

# Sequential Learning from Noisy Data: Data-Assimilation Meets The Echo-State Network

Debdipta Goswami

**Abstract**—This paper explores the problem of training a recurrent neural network from noisy data. While neural network based dynamic predictors perform well with noise-free training data, prediction with noisy inputs during training phase poses a significant challenge. Here a sequential training algorithm is developed for an echo-state network (ESN) by incorporating noisy observations using an ensemble Kalman filter. The resultant Kalman-trained echo-state network (KalT-ESN) outperforms the traditionally trained ESN with least square algorithm while still being computationally cheap. The proposed method is demonstrated on noisy observations from three systems: two synthetic datasets from chaotic dynamical systems and a set of real-time traffic data.

## I. INTRODUCTION

The current interest in data-driven modeling and forecast of complex systems has motivated a wide array of machine-learning algorithms for different problems, e.g., classification, speech recognition [1], board games [2], and even discovering mathematical algorithms [3]. In particular, recurrent neural networks (RNNs) have proved to be useful in dynamical systems and time-series prediction. For example, an echo-state network (ESN) [4] can model chaotic systems very effectively [5], [6]. However, these prediction techniques rely on relatively noise free training data in order to effectively tune the model parameters. But in many practical cases, the training data is noisy. Such scenarios include atmospheric modeling and traffic systems.

Neural network predictors, instead of using a physics-based handcrafted dynamic model, utilize the rich training dataset to build a parametric surrogate model, and then utilizes it to predict the future states. An ESN is a special type of RNN that uses a reservoir of nonlinear, randomly connected neurons to process time-varying input signal. Such a network with a convergence property, known to the ESN literature as echo-state property (ESP), can uniformly approximate any nonlinear fading memory filter [7]. The ESN is attractive as a neural predictor due to its ability to be tuned via output connections (also called the readout map) with minimal computing resources. Also, a reservoir can be directly implemented by hardwares using field programmable gate arrays (FPGAs) or a photonic reservoir, thereby increasing efficiency and reducing computational overhead [8]. It is also extended to quantum computing realm via quantum reservoir computers (QRCs) [9], [10]. The effectiveness of ESN-based approaches for sparse estimation of chaotic systems and traffic network prediction is shown in [5], [11], and [12].

Debdipta Goswami is an Assistant Professor in the Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, OH 43210. goswami.78@osu.edu

However, the training of an ESN implicitly assumes a high fidelity training dataset representing the actual system states and therefore sensitive to the measurement noise. Although ESN-based methods [11] are developed to utilize noisy measurements to predict the unmeasured states during the testing phase, they rely on relatively clean data for training.

On the other hand, in a model-based estimation problems from noisy measurements, the state estimate is computed in two steps. First, the motion update step facilitated by the model yields a forecast estimate. Then, the final estimate is generated by assimilating a noisy measurement via Bayesian update. For a linear system with Gaussian process and measurement noise, the optimal estimator is given by the celebrated Kalman filter [13], whereas for a nonlinear system, optimal filtering is usually infinite dimensional requiring the solution of a stochastic partial differential equation. A variety of suboptimal finite dimensional techniques are usually employed for nonlinear estimation, e.g., the extended Kalman filter (EKF) [14], unscented Kalman filter (UKF) [15], ensemble Kalman filter (EnKF), and particle filter. However, these methods require a dynamic model to perform the motion update of the state estimate.

This paper develops a sequential training method for an ESN by combining its strength of modeling an unknown dynamical system with a nonlinear filtering algorithm. An ESN is particularly well suited for modeling a dynamical system with faster and cheaper training that does not require backpropagation through time (BPTT). An ESN with fading memory can universally model nonlinear dynamics [16], [7]. The ESN architecture adopts an input-output neural network with a randomly generated recurrent reservoir where only output layer is trained. The training is usually done via a least square linear regression that implicitly assumes noise free training data. The proposed method transforms the training problem of an ESN to a combined state-parameter estimation problem where the output weights are sequentially updated with the incoming noisy observations along with the system states.

The combined estimation problem is, however, no longer linear. Hence, a nonlinear data-assimilation method is required for the measurement update. While the EKF and UKF perform well in model-based scenarios, the computation of the linearized dynamics is challenging for an ESN. The ensemble Kalman filter [17] is thus chosen for the measurement update for its strength in representing the posterior distribution of states by its sample means and covariances. The resulting algorithm is called Kalman training of the echo-state network (KalT-ESN). The algorithm is also extended

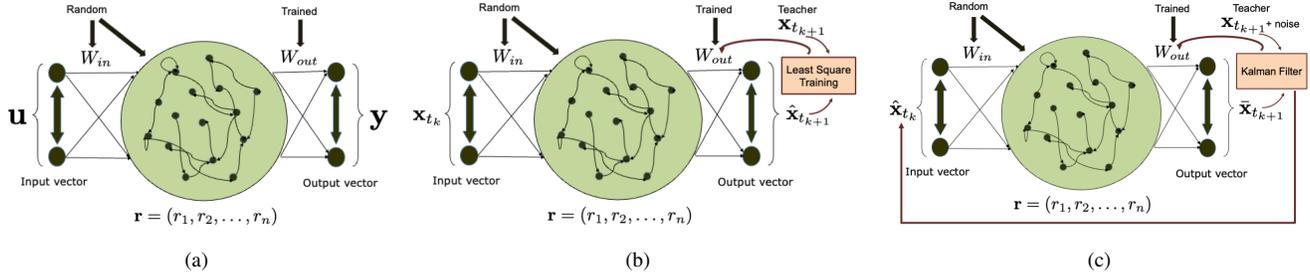


Fig. 1: Architecture and training of an ESN: (a) the basic ESN, (b) least square training from a batch of noiseless data, and (c) Kalman filter based sequential training from noisy data

to partial noisy measurements as training data by using a delay-embedding at the input layer [18].

The contributions of this paper are (1) providing a sequential training algorithm for an ESN via combined state and parameter estimation when the training data is noisy; (2) combining the prediction power of a recurrent neural network with the traditional Bayesian measurement update model of an ensemble Kalman filter; (3) improving the prediction accuracy over time for a chaotic nonlinear system from its noisy measurements; (4) extension of the algorithm for noisy partial measurements as the training data by delay-embedding; and (5) application of the proposed method to a real set of mobility data in order to predict daily cycles of traffic volume. The model-free estimation algorithm developed here has wide applications for estimation of complex dynamics from noisy observations when a reliable model is unavailable.

The paper is organized as follows. Section II provides a brief overview of the echo-state network (ESN). Section III presents the combined state and parameter estimation problem and the KalT-ESN algorithm. Section IV illustrates the applications to three different problems: two synthetic data streams generated by chaotic nonlinear systems and one real set of data traffic sensor data. Section V concludes the manuscript and discusses ongoing and future work.

## II. ECHO-STATE NETWORKS: A UNIVERSAL PREDICTOR

Echo-state networks are a type of recurrent neural network used for dynamical systems prediction. It consists of a large dynamic reservoir of randomly connected neurons driven nonlinearly by input signals. These neuronal responses are then linearly combined to match a desired output signal. Due to its dependence on the richness of the dynamic reservoir, it is also called a reservoir computer (RC). An ESN consists of an input layer  $\mathbf{u} \in \mathbb{R}^m$ , coupled through input coupling matrix  $W_{in} \in \mathbb{R}^{n \times m}$  with a recurrent nonlinear reservoir  $\mathbf{r} \in \mathbb{R}^n$ . The output  $\mathbf{y} \in \mathbb{R}^p$  is generated from  $n$  neurons of the reservoir via a readout matrix  $W_{out} \in \mathbb{R}^{n \times p}$ . The reservoir network evolves nonlinearly in following fashion [16], [11]

$$\mathbf{r}(t + \Delta t) = (1 - \alpha)\mathbf{r}(t) + \alpha\psi(W\mathbf{r}(t) + W_{in}\mathbf{u}(t)). \quad (1)$$

The time-step  $\Delta t$  denotes the sampling interval of the training data. The leakage rate parameter  $\alpha \in (0, 1]$  helps slowing down the evolution of the reservoir states as  $\alpha \rightarrow 0$ . The nonlinear activation function  $\psi(\cdot)$  is usually a sigmoid function, e.g.,  $\tanh(\cdot)$ . The output  $\mathbf{y}(t)$  is linearly read out from the reservoir states [16], [11], i.e.,

$$\mathbf{y}(t) = W_{out}\mathbf{r}(t). \quad (2)$$

The weights  $W_{in}$  and  $W$  are initially randomly drawn and then held fixed. The weight  $W_{out}$  is adjusted during the training process. The reservoir weight matrix  $W$  is usually kept sparse for computational efficiency.

During the training phase, an ESN is driven by an input sequence  $\{\mathbf{u}(t_1), \dots, \mathbf{u}(t_N)\}$  that yields a sequence of reservoir states  $\{\mathbf{r}(t_1), \dots, \mathbf{r}(t_N)\}$ . The reservoir states are stored in a matrix  $\mathbf{R} = [\mathbf{r}(t_1), \dots, \mathbf{r}(t_N)]$ . The correct outputs  $\{\mathbf{y}(t_1), \dots, \mathbf{y}(t_N)\}$ , which are part of the training data, are also arranged in a matrix  $\mathbf{Y} = [\mathbf{y}(t_1), \dots, \mathbf{y}(t_N)]$ . The training is carried out by a linear regression with Tikhonov regularization as follows [4]:

$$W_{out} = \mathbf{Y}\mathbf{R}^T(\mathbf{R}\mathbf{R}^T + \beta\mathbf{I})^{-1}, \quad (3)$$

where  $\beta > 0$  is a regularization parameter that ensures non-singularity.

**Remark 1.** For an ESN to be a universal approximator, i.e., to be able to realize any nonlinear operator with bounded memory arbitrarily accurately, it must satisfy the echo-state property (ESP) [4]. An ESN is said to have the ESP if the reservoir asymptotically washes out any information from the initial conditions. For the  $\tanh(\cdot)$  activation function, it is empirically observed that the ESP holds for any input if the spectral radius of  $W$  is smaller than unity [4]. To ensure this condition,  $W$  is normalized by its spectral radius.

## III. SEQUENTIAL TRAINING OF AN ECHO-STATE NETWORK: AN ENSEMBLE KALMAN FILTER APPROACH

An ESN can be trained to predict a time-series  $\{\mathbf{x}(t_i) \in \mathbb{R}^d : i \in \mathbb{N}\}$  generated by a dynamical system by setting  $\mathbf{u}(t)$  and  $\mathbf{y}(t)$  as the current and next state value (i.e.,  $\mathbf{x}(t_k)$  and  $\mathbf{x}(t_{k+1})$ ) respectively. The network is trained for a certain training length  $N$  of the time-series data  $\{\mathbf{x}(t_i, i = 1, \dots, N)\}$ , which can then run freely by feeding the output

---

**Algorithm 1** KalT-ESN: Sequential Training of an ESN with EnKF
 

---

**Input:** Noisy training data  $\{\mathbf{y}(t_1), \dots, \mathbf{y}(t_N)\}$ ,  $\mathbf{y}(t_i) \in \mathbb{R}^d$

**Hyperparameters:** Training length  $N$ , leaking rate  $\alpha$ , regularization parameter  $\beta$ , reservoir connection probability  $p \in (0, 1)$ , reservoir size  $n$ , activation  $\psi$ , ensemble size  $M$ , initial data covariance  $\Sigma_x \in \mathbb{R}^{d \times d}$ , initial weight covariance  $\Sigma_w \in \mathbb{R}^{dn \times dn}$ , measurement noise covariance  $\Sigma_v \in \mathbb{R}^{d \times d}$

**Output:**  $W_{in}, W, W_{out}$

```

1: procedure TRAIN( $\{\mathbf{y}(t_1), \dots, \mathbf{y}(t_N)\}$ ;  $\alpha, \beta, p, n, \psi, M, \Sigma_x, \Sigma_w, \Sigma_v$ )
2:   Generate  $W \sim G(n, p)$   $\triangleright$  Adjacency matrix of an Erdős-Renyi random graph
3:   Generate  $W_{in} \in \mathbb{R}^{n \times d}$  random matrix
4:   Generate  $\hat{\mathbf{X}}(t_1) = [\hat{\mathbf{x}}^{(1)}(t_1), \dots, \hat{\mathbf{x}}^{(M)}(t_1)] \sim \mathcal{N}(\mathbf{y}(t_1), \Sigma_x)$   $\triangleright$  Data ensemble generation
5:   Generate  $\mathbf{R}(t_1) = [\mathbf{r}^{(1)}(t_1), \dots, \mathbf{r}^{(M)}(t_1)] = \mathbf{0}_{n \times M}$ 
6:   Generate  $\hat{\mathbf{W}}_o(t_1) = [\hat{\mathbf{w}}_o^{(1)}(t_1), \dots, \hat{\mathbf{w}}_o^{(M)}(t_1)] \sim \mathcal{N}(\mathbf{0}_{dn}, \Sigma_w)$   $\triangleright$  Weight ensemble generation
7:   for  $k = 1$  to  $N - 1$  do
8:     for  $i = 1$  to  $M$  do
9:        $\hat{W}_o^{(i)}(t_k) \leftarrow \text{reshape}(\hat{\mathbf{w}}_o^{(i)}(t_k), d, n)$ 
10:       $\bar{\mathbf{x}}^{(i)}(t_{k+1}) \leftarrow \hat{W}_o^{(i)}(t_k) \psi(W \mathbf{r}^{(i)}(t_k) + W_{in} \hat{\mathbf{x}}^{(i)}(t_k))$ 
11:    end for
12:     $\mathbf{X}^f(t_{k+1}) \leftarrow [\bar{\mathbf{x}}^{(1)}(t_{k+1}), \dots, \bar{\mathbf{x}}^{(M)}(t_{k+1})]$ 
13:     $\mathbf{W}_o^f(t_{k+1}) \leftarrow \mathbf{W}_o(t_k)$ 
14:     $\mathbf{E}_w(t_{k+1}) = (\mathbf{W}_o^f(t_{k+1}) - \bar{\mathbf{W}}_o^f(t_{k+1}))$ 
15:     $\mathbf{E}_x(t_{k+1}) = (\mathbf{X}^f(t_{k+1}) - \bar{\mathbf{X}}^f(t_{k+1}))$ 
16:     $P_{wx}(t_{k+1}) \leftarrow \mathbf{E}_w(t_{k+1}) \mathbf{E}_x(t_{k+1})^T / (M - 1)$ 
17:     $P_{xx}(t_{k+1}) \leftarrow \mathbf{E}_x(t_{k+1}) \mathbf{E}_x(t_{k+1})^T / (M - 1)$ 
18:     $K^x(t_{k+1}) \leftarrow P_{xx}(t_{k+1}) [P_{xx}(t_{k+1}) + \Sigma_v]^{-1}$ 
19:     $K^w(t_{k+1}) \leftarrow P_{wx}(t_{k+1}) [P_{xx}(t_{k+1}) + \Sigma_v]^{-1}$ 
20:     $\hat{\mathbf{X}}(t_{k+1}) \leftarrow \mathbf{X}^f(t_{k+1}) + K^x(t_{k+1}) (\mathbf{y}(t_{k+1}) - \mathbf{X}^f(t_{k+1}))$ 
21:     $\hat{\mathbf{W}}_o(t_{k+1}) \leftarrow \mathbf{W}_o^f(t_{k+1}) + K^w(t_{k+1}) (\mathbf{y}(t_{k+1}) - \mathbf{X}^f(t_{k+1}))$ 
22:  end for
23:   $W_{out} \leftarrow \text{reshape}(\bar{\mathbf{W}}_o(t_N), d, n)$   $\triangleright$  Final output weights
24: end procedure

```

---

$\mathbf{y}(t_k)$  back to the input  $\mathbf{u}(t_{k+1})$  of the reservoir. In this case, both  $\mathbf{u}$  and  $\mathbf{y}$  have the same dimension  $d$  as that of the time-series data. This setup is shown in Fig. 1(a)-(b).

An ESN proves to be a powerful tool for dynamical systems prediction when trained with noiseless data [5], [6]. Its performance is significantly improved when partial observations are available during the testing phase by assimilating them through an ensemble Kalman filter [11]. It can also be modified to accommodate partial state measurements as training data by a higher dimensional delay-embedding in the input layer [18]. However, a significant challenge is posed when noisy data is present during the training phase. A simple approach is to ignore the noise and train the network using the least square method (3). But this simplistic approach disregards the fact that the training data might contain rapid variations and outliers that might impart large errors in the weights of the ESN. The problem becomes particularly worse for chaotic time-series where the training noise can drive the network away from the chaotic attractor.

This paper proposes an alternative method for training the ESN from noisy data by incorporating a filtering step

recursively in the training process. We assume that the training data  $[\mathbf{y}(t_1), \dots, \mathbf{y}(t_N)]$  can be modeled as  $\mathbf{y}(t_k) = \mathbf{x}(t_k) + \mathbf{v}_k$  with  $\mathbf{v}_k$  as a zero-mean i.i.d. Gaussian random variable with covariance  $\Sigma_v$ . A combined state and parameter estimation method is then used via state augmentation. The forecast model is formulated as

$$\mathbf{r}(t_{k+1}) = (1 - \alpha) \mathbf{r}(t_k) + \alpha \psi(W \mathbf{r}(t_k) + W_{in} \hat{\mathbf{x}}(t_k)) \quad (4)$$

$$\begin{aligned} \bar{\mathbf{x}}(t_{k+1}) &= \hat{W}_{out}(t_k) \mathbf{r}(t_{k+1}) \\ \bar{W}_{out}(t_{k+1}) &= \hat{W}_{out}(t_k) \end{aligned}$$

where  $\mathbf{r}(t_k)$  is the known reservoir state,  $\hat{\mathbf{x}}(t_k)$  and  $\hat{W}_{out}(t_k)$  are the estimates of the true data and output weight at time  $t_k$ . Now, since we can observe the true  $\mathbf{r}(t_k)$ , we can define  $\Phi_{t_k}^x(\cdot) \triangleq (1 - \alpha) \mathbf{r}(t_k) + \alpha \psi(W \mathbf{r}(t_k) + W_{in}(\cdot))$  so that the forecast model (4) becomes

$$\begin{aligned} \bar{\mathbf{x}}(t_{k+1}) &= \hat{W}_{out}(t_k) \Phi_{t_k}^x(\hat{\mathbf{x}}(t_k)) \\ \bar{W}_{out}(t_{k+1}) &= \hat{W}_{out}(t_k), \end{aligned} \quad (5)$$

with  $\bar{\mathbf{x}}(t_{k+1})$  and  $\bar{W}_{out}(t_{k+1})$  as the data and output weight forecast at  $t_{k+1}$  respectively. Thus, we now have a dynamic forecast model (5) which can be formulated into a Kalman filter step for the analysis part. First, we vectorize the  $W_{out}$  matrix as  $\underline{\mathbf{w}}_o \in \mathbb{R}^{nd}$  and define the new augmented state variable  $\mathbf{z} = [\mathbf{x}^T \quad \underline{\mathbf{w}}_o^T]^T$ . The forecast model (5) with this augmented variable  $\mathbf{z}$  becomes

$$\bar{\mathbf{z}}(t_{k+1}) = \Phi_{t_k}(\hat{\mathbf{z}}(t_k)), \quad (6)$$

where

$$\Phi_{t_k}(\mathbf{z}) = \begin{bmatrix} W_{out} \Phi_{t_k}^x(\mathbf{x}) \\ \underline{\mathbf{w}}_o \end{bmatrix},$$

with  $W_{out}$  as  $\underline{\mathbf{w}}_o$  reshaped into a  $d \times n$  matrix. Since this is a nonlinear forecast model, a nonlinear modification of the Kalman filter algorithm is employed to jointly estimate the true training data  $\mathbf{x}(t_k)$  and the output weight  $W_{out}$ . In particular, an ensemble Kalman filter (EnKF) is utilized to estimate the covariance of  $\mathbf{z}(t_k)$  from a Gaussian ensemble realization. An initial ensemble  $\hat{\mathbf{Z}}(t_1) = [\hat{\mathbf{z}}^{(1)}(t_1), \dots, \hat{\mathbf{z}}^{(M)}(t_1)]$  is generated with members drawn from  $\mathcal{N}([\mathbf{y}(t_1)^T \quad \mathbf{0}_{nd}^T]^T, \Sigma_z)$ . The initial covariance  $\Sigma_z$  is a hyperparameter and assumed to be block diagonal, i.e.,  $\Sigma_z = \text{bdiag}(\Sigma_x, \Sigma_w)$ . Algorithm 1 presents the procedure for the sequential training of an ESN with EnKF. The inputs are the noisy training data  $[\mathbf{y}(t_1), \dots, \mathbf{y}(t_N)]$ . After  $N$  training steps, the procedure returns the trained output matrix  $W_{out}$ .

**Remark 2.** A block diagonal  $\Sigma_z$  with only direct measurements of the state  $\mathbf{x}$  explicitly separates the Kalman update step between the state  $\mathbf{x}$  and parameters  $\underline{\mathbf{w}}_o$  as shown in Algorithm 1, line 18-19.

**Remark 3.** In addition to noisy training data, Algorithm 1 works with sequentially available data in a recursive fashion, unlike the static least square training (3).

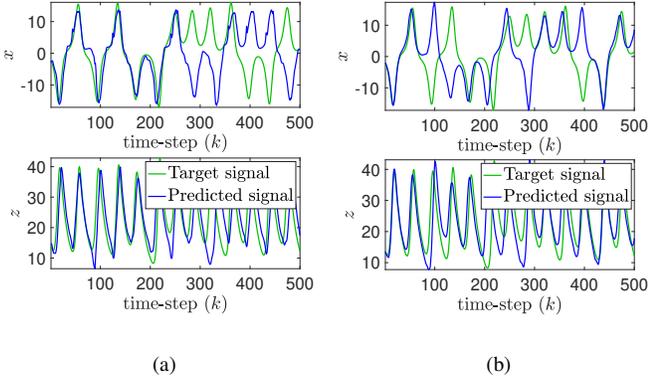


Fig. 2: prediction of the noisy time-series  $x(t_k)$  and  $z(t_k)$  from Lorenz system (7) with  $\sigma_v^2 = 1.0$ : (a) true and predicted signal with KalT-ESN, (b) true and predicted signal with least square training

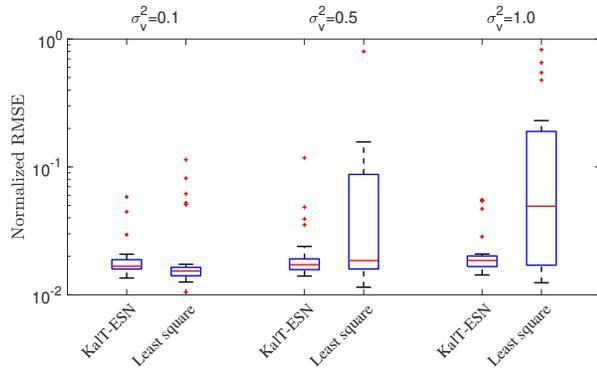


Fig. 3: Error profile of Lorenz time-series prediction: NRMSE with different measurement noise covariance  $\sigma_v^2$

**Remark 4.** The Kalman based training of an ESN can assimilate the partial state measurements during the testing phase to keep improving the output weights  $W_{out}$  and the subsequent time series estimates  $\hat{\mathbf{x}}(t_k)$ .

#### IV. NUMERICAL EXAMPLES

This section illustrates the performance and ablation study of the KalT-ESN algorithm on three time series data corrupted with noise. The first two are time-series generated by chaotic dynamical systems and the last one is a real-time traffic flow data obtained by Numina sensor nodes [19] installed on the University of Maryland campus.

##### A. Lorenz System

The KalT-ESN algorithm is tested on a time-series  $[x(t_k) \ y(t_k) \ z(t_k)]$  generated by the Lorenz system:

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z) - y \\ \dot{z} &= xy - \beta z, \end{aligned} \quad (7)$$

where  $\sigma = 10$ ,  $\rho = 28$ , and  $\beta = 8/3$  produces chaotic behavior. The training data is corrupted by a measurement noise  $\mathbf{v}(t_k) \sim \mathcal{N}(0, \sigma_v^2 I_{3 \times 3})$ . Table I lists the hyperparameters used to train the ESN. The prediction via KalT-ESN and least square training is depicted in Fig. 2. Fig. 3 provides a

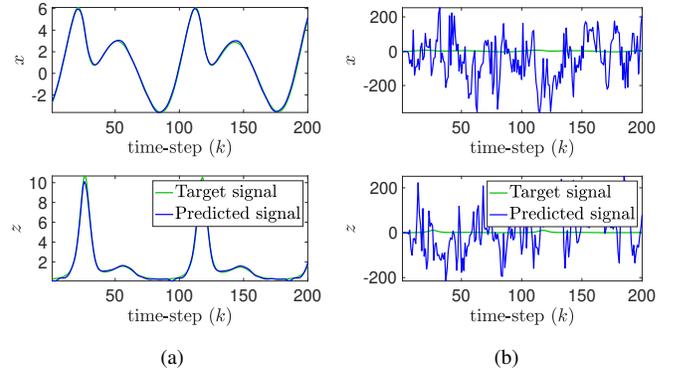


Fig. 4: prediction of the noisy time-series  $x(t_k)$  and  $z(t_k)$  from Rössler system (9) with  $\sigma_v^2 = 0.1$ : (a) true and predicted signal with KalT-ESN, (b) true and predicted signal with least square training

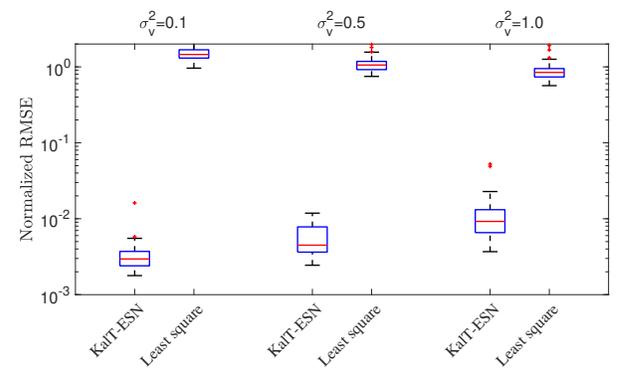


Fig. 5: Error profile of Rössler time-series prediction: NRMSE with different measurement noise covariance  $\sigma_v^2$

detailed error profile for measurement noise covariance  $\sigma_v^2$ . The normalized root mean square error (NRMSE) between the true sequence  $\{\mathbf{x}(t_k) : i = 1, \dots, l\}$  and the predicted sequence  $\{\hat{\mathbf{x}}(t_k) : i = 1, \dots, l\}$  is given by

$$\text{NRMSE}(\mathbf{x}, \hat{\mathbf{x}}) = \sqrt{\frac{\sum_{k=1}^l \|\mathbf{x}(t_k) - \hat{\mathbf{x}}(t_k)\|^2}{\sum_{k=1}^l \|\mathbf{x}(t_k)\|^2}}, \quad (8)$$

where  $l$  is the prediction length. The prediction NRMSEs for different measurement noise covariances over 30 independent Monte-Carlo trials are plotted in Fig. 3. The performance of KalT-ESN algorithm remains consistent with different  $\sigma_v$  while the performance of the least square training degrades heavily with higher noise covariance.

##### B. Rössler System

Next, KalT-ESN is utilized to predict the noisy state measurement generated by the Rössler system described in [20]:

$$\begin{aligned} \dot{x} &= -y - x \\ \dot{y} &= x + ay \\ \dot{z} &= b + z(x - c), \end{aligned} \quad (9)$$

TABLE I: ESN hyperparameters

Hyperparameter	Value		
	Lorenz system (7)	Rössler system (9)	Traffic Volume
Time step $\Delta t$	0.02s	0.1s	1h
Reservoir size $n$	500	500	4000
Reservoir connection probability $p$	0.01	0.01	0.01
Training length $N$	6000	1000	500
Activation $\psi(\cdot)$	$\tanh(\cdot)$	$\tanh(\cdot)$	$\tanh(\cdot)$
Leaking rate $\alpha$	0.3	0.3	0.7
Regularization $\beta$	$10^{-6}$	$10^{-6}$	$10^{-6}$
Ensemble size $M$	300	300	300
Initial state covariance $\Sigma_x = \sigma_x^2 I$	$\sigma_x^2 = 0.2$	$\sigma_x^2 = 0.2$	$\sigma_x^2 = 10$
Initial weight covariance $\Sigma_w = \sigma_w^2 I$	$\sigma_w^2 = 0.2$	$\sigma_w^2 = 0.2$	$\sigma_w^2 = 1$

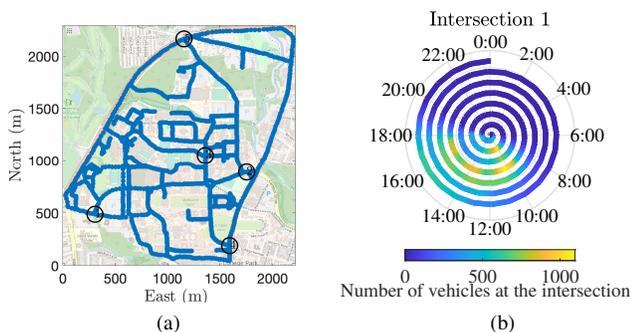


Fig. 6: Schematic diagram of traffic data: (a) University of Maryland road network with Numina sensors, (b) Traffic congestion pattern of an intersection over a single week, each revolution denotes a day of the week with times marked as angles; the number of vehicles is denoted by the colormap. The daily pattern of peak congestion between mornings and afternoons is evident.

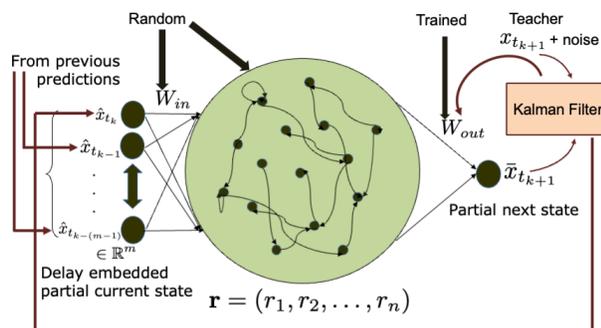


Fig. 7:  $m$ -dimensional delay-embedding of the scalar partial observation for KalT-ESN

with  $a = 0.5$ ,  $b = 2$ , and  $c = 4$  to produce chaotic behavior. Similar to the Lorenz system example, the training data is corrupted by a measurement noise  $\mathbf{v}(t_k) \sim \mathcal{N}(0, \sigma_v^2 I_{3 \times 3})$ . Table I lists the hyperparameters used to train the ESN. The prediction via KalT-ESN and least square training is depicted in Fig. 4. Fig. 5(b) plots the detailed error profile for measurement noise covariance  $\sigma_v^2$ . The results are generated by 30 independent Monte-Carlo trials for training and testing the ESNs.

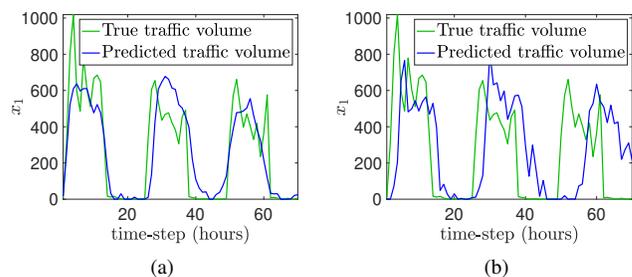


Fig. 8: prediction of the noisy time-series of traffic volume recorded in Numina sensor 1 with  $\sigma_v^2 = 500$ : (a) true and predicted traffic volume with KalT-ESN, (b) true and predicted traffic volume with least square training

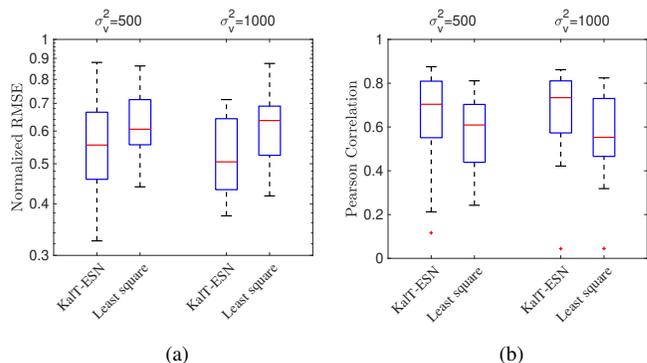


Fig. 9: Error and correlation profile of traffic volume prediction: (a) NRMSE and (b) Pearson correlation with different measurement noise covariance  $\sigma_v^2$ . The delay embedding dimension is  $m = 10$ .

### C. Prediction of Traffic Volume on an Intersection of a Road Network

KalT-ESN is now applied to a dataset of traffic volumes obtained from Numina [19] sensors at five different intersections on the University of Maryland campus. Fig. 6(a) represents the road network marked with sensor locations. Each sensor counts the number of pedestrians, bicycles, and vehicles at the respective intersections and store them in a server. The time series data of hourly vehicle traffic volume for two months is used. Fig. 6(b) represents the hourly vehicle traffic volume over a week with a clear daily pattern.

The traffic dynamics is an infinite-dimensional spatio-temporal dynamical system evolving over a road network, and hence, the traffic volumes recorded from each sensor provides a partial measurement. An ESN usually requires full state measurements in the training phase [11], [5]. To mitigate this problem, a delay-embedding in the input layer [18] is used for training. The delay-embedding for KalT-ESN is demonstrated in Fig. 7. Both least square training and KalT-ESN is applied on this delay-embedded time series. Only a noisy scalar measurement from sensor 1 is used in this paper with embedding dimension  $m = 10$ .

The ESN is trained on 500 hours of traffic volume data and tested for 70 hours, i.e., approximately three days. The training hyperparameters are listed in Table I. Fig. 8 shows the traffic volume prediction by KalT-ESN and least square training. Fig. 9 shows the NRMSE and Pearson correlation coefficient between predicted and true traffic volumes with sensor data from intersection 1 corrupted with noise. The results are similar for the other four intersections and not included here. The Pearson correlation coefficient between true and predicted sequences ( $\{x(i) : i = 1, \dots, l\}$  and  $\{\hat{x}(i) : i = 1, \dots, l\}$  respectively) measures their normalized linear correlation. It is given by

$$r(x, \hat{x}) = \frac{\sum_k (x(t_k) - \bar{x})^T (\hat{x}(t_k) - \bar{\hat{x}})}{\sqrt{\sum_i \|x(t_k) - \bar{x}\|^2} \sqrt{\sum_k \|\hat{x}(t_k) - \bar{\hat{x}}\|^2}}, \quad (10)$$

where  $\bar{x}$  and  $\bar{\hat{x}}$  denotes the time-average values of  $x(t_k)$  and  $\hat{x}(t_k)$ . KalT-ESN yields improved NRMSE and higher Pearson correlation coefficient with an embedding dimension of  $m = 10$  only.

**Remark 5.** The delay embedded KalT-ESN provides a practical way to train an ESN from both noisy and partial state measurement, and hence, presents a natural continuation of the delay-embedded ESN presented in [18].

## V. CONCLUSION

This paper proposes a sequential training algorithm for an echo-state network (ESN) that combines the power of universal prediction by an ESN with data-assimilation by ensemble Kalman filter (EnKF). The algorithm, called Kalman training of the echo-state network (KalT-ESN), recursively updates the output weights of an ESN using an EnKF from the noisy training data. The proposed training algorithm demonstrates improved performance in presence of additive noise in the training dataset. It is extended for partial noisy measurements in the training phase using a time-delay embedding at the input layer. The method is then applied to a real data set of traffic patterns on the road network of the University of Maryland, College Park campus to predict the traffic volume at various intersections. For ongoing and future work, inference of unobserved states via time-delay embedded ESN with surrogate spatial interpolation model will be investigated.

## ACKNOWLEDGEMENT

The author thanks Dr. Derek A. Paley and the University of Maryland Department of Transportation for the Numina sensor data.

## REFERENCES

- [1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.
- [2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [3] A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatin, A. Novikov, F. J. R. Ruiz, J. Schrittwieser, G. Swirszcz, D. Hassabis, and P. Kohli, "Discovering faster matrix multiplication algorithms with reinforcement learning," *Nature*, vol. 610, no. 7930, pp. 47–53, 2022.
- [4] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science*, vol. 304, no. 5667, pp. 78–80, 2004.
- [5] Z. Lu, J. Pathak, B. Hunt, M. Girvan, R. Brockett, and E. Ott, "Reservoir observers: Model-free inference of unmeasured variables in chaotic systems," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 4, pp. 1054–1500, 2017.
- [6] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, "Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach," *Phys. Rev. Lett.*, vol. 120, p. 024102, Jan 2018.
- [7] L. Grigoryeva and J.-P. Ortega, "Echo state networks are universal," *Neural Networks*, vol. 108, pp. 495 – 508, 2018.
- [8] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, "Recent advances in physical reservoir computing: A review," *Neural Networks*, vol. 115, pp. 100–123, 2019.
- [9] K. Fujii and K. Nakajima, "Harnessing disordered-ensemble quantum dynamics for machine learning," *Phys. Rev. Applied*, vol. 8, p. 024030, Aug 2017.
- [10] J. Chen, H. I. Nurdin, and N. Yamamoto, "Towards single-input single-output nonlinear system identification and signal processing on near-term quantum computers," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 401–406.
- [11] D. Goswami, A. Wolek, and D. A. Paley, "Data-driven estimation using an echo-state neural network equipped with an ensemble kalman filter," in *2021 American Control Conference (ACC)*, 2021, pp. 2549–2554.
- [12] D. Goswami, A. Riggins, and D. A. Paley, "Data-driven prediction of urban micromobility: A study of dockless electric scooters [applications of control]," *IEEE Control Systems Magazine*, vol. 42, no. 5, pp. 18–31, 2022.
- [13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME—Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.
- [14] H. K. Khalil, *Nonlinear Systems*. Upper Saddle River, (N.J.): Prentice Hall, 1996.
- [15] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.
- [16] W. Maass and H. Markram, "On the computational power of circuits of spiking neurons," *Journal of Computer and System Sciences*, vol. 69, no. 4, pp. 593 – 616, 2004.
- [17] T. Bengtsson, C. Snyder, and D. Nychka, "Toward a nonlinear ensemble filter for high-dimensional systems," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D24, pp. n/a–n/a, 2003, 8775.
- [18] D. Goswami, "Delay embedded echo-state network: A predictor for partially observed systems," *arXiv preprint arXiv:2211.05992*, 2022.
- [19] Numina, "Numina mobility solutions," <https://numina.co>, 2019.
- [20] O. RöSSLer, "An equation for continuous chaos," *Physics Letters A*, vol. 57, no. 5, pp. 397–398, 1976.