Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.160

An ensemble quadratic echo state network for non-linear spatio-temporal forecasting

Patrick L. McDermott* and Christopher K. Wikle

Received 23 August 2017; Accepted 30 August 2017

Spatio-temporal data and processes are prevalent across a wide variety of scientific disciplines. These processes are often characterized by non-linear time dynamics that include interactions across multiple scales of spatial and temporal variability. The datasets associated with many of these processes are increasing in size because of advances in automated data measurement, management and numerical simulator output. Non-linear spatio-temporal models have only recently seen interest in statistics, but there are many classes of such models in the engineering and geophysical sciences. Traditionally, these models are more heuristic than those that have been presented in the statistics literature but are often intuitive and quite efficient computationally. We show here that with fairly simple, but important, enhancements, the echo state network machine learning approach can be used to generate long-lead forecasts of non-linear spatio-temporal processes, with reasonable uncertainty quantification, and at a fraction of the computational expense of a traditional parametric non-linear spatio-temporal models. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: general quadratic non-linearity; long-lead forecasting; recurrent neural network; reservoir computing; sea surface temperature

1 Introduction

Spatio-temporal processes in the natural world often exhibit non-linear behaviour such as growth through time, frontal boundaries, density dependence, shock waves, repulsion, non-linear advection and predator–prey interactions, to name a few. Although at some scales in space and time, linear models can be quite effective for predicting these processes, it is sometimes essential to consider the non-linearity explicitly. This is especially true when forecasting at very short or very long timescales or when interpolating data where there are large gaps and active dynamical processes (e.g. ocean eddies). Indeed, non-linearity is often the cause of marginal non-Gaussianity and extremes in such data. Historically, there has not been much development of parametric models for non-linear dynamic spatio-temporal processes in the statistics literature, but it has recently seen increasing interest (e.g. Wikle & Hooten, 2010; Wikle & Holan, 2011; Gladish & Wikle, 2014; Richardson, 2017). These models have emphasized state-dependent transition operators and the importance of quadratic interactions, which can be shown to be a fundamental property of many physical and biological processes. Although these frameworks are quite general and can be effective, they suffer from a curse of dimensionality in parameter space and require careful attention to reduce the effective number of parameters through mechanistically motivated (hard) shrinkage, regularization and/or state reduction (see Wikle, 2015, for an overview). In addition, even when one controls for the large number of parameters in these models,

Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, USA
***Email: plmyt7@mail.missouri.edu**

they can still be quite expensive computationally. Thus, it is of interest to consider the so-called black-box parametric approaches for spatio-temporal models that can accommodate non-linearity, while retaining computational efficiency.

The growth of "deep learning" methods in the machine learning literature suggests that these approaches may be suitable as an efficient "black-box" model to accommodate non-linear spatio-temporal dynamics. Although the standard feedforward neural network is not able to account for the time dependence present in such processes, the recurrent neural network (RNN) approach developed in the engineering literature in the 1980s (e.g. see the review in Lukoševičius & Jaeger, 2009), and popularized more recently in machine learning, is designed to allow cycles and sequences in their hidden layers. These methods have not been used much for spatio-temporal prediction, but they have been used extensively in natural language processing applications, where the sequence of words, and hence their temporal dependence, is fundamentally important.

Traditionally, RNN models have been fairly difficult to fit in settings such as the typical spatio-temporal forecasting problem due to the so-called vanishing gradient problem in the back-propagation algorithm used to obtain weights in the hidden layers. Because of this, two varieties of RNNs have been developed to minimize the number of weights that need to be trained, the so-called echo state networks (ESNs; Jaeger, 2007) and liquid state machines (Maass et al., 2002). These approaches, which are now often labelled more generally as "reservoir computing" methods, consider sparsely connected hidden layers that allow for sequential interactions. In addition, a crucial component of such reservoir models is that the connectivity and the weights for the hidden units are *fixed* yet *randomly assigned*! That is, the input data go into a hidden fixed "reservoir" that contains sequential linkages. This reservoir is typically of higher dimension than the input, so there is a dynamical expansion of the input, thus adding model flexibility. The reservoir states are then mapped to the desired output, and importantly, *only the weights at this mapping phase are estimated*. Figure 1 shows a schematic of a typical ESN.

ESN models have not been used extensively for spatio-temporal applications, nor have they been considered from a statistical perspective for such processes. Statistical forecast methods should present reasonable uncertainty metrics for the forecasts they generate. Traditional ESN models, along with most extensions of the ESN model, do not typically provide formal uncertainty measures on the forecasts. However, there have been a few attempts to quantify uncertainty through the consideration of ensembles or bootstrap samples utilizing different random reservoir weights

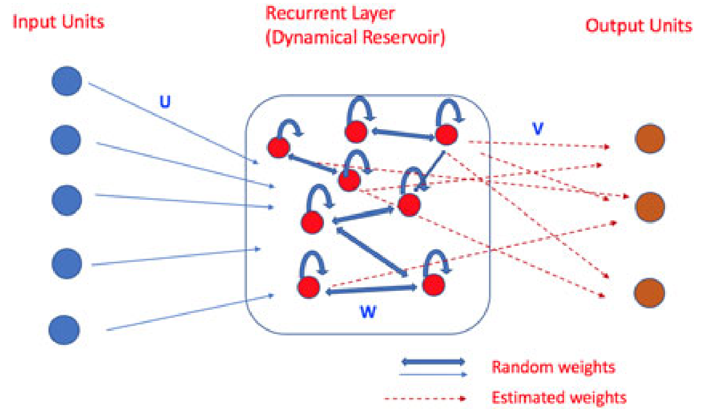

Figure 1. Schematic diagram showing the basic echo state network. The inputs are fed (via weights **U**) into a dimension expanded recurrent layer (the dynamical reservoir) with sparse interactions given by the weights **W**. The hidden units from this reservoir are then mapped into the output layer through weights **V**. Importantly, only these output weights (given by the dashed red lines) are estimated. The input and reservoir weights (given by thin and thick blue lines, respectively) are fixed, sparse and (importantly) randomly specified.

(e.g. Yao et al., 2013; Sheng et al., 2013). Despite the introduction of these uncertainty quantification methods, many of the recent developments in the ESN literature have continued to use only point estimate forecasts. In addition, other heuristic non-linear forecasting methods (e.g. analogue forecasting—see McDermott & Wikle, 2016, for a recent example) often utilize embeddings, in which lagged values of the inputs are used in the forecast, as motivated by the theorem of Takens (1981). This is far from standard practice in the ESN literature. Furthermore, ESN methods do not typically consider quadratic interactions in the mapping between the reservoir and the output.

We show here that a simple ensemble ESN, with embedded inputs and quadratic reservoir-to-output interactions, can provide very effective forecasts (with uncertainty measures) for notoriously difficult non-linear spatio-temporal dynamical systems. In particular, we consider a simulation experiment with the classic Lorenz (1996) 40-variable non-linear system and a real-world example for long-lead forecasting of tropical Pacific sea surface temperature (SST), an important problem due to the importance of the impacts of the El Niño and La Niña phenomena on weather conditions across the globe.

2 Spatio-temporal echo state network

A very basic RNN (e.g. Lukoševičius & Jaeger, 2009) can be specified as follows for time $t = 1, \ldots, T$:

$$\begin{aligned} \text{response:} \quad & \mathbf{Y}_t = g_o(\mathbf{o}_t) \\ \text{output:} \quad & \mathbf{o}_t = \mathbf{V}\mathbf{h}_t \end{aligned} \tag{1}$$

$$\text{hidden state: } \mathbf{h}_t = (1-\alpha)\, \mathbf{h}_{t-1} + \alpha\, \tilde{\mathbf{h}}_t, \tag{2}$$

$$\tilde{\mathbf{h}}_t = g_h(\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{x}_t), \tag{3}$$

where $\mathbf{Y}_t$ is an n_y vector of responses at time t; $\mathbf{x}_t$ is an n_x-dimensional input vector (typically assumed to include a 1 in the first position for an "intercept"); $\mathbf{o}_t$ is an n_y vector of outputs that are associated with a linear transformation of the n_h-dimensional hidden unit vector, $\mathbf{h}_t$ (with $\tilde{\mathbf{h}}_t$ as its update); $\mathbf{W}$ and $\mathbf{U}$ are $n_h \times n_h$ and $n_h \times n_x$ hidden-layer weight matrices, respectively; $\mathbf{V}$ is the $n_y \times n_h$ output weight matrix; and $g_o(\cdot)$ and $g_h(\cdot)$ are specified activation functions. The α parameter in (2) takes a value $(0, 1]$ and is known as the "leaking rate." The ESN version of this simple RNN then considers the hidden-layer weight matrices $\mathbf{W}$ and $\mathbf{U}$ (the reservoir weights) to be fixed; they are just drawn once from a distribution centred around zero, with added sparsity (see Section 2.1 for details). Only the output matrix $\mathbf{V}$ is estimated! This presents a huge computational cost savings because there are relatively few output weight parameters, and they can be estimated through standard regularization-based statistical estimation approaches (e.g. if $g_o(\cdot)$ is the identity function, then a simple ridge regression estimation of $\mathbf{V}$ is typically used). The ESN model has gained much popularity in large part due to this computational advantage.

It is important to understand the role of the reservoir given by Equation (3) in the basic ESN framework. As described by Lukoševičius (2012), the hidden units in the reservoir act as a non-linear expansion of the input vector, $\mathbf{x}_t$, and, perhaps more importantly, as a way to establish "memory" or account for the sequential nature of the dependence in the input vectors and, ultimately, the response. The idea of a non-linear expansion in a high dimension helps to magnify potentially salient dynamic features of the input, and thus, the output weights ($\mathbf{V}$ in Equation (1)) provide a way to select those expanded states that are important for the response.

In Section 2.1, we describe this model in more detail and include our modifications for spatio-temporal prediction.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.160

2.1 Quadratic echo state network

As described in Section 1, when non-linear spatio-temporal processes are predicted, it is often quite important to include quadratic interactions between hidden processes and the response, as well as embeddings (lagged values) of the input. Very simple modifications to the basic ESN from Section 2 allow for these important model components. In the following, we show one such modified model, a basic quadratic ESN (QESN) for continuous output (i.e. where $g_o(\cdot)$ is the identity function). We note that it is straightforward to include non-linear activation functions at this stage (e.g. softmax) depending on the response support and the goal of the analysis (i.e. classification versus regression).

For $t = 1, \ldots, T$, the QESN model is given by

$$\text{response:} \quad \mathbf{Y}_t = \mathbf{V}_1 \mathbf{h}_t + \mathbf{V}_2 \mathbf{h}_t^2 + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{R}), \tag{4}$$

$$\text{hidden state:} \quad \mathbf{h}_t = (1-\alpha)\, \mathbf{h}_{t-1} + \alpha\, \tilde{\mathbf{h}}_t, \tag{5}$$

$$\tilde{\mathbf{h}}_t = g_h \left(\frac{\nu}{|\lambda_w|} \mathbf{W} \mathbf{h}_{t-1} + \mathbf{U} \tilde{\mathbf{x}}_t \right), \tag{6}$$

$$\text{parameters:} \quad \mathbf{W} = [w_{i,\ell}]_{i,\ell} : w_{i,\ell} = \gamma_{i,\ell}^{w}\ \mathbf{Unif}(-a_w, a_w) + (1-\gamma_{i,\ell}^{w})\ \delta_0, \tag{7}$$

$$\begin{aligned} &\mathbf{U} = [u_{i,j}]_{i,j} : u_{i,j} = \gamma_{i,j}^{u}\ \mathbf{Unif}(-a_u, a_u) + (1-\gamma_{i,j}^{u})\ \delta_0, \\ &\gamma_{i,\ell}^{w} \sim Bern(\pi_w), \\ &\gamma_{i,j}^{u} \sim Bern(\pi_u), \end{aligned} \tag{8}$$

where $\mathbf{Y}_t$ is the n_y-dimensional response vector at time t, $\mathbf{h}_t$ is the n_h-dimensional hidden state vector and

$$\tilde{\mathbf{x}}_t = \left[\mathbf{x}_t', \mathbf{x}_{t-\tau}', \mathbf{x}_{t-2\tau}', \ldots, \mathbf{x}_{t-m\tau}' \right]' \tag{9}$$

is the $n_{\tilde{x}} = (m+1)n_x$-dimensional embedding input vector, containing embeddings for time periods $t-\tau$ through $t-m\tau$, where τ is often the forecast lead time (although generally τ can be any integer). Furthermore, δ_0 is a Dirac function at zero, and λ_w corresponds to the largest eigenvalue of $\mathbf{W}$ (i.e. the "spectral radius" of $\mathbf{W}$). The only parameters that are estimated in this model are $\mathbf{V}_1$ and $\mathbf{V}_2$, and $\mathbf{R}$ from Equation (4), for which we use a ridge penalty hyperparameter, r_v. Importantly, note that the $\mathbf{W}$ and $\mathbf{U}$ matrices are simulated from a mixture distribution of small values (uniformly sampled in a range between $(-a_w, a_w)$ and $(-a_u, a_u)$) with, on average, $(1-\pi_w)$ and $(1-\pi_u)$ elements set equal to zero and then assumed to be fixed. Finally, the hyperparameters $\{\nu, n_h, r_v, \pi_w, \pi_u, a_w, a_u\}$ are specified. Note that in the case of long-lead forecasting (say, at lead time τ_F), $\mathbf{x}_t$ may consist of inputs at time $t - \tau_F$ (e.g. see the application in Section 4).

2.2 Ensemble quadratic echo state network

Most traditional ESN applications do not include a mechanism to quantify the uncertainty of the model predictions. This is perhaps somewhat surprising given that the reservoir weight parameters are not estimated but are chosen at random. We would expect that the model is likely to behave differently with a different set of **W** and **U** weights. This is especially true when the number of hidden units is fairly small. Although traditional ESN models have a large number of hidden units, it can be desirable to have many ensemble members with a smaller number of units. This tends to prevent overfitting, allows the ensemble members to behave as a committee of relatively weak learners and gives a more realistic sense of the prediction uncertainty for out-of-sample forecasts. Thus, we could generate an ensemble or bootstrap sample of forecasts (e.g. Yao et al., 2013; Sheng et al., 2013). This ensemble approach can easily be implemented with our model, as given in Algorithm 1, to make out-of-sample forecasts for n_f periods (note that the R code is included in the Supporting Information, Codes S1–S4). For the sake of clarity, the set $\{\mathbf{Y}_t, \tilde{\mathbf{x}}_t \colon t = 1, \ldots, T\}$ in Algorithm 1 represents in-sample or observed data, while $\widehat{\mathbf{Y}}$ represents out-of-sample predictions.

```
Data: $\{\mathbf{Y}_t, \tilde{\mathbf{x}}_t \colon t = 1, \ldots, T\}$
Result: Ensemble of predictions: $\{\widehat{\mathbf{Y}}_t^k : t = T+1, \ldots, T+n_f; k = 1, \ldots, K\}$
Initialize: Select tuning parameters $\{n_h, \nu, r_v, \pi_w, \pi_u, a_w, a_u, \alpha\}$ (e.g. by validation or cross-validation) ;
for $k = 1, \ldots, K$ do
    Simulate $\mathbf{W}^k$, $\mathbf{U}^k$ using (7) and (8)
    Calculate $\{\mathbf{h}_t^k : t = 1 : T\}$ using (5) and (6)
    Use ridge regression to estimate $\mathbf{V}_1^k$, $\mathbf{V}_2^k$
    Calculate out-of-sample forecasts $\{\widehat{\mathbf{Y}}_t^k : t = T+1, \ldots, T+n_f\}$ (requires calculating $\{\widehat{\mathbf{h}}_t^k : t = T+1, \ldots, T+n_f\}$
    from the reservoir)
end
Use ensemble of forecasts $\{\widehat{\mathbf{Y}}_t^k : t = T+1, \ldots, T+n_f; k = 1, \ldots, K\}$ to calculate moments, prediction intervals (P.I.s),
etc.
```

Algorithm 1: Simple ensemble quadratic echo state network algorithm

2.3 Model parameterizations and hyperparameters

The model presented in Section 2.1 depends on several hyperparameters, some of which are typically more important than others in ESN applications. An excellent and detailed summary of the practical issues associated with traditional ESN implementation is given by Lukoševičius (2012). We discuss some of these notions here, along with our experience in the context of spatio-temporal forecasting with this specific model.

The size of the reservoir, n_h, is traditionally one of the more important hyperparameters. In most implementations, one seeks a reservoir with a large number of hidden states (n_h large) and assumes that regularization will mitigate the potential to overfit. A rule of thumb in traditional ESN settings is to make n_h as large as possible, so that $T > n_x + n_h$. In the spatio-temporal context, we typically do not have extremely large values of T, but we make up for that with the embedding input, $\tilde{\mathbf{x}}_t$. Thus, we have found in our settings that smaller values of n_h are often sufficient. Also, as described in Section 2.2, we generally prefer n_h to be relatively small so that our ensemble acts more as a committee of weak learners, which helps prevent against overfitting. In practice, we typically select n_h through a validation or cross-validation procedure (Sections 3 and 4).

In traditional ESN applications, the leaking rate parameter, α, is often quite important. We have found that in our QESN model this is not typically the case. That is, in almost every application we have considered, validation and cross-validation have suggested that $\alpha = 1$ is the best setting for this parameter, in which case, $\tilde{\mathbf{h}}_t = \mathbf{h}_t$. However, this need not be the case because small leaking rates can be helpful for slowly varying systems (Lukoševičius, 2012). Leaking rates are therefore application dependent, and so we recommend evaluating whether $\alpha \neq 1$ improves out-of-sample prediction.

The scaling of the hidden state reservoir weighting matrix, $\mathbf{W}$, in (6) is quite important. In general, the spectral radius (largest eigenvalue) of $\mathbf{W}$ must be less than 1 to ensure what is known in the literature as the "echo state property." This is a property that allows, with large enough time increments, for the hidden states to lose their dependence on the initial input conditions. Practically, when the spectral radius is not less than 1, the hidden state can experience complex non-linear dynamics (e.g. multiple fixed points, periodicities and chaotic behaviour), which destroys the echo

(wileyonlinelibrary.com) DOI: 10.1002/sta4.160

state property (e.g. see the discussion in Lukoševičius, 2012). A rule of thumb is that a smaller spectral radius should be used if the responses are more dependent on the input at recent times and a larger value (but still less than 1) should be used if the responses depend more on the past. For our purposes, this means that the parameter ν should be quite important as it controls the spectral radius. That is, dividing **W** by $|\lambda_w|$ gives a spectral radius of 1, and so multiplying by $\nu < 1$ gives us the flexibility to control the overall spectral radius of the hidden state weighting parameters. We typically choose the specific value of ν by validation or cross-validation.

There is some debate in the literature about the importance of sparsity on the reservoir weight matrices, **W** and **U**. We have found in our applications that it is important that these be quite sparse (e.g. 80–95% zeros). Both π_w and π_u can also be selected through validation or cross-validation, although the model is typically not very sensitive to the specific (small) value of either parameter. For the applications presented here, the model was moderately sensitive to the choice of ridge parameter, r_v, used for estimating $\mathbf{V}_1$ and $\mathbf{V}_2$. We typically use validation or cross-validation to select this parameter. We note that other forms of regularization (e.g. L_1 penalties or hybrid L_1, L_2 penalties) could be used here as well.

Lastly, it is typically the case that inputs are normalized in ESN applications, although in principle, this can be accommodated through the parameter scaling (i.e. the a_w and a_u hyperparameters). Note also that it is known that the reservoir acts to compress the variability of the principal components of the inputs, $\mathbf{x}_t$. Thus, it is often recommended that if one uses principal components as inputs, then it may be best to not include those that are associated with the smallest variability, as their importance and influence will be exaggerated in the reservoir (e.g. Lukoševičius, 2012). In spatio-temporal applications, we typically use some form of dimension reduction on the input $\mathbf{x}_t$ vectors (e.g. empirical orthogonal functions (EOFs)—which are just principal components for spatio-temporal data, e.g. Cressie & Wikle, 2011; but other dimension reduction approaches can be used). We also typically work with response vectors that have been projected onto their leading EOFs given the large number of spatial locations that are often present in spatio-temporal prediction problems (Appendix S1). Therefore, we have found that when using EOF coefficients as inputs, it is best to normalize the inputs by the overall mean and standard deviation (across all EOFs).

In summary, we have found in our spatio-temporal applications that the results tend to be more sensitive to the choice of $\{n_h, \nu, r_v\}$ than $\{\pi_w, \pi_u, a_w, a_u\}$ but recommend evaluating the sensitivity of the forecasts to these parameters in new applications.

3 Simulated data: the 40-variable Lorenz system

In order to evaluate the model and the importance of the modifications considered in Section 2.1 (i.e. embedding inputs, quadratic outputs and ensemble uncertainty quantification), we consider simulated data from the classic 40-variable non-linear model of Lorenz (1996), referred to as the "Lorenz-96" model. Although the classic three-variable Lorenz model is often cited in the ESN literature, to our knowledge, this more spatially relevant 40-variable model has not been evaluated in the ESN context. In particular, the system is governed by the set of 40 ordinary differential equations given by

$$\frac{\mathrm{d}z_{i,t}}{\mathrm{d}t} = (z_{i+1,t} - z_{i-2,t})z_{i-1,t} - z_{i,t} + F_t, \quad i = 1, \ldots, 40, \tag{10}$$

where F_t is a forcing variable and the variables $z_{i,t}$ correspond to state variables at 40 equally spaced locations on a circular (e.g. latitude circle) one-dimensional spatial domain with periodic boundary conditions (e.g. $z_{41,t} = z_{1,t}$). The non-linearity in (10) comes from the quadratic interactions across locations, and this was originally designed to mimic non-linear advection processes in geophysical fluids.

The data were simulated by using a simple Euler solver with a time step of $\Delta = 0.10$ and forcing value of $F_t = 5$. Moreover, the lead time is set to six periods in order to increase the non-linearity and create a more realistic forecasting scenario. Gaussian white noise error was added to each realization, so that $Y_{i,t} = z_{i,t} + \eta_{i,t}$, where $\eta_{i,t} \sim N(0, \sigma^2_\eta)$ for $i = 1, \ldots, 40$, with $\sigma_\eta = 0.5$. We considered 750 time periods (post burn-in), with the last 99 time periods held out for out-of-sample prediction.

3.1 Lorenz-96 model validation

A validation approach was used to select the hyperparameters for the QESN model. Specifically, we consider the following grid of hyperparameters: $n_h = \{30 + 15 \times q_h : q_h = 0, \ldots, 5\}$, $\nu = \{0.05 \times q_\nu : q_\nu = 1, \ldots, 20\}$ and $r_v = \{0.001, 0.005, 0.01\}$. Furthermore, we had little *a priori* information about how the input should be embedded; hence, we also included the number of embeddings (m) as a hyperparameter in the grid search. Specifically, with a value of $\tau = 1$, which produced more accurate forecasts than using the lead time of 6, the parameter m was evaluated over a grid of $\{2 \times q_m : q_m = 1, \ldots, 5\}$. We should note that when embeddings have been used in the ESN literature, the embedding dimension is often set heuristically without considering different possible values. Finally, the model was not very sensitive to the hyperparameters in the set $\{\pi_w, \pi_u, a_w, a_u\}$, and so all four were set to 0.1. All models were evaluated via mean square error (MSE) calculated over all validation periods and locations, using an ensemble size of 500 (i.e. K =500 in Algorithm 1). The parameters associated with the lowest validation sample MSE were $n_h = 60$, $\nu = 0.55$, $r_v = 0.001$ and $m = 4$.

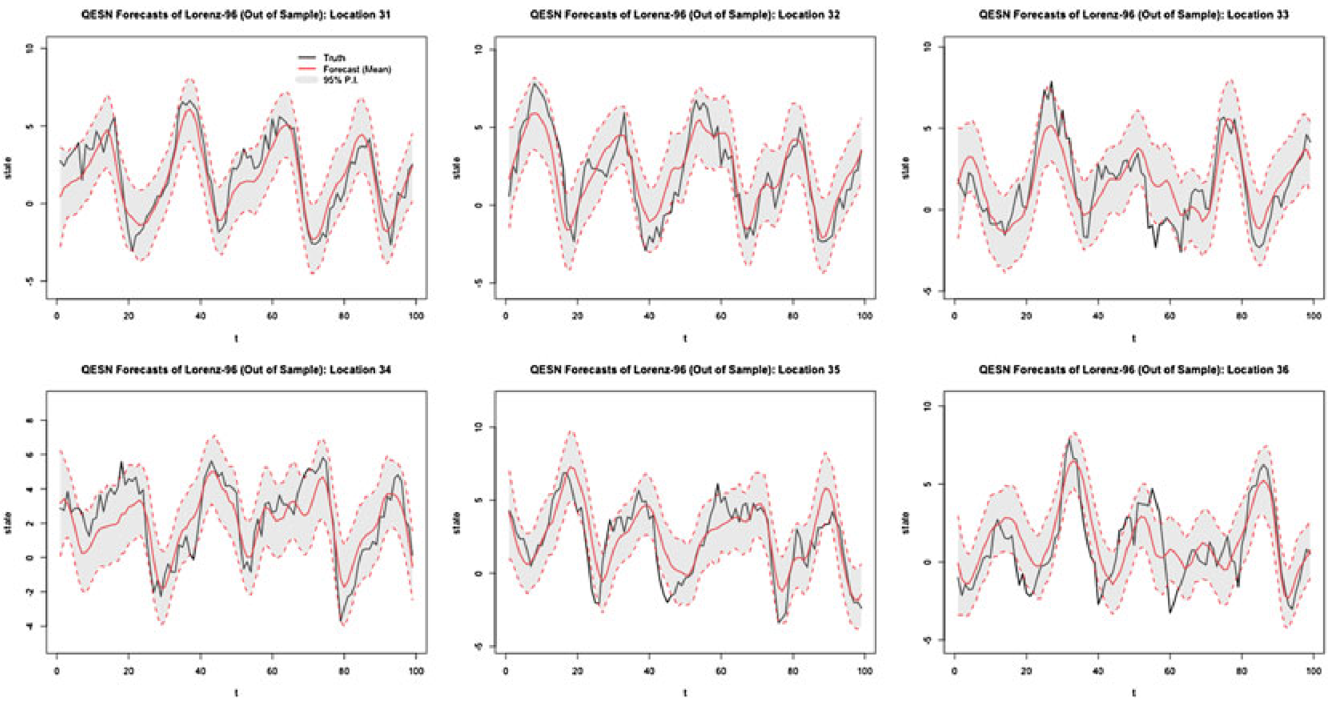

Figure 2. Out-of-sample forecasts for six locations from the simulated Lorenz-96 system over 99 periods with a lead time of six periods. The black line represents the truth, while the solid red line denotes the forecasted mean from the ensemble quadratic echo state network (QESN) model using hyperparameters found through validation. The shaded grey area signifies 95% prediction intervals.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.160

3.2 *Lorenz-96 model out-of-sample forecasts*

Using the validation parameters given in Section 3.1, the model was trained using the first 651 observations (which are considered in-sample observations here), with the last 99 observations held out for out-of-sample forecasts. These forecasts and associated 95% P.I.s are given in Figure 2 for six of the 40 locations. Considering the difficulty of forecasting a non-linear system six time periods forward, Figure 2 shows that the forecasts for the QESN model generally correctly forecast the time evolution of the system. Perhaps of more consequence, across all locations, the model has good PI coverage probabilities, with 95.4% of the true values falling within the 95% P.I.s. This example demonstrates the potential for the ensemble QESN model to successfully forecast spatio-temporal systems at non-linear timescales, and to capture reasonable forecast uncertainty.

4 Application: long-lead forecasting of Pacific sea surface temperature

The anomalous warming (El Niño) and cooling (La Niña) of the tropical Pacific ocean that occurs quasi-periodically on timescales of 3–5 years accounts for one of the largest sources of variability in weather systems across the globe. These phenomena are sometimes collectively known as the El Niño Southern Oscillation (ENSO). The effects of ENSO variability can be quite serious in terms of heat waves, drought, flooding and increased potential for other types of severe weather. For this reason, there has long been interest in forecasting the state of the tropical Pacific ocean surface temperature many months into the future in order to facilitate resource planning. ENSO is a very complicated multivariate process that operates on many spatio-temporal scales of variability, and it is known to exhibit non-linear evolution at certain timescales. In addition, many forecast methods that have performed reasonably well in past cycles of ENSO did not work very effectively in the year leading up to the last major ENSO cycle in 2015–6 (L'Heureux et al., 2016; Hu & Fedorov, 2017). Thus, we consider the 6-month long-lead forecasts of Pacific SST over this period to illustrate our ensemble QESN model.

Long-lead prediction is one of the few scenarios in modern climatological and weather forecasting where statistical methods can do as well or better than deterministic methods (i.e. numerical solutions to partial differential equations that govern the ocean and atmosphere), as summarized by Barnston et al. (1999) and Jan van Oldenborgh et al. (2005). Although linear models have been shown to produce skillful forecasts of ENSO (e.g. Penland & Magorian, 1993; Knaff & Landsea, 1997), it has been established that models that can accommodate non-linear interactions often work better, particularly in forecasting the evolution of the El Niño phase of the ENSO cycle. In particular, successful non-linear models that have been used in the past include the non-linear analogue approach of Drosdowsky (1994), classical neural network models of Tangang et al. (1998) and Tang et al. (2000), switching Markov model of Berliner et al. (2000), empirical non-linear inverse models of Timmermann et al. (2001), empirical model reduction methods of Kondrashov et al. (2005) and Kravtsov et al. (2005, 2009), and the general quadratic non-linear (GQN) models of Wikle & Hooten (2010), Wikle & Holan (2011) and Gladish & Wikle (2014). However, most of these methods were developed outside of statistics and are heuristic with little or no formal uncertainty quantification. Exceptions include the switching Markov and GQN models, which were developed in the statistics literature and include a formal (hierarchical Bayesian) uncertainty quantification. Indeed, one of the strengths of the formal GQN approach is that even when its forecast means are not significantly better than these other approaches, its quantification of uncertainty in terms of P.I.s tends to be much more realistic for the unusually strong ENSO events. Yet these formal hierarchical non-linear statistical SST forecast models can be quite expensive computationally, thus suggesting the possible utility of more computationally efficient models such as the ensemble ESN model presented here.

4.1 Data

The long-lead SST forecasting application consists the publicly available Extended Reconstruction Sea Surface Temperature data provided by the National Ocean and Atmospheric Administration (http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/). In particular, we use monthly data of SST anomalies over a period from 1970 to 2016. The spatial domain of the data is over 29°S–29°N latitude and 124°E–70°W longitude, with a resolution of 2° × 2° (i.e. 2229 oceanic spatial locations). Anomalies were calculated by subtracting the monthly climatological means calculated over the period 1981–2010.

The ENSO trajectory can be seen more concisely by using the common univariate summary measure for ENSO, the Niño 3.4 index, which represents the average of the SST anomalies over the so-called Niño 3.4 region (5°S–5°N, 120°–70°W). The Niño 3.4 index time series for the period from December 2013 to December 2016 is plotted in Figure 3. Because of the overall importance of the Niño 3.4 region, it is common to use this index to evaluate long-lead SST forecasts (Barnston et al., 2012).

4.2 Ensemble quadratic echo state network implementation

To select the hyperparameters for the QESN model, we conducted a validation study by using periods from the ENSO event that occurred from 1997 to 1999 as a holdout sample. In particular, all of the data from January 1970 through December 1996 were used to train the QESN model, while a sequence of months from May 1997 through August 1999 made up the holdout sample. Both the output and the embedded input consisted of principal component time series associated with the first 10 EOFs (note that spatial fields correspond to the product of these time series and the EOF basis functions; these spatial fields are used for plotting and for calculating the Niño 3.4 index average of the forecast). A more detailed description of the dimension reduction performed here can be found in the Appendix

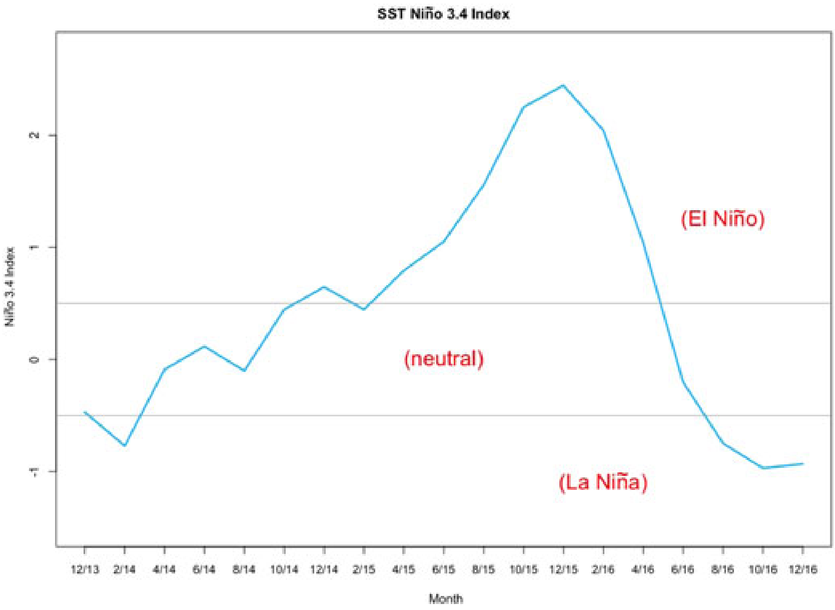

Figure 3. Time series plot of the average sea surface temperature (SST) for all grid locations in the Niño 3.4 region (5°S–5°N, 120°–70°W) from December 2013 through December 2016. For purposes of discussion, values above 0.5 represent El Niño periods, and periods falling below -0.5 are considered La Niña periods.

S1. In this case the input vectors are just lagged values of the response data (i.e. $\mathbf{x}_t = \mathbf{Y}_{t-\tau_F}$). The validation study uses the same grid of values for the n_h, ν and r_v parameters as the Lorenz example. In addition, for the embedded input vectors, τ is set to the lead time (i.e. 6 months) and m was varied over the set $\{1, 2, 3, 4, 5\}$. Once again, we found that the model forecasts were not very sensitive to the particular value for the hyperparameters $\{\pi_w, \pi_u, a_w, a_u\}$, and so we used the same values as in the Lorenz example. For the validation study, the evaluation of a particular set of hyperparameters was based on the MSE of the Niño 3.4 region, calculated by using Algorithm 1 with 500 ensemble members. The lowest MSE for the validation study was associated with the following hyperparameters: $n_h = 120, \nu = .35, r_v = .01$ and $m = 4$.

Evaluation of the ensemble QESN model was conducted by making out-of-sample forecasts for the 2015–6 ENSO cycle. That is, after the model was trained from January 1970 to August 2014 (in-sample observations) using the hyperparameters found in the validation study, 6-month lead time out-of-sample forecasts were made for every 2 months of the 2015–6 ENSO cycle from February 2015 to December 2016 (as shown in Figure 4). After validation, the R program version of this model (Supporting Information) took less than 15 seconds to train and generate predictions using a 2.3-GHz laptop.

4.3 Ensemble quadratic echo state network results

As previously discussed, forecasting the Niño 3.4 region is both vital from a planning and management perspective and challenging forecasting problem. Forecast means of the Niño 3.4 index during the 2015–6 ENSO cycle, along with their 95% P.I.s, are shown in Figure 4. Overall, the QESN model produces uncertainty bounds that cover the observed Niño 3.4 index for the entire 2015–6 ENSO cycle. Similar to almost every other statistical and deterministic model, the QESN model underestimated the peak of the cycle during December 2015, yet the uncertainty bounds appear very realistic (and superior to many of those used for operational forecasts).

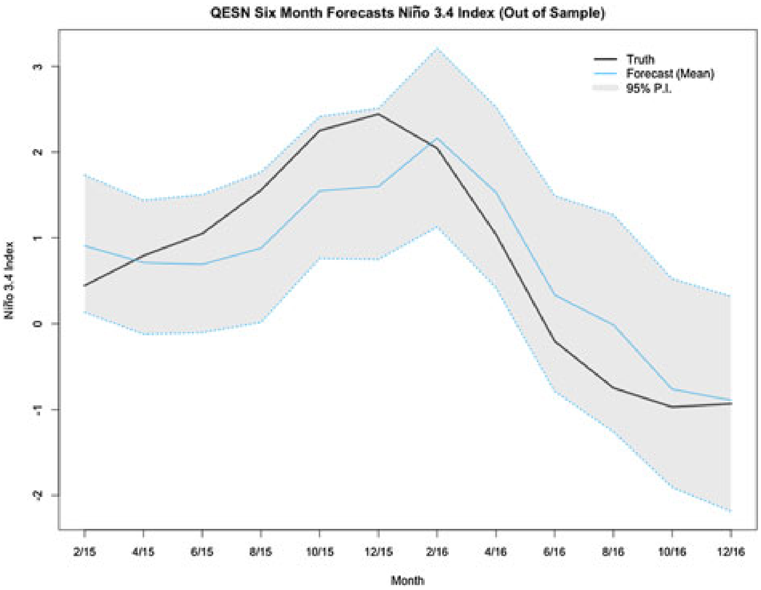

Figure 4. Out-of-sample 6-month lead time forecasts and prediction intervals (P.I.s) for every 2 months from February 2015 to December 2016 for the average sea surface temperature anomalies in the Niño 3.4 region. The solid black line represents the truth calculated from the data, and the corresponding forecast mean from the ensemble quadratic echo state network (QESN) is denoted by the solid blue line. The shaded grey area represents 95% P.I.s from the ensemble QESN model.

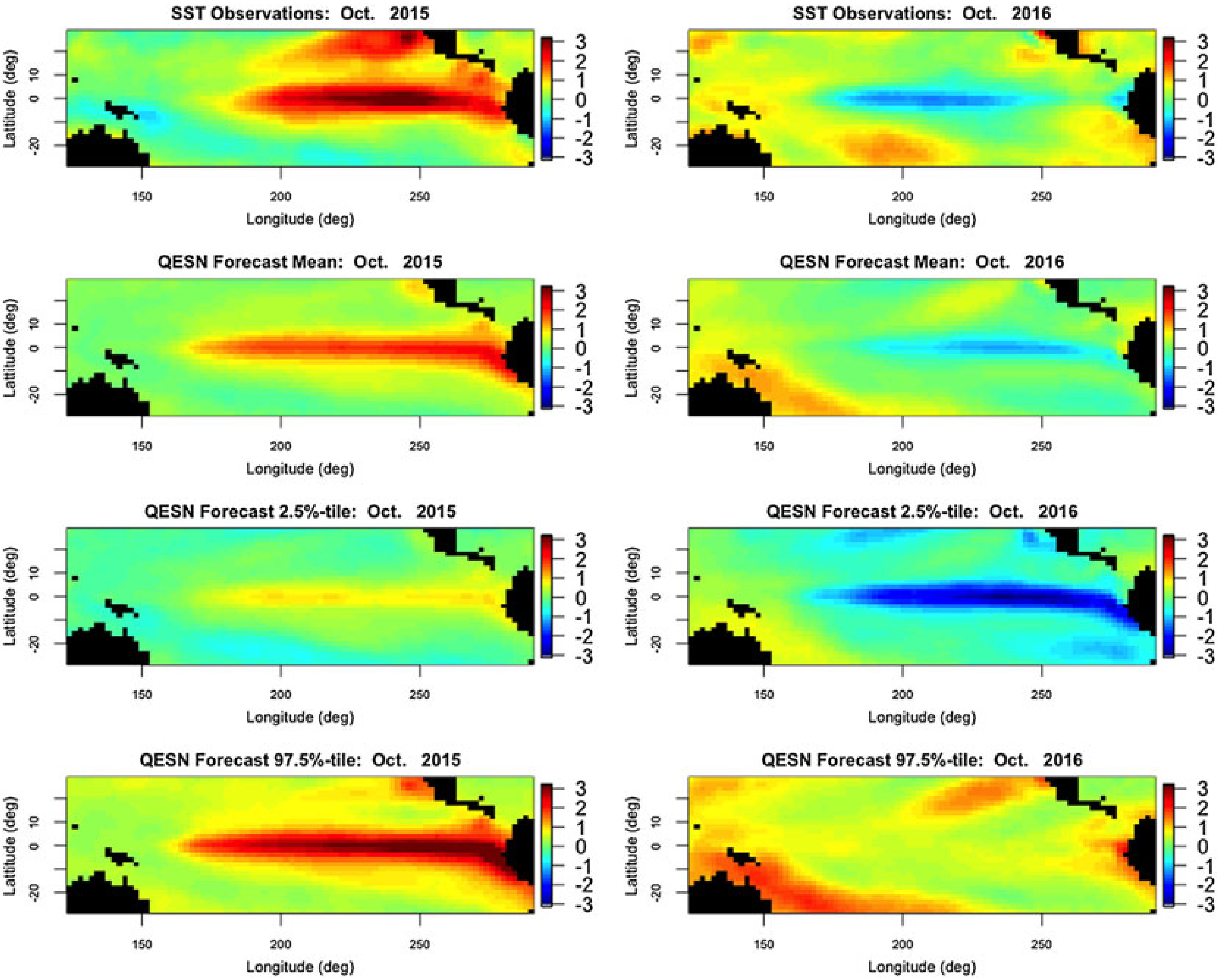
SST Observations: Oct. 2015
SST Observations: Oct. 2016
QESN Forecast Mean: Oct. 2015
QESN Forecast Mean: Oct. 2016
QESN Forecast 2.5%-tile: Oct. 2015
QESN Forecast 2.5%-tile: Oct. 2016
QESN Forecast 97.5%-tile: Oct. 2015
QESN Forecast 97.5%-tile: Oct. 2016
Latitude (deg)
-20
0
10
Longitude (deg)
150
200
250
3
2
1
0
-1
-2
-3

Figure 5. Forecast summary maps for all spatial locations for 6-month forecasts valid on October 2015 and October 2016. The top row denotes the true sea surface temperature (SST) anomalies for the respective years. Forecasted means from the ensemble quadratic echo state network (QESN) model for each year are given in the second to top row. The bottom two rows capture the lower and upper quantiles for a 95% prediction interval calculated in each grid box for each year.

Although the Niño 3.4 index can provide valuable information and is a useful summary of forecast accuracy, there is also an interest in examining model forecast performance over the entire spatial domain. Out-of-sample 6-month forecasts for all 2229 grid locations are shown in Figure 5 for October 2015 and 2016. For both time periods, the model forecast largely picks up the intensity of the phenomenon, while also capturing much of the true values within the 95% P.I.s. This later point is very important, as many of the heuristic forecast methods in use do not provide such a formal uncertainty quantification.

To examine the usefulness of the embedding and quadratic extensions of the ESN model, three additional models were run with the hyperparameters selected in the validation study. In particular, we ran the model without any embeddings

(wileyonlinelibrary.com) DOI: 10.1002/sta4.160

Table I. Results for the ensemble quadratic echo state network (QESN) and models M1–M3 in terms of mean squared error (MSE) over all 2229 locations (overall MSE), the MSE for the Niño 3.4 region (Niño 3.4 MSE) and the continuous ranked probability score (CRPS) over all locations.

Model	Overall MSE	Niño 3.4 MSE	CRPS
QESN	**0.288**	**0.261**	**3.722**
M1	0.343	0.545	4.570
M2	0.328	0.586	4.595
M3	0.345	0.741	4.845

M1 denotes a model without any embeddings, M2 is a model without any quadratic output terms and M3 is a model without embeddings or quadratic output terms. For each metric, the model that produced the lowest value is bolded.

(labelled M1), without any quadratic terms (labelled M2) and without both embeddings and quadratic terms (labelled M3). Along with evaluating these models with the overall MSE (for all ocean grid locations) and the Niño 3.4 index MSE, we also considered the continuous ranked probability score (CRPS) over all locations. The CRPS summary is useful for evaluating the prediction accuracy of a forecast, while also considering the distribution of the forecast, and thus the quantification of uncertainty (Gneiting & Katzfuss, 2014). For all three summary metrics displayed in Table I, the ensemble QESN model clearly outperforms the other models, suggesting that embedding inputs and quadratic output components are helpful in producing better forecast distributions.

Finally, it is useful to compare the ensemble QESN model to the GQN model of Wikle & Hooten (2010), which has been shown to be a useful non-linear dynamical spatio-temporal model that incorporates formal uncertainty quantification. In particular, we compared the 6-month lead time forecast distributions from the ensemble QESN model and the GQN model for forecasts valid in October 2015 and October 2016 (near the most intense portions of the El Niño and La Niña in this ENSO cycle). Note that while it took less than 15 seconds to generate forecasts with the ensemble QESN model, it took over 1 hour to generate an equivalent number of posterior samples from the GQN model!

The distributional comparison for the Niño 3.4 region averages is shown in Figure 6. The left panel shows the forecast distribution for October 2015 (El Niño), and the right panel shows the forecast distribution for October 2016 (La Niña). In the case of the El Niño period forecast, both model forecast distributions contain the true value, but the GQN model forecast central tendency is closer to the truth than the ensemble QESN, suggesting it was a better forecast distribution. However, the La Niña forecasts tell a much different story. The GQN forecast distribution shows a substantial warm bias (and does not include the truth), whereas the ensemble QESN forecast distribution is quite good, with its central tendency close to the truth and a reasonable uncertainty range. Previously published examples of GQN long-lead forecasts for earlier ENSO events have shown that it tends to perform better for the El Niño phase than it does for the La Niña phase, most likely owing to the fact that the El Niño phase evolution is more non-linear (although, in the past, it has performed better than it did here for the La Niña case; Wikle & Hooten, 2010; Wikle & Holan, 2011; Gladish & Wikle, 2014). It is very encouraging that the ensemble ESN model shows high-quality forecast distributions for both periods for the 2015–6 ENSO, especially given how efficiently it can be computed.

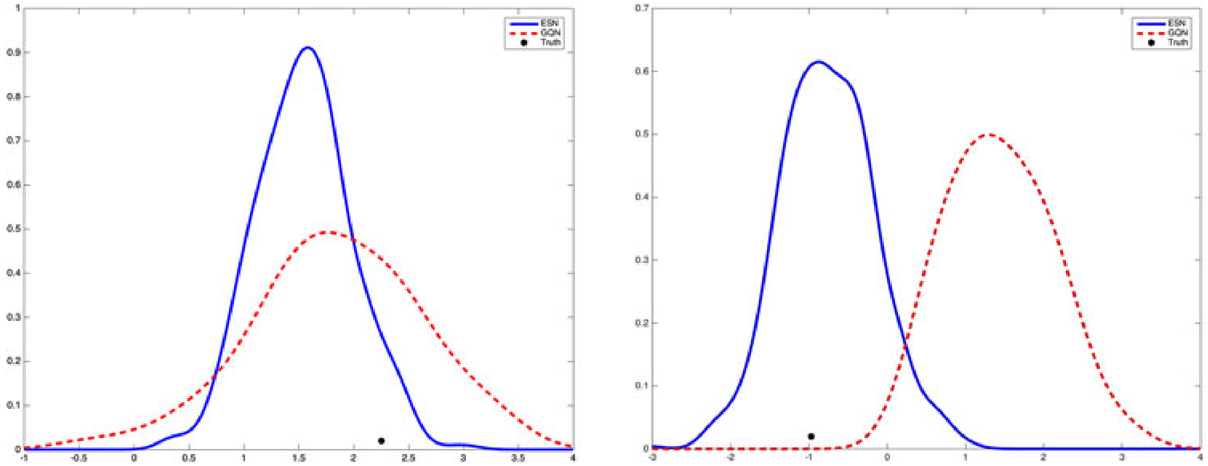

Figure 6. Comparison of ensemble quadratic echo state network (QESN) 6-month lead forecasts of sea surface temperature averaged over the Niño 3.4 index region to the general quadratic non-linearity (GQN; Wikle & Hooten, 2010) model forecasts. Left panel: Forecast for October 2015. Right panel: Forecast for October 2016. The solid blue line represents the kernel smooth of samples from the echo state network (ESN) model, and the dashed red line represents the kernel smooth of samples from the GQN model. The dark circle corresponds to the value of the truth index for the verification time.

5 Discussion and conclusion

Many spatio-temporal processes are complex dynamical systems with non-linear interactions across process components and spatio-temporal scales of variability. Depending on the scales of interest, it is often crucial to include these non-linear interactions in forecasting situations in order to obtain realistic predictions and uncertainty measures. Only relatively recently have parametric statistical models been developed for non-linear spatio-temporal processes. These models tend to have quadratic non-linear structure and “deep” (multi-level) parameter levels and have been shown to be flexible and useful in a variety of applications with large datasets. However, in these situations, the models can be quite expensive to implement in terms of computational resources. As an alternative, we consider a spatio-temporal extension of the ESN models that have been used in the machine learning context for non-linear time series applications, in order to exploit their efficient reservoir computing framework.

In particular, for our spatio-temporal extension, we consider three simple modifications of the traditional ESN, which we call the ensemble quadratic ESN, or ensemble QESN. This includes the addition of embeddings in the input vector, a quadratic term in mapping the hidden state to the response vector and an ensemble implementation that accounts for the uncertainty associated with the fixed, yet randomly generated, reservoir weight matrices. We show that these components allow for reasonable forecasts of non-linear spatio-temporal processes at a fraction of the computational time associated with more formal statistical methods that accommodate realistic uncertainty quantification. This is demonstrated on both a classic simulated non-linear system from Lorenz (1996) and the challenging problem of long-lead forecasting of Pacific SST during the most recent intense ENSO cycle.

Although the results presented here are encouraging in the quality of the forecasts relative to the computation time, there remain issues related to the choice of model hyperparameters. That is, we rely on a validation sample, drawn from a contiguous time period in the past (e.g. the 1997–8 ENSO events) to obtain these hyperparameters. If these methods are to be used for spatio-temporal forecasting in practice, then it would be useful to consider more formally the uncertainty associated with these choices as well. In addition, it is of interest to consider this simple methodology

(wileyonlinelibrary.com) DOI: 10.1002/sta4.160

on a wider range of spatio-temporal prediction problems and to consider its utility in spatio-temporal classification problems.

In conclusion, with slight modification, relatively simple "off-the-shelf" machine learning methods for complex sequential data can be effective in spatio-temporal prediction. This suggests that other black-box learning tools for dependent data may also be useful to help motivate more formal statistical models for spatio-temporal data.

Acknowledgements

This work was partially supported by the US National Science Foundation (NSF) and the US Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) programme.

References

Barnston, AG, He, Y & Glantz, MH (1999), 'Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997–1998 El Niño episode and the 1998 La Niña onset', *Bulletin of the American Meteorological Society*, **80**(2), 217–243.

Barnston, AG, Tippett, MK, L'Heureux, ML, Li, S & DeWitt, DG (2012), 'Skill of real-time seasonal ENSO model predictions during 2002–2011: is our capability increasing?', *Bulletin of the American Meteorological Society*, **93**(5), 631–651.

Berliner, LM, Wikle, CK & Cressie, N (2000), 'Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling', *Journal of Climate*, **13**(22), 3953–3968.

Cressie, N & Wikle, C (2011), *Statistics for Spatio-temporal Data*, *John Wiley & Sons*, New York.

Drosdowsky, W (1994), 'Analog (nonlinear) forecasts of the Southern Oscillation index time series', *Weather and Forecasting*, **9**(1), 78–84.

Gladish, DW & Wikle, CK (2014), 'Physically motivated scale interaction parameterization in reduced rank quadratic nonlinear dynamic spatio-temporal models', *Environmetrics*, **25**(4), 230–244.

Gneiting, T & Katzfuss, M (2014), 'Probabilistic forecasting', *Annual Review of Statistics and Its Application*, **1**, 125–151.

Hu, S & Fedorov, AV (2017), 'The extreme El Niño of 2015–2016: the role of westerly and easterly wind bursts, and preconditioning by the failed 2014 event', *Climate Dynamics*, **43**, 1–19.

Jaeger, H (2007), 'Echo state network', *Scholarpedia*, **2**(9), 2330.

Jan van Oldenborgh, G, Balmaseda, MA, Ferranti, L, Stockdale, TN & Anderson, DLT (2005), 'Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years?', *Journal of Climate*, **18**(16), 3240–3249.

Knaff, JA & Landsea, CW (1997), 'An El Niño–Southern Oscillation climatology and persistence (CLIPER) forecasting scheme', *Weather and Forecasting*, **12**(3), 633–652.

Kondrashov, D, Kravtsov, S, Robertson, AW & Ghil, M (2005), 'A hierarchy of data-based ENSO models', *Journal of Climate*, **18**(21), 4425–4444.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.160

Kravtsov, S, Kondrashov, D & Ghil, M (2005), 'Multilevel regression modeling of nonlinear processes: derivation and applications to climatic variability', *Journal of Climate*, **18**(21), 4404–4424.

Kravtsov, S, Kondrashov, D & Ghil, M (2009), Empirical model reduction and the modelling hierarchy in climate dynamics and the geosciences, *Stochastic Physics and Climate Modeling*, *Cambridge University Press*, Cambridge, 35–72.

L'Heureux, ML, Takahashi, K, Watkins, AB, Barnston, AG, Becker, EJ, Di Liberto, TE, Gamble, F, Gottschalck, J, Halpert, MS, Huang, B, Mosquera-Vasquez, K & Wittenberg, AT (2016), 'Observing and predicting the 2015–16 El Niño', *Bulletin of the American Meteorological Society*, **98**, 1363–1382.

Lorenz, E (1996), Predictability: a problem partially solved, *Proceedings Seminar on Predictability*, *ECMWF*, Reading, Berkshire, UK, 1–18.

Lukoševičius, M (2012), *A practical guide to applying echo state networks*, Neural Networks: Tricks of the Trade, *Springer*, Berlin, Heidelberg, 659–686.

Lukoševičius, M & Jaeger, H (2009), 'Reservoir computing approaches to recurrent neural network training', *Computer Science Review*, **3**(3), 127–149.

Maass, W, Natschläger, T & Markram, H (2002), 'Real-time computing without stable states: a new framework for neural computation based on perturbations', *Neural Computation*, **14**(11), 2531–2560.

McDermott, PL & Wikle, CK (2016), 'A model-based approach for analog spatio-temporal dynamic forecasting', *Environmetrics*, **27**(2), 70–82.

Penland, C & Magorian, T (1993), 'Prediction of Nino 3 sea surface temperatures using linear inverse modeling', *Journal of Climate*, **6**(6), 1067–1076.

Richardson, RA (2017), 'Sparsity in nonlinear dynamic spatiotemporal models using implied advection', *Environmetrics*, **2**, e2456. https://doi.org/10.1002/env.2456.

Sheng, C, Zhao, J, Wang, W & Leung, H (2013), 'Prediction intervals for a noisy nonlinear time series based on a bootstrapping reservoir computing network ensemble', *IEEE Transactions on Neural Networks and Learning Systems*, **24**(7), 1036–1048.

Takens, F (1981), 'Detecting strange attractors in turbulence', *Lecture Notes in Mathematics*, **898**(1), 366–381.

Tang, B, Hsieh, WW, Monahan, AH & Tangang, FT (2000), 'Skill comparisons between neural networks and canonical correlation analysis in predicting the equatorial pacific sea surface temperatures', *Journal of Climate*, **13**(1), 287–293.

Tangang, FT, Tang, B, Monahan, AH & Hsieh, WW (1998), 'Forecasting ENSO events: a neural network-extended EOF approach', *Journal of Climate*, **11**(1), 29–41.

Timmermann, A, Voss, H & Pasmanter, R (2001), 'Empirical dynamical system modeling of ENSO using nonlinear inverse techniques', *Journal of Physical Oceanography*, **31**(6), 1579–1598.

Wikle, C (2015), 'Modern perspectives on statistics for spatio-temporal data', *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**(1), 86–98.

Wikle, CK & Holan, SH (2011), 'Polynomial nonlinear spatio-temporal integro-difference equation models', *Journal of Time Series Analysis*, **32**(4), 339–350.

Wikle, CK & Hooten, MB (2010), 'A general science-based framework for dynamical spatio-temporal models', *TEST*, **19**(3), 417–451.

Yao, W, Zeng, Z, Lian, C & Tang, H (2013), Ensembles of echo state networks for time series prediction, *Advanced Computational Intelligence (ICACI), 2013 Sixth International Conference on*, *IEEE*, Hangzhou, China, 299–304.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.