Agricultural price forecasting using NARX model for soybean oil

Ronit Jaiswal[1], Girish K. Jha[1,*], Rajeev Ranjan Kumar[2] and Achal Lama[2]

[1]ICAR-Indian Agricultural Research Institute, PUSA, New Delhi 110 012, India
[2]ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110 012, India

The non-linear, non-stationary and complicated nature of agricultural price series makes their accurate forecasting extremely challenging. In comparison to standard statistical methods, artificial neural networks (ANN) have demonstrated promising results for predicting such series. However, the incorporation of auxiliary information can improve prediction accuracy if it is closely linked to the target series. A dynamical neural architecture called a non-linear autoregressive model with exogenous input (NARX) carefully makes use of the auxiliary information to construct a data-dependent non-linear forecasting model. The study explores the performance of NARX model for the real price series of soybean oil (soybean) using soybean (soybean oil) price as exogenous inputs. NARX models outperform ARIMA, ARIMAX and ANN models in terms of RMSE, MAPE, MASE and directional statistics as evaluation criteria. Further, the Diebold-Mariano test confirms a significant improvement in its predictive accuracy.

Keywords: Artificial neural networks, mean absolute scaled error, NARX, price forecasting, soybean oil.

PRICE forecasting for agricultural commodities is a challenging field of time series analysis. Production and price of agricultural commodities are not only governed by market forces and state regulations but are also affected by several abiotic factors like natural disasters, extreme weather conditions, etc. and biotic factors like pests, diseases, etc.[1]. Reliable and accurate price forecasting is not only necessary for the food security of consumers but also for the producers, as it plays a key role in affecting their income. A thorough review of the literature affirms that many research works have endeavoured to construct models for price series intricacies with the aim of enhancing forecasting accuracy[2]. These works have utilized broadly two categories of forecasting models, namely statistical and machine learning models. Statistical models include linear models, such as autoregressive integrated moving average (ARIMA) and its submodels[3], and non-linear models such as smooth transition autoregressive (STAR), self-exciting threshold autoregressive (SETAR), and others[4]. However, due to prerequisite assumptions and the demand for a precise relationship among data, these statistical models cannot efficiently capture the non-linearity and complexity inherent in the price series.

*For correspondence. (e-mail: girish.stat@gmail.com)

Among the machine learning models, artificial neural networks (ANN), specifically, time delay neural networks (TDNN)[5], have been dominant over statistical techniques for the last two decades[6]. An important strategy for improving the performance of the TDNN model is the use of auxiliary information. The present study attempts to apply this strategy to non-linear agricultural price series by building non-linear autoregressive with exogenous inputs (NARX) models[7,8] with the lags of exogenous series. Compared to other machine learning approaches, NARX models allow for a simpler explanation of the parameter being studied, based on its relationship with exogenous inputs. NARX models also offer improved short-term forecasting due to the iterative updating of predictors and model parameters[9]. NARX networks offer a significant benefit over other ANN techniques as they are able to rapidly attain optimal weights for the connections between neurons and input parameters, requiring fewer iterations to construct an effective model[10].

Soybean, scientifically known as *Glycine max* (L.) Merr., is an important commodity worldwide due to its versatility as a low-cost source of protein, unsaturated fats, carbohydrates, livestock feed, and biofuel. Recently, the health advantages associated with soybean oil have increased its global demand. As a result, the demand and production of soybeans are expected to grow in the upcoming years, providing benefits to farmers, purchasers, animal feed and biofuel manufacturers, as well as food producers. Therefore, it is crucial to have precise and reliable forecasting of soybean prices. The contribution of this study is that we develop optimized NARX models for price forecasting of soybean oil (soybean) using soybean (soybean oil) price as exogenous inputs. Furthermore, we conduct a comprehensive assessment of the predictive accuracy of the NARX model in comparison to the ARIMA, autoregressive integrated moving average with exogenous variables (ARIMAX) and TDNN models for both the target series and exogenous price series.

NARX is a non-linear system for discrete-time input–output modelling that may be described mathematically as

$$y(t+1) = f\,[y(t), \ldots, y(t-d_y+1);\; x(t), \ldots, x(t-d_x+1)], \tag{1}$$

where $y(t) \in \mathbb{R}$ and $x(t) \in \mathbb{R}$ indicate the output and input of the model at discrete time-step t respectively, while $d_x \geq 1$ and $d_y \geq 1$, are lags of the exogenous series and the target series respectively, where $d_x \geq d_y$. The model in vector form is as

$$y(t+1) = f\,[\mathbf{y}(t); \mathbf{x}(t)], \tag{2}$$

where the vectors $\mathbf{x}(t)$ and $\mathbf{y}(t)$ represent respectively, the input and output regressors. Typically, the non-linear mapping $f(.)$ is unknown and must be approximated. The resulting architecture is referred to as a NARX network[11], a powerful dynamical model that is comparable computationally to Turing machines[12]. The architecture of a single hidden layer NARX network is depicted in Figure 1. NARX networks are frequently more effective than traditional neural networks in detecting long-term dependencies. Back propagation through time (BPTT) is used to train these networks. BPTT consists of two phases: time-unfolding the network and error backpropagation over the unfolded network. A NARX network's output delays during the initial stage appear as jump-ahead connections, which reduce the network's sensitivity to long-term dependencies by giving gradient information a shorter path. As a result, a NARX network may replace any traditional network without sacrificing processing capability.

The ARIMA model is a group of models utilized to analyse time series data. The general ARIMA (p, d, q) model represents p order of the autoregressive (AR) part, d degree of first differencing, and q order of the moving average (MA) part. The model can be represented as $\varphi(B)\Delta^d y_t = \theta(B)u_t$, where y_t denotes the price series at time t and u_t is an error term assumed to be independently and identically distributed with zero mean and variance of σ^2. B is the backshift operator defined as $By_t = y_{t-1}$ and $\Delta = (1 - B)$ represents the differencing operator. Additionally, $\varphi(B)$ is polynomial of degree p in B and $\theta(B)$ is polynomial of degree q in B. ARIMAX is a generalized version of the ARIMA model. ARIMA is appropriate for univariate datasets, whereas ARIMAX may include extra explanatory factors. Consider two stationary time series, y_t and x_t, then the transfer function model (TFM) may be expressed as follows

$$y_t = C + \mu(B)x_t + u_t, \tag{3}$$

where x_t is input series, y_t output series and C is the constant term. $\mu(B)x_t$ denotes response function that permits x to impact y through lags and thus we can write

$$\mu(B)x_t = (\mu_0 + \mu_1 B + \mu_2 B^2 + \cdots)x_t. \tag{4}$$

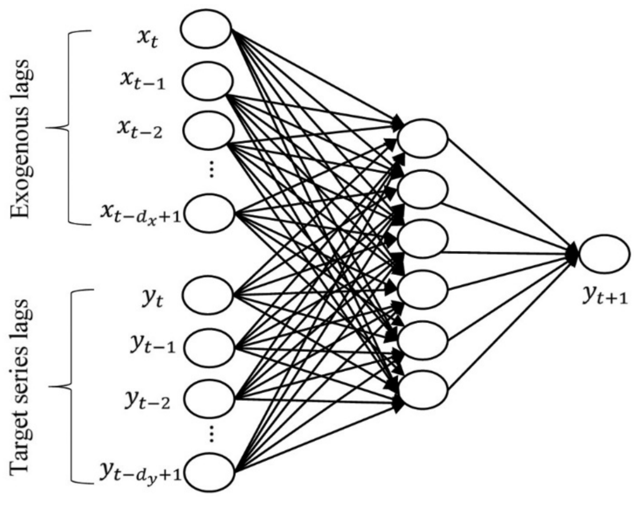

Figure 1. Topology of a single hidden layer NARX network with d_x delayed inputs and d_y delayed outputs.

Equation (3) is known as the ARIMAX model when x_t and u_t follow the ARIMA model. The coefficients μ_j are the impulse response weights which may be positive or negative.

Time delay neural network (TDNN)[13] utilizes the time lags of a univariate series to portray its temporal dimension as short-term memory[5]. The TDNN model[14] may be stated mathematically as

$$y_{t+1} = \gamma\left(\sum_{m=1}^{q} \partial_m \psi\left(\sum_{j=0}^{p} w_{mj} y_{t-j} + b_m\right)\right), \tag{5}$$

where $y_t, y_{t-1}, \ldots, y_{t-p}$ represents the inputs for a neuron. The output of the neuron at time step $t + 1$ is denoted as y_{t+1}, while $\psi(.)$ and $\gamma(.)$ represent the activation functions of the hidden and output nodes respectively. ∂_m is the weight between the mth hidden and output neurons, and w_{mj} is the synaptic weight between the jth input and the mth hidden neurons. Finally, b_m represents the bias term.

As there is no unique criterion for measuring accuracy that can reflect the complete distributional characteristics of the errors, four distinct metrics, i.e. root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE) and directional statistics (D_{stat}) are used to compare the prediction output of the developed model to that of the other existing models. We used both RMSE and MAPE because the first is a scale-dependent measure, whereas the latter is unit-free. On the other hand, MASE is a metric that can be computed as the average of the scaled error[15]. D_{stat} is employed to assess the precision of predicting the direction of price movement changes in the forecasts generated by each model. The mathematical formulae for the four metrics are presented below

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y(t) - \hat{y}(t))^2},$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y(t) - \hat{y}(t)}{y(t)}\right| \times 100\%,$$

$$\text{MASE} = \frac{\frac{1}{n}\sum |y(t) - \hat{y}(t)|}{\frac{1}{T-1}\sum_{t=2}^{T}|y(t) - y(t-1)|},$$

$$D_{\text{stat}} = \frac{1}{n}\sum_{t=1}^{n} a(t) \times 100\%,$$

where $y(t)$ and $\hat{y}(t)$ indicate the actual and predicted values respectively; the prediction size is denoted by n, and $a(t) = 1$ if $(y(t+1) - y(t)) \times (\hat{y}(t+1) - \hat{y}(t)) \geq 0$ or $a(t) = 0$ otherwise. For the MASE calculation, the denominator corresponds to the mean absolute error of the one step 'naive forecast method' applied to the training data set (t = 1, …, T). Furthermore, to evaluate the notable enhancement in the forecast accuracy of all models, the Diebold-Mariano (DM) test is utilized[16]. The null hypothesis of the test assumes the loss differential, $d_{(t)} = f(e_{1(t)}) - f(e_{2(t)})$; $t = 1, 2, 3, \ldots, n$ has zero expectation, i.e. both predictions are equal in accuracy, where ($e_{1(t)}$) and ($e_{2(t)}$) denote the error series from the forecasts of any two models. The DM test statistic is

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}},$$

where n denotes the prediction size, $\bar{d} = \frac{1}{n}\sum_{t=1}^{n} d_{(t)}$ is the sample mean of the loss differential, $\hat{V}(\bar{d}) = \frac{1}{n}[\gamma_0 + 2\sum_{k=1}^{h-1}\hat{\gamma}_k]$ is variance estimate of the mean using h-step forecasts, and $\hat{\gamma}_k = \frac{1}{n}[\sum_{t=k+1}^{n}(d_{(t)} - \bar{d})(d_{(t-k)} - \bar{d})]$ is an estimate of kth autocovariance of $d_{(t)}$.

The NARX model has been evaluated empirically using monthly international price ($/MT) data for soybean oil and soybean, which have been sourced from the 'World Bank Commodity Price Data' for the period between January 1980 and November 2021. Both price series comprise 503 observations and Figure 2 shows their time plots. Table 1 depicts the basic statistics of the price series, where it can be seen that the standard deviations of both series are high. Both series are leptokurtic and positively skewed in nature, while the non-normal nature of each series is confirmed using Jarque-Bera test[17].

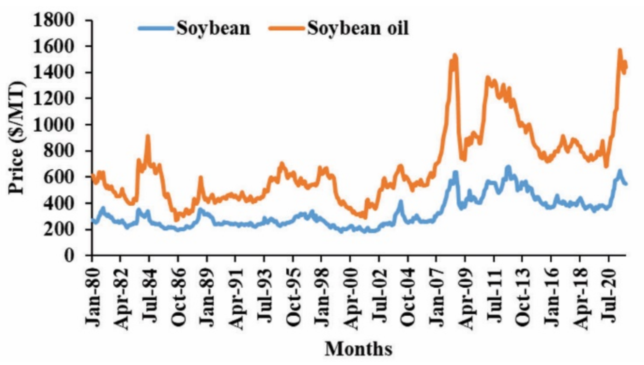

Figure 2. Monthly international price of soybean oil and soybean.

Table 1. Various descriptive statistics of monthly international price ($/MT) of soybean oil and soybean

Statistics	Soybean oil	Soybean
Average	672.9	327.60
Minimum	271.0	183.00
Maximum	1574.7	684.00
Standard deviation	279.52	115.51
Kurtosis	0.713	0.11
Skewness	1.10	1.01
Jarque-Bera (p-value)	112.66 (<0.001)	86.69 (<0.001)

The non-stationarity feature of two price series is confirmed using augmented Dickey-Fuller (ADF)[18] and Phillip–Perron (PP)[19] tests, as shown in Table 2. Additionally, the non-linearity of two data series is confirmed through the Brock–Dechert–Scheinkman (BDS)[20] test, as presented in Table 3. The dataset comprising 503 data points was divided into training and testing sets for model fitting, where the testing set contained 12 data points (one year), and the remaining formed the training set for building models and in-sample prediction. R statistical software version 3.6.1 is used to develop the models and perform statistical analysis in this study.

In order to fit NARX models for the price series of soybean oil, the price series of soybean is considered an exogenous series, and vice versa. The time plots and descriptive statistics of the data show that each series is unstable and has high variability. Hence, before implementing models on the datasets, we employed a natural logarithmic modification to stabilize the variance of the data. The process of modelling begins with data pre-processing and converting it to a supervised learning format. We used the NARX model architecture, which included an input layer, a hidden layer and an output layer with a single output node. Hidden layer and output layer employ a sigmoid and an identity activation function respectively. The NARX network is constructed in two phases. The training phase is carried out using a series–parallel design, sometimes known as an open loop architecture, because it resembles pure feedforward. This particular architecture takes into account the actual output instead of using the estimated output, which allows for a more precise input to be fed into the feedforward network. In the second phase, the configuration of series-parallel is altered to a parallel configuration for testing. This closed-loop setup is ideal for forecasting multiple steps. As a result, during the testing phase, the estimated result is used as the input value for the next computation iteration. The NARX model's training necessitates the adjustment of several hyperparameters, including the number of input nodes, hidden nodes, output nodes and epochs. We set the number of epochs at 500 and the output node at 1 (ref. 21).

We utilized the grid search technique to explore all feasible combinations of input and hidden nodes by creating a two-dimensional grid to adjust them. The allowable range for the number of nodes at hidden layer and lags is between 2 and 20 for both the target and exogenous series. Each setup is repeated 50 times, and the average outcome is utilized to identify the most precise configuration. The findings indicate that when using soybean as an exogenous series, the NARX model with 4 and 6 nodes at input and hidden layers respectively, performs the best for soybean oil. For soybean, however, the NARX model with 6 and 8 nodes at input and hidden layers respectively, utilizing soybean oil as an exogenous series exhibits better performance in terms of RMSE and MAPE. The ARIMA, ARIMAX and TDNN models are fitted as per the standard procedure[14,22].

Table 2. Results of Philips Perron (PP) and Augmented Dickey-Fuller (ADF) tests on soybean oil and soybean price series

	PP unit root test		ADF test		
Price series	Statistic	*p*-value	Statistic	*p*-value	Remarks
Soybean oil	–16.22	0.21	–3.42	0.05	Non-stationary
Soybean	–18.17	0.10	–3.51	0.04	Non-stationary

Table 3. Results of Brock–Dechert–Scheinkman (BDS) test on soybean oil and soybean price series

	Soybean oil		Soybean		
	Embedding dimensions		Embedding dimensions		
Epsilon	2	3	2	3	*p*-value
0.5σ	109.36	177.70	79.88	124.01	<0.0001
1.0σ	59.98	70.74	61.84	74.62	<0.0001
1.5σ	46.54	48.04	50.09	52.31	<0.0001
2.0σ	40.34	39.10	42.92	42.01	<0.0001

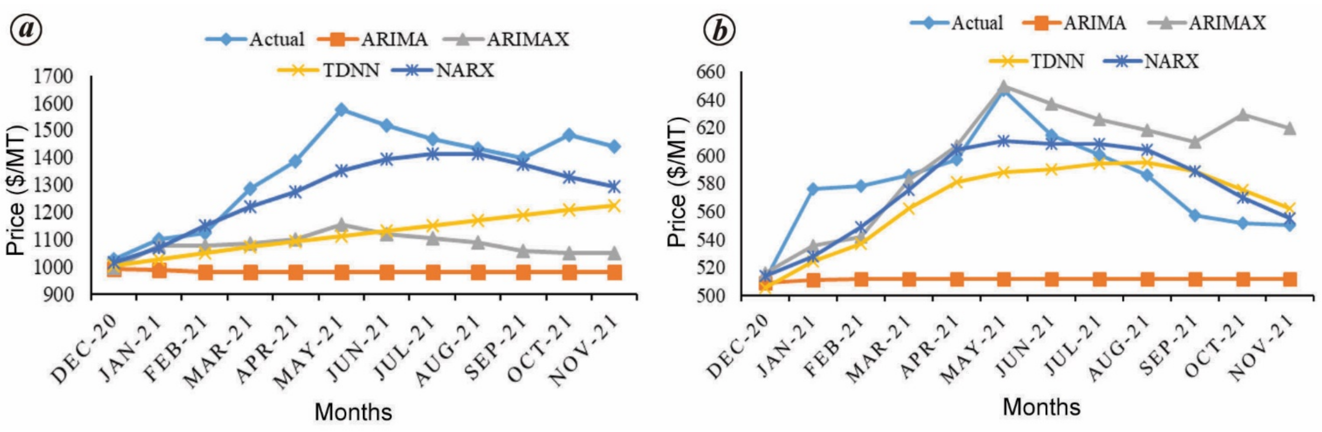

Figure 3. Price forecasts versus actual price of different models for (***a***) soybean oil and (***b***) soybean.

In case of fitting soybean price series, both ARIMA and ARIMAX models were optimized at $p = 1$, $d = 1$ and $q = 0$, while the number of lags for the exogenous series for ARIMAX model was 1. In case of fitting soybean oil price series, both ARIMA and ARIMAX models were optimized at $p = 3$, $d = 1$ and $q = 1$, while the number of lags for the exogenous series for ARIMAX model was 3. In case of soybean series, TDNN model was optimized at 4 input nodes and 8 hidden nodes, while in case of soybean oil series TDNN model was optimized at 4 input nodes and 6 hidden nodes. We were interested in short-term forecasting and hence evaluated forecast horizons of up to 12 months. Figure 3 depicts the forecasted series and the level series for both soybean oil and soybean price series using ARIMA, ARIMAX, TDNN and NARX models. From the figure, it can be seen that the forecasts are more accurate when obtained using NARX models. Table 4 confirms the superiority of these models across both evaluation criteria and Figure 4 presents the bar diagrams of the RMSE of each model for both series. The empirical results show that NARX model gets the lowest scores for the three evaluation criteria RMSE, MAPE and MASE. From Table 4, it is observed that NARX models perform better than ARIMAX models as it is obvious that ARIMAX models are inefficient at capturing non-linearity. Further, it was observed that the use of exogenous series improved both the statistical and neural network models in terms of RMSE and MAPE. While comparing the NARX model among other models through MASE, we used ARIMA model as the benchmark model. Thus, each model was scaled from the ARIMA model while calculating MASE and the result is shown in Table 4. The values of MASE also confirmed that NARX was better than TDNN which was better than ARMAX. The directional statistics evaluated from the forecasts obtained from every model are also presented in Table 4. We found that NARX was capturing the direction of price movement better as compared to other models. Further, we also observed that the models that utilized exogenous series (ARIMAX

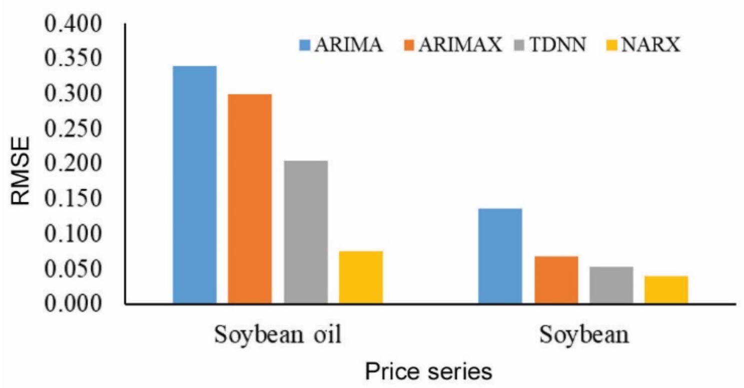

Figure 4. Bar diagram of RMSE of soybean oil and soybean for different models.

Table 4. Optimized results in terms of RMSE, MAPE and D_{stat} (%) of every model for price series of soybean oil and soybean

Criteria	Series	ARIMA	ARIMAX	TDNN	NARX
RMSE	Soybean oil	0.339	0.300	0.205	0.075
	Soybean	0.136	0.068	0.053	0.040
MAPE	Soybean oil	4.278	4.119	2.516	0.828
	Soybean	1.930	0.845	0.689	0.499
MASE	Soybean oil	1.000	0.699	0.588	0.194
	Soybean	1.000	0.435	0.356	0.258
D_{stat} (%)	Soybean oil	27	73	55	91
	Soybean	91	91	73	100

Table 5. Diebold-Mariano test results showing test statistics of different models for 12 forecast horizons with *p* values (in bracket)

	Soybean oil			Soybean		
	Benchmark models			Benchmark models		
Tested models	ARIMA	ARIMAX	TDNN	ARIMA	ARIMAX	TDNN
ARIMAX	4.80 (<0.001)			2.32 (0.02)		
TDNN	5.26 (<0.001)	1.94 (0.04)		3.73 (0.001)	0.92 (0.19)	
NARX	5.23 (<0.001)	4.86 (<0.000)	3.95 (0.001)	3.68 (0.001)	1.69 (0.05)	2.16 (0.03)

and NARX) had a better capability of capturing the price movement direction. The DM test was employed to investigate the improvement in predictive accuracy among the NARX models over other models. The DM test results are presented in Table 5 for both soybean oil and soybean series, which demonstrate that the NARX model significantly outperformed the other benchmark models and has superior accuracy in forecasts over them.

The key benefit of this study is that we have used a strategy by utilizing the price series of related commodities as exogenous series to improve the modelling ability of the neural network model which gives a better prediction of prices of such irregular agricultural price series. Such reliable, robust, and more accurate predictions may guide the farmers and various stakeholders in analysing and taking proper decisions on production, timing to conduct open market operations and other policy decisions well in advance.

A promising forecasting model called the NARX model is empirically evaluated and its performance is compared with ARIMA, TDNN and ARIMAX models using monthly international price series of soybean oil and soybean. We developed the NARX model for the price series of soybean oil (soybean) considering the price series of soybean (soybean oil) as an exogenous series. From the empirical evaluation, we confirm that we can improve the neural network models for agriculture price series by utilizing the series of correlated exogenous series, and thus NARX models outperform other competing models. This shows that the prices of closely linked commodities may be valuable for better predictive capacity. Results can be utilized as technical forecasts on their own or in conjunction with fundamental projections to create perspectives on price patterns and conduct policy research. However, even if neural network models (particularly NARX) are more effective at capturing the patterns of a non-linear series, training these models requires adept data management, vigorous coding skills, and adequate amounts of time. Future work may involve investigating the performance of models utilizing any decomposition techniques before applying the NARX model or multilayer NARX network for agricultural price forecasting.

Conflict of interest: The authors declare that they have no conflict of interest.

Code and data availability: The R code and the data can be made available on request.

1. Wang, J., Wang, Z., Li, X. and Zhou, H., Artificial bee colony-based combination approach to forecasting agricultural commodity prices. *Int. J. Forecast.*, 2022, **38**, 21–34.
2. Wang, L., Feng, J., Sui, X., Chu, X. and Mu, W., Agricultural product price forecasting methods: research advances and trend. *Br. Food J.*, 2020, **122**(7), 2121–2138.
3. Box, G. E. P., Reinsel, G. C., Jenkins, M. and Ljung, G. M., *Time Series Analysis: Forecasting and Control*, Wiley, Hoboken, 2015, 5th edn.
4. Weron, R., Electricity price forecasting: a review of the state-of-the-art with a look into the future. *Int. J. Forecast.*, 2014, **30**, 1030–1081.
5. Zhang, G., Patuwo, B. E. and Hu, M. Y., Forecasting with artificial neural networks: the state of the art. *Int. J. Forecast.*, 1998, **14**, 35–62.
6. Makirdakis, S., Spiliotis, E. and Assimakopoulos, V., The accuracy of machine learning (ML) forecasting methods versus statistical ones: extending the results of the M3-competition. Working Paper, University of Nicosia, Institute for the Future, Greece, 2017.
7. Leontaritis, I. J. and Billings, S. A., Input-output parametric models for non-linear systems Part I: Deterministic non-linear systems. *Int. J. Control*, 1985, **41**, 303–328.
8. Narendra, K. S. and Parthasarathy, K., Identification and control of dynamical systems using neural networks. *IEEE Trans. Neural Networks*, 1990, **I**, 4–27.
9. Di Nunno, F., Granata, F., Gargano, R. and de Marinis, G., Prediction of spring flows using nonlinear autoregressive exogenous (NARX) neural network models. *Environ. Monit. Assess.*, 2021, **193**(6), 1–17.
10. Guzman, S. M., Paz, J. O. and Tagert, M. L. M., The use of NARX neural networks to forecast daily groundwater levels. *Water Resour. Manage.*, 2017, **31**(5), 1591–1603.
11. Marcjasz, G., Uniejewski, B. and Weron, R., On the importance of the long-term seasonal component in day-ahead electricity price forecasting with NARX neural networks. *Int. J. Forecast.*, 2019, **35**, 1520–1532.
12. Siegelmann, H. T., Horne, B. G. and Giles, C. L., Computational capabilities of recurrent NARX neural networks. *IEEE Trans. Syst. Man Cybernet. Part B: Cybernetics*, 1997, **27**, 208–215.
13. Haykin, S., *Neural Networks and Learning Machines Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, Pearson Education, India, 2018, 3rd edn.
14. Jha, G. K. and Sinha, K., Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India. *Neural Comput. Appl.*, 2014, **24**, 563–571.
15. Hyndman, R. J. and Koehler, A. B., Another look at measures of forecast accuracy. *Int. J. Forecast.*, 2006, **22**, 679–688.
16. Diebold, F. X. and Mariano, R. S., Comparing predictive accuracy. *J. Business Econ. Stat.*, 2002, **20**, 134–144.
17. Jarque, C. M. and Bera, A. K., Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ. Lett.*, 1980, **6**, 255–259.
18. Said, S. E. and Dickey, D. A., Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 1984, **71**, 599–607.
19. Escobari, D., Garcia, S. and Mellado, C., Identifying bubbles in Latin American equity markets: Phillips-Perron-based tests and linkages. *Emer. Markets Rev.*, 2017, **33**, 90–101.
20. Cromwell, J., Labys, W. and Terraza, M., *Univariate Tests for Time Series Models Univariate Tests for Time Series Models*, Sage, California, USA, 2011.
21. Jaiswal, R., Jha, G. K., Kumar, R. R. and Choudhary, K., Deep long short-term memory based model for agricultural price forecasting. *Neural Comput. Appl.*, 2022, **34**, 4661–4676.
22. Jha, G. K. and Sinha, K., Agricultural price forecasting using neural network model: an innovative information delivery system. *Agric. Econ. Res. Rev.*, 2013, **26**, 229–239.

ACKNOWLEDGEMENT. The first author thanks the University Grants Commission (UGC) for providing financial support and PG School, Indian Agricultural Research Institute (IARI), New Delhi, for providing necessary facilities.

Received 8 December 2022; revised accepted 24 March 2023

doi: 10.18520/cs/v125/i1/79-84