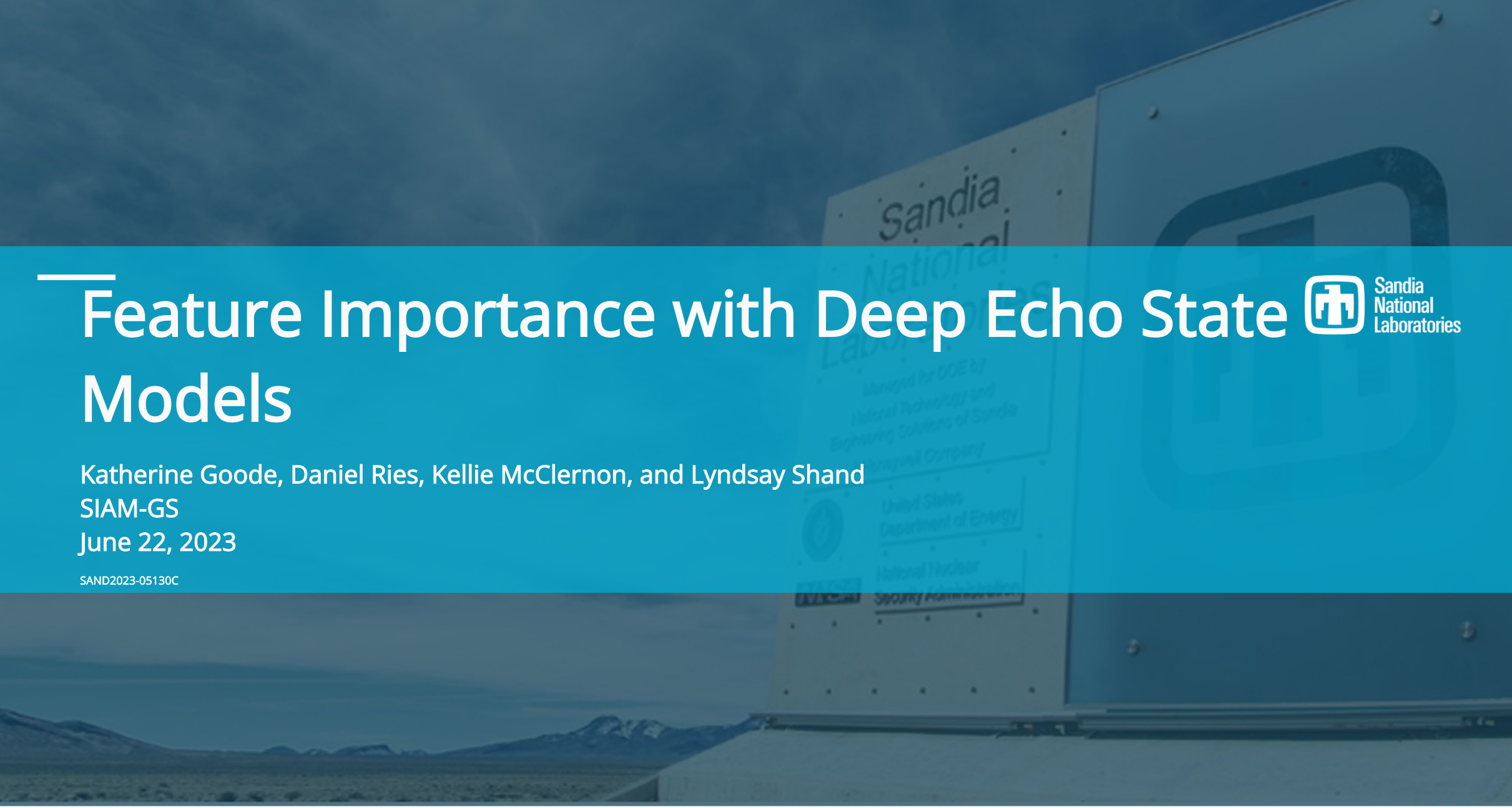


Feature Importance with Deep Echo State Models

Katherine Goode, Daniel Ries, Kellie McClernon, and Lyndsay Shand
SIAM-GS
June 22, 2023

SAND2023-05130C

Outline

- Motivation: Climate Interventions and Pathways
- Approach: Echo State Networks and Feature Importance
- Climate Application: Mount Pinatubo
- Conclusions and Future Work

Motivation

Climate Interventions and Pathways

Climate Interventions

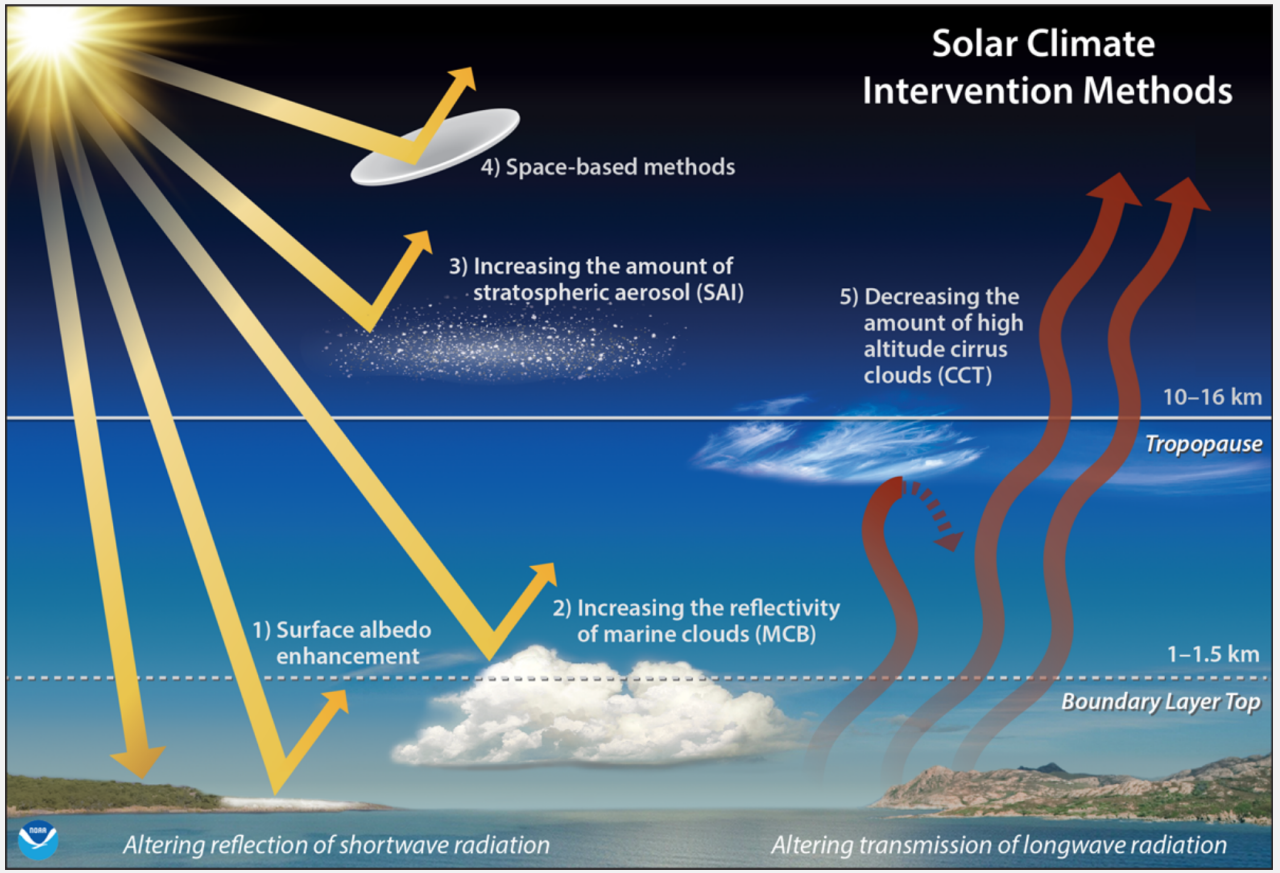


Image source: https://eos.org/science-updates/improving-models-for-solar-climate-intervention-research

Threat of climate change has led to...

- Proposed possible interventions
  - Stratospheric aerosol injections
  - Marine cloud brightening
  - Cirrus cloud thinning
  - etc.

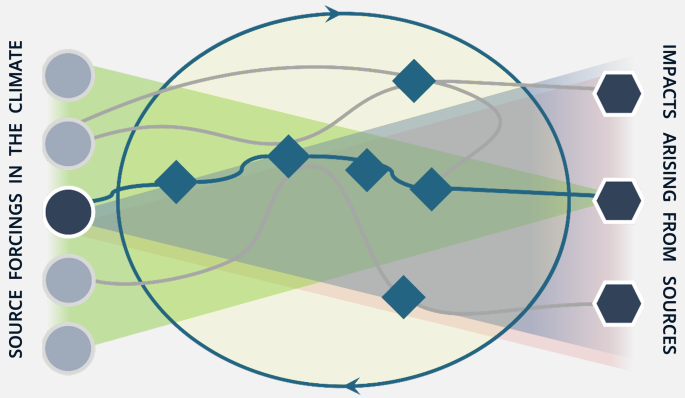
What are the downstream effects of such mitigation strategies?

Our Objective

Develop algorithms to characterize (i.e., quantify) relationships between climate variables related to a climate event (with observed data)

Climate Pathway (associated with a climate event)

- Source variable
- Intermediate variables
- Impact variable



Example

- Mount Pinatubo eruption in 1991
- Released 18-19 Tg of sulfur dioxide
- Proxy for anthropogenic stratospheric aerosol injection

Mount Pintabuo Example Pathway

Sulfur dioxide (Source)

- Injection of sulfur dioxide (18-19 Tg) into atmosphere [1]

⬇

Aerosol optical depth (AOD) (Intermediate)

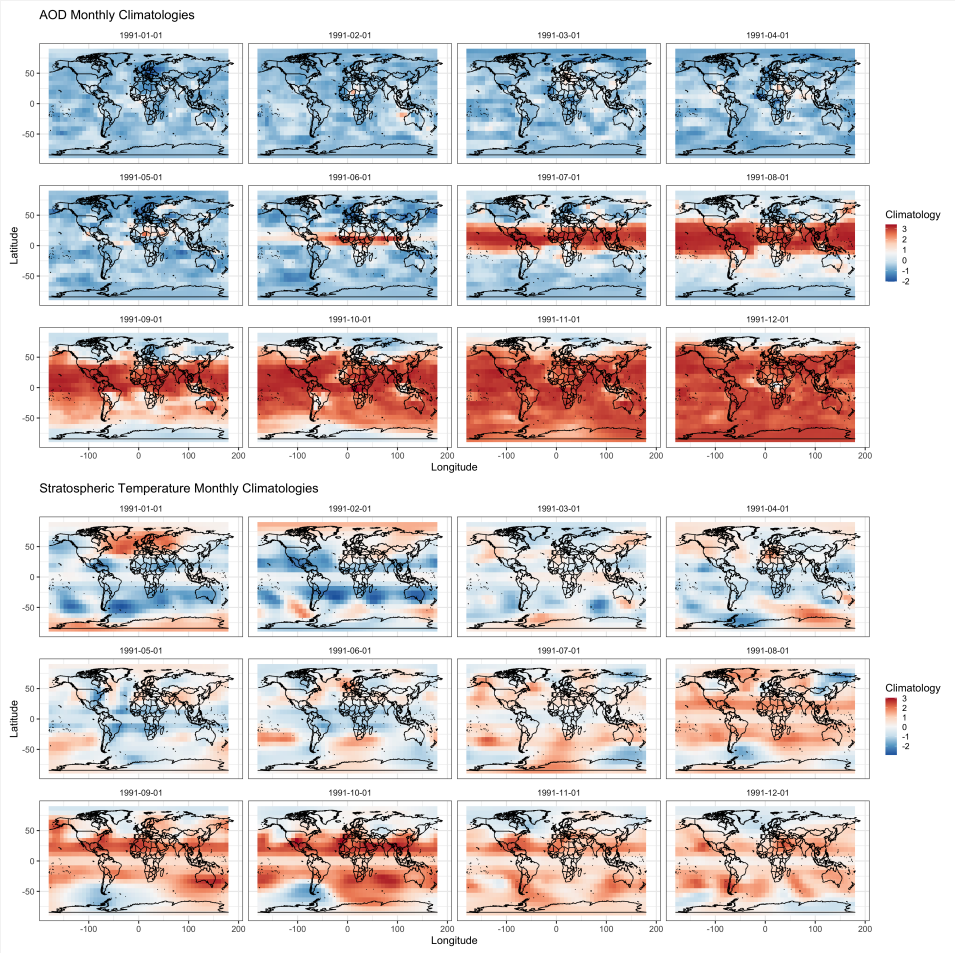
- Vertically integrated measure of aerosols in air from surface to stratosphere [2]

⬇

Stratospheric temperature (Impact)

- Temperatures at pressure levels of 30-50 mb rose 2.5-3.5 degrees centigrade compared to 20-year mean [3]

Figure generated using Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA- 2) data [4]



Our Approach

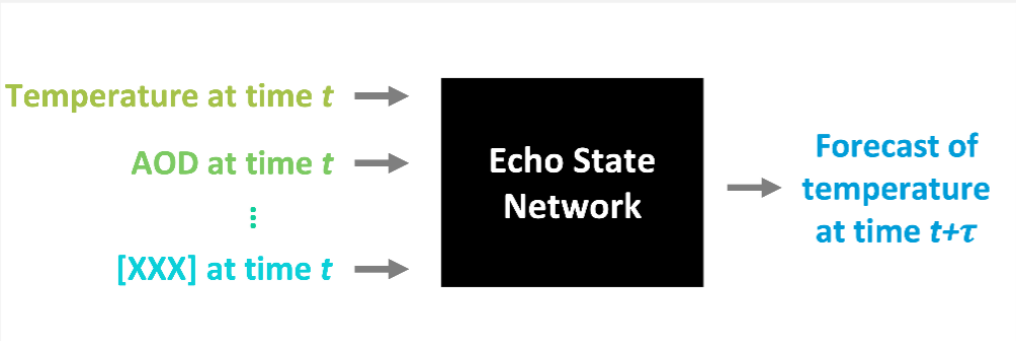
Use machine learning…

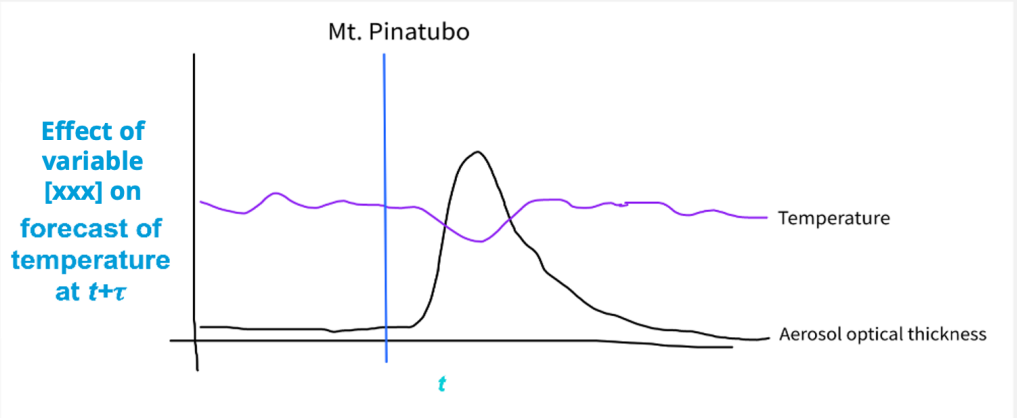
Step 1: Model pathway variables with echo state network

- Allow complex machine learning model to capture complex pathway variable relationships

Step 2: Understand pathways via explainability

- Apply explainability techniques (feature importance) to understand pathways captured by model





Approach

Echo State Networks and Feature Importance

Echo-State Networks

- Machine learning model for temporal data
  - Sibling to recurrent neural network (RNN)
- Computationally efficient
  - Compared to RNNs and spatio-temporal statistical models
  - ESN reservoir parameters randomly sampled instead of estimated
- Previous work demonstrated use of ESN for long-term spatio-temporal forecasting (McDermott and Wikle [5])

Output stage (ridge regression):

$$\mathbf{y}_t = \mathbf{V}\mathbf{h}_t + \boldsymbol{\epsilon}_t$$

$$\epsilon_\mathbf{t} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$$

Hidden stage (nonlinear stochastic transformation):

$$\mathbf{h}_t = g_h\left(\frac{\nu}{|\lambda_w|}\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\tilde{\mathbf{x}}_{t-\tau}\right)$$

$$\tilde{\mathbf{x}}_{t-\tau} = \left[\mathbf{x}'_{t-\tau}, \mathbf{x}'_{t-\tau-\tau^*}, \ldots, \mathbf{x}'_{t-\tau-m\tau^*}\right]'$$

Note: Only parameters estimated are in $\mathbf{V}$.

Echo-State Networks: Spatio-Temporal Context

Spatio-temporal processes at spatial locations $\{\mathbf{s}_i \in \mathcal{D} \subset \mathbb{R}^2; i = 1, \ldots, N\}$ over times $t = 1, \ldots, T$...

Impact variable (e.g., stratospheric temperature):

$$\mathbf{Z}_{Y,t} = (Z_{Y,t}(\mathbf{s}_1), Z_{Y,t}(\mathbf{s}_2), \ldots, Z_{Y,t}(\mathbf{s}_N))'$$

Source/intermediate variables (e.g., aerosol optical depth):

$$\mathbf{Z}_{k,t} = (Z_{k,t}(\mathbf{s}_1), Z_{k,t}(\mathbf{s}_2), \ldots, Z_{k,t}(\mathbf{s}_N))'$$

$$\text{for } k = 1, \ldots, K$$

Stage	Formula	Description
Data stage (outputs)	$\mathbf{Z}_{Y,t} \approx \mathbf{\Phi}_Y \mathbf{y}_t$	Basis function decomposition (e.g., PCA)
Output stage	$\mathbf{y}_t = \mathbf{V}\mathbf{h}_t + \boldsymbol{\epsilon}_t$	Ridge regression
Hidden stage	$\mathbf{h}_t = g_h\left(\frac{\nu}{\lvert\lambda_w\rvert}\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\tilde{\mathbf{x}}_{t-\tau}\right)$	Nonlinear stochastic transformation
Data stage (inputs)	$\mathbf{Z}_{k,t} \approx \mathbf{\Phi}_k \mathbf{x}_{k,t} \quad \text{where } \mathbf{x}_t = [\mathbf{x}'_{1,t}, \ldots, \mathbf{x}'_{K,t}]'$	Basis function decomposition (e.g., PCA)

Feature Importance

Goal

- Feature importance aims to quantify effect of input variable on a model's predictions

Background

- Permutation feature importance [6]
- Pixel absence affect with ESNs [7]
- Temporal permutation feature importance [8]

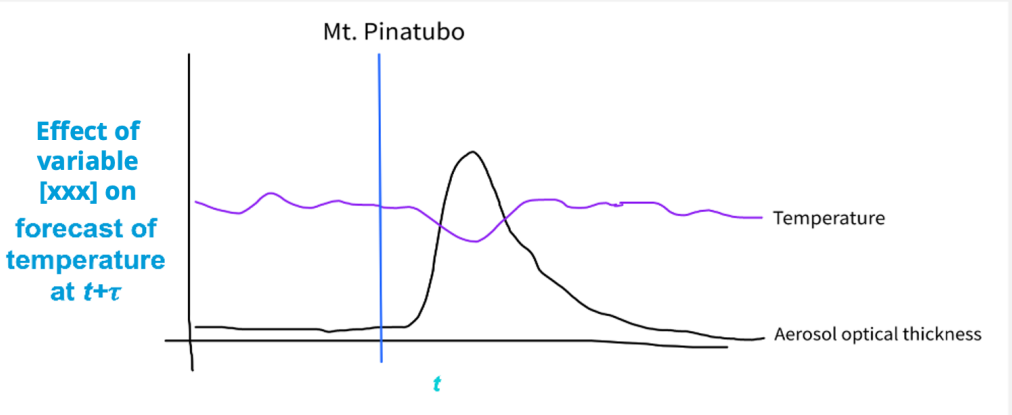
Our Work

- Adapt for ESNs in context of spatio-temporal data

In particular...

Compute feature importance on trained ESN model for:

- input variable over block of times
- on forecasts of response variable at a time



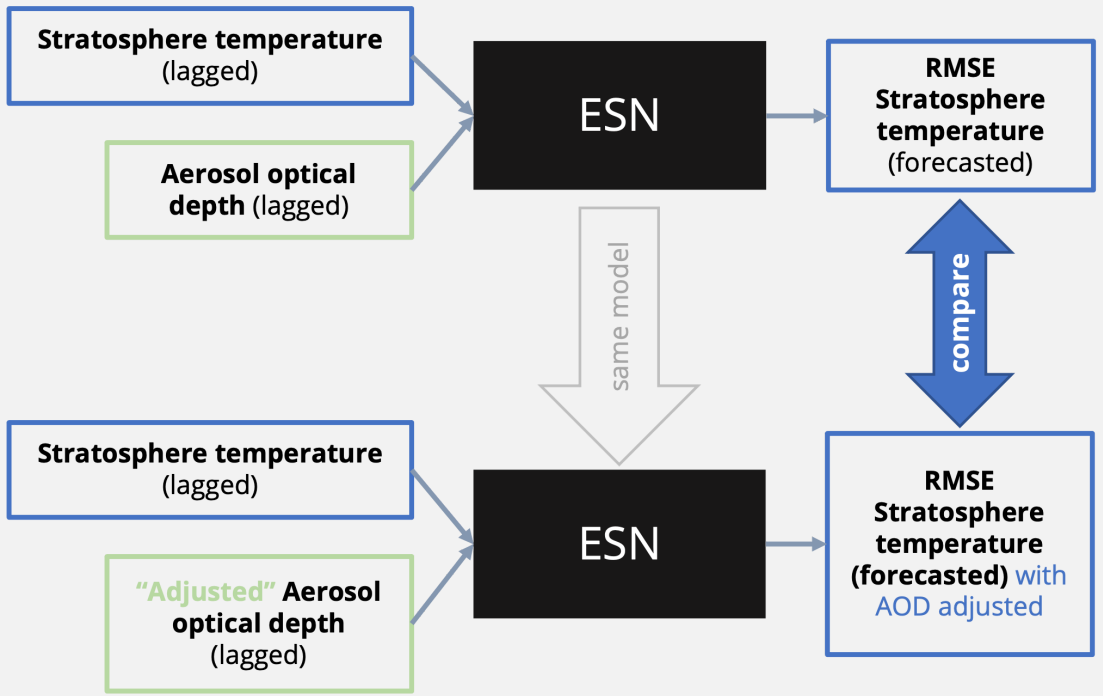
Feature Importance for ESNs

Concept

- "Adjust" inputs at times(s) of interest
- Quantify effect on model performance
- Large decrease in performance indicates important time(s)

Two Approaches: "Adjust" inputs by either

- Permute values: spatio-temporal permutation feature importance (stPFI)
- Set values to zero: spatio-temporal zeroed feature importance (stZFI)



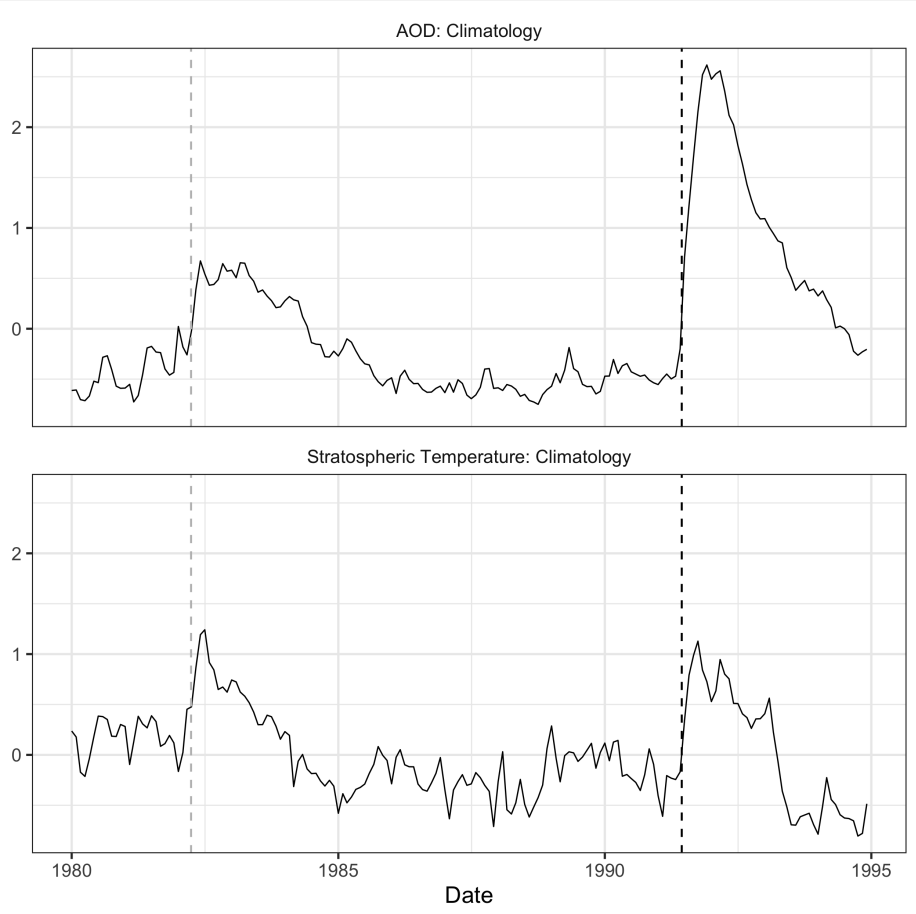
Climate Application

Mount Pinatubo

Mount Pinatubo Example: Data

Data

- Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA- 2)
- Training Years: 1980 to 1995
    - Includes eruptions of Mount Pinatubo (1991) and El Chichón (1982)
- Time Interval: Monthly
- Latitudes: -86 to 86 degrees



Mount Pinatubo Example: Model

ESN Output

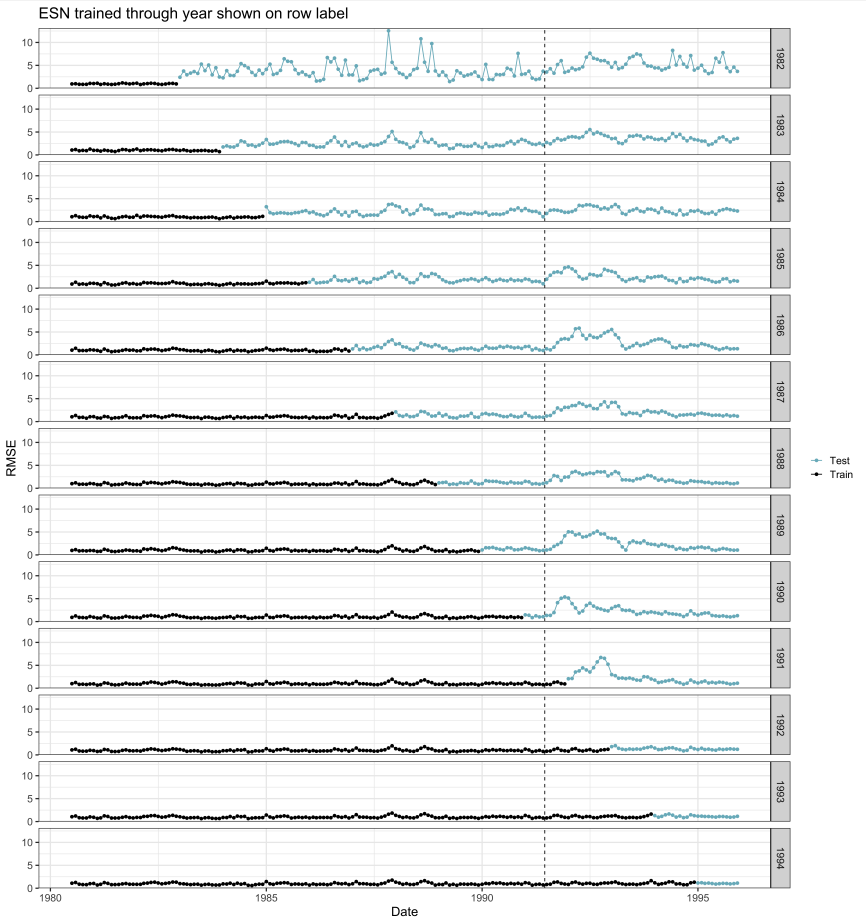
- Stratospheric Temperature (50mb)

ESN Inputs

- Lagged Stratospheric Temperature (50mb; one month lag)
- Lagged AOD (one month lag)

Preprocessing (all variables)

- Climatologies
- Principal components (first 5)



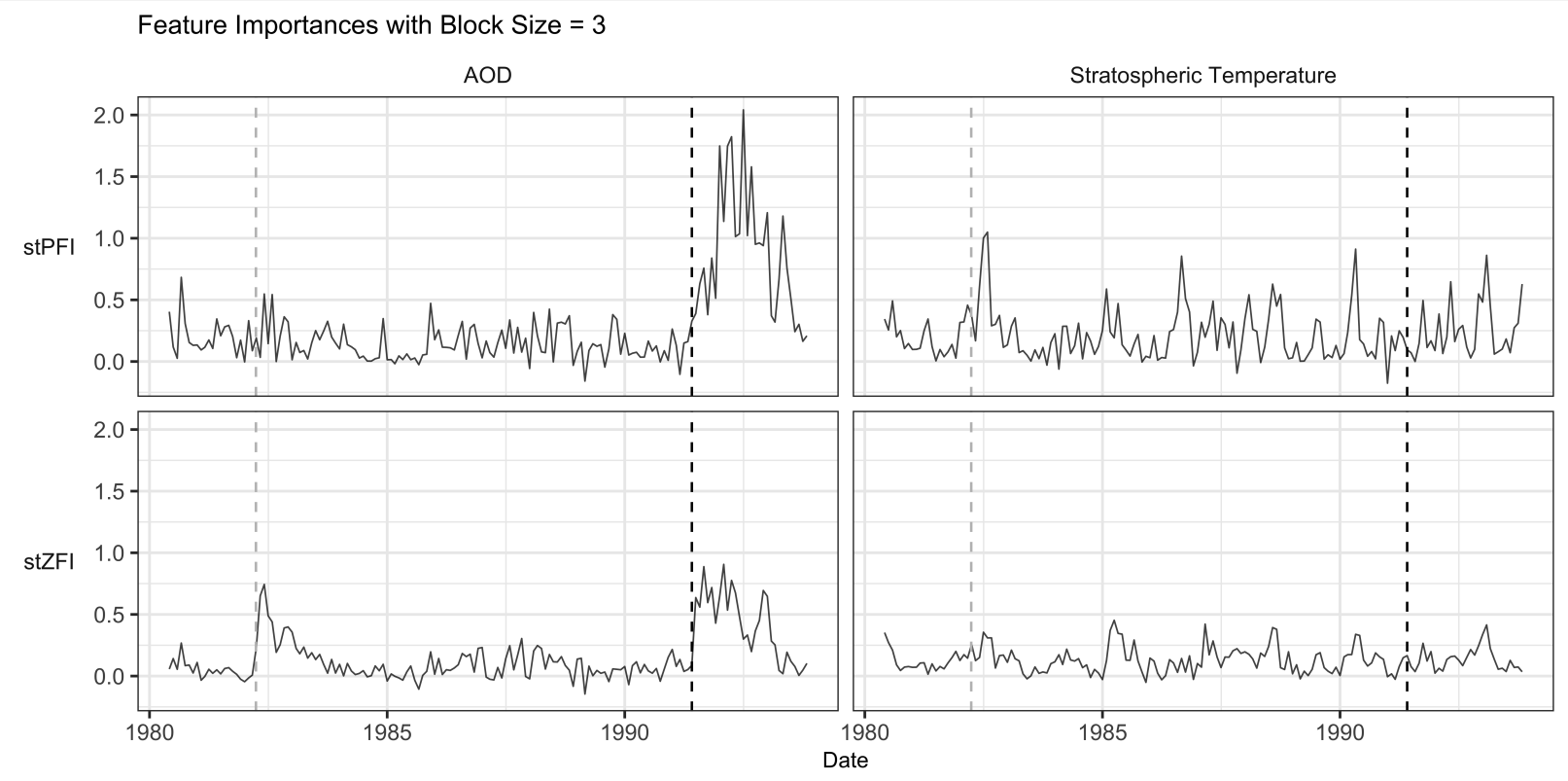
Mount Pinatubo Example: Feature Importance

Key Point

Peak of importance for AOD (and lack of peak of importance for lagged stratospheric temperatures), provides evidence that volcanic eruption impact on temperature can be traced through AOD

FI Metric

Weighted RMSE (weighted by cosine of the latitude)



Conclusions and Future Work

Summary and Conclusions

Summary

- Interested in quantifying relationships between climate variables associated with pathway of climate event
- Motivated by increasing possibility of climate interventions
- Our machine learning approach:
    - Use ESN to model variable relationships
    - Understand variable relationships using proposed spatio-temporal feature importance

Conclusion

- Approach provided evidence of AOD being an intermediate variable in Mount Pinatubo climate pathway affecting stratospheric temperature

Future (Current) Work

ESN extensions

- Addition of multiple layers
- ESN ensembles
- Bayesian ESNs

Spatio-temporal feature importance

- Implement proposed retraining technique [9] to lessen detection of spurious relationships
- Adapt to visualize on spatial scale
- Comparison to other newly proposed explainability techniques for ESNs (layer-wise relevance propagation) [10]

Mount Pinatubo application

- Inclusion of additional pathway variables (e.g., SO2, radiative flux, surface temperature)
- Importance of grouped variables

References

[1] S. Guo, G. J. Bluth, W. I. Rose, et al. "Re-evaluation of SO$*2$ release of the 15 June 1991 Pinatubo eruption using ultraviolet and infrared satellite sensors". In: _Geochemistry, Geophysics, Geosystems* 5 (4 2004), pp. 1-31. DOI: 10.1029/2003GC000654.

[2] M. Sato, J. E. Hansen, M. P. McCormick, et al. "Stratospheric aerosol optical depths, 1850-1990". In: *Journal of Geophysical Research: Atmospheres* 98.D12 (1993), pp. 22987-22994. DOI: https://doi.org/10.1029/93JD02553. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/93JD02553. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/93JD02553.

[3] K. Labitzke and M. McCormick. "Stratospheric temperature increases due to Pinatubo aerosols". In: *Geophysical Research Letters* 19 (2 1992), pp. 207-210. DOI: 10.1029/91GL02940.

[4] R. Gelaro, W. McCarty, M. J. Suarez, et al. "The ModernEra Retrospective Analysis for Research and Applications, Version 2 (MERRA-2)". In: *Journal of Climate* 30 (14 2017), pp. 5419-5454. DOI: 10.1175/JCLI-D-16-0758.1.

[5] P. L. McDermott and C. K. Wikle. "Deep echo state networks with uncertainty quantification for spatio-temporal forecasting". In: *Environmetrics* 30.3 (2019). ISSN: 1180-4009. DOI: 10.1002/env.2553.

[6] A. Fisher, C. Rudin, and F. Dominici. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously". In: *Journal of Machine Learning Research*. 177 20 (2019), pp. 1-81. eprint: 1801.01489. URL: http://jmlr.org/papers/v20/18-760.html.

[7] A. B. Arrieta, S. Gil-Lopez, I. Laña, et al. "On the post-hoc explainability of deep echo state networks for time series forecasting, image and video classification". In: *Neural Computing and Applications* 34.13 (2022), pp. 10257-10277. ISSN: 0941-0643. DOI: 10.1007/s00521-021-06359-y.

[8] A. Sood and M. Craven. "Feature Importance Explanations for Temporal Black-Box Models". In: *arXiv* (2021). DOI: 10.48550/arxiv.2102.11934. eprint: 2102.11934.

[9] G. Hooker, L. Mentch, and S. Zhou. "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance". In: *Statistics and Computing* 31 (2021), pp. 1-16.

[10] M. Landt-Hayen, P. Kröger, M. Claus, et al. "Layer-Wise Relevance Propagation for Echo State Networks Applied to Earth System Variability". In: *Signal, Image Processing and Embedded Systems Trends*. Ed. by D. C. Wyld. Computer Science & Information Technology (CS & IT): Conference Proceedings 20. ARRAY(0x55588c8d8680), 2022, pp. 115-130. ISBN: 978-1-925953-80-0. DOI: doi:10.5121/csit.2022.122008. URL: https://doi.org/10.5121/csit.2022.122008.

Thank you

Katherine Goode

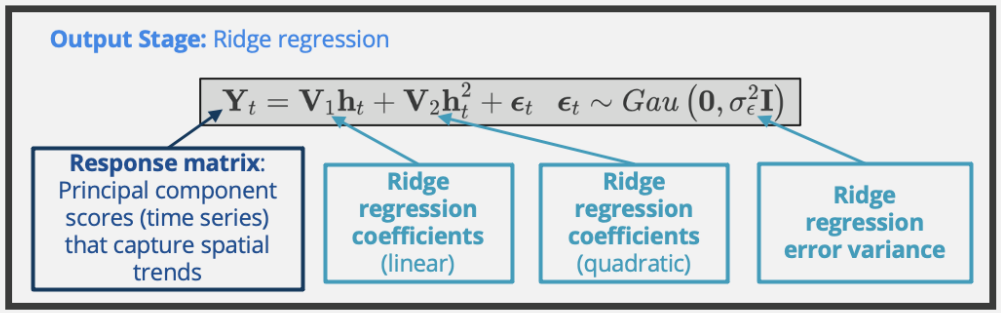
kjgoode@sandia.gov

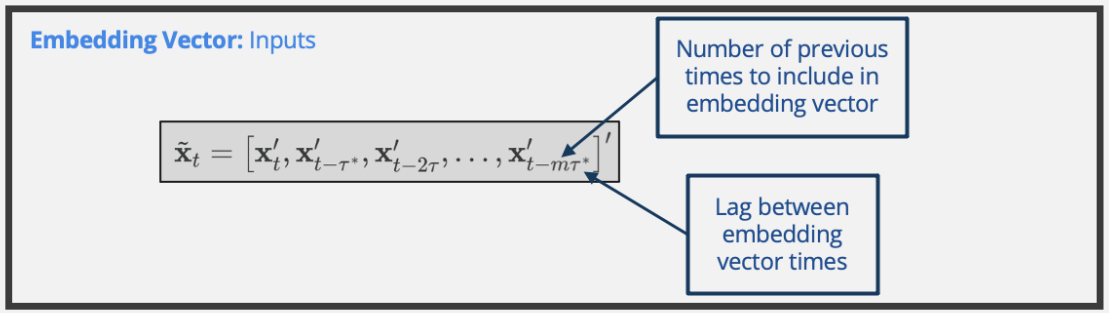
goodekat.github.io

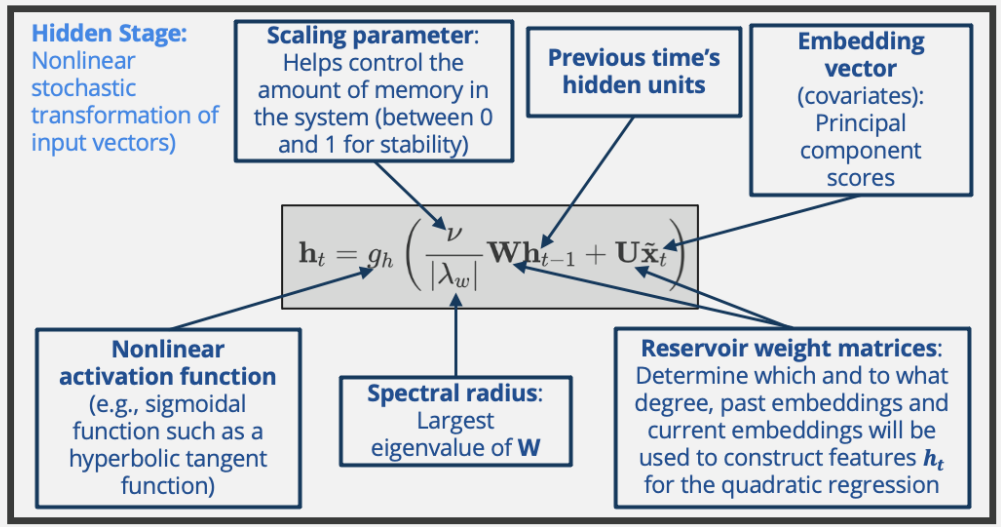
Back-Up Slides

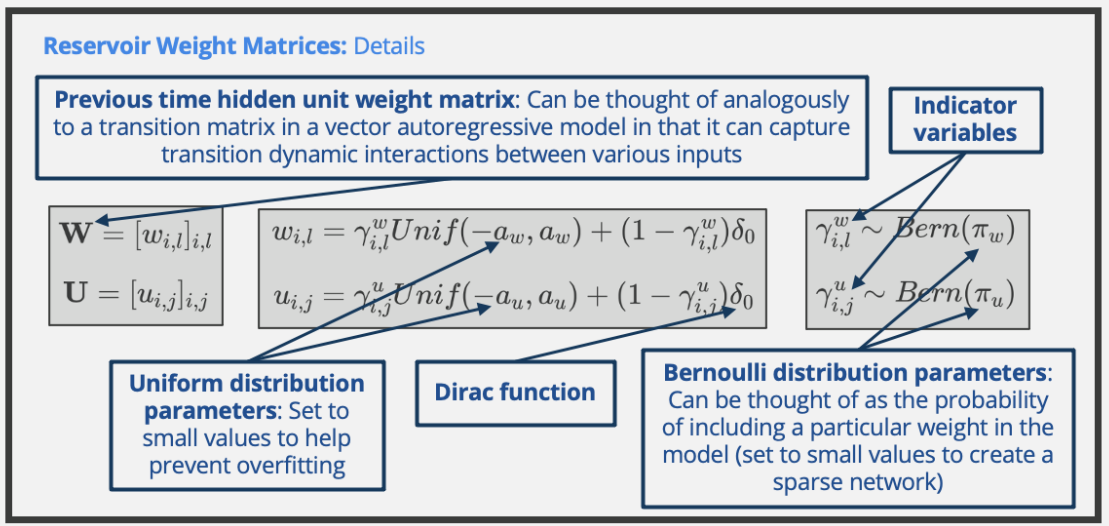
ESN Details

Quadratic Echo State Network



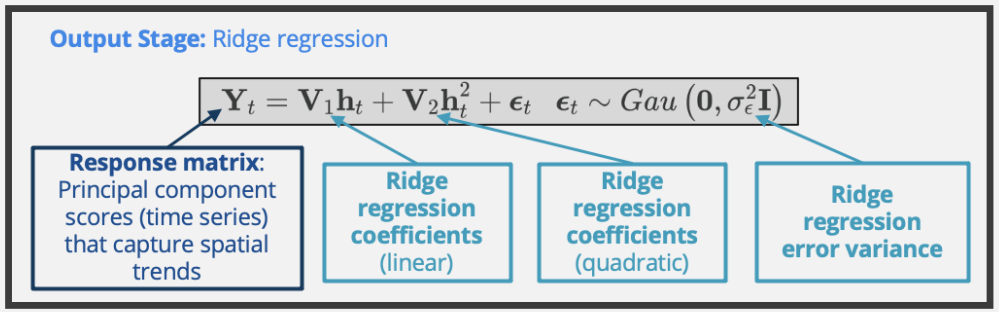


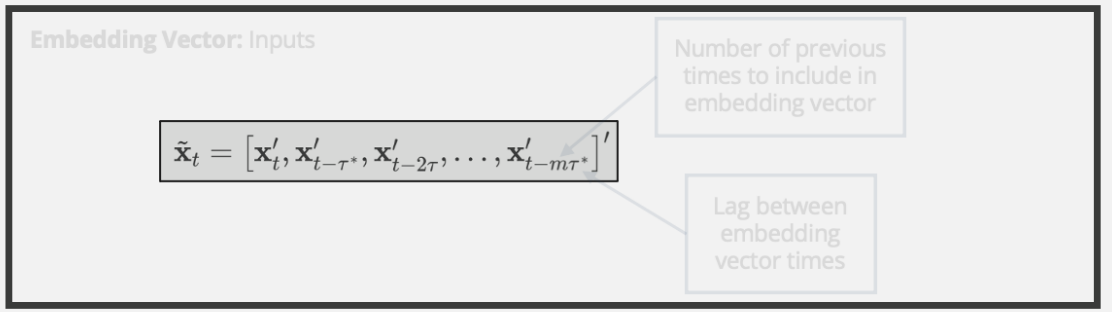


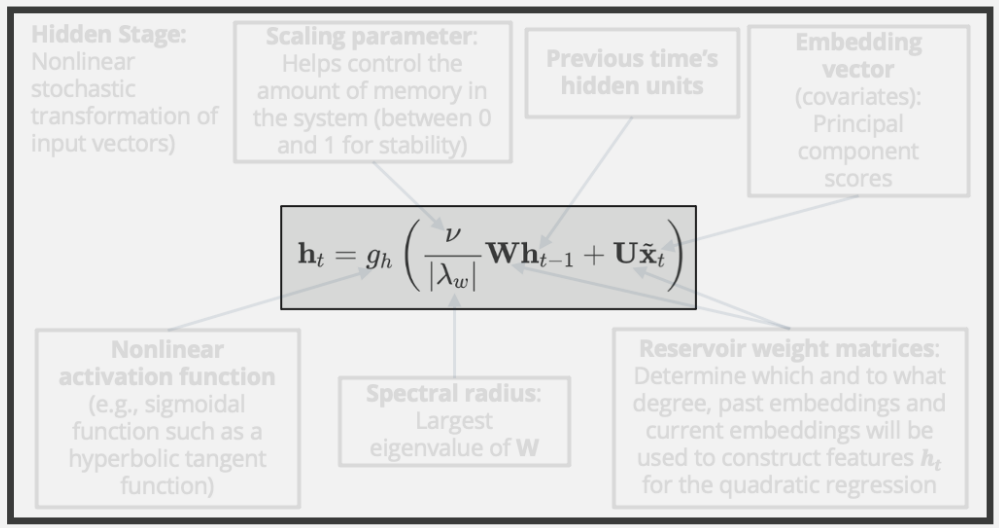


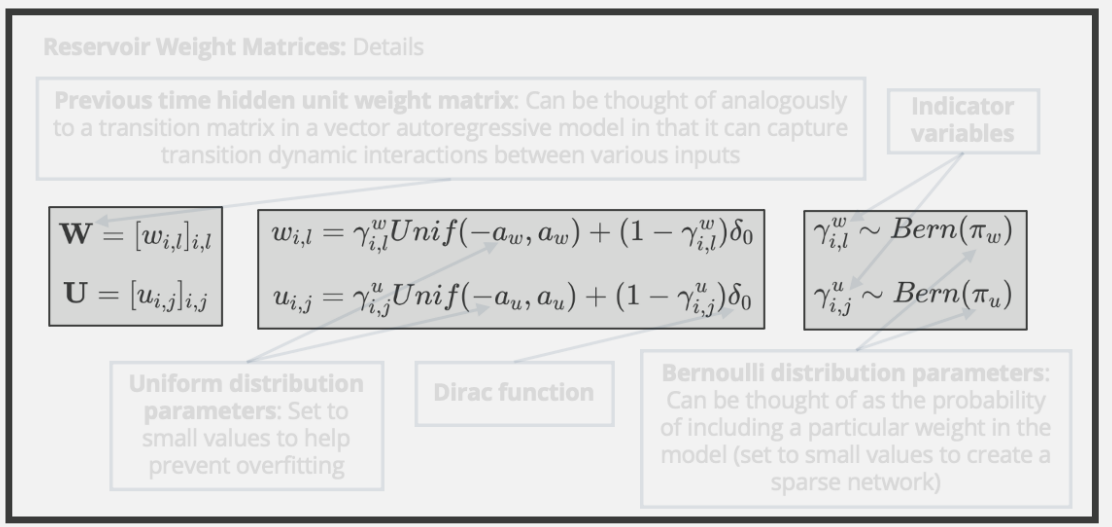
ESN Details

Quadratic Echo State Network









ESN Details

Quadratic Echo State Network

Output Stage: Ridge regression

$$\mathbf{Y}_t = \mathbf{V}_1\mathbf{h}_t + \mathbf{V}_2\mathbf{h}_t^2 + \boldsymbol{\epsilon}_t \quad \boldsymbol{\epsilon}_t \sim Gau\left(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}\right)$$

- **Response matrix**: Principal component scores (time series) that capture spatial trends
- **Ridge regression coefficients** (linear)
- **Ridge regression coefficients** (quadratic)
- **Ridge regression error variance**

Embedding Vector: Inputs

$$\tilde{\mathbf{x}}_t = \left[\mathbf{x}_t', \mathbf{x}_{t-\tau^*}', \mathbf{x}_{t-2\tau}', \ldots, \mathbf{x}_{t-m\tau^*}'\right]'$$

- Number of previous times to include in embedding vector
- Lag between embedding vector times

Hidden Stage: Nonlinear stochastic transformation of input vectors)

$$\mathbf{h}_t = g_h\left(\frac{\nu}{|\lambda_w|}\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\tilde{\mathbf{x}}_t\right)$$

- **Scaling parameter**: Helps control the amount of memory in the system (between 0 and 1 for stability)
- **Previous time's hidden units**
- **Embedding vector** (covariates): Principal component scores
- **Nonlinear activation function** (e.g., sigmoidal function such as a hyperbolic tangent function)
- **Spectral radius**: Largest eigenvalue of **W**
- **Reservoir weight matrices**: Determine which and to what degree, past embeddings and current embeddings will be used to construct features $\boldsymbol{h}_t$ for the quadratic regression

Reservoir Weight Matrices: Details

Previous time hidden unit weight matrix: Can be thought of analogously to a transition matrix in a vector autoregressive model in that it can capture transition dynamic interactions between various inputs

Indicator variables

$$\mathbf{W} = [w_{i,l}]_{i,l} \qquad \mathbf{U} = [u_{i,j}]_{i,j}$$

$$w_{i,l} = \gamma_{i,l}^w Unif(-a_w, a_w) + (1-\gamma_{i,l}^w)\delta_0$$

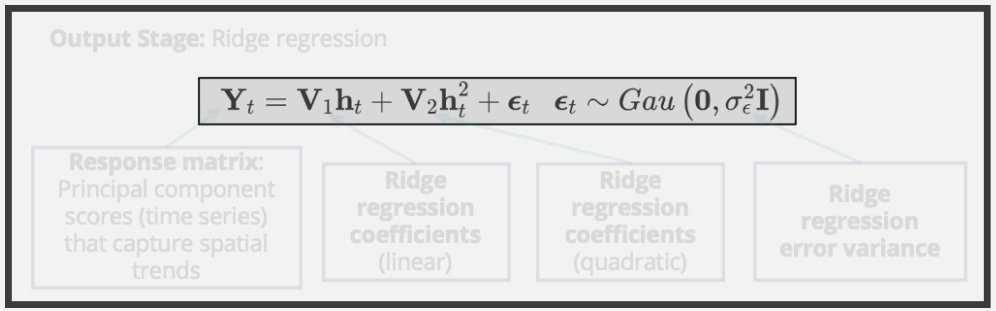
$$u_{i,j} = \gamma_{i,j}^u Unif(-a_u, a_u) + (1-\gamma_{i,j}^u)\delta_0$$

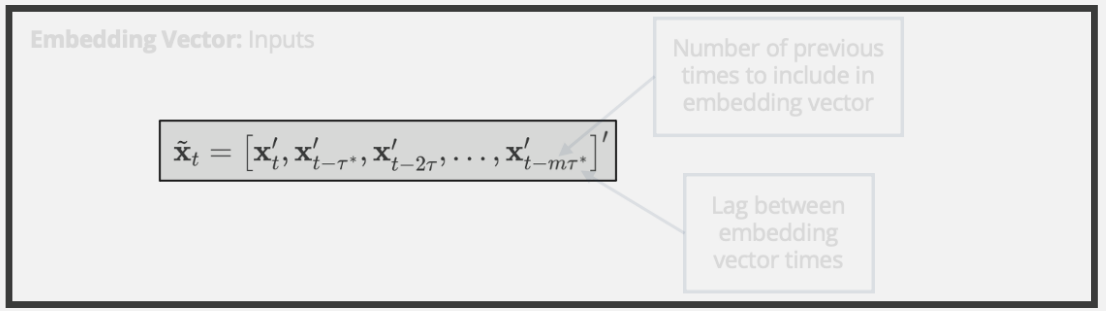
$$\gamma_{i,l}^w \sim Bern(\pi_w) \qquad \gamma_{i,j}^u \sim Bern(\pi_u)$$

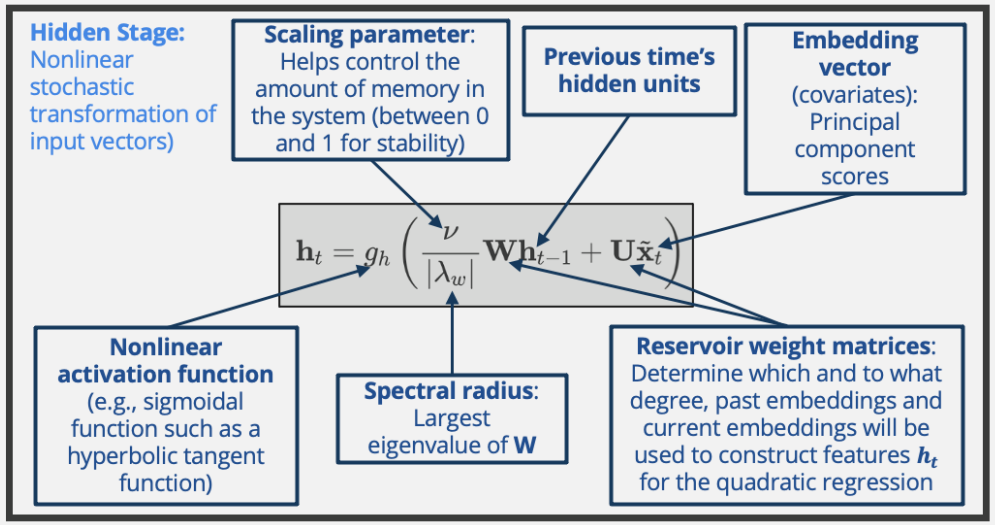
- **Uniform distribution parameters**: Set to small values to help prevent overfitting
- **Dirac function**
- **Bernoulli distribution parameters**: Can be thought of as the probability of including a particular weight in the model (set to small values to create a sparse network)

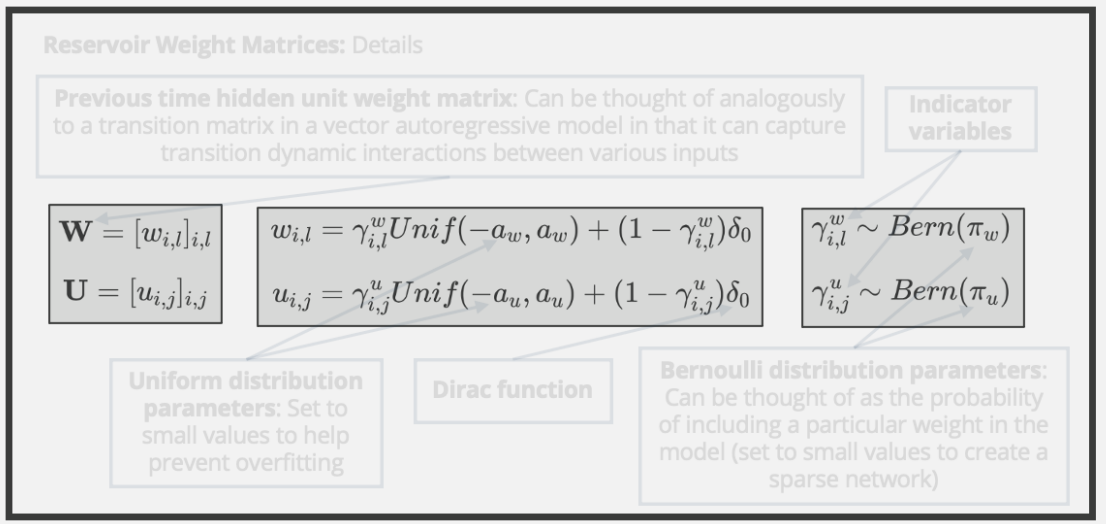
ESN Details

Quadratic Echo State Network









ESN Details

Quadratic Echo State Network

Output Stage: Ridge regression

$$\mathbf{Y}_t = \mathbf{V}_1\mathbf{h}_t + \mathbf{V}_2\mathbf{h}_t^2 + \boldsymbol{\epsilon}_t \quad \boldsymbol{\epsilon}_t \sim Gau\left(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}\right)$$

- **Response matrix**: Principal component scores (time series) that capture spatial trends
- **Ridge regression coefficients** (linear)
- **Ridge regression coefficients** (quadratic)
- **Ridge regression error variance**

Embedding Vector: Inputs

$$\tilde{\mathbf{x}}_t = \left[\mathbf{x}_t', \mathbf{x}_{t-\tau^*}', \mathbf{x}_{t-2\tau}', \ldots, \mathbf{x}_{t-m\tau^*}'\right]'$$

- Number of previous times to include in embedding vector
- Lag between embedding vector times

Hidden Stage: Nonlinear stochastic transformation of input vectors)

$$\mathbf{h}_t = g_h\left(\frac{\nu}{|\lambda_w|}\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\tilde{\mathbf{x}}_t\right)$$

- **Scaling parameter**: Helps control the amount of memory in the system (between 0 and 1 for stability)
- **Previous time's hidden units**
- **Embedding vector** (covariates): Principal component scores
- **Nonlinear activation function** (e.g., sigmoidal function such as a hyperbolic tangent function)
- **Spectral radius**: Largest eigenvalue of **W**
- **Reservoir weight matrices**: Determine which and to what degree, past embeddings and current embeddings will be used to construct features $\boldsymbol{h}_t$ for the quadratic regression

Reservoir Weight Matrices: Details

Previous time hidden unit weight matrix: Can be thought of analogously to a transition matrix in a vector autoregressive model in that it can capture transition dynamic interactions between various inputs

Indicator variables

$$\mathbf{W} = [w_{i,l}]_{i,l} \qquad \mathbf{U} = [u_{i,j}]_{i,j}$$

$$w_{i,l} = \gamma_{i,l}^w Unif(-a_w, a_w) + (1-\gamma_{i,l}^w)\delta_0$$

$$u_{i,j} = \gamma_{i,j}^u Unif(-a_u, a_u) + (1-\gamma_{i,j}^u)\delta_0$$

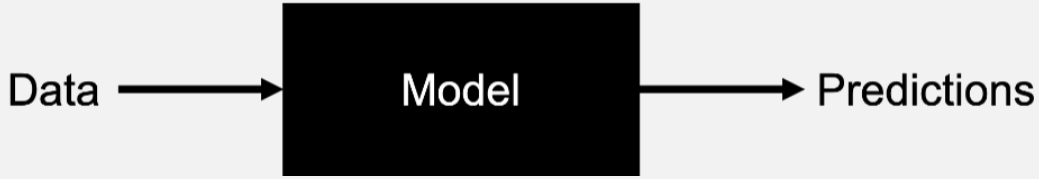
$$\gamma_{i,l}^w \sim Bern(\pi_w) \qquad \gamma_{i,j}^u \sim Bern(\pi_u)$$

- **Uniform distribution parameters**: Set to small values to help prevent overfitting
- **Dirac function**
- **Bernoulli distribution parameters**: Can be thought of as the probability of including a particular weight in the model (set to small values to create a sparse network)

Issue and Solution

Black-box:

- ESN parameters NOT interpretable (unlike spatio-temporal statistical models)
- Objective is to quantify variable relationships...



Interpretable: A model is interpretable if it is possible to assign meaning to the model's parameters in the context of the application, which provides insight into how the model inputs relate to the model outputs.

- Consider a linear model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$. We can interpret the coefficient $\hat{\beta}_1$ as the amount the response variable $\hat{y}$ increases for a one unit increase in the predictor variable x_1.

Explainable: A model is explainable if it is possible to implement post hoc investigations on a trained model that infer how the model inputs relate to the model outputs.

- Feature importance: Aims to quantify the effect of an input variable on a model's predictions. Various techniques have been proposed for computing FI

Feature Importance: Spatio-Temporal Context

Compute FI on the trained ESN model for...

- spatio-temporal input variable k
- over the block of times $\{t, t-1, \ldots, t-b+1\}$
- on the forecasts of the spatio-temporal response variable at time $t+\tau$.

	$x_{1,t,1}$	...	x_{1,t,P_1}	$x_{2,t,1}$	...	x_{2,t,P_2}	...	$x_{K,t,1}$	...	x_{K,t,P_K}
$t=1$										
$t=2$										
$t=3$										
$t=4$										
$t=5$										
...										
$t=T$										

	$y_{1,t}$	...	$y_{Q,t}$
$t=1$			
$t=2$			
$t=3$			
$t=4$			
$t=5$			
...			
$t=T$			

Feature Importance: Spatio-Temporal Context

	$x_{1,t,1}$	...	x_{1,t,P_1}	$x_{2,t,1}$	...	x_{2,t,P_2}	...	$x_{K,t,1}$	...	x_{K,t,P_K}
$t = 1$										
$t = 2$										
$t = 3$										
$t = 4$										
$t = 5$										
...										
t = T										

	$y_{1,t}$	...	$y_{Q,t}$
$t = 1$			
$t = 2$			
$t = 3$			
$t = 4$			
$t = 5$			
...			
t = T			

Two Approaches: "Adjust" inputs by either

- Permutation: spatio-temporal permutation feature importance (stPFI)
- Set values to zero: spatio-temporal zeroed feature importance (stZFI)

Feature Importance: Difference in RMSEs from observed and "adjusted" spatial predictions

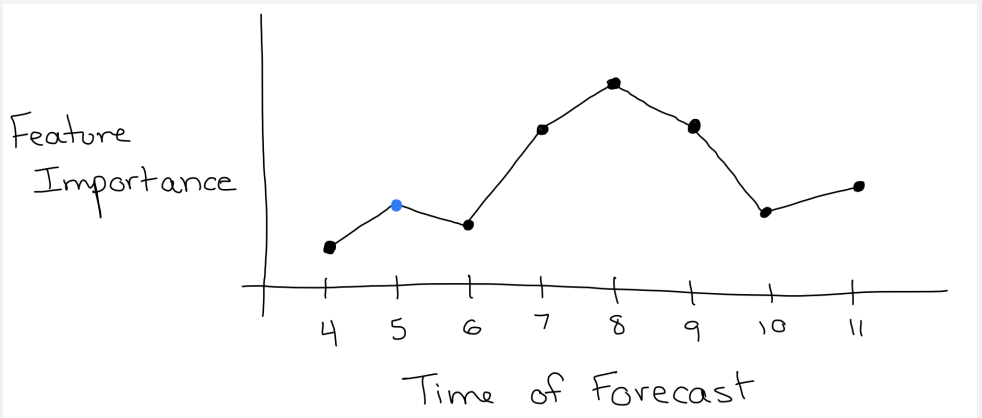
$$\mathcal{I}_{t,t+\tau}^{(k,b)} = \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b)}\right) - \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}\right)$$

Feature Importance: Spatio-Temporal Context

	$x_{1,t,1}$	...	x_{1,t,P_1}	$x_{2,t,1}$	...	x_{2,t,P_2}	...	$x_{K,t,1}$	...	x_{K,t,P_K}
$t = 1$										
$t = 2$										
$t = 3$										
$t = 4$										
$t = 5$										
...										
$t = T$										

	$y_{1,t}$	...	$y_{Q,t}$
$t = 1$			
$t = 2$			
$t = 3$			
$t = 4$			
$t = 5$			
...			
$t = T$			

Visualization: Feature importance of $\mathbf{x}_1$ during times $\{t, t-1, t-2\}$ on forecast of $\mathbf{y}_t$ at time $t+1$:



Feature Importance Details

Let $\mathcal{I}_{t,t+\tau}^{(k,b)}$ denote the FI on the trained ESN model f for

- spatio-temporal input variable k
- over the block of times $\{t, t-1, \ldots, t-b+1\}$
- on the forecasts of the spatio-temporal response variable at time $t+\tau$.

We compute the FI $\mathcal{I}_{t,t+\tau}^{(k,b)}$ as follows:

Step 1: Obtain forecasts $f(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1) = \hat{\mathbf{y}}_{t+\tau}$ at time $t+\tau$.

Step 2: Let $\mathcal{M}$ be a model prediction performance metric comparing observed to predicted values with the constraint that smaller values indicated better model performance (e.g., root mean squared error). Compute the performance metric on the trained model f at time $t+\tau$ as:

$$\mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}\right).$$

Feature Importance Details

Step 3: Generate *adjusted* forecasts using one of the following two methods:

- Permutation (stPFI): For replicate $r = 1, 2, \ldots, R$, randomly permute the values within each vector $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, \ldots, \mathbf{x}_{k,t-b+1}$. Replace the corresponding observed values within $\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-b+1}$ with the permuted versions. Let the versions of $\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-b+1}$ containing the permuted values associated with variable k and replicate r be denoted as

$$\mathbf{x}_t^{(k,r)}, \mathbf{x}_{t-1}^{(k,r)}, \ldots, \mathbf{x}_{t-b+1}^{(k,r)},$$

  respectively. Then obtain forecasts at time $t + \tau$ as

$$f\left(\mathbf{x}_t^{(k,r)}, \mathbf{x}_{t-1}^{(k,r)}, \ldots \mathbf{x}_{t-b+1}^{(k,r)}, \mathbf{x}_{t-b}, \ldots, \mathbf{x}_1\right) = \hat{\mathbf{y}}_{t+\tau}^{(k,b,r)}.$$

  The R replications are implemented to account for variability among permutations.

Feature Importance Details

Step 3: Generate *adjusted* forecasts using one of the following two methods:

- Zeroing (stZFI): Replace the vectors of $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, \ldots, \mathbf{x}_{k,t-b+1}$ within $\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-b+1}$ with zeros. Let the versions of $\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-b+1}$ containing the inserted zeros associated with variable k be denoted as

$$\mathbf{x}_t^{(k)}, \mathbf{x}_{t-1}^{(k)}, \ldots, \mathbf{x}_{t-b+1}^{(k)},$$

  respectively. Then obtain forecasts at time $t + \tau$ as

$$f\left(\mathbf{x}_t^{(k)}, \mathbf{x}_{t-1}^{(k)}, \ldots \mathbf{x}_{t-b+1}^{(k)}, \mathbf{x}_{t-b}, \ldots, \mathbf{x}_1\right) = \hat{\mathbf{y}}_{t+\tau}^{(k,b)}.$$

  Note that no replications are needed to account for variability with zeroing.

Feature Importance Details

Step 4: Compute the prediction performance metric on the forecasts obtained by inputting the adjusted predictions into the trained model f. That is, with stPFI compute

$$\mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b,r)}\right),$$

for $r = 1, \ldots, R$, and with stZFI compute

$$\mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b)}\right).$$

Feature Importance Details

Step 5: Finally, compute:

stPFI at time $t+\tau$ as the average change in model prediction performance when inputs $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, \ldots, \mathbf{x}_{k,t-b+1}$ are permuted:

$$\mathcal{I}_{t,t+\tau}^{(k,b)} = \left[\frac{1}{R} \sum_{r=1}^{R} \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b,r)}\right) \right] - \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}\right),$$

or stZFI at time $t+\tau$ as the change in model prediction performance when inputs $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, \ldots, \mathbf{x}_{k,t-b+1}$ are set to 0:

$$\mathcal{I}_{t,t+\tau}^{(k,b)} = \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b)}\right) - \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}\right).$$

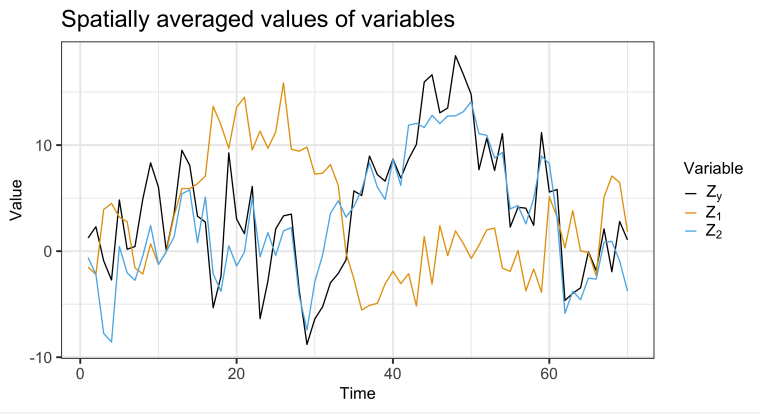
Simulated Data Demonstration

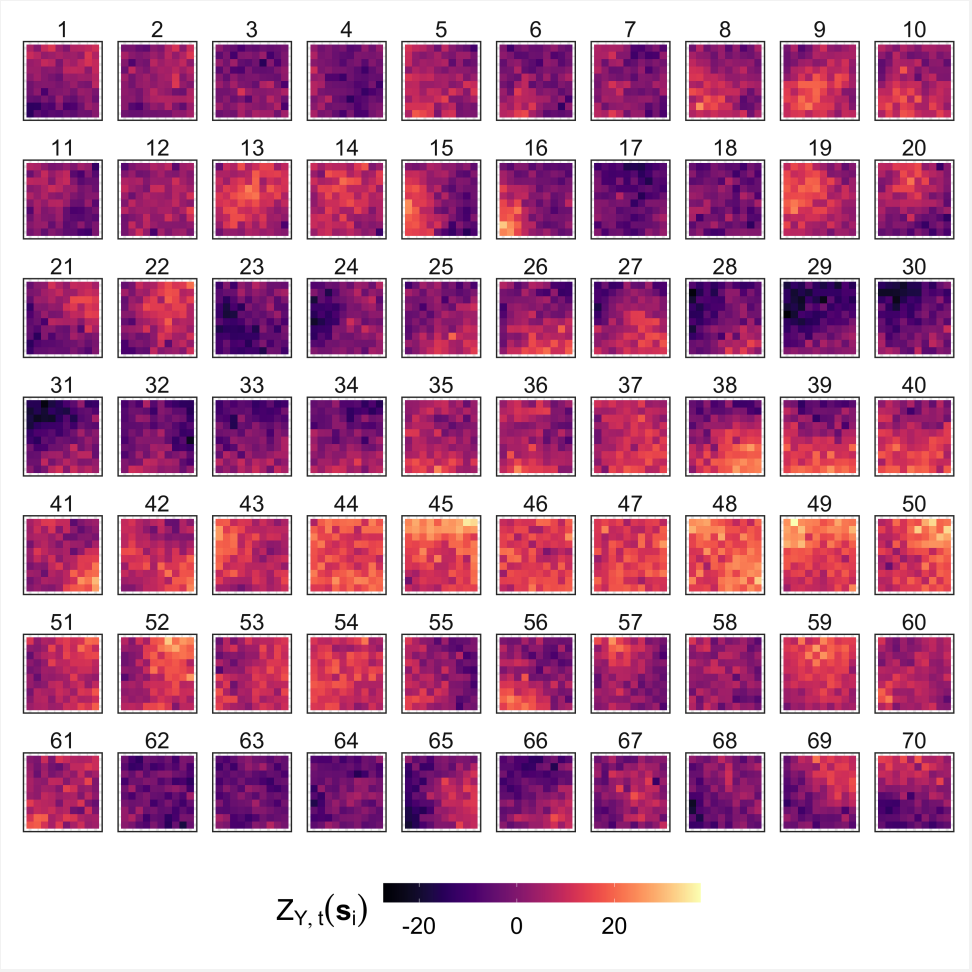
Simulated response

$$Z_{Y,t}(\mathbf{s}_i) = Z_{2,t}(\mathbf{s}_i)\beta + \delta_t(\mathbf{s}_i) + \epsilon_t(\mathbf{s}_i)$$

where

- $Z_{2,t}$ spatio-temporal covariate
- $\delta_t(\mathbf{s}_i)$ spatio-temporal random effect
- $\epsilon_t(\mathbf{s}_i) \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$





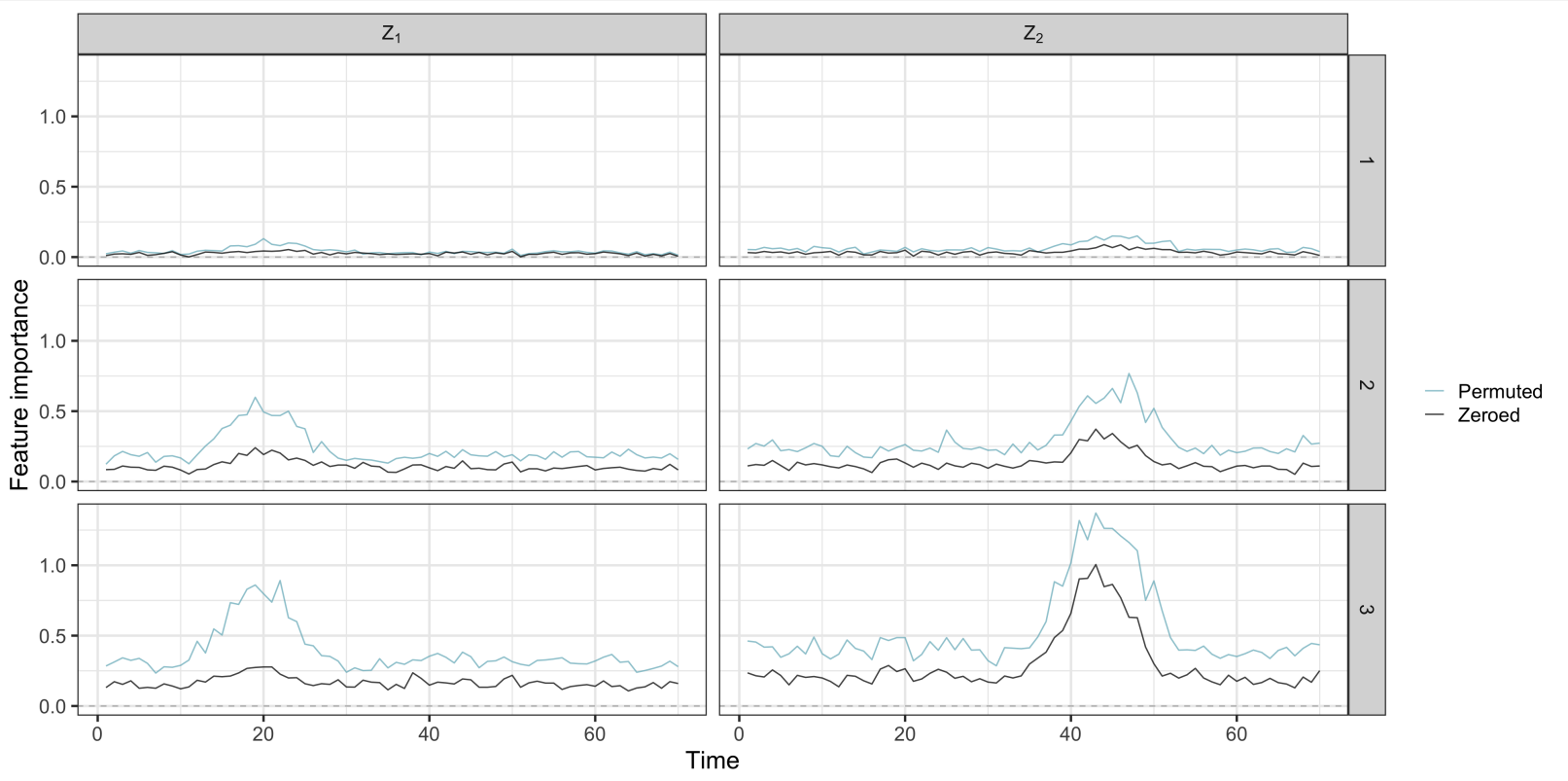
Simulated Data Demonstration

Fit an ESN

- Forecast $Z_{Y,t}$
- Inputs $Z_{1,t-\tau}$ and $Z_{2,t-\tau}$

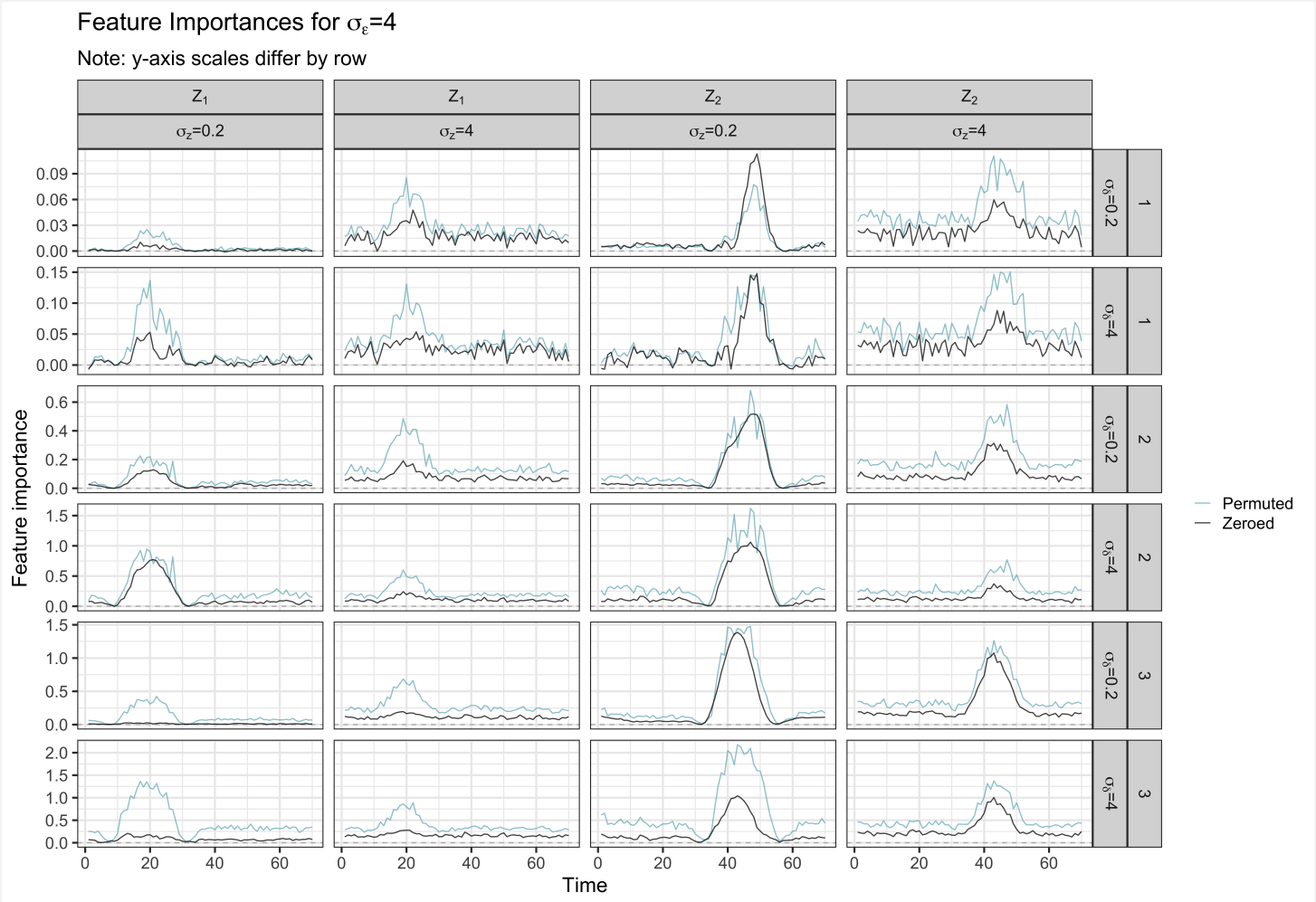
Compute stPFI and stZFI

- Blocks of size 1 to 3

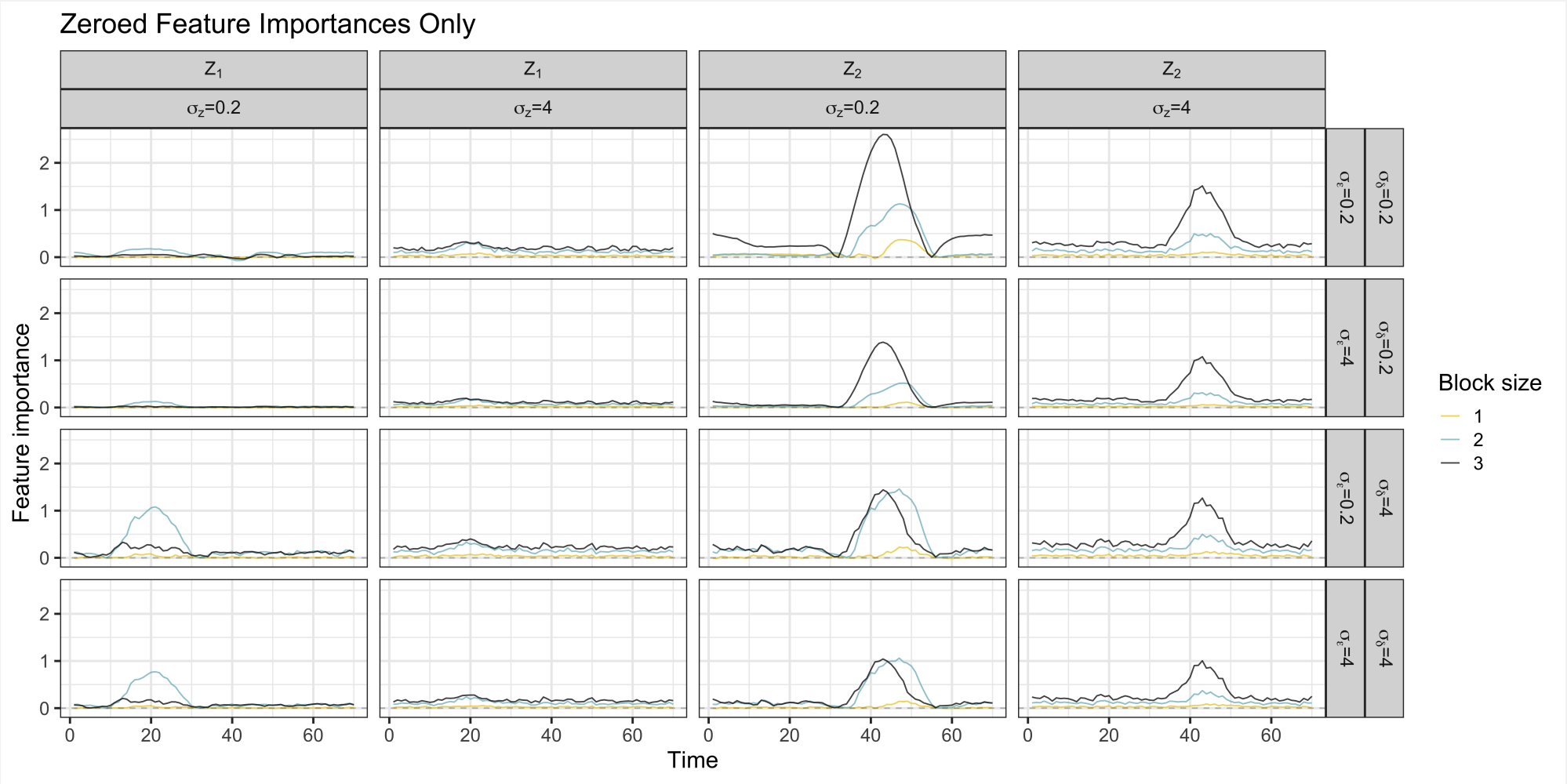


Each line represents the importance of the block of lagged times of an input variable on the forecast at time t

Simulated Data: Effect of Variability on FI



Simulated Data: Effect of Variability on FI



Effect of Correlation on FI

Effect of Correlation on PFI

Correlation between features can lead to biased PFI values dues to the model being forced to extrapolate

- When a correlated variable is permuted, it can lead to observations not in the training data
- Model is forced to extrapolate for that observation
- **Extrapolation can lead to a major effect on prediction making a variable seem more important than it is**

Example

Data is simulated so that X1 affects Y but X2 does not:

(Left) Within training data (stars) random forest correctly determines relationship between X1, X2, and Y (contour lines) but incorrect outside of training data

(Right) When X2 is permuted, observation could land outside training data and lead to change in prediction (i.e., large PFI)

Source: Hooker, Mentch, and Zhou (2021)