RIT

Analysis of Wide and Deep Echo State Networks for Multiscale Spatiotemporal Time Series Forecasting

Zachariah Carmichael, Humza Syed, Dhireesha Kudithipudi
Neuromorphic AI Lab
Rochester Institute of Technology

Supported By:

Outline

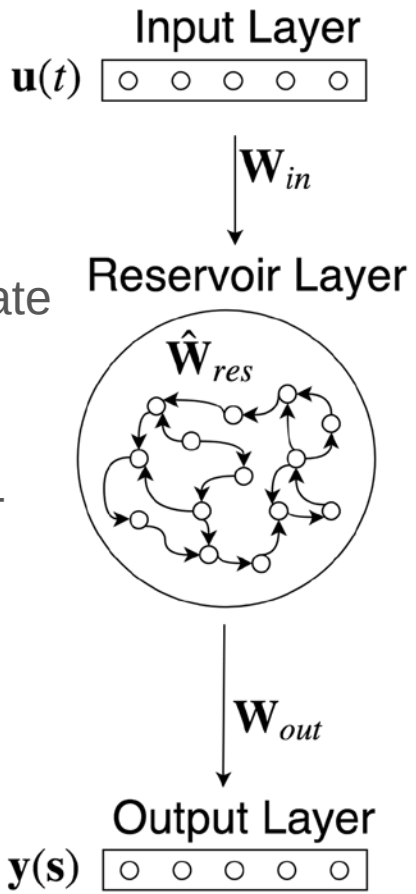
- Echo State Networks
- DeepESN Architecture
- Mod-DeepESN Architecture
- Analytical Measures
- Experiments and Results

Echo State Network

- Reservoir Computing: Recurrent neural networks with fixed random recurrent weights
- Reservoirs map inputs to a higher dimension and integrate inputs over time
- Readout layer trained using ridge regression with the Moore-Penrose pseudoinverse to approximate the least-squares solution
- χ(s) = state matrix from output of reservoir layer

$$\mathbf{y}(\mathbf{s}) = \mathbf{W}_{out}\boldsymbol{\chi}(\mathbf{s})$$

$$\mathbf{W}_{out} = \mathbf{y}(\mathbf{s}) \cdot (\boldsymbol{\chi}(\mathbf{s})^{\mathsf{T}} \cdot \boldsymbol{\chi}(\mathbf{s}) + \beta\mathbb{I})^{-1} \cdot \boldsymbol{\chi}(\mathbf{s})^{\mathsf{T}}$$



Why Deep or Wide ESNs?

- Baseline ESN performs well on small spatio-temporal tasks but underperforms for complex tasks with multi-scale dynamics
- Adding depth supports hierarchical representations of the temporal input at multiple time scales
- Adding width allows for diverse features to be extracted by an ensemble of reservoir pathways

DeepESN

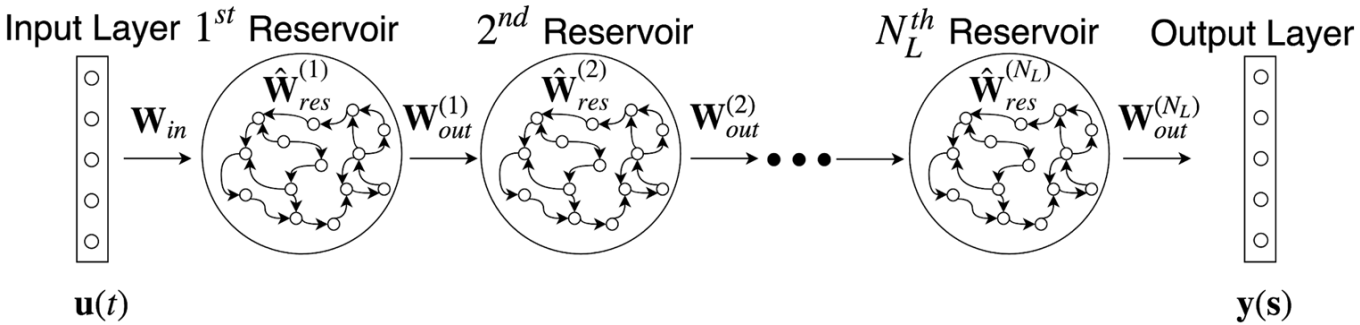
- Extracts temporal features in a hierarchical manner through propagation of data through each reservoir layer
- All reservoirs (and the input) are connected to the readout layer

$$\mathbf{x}^{(l)}(t) = \left(1 - a^{(l)}\right)\mathbf{x}^{(l)}(t-1) + a^{(l)}\tanh(\mathbf{W}_{in}^{(l)}\mathbf{i}^{(l)}(t) + \boldsymbol{\theta}^{(l)} + \widehat{\mathbf{W}}^{(l)}\mathbf{x}^{(l)}(t-1)),$$

α = leaky rate

$x^{(l)}(t)$ = output of reservoir at layer l and time step t

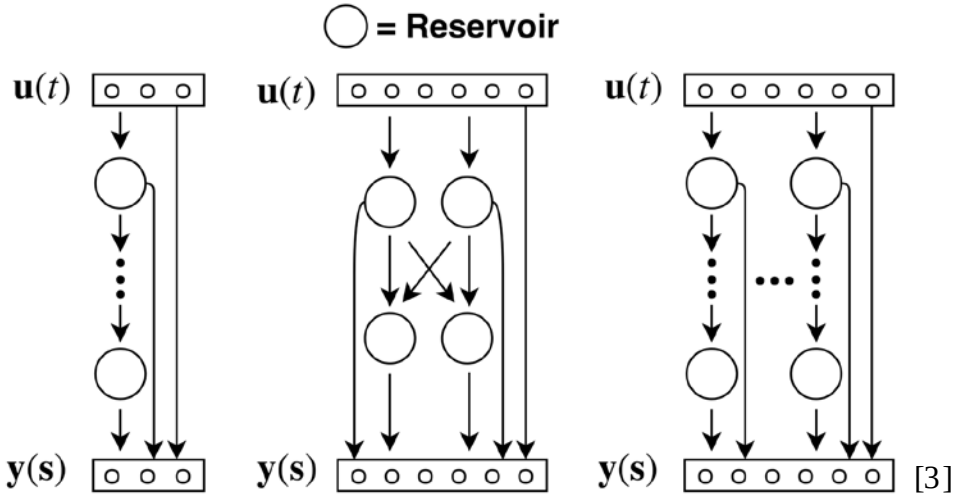
$\hat{W}^{(l)}$ = reservoir weights at layer l



Mod-DeepESN - Modular Deep Echo State Network

- Utilization of heterogeneous topologies and connectivity to capture multi-scale dynamics of input temporal data

- Intrinsic plasticity (IP) employed to fit reservoir responses towards a Gaussian



[3]

Mod-DeepESN - Modular Deep Echo State Network

- State matrix (χ) definition (for time series) and ridge regression equations

$$\mathbf{x}^{(l)}(t) = (1 - a^{(l)})\mathbf{x}^{(l)}(t-1) + a^{(l)} \tanh\left(\mathbf{W}_{res}^{(l)}\mathbf{i}^{(l)}(t) + \hat{\mathbf{W}}_{res}^{(l)}\mathbf{x}^{(l)}(t-1)\right)$$

$$\mathbf{x}(t) = (\mathbf{u}(t), \mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(N_L)}(t)) \in \mathbb{R}^{N_U + N_L N_R}$$

$$\chi(\mathbf{s}) = (\mathbf{x}(0), \ldots, \mathbf{x}(N_t - 1)) \in \mathbb{R}^{(N_U + N_L N_R) \times N_t}$$

$$\mathbf{W}_{out} = \mathbf{y}(\mathbf{s}) \cdot (\chi(\mathbf{s})^{\intercal} \cdot \chi(\mathbf{s}) + \beta\mathbb{I})^{-1} \cdot \chi(\mathbf{s})^{\intercal}$$

$N_{U/L/R/t}$ = *# of Inputs /* # of Layers / # of Reservoir Neurons / # of Time Steps

α = leaky rate

$\mathbf{u}(t)$ = input at time step t

$\mathbf{x}(t)$ = output of reservoir at time step t

$\hat{\mathbf{W}}_{res}^{(l)}$ = reservoir weights at layer l

Metrics

- Maximum Lyapunov Exponent (MLE) [4]
  - Characterizes the rate of separation of each reservoir given input sequences
  - Used as an indicator of the stability of the reservoir
    - Are the dynamics of the reservoir closer to a chaotic state or a stable state?
  - We adapt to the breadth of our network

$$\lambda_{max} = \max_{\substack{i=1,\dots,N_B \\ j=1,\dots,N_D \\ k=1,\dots,N_R}} \frac{1}{N_S} \sum_{t=1}^{N_S} \ln\left(\left|\mathrm{eig}_k\left((1-\alpha)\mathbb{I} + \alpha \mathbf{D}^{(i,j)}(t)\hat{\mathbf{W}}^{(i,j)}\right)\right|\right)$$

$$\mathbf{D}^{(i,j)}(t) = \begin{bmatrix} 1-\left(\tilde{x}_1^{(i,j)}(t)\right)^2 & 0 & \dots & 0 \\ 0 & 1-\left(\tilde{x}_2^{(i,j)}(t)\right)^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1-\left(\tilde{x}_{N_R}^{(i,j)}(t)\right)^2 \end{bmatrix}$$

$N_{B/D/R/S}$ = Breadth / Depth / # of Reservoir neurons / # of time steps

α = leaky rate

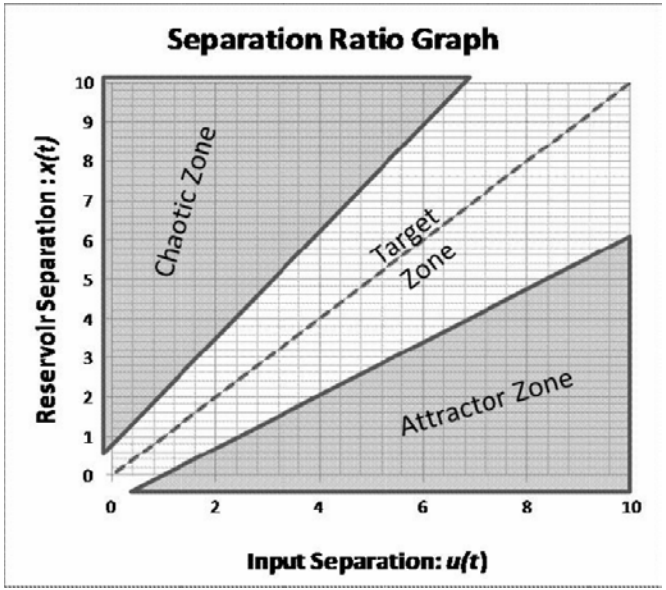
$D^{(i,j)}(t)$ = Diagonal matrix, whose non-zero entries are reservoir outputs at each time step ($\tilde{x}^{(i,j)}(t)$)

‹ $^{(i,j)}$ = reservoir weights of reservoir i, j

Metrics

- Separation Ratio Graphs [5]
  - Ratio between the distance of the input vectors and their corresponding reservoir output vectors
  - Ideal case is when y = x; slope = 1, y-intercept = 0

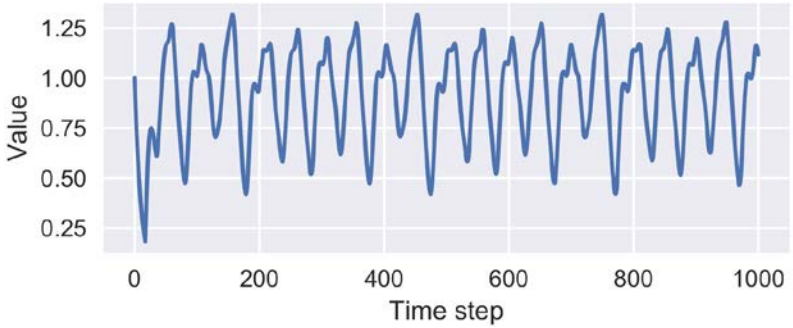
$$\frac{||\mathbf{x_i}(\mathbf{t}) - \mathbf{x_j}(\mathbf{t})||}{||\mathbf{u_i}(\mathbf{t}) - \mathbf{u_j}(\mathbf{t})||}$$



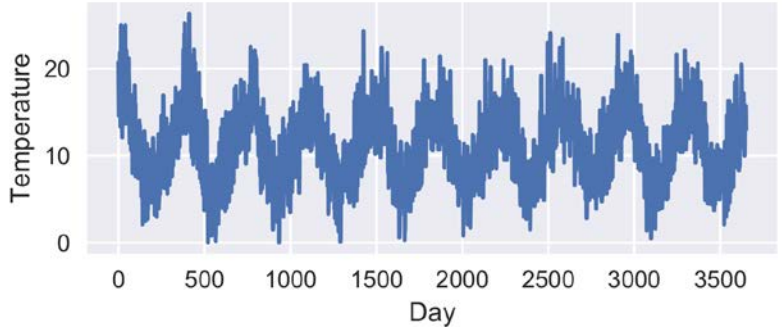
[5]

Time Series Tasks

- Mackey-Glass chaotic time series
  - 84-step-ahead forecasting



- Melbourne daily minimum temperature series [6]
  - 1-step-ahead forecasting



Forecasting Results

$$\text{RMSE} = \sqrt{\frac{1}{N_t}\sum_{t=1}^{N_t}[\mathbf{u}(t) - \hat{\mathbf{u}}(t)]^2}$$

$$\text{NRMSE} = \sqrt{\sum_{t=1}^{N_t}[\mathbf{u}(t) - \hat{\mathbf{u}}(t)]^2 / \left(\sum_{t=1}^{N_t}[\mathbf{u}(t) - \bar{\mathbf{u}}]^2\right)}$$

$$\text{MAPE} = \frac{1}{N_t}\sum_{t=1}^{N_t}\frac{|\mathbf{u}(t) - \hat{\mathbf{u}}(t)|}{\mathbf{u}(t)} \times 100\%$$

- Competitive with ESNs for Mackey Glass
- Outperforms in Melbourne forecasting
- IP plays a large role in reducing Mod-DeepESN error

	Method	N_L	N_R	IP	RMSE	NRMSE	MAPE
Baseline	ESN [11]	1	-	-	43.7	201	7.03
	ϕ-ESN [4]	2	-	-	8.60	39.6	1.00
	R^2SP [2]	2	-	-	27.2	125	1.00
	MESM [14]	7	-	-	12.7	58.6	1.91
	Deep-ESN [12]	2	-	-	**1.12**	**5.17**	**.151**
Ours	*Wide*	3	256	N	56.1	54.4	8.94
	Layered	3	256	N	57.8	96.2	19.9
	Wide+Layered	6 (3)	256	N	41.1	55.4	11.2
	Wide	3	256	Y	7.22	27.5	5.55

[3]

Mackey Glass 84step ahead prediction

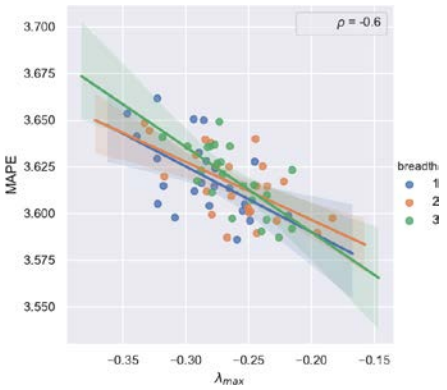
	Method	N_L	N_R	IP	RMSE	NRMSE	MAPE
Baseline	ESN [11]	1	-	-	501	139	39.5
	ϕ-ESN [4]	2	-	-	493	141	39.6
	R^2SP [2]	2	-	-	495	137	39.3
	MESM [14]	7	-	-	478	136	37.7
	Deep-ESN [12]	2	-	-	473	135	**37.0**
Ours	*Wide*	2	1024	N	473	135	38.6
	Layered	2	1024	N	470	134	38.2
	Wide+Layered	4 (2)	1024	N	471	135	38.6
	Wide+Layered	4 (2)	1024	Y	**459**	**132**	37.1

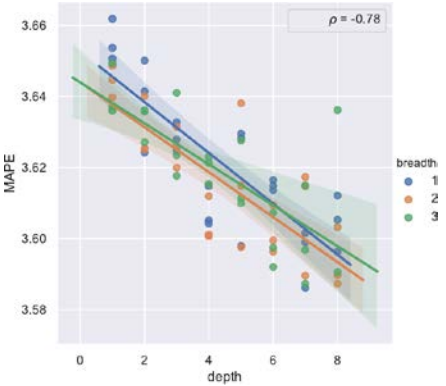
[3]

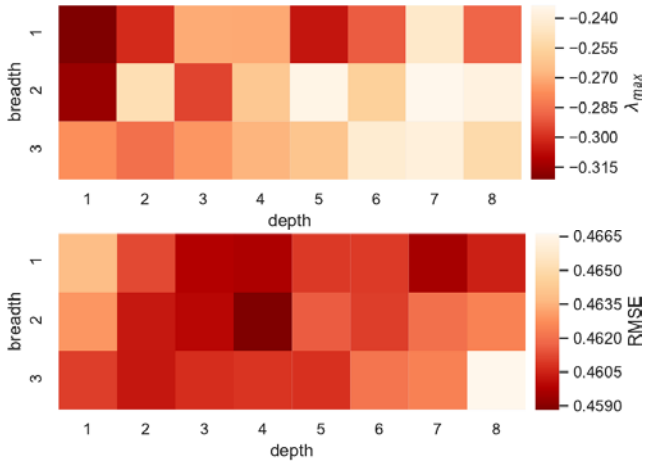
Melbourne 1-step ahead prediction

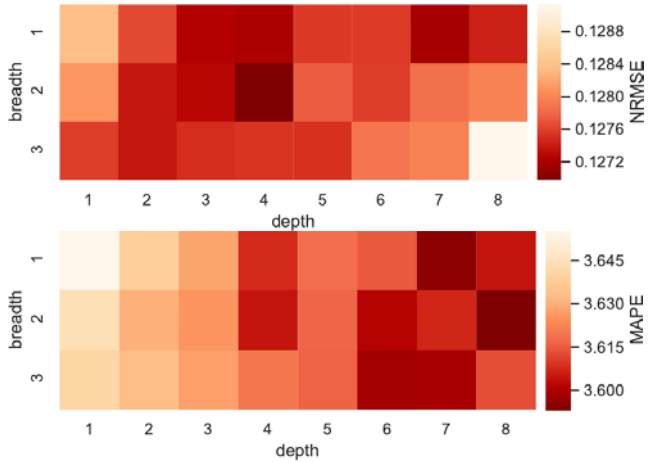
Melbourne Analysis

- MAPE:
  - Loosely correlated with MLE
  - Decreases consistently with depth
- Both breadth and depth bring MLE towards the “edge of chaos”

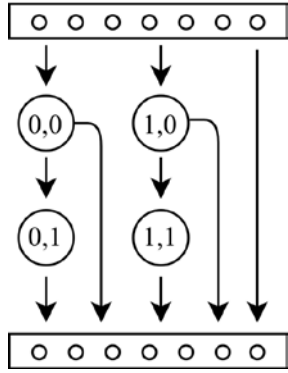


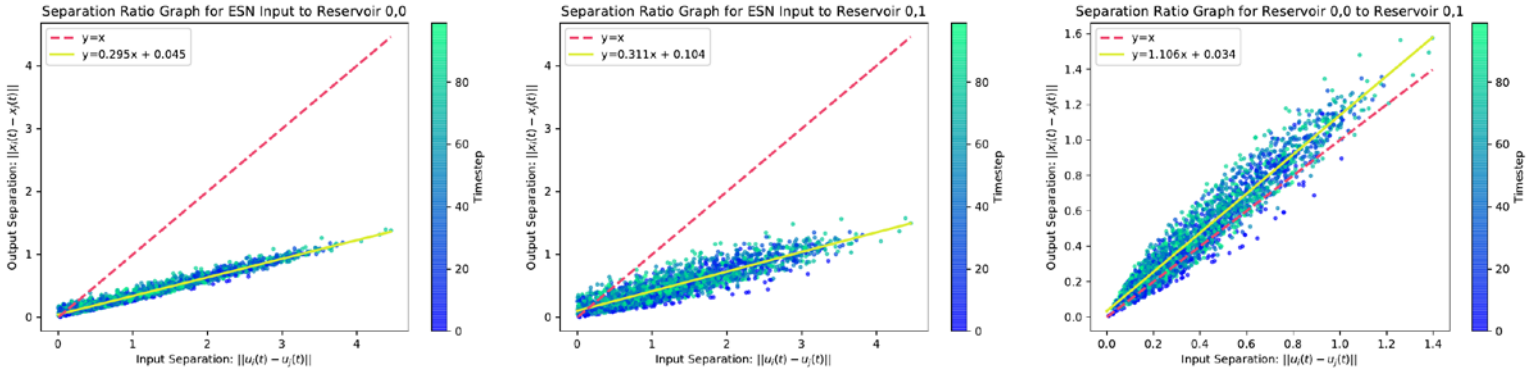


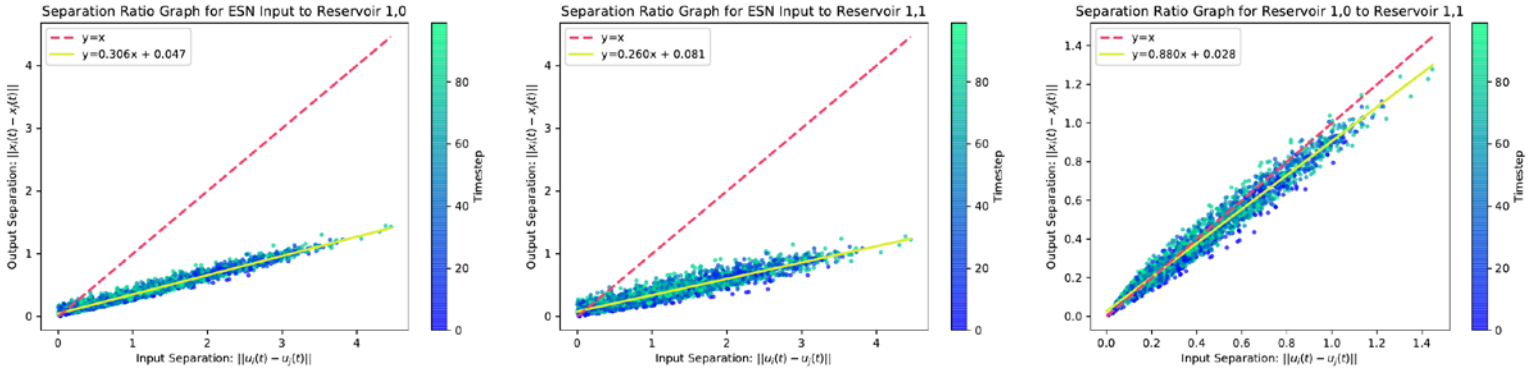




Melbourne Separation Ratio Graphs







Conclusions

- Introduce a network with heterogeneous topology and modular connectivity
- Explore the effects of breadth and depth within Mod-DeepESN
- Illustrate the hierarchical reservoir dynamics throughout the network
- Demonstrate that network topology and MLE can anticipate network performance

References

[1] H. Jaeger. 2001.The "echo state" approach to analysing and training recurrent neural networks. GMD Report 148. GMD - German National Research Institute for Computer Science. http://www.faculty.jacobs-university.de/hjaeger/pubs/EchoStatesTechRep.pdf

[2] Claudio Gallicchio, Alessio Micheli, Luca Pedrelli, Deep reservoir computing: A critical experimental analysis, Neurocomputing, Volume 268, 2017, Pages 87-99, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2016.12.089.

[3] Zachariah Carmichael,Humza Syed, Stuart Burtner, Dhireesha Kudithipudi. (2018). Mod-DeepESN: Modular Deep Echo State Network. 10.32470/CCN.2018.1239-0.

[4] Claudio Gallicchio, Alessio Micheli, Luca Silvestri, Local Lyapunov exponents of deep echo state networks, Neurocomputing, Volume 298, 2018, Pages 34-45, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2017.11.073.

[5] T. E. Gibbons, "Unifying quality metrics for reservoir networks," The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, 2010, pp. 1-7. doi: 10.1109/IJCNN.2010.5596307

[6] Time Series Data Library. (1981–1990).Daily minimum temperatures in melbourne,australia.Retrieved from https://datamarket.com/data/set/2324/daily-minimum-minimum-temperatures-in-melbourne-australia-1981-1990

Questions?