

# Multi-scale Dynamics by Adjusting Leaking Rate to Enhance Performance of Deep Echo State Networks

Shuichi Inoue<sup>1</sup>, Sou Nobukawa<sup>1,2,3,4,\*</sup>, Haruhiko Nishimura<sup>5</sup>, Eiji Watanabe<sup>6,7</sup> and Teijiro Isokawa<sup>8</sup>

<sup>1</sup>Graduate School of Information and Computer Science, Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, 275-0016, Chiba, Japan

<sup>2</sup>Department of Computer Science, Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, 275-0016, Chiba, Japan

<sup>3</sup>Research Center for Mathematical Engineering, Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, 275-0016, Chiba, Japan

<sup>4</sup>Department of Preventive Intervention for Psychiatric Disorders, National Institute of Mental Health, National Center of Neurology and Psychiatry, 4-1-1 Ogawa-Higashi, Kodaira, 187-8661, Tokyo, Japan

<sup>5</sup>Faculty of Informatics, Yamato University, 564-0082, Osaka, Japan

<sup>6</sup>National Institute for Basic Biology, 444-0867, Okazaki, Japan

<sup>7</sup>The Graduate University for Advanced Studies, SOKENDAI, 240-0193, Hayama, Japan

<sup>8</sup>Graduate School of Engineering, University of Hyogo, 671-2201, Himeji, Japan

Correspondence\*:

Sou Nobukawa

[nobukawa@cs.it-chiba.ac.jp](mailto:nobukawa@cs.it-chiba.ac.jp)

## 2 ABSTRACT

The deep echo state network (Deep-ESN) architecture, which comprises a multi-layered reservoir layer, exhibits superior performance compared to conventional echo state network (ESN) driven by the divergent layered-specific time-scale responses in the Deep-ESN. Although researchers have attempted to use experimental trial-and-error grid searches and Bayesian optimization methods to adjust the hyperparameters, suitable guidelines for setting hyperparameters to adjust the time-scale of dynamics in each layer from the viewpoint of dynamical characteristics have not been established. In this context, we hypothesized that evaluating the dependence of the multi-time-scale dynamical response on the leaking rate as a typical hyperparameter of time-scale in each neuron would help achieve a guideline for optimizing the hyperparameters of Deep-ESN. First, we set several leaking rates for each layer of the Deep-ESN and performed a multi-scale entropy (MSE) analysis to analyze the impact of leaking rate on the dynamics in each layer. Second, we performed layer-by-layer MSE analysis and cross-correlation analysis between adjacent layers to elucidate the structural mechanisms to enhance the performance. As a result, an optimum task-specific leaking rate value for producing layer-specific multiple time-scale responses and a queue structure with layer-to-layer signal transmission delays for retaining past applied input enhance the prediction performance of Deep-ESNs. These findings help establish ideal design guidelines for setting the hyper-parameters of Deep-ESN.

21 **Keywords:** Multi-scale dynamics, Machine learning, Reservoir computing, Echo State Network, Deep Echo State Network

## 1 INTRODUCTION

22 Reservoir computing, a branch of recurrent neural networks (RNNs), has garnered significant interest  
23 in terms of applications in machine learning (Lukoševičius and Jaeger, 2009; Tanaka et al., 2019;  
24 Gallicchio and Micheli, 2021). A neural network for reservoir computing consists of three layers: an  
25 input layer, a reservoir layer, and an output layer (Jaeger, 2001; Lukoševičius and Jaeger, 2009). In  
26 reservoir computing, the input time-series data are transformed into spatio-temporal patterns at the  
27 reservoir layer. For the responses of these transformed patterns, the responses of individual neurons act as  
28 a kernel, representing the desired output signal, which is consequently used for time-series prediction and  
29 classification (Gallicchio and Micheli, 2021; Jaeger, 2001).

30 Echo state network (ESN), a representative model for reservoir computing, operates based on the  
31 response of firing rate (Jaeger, 2007). In an ESN, depicted in Fig.1, the synaptic connections of reservoir  
32 weights are fixed, and only the synaptic connections of the output weights matrix are adjusted in the  
33 learning process (Tanaka et al., 2019). This architecture differs from other RNNs wherein all synaptic  
34 connections within the network undergo updates during learning (Williams and Zipser, 1989). Thus, ESNs  
35 are more learning-efficient (Werbos, 1990) than the widely-utilized LSTM model, still demonstrating  
36 lower accuracy (Gallicchio et al., 2018; Salehinejad et al., 2017). Such efficient learning architectures  
37 may open up the potential for applications in areas characterized by resource-limited hardware, such as  
38 edge devices (Tanaka et al., 2019; Sakemi et al., 2024).

39 The deep echo state network (Deep-ESN) model, which possesses the multi-layered reservoir layer  
40 as illustrated in Fig. 2, has been proposed. This model performs better than the conventional ESN,  
41 consisting of a single-layered reservoir (Deng et al., 2012; Gallicchio et al., 2017; Gallicchio and Micheli,  
42 2021). The divergent responses in each layer of Deep-ESN, which exhibits multiple time-scale  
43 dynamics, enhance the memory capacity and feature representation compared to its single-layered  
44 counterpart (Gallicchio et al., 2017; Malik et al., 2016; Tchakoucht and Ezziyyani, 2018; Long et al.,  
45 2019; Gallicchio and Micheli, 2019; Kanda and Nobukawa, 2022). These advantages of Deep-ESN may  
46 enable applications in tasks involving non-linear dynamic signals that exhibit multi-time-scale behaviors  
47 in various types of systems such as biological systems, power systems, and financial markets (Costa et al.,  
48 2002; Venkatasubramanian et al., 1995; Yan and He, 2021; Bhandari, 2017; Chen and Shang, 2020).

49 The adjustment of numerous hyperparameters in Deep-ESN is often based on experimental  
50 measurements, trial-and-error grid searches, and Bayesian optimization methods (Lukoševičius and Uzelis,  
51 2019; Adeleke, 2019; Viehweg et al., 2023; Bai et al., 2023). Regarding the Bayesian optimization in  
52 Deep-ESN, the emphasis is mainly placed on performance, and the analysis of reservoir dynamics  
53 and mechanisms of functionality enhancement is often overlooked; consequently, there are no specific  
54 design guidelines on dynamics (Bai et al., 2023). To achieve the establishment of guidelines, in terms of  
55 hyperparameters that adjust the time scale of each layer, several studies have applied scaling methods  
56 to the input weights matrix,  $W^{(l)}$ , as illustrated in Fig. 2 (Kanda and Nobukawa, 2022). By integrating  
57 this method, the signal strength between layers decreases as the depth increases. This results in a layered-  
58 specific time-scale response for each layer (Kanda and Nobukawa, 2022). This characteristic was revealed  
59 by analyzing the layer dynamics using the multiscale entropy (MSE) method (Humeau-Heurtier, 2015;  
60 Kanda and Nobukawa, 2022). However, this approach, which is aimed at scaling input weights comes,  
61 possesses a major drawback. The signal strength in the reservoir layer diminishes quickly due to the small

62 scaling rate between layers. On the other hand, the leaking rate is a hyperparameter that influences the  
 63 temporal history effect of the dynamics in  $\mathbf{x}^{(l)}(t)$  of Fig.2 (Schrauwen et al., 2007; Jaeger et al., 2007).  
 64 Essentially, the leaking rate adjusts the decay factor of dynamics in each neuron (Schrauwen et al., 2007;  
 65 Jaeger et al., 2007). Therefore, the method for adjusting the leaking rate is another suitable candidate to  
 66 achieve the layered-specific dynamical response in Deep-ESN.

67 In this context, we hypothesized that evaluating the dependence of dynamical response in the multi-  
 68 layered reservoir in terms of adjusting the leaking rate would provide insights into achieving a guideline  
 69 for optimizing the hyper-parameters of Deep-ESN. For the preliminary investigation, we set the same  
 70 leaking rate for each layer of the Deep-ESN and performed an MSE analysis to analyze the impact  
 71 of the leaking rate on the dynamics in each layer. The results confirm that each layer of the Deep-  
 72 ESN generates dynamics at different time scales, which induces a queue-like property where the delay  
 73 response is preserved by the hierarchical structure (Inoue et al., 2023). However, the tendencies of  
 74 performance for more diverse time-series prediction tasks remain unclear; the evaluation in the case of  
 75 setting heterogeneous leaking rates in the multi-layered reservoir was not conducted. Therefore, in this  
 76 study, based on our preliminary outcomes of a previous study (Inoue et al., 2023), we further revealed  
 77 these points. Specifically, Deep-ESNs with homogenous and heterogeneous leaking rates for each layer  
 78 are used to perform and evaluate a time series prediction task using three time series signals, Lorenz,  
 79 Rössler, and Mackey-Glass models. Furthermore, we perform layer-by-layer MSE analysis and cross-  
 80 correlation analysis between adjacent layers to elucidate the mechanisms behind functional enhancement  
 81 achieved through the leaking rates.

## 2 MATERIAL AND METHODS

### 82 2.1 Echo State Network

83 Figure 1 shows an architecture of ESN. The input signal is defined as  $\mathbf{u}(t) \in \mathbb{R}^{N_u}$  with  $N_u$ -dimensional  
 84 inputs. The reservoir state is defined by  $\mathbf{x}(t) \in \mathbb{R}^{N_x}$ , where  $N_x$  is the number of neurons in the reservoir  
 85 layer. The reservoir state  $\mathbf{x}(t)$  is defined by

$$\mathbf{x}(t) = (1 - a)\mathbf{x}(t - 1) + a \tanh(\mathbf{W}_{\text{in}}\mathbf{u}(t) + \hat{\mathbf{W}}\mathbf{x}(t - 1)), \quad (1)$$

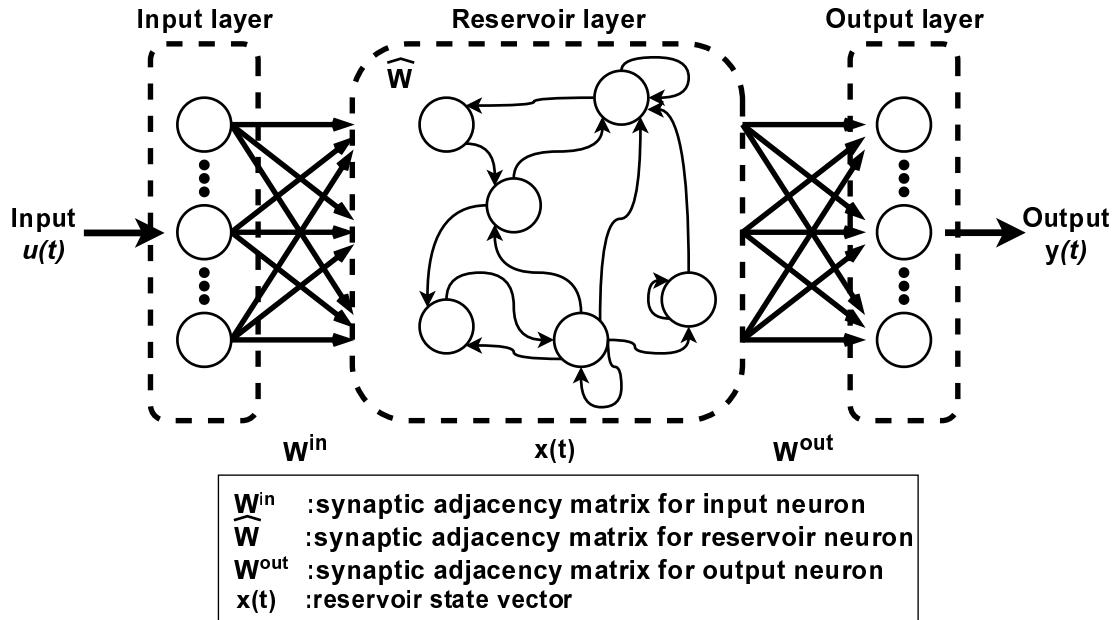
86 where  $a \in [0, 1]$  is the leaking rate, and  $\mathbf{W}_{\text{in}} \in \mathbb{R}^{N_x \times N_u}$  is the input weight matrix. Each component  
 87 of  $\mathbf{W}_{\text{in}}$  is given by a uniform random value, the range of which is  $[-s_{\text{in}}, s_{\text{in}}]$ . The  $\hat{\mathbf{W}} \in \mathbb{R}^{N_x \times N_x}$  is the  
 88 recurrent synaptic weight matrix, which is a random matrix with uniform random numbers, and its spectral  
 89 radius is set to  $\rho$ . The output of the ESN at time  $t$  is determined using

$$\mathbf{y}(t) = \mathbf{W}_{\text{out}}\mathbf{x}(t), \quad (2)$$

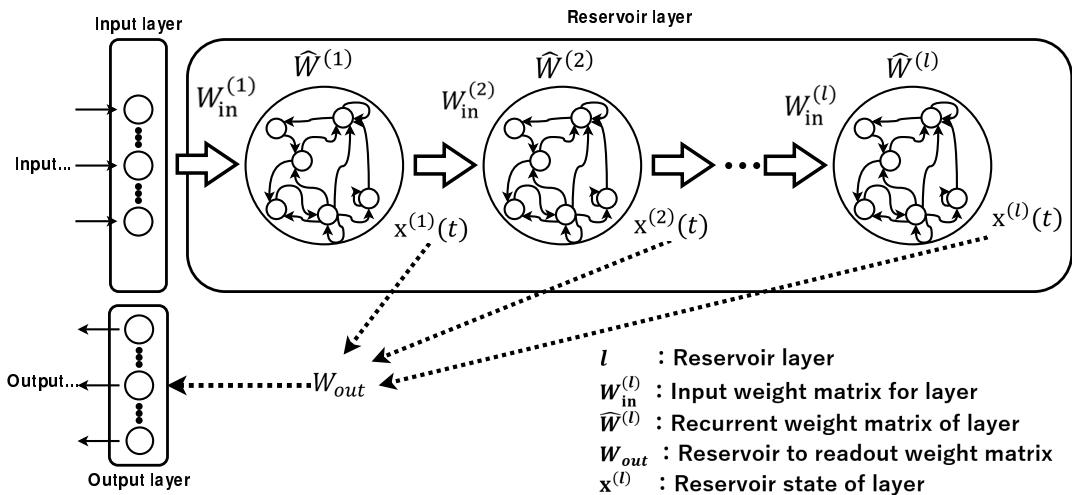
90 where  $\mathbf{W}_{\text{out}} \in \mathbb{R}^{N_y \times N_x}$  is defined as the output weight matrix and  $\mathbf{y}(t) \in \mathbb{R}^{N_y}$  is the  $N_y$ -dimensional  
 91 output. The initial value of  $\mathbf{W}_{\text{out}}$  is a random matrix of uniform random numbers.

### 92 2.2 Deep Echo State Network

93 The multi-layered Deep-ESN was constructed based on the single-layered ESN; Fig.2 shows a diagram  
 94 of the Deep-ESN (Deng et al., 2012; Gallicchio et al., 2017; Gallicchio and Micheli, 2021). The only  
 95 difference with ESN is that the reservoir layer is multi-layered. The reservoir state vector of Deep-ESN is



**Figure 1.** Architecture: Echo State Network (ESN).



**Figure 2.** Architecture: Deep Echo State Network (Deep-ESN).

96 defined by

$$\mathbf{x}^{(l)}(t) = (1 - a^{(l)})\mathbf{x}^{(l)}(t-1) + a^{(l)} \tanh(\mathbf{W}_{\text{in}}^{(l)} \mathbf{i}^{(l)}(t) + \boldsymbol{\theta}^{(l)} + \hat{\mathbf{W}}^{(l)} \mathbf{x}^{(l)}(t-1)), \quad (3)$$

97 where variables consisting  $l$  refer to  $l$ -layer parameters.  $\boldsymbol{\theta}^{(l)} \in \mathbb{R}^{N_x}$  is the bias in the reservoir coupling  
 98 weight matrix.  $\mathbf{W}_{\text{in}}^{(l)} \in \mathbb{R}^{N_x \times N_u}$  is the input weight matrix for each layer, and  $\hat{\mathbf{W}}^{(l)} \in \mathbb{R}^{N_x \times N_x}$  is the  
 99 recurrent weight matrix for each layer. The dynamics of each reservoir layer can be defined as  $\overline{x^{(l)}(t)}$ ,  
 100 averaging the reservoir state vector  $\mathbf{x}^{(l)}(t)$  over all neurons. Further,  $\mathbf{i}^{(l)}(t)$  represents the input to the  $l$ -th

101 layer in the Deep-ESN and is given by

$$\mathbf{i}^{(l)}(t) = \begin{cases} \mathbf{u}(t) & \text{if } l = 1, \\ \mathbf{x}^{(l-1)}(t) & \text{if } l > 1. \end{cases} \quad (4)$$

102 The output of the Deep-ESN at time  $t$  is performed by

$$\mathbf{y}(t) = \mathbf{W}_{\text{out}}[\mathbf{x}^{(1)}(t)\mathbf{x}^{(2)}(t)\dots\mathbf{x}^{(N_L)}(t)]^\top + \boldsymbol{\theta}_{\text{out}}, \quad (5)$$

103 where  $N_L$  is defined as the total number of reservoir layers and  $\mathbf{y}(t) \in \mathbb{R}^{N_y}$  is the output in  $N_y$   
 104 dimensions.  $\mathbf{W}_{\text{out}} \in \mathbb{R}^{N_y \times N_L N_x}$  is the output weight matrix. The bias in the output layer was set to  
 105  $\boldsymbol{\theta}_{\text{out}} = [1, 1, \dots, 1]^\top$ .

106 In this study, the total number of layers was set to  $N_L = 10$ , the number of neurons in each layer was  
 107 set to  $N_x = 100$ , the scaling parameter of the input weight matrix  $\mathbf{W}_{\text{in}}$  was set to  $s_{\text{in}} = 1$  and the spectral  
 108 radius set to  $\rho = 1.0, 0.9, 0.8$ , and ridge regression was used as the learning algorithm. For the leaking  
 109 rate  $a^{(l)}$  settings used in this experiment, a homogeneous model with the same leaking rate in all layers,  
 110 and a heterogeneous model with different leaking rates in each layer, are used. The leaking rate  $a^{(l)}$  of the  
 111 homogeneous model was commonly set to be  $a^{(l)} = 1.0, 0.9, \dots, 0.1$  in all layers. In the heterogeneous  
 112 model, the leaking rate  $a^{(l)}$  was set to decrease incrementally by 0.1 across each layer, starting from 1.0  
 113 and going down to 0.1. For simplicity, the same values were set for the hyperparameters  $N_L$ ,  $N_x$ ,  $s_{\text{in}}$ ,  $\rho$ ,  
 114 and  $a^{(l)}$  for each layer in the reservoir. In addition,  $\mathbf{W}_{\text{in}}$ ,  $\hat{\mathbf{W}}^{(l)}$ ,  $\mathbf{W}_{\text{out}}$  and  $\mathbf{x}(0)$  were initialized with different  
 115 random seeds for each trial. During the execution of the time-series prediction task, the seed values were  
 116 changed and 100 trials were performed.

## 117 2.3 Time-series prediction task

118 In terms of the impact of leaking rate on the performance, as typical non-linear signals, Mackey–  
 119 Glass equation, Lorenz equation and Rössler equation were prepared as time series signals with different  
 120 dynamical characteristics. In this study, homogeneous and heterogeneous models are used in a time series  
 121 prediction task, which is evaluated by predicting five subsequent steps. Each task involves 100 trials with  
 122 different initial values. The normalized root-mean-square deviation (NRMSE) was used to evaluate the  
 123 prediction accuracy for each task in the homogeneous and heterogeneous models.

### 124 2.3.1 Mackie-Grass equation

125 For time-series prediction, time-series were generated from the Mackey–Glass equation (Glass and Mackey,  
 126 2010):

$$\frac{dx_{\text{mg}}}{dt} = \frac{0.2x_{\text{mg}}(t - \tau)}{1 + x_{\text{mg}}(t - \tau)^{10}} - 0.1x_{\text{mg}}(t), \quad (6)$$

127 where  $\tau$  is a constant representing the delay. In this study,  $\tau = 32$ , the solution was obtained using the  
 128 fourth-order Runge-Kutta method, and the trajectories were sampled in a time window  $\Delta t = 10$ .

---

 129 2.3.2 Lorenz equation

 130 The Lorenz equation is represented by a system of non-linear differential equations of the form  
 131 (Manneville and Pomeau, 1979),

$$\begin{aligned}\frac{dx_l}{dt} &= \sigma(y_l - x_l), \\ \frac{dy_l}{dt} &= x_l(\rho - z_l) - y_l, \\ \frac{dz_l}{dt} &= x_l y_l - \beta z_l.\end{aligned}\quad (7)$$

 132 The parameters for the Lorenz equation were  $\sigma = 10$ ,  $r = 28$  and  $b = 8/3$ . These values are known  
 133 to exhibit chaotic behavior. In this study, the solution was obtained using the fourth-order Runge-Kutta  
 134 method, and the trajectories were sampled in a time window  $\Delta t = 0.02$ .

## 135 2.3.3 Rössler equation

 136 The Rössler system, which is a non-linear dynamical system (Rössler, 1983), was adopted to generate  
 137 chaotic time-series data. The system is defined by the following set of three non-linear ordinary differential  
 138 equations:

$$\frac{dx_r}{dt} = -y_r - z_r, \quad (8)$$

$$\frac{dy_r}{dt} = x_r + ay_r, \quad (9)$$

$$\frac{dz_r}{dt} = b + z_r(x_r - c). \quad (10)$$

 139 In these equations,  $x_r$ ,  $y_r$ , and  $z_r$  represent the system states. The parameters  $a$ ,  $b$ , and  $c$  directly affect the  
 140 behavior of the system. For our experiments, the parameters were set to  $a = 0.2$ ,  $b = 0.2$ , and  $c = 5.7$ . In  
 141 this study, the solution was obtained using the fourth-order Runge-Kutta method and the trajectories were  
 142 sampled in a time window  $\Delta t = 0.02$ .

## 143 2.4 Multiscale entropy analysis

 144 The MSE analysis is a method for performing a coarse-graining of the time-series of interest and  
 145 quantitatively evaluating the complexity across multiple time scales (Humeau-Heurtier, 2015). As an  
 146 analytical procedure, the first step is to coarse-grain the dynamics of each reservoir layer  $\overline{x^{(l)}(t)}$ , with a  
 147 time scale factor  $\tau_s$  using,

$$z_j^{(\tau_s)} = \left(\frac{1}{\tau_s}\right) \sum_{i=(j-1)\tau_s+1}^{j\tau_s} \overline{x^{(l)}(i)}, \quad (1 \leq j \leq \frac{N}{\tau_s}). \quad (11)$$

 148 In case  $\tau_s = 1$ , the original time-series is coarse-grained to a longer time scale dynamics as  $\tau_s$  increases.  
 149 At each time scale  $\tau_s$  and layer  $l$ , complexity of the coarse-grained time-series is then quantified by the

150 sample entropy (SampEn). SampEn is given by

$$\text{SampEn}(r, m, N) = -\log \frac{U_{m+1}(r)}{U_m(r)}, \quad (12)$$

151 where  $U_m(r)$  represents the probability of being  $|\mathbf{z}_i^m - \mathbf{z}_j^m| < r$  ( $i \neq j, i, j = 1, 2, \dots$ ) and  $\mathbf{z}_i^m$  represents  
 152 the  $m$ -dimensional vector  $\mathbf{z}_i^m = \{z_i^{(\tau_s)}, z_{i+1}^{(\tau_s)}, \dots, z_{i+m-1}^{(\tau_s)}\}$ . Thus, the complexity of the dynamics of each  
 153 layer can be analyzed and evaluated from different time-scale perspectives.

## 154 2.5 Normalised root mean square error

155 The NRMSE is a statistical measure used to assess the accuracy of a model's predictions, and it is  
 156 defined as follows:

$$\text{NRMSE} = \sqrt{\frac{\sum_{t=1}^T (y(t) - y_d(t))^2}{T\sigma^2(y_d)}}. \quad (13)$$

157 In this study, the task inputs and outputs are one-dimensional. Therefore,  $y(t)$  is the output of the ESN, in  
 158 the case with  $N_y = 1$ , at time  $t$ ,  $y_d(t)$  is the teacher signal at time  $t$ ,  $\sigma^2(y(t))$  is the variance of the teacher  
 159 signal, and  $T$  is the evaluation period (number of data points). Here, the NRMSE was evaluated among  
 160 100 trials with different initial conditions.

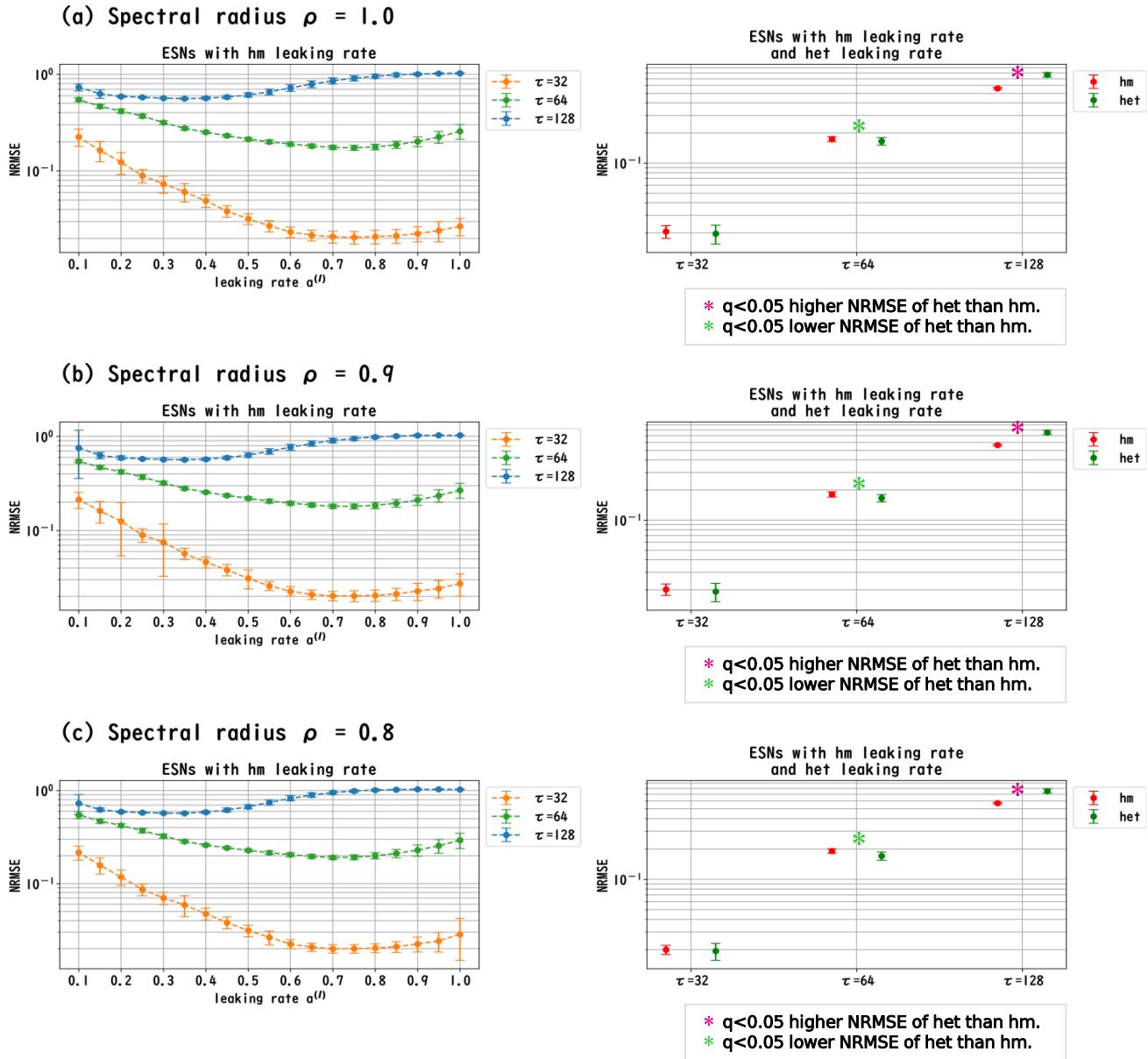
## 161 2.6 Cross-correlation

162 Cross-correlation is a method used to evaluate synchronization with delay between two time-series  
 163 signals. We used the cross-correlation  $\text{Corr}(k)$  between the dynamics of reservoir state at  $l$ -layer and  
 164  $l + 1$  layer, i.e., the time-series  $\overline{x^{(l)}(t)}$  and  $\overline{x^{(l+1)}(t - k)}$ , where  $k$  is the delay time and each time-series is  
 165 z-score transformed.

## 3 RESULTS

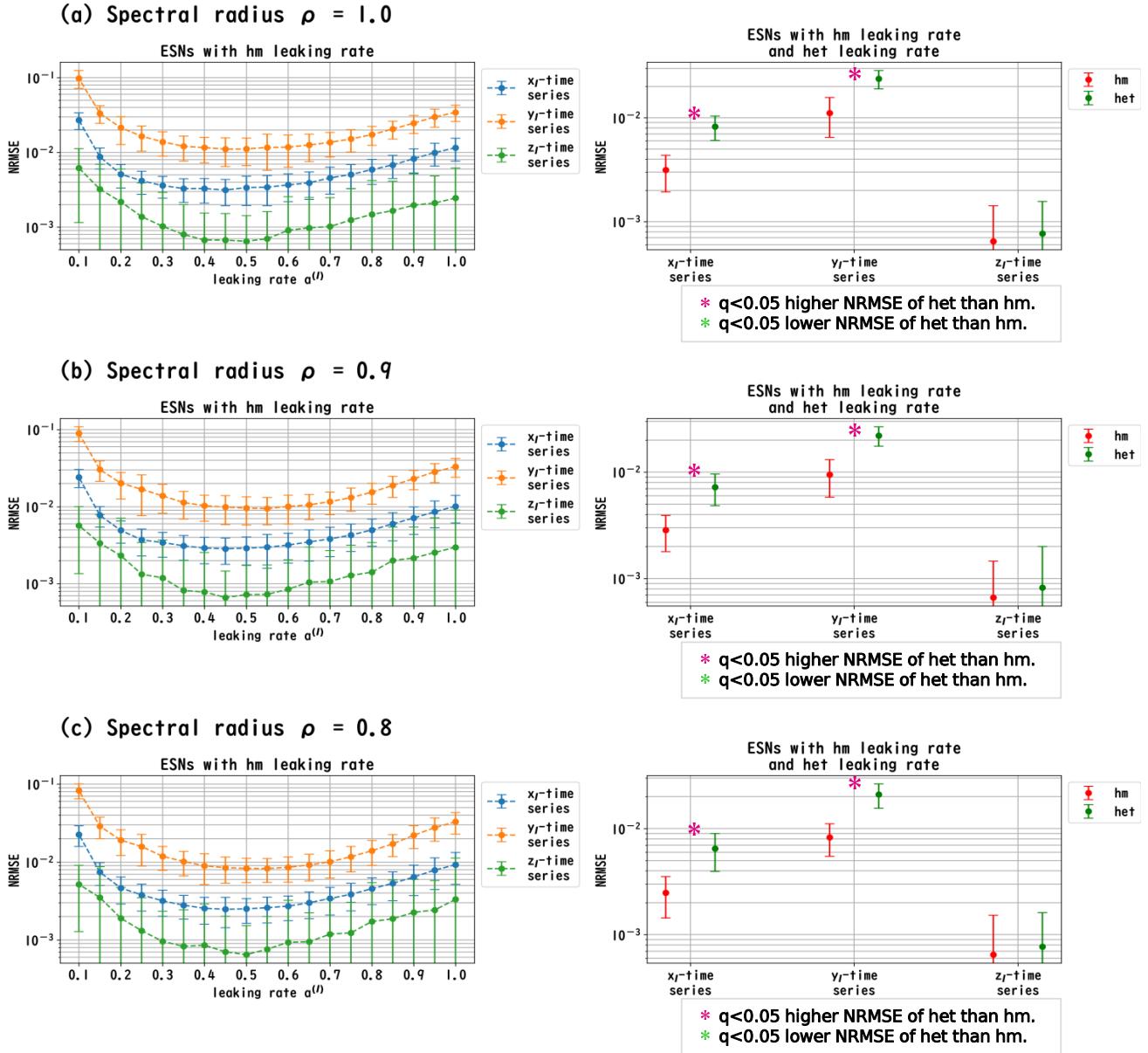
### 166 3.1 Time-series prediction task

167 We evaluated the dependence of the performance of Deep-ESNs on leaking rates  $a^{(l)}$  in the time-series  
 168 prediction tasks in the non-linear dynamical signals of the Mackey-Glass equation, Lorenz equation, and  
 169 Rössler equation. Figures 3, 4, and 5 show the results of the evaluation of homogeneous and heterogeneous  
 170 models in each time-series prediction task. In the homogeneous model, the leaking rate was set to be  
 171  $a^{(l)} = 1.0, 0.9, \dots, 0.1$  for all layers in common. In the heterogeneous model, the leaking rate decreased  
 172 by 0.1 in each layer, from 1.0 to 0.1. In the left panels of Figs. 3, 4, and 5, the dependences of NRMSE  
 173 on the leaking rate in the homogeneous model for the spectral radii  $\rho = 0.8, 0.9, 1.0$  are shown. The  
 174 results demonstrated that the profiles of NRMSEs in all tasks and the spectral radii exhibited a U-shape  
 175 in response to the leaking rate, indicating the presence of an optimal leaking rate for the prediction tasks.  
 176 Furthermore, the right panels of Figs. 3, 4, and 5 demonstrate a comparison of the NRMSEs of the most  
 177 superior cases in the homogeneous model, which are obtained from the profile of dependence on the  
 178 leaking rate in the left panels, with the heterogeneous model leaking rate. The results demonstrated that  
 179 the NRMSE of the heterogeneous model is significantly low (based on paired- $t$  test using Bonferroni false  
 180 discovery rate with  $q < 0.05$ ) only for the Mackey-Glass time series ( $\tau = 64$ ) across all spectral radii.



**Figure 3.** Prediction performance in the Mackey-Glass time series. (Left panel) Dependence of the performance of Deep-ESNs on leaking rates ( $a^{(l)} = 1.0, 0.9, \dots, 0.1$  for all layers in common) in the homogeneous (hm) models with spectral radii  $\rho = 1.0$  (a), 0.9 (b), and 0.8 (c). The profiles of NRMSEs in all spectral radii exhibited a U-shape against the leaking rate, indicating the presence of an optimal leaking rate for the prediction tasks. (Right panel) NRMSEs for most superior cases in the homogeneous model (corresponding to the cases with minimum NRMSEs shown in the left panel) and for the cases of the heterogeneous (het) model where the leaking rate was set to decrease by 0.1 in each layer, from 1.0 to 0.1. The NRMSE of the heterogeneous model is low (based on paired-*t* test using Bonferroni false discovery rate with  $q < 0.05$  ( $p < 0.05/9$ )) only for the Mackey-Glass time series ( $\tau = 64$ ) throughout all spectral radii.

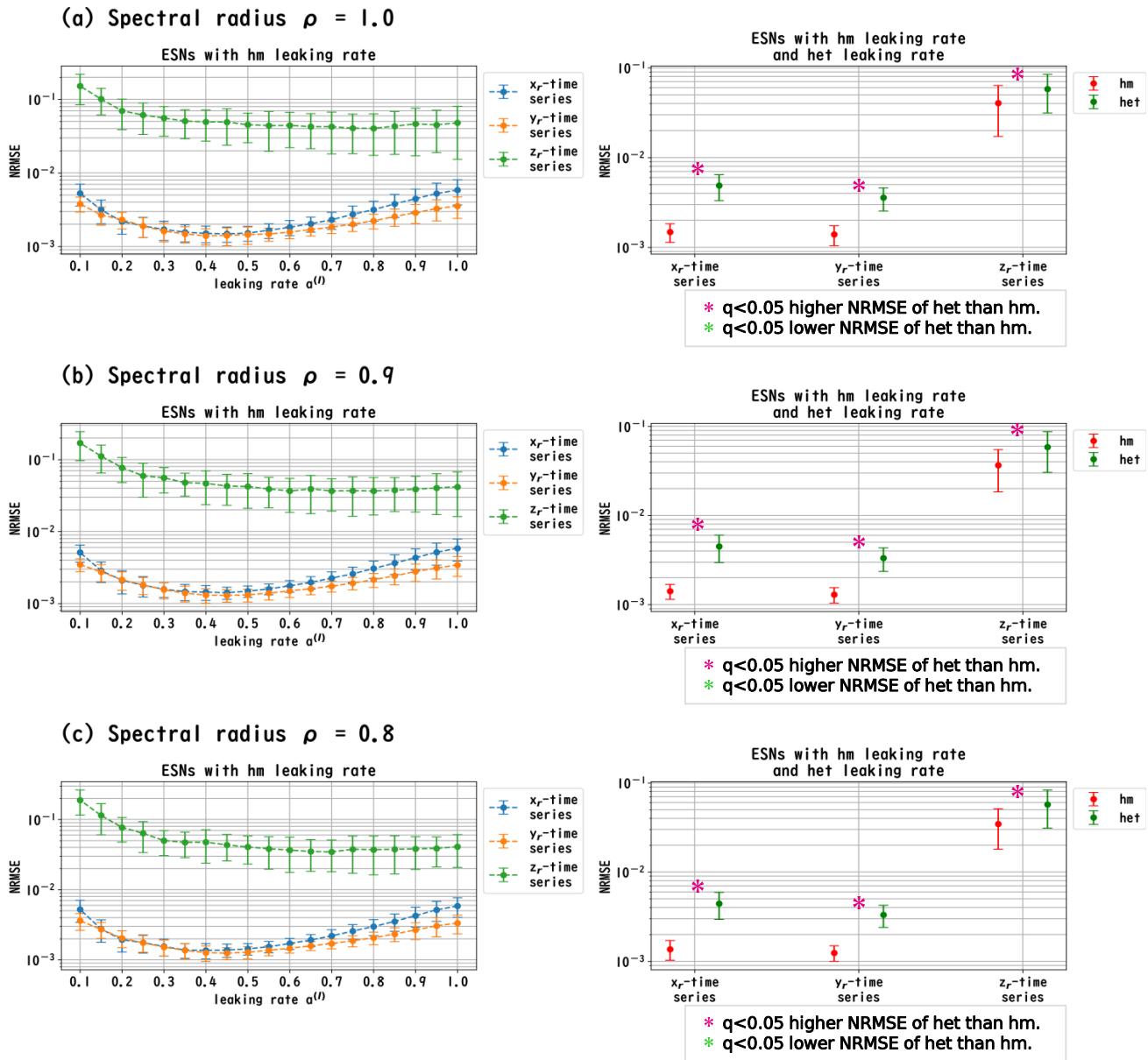
181 This tendency implied that the heterogeneous models can precisely respond to strong multi-temporal-  
 182 scale dynamics, similar to the Mackey-Glass time series ( $\tau = 64$ ) exhibited wide multi-temporal-scale  
 183 components in this dynamics (see section 1 in “Supplemental Material”).



**Figure 4.** Prediction performance in the Lorenz time series. (Left panel) Dependence of the performance of Deep-ESNs on leaking rates ( $\alpha^{(l)} = 1.0, 0.9, \dots, 0.1$  for all layers in common) in the homogeneous (hm) models with spectral radii  $\rho = 1.0$  (a), 0.9 (b), and 0.8 (c). The profiles of NRMSEs in all spectral radii exhibited a U-shape in response to the leaking rate, indicating the presence of an optimal leaking rate for the prediction tasks. (Right panel) NRMSEs for most superior cases in the homogeneous model (corresponding to the cases with minimum NRMSEs shown in the left panel) and for the cases of the heterogeneous (het) model where the leaking rate was set to decrease by 0.1 in each layer, from 1.0 to 0.1. The NRMSE of the homogeneous model is low (based on paired-*t* test using Bonferroni false discovery rate with  $q < 0.05$  ( $p < 0.05/9$ )) for the Lorenz time series ( $x_l(t)$  and  $y_l(t)$ ) throughout all spectral radii.

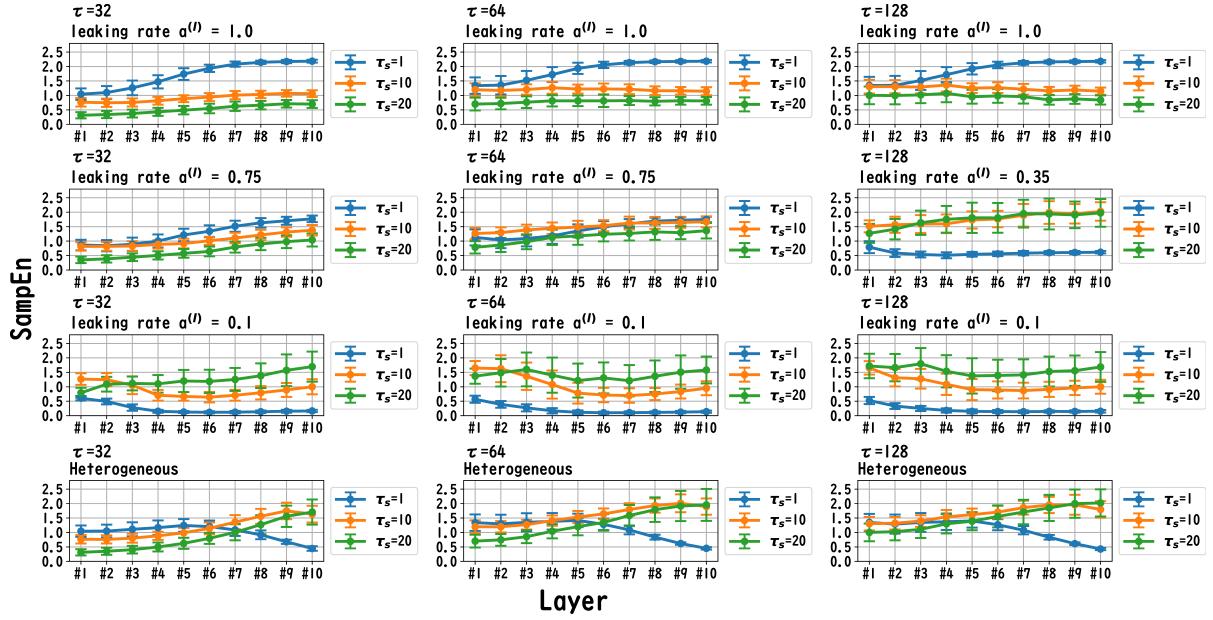
### 184 3.2 Multiscale entropy analysis

185 The MSE analysis was used to investigate the behavior of the dynamics of each layer of the reservoir.  
 186 Figures 6, 7 and 8 show the results of the MSE analysis for tasks Mackey–Glass, Lorenz, and Rössler  
 187 respectively. The temporal scales  $\tau_s$  were set to 1, 10, and 20. In the case of the homogeneous model, a



**Figure 5.** Prediction performance in the Rössler time series. (Left panel) Dependence of the performance of Deep-ESNs on leaking rates ( $a^{(l)} = 1.0, 0.9, \dots, 0.1$  for all layers in common) in the homogeneous (hm) models with spectral radii  $\rho = 1.0$  (a), 0.9 (b), and 0.8 (c). The profiles of NRMSEs in all spectral radii exhibited a U-shape against the leaking rate, indicating the presence of an optimal leaking rate for the prediction tasks. (Right panel) NRMSEs for most superior cases in the homogeneous model (corresponding to the cases with minimum NRMSEs shown in the left panel) and for the cases of the heterogeneous (het) model where the leaking rate was set to decrease by 0.1 in each layer, from 1.0 to 0.1. The NRMSE of the homogeneous model is low (based on paired-*t* test using Bonferroni false discovery rate with  $q < 0.05$  ( $p < 0.05/9$ )) for the Rössler time series ( $x_1(t)$ ,  $y_1(t)$  and  $z_1(t)$ ) throughout all spectral radii.

188 trend towards reduced complexity on the fast time scale ( $\tau_s = 1$ ) was observed as the leaking rate was  
 189 reduced. The complexity on slow time scales ( $\tau_s = 10, 20$ ) may vary through layers or be nearly constant,  
 190 depending on the time scale of the task and leaking rate. In the heterogeneous model case, the complexity



**Figure 6.** Multiscale entropy (MSE) analysis of reservoir dynamics (temporal scale  $\tau_s = 1, 10, 20$ ) with spectral radius  $\rho = 1.0$  in the case of Mackey–Glass task. The first, second, and third columns list the cases with Mackey-Glass delay time constants  $\tau = 32, 64, 128$ , respectively. The first, second, and third columns correspond to maximum value of the leaking rate ( $a^{(l)} = 1.0$ ), leaking rate when the NRMSE was most outstanding, and the minimum leaking rate ( $a^{(l)} = 0.1$ ), respectively. In the case of the homogeneous model, a trend towards reduced complexity on the fast time scale ( $\tau_s = 1$ ) was observed as the leaking rate was reduced. The complexity on slow time scales ( $\tau_s = 10, 20$ ) tends to vary through layers or to be almost constant depending on the time scale of the task and the leaking rate. In the heterogeneous model case, the complexity on fast time scales ( $\tau_s = 1$ ) tends to decrease according to the layer depth. The complexity of the slow time scale ( $\tau_s = 10, 20$ ) tends to increase layer by layer.

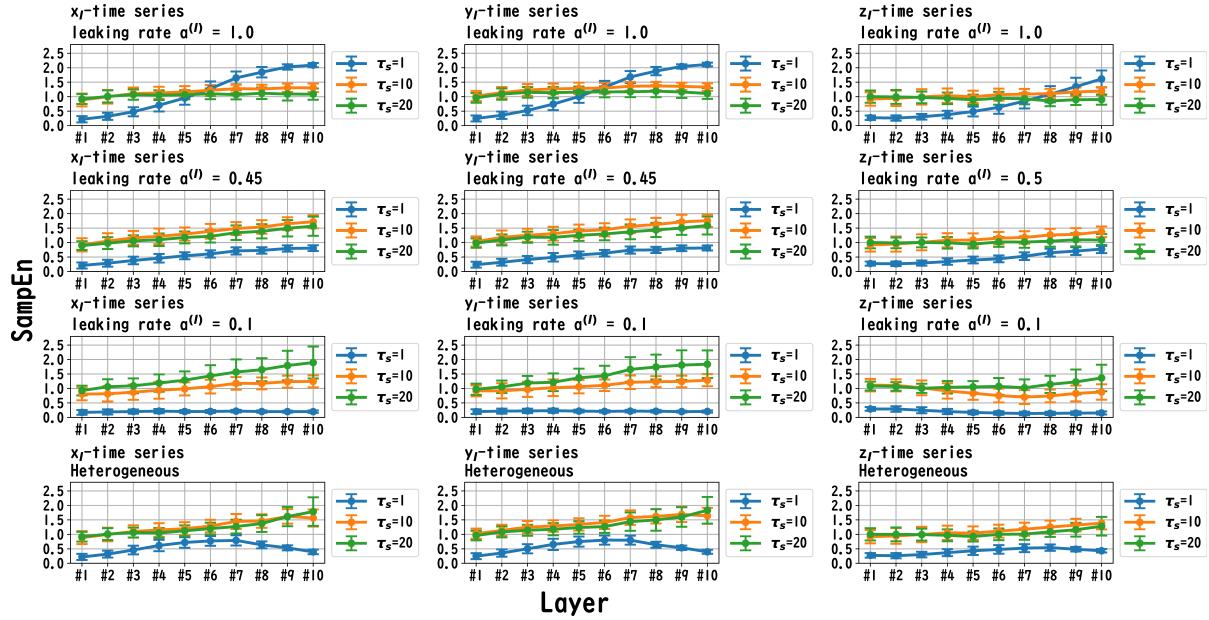
191 on fast time scales ( $\tau_s = 1$ ) tends to decrease with layer depth. The complexity of the slow time scale  
 192 ( $\tau_s = 10, 20$ ) tends to increase layer by layer.

### 193 3.3 Cross-correlation

194 The signal transmissions in the dynamics of the reservoir state among layers were evaluated using cross-  
 195 correlation analysis. Figure 9 illustrates the dynamics of the reservoir states between adjacent reservoir  
 196 layers ( $l$ - and  $l + 1$  th layers):  $|\text{Corr}(k)|$  in the case with the heterogeneous model (spectral radius  $\rho = 1.0$ )  
 197 for Mackey-Glass ( $\tau = 64$ ) task. Here, this setting corresponds to the highest accuracy in Fig. 3 (a) for  
 198 Mackey-Glass ( $\tau = 64$ ) task. As the results,  $|\text{Corr}(k)|$  maximized at the positive lag ( $k \geq 0$ ) in the specific  
 199 between-layers (#3&#4, #4&#5, #5&#6, #6&#7, #8&#9, #9&#10), i.e., the delays of signal transmission  
 200 from  $l$ -th layer to  $l + 1$ -th layer exist. To evaluate this tendency against the different tasks used in this  
 201 study, Fig. 10 represented the  $k$  values where  $|\text{Corr}(k)|$  was maximized for adjacent pair-wise layers. The  
 202 results showed that a major part of pair-wise layers in all tasks exhibits positive  $k$  values ( $\geq 1$ ), that is,  
 203 delays of signal transmission between adjacent layers exist. This helps retain past information.

## 4 DISCUSSION

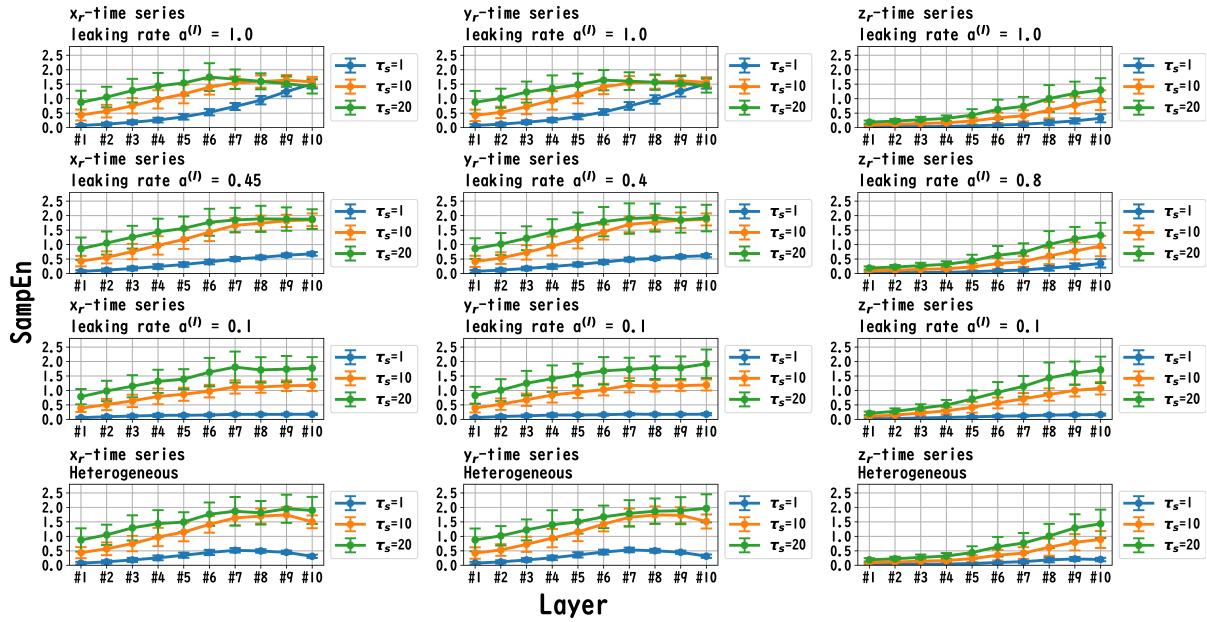
204 In the discussion, we first recapitulate the key findings derived from the results. Our investigation was  
 205 primarily aimed at understanding the mechanism of the Deep-ESN function enhancement and guidelines  
 206 by adjusting the leaking rate  $a^{(l)}$ . Across three different experiments, we validated our hypothesis that



**Figure 7.** The MSE analysis of reservoir dynamics (temporal scale  $\tau_s = 1, 10, 20$ ) with spectral radius  $\rho = 1.0$  in the case of Lorenz task. The first, second, and third columns show the cases with Lorenz task  $x_l(t)$ ,  $y_l(t)$ ,  $z_l(t)$ , respectively. The first, second, and third columns correspond to maximum value of the leaking rate ( $a^{(l)} = 1.0$ ), leaking rate when the NRMSE was the most outstanding, and minimum value of the leaking rate ( $a^{(l)} = 0.1$ ), respectively. In the case of the homogeneous model, a trend towards reduced complexity on the fast time scale ( $\tau_s = 1$ ) was observed as the leaking rate was reduced. The complexity on slow time scales ( $\tau_s = 10, 20$ ) tends to vary through layers or to be almost constant, depending on the  $x_l/y_l/z_l$  and the leaking rate. In the heterogeneous model case, the complexity on fast time scales ( $\tau_s = 1$ ) tends to decrease according to the layer depth. The complexity of the slow time scale ( $\tau_s = 10, 20$ ) tends to increase layer by layer.

the evaluation for dependence of dynamical response in the multi-layered reservoir on adjusting leaking rate would provide insights into achieving a guideline for optimizing the hyper-parameters of the Deep-ESN. Specifically, the first experiment (time-series prediction tasks) with the homogeneous model showed that the profiles of NRMSEs in all tasks and the spectral radii exhibited a U-shape against the leaking rate, indicating an optimal leaking rate for each prediction task. The heterogeneous model was effective for the Mackey-Glass time series ( $\tau = 64$ ), which contains wide multi-temporal-scale components in its dynamics. The second experiment (MSE analysis) with the homogeneous model indicated that the complexity associated with fast time scales tends to decrease with the leaking rate, while that on slow time scales tends to vary or remain almost constant according to the number of layers, depending on the task timescale and leaking rate. In heterogeneous models, fast time scale complexity tends to decrease with layer depth, while slow time scale complexity tends to increase layer by layer. Finally, the third experiment (cross-correlation analysis) with homogeneous and heterogeneous models showed that the delay in layer-to-layer signal transmission in all tasks, which helps retain the past information.

First, we discuss the reasons for the presence of an optimum leaking rate (see right panel of Figs. 3, 4 and 5). In the single neural dynamics, as the leaking rate increases, the dynamics of neuron becomes faster. Based on this effect, the complexity of fast scale dynamics in the case with a large leaking rate increases further, especially, in deep layers through the multiple-layered propagation (see the tendency of SampEn against increasing the leaking rate in the case with  $\tau_s = 1$  in Figs. 6, 7 and 8). This tendency was observed in the homogeneous case with a large leaking rate and heterogeneous case. Meanwhile, the

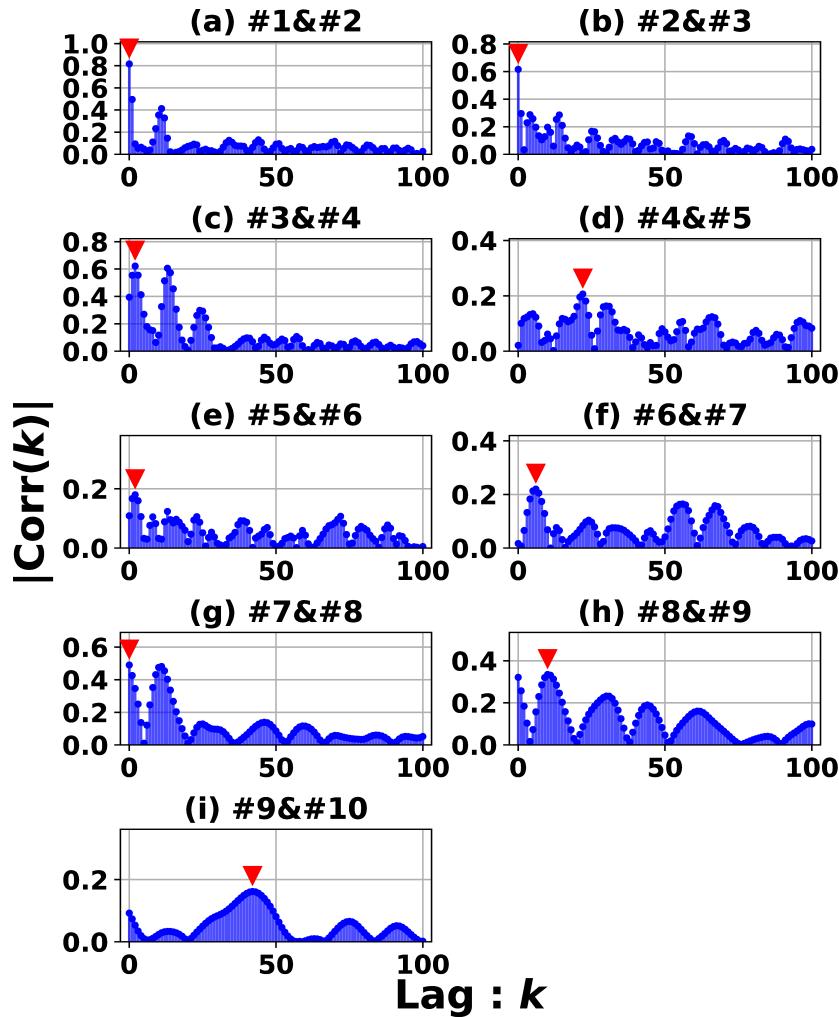


**Figure 8.** MSE analysis of reservoir dynamics (temporal scale  $\tau_s = 1, 10, 20$ ) with spectral radius  $\rho = 1.0$  in the case for Rössler task. The first, second, and third columns show the cases with Rössler task  $x_r(t)$ ,  $y_r(t)$ ,  $z_r(t)$ , respectively. The first, second, and third columns correspond to maximum value of the leaking rate ( $a^{(l)} = 1.0$ ), leaking rate when the NRMSE was the most outstanding, and minimum value of the leaking rate ( $a^{(l)} = 0.1$ ), respectively. In the homogeneous case with task  $x_r(t)$ ,  $y_r(t)$ , in the case with high leaking rate, the complexity on fast time scales ( $\tau_s = 1$ ) increases with the layer depths. The complexity on slow time scales ( $\tau_s = 10, 20$ ) tends to vary through layers or to be almost constant, depending on the  $x_r/y_r/z_r$  and the leaking rate. In the heterogeneous model case with  $x_r(t)$  and  $y_r(t)$ , the complexity on fast time scales ( $\tau_s = 1$ ) tends to decrease with layer depth. The complexity of the slow time scale ( $\tau_s = 10, 20$ ) tends to increase layer by layer.

complexity at slow time scales exhibited a diverse layer-specific SampEn profile, depending on leaking rate and task (see the SampEn of  $\tau_s = 10, 20$  in Figs. 6, 7 and 8). This tendency may be attributed to the complex interactions in the layer-to-layer signal propagation. Generally, to achieve high performance of ESN, the representation of complex desired signals requires the combination of diverse time-scale dynamical responses in the readout (Tanaka et al., 2022). The layer-specific time-scale dynamic response obtained using the Deep-ESN can satisfy this requirement, and this may be achieved by adjusting the leaking rate.

Next, we discuss the structural effectiveness of the Deep-ESN. The hierarchical structure makes delay in the layer-to-layer signal transmission (see Figs. 9 and 10); consequently, this delay helps retain the information used in the past, specifically that resembling the queue structure. This characteristic contributes significantly to good performance (Gallicchio et al., 2017) of Deep-ESN.

Although this study revealed the presence of an optimal leaking rate, the grid-search is still required for the concrete set for the leaking rate in accordance with the tasks. This facilitates the need to develop an approach to find the optimal leaking rate based on the dynamic characteristics used in this study. This point must be dealt with in the future works.



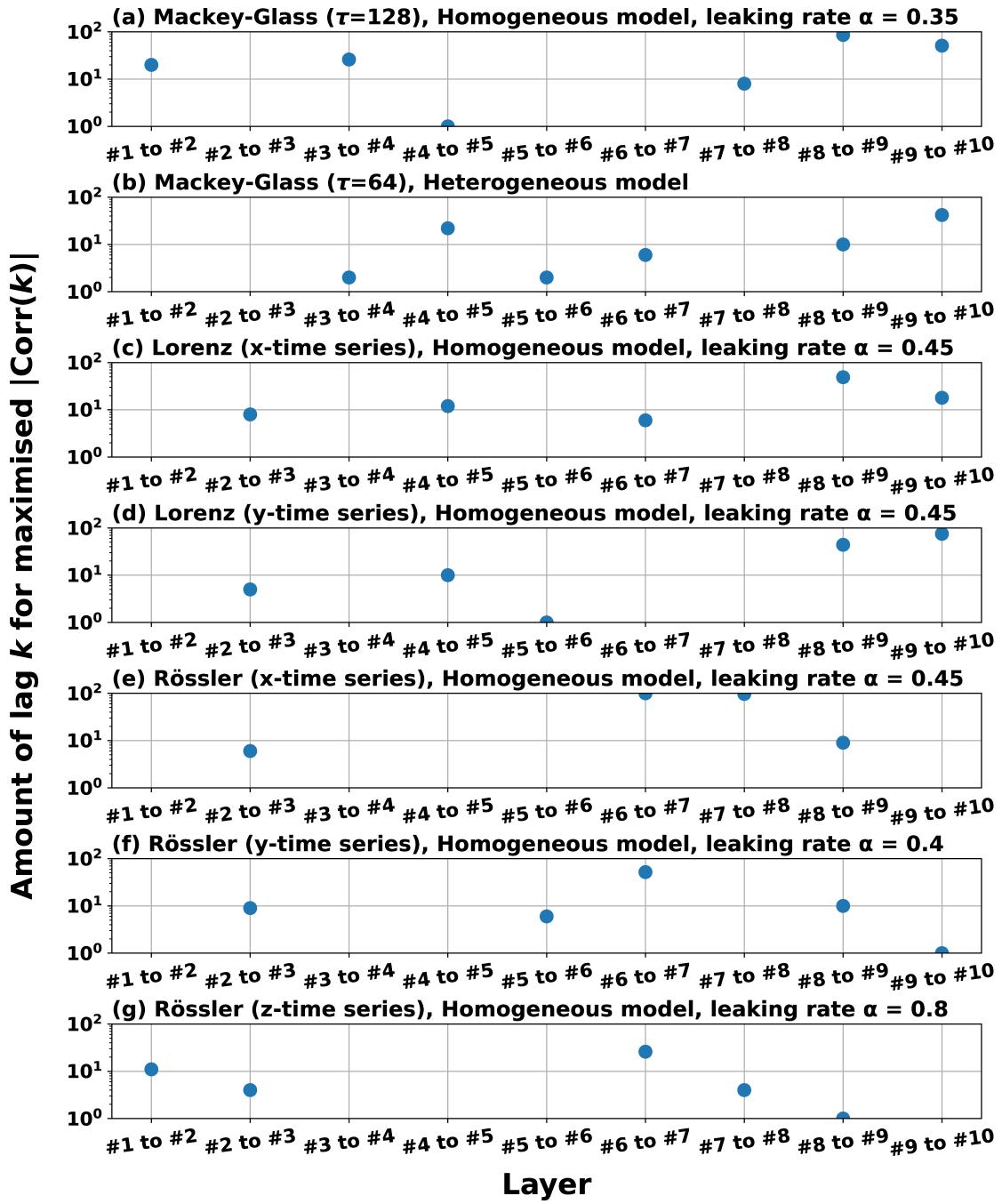
**Figure 9.** Absolute value of cross-correlation for the dynamics of the reservoir states between adjacent reservoir layers ( $l$ - and  $l+1$  th layers):  $|\text{Corr}(k)|$  in the case with the heterogeneous model (spectral radius  $\rho = 1.0$ ) for Mackey-Glass ( $\tau = 64$ ) task. Here, this setting corresponds to the highest accuracy in Fig. 3 (a) for Mackey-Glass ( $\tau = 64$ ) task. Lag  $k$  where the maximized  $|\text{Corr}(k)|$  achieves (represented by red arrow) shows the delay in signal transmission from  $l$ -th layer to  $l+1$ -th layer. In the panels of (c), (d), (e), (f), (h) and (i), peaks are  $k \geq 1$ , hence reservoir dynamics are delayed between layers.

## 5 CONCLUSIONS

As conclusions, through MSE and cross-correlation analyses, this study revealed the presence of optimal leaking rate to represent the complex desired signal and the mechanism to retain the past information in Deep-ESN. Despite some limitations, these findings contribute to establishing optimum design guidelines for setting hyper-parameters of the Deep-ESN.

## AUTHOR CONTRIBUTIONS

SI, SN and HN conceived the methods and wrote the main manuscript, and prepared all figures. SI, SN, HN, EW, and TI analyzed and discussed the results. SI and SN conducted the experiments. All authors contributed to the manuscript revision and have read and approved the submitted version.



**Figure 10.** Lag  $k$  values where  $|\text{Corr}(k)|$  was maximized for adjacent pair-wise layers in the cases with parameter setting (homogeneous/heterogeneous models) to achieve the superior NRMSE in the time series prediction task (corresponding to the right panel of (a) in Figs. 3, 4 and 5). The values are not plotted when the reservoir dynamics of the  $l$  and  $l + 1$  layers are zero-lag synchronized (peak of correlation is  $k = 0$ ), but plotted when the  $l + 1$  layer is delayed against the  $l$  layer (peak of correlation is  $k \geq 1$ ). The presence of between four and six plot points in each subplot confirms that reservoir dynamics delays occur in the  $l$  and  $l + 1$  layers.

## FUNDING

248 This work was supported by JSPS KAKENHI for the Grant-in-Aid for Scientific Research (C) (Grant  
 249 Number JP22K12183) to SN. The funder was not involved in the study design, collection, analysis,  
 250 interpretation of data, the writing of this article, or the decision to submit it for publication.

## CONFLICT OF INTEREST

251 The authors declare that the research was conducted in the absence of any commercial or financial  
252 relationships that could be construed as potential conflicts of interest.

## REFERENCES

- 253 Lukoševičius M, Jaeger H. Reservoir computing approaches to recurrent neural network training.  
254 *Computer Science Review* **3** (2009) 127–149.
- 255 Tanaka G, Yamane T, Héroux JB, Nakane R, Kanazawa N, Takeda S, et al. Recent advances in physical  
256 reservoir computing: A review. *Neural Networks* **115** (2019) 100–123.
- 257 Gallicchio C, Micheli A. Deep reservoir computing. *Reservoir Computing: Theory, Physical  
258 Implementations, and Applications* (2021) 77–95.
- 259 Jaeger H. The “echo state” approach to analysing and training recurrent neural networks-with an erratum  
260 note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical  
261 Report* **148** (2001) 13.
- 262 Jaeger H. Echo state network. *scholarpedia* **2** (2007) 2330.
- 263 Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks.  
264 *Neural computation* **1** (1989) 270–280.
- 265 Werbos PJ. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE* **78**  
266 (1990) 1550–1560.
- 267 Gallicchio C, Micheli A, Pedrelli L. Comparison between deepesns and gated rnns on multivariate time-  
268 series prediction. *arXiv preprint arXiv:1812.11527* (2018).
- 269 Salehinejad H, Sankar S, Barfett J, Colak E, Valaee S. Recent advances in recurrent neural networks.  
270 *arXiv preprint arXiv:1801.01078* (2017).
- 271 Sakemi Y, Nobukawa S, Matsuki T, Morie T, Aihara K. Learning reservoir dynamics with temporal  
272 self-modulation. *Communications Physics* **7** (2024) 29.
- 273 Deng L, Yu D, Platt J. Scalable stacking and learning for building deep architectures. *2012  
274 IEEE International conference on Acoustics, speech and signal processing (ICASSP)* (IEEE) (2012),  
275 2133–2136.
- 276 Gallicchio C, Micheli A, Pedrelli L. Deep reservoir computing: A critical experimental analysis.  
277 *Neurocomputing* **268** (2017) 87–99.
- 278 Malik ZK, Hussain A, Wu QJ. Multilayered echo state machine: A novel architecture and algorithm.  
279 *IEEE Transactions on cybernetics* **47** (2016) 946–959.
- 280 Tchakoucht TA, Ezziyyani M. Multilayered echo-state machine: A novel architecture for efficient  
281 intrusion detection. *IEEE Access* **6** (2018) 72458–72468.
- 282 Long J, Zhang S, Li C. Evolving deep echo state networks for intelligent fault diagnosis. *IEEE  
283 Transactions on Industrial Informatics* **16** (2019) 4928–4937.
- 284 Gallicchio C, Micheli A. Richness of deep echo state network dynamics. *International work-conference  
285 on artificial neural networks* (Springer) (2019), 480–491.

- 286 Kanda K, Nobukawa S. Feature extraction mechanism for each layer of deep echo state network. 2022  
287 *International Conference on Emerging Techniques in Computational Intelligence (ICETCI)* (IEEE)  
288 (2022), 65–70.
- 289 Costa M, Goldberger AL, Peng CK. Multiscale entropy analysis of complex physiologic time series.  
290 *Physical review letters* **89** (2002) 068102.
- 291 Venkatasubramanian V, Schattler H, Zaborsky J. Dynamics of large constrained nonlinear systems-a  
292 taxonomy theory [power system stability]. *Proceedings of the IEEE* **83** (1995) 1530–1561.
- 293 Yan B, He S. Dynamics and complexity analysis of the conformable fractional-order two-machine  
294 interconnected power system. *Mathematical Methods in the Applied Sciences* **44** (2021) 2439–2454.
- 295 Bhandari A. Wavelets based multi-scale analysis of select global equity returns. *Theoretical and Applied  
Economics* **24** (2017) 613.
- 297 Chen S, Shang P. Financial time series analysis using the relation between mpe and mwpe. *Physica A:  
Statistical Mechanics and its Applications* **537** (2020) 122716.
- 299 Lukoševičius M, Uselis A. Efficient cross-validation of echo state networks. *Artificial Neural Networks  
and Machine Learning—ICANN 2019: Workshop and Special Sessions: 28th International Conference  
on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings* 28 (Springer)  
300 (2019), 121–133.
- 302 Adeleke OA. Echo-state networks for network traffic prediction. *2019 IEEE 10th annual information  
technology, electronics and mobile communication conference (IEMCON)* (IEEE) (2019), 0202–0206.
- 305 Viehweg J, Worthmann K, Mäder P. Parameterizing echo state networks for multi-step time series  
306 prediction. *Neurocomputing* **522** (2023) 214–228.
- 307 Bai YT, Jia W, Jin XB, Su TL, Kong JL, Shi ZG. Nonstationary time series prediction based on deep echo  
308 state network tuned by bayesian optimization. *Mathematics* **11** (2023) 1503.
- 309 Humeau-Heurtier A. The multiscale entropy algorithm and its variants: A review. *Entropy* **17** (2015)  
310 3110–3123.
- 311 Schrauwen B, Defour J, Verstraeten D, Van Campenhout J. The introduction of time-scales in reservoir  
312 computing, applied to isolated digits recognition. *Artificial Neural Networks—ICANN 2007: 17th  
313 International Conference, Porto, Portugal, September 9–13, 2007, Proceedings, Part I* 17 (Springer)  
314 (2007), 471–479.
- 315 Jaeger H, Lukoševičius M, Popovici D, Siewert U. Optimization and applications of echo state networks  
316 with leaky-integrator neurons. *Neural networks* **20** (2007) 335–352.
- 317 Inoue S, Nobukawa S, Nishimura H, Watanabe E, Isokawa T. Mechanism for enhancement of functionality  
318 in deep echo state network by optimizing leaking rate. *2023 International Conference on Emerging  
319 Techniques in Computational Intelligence (ICETCI)* (2023).
- 320 Glass L, Mackey M. Mackey-glass equation. *Scholarpedia* **5** (2010) 6908.
- 321 Manneville P, Pomeau Y. Intermittency and the lorenz model. *Physics Letters A* **75** (1979) 1–2.
- 322 Rössler OE. The chaotic hierarchy. *Zeitschrift für Naturforschung A* **38** (1983) 788–801.
- 323 Tanaka G, Matsumori T, Yoshida H, Aihara K. Reservoir computing with diverse timescales for prediction  
324 of multiscale dynamics. *Physical Review Research* **4** (2022) L032014.