

Temporal Convolution Derived Multi-Layered Reservoir Computing

Johannes Viehweg^{a,1,*}, Dominik Walther^{a,1}, Patrick Mäder^{a,b,1,2}

^a*Helmholtzplatz 5, 98693 Ilmenau, Germany*

^b*Philosophenweg 16, 07743 Jena, Germany*

Abstract

The prediction of time series is a challenging task relevant in such diverse applications as analyzing financial data, forecasting flow dynamics or understanding biological processes. Especially chaotic time series that depend on a long history pose an exceptionally difficult problem. While machine learning has shown to be a promising approach for predicting such time series, it either demands long training time and much training data when using deep Recurrent Neural Networks. Alternative, when using a Reservoir Computing approach it comes with high uncertainty and typically a high number of random initializations and extensive hyper-parameter tuning. In this paper, we focus on the Reservoir Computing approach and propose a new mapping of input data into the reservoir's state space. Furthermore, we incorporate this method in two novel network architectures increasing parallelizability, depth and predictive capabilities of the neural network while reducing the dependence on randomness. For the evaluation, we approximate a set of time series from the Mackey-Glass equation, inhabiting non-chaotic as well as chaotic behavior as well as the SantaFe Laser dataset and compare our approaches in regard to their predictive capabilities to Echo State Networks, Autoencoder connected Echo State Networks and Gated Recurrent Units. For the chaotic time series, we observe an error reduction of up to 85.45% compared to Echo State Networks and 90.72% compared to Gated Recurrent Units. Furthermore, we also observe tremendous improvements for non-chaotic time series of up to 99.99% in contrast to the existing approaches.

1. Introduction

Time series prediction plays a crucial role in various fields, encompassing biologically inspired systems such as population dynamics [22] and financial stock market metrics [32] up to climate and global flow models [58, 36]. However, predicting time series is a challenging task, particularly when dealing with chaotic, non-periodic dynamics. In the pursuit of accurate predictions, machine learning techniques, particularly Recurrent Neural Networks (RNNs) such as Long-Short Term Memory Networks (LSTM) [15] and the subsequent Gated Recurrent Units (GRU) [7], have emerged as a viable approach for predicting dynamical systems [45] but suffer from their demand of time and resources for the training [50]. Feed-forward-based Transformer architectures [47] allowed to overcome these problems, e.g., by parallelizing the training process, especially in the area of natural language processing. Another approach to be named here is the idea of convolutions in time, namely Temporal Convolutional Networks (TCN) [2].

However, they have not been found superior when predicting chaotic time series, e.g., shown in [39, 40, 10].

In recent years, Reservoir Computing (RC) [48], especially Echo State Networks (ESN) [17] with only three layers, namely the input layer, a single recurrent reservoir and the output layer, have garnered significant attention due to their effectiveness in time series prediction and their low resource demand in training and prediction phase [38, 27]. A problem of the approach is the high variance between training results

*Corresponding author: johannes.viehweg@tu-ilmenau.de

¹Johannes Viehweg, Dominik Walther, and Patrick Mäder are with the Data-intensive Systems and Visualization group (dAISY), Technische Universität Ilmenau, 98693 Ilmenau, Germany.

²Patrick Mäder is with the Faculty of Biological Sciences, Friedrich Schiller University, 07743 Jena, Germany.

due to the weights being chosen randomly and, except for the ones of the output layer, not being further optimized. This can lead to a heterogeneous situation with some trained networks being well fitted and others not resembling the target in a sufficient manner while being trained with the same hyper-parameters but different drawn weights. One approach of handling this problem is the adaptation of the structure of the reservoir to a specific problem [42] or utilizing additional methods of regularization [56]. Several previous studies propose the concept of multiple reservoir layers rather than using a single one in order to improve ESN’s prediction performance, e.g. [33, 34]. However, we argue that despite improved prediction performance for luckily drawn configuration, this approach further increases the randomness due to multiple randomly chosen reservoirs. Other recent works in the field of RC deliver a new theoretical foundation and explanation of the approach [4] and argue that their effectiveness could be explained by showing their connection to Vector Autoregressive Average (VAR) for the identity activation function. This allows for a predictor without the need to rely on favorable randomness for the weights. It also allows for an analysis of the ESN, which would be unfeasible for the basic case. According to [12], an optimized RC network is hereby equivalent to non-linear VAR, which allows for construction of a predictor without randomness. This understanding is used to propose novel architectures such as the Next Generation Reservoir Computing (NG-RC) [12] reducing the needed randomness. In conclusion, previous studies observe instabilities for long delays, i.e., mapping a dependence of a long time interval of the time series into the network’s reservoir, and a dependence on randomness in the initialization of the network.

We adapt the theoretical foundation of reservoir computing discussed above and propose a new architecture and mapping to overcome the discussed problems with randomness and long delays both encouraging an unstable autoregressive prediction resulting in a constant predicted value or an non-computable error. We show a first stable auto-regressive prediction, staying close to the ground truth, for multiple Lyapunov times for those time series while exploiting the benefits of parallelizability offered by its inspiration, the TCNs and further increasing the predictive capabilities by coupling it with an Extreme Learning Machine (ELM) of [16] into a state space which we argue benefits from being in the field of so called hyper-dimensional computing [20] similar to [11].

Our research aims to answer the following questions:

1. How can the problem of randomness be mitigated without losing the ability to handle time series with an importance of memory?
2. Is the adaptation of the idea of Temporal Convolutional Networks suitable for the handling of chaotic time series?
3. Is the introduction of randomness beneficial in the case of chaos?

To answer those we evaluate our approach on multiple configurations for well-established benchmark datasets. Those are chosen for assessing the predictive capabilities of algorithms in chaotic, history-dependent systems. We generated them based on the the Mackey-Glass equation [31, 13], which has been widely used in studies such as [3, 55, 25, 24, 35] as well as utilizing the SantaFe Laser dataset [53]. This paper is structured as follows: In Section 2 we discuss the different baseline varieties of RC and ESN used in this work. Our proposal and background to it are explained in Section 3. The baseline of AEESN and GRUs as well as the data is shown in Section 4 together with the evaluation and results of our tests. We close the paper with a discussion of its limitation in Section 5 and conclude in Section 6.

2. Reservoir Computing

The core principle of RC involves reducing the complexity of training by focusing on a limited set of weights. These weights are crucial for linearly mapping a high-dimensional state to a specific output in a singular computation step. Within this framework, an autoregressive model for time series prediction is established over a test interval of length S_P , following a training and initialization phase of lengths S_T and S_I , respectively. This model handles inputs x and targeted outputs y , for the single step prediction as $x^{(t'+1)} = y^{(t')}$, $t' \geq S_I + S_T$. In case of the multi step prediction and an computed output \hat{y} we assume from this point forward the equation $x^{(t'+1)} = \hat{y}^{(t')}$.

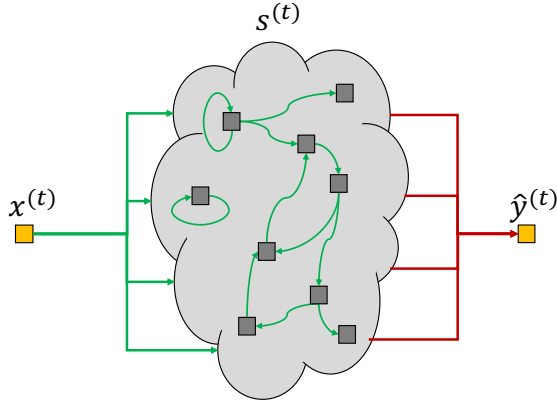


Figure 1: Basic architecture of the ESN, where green arrows (\rightarrow) refer to the randomly initialized set of connections from the input $x^{(t)}$ into the reservoir and red arrows (\rightarrow) refer to the trained mapping from the reservoir to the output $\hat{y}^{(t)}$

2.1. Echo State Networks

In this paper, our focus is on the extensively studied ESN architecture as most well known and used example of RC. In its basic form, depicted in Figure 1, the ESN consists of three layers with respective weights $W^{\text{in}} \in \mathbb{R}^{N^r \times N^{\text{in}}}$, $W^r \in \mathbb{R}^{N^r \times N^r}$ and $W^{\text{out}} \in \mathbb{R}^{N^{\text{out}} \times N^r}$. Here, $N^{\text{in}} = \hat{\delta}N^{\text{out}}$ with $\hat{\delta}$ denoting an input delay, typically chosen as $\hat{\delta} = 1$, and $N^{\text{in}} \ll N^r$. The weights W^{in} and W^r are randomly generated but fixed, as suggested in [17, 29, 28, 18]. The weights are drawn from a uniform distribution: $\forall w \in W^{[\text{in}, r]}, w \sim \mathcal{U}(-\delta, \delta)$, with $\delta \in \mathbb{R}$ and a connectivity between the positions in the state space $\sigma(W^r) \in \mathbb{R}_{w^r \in [0,1]}$. Only the output mapping using the weights W^{out} is learned. The recurrent connections enable memory in the network, typically scaled by the number of neurons $N^r \in \mathbb{N}$ and the spectral radius $\rho \in \mathbb{R}_{\geq 0}$ of the recurrent weights W^r . The internal weights in the matrix W^r are scaled by the spectral radius ρ , often chosen to be slightly below 1 [19, 28]. For the RC approaches $x^{(t)} \in \mathbb{R}^{N^{\text{in}}}$ is the input at time step $t \in \mathbb{N}_{> 0}$, $s^{(t)} \in \mathbb{R}^{N^r}$ the state, and $\hat{y}^{(t)} \in \mathbb{R}^{N^{\text{out}}}$ the output as shown in Fig. 1. The state $s^{(t)}$ is calculated as follows:

$$s^{(t)} = f(W^{\text{in}}x^{(t)} + W^r s^{(t-1)}), \quad (1)$$

where $f(\cdot)$ represents a non-linear activation function $f(\cdot) = \tanh(\cdot)$.

The output weights W^{out} are the only weights that are learned. This is accomplished in a single step using ridge regression, also known as Tikhonov regularization, as proposed in [17]. The initial state is used to compute the first S_I steps, which are then discarded during the learning process. Subsequently, the targets $Y \in \mathbb{R}^{N^{\text{out}} \times S_T}$ and states $S \in \mathbb{R}^{N^r \times S_T}$ for the next S_T time steps are collected. The computation of W^{out} is done as:

$$W^{\text{out}} = Y S^T (S S^T + \beta \mathbb{I})^\dagger, \quad (2)$$

where β is the regularization coefficient, \dagger denotes the Moore-Penrose pseudoinverse, and \mathbb{I} represents the identity matrix of size $N^r \times N^r$. Alternative approaches to learning W^{out} , such as the Least Absolute Selection and Shrinkage Operator (LASSO) [46] or Support Vector Machines (SVM) [5, 8], are also viable options. Other methods of regularization are discussed in the literature, such as the introduction of noise [29] but are rarely observed to be used in application studies [49]. Therefore, we limit the regularization of the ESN to the use of β in this work. The random initialization and computation of states introduce variance in the values of the weights and predictions. Although hyper-parameter optimization can reduce the variance between initializations [49], multiple runs per task are necessary. To optimize the networks for a specific task a set of hyper-parameters must be chosen $\Omega = \{N^r, \rho, \gamma, \beta, S_T, S_I\}$. In the frame of this work we refer to this approach as Basic ESN.

2.2. Extreme Learning Machines

A method related to RC, but without recurrence and holding a network’s state, aka memory, are Extreme Learning Machines (ELM) [16]. An analogous method of learning the output layer’s weights W^{out} is used. In contrast to the Basic ESN the state is computed as:

$$s^{(t)} = f(W^{\text{in}}x^{(t)}). \quad (3)$$

Still multiple runs with different random seeds are necessary to determine the error of the prediction with certainty. Approaches of multiple layers, while used in other neural networks [43], are still an active area of research for RC [33, 30].

2.3. Next Generation Reservoir Computing

An obvious problem of the prior discussed networks is their initialization based on random distributions. This typically results in significant performance variations between differently initialized models, subsequently implying the need for multiple training runs to determine the most suitable model and its predictive capabilities. Several approaches have been proposed to consider this randomness and to offer more predictable training results, most notable Gauthier et al.’s Next Generation Reservoir Computing (NG-RC) [12]. They substitute the randomly initialized mapping into the state space by a deterministic non-linear function defined e.g. as:

$$s^{(t)} = \hat{f}(g(\#_{i=0}^{\hat{\delta}} x^{(t-i)})). \quad (4)$$

The authors propose e.g. a concatenation of $\hat{\delta}$ elements to a vector $s^{(t)} \in \mathbb{R}^{N^r \times 1}$ as $\#_{i=0}^j$. Here, $g(\cdot)$ refers to the calculation on the inputs and $\hat{f}(\cdot)$ is the activation function, chosen as identity function. Analogous to the ESN, the states $s^{(k)}$ for $0 \leq k < S_T$ are saved in a matrix for all S_T time steps of training, but due to the feed-forward nature without the need to discard an initial wash-out phase of S_I steps. The training of the output weights remains analogous to the standard ESN training (cp. Eq. 2). The novel idea proposed by NG-RC is a mapping constructed from a linear part $\mathbb{O}_{lin}^{(\cdot)}$ and a non-linear one $\mathbb{O}_{nonlin}^{(\cdot)}$. The linear part is a concatenation of all delayed inputs as

$$\mathbb{O}_{lin}^{(t)} = \#_{i=0}^{\hat{\delta}} x^{(t-i)}. \quad (5)$$

For the non-linear part, the upper right triangle matrix of the outer product is used, which is given as $(\cdot) \otimes (\cdot)$ and concatenated with the squared values of the linear part as

$$\mathbb{O}_{nonlin}^{(t)} = [\#_{i=0}^{\hat{\delta}} x^{(t-i)^2}; \mathbb{O}_{lin}^{(t)} \otimes \mathbb{O}_{lin}^{(t)}]. \quad (6)$$

For $g(\cdot)$ a mapping is assumed in the form

$$g(\#_{i=0}^{\hat{\delta}} x^{(t-i)}) = [\mathbb{O}_{lin}; \mathbb{O}_{nonlin}], \quad (7)$$

with $[\cdot; \cdot]$ being the concatenation of the given inputs.

3. Adapting Temporal Convolutions for Reservoir Computing

Temporal convolutions emerge as a significant advancement within feed-forward networks, reignited by the successes in Convolutional Neural Networks (CNNs) for image processing [23] and Transformers for language processing tasks [47]. Historically, time series analysis has been dominated by Recurrent Neural Networks (RNNs). However, recent empirical studies [2] have demonstrated the efficacy of Temporal Convolutional Networks (TCNs) in this domain, particularly in tasks such as classification [51], where they have achieved notable success. TCNs operate by processing inputs within a defined window size in a pairwise fashion to generate outputs, with the stride between inputs incrementally increasing as the network deepens. Importantly, the process of optimizing the weights assigned to each input for subsequent layers is conducted through backpropagation, ensuring effective learning and adaptation to time series data.

3.1. The Temporal Convolution Concept

Temporal Convolutions [52] combine the benefits of a feed-forward network, i.e., parallelizable computation and efficient training, with the benefits of recurrent networks, i.e., maintaining a history of previous networks states facilitating the analysis of long-term dependencies in the input data. The inputs $x^{(t)}, x^{(t-1)}$ are weighted together in the first layer, which results in one value in the state space. We refer to this as a token $\tilde{x}_k \in {}_l s_k^{(t)}$ in this work, i.e. the k -th element of the l -th layer. In the succeeding layer, two such computed tokens $\tilde{x}_{(k)}$ and $\tilde{x}_{(k-2)}$ are used, skipping $\tilde{x}_{(k-1)}$. For each layer $l \geq 1$ this implies a computation of tokens from the former layer, generalized as $\hat{x} = {}_l s^{(t)}$, in form of the use of $\hat{x}_{(k)}$ and $\hat{x}_{(k-2^{l-1})}$, respectively with $l = 0$ being the original input. Figure 2 visualizes that mechanism and shows that convolutions only have access to current and previous time steps. Hereby, Figure 2 shows a single block χ_i whereas a TCN is build from m such block, $1 \leq i \leq m$. In this direct mapping the history of the time series is free of its sequential order for $\hat{\delta}$ time steps. This allows for a parallelization of the computation, resulting in TCNs being able to exploit the benefits of feed-forward networks.

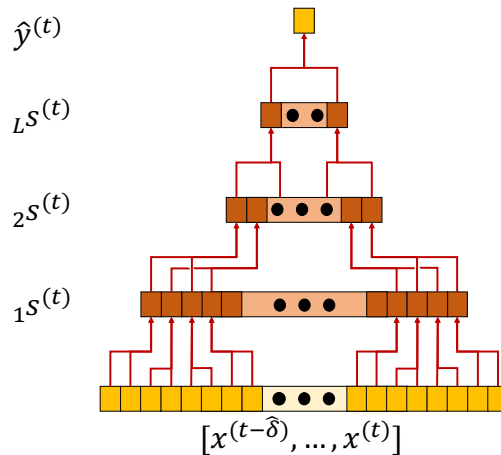


Figure 2: Single block of a TCN χ_i , where red arrow-style edges (\rightarrow) refer to the information flow from the input through multiple layers of convolutions to each subsequent layer as well as from the last layer L to the output $\hat{y}^{(t)}$.

3.2. Temporal Convolution Derived Reservoir Computing (TCRC)

The basic idea of the approach is to map the input $x^{(\cdot)}$ into the state $s^{(\cdot)}$ by a pairwise multiplication of $x^{(t-i)} \cdot x^{(t-i-1)}$. The similarity to NG-RC is easy to see, but the multiplication scheme simplified, the resulting state space subsequently reduced. In this paper we will refer to this as Temporal Convolution Derived Reservoir Computing (TCRC). The idea is comparable albeit more limited in the resulting connections to the NG-RC approach (cp. Sec. 2.3) by accessing only inputs of direct neighborhoods rather than all inputs in the range of the delay. We assume for the first network layer

$${}_1 s_i^{(t)} = x^{(t-i)} x^{(t-i-1)}. \quad (8)$$

Furthermore, we assume $0 \leq i < \hat{\delta} < S_T$, with $\hat{\delta}$ being the maximum delay applied to the input. The state vector of the first layer ${}_1 s^{(t)}$ is calculated as:

$${}_1 s^{(t)} = \#_{i=1}^{\hat{\delta}} s_i^{(t)} \quad (9)$$

Here, $\#_{i=1}^k$ is the concatenation over all indices in the given interval i to k . We introduce multiple layers to our approach. This was inspired by TCNs and the use of multiple layers in neural networks at large [14],

we introduce multiple layers to our approach. In the case of $L \in \mathbb{N}$ layers of this mapping, we propose for a layer $j > 1$

$${}_j s_i^{(t)} = {}_{j-1} s_i^{(t)} \cdot {}_{j-1} s_{i-1}^{(t)} = \hat{x}_i \cdot \hat{x}_{i-1}. \quad (10)$$

For mapping an accumulated state space over each layer to the output, we propose for the final state of the TCRC at the time step t for ${}_j s^{(t)}$ analogous to Eq. 9:

$$s^{(t)} = f([\#_{i=1}^{\hat{\delta}} x^{(t-1)}; \#_{j=1}^L {}_j s^{(t)}]), \quad (11)$$

for L layers as shown in Fig. 3 and the activation function $f(\cdot)$ chosen as $\tanh(\cdot)$, with $[\cdot; \cdot]$ being the concatenation of its given elements. In conclusion, our proposed approach results in the set of hyper-parameters $\Omega' = \{\hat{\delta}, L, S_T, \beta\}$, with $|\Omega'| < |\Omega|$ and $|\cdot|$ being the cardinality of the set. We allow for a zero padding, increasing the size of the state space, for potential noise as an alternative method of regularization as proposed in [29].

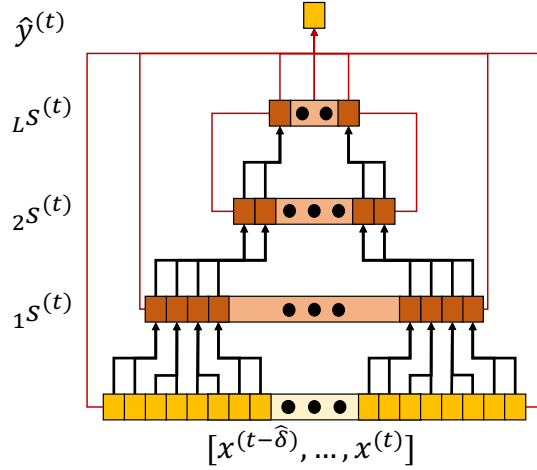


Figure 3: Proposed TCRC architecture, where black arrows (\rightarrow) refer to the multiplied tokens and red arrows (\rightarrow) refer to the learned mapping from the state space of each layer ${}_i s^{(t)}$ to the output $\hat{y}^{(t)}$.

3.3. TC-Extreme Layer (TCRC-ELM)

For combining the temporal convolution concept with Extreme Learning Machines (ELM) [16], we propose to use a single randomly generated weight matrix and the state $s^{(t)}$ of the TCRC as input to an ELM (cp. Fig. 4), similar to the work of [11]. We refer to this approach as TCRC-ELM. An input layer $W^{\text{in}} \in \mathbb{R}^{N^{r'} \times N^r}$ is drawn from a uniform distribution. We assume that $W^{\text{in}} \sim \mathcal{U}(-\delta, \delta)$ with $\delta = \frac{1}{2}$ and, furthermore, we assume that $s^{(t)} \in \mathbb{R}^{N^r}$, $\hat{s}^{(t)} \in \mathbb{R}^{N^{r'}}$ and $N^{r'} = n \cdot N^r$ with $n, N^{r'} \in \mathbb{N}$. To compute the additional state based on the ELM-like layer, we denote

$$\hat{s}^{(t)} = \hat{f}(W^{\text{in}}[\#_{i=1}^{\hat{\delta}} x^{(t-1)}; \#_{j=1}^L {}_j s^{(t)}]). \quad (12)$$

We utilize $\tanh(\cdot)$ as activation function $\hat{f}(\cdot)$. In conclusion, this architecture has a slightly increased number of hyper-parameters $\Omega'' = \{n, \hat{\delta}, L, S_T\}$ with $|\Omega'| < |\Omega''| < |\Omega|$.

Concluding the explanation of our proposal we want to give a more implementation focused description. For this we refer to Code 1.

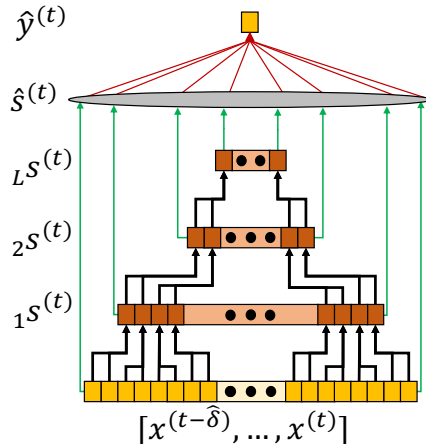


Figure 4: Proposed TCRC-ELM architecture, where black arrows (\rightarrow) refer to the multiplied tokens, green arrows (\rightarrow) refer to the randomly drawn weights of W^{in} , and red arrows (\rightarrow) refer to the learned mapping from the state space $\hat{s}^{(t)}$ to the output $\hat{y}^{(t)}$. For the sake of simplicity we have not shown the optional learned connections from all $j s^{(t)}$ to $\hat{y}^{(t)}$.

4. Evaluation and Results

In this section, we introduce the datasets used for evaluation as well as their preprocessing and the metrics used for evaluating the different approaches. Further, we evaluate the performance of state-of-the-art models, i.e. ESN, GRU and NG-RC, to establish a baseline on the selected datasets. Subsequently, we compare the performance of shallow TCRC and TCRC-ELM models to optimized ESN, DeepESN and GRU before also discussing results of deep TCRC architectures. We evaluate the complexity of our approach by the measured wall clock time and memory allocation at the end of the section.

Dataset. For the evaluation of the proposed methods, we chose the Mackey-Glass equation [13]. The selection of this dataset is motivated by its wide use in the RC literature to test ESNs, up to the proposal of the architecture in [17] and its in some cases chaotic nature. Albeit univariate, this makes it to an example widely comparable and simple enough to explain the relations between good performing predictors and the selected parameters yet still complex enough to be in need of a sophisticated predictive approach. We approximated the time series with the delays $\tau \in \{5, 10, 15, 17, 20, 25\}$ by solving

$$\dot{x}^{(t)} = \frac{\beta_{MG} x^{(t-\tau)}}{\Theta^n + x^{(t-\tau)^n}} - \gamma x^{(t)}, \quad (13)$$

with $[\beta_{MG}, \Theta, \gamma, n] = [0.2, 1, 0.1, 10]$ (cp. Fig. 5). The parameter τ defines the number of time steps in the past, which still have to be held in memory to calculate the current time step. The figure shows how the time series become more complex with an increasing τ . Thereby, $\tau = 17$ is the smallest delay resulting in chaotic behavior as explained in [34]. We train with $S_T = 2000$ time steps to be comparable with, e.g. [49]. For the test, we have chosen $S_P = 286$ time steps, representing two Lyapunov times for the $\tau = 17$ parametrization [9]. The approximation was done with an adaptation of the Euler method, with a step width of one. As secondary dataset we utilize the SantaFe Laser time series [1], as well-established benchmark dataset in the field of time series forecasting. The data points were measured with the time series showing chaotic behaviour, for explanation and further discussion we refer to the work of Weiss et. al [54]. We have limited the prediction of this dataset to 100 times steps according to [53]. We normalize all datasets to a zero mean value $\mu(\cdot)$ and a standard deviation $\hat{\sigma}(\cdot)$ of one. This is motivated by the utilized $\tanh(\cdot)$ non-linearity which saturates at values beyond the $[-1, 1]$ range. With the input being given to the output layer with a residual connection this limits the variance in the states, reducing the expressiveness of the $s^{(\cdot)}$

Algorithm 1 Pseudocode of the TCRC-ELM method

Require: $\hat{\delta} \geq 1$
Require: $t \geq \hat{\delta}$
Require: $L \leq \hat{\delta}$
 $W^{\text{in}} \sim \mathcal{U}(-\delta, \delta)$
while $t < S_T$ **do**
 ${}_0s^{(t)} \leftarrow \#_{i=0}^{\hat{\delta}} s_i^{(t-i)}$
 while $l < L$ **do**
 ${}_l s^{(t)} \leftarrow \#_{i=1}^{\hat{\delta}-l} s_{i-1}^{(t)} \#_{l-1} s_i^{(t)}$
 end while
 $\hat{s}^{(t)} \leftarrow W^{\text{in}} \#_{l=0}^L {}_l s_i^{(t)}$
end while
 $W^{\text{out}} \leftarrow T(f(\#_{t=0}^{S_T} \hat{s}^{(t)}, \#_{l=0}^L \#_{t=0}^{S_T} {}_l s^{(t)}, \#_{t=\hat{\delta}}^{S_T} \#_{i=0}^{\hat{\delta}} x^{t-i}))$ $\triangleright T(\cdot)$ is the application of the \triangleright Tikhonov
Regularization
while $\hat{t} \leq S_P$ **do**
 ${}_0s^{(S_T+\hat{t})} \leftarrow \#_{i=0}^{\hat{\delta}} s_i^{(S_T+\hat{t}-i)}$
 while $l < L$ **do**
 ${}_l s^{(S_T+\hat{t})} \leftarrow \#_{i=1}^{\hat{\delta}-l} s_{i-1}^{(S_T+\hat{t})} \#_{l-1} s_i^{(S_T+\hat{t})}$
 end while
 $\hat{s}^{(S_T+\hat{t})} \leftarrow W^{\text{in}} \#_{l=0}^L {}_l s_i^{(S_T+\hat{t})}$
 $\hat{y}^{(\hat{t})} \leftarrow W^{\text{out}} f([\hat{s}^{(S_T+\hat{t})}, \dots])$ \triangleright Compare stacking W^{out}
 $x^{(S_T+\hat{t})} \leftarrow \hat{y}^{(\hat{t})}$
end while

and $\hat{s}^{(\cdot)}$ respectively for the learning task. For the benefits of normalization, we also reference to [14] and to empirical observations such as those of [44]. We normalized the input as well as the targets by use of the z-score [6]:

$$X = \frac{X_0 - \mu(X_0)}{\hat{\sigma}(X_0)}. \quad (14)$$

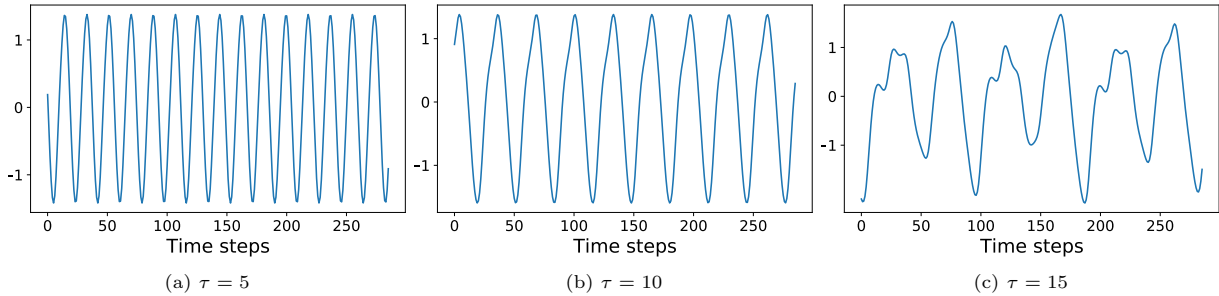
Here we use X_0 as the original input (cp. Eq. 13), i.e. because of the characteristic of all time series to have values in a fixed range, all time steps computed. $X \in \mathbb{R}^{N^{\text{in}} \times (S_I + S_T)}$ is the matrix from which the inputs are drawn. The target outputs are drawn from $Y, Y_0 \in \mathbb{R}^{N^{\text{out}} \times (S_I + S_T + S_P)}$ analogous to the inputs (cp. Eq. 14).

Training and Evaluation. We limit the shown results to the tested set of hyper-parameters, i.e. the best performing set for our approaches. For the ESN and TCRC-ELM, which use randomness in their initialization, we report the mean error over 15 runs. In regard to our own approach we tested multiple delays $\hat{\delta}$ of each tested τ in the range $\hat{\delta} \in [2, 1900]$, substantially larger than any potential τ used for the Mackey-Glass equations. As training objective and for comparison of trained models, we use the mean squared error MSE

$$\mathcal{L} = \frac{\sum_{t'=0}^{S_P} (y^{(t')} - \hat{y}^{(t')})^2}{S_P}. \quad (15)$$

For the evaluation itself, we test the TCRC and TCRC-ELM with only the regularization parameter β and the delay $\hat{\delta}$ varied at first. In a second step we show the influence of allowing for a deep TCRC, using multiple layers as shown in Figure 3 both for the TCRC itself as well as the TCRC-ELM. In both evaluation steps, we compare them to an ESN with optimized hyper-parameters as well as an GRU with optimized input size. We additionally report the number of valid time steps, whereby we have chosen a time steps as valid, for as long as it stays in an absolute difference of below one standard deviation of the data around the ground truth, reasoned by it staying inside the orbit of the attractor.

Non-chaotic variation of the Mackey-Glass time series



Chaotic variation of the Mackey-Glass time series

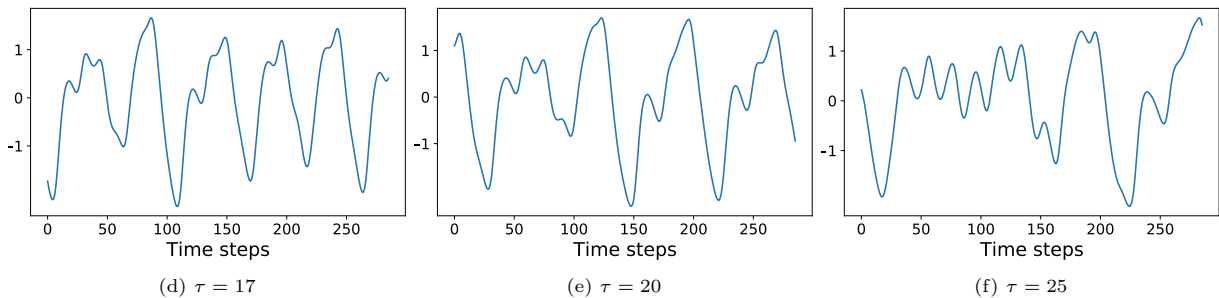


Figure 5: Visualization of the Mackey-Glass equation for 286 time steps and different values of τ , shifting from non-chaotic to chaotic with $\tau \geq 17$.

Parametrization. The baseline for the ESN was the default setting proposed by [49], from which we used a grid search approach with $N^{\text{res}} \in [1000, 3000]$, $\rho \in [0.5, 1.5]$, density $\sigma(W^r) \in [0.1, 1]$ as well as the regularization coefficient $\beta \in [10^{-10}, 10^{-3}]$. Other approaches to optimize the parameter selection are possible as shown, e.g. by [41, 57] with a comparison of optimization methods used in literature. In case of the AEESN we started from the parametrization utilized in the proposing literature of [30] and a grid search. For our own proposal we used a grid search approach with a delay $\hat{\delta} \in [1, 1900]$ time steps, the number of layers $l \in [1, 6]$ and a regularization coefficient $\beta \in [10^{-6}, 1]$. For the GRU we utilized a grid search with a state space size from four and so smaller than our smallest delay up to 2048, a history and input of up to 150 time steps and allowing for a number of layers $l \in [1, 5]$, while training for a number of epochs $e \in [100, 2000]$ with the Adam optimizer [21, 37].

4.1. Baselines

In this section we discuss the established models for time series prediction used in this work. They are used here as baseline to measure the results of our approaches TCRC and TCRC-ELM against.

Echo State Network (ESN). We use a standard Echo State Network (ESN) to establish a baseline for our evaluation on the generated datasets (cp. Tab. 1). For the ESN, we have tested spectral radii from $\rho = 0$, basically resulting in an ELM since the ESN’s memory has no effect for that case, up to $\rho = 1.5$ and state space sizes $N^r \in \{1000; 2000; 3000\}$ with a connectivity of $\sigma(W^r)$ between 0.1 and 1.0 and no leakage rate applied. We observe the error to increase together with τ as to be expected, from $5.72 \cdot 10^{-3}$ to 1.06. The ESN is the only predictor with a substantial increase between the two lowest values of τ while all other predictors are comparable for this values (GRU) or even decrease from $\tau = 5$ to $\tau = 10$. A substantial increase we observe between $\tau = 15$ and $\tau = 17$. We argue that this is expected due to the Mackey-Glass equation switching between those values from non-chaotic to chaotic behavior. The predictive capabilities of the ESN are insufficient in the bounds of hyper-parameters, chosen to be comparable with our proposed

Table 1: MSE of best performing Basic ESN, AEESN, GRU and TCRC for prediction of 286 time steps for different τ of the Mackey-Glass equation and 100 time steps for the Santa Fe Laser dataset. We present the best tested $\rho \in [0, 1.5]$ and $N^T \in \{1000, 2000, 3000\}$ as well as best tested input lengths $\hat{\delta}$ for the GRU and TCRC. The green values show a decrease while an increase is marked by red values which relative to the ESN for each dataset. The best performing model is marked in bold. Additional regularization is not used.

Dataset	ESN	AEESN	GRU	TCRC		TCRC-ELM	
				MSE	rel. impr.	MSE	rel. impr.
Mackey-Glass, $\tau =$							
5	$5.72 \cdot 10^{-3}$	$2.37 \cdot 10^1$	$1.90 \cdot 10^{-4}$	$1.25 \cdot 10^{-6}$	-99.98%	$4.50 \cdot 10^{-6}$	-99.92%
10	$3.59 \cdot 10^{-2}$	$1.26 \cdot 10^{-1}$	$7.61 \cdot 10^{-3}$	$2.31 \cdot 10^{-9}$	-99.99%	$2.25 \cdot 10^{-7}$	-99.99%
15	$5.41 \cdot 10^{-2}$	$2.45 \cdot 10^{-2}$	$2.50 \cdot 10^{-1}$	$4.79 \cdot 10^{-4}$	-99.11%	$7.36 \cdot 10^{-3}$	-86.40%
17	$2.99 \cdot 10^{-1}$	$7.95 \cdot 10^{-2}$	$4.69 \cdot 10^{-1}$	$9.33 \cdot 10^{-1}$	+212.04%	$4.68 \cdot 10^{-2}$	-84.35%
20	$3.20 \cdot 10^{-1}$	$3.88 \cdot 10^{-1}$	$6.10 \cdot 10^{-1}$	$5.58 \cdot 10^{-1}$	+74.38%	$3.24 \cdot 10^{-1}$	+1.25%
25	1.06	$1.70 \cdot 10^{-1}$	$8.15 \cdot 10^{-1}$	1.25	+17.92%	$5.33 \cdot 10^{-1}$	-49.72%
Laser							
	$6.31 \cdot 10^{-2}$	$9.13 \cdot 10^0$	$5.92 \cdot 10^{-1}$	$4.33 \cdot 10^{-1}$	+581.21%	$9.48 \cdot 10^{-1}$	+1402.38%

approach for the most chaotic dataset with an error above one for values normalized to a standard deviation of one. The deviation across multiple random initialization led to the observation of the ESNs prediction varying by an order of magnitude between the runs, albeit only for the chaotic time series. The ESN achieved the best prediction for the SantaFe Laser dataset, one order of magnitude better compared to the proposed approaches and other baselines.

Gated Recurrent Unit (GRU). We tested a Gated Recurrent Unit (GRU) with a single up to five layers to be comparable to the ESN, while still allowing for the advantages of a deep neural network, with an input length $\hat{\delta} \in [1, 150]$ and a size of the GRU cell of up to 2048. We observe the prediction of a fixed value with a single time step instead of showing dynamic behavior. This seems surprising but is comparable to results in the literature [40]. We observe a benefit of increasing the input by use of delays across multiple time steps as shown in Tab. 1. Hereby we tested delays up to $\hat{\delta} = 150$, larger than in some literature, e.g. [45] but still observed the best performing, and here reported GRU to utilize a single layer. The error increases with an increase of τ . In comparison to the ESN the single layered GRU is better by up to 78.80% for the non-chaotic time series of $\tau = 10$. Using the GRU results in a worse prediction for $\tau = 15$ to $\tau = 20$, while in a better one for $\tau = 25$ and also the hyperchaotic parametrization of $\tau = 30$, shown in Tab. A.4. As shown in Tab. C.5 the GRU can achieve a valid prediction of up to the full 286 time steps with missing the prediction for $\tau = 15$. We argue for this prediction to be an outlier, our approach to the approximation of the dataset being to inaccurate, leading to a behaviour, more shifted towards chaotic than represented in the training data. For the SantaFe Laser time series we observe the GRU to be a better predictor than the AEESN and TCRC-ELM, but worse than TCRC and ESN.

Next Generation Reservoir Computing (NG-RC). Next Generation Reservoir Computing (NG-RC) is described by its authors as comparable to ESNs with hundreds to thousands of nodes [12]. Therefore, we consider the evaluated reservoir sizes for ESN in our study to be a valid comparison. For the lowest tested value of $\tau = 5$ we observe the best prediction with a delay of $\hat{\delta} = 10$, the error decreases by above 99.99%. In case of $\tau = 10$ the selection of $\hat{\delta} = 10$ compared to the ESN leads to a decrease of 64.90%. The best MSE for $\tau \geq 15$ is observed for a very low delay $\hat{\delta} = 2$ which leads to a constant prediction. An increase of the $\hat{\delta}$ leads to the network seemingly being unable to learn. Due to this, we do not further use the NG-RC for comparisons and do not discuss the prediction of the chaotic time series. We observe a trend to a large N^{in} with $\tau \ll N^{\text{in}}$. Also we observe, compared to the ESN, a more drastic increase of the MSE from non-chaotic to chaotic variants of the system. We observe a similar result of not being close to the targeted trajectory for the SantaFe laser dataset, showing the need of an alternate non-linear function for this time series, than e.g. the Lorenz'63 attractor of [26] utilized in [12].

DeepR-ESN (AEESN). As further baselines we used the Deep ESN with Autoencoders between the reservoirs (AEESN) proposed by [30]. We found the AEESN for our preprocessing of our datasets and chosen time series to vary between results substantially with in some cases, especially for $\tau = 5$ with an error of up to more than 225, which results in such an incomparable large mean reported in Tab. 2. The error for $\tau = 30$ is reported in the appendix (cp. Tab. A.4). Reporting only the best initializations, e.g. in Tab. 2 and Tab. C.5 shows the AEESN as beneficial compared to the ESN in some cases, even when measured in the number of valid time steps of the prediction. We observe it to be superior to both of our proposed approaches for $\tau = 25$, outperforming the TCRC for all chaotic time variations of Mackey-Glass.

In case of the Santa Fe Laser time series the AEESN is the worst performing baseline, as shown in Tab. 2 and 1 because of the utilization of the mean across 15 initializations.

Table 2: MSE of best performing configurations of TCRC and TCRC-ELM as well as best baseline for prediction of 286 time steps for different τ of the Mackey-Glass equation as well as 100 time steps for the SantaFe Laser dataset. We present the best tested input lengths δ for our TCRC and TCRC-ELM with the best tested number of layers L . The green values show a decrease while an increase is marked by red values relative to the best baseline for each dataset. The best performing model is marked in bold.

Dataset		Baseline	TCRC			TCRC-ELM		
Mackey-Glass	τ		L	MSE	rel. impr.	L	MSE	rel. impr.
non-chaotic	5	$1.90 \cdot 10^{-4}$	4	$1.15 \cdot 10^{-6}$	-99.39%	2	$2.81 \cdot 10^{-6}$	-98.52%
	10	$7.61 \cdot 10^{-3}$	1	$2.31 \cdot 10^{-9}$	-99.99%	1	$1.37 \cdot 10^{-7}$	-99.99%
	15	$2.45 \cdot 10^{-3}$	3	$4.07 \cdot 10^{-4}$	-98.38%	1	$2.22 \cdot 10^{-3}$	-90.94%
chaotic	17	$7.95 \cdot 10^{-2}$	3	$2.22 \cdot 10^{-1}$	$+79.25\%$	1	$4.35 \cdot 10^{-2}$	-45.28%
	20	$3.20 \cdot 10^{-1}$	4	$1.70 \cdot 10^{-1}$	-46.88%	1	$1.16 \cdot 10^{-1}$	-63.75%
	25	$1.70 \cdot 10^{-1}$	1	$5.67 \cdot 10^{-1}$	$+233.53\%$	1	$3.98 \cdot 10^{-1}$	$+134.12\%$
Laser								
		$6.31 \cdot 10^{-2}$	1	$4.33 \cdot 10^{-1}$	$+581.46\%$	1	$9.48 \cdot 10^{-1}$	$+1402.38\%$

4.2. Proposed approaches

Here we want to discuss both of our approaches. We start by discussion of the the shallow TCRC and TCRC-ELM. After this the influence of multiple layers in the TCRC and subsequently the TCRC-ELM will be evaluated.

TCRC. For the TCRC the predictions show a trend to an increased error from $\tau = 5$ to $\tau = 25$ as shown in 1 for the default set of hyper-parameters, using a single layer. Unintuitive we argue seems the decrease of the error from $\tau = 5$ to $\tau = 10$. We argue this to be due to the periodicity of both time series with $\tau = 10$ evolving slower and showing less dynamics in 286 time steps than $\tau = 5$. Overall we observe the TCRC to result in the best predictions for the non-chaotic variations of the Mackey-Glass equation up to $\tau \leq 15$. Thereby, the error compared to the ESN is reduced by up to 99.99% for $\tau = 10$. For all non-chaotic time series we observe a reduction of the error of over 99.10%. For the chaotic variations we observe the TCRC with a default set of hyper-parameters to be an inferior predictor compared to the ESN and AEESN. Especially for $\tau = 17$ the error is increased in comparison to the ESN by 212.04%. With a further increase in τ the relative decrease in performance declines to 74.38% and 17.92% respectively, but we argue this to be because of the worsening performance of the ESN. In regard to the number of valid time steps we have shown in Tab. C.5 that our proposed approach is able to achieve a valid prediction for the complete interval of our test data, staying in the attractor, for the non-chaotic time series. The outlier of $\tau = 20$ is repeated not only in regard to the MSE but also the number of valid time steps, showing a better prediction than for $\tau = 17$. For the most chaotic tested time series, outside hyperchaos, the TCRC still achieves 100 time

steps, before the shift in the timing is too large, albeit still being inside the value space of the attractor, as shown in Fig. B.10.

We observe the TCRC to be the second best predictor in case of the Santa Fe Laser dataset, being 26.86% better than the third best performing baseline of GRU, as shown in Tab. 2 and Tab. 1, but worse than the ESN.

TCRC-ELM. For a single layer, the TCRC-ELM yields superior results, similar to the TCRC, in comparison to the ESN for all datasets with the exception of $\tau = 20$ (cp. Tab. 1). Like for the TCRC, the error decreases from $\tau = 5$ to $\tau = 10$ by 95.00%. For the non-chaotic variations, up to including $\tau = 15$, we observe a substantial increase in the error compared to the TCRC, albeit still being better than the ESN, GRU and AEESN. For the least chaotic time series, we observe the TCRC-ELM to be the overall best tested predictor. The variation of $\tau = 20$ is the only case in which the shallow TCRC-ELM yields 1.25% worse results than the ESN. For the most chaotic time series with $\tau = 25$, the TCRC-ELM is the second best performing predictor, after the AEESN. We see, that the TCRC-ELM achieves the same number of valid time steps as the TCRC for the non-chaotic time series, while outperforming it for the chaotic ones, as shown in Tab. C.5 and exemplary visualized in Fig. B.11. The TCRC-ELM is the second worst performing predictor for the Santa Fe Laser dataset in our tests, drifting away from the dynamics of the trajectory, increasing the error by still seemingly correct behavior.

In total we argue to have managed to mitigate the negative influence of randomness, while preserving the ability to handle memory-dependent time series, as we set as a problem in Sec. 1, more specifically as our first research question. Furthermore, we have shown the ability of our proposed approach to handle chaotic time series, which we have set as our second research question.

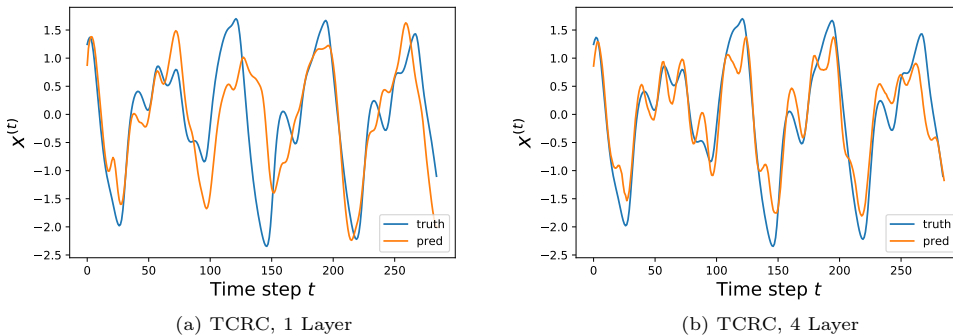


Figure 6: Exemplary visualization of the prediction and ground truth for $\tau = 20$ for the TCRC

Multi-Layered TCRC. With the introduction of multiple layers and so a deep TCRC we observe a further reduction of the error (cp. Tab. 2). Even for the single layered case, we now allow for noise as additional regularization mechanism (cp. Sec. 3.2). For $\tau = 5$, we observe the lowest error for the TCRC, in this case with four layers and a reduction of 8.00%, compared to the shallow TCRC. The TCRC-ELM benefits from a second layer with a decrease of 37.56% compared to its single layered variation. In difference to this, the time series based on $\tau = 10$ does not benefit from a deep TCRC, and neither does the TCRC-ELM, which solely decreases in error due to regularization. For the most dynamic but non-chaotic time series we observe the TCRC as best performing predictor benefitting from three layers, while the TCRC-ELM benefits from the allowed noise and larger state space. For the chaotic time series, the TCRC shows clearly decreased error with multiple layers and noise outperforming the ESN for all $\tau \geq 17$. In the least chaotic case the TCRC uses three layers as for $\tau = 15$. For $\tau = 17$ the error decreases compared to the ESN by 25.75%. The TCRC-ELM benefits from noise for $\tau = 17$, resulting in a decrease compared to the shallow network of 7.05%. The time series approximated with $\tau = 20$ results in a four layered TCRC and a noisy TCRC-ELM being able to outperform the ESN by 46.88% and 63.75% respectively. Also for $\tau = 25$, our

proposed approaches outperform the ESN, while being worse predictors than the AEESN, both using noise and a single layer. As a difference to the TCRC, we observe no substantial benefit of multiple layers for the TCRC-ELM in this study. The best performing TCRC-ELMs uses a single layer for $\tau \geq 10$ with only $\tau = 5$ benefiting from two layers. In case of the hyperchaotic Mackey-Glass variation with $\tau = 30$ reported in Tab. A.4 we observe the TCRC-ELM to be slightly better than the TCRC, similar to the behavior on chaotic datasets of Mackey-Glass with an decrease in error of 5.61%. With the introduction of the TCRC-ELM we observe a benefit of the reintroduction of randomness, allowing for better resulting the case of chaotic time series across all approximated time series, with a disadvantage in case of the Santa Fe Laser dataset.

4.3. Complexity

Table 3: Runtime in seconds and maximum memory allocation in MBit of each tested model

Metric	$\simeq N^r$	ESN	AEESN	GRU	TCRC	TCRC-ELM
Runtime						
	300	0.97	97.29	10.18	6.88	5.18
	2000	22.22	305.48	59.17	693.02	176.92
Memory Alloc.						
	2000	2447	1424	3108	2864	2834

In terms of runtime we observe an increase for a single run with a large N^r for the TCRC and TCRC-ELM as shown in Tab. 3 compared to the ESN. The smaller delay for the TCRC-ELM of the same state space size as the TCRC results in it being faster than the TCRC. The ESN is per run substantially faster, in need of only 18.73% of the runtime of the TCRC-ELM in case of the smaller compared state space. In case of the larger state space the ESNs runtime increases by factor 22.91. In case of TCRC and TCRC-ELM it increases by 100.73 and 34.15, respectively. We observe our proposed approaches to be faster than AEESN and GRU for the smaller state space, but scaling worse with an increase in state space size.

In case of allocated memory the ESN is the second smallest model, saving 417 MBit, 17.04%, respectively compared to the TCRC as the largest model (cp. Tab. 3). In comparison to the TCRC-ELM the ESN is 387 MBit, and such 15.82% smaller. The AEESN is the overall smallest model, needing around half the memory compared to TCRC and TCRC-ELM, a reduction of 50.28% and 49.75% compared to TCRC and TCRC-ELM, respectively. The GRU is the most memory intensive predictor, increasing the needed memory compared to TCRC and TCRC-ELM by 8.52% and 9.67%, respectively.

5. Discussion

In this work, we propose a new mapping and presented its use with multiple layers as well as delays. We observe our approach in case of optimized hyper-parameters to result in an up to 99.99% and 85.45% lower error than the also optimized ESN for non-chaotic and chaotic time series respectively. The introduction of the random mapping to the TCRC, resulting in the TCRC-ELM improves the predictive capabilities of the TCRC for chaotic time series but decreases it for non-chaotic time series. Still, both of our approaches decrease the error for the time series with $\tau \leq 15$ by over 95.00% in comparison to the ESN and by over 90.00% compared to the best baseline of ESN, AEESN and GRU. For the chaotic variations, the TCRC-ELM still achieves a reduction by 85.45%, the TCRC up to 46.88% compared to the ESN. In conclusion, our approaches outperform all baselines except AEESN for the most chaotic approximated time series $\tau = 25$. We argue that this result shows that the predictive capabilities of reservoir computing can be achieved without the drawback of randomness in the weights but also shows still existing limits compared to the more learned GRU and more complex AEESN, the later of which outperforms our approaches for the hyperchaotic $\tau = 30$. In case of the measured SantaFe Laser dataset the ESN outperforms both of our approaches with the TCRC being the second best predictor. In regard to our research questions, we have answered all of them positively. The results of our TCRC shows the possibility to handle chaotic time series even with a need for memory in

the framework of reservoir computing without the need of random initializations. Still our results show the benefit of reintroduction of randomness as W^{in} in the TCRC-ELM for the prediction of chaotic time series.

Limitations of the study. The utilized time series, i.e. approximated variations of the Mackey-Glass equations, are not representative for all problems. For example, the data do not suffer from noise or faulty measurements. We have limited ourselves to approximated time series to allow for interpretability of the results because of known characteristics of the time series. Furthermore, we did not use multivariate time series due to the rare occurrence of delayed differential equations in multiple dimensions in general and specifically we are not aware of a single one that is used as benchmark in previous RC studies. The training length S_T was set to a fixed value for the sake of comparability. Several studies propose the use of a leakage rate as addition to the initial ESN and also demonstrate that the concept can yield substantially better performing networks for some datasets. We aim to compare to the fundamental version of the ESN as baseline and therefore decided to not utilize a leakage. Furthermore, recent studies also proposed alternative approaches to ESNs regarding the initialization of the random matrices or alternative topologies of these were also not used as baselines because of the limited comparability in literature and their partially specialised architecture for certain characteristics of time series. For the Mackey-Glass-based datasets utilized for testing our approaches we have chosen an approximation with a step width of one. This leads to the trajectory itself becoming not-periodic for $\tau = 15$, which can lead to the results varying, especially for this delay τ , between approximations. We also remark that differences in the implementation allow different runtimes. An extensive comparison of methods of the software engineering as it would be necessary to show influences of such is the focus of currently conducted research.

6. Conclusion

The prediction of time series is a challenging task even for modern neural networks, becoming even harder time series with long historic dependences and chaotic behavior. An architecture that has been demonstrated well suited for this kind of task is the mostly randomly initialized RC. In the last years, also TCN have been demonstrated as an alternative approach to handle history dependent time series, which in contrast need a more complex iterative learning mechanism. Inspired by the learning mechanism of RC and the history handling of TCN, we propose two novel network architectures. The architectures utilize a novel deterministic mapping of inputs into the network’s state space, a method to create deep, multi-layered RC networks based on the deterministic mapping, and finally the additional use of a random mapping based on the size of the non-random network prior. The proposed methods improve the predictive accuracy compared to ESNs, as the most prominent RC architectures, by several orders of magnitude. We also observe a benefit of depth, i.e. multi-layered architectures, in case of the TCRC yielding results that outperform the ESN for all tested time series by up to 99.99% for non-chaotic and 46.88% for chaotic time series. If optimized in regard to its number of layers the TCRC outperforms even the GRU with the exceptions of $\tau \in \{30\}$. The TCRC-ELM outperforms the ESN for non-chaotic time series by a similar relative improvement, for the chaotic ones up 85.45%. With the exception of $\tau = 20$ the TCRC-ELM outperforms all other networks excepts the AEESN, including the GRU for the chaotic time series. These results were achieved with an optimized ESN but without the optional usage of a leakage rate as additional hyper-parameter.

We plan to study additional scalings and a linear blending between state spaces at different time steps for our proposed method. An extensive and expensive study with delayed ELMs, NG-RCs with large $\hat{\delta}$ and comparatively scaled states spaces for TCRC-ELMs is also seen as future work. Further, we plan to explore the possibilities of non-random mappings for the TCRC-ELM to extinguish the randomness and need for multiple runs, reducing the runtime substantially.

Funding

We are funded by the Carl Zeiss Foundation’s grants: P2018-02-001 and P2017-01-005 and by the German Federal Ministry of Education and Research (BMBF) grant: 02P22A040.

References

- [1] . . Dynaml. URL: <https://github.com/transcendent-ai-labs/DynaML/tree/master/data>.
- [2] Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 .
- [3] Bianchi, F.M., Livi, L., Alippi, C., 2016. Investigating echo-state networks dynamics by means of recurrence analysis. *IEEE transactions on neural networks and learning systems* 29, 427–439.
- [4] Bolt, E., 2020. On explaining the surprising success of reservoir computing forecaster of chaos? the universal machine learning dynamical system with contrasts to var and dmd. arXiv preprint arXiv:2008.06530 .
- [5] Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, in: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152.
- [6] Carey, J.J., Delaney, M.F., 2010. T-scores and z-scores. *Clinical reviews in bone and mineral metabolism* 8, 113–121.
- [7] Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 .
- [8] Cortes, C., Vapnik, V., 1995. Support-vector networks. *Machine learning* 20, 273–297.
- [9] Farmer, J.D., 1982. Chaotic attractors of an infinite-dimensional dynamical system. *Physica D: Nonlinear Phenomena* 4, 366–393.
- [10] Fu, S., Tang, Z., Li, J., 2023. Ibm stock forecast using lstm, gru, attention and transformer models, in: *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, IEEE. pp. 167–172.
- [11] Gallicchio, C., Micheli, A., 2011. Architectural and markovian factors of echo state networks. *Neural Networks* 24, 440–456.
- [12] Gauthier, D.J., Bolt, E., Griffith, A., Barbosa, W.A., 2021. Next generation reservoir computing. *Nature communications* 12, 1–8.
- [13] Glass, L., Mackey, M., 2010. Mackey-glass equation. *Scholarpedia* 5, 6908.
- [14] Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- [15] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural computation* 9, 1735–1780.
- [16] Huang, G.B., Zhu, Q.Y., Siew, C.K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks, in: *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*, Ieee. pp. 985–990.
- [17] Jaeger, H., 2001. The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148, 13.
- [18] Jaeger, H., 2007. Echo state network. *Scholarpedia* 2, 2330.
- [19] Jaeger, H., Lukoševičius, M., Popovici, D., Siewert, U., 2007. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural networks* 20, 335–352.
- [20] Kanerva, P., 2009. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive computation* 1, 139–159.
- [21] Kingma, D.P., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .
- [22] Kuang, Y., 1993. *Delay differential equations: with applications in population dynamics*. Academic press.
- [23] LeCun, Y., et al., 1989. Generalization and network design strategies. *Connectionism in perspective* 19, 18.
- [24] Li, Q., Chen, Y., Ao, N., Han, X., Wu, Z., 2018. Echo state network-based visibility graph method for nonlinear time series prediction, in: *2018 Chinese Control And Decision Conference (CCDC)*, IEEE. pp. 1854–1859.
- [25] Li, Z., Tanaka, G., 2020. Deep echo state networks with multi-span features for nonlinear time series prediction, in: *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE. pp. 1–9.
- [26] Lorenz, E.N., 1963. Deterministic nonperiodic flow. *Journal of the atmospheric sciences* 20, 130–141.
- [27] Lu, Z., Pathak, J., Hunt, B., Girvan, M., Brouckert, R., Ott, E., 2017. Reservoir observers: Model-free inference of unmeasured variables in chaotic systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 27, 041102.
- [28] Lukoševičius, M., 2012. A practical guide to applying echo state networks, in: *Neural networks: Tricks of the trade*. Springer, pp. 659–686.
- [29] Lukoševičius, M., Jaeger, H., 2009. Reservoir computing approaches to recurrent neural network training. *Computer Science Review* 3, 127–149.
- [30] Ma, Q., Shen, L., Cottrell, G.W., 2020. Deep-esn: A deep projection-encoding echo-state network. *Information Sciences* 511, 152–171.
- [31] Mackey, M.C., Glass, L., 1977. Oscillation and chaos in physiological control systems. *Science* 197, 287–289.
- [32] Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* 36, 54–74.
- [33] Malik, Z.K., Hussain, A., Wu, Q.J., 2016. Multilayered echo state machine: a novel architecture and algorithm. *IEEE Transactions on cybernetics* 47, 946–959.
- [34] Moon, J., Wu, Y., Lu, W.D., 2021. Hierarchical architectures in reservoir computing systems. *Neuromorphic Computing and Engineering* 1, 014006.
- [35] Morales, G.B., Mirasso, C.R., Soriano, M.C., 2021. Unveiling the role of plasticity rules in reservoir computing. *Neuro-computing* .
- [36] Panahi, S., Kong, L.W., Moradi, M., Zhai, Z.M., Glaz, B., Haile, M., Lai, Y.C., 2024. Machine-learning prediction of tipping and collapse of the atlantic meridional overturning circulation. arXiv preprint arXiv:2402.14877 .
- [37] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch, in: *NIPS-W*.
- [38] Pathak, J., Wikner, A., Fussell, R., Chandra, S., Hunt, B.R., Girvan, M., Ott, E., 2018. Hybrid forecasting of chaotic

- processes: Using machine learning in conjunction with a knowledge-based model. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28, 041101.
- [39] Qin, H., 2019. Comparison of deep learning models on time series forecasting: A case study of dissolved oxygen prediction. arXiv preprint arXiv:1911.08414 .
- [40] Shahi, S., Fenton, F.H., Cherry, E.M., 2022. Prediction of chaotic time series using recurrent neural networks and reservoir computing techniques: A comparative study. *Machine learning with applications* 8, 100300.
- [41] Soltani, R., Benmohamed, E., Ltifi, H., 2023. Echo state network optimization: A systematic literature review. *Neural Processing Letters* 55, 10251–10285.
- [42] Sun, C., Song, M., Cai, D., Zhang, B., Hong, S., Li, H., 2022. A systematic review of echo state networks from design to application. *IEEE Transactions on Artificial Intelligence* .
- [43] Teo, K.K., Wang, L., Lin, Z., 2001. Wavelet packet multi-layer perceptron for chaotic time series prediction: effects of weight initialization, in: *International Conference on Computational Science*, Springer. pp. 310–317.
- [44] Teutsch, P., Käufer, T., Mäder, P., Cierpka, C., 2023. Data-driven estimation of scalar quantities from planar velocity measurements by deep learning applied to temperature in thermal convection. *Experiments in Fluids* 64, 1–18.
- [45] Teutsch, P., Mäder, P., 2022. Flipped classroom: Effective teaching for time series forecasting. *Transactions on Machine Learning Research* URL: <https://openreview.net/forum?id=w3x20YecQK>.
- [46] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 267–288.
- [47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. arXiv preprint arXiv:1706.03762 .
- [48] Verstraeten, D., Schrauwen, B., D’Haene, M., Stroobandt, D., 2007. An experimental unification of reservoir computing methods. *Neural Networks* 20, 391–403.
- [49] Viehweg, J., Worthmann, K., Mäder, P., 2022. Parameterizing echo state networks for multi-step time series prediction. *Neurocomputing* URL: <https://www.sciencedirect.com/science/article/pii/S0925231222014291>, doi:<https://doi.org/10.1016/j.neucom.2022.11.044>.
- [50] Vlachas, P.R., Pathak, J., Hunt, B.R., Sapsis, T.P., Girvan, M., Ott, E., Koumoutsakos, P., 2020. Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks* .
- [51] Walther, D., Schmidt, L., Schricker, K., Junger, C., Bergmann, J.P., Notni, G., Mäder, P., 2022. Automatic detection and prediction of discontinuities in laser beam butt welding utilizing deep learning. *Journal of Advanced Joining Processes* 6, 100119.
- [52] Walther, D., Viehweg, J., Haueisen, J., Mäder, P., 2023. A systematic comparison of deep learning methods for eeg time series analysis. *Frontiers in Neuroinformatics* 17, 1067095.
- [53] Weigend, A.S., Gershenfeld, N.A., 1993. Results of the time series prediction competition at the santa fe institute, in: *IEEE international conference on neural networks*, IEEE. pp. 1786–1793.
- [54] Weiss, C.O., Hübner, U., Abraham, N.B., Tang, D., 1995. Lorenz-like chaos in nh3-fir lasers. *Infrared Physics & Technology* 36, 489–512.
- [55] Yao, X., Wang, Z., 2020. Fractional order echo state network for time series prediction. *Neural Processing Letters* 52, 603–614.
- [56] Zhai, Z.M., Kong, L.W., Lai, Y.C., 2023a. Emergence of a resonance in machine learning. *Physical Review Research* 5, 033127.
- [57] Zhai, Z.M., Moradi, M., Kong, L.W., Glaz, B., Haile, M., Lai, Y.C., 2023b. Model-free tracking control of complex dynamical trajectories with machine learning. *Nature communications* 14, 5698.
- [58] Zhai, Z.M., Moradi, M., Panahi, S., Wang, Z.H., Lai, Y.C., 2024. Machine-learning nowcasting of the atlantic meridional overturning circulation. *APL Machine Learning* 2.

Table A.4: MSE of best performing parametrizations for the tested networks for prediction of 286 time steps for $\tau = 30$ of the Mackey-Glass equation.

τ	ESN	AEESN	GRU	TCRC	TCRC-ELM
30	$1.15 \cdot 10^1$	$3.11 \cdot 10^{-2}$	$3.60 \cdot 10^{-1}$	$7.13 \cdot 10^{-2}$	$6.73 \cdot 10^{-2}$

Appendix A. Prediction of Hyperchaotic Mackey-Glass Equation

The hyper-parameters of the ESN, GRU, AEESN, TCRC and TCRC-ELM were chosen as in their respective proposals [30], with testing a variational autoencoder as well as a Primary Component Analysis as substitute of an autoencoder. We reported the best combination of hyper-parameters, found as described in Sec. 4 in Tab. A.4. The reported value of $\tau = 30$ marks the threshold for the step from chaotic to hyperchaotic behavior with a second positive Lyapunov-exponent.

Appendix B. Visual Comparison of predictions

We show an exemplary comparison, limited to one run of each predictor, for the best prediction, measured by MSE.

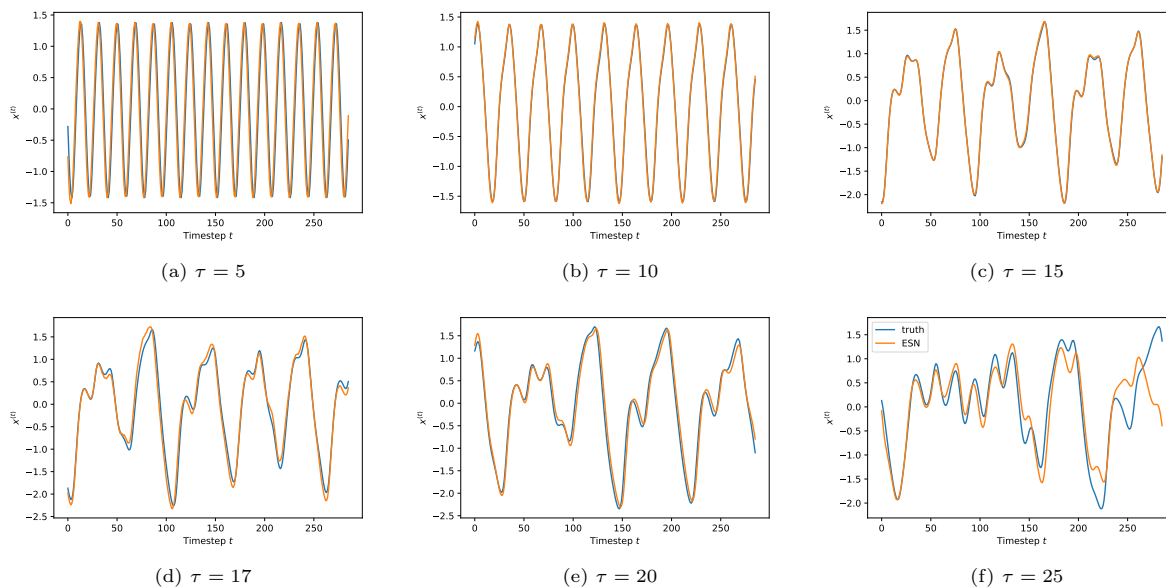


Figure B.7: Exemplary visualization of the predicted Mackey-Glass equation for 286 time steps and different values of τ with ESN.

Appendix C. Number of Valid Time Steps

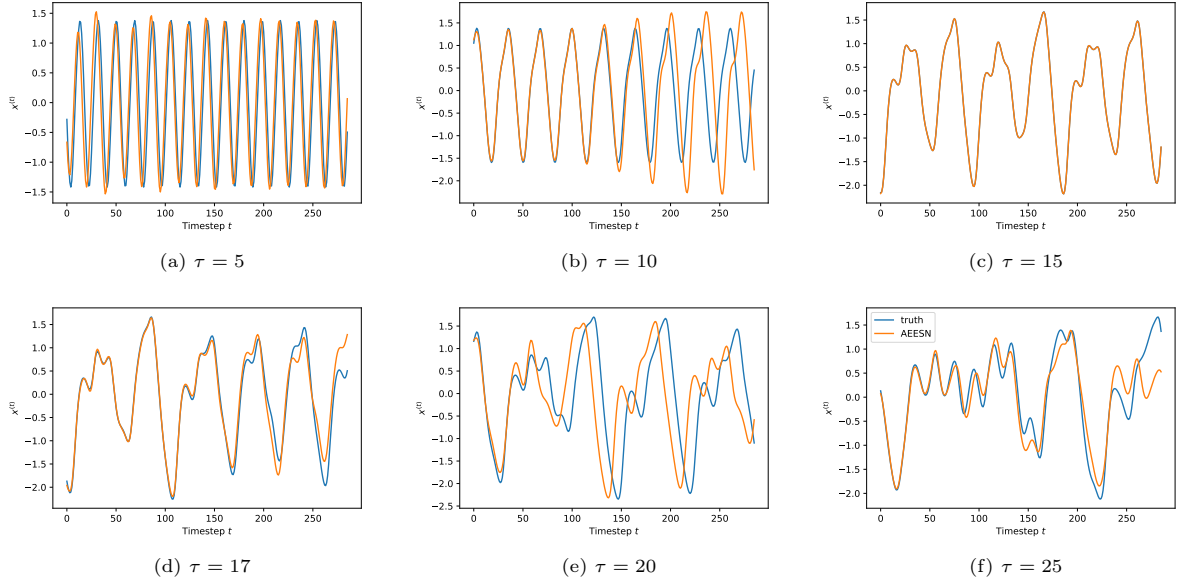


Figure B.8: Exemplary visualization of the predicted Mackey-Glass equation for 286 time steps and different values of τ with AEESN.

Table C.5: Number of valid time steps of best performing Basic ESN, AEESN, GRU and TCRC for different τ of the Mackey-Glass equation, with valid being predictions within one standard deviation difference to the ground truth. We present the best tested $\rho \in [0, 1.5]$ and $N^r \in \{1000, 2000, 3000\}$ as well as best tested input lengths δ for the GRU and TCRC.

τ	ESN	AEESN	GRU	TCRC	TCRC-ELM
5	286	16	286	286	286
10	286	286	286	286	286
15	286	286	0	286	286
17	286	286	115	209	286
20	286	286	30	267	286
25	273	286	286	100	170

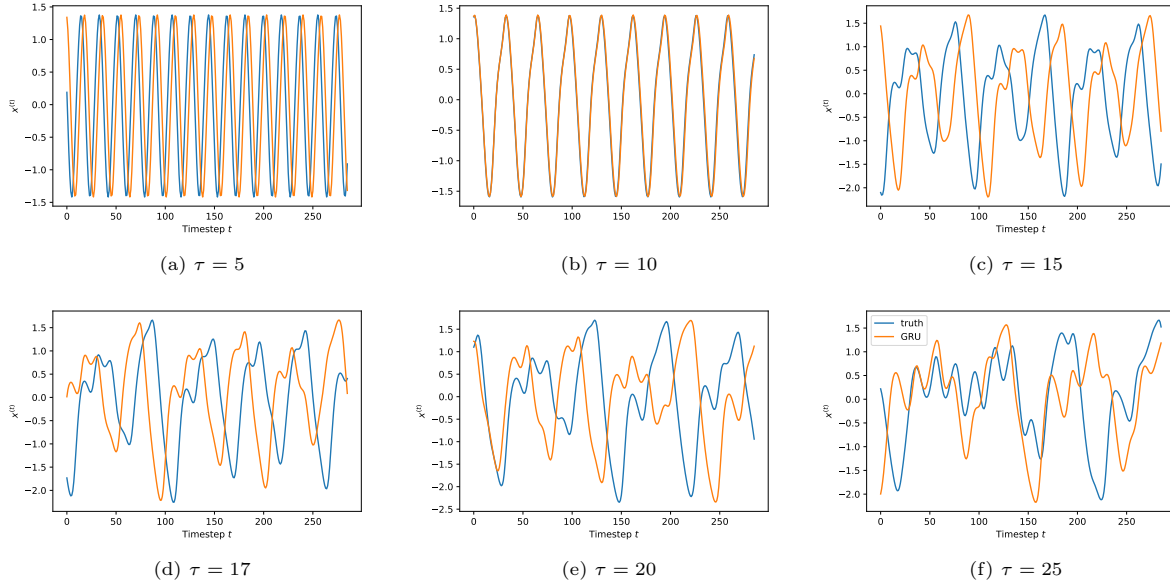


Figure B.9: Exemplary visualization of the predicted Mackey-Glass equation for 286 time steps and different values of τ with GRU.

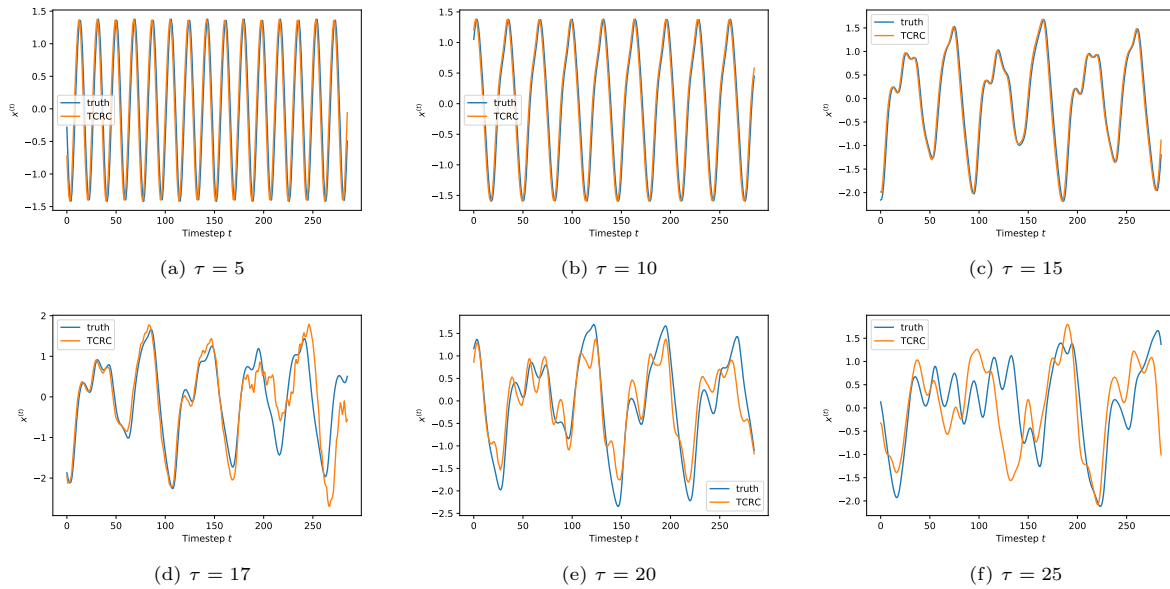


Figure B.10: Exemplary visualization of the predicted Mackey-Glass equation for 286 time steps and different values of τ with TCRC.

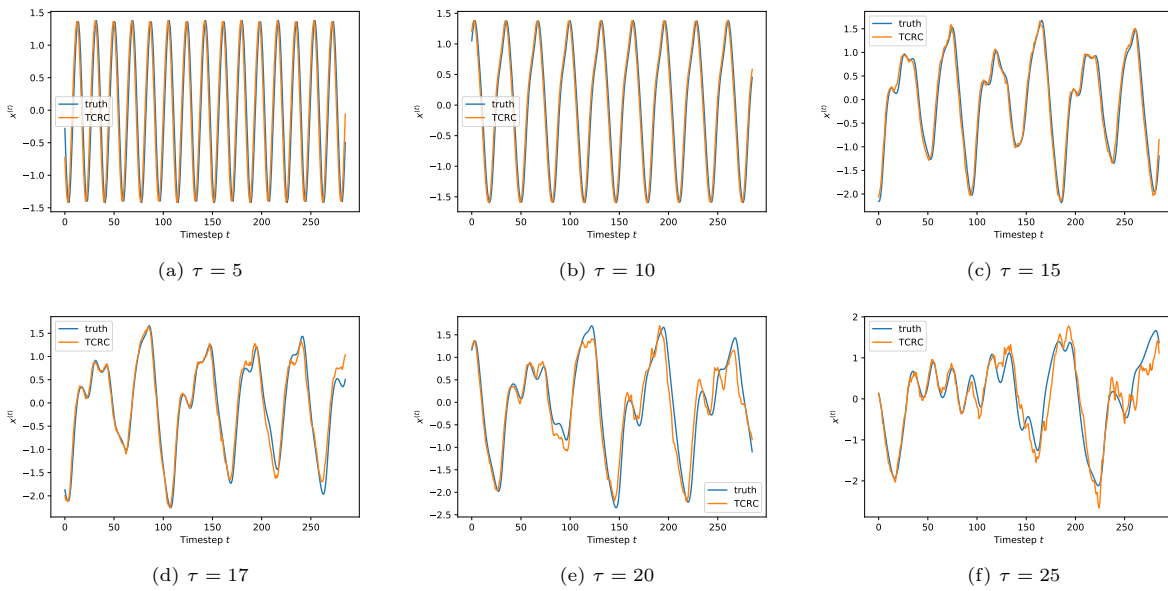


Figure B.11: Exemplary visualization of the predicted Mackey-Glass equation for 286 time steps and different values of τ with TCRC-ELM.