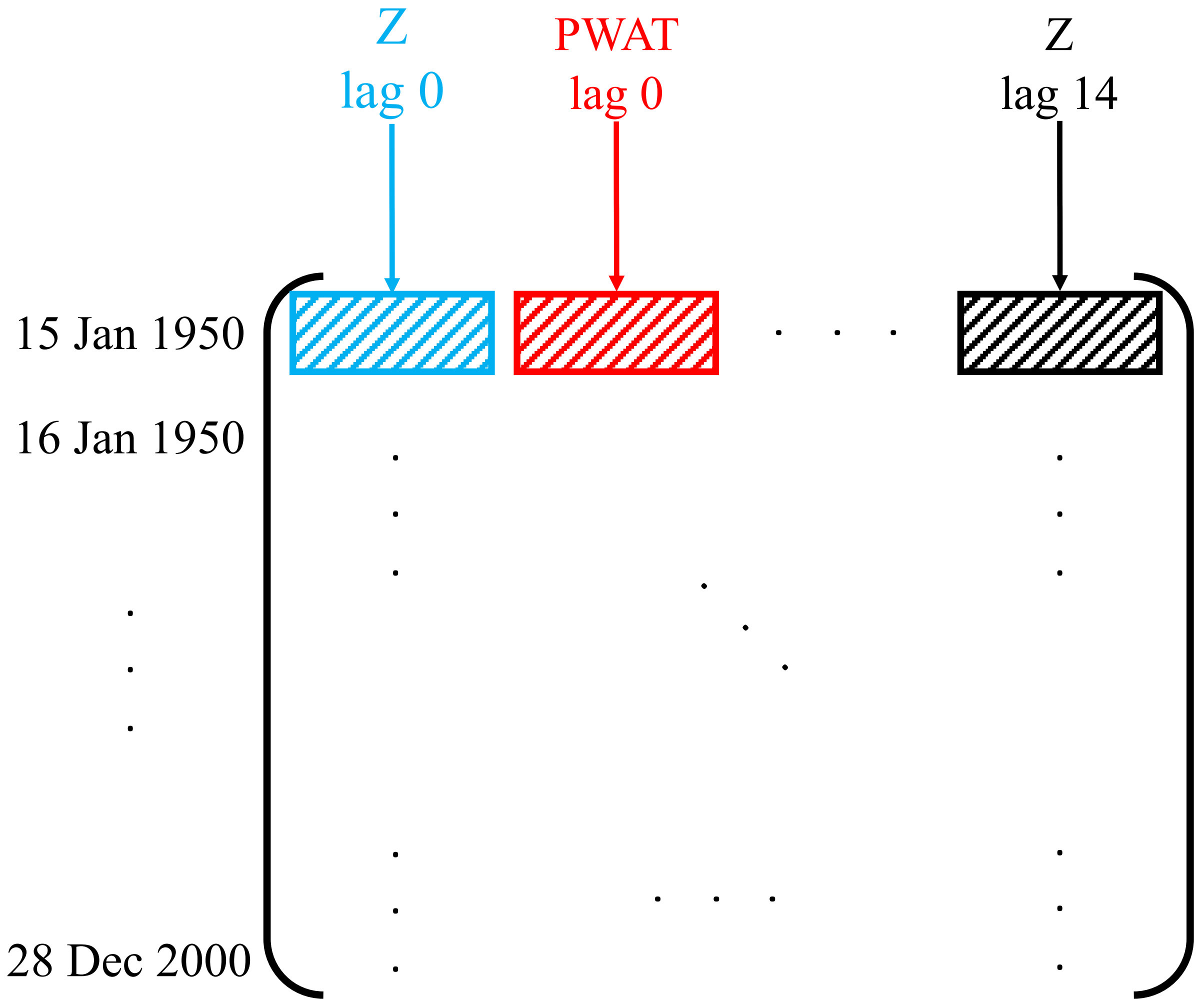
Data

- ECMWF ERA-5 reanalysis (Hersbach et al. 2020); daily data on 1.5° lat/lon grid.
  - Standardize via 1981-2010 climatology.
- Predictors (7-day centered running mean):
  - Geopotential height
    - Averaged in [850, 300] hPa column
  - Precipitable water
  - Zonal, meridional wind components
    - Averaged in [850, 300] hPa column
  - Sea-level pressure

Lags 7,14

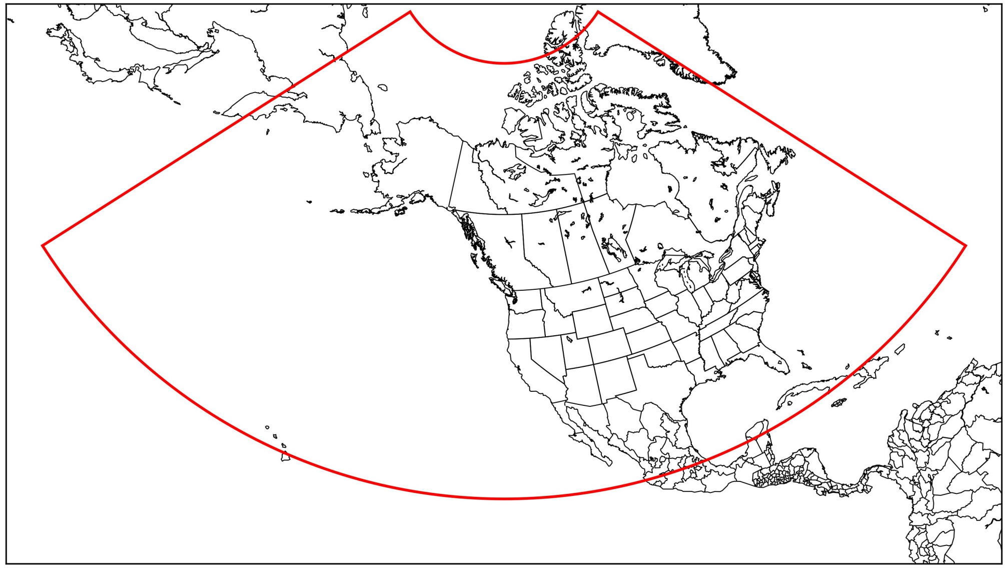
Lag 0

- Predictand: 1 if day in extreme event, 0 otherwise.



Workflow

- Domain: [20°, 80°] N; [160°, 310°] E
- RF hyperparameters held constant:
  - Gini impurity criterion
  - Number of considered predictors at each split: square root of total predictors
  - Label weight: inversely proportional to frequency
- Optimized RF hyperparameters:
  - Number of trees: [100, 1000]
  - Minimum samples to split leaf: [1, 4, 16, 32, 64, 128, 256, 500]
  - 10-fold cross validation



$$CSI = \frac{A}{A + B + C}$$

		Forecast	
		Yes	No
Observed	Yes	A	B
	No	C	D

Optimized Models and Accuracies

Region	Trees	Split Leaf Samples	Training CSI	Testing CSI
Southern Plains	1000	256	0.475	0.0924
Ohio River Valley	100	64	0.819	0.0408

Southern Plains

Training

		Forecast	
		Yes	No
Observed	Yes	270	1
	No	297	18030

Testing

		Forecast	
		Yes	No
Observed	Yes	22	50
	No	166	6680

Ohio River Valley

Training

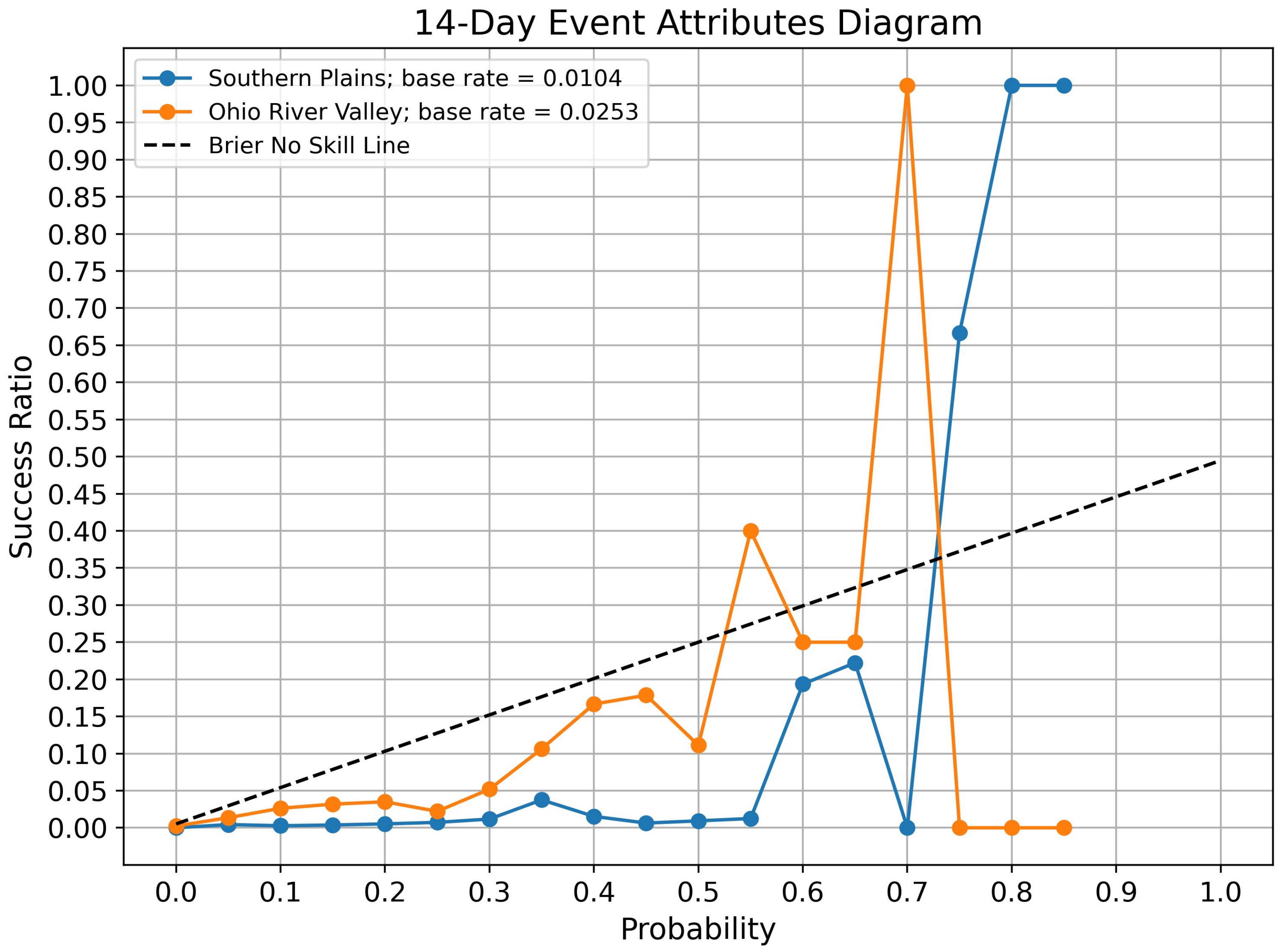
		Forecast	
		Yes	No
Observed	Yes	322	1
	No	70	18205

Testing

		Forecast	
		Yes	No
Observed	Yes	8	167
	No	21	6722

Optimized Models and Accuracies

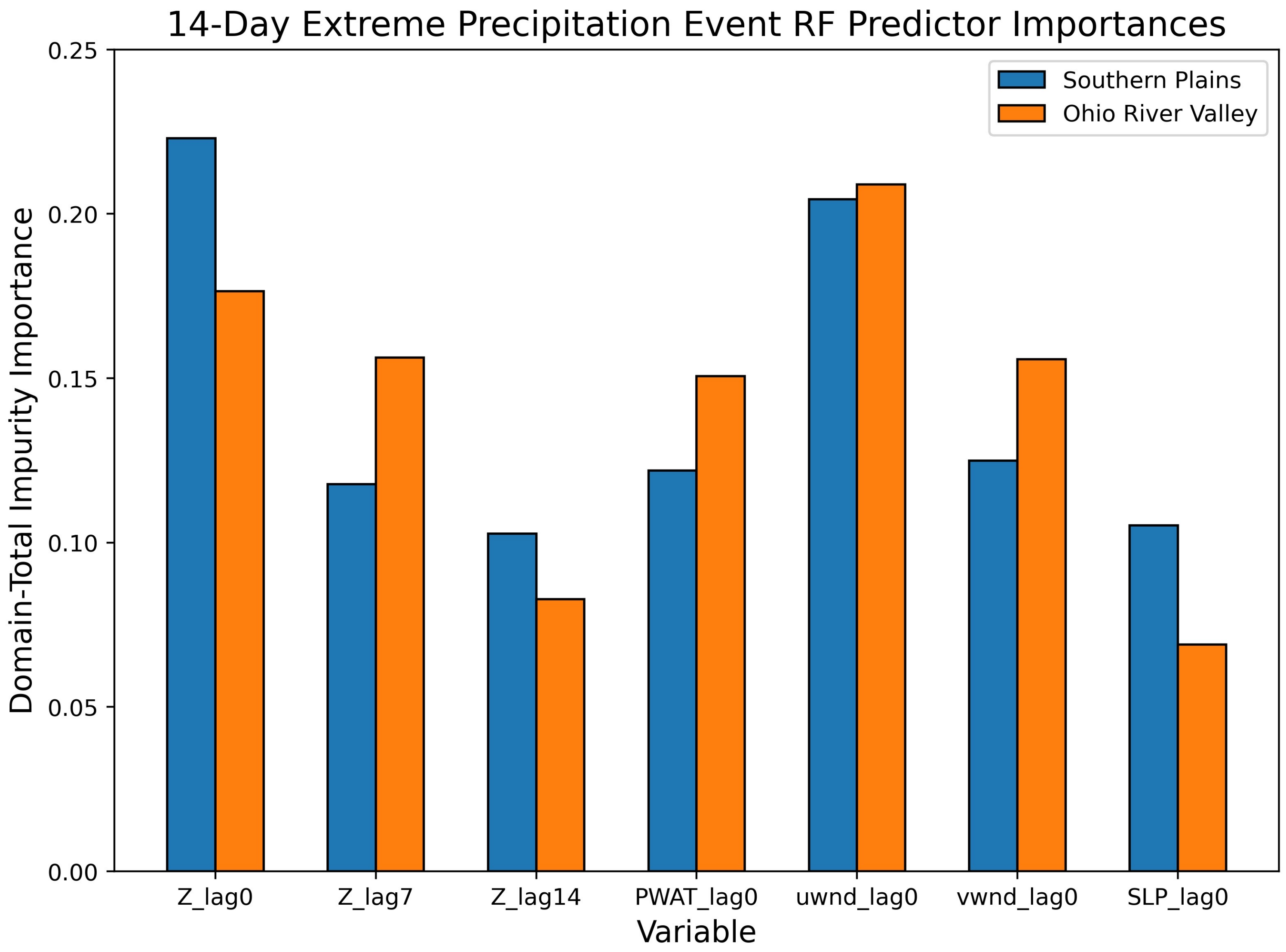
- Models struggle to match forecast probability to true probability
  - **Source** of error differs for two regions



PRES²iP

Gini Impurity Importance

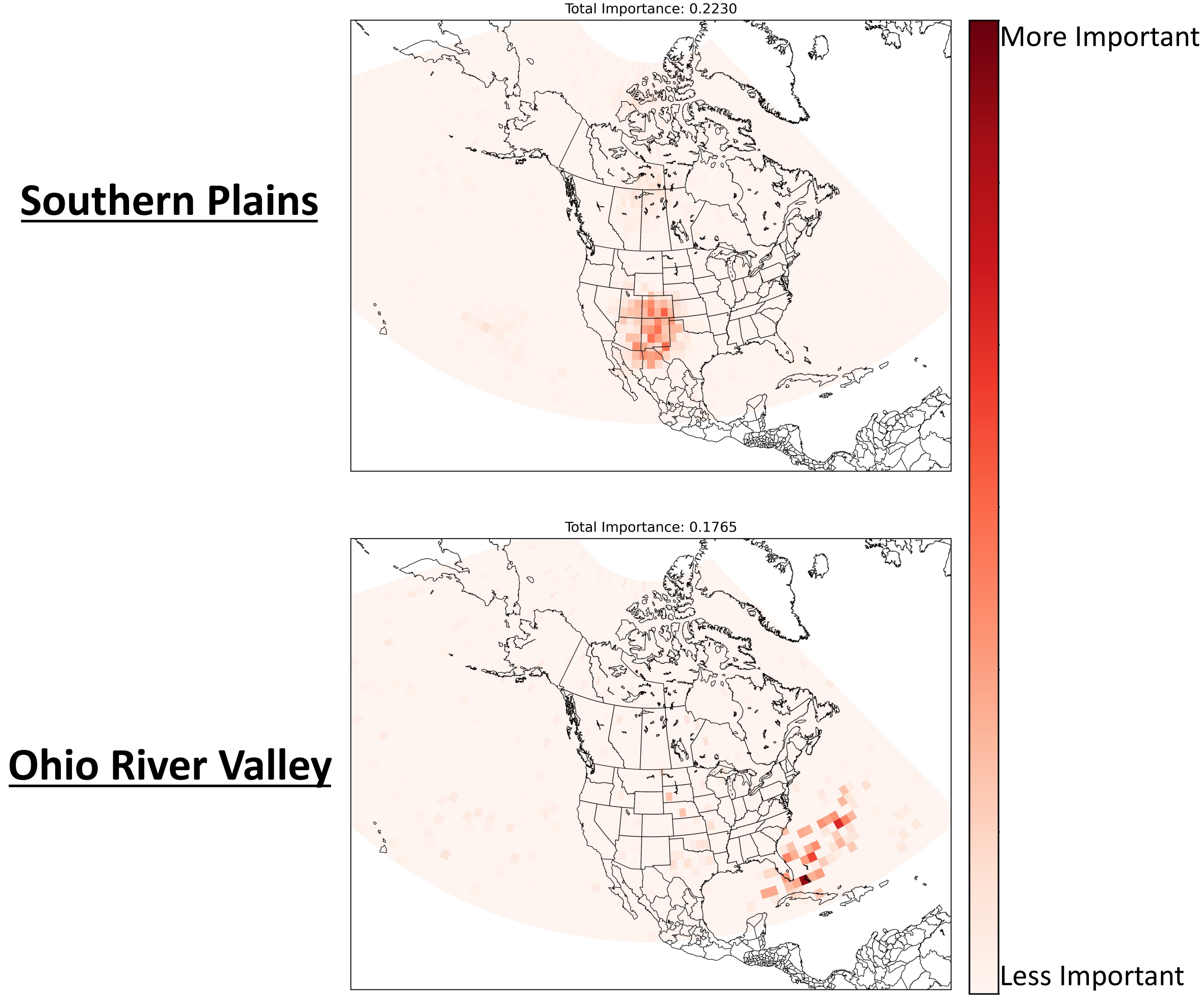
- Importance of predictors calculated via their effectiveness in splitting samples higher in tree.
- Lag 0 geopotential height, zonal wind component most important.

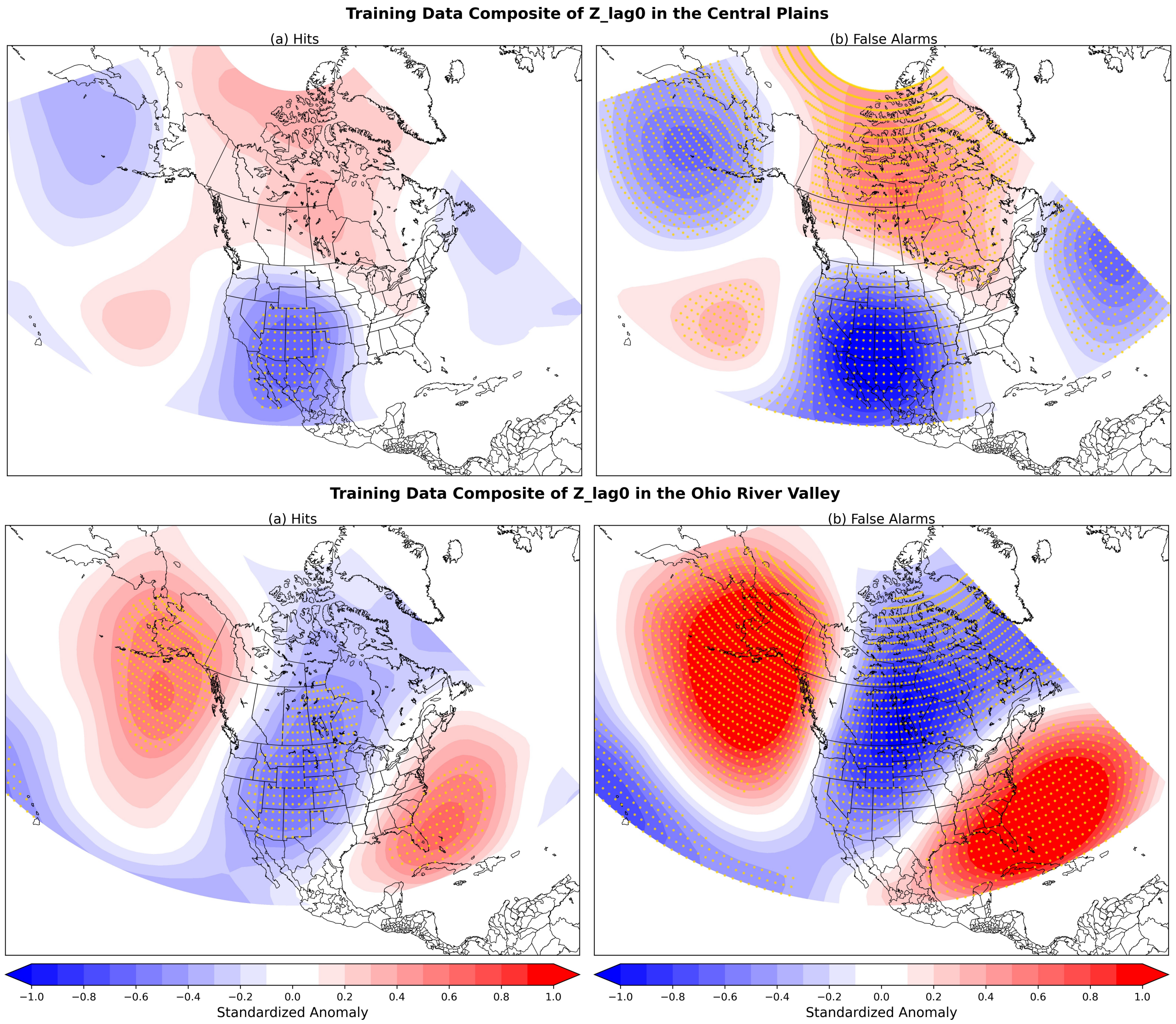


PRES²iP

Geopotential Height Importance

Gini Importance





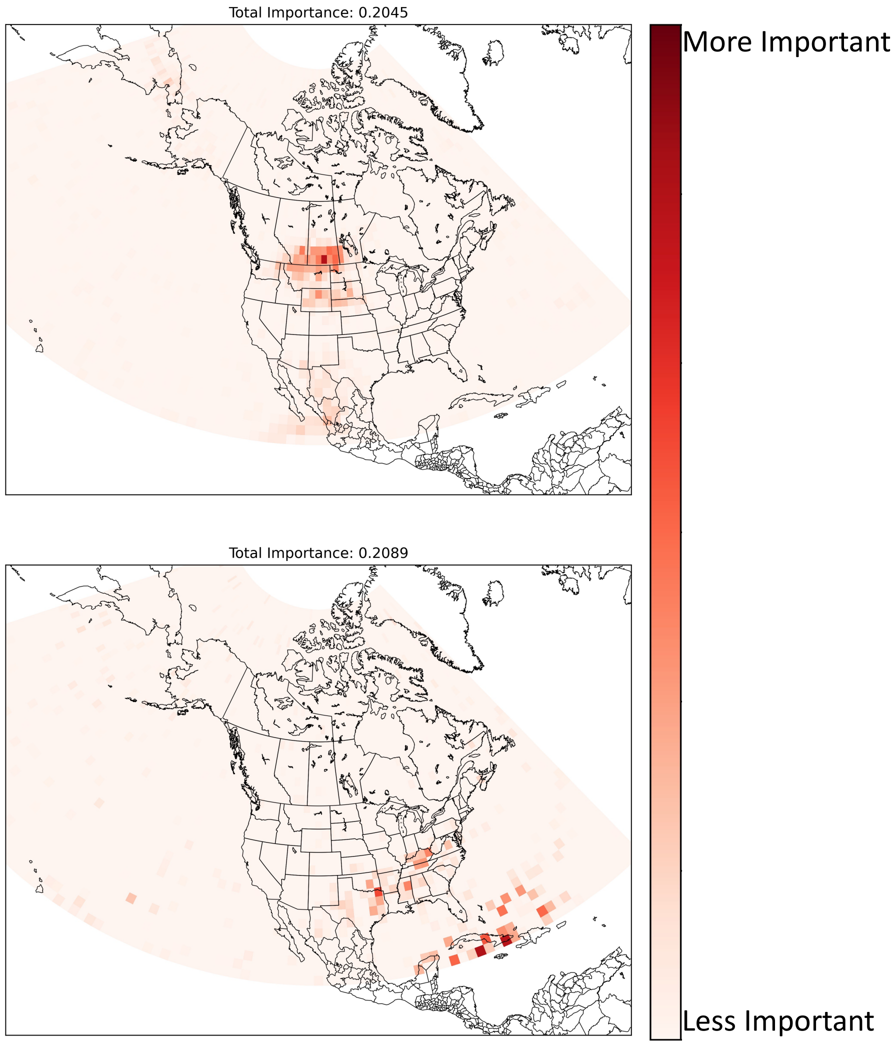
PRES²iP

Zonal Wind Importance

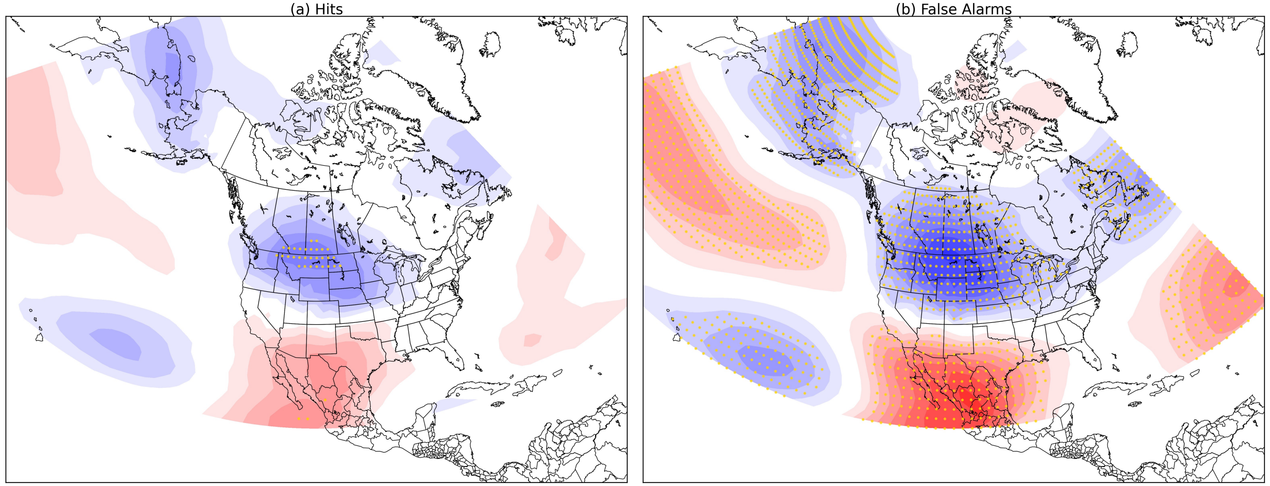
Gini Importance

Southern Plains

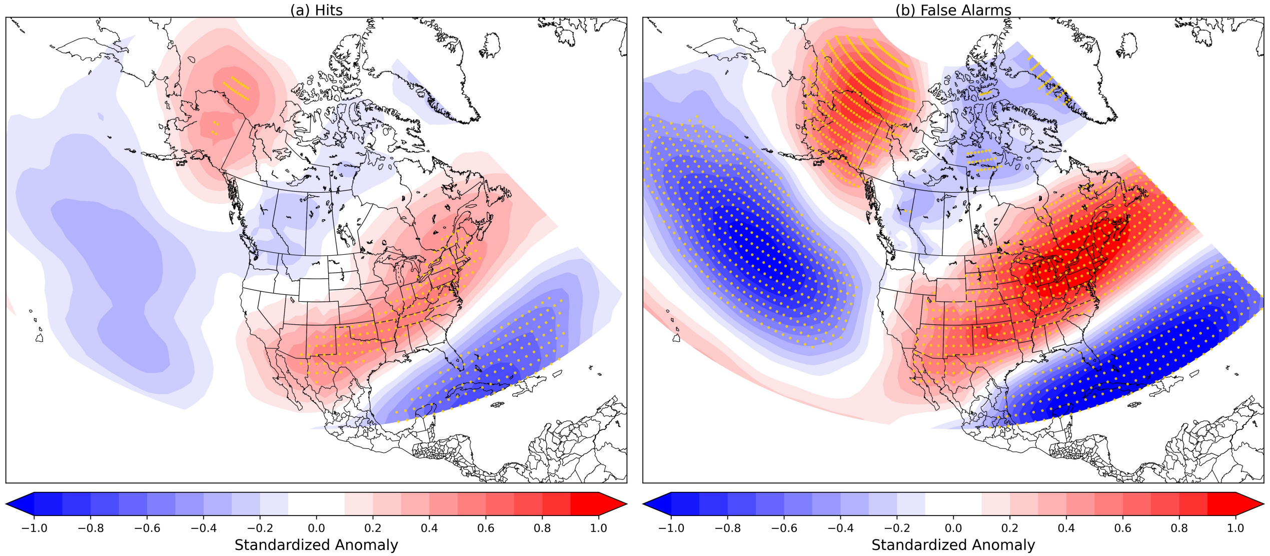
Ohio River Valley



Training Data Composite of uwnd_lag0 in the Central Plains



Training Data Composite of uwnd_lag0 in the Ohio River Valley



PRES²iP

Conclusions and Future Work

- Developed RFs currently have minimal skill in differentiating 14-day event and non-event days.
- Geopotential height, zonal wind component most important in both regions (e.g., Jennrich et al. 2020).
  - ➢Continue physical analysis to choose lags that will be most predictive.
- Future experiments:
  - Use detrended anomalies.
  - Add further predictors, e.g., omega.
  - Employ PCA to reduce dimensionality, further increase signal-to-noise ratio.
  - Weight days adjacent to event.
- **Other PRES²iP presentations: Melanie Schroers, Olivia VanBuskirk, and Devin McAfee.**

Acknowledgements

- Coauthors: Dr. Jason Furtado and Dr. Michael Richman
- Melanie Schroers and the rest of the amazing PRES^{2}iP team!
- Ollie Millin
- NSF PREEVENTS program, Grant ICER-1663840

Works Cited

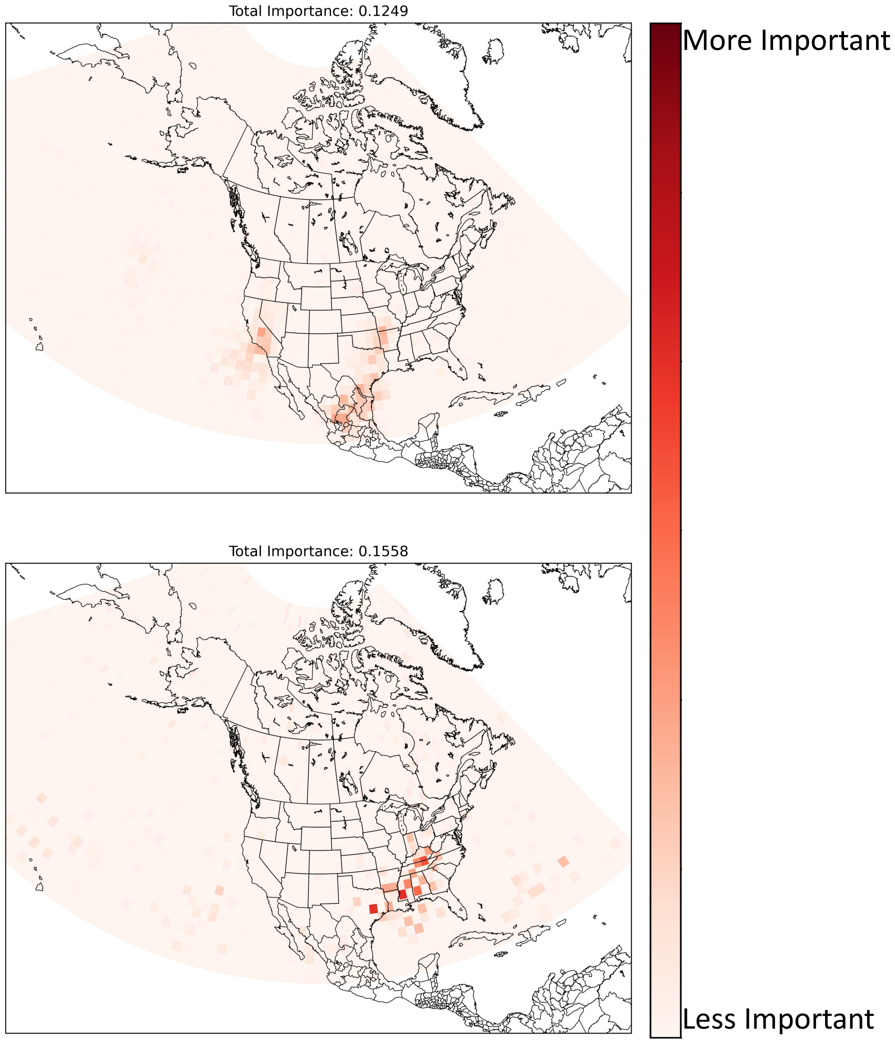
- Barlow, W., and Coauthors, 2019: North American extreme precipitation events and related large-scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics*, doi: 10.1007/s00382-019-04958-z.
- Dickinson, T. D., M. B. Richman, and J. C. Furtado, 2021: Subseasonal to Seasonal Extreme Precipitation Events in the Contiguous United States: Generation of a Database and Climatology. *J. Climate*, **34**, 7571-7586, doi:10.1175/JCLI-D-20-0580.1.
- Herman, G. R., and R. S. Schumacher, 2018: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev.*, **146**, 1571-1600, doi: 10.1175/MWR-D-17-0250.1.
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, doi:10.1002/qj.3803.
- Jennrich, G. C., J. C. Furtado, J. B. Basara, and E. R. Martin, 2020: Synoptic Characteristics of 14-Day Extreme Precipitation Events Across the United States. *J. Climate*, **33**, 6423–6440, doi:10.1175/JCLI-D-19-0563.1.
- Novak, D., K. Brill, M. Eckert, R. Oravec, B. Sullivan, R. Bann, F. Barthold, and M. Bodner, 2011: Quantifying extreme rainfall threats at the Hydrometeorological Prediction Center. *24th Conf. on Weather and Forecasting/20th Conf. on NWP*, Seattle, WA, **Amer. Meteor. Soc., 14A.4.**, [Available online at https://ams.confex.com/ams/91Annual/webprogram/Paper179714.html].

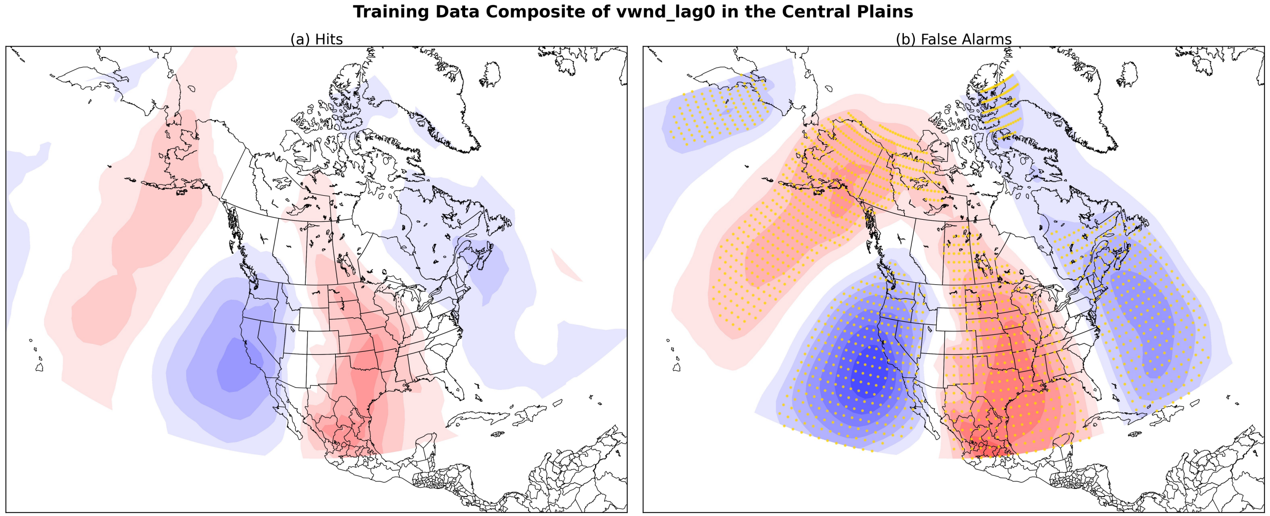
Meridional Wind Importance

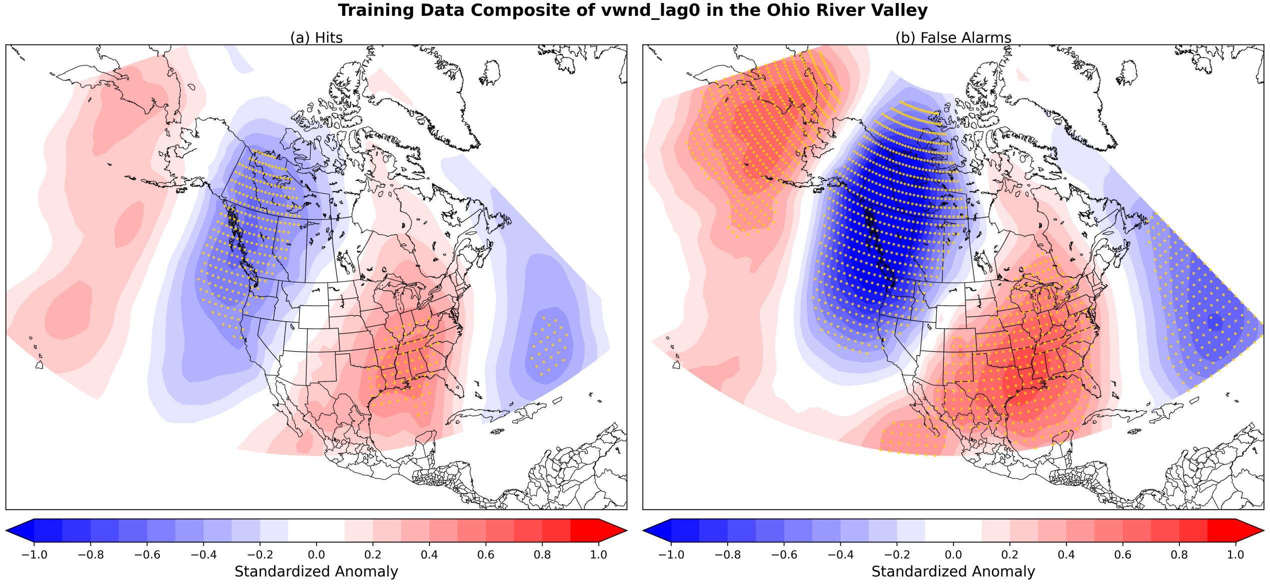
Gini Importance

Southern Plains

Ohio River Valley





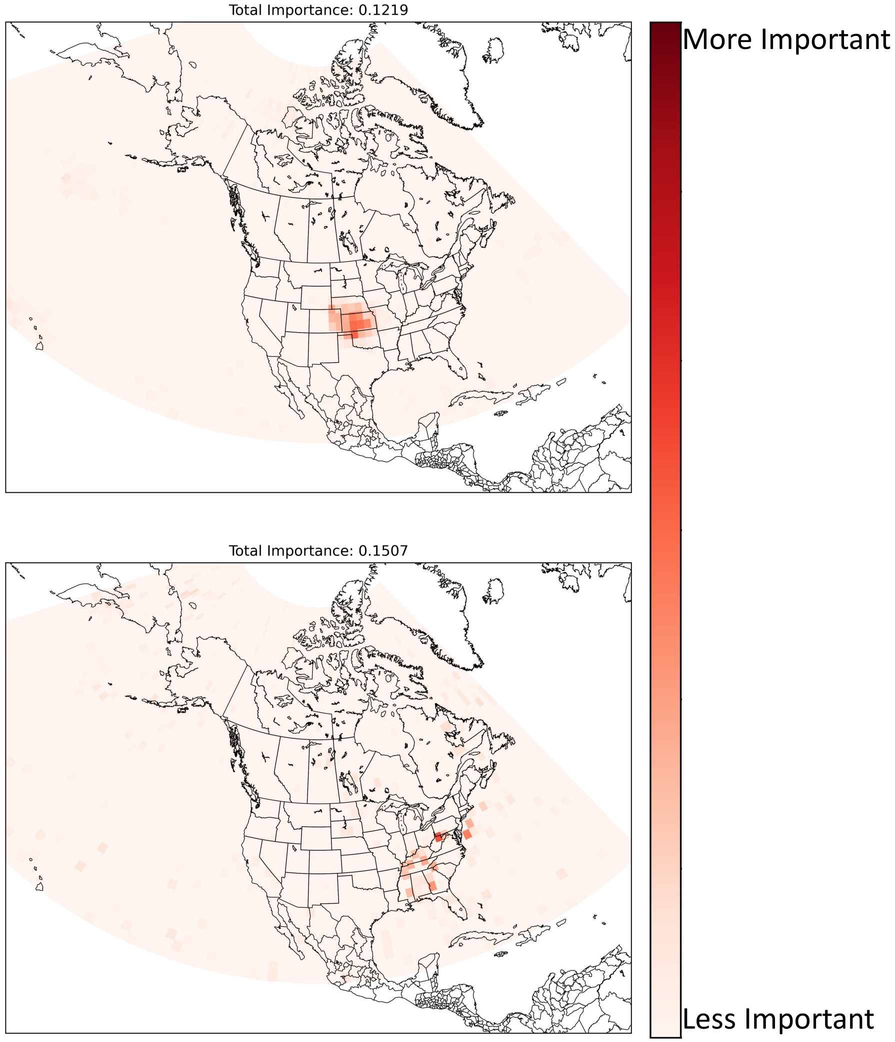


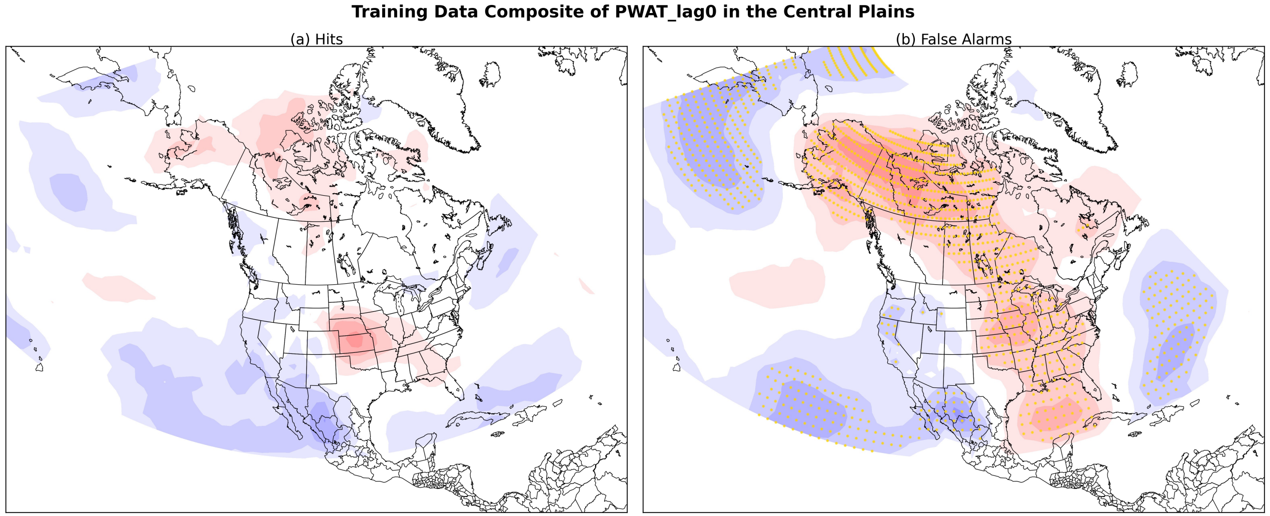
Precipitable Water Importance

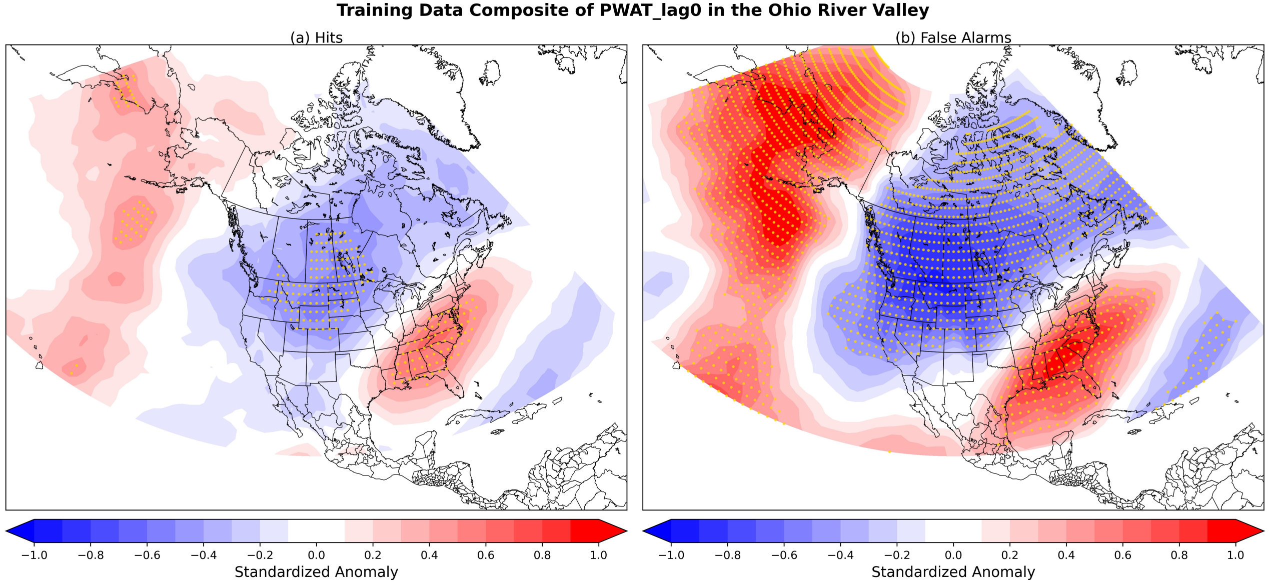
Gini Importance

Southern Plains

Ohio River Valley





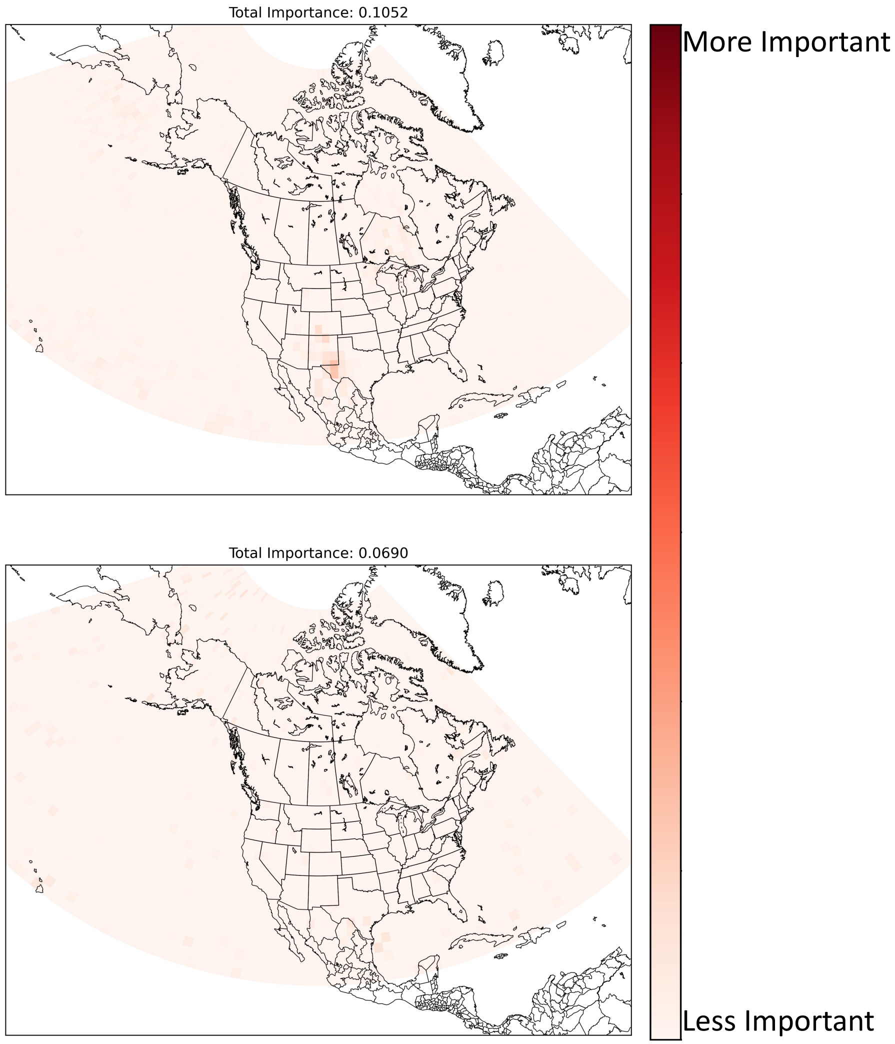


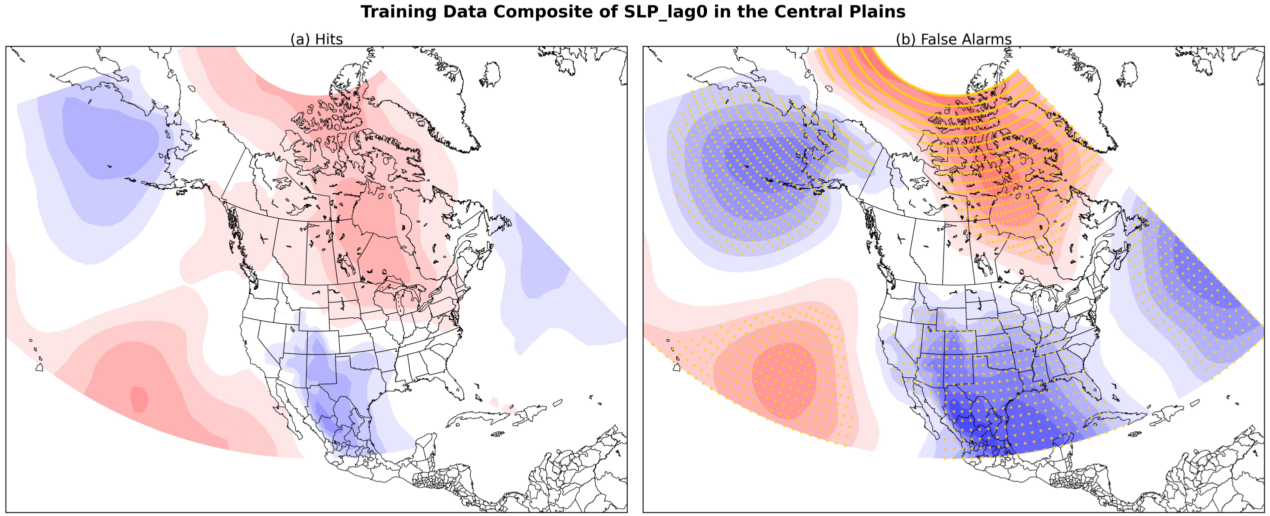
Sea-Level Pressure Importance

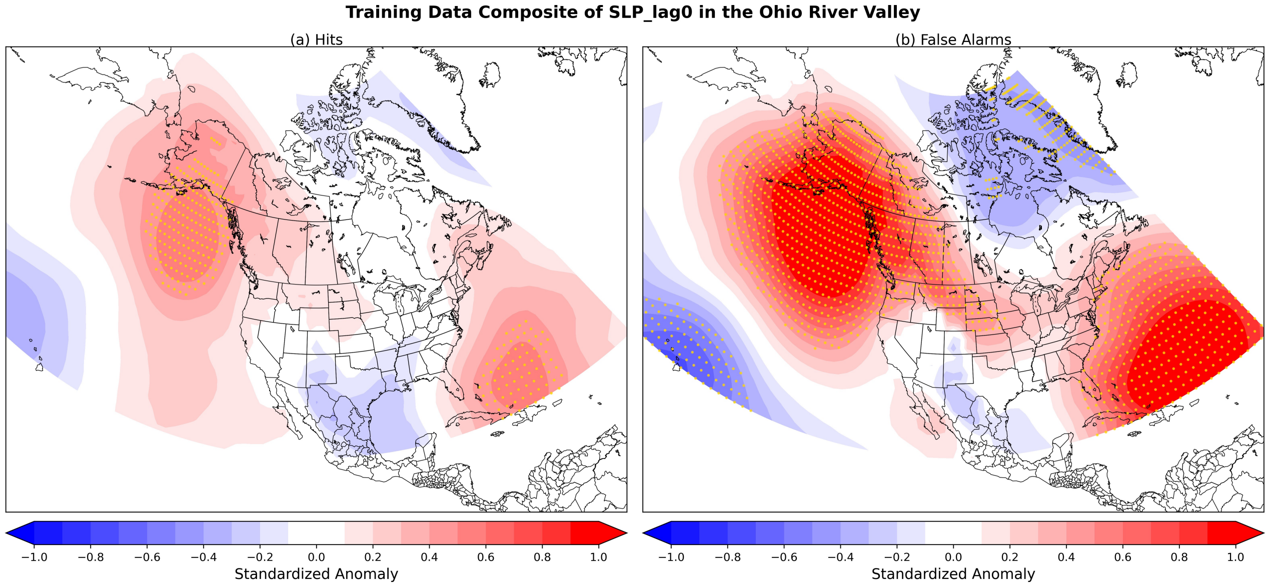
Gini Importance

Southern Plains

Ohio River Valley







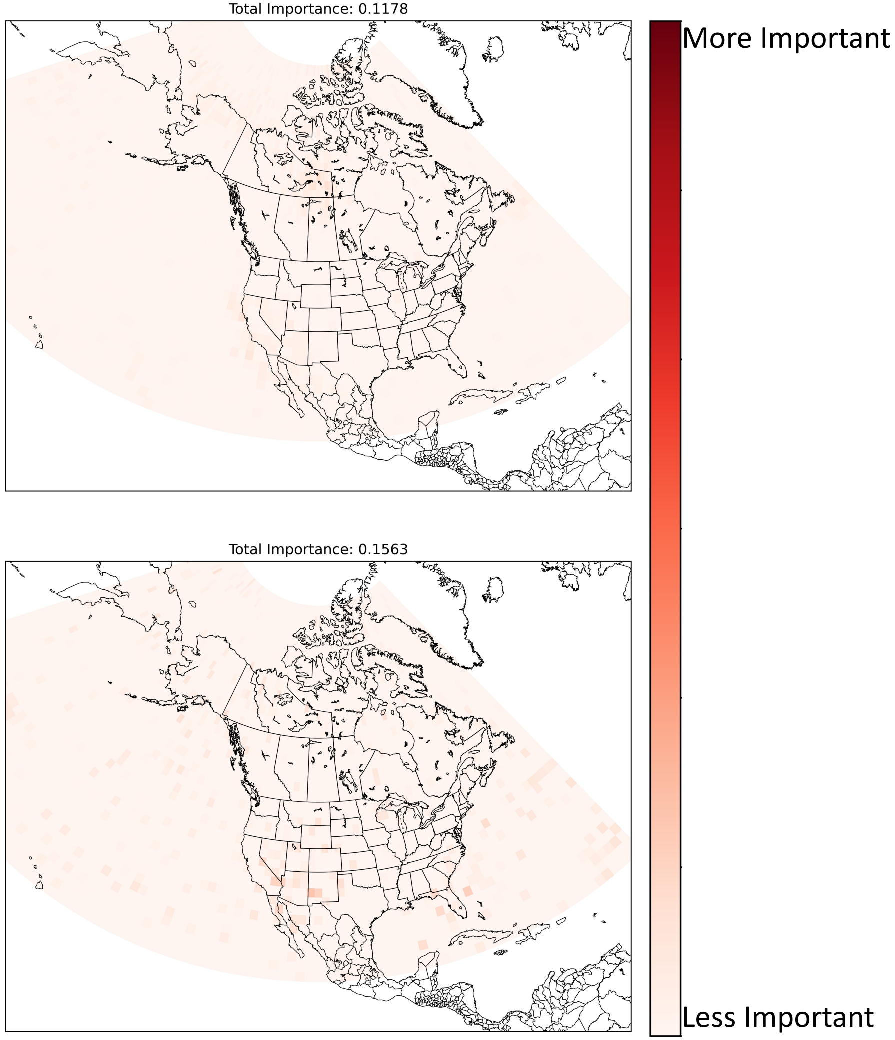
PRES²iP

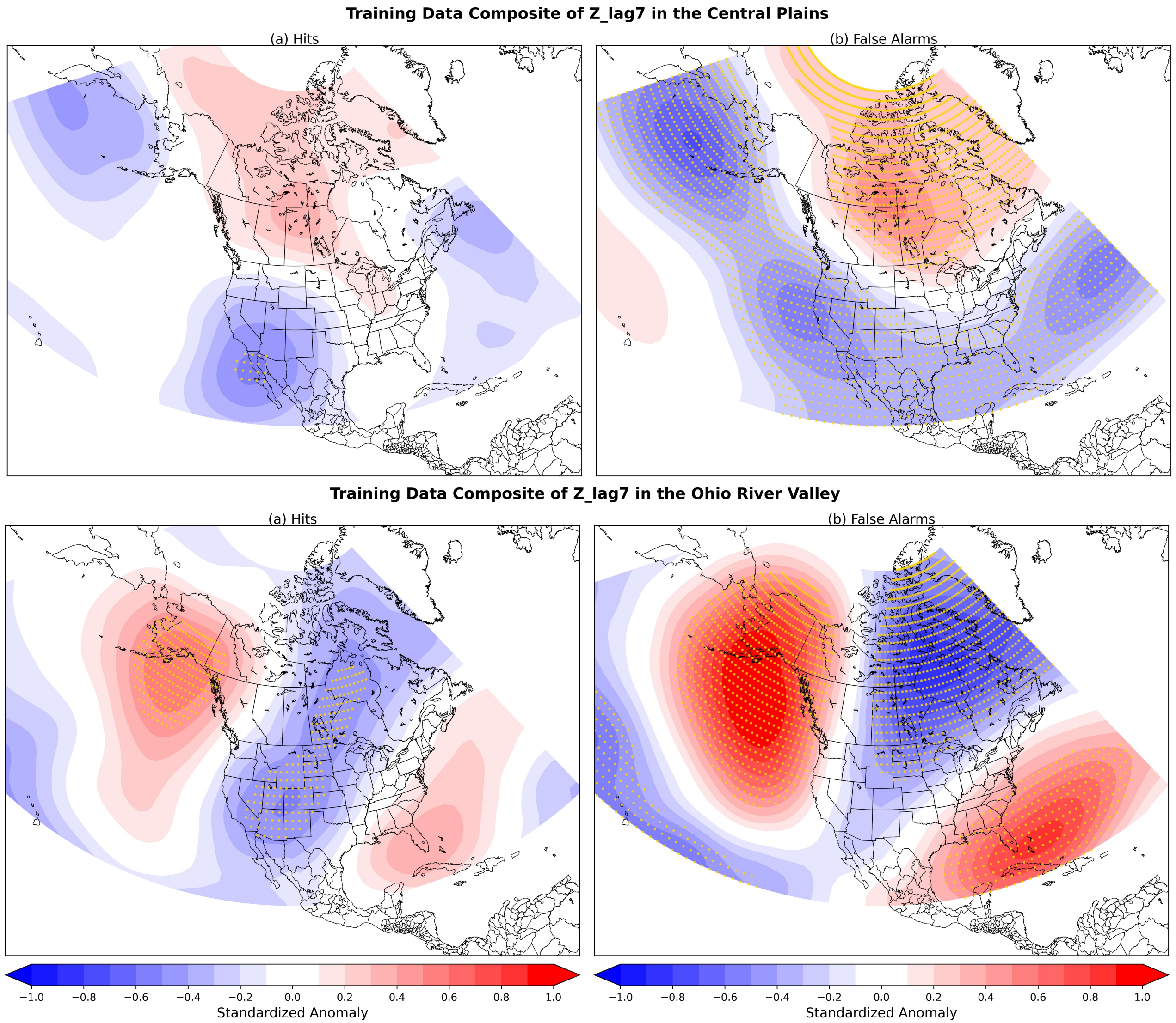
Geopotential Height Lag 7 Importance

Gini Importance

Southern Plains

Ohio River Valley





PRES²iP

Geopotential Height Lag 14 Importance

Gini Importance

Southern Plains

Ohio River Valley