frontiers | Frontiers in Agronomy

TYPE Original Research
PUBLISHED 27 October 2025
DOI 10.3389/fagro.2025.1687988

Check for updates

OPEN ACCESS

EDITED BY
Rajeev Ranjan Kumar,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Sadikul Islam,
Indian Institute of Soil and Water Conservation (ICAR), India
Nobin Chandra Paul,
National Institute of Abiotic Stress Management (ICAR), India

*CORRESPONDENCE
Pradip Basak
pradipbasak.99@gmail.com

RECEIVED 18 August 2025
ACCEPTED 22 September 2025
PUBLISHED 27 October 2025

CITATION
Basak P, Sultana S, Gupta DS, Paul T, Debnath MK, Sarkar P, Hembram S and Kheroar S (2025) Integrating weather variables and AI models for forecasting major pests in jute: applications in climate-smart crop management. *Front. Agron.* 7:1687988. doi: 10.3389/fagro.2025.1687988

Integrating weather variables and AI models for forecasting major pests in jute: applications in climate-smart crop management

Pradip Basak[1]*, Shifat Sultana[1], Deb Sankar Gupta[1], Tarun Paul[2], Manoj Kanti Debnath[1], Prahlad Sarkar[3], Satyajit Hembram[4] and Shyamal Kheroar[3]

[1]Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal, India, [2]Department of Agronomy, Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal, India, [3]AINP-Jute and Allied Fibres, Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal, India, [4]RRS, Terai Zone, Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal, India

Jute crop suffers a substantial amount of physical and economic loss every year due to the infestation of several insect pests, such as yellow mite (*Polyphagotarsonemus latus* Banks) and jute semilooper (*Anomis sabulifera* Guen), at different stages of crop growth. This study utilizes data on the mean incidence of yellow mite and jute semilooper at different days after sowing (DAS) from 2013 to 2023, along with weather variables, collected at the AINP-JAF, UBKV Centre, Cooch Behar, West Bengal. The results indicate that the incidence of jute semilooper follows a seasonal pattern, with most peaks occurring at approximately 45 DAS. Additionally, the mean incidence of yellow mite is found to be significantly positively correlated with maximum temperature and negatively correlated with minimum and maximum relative humidity at a 2-week lag. This suggests that dry weather with high temperatures 2 weeks prior contributes to higher yellow mite infestations at the current time. A similar correlation is observed for jute semilooper infestation. Various time series and machine learning models, including Autoregressive Integrated Moving Average (ARIMA), ARIMA-T, Seasonal ARIMA (SARIMA), SARIMA-T, ARIMA with exogenous variables (ARIMAX), SARIMA with exogenous variables (SARIMAX)-T, Random Forest, Support Vector Regression (SVR), and TDNNX, are applied to the training dataset from 2013 to 2022. The models are validated using the test data for the year 2023, based on root mean square error (RMSE) and root median square error (RMdSE) values. For yellow mite, TDNNX is found to be the best fitted model followed by SVR and SARIMAX-T in terms of RMSE and RMdSE values. Similarly, for jute semilooper, TDNNX is found to be the best fitted model followed by Random Forest and SARIMA. Finally, pest incidence forecasts for yellow mite and jute semilooper are obtained for 2024 using the forecasted and average weather data, applying the TDNNX model.

KEYWORDS

weather variables, SARIMA, SARIMAX, SVR, random forest, TDNNX, major pests, jute

Introduction

Jute, a cost-effective natural fiber, ranks second only to cotton in terms of production and versatility. In India, jute farming covers 646.11 thousand hectares, yielding 94.49 lakh tons with a productivity of 2.62 tons per hectare in 2023–2024 (Directorate of Economics and Statistics, MoAFW, Govt. of India). The raw jute sector provides significant employment to the rural population due to its labor-intensive nature. West Bengal, which accounts for 78.94% of the cultivated jute land and 82% of the production, is the leading jute-growing state in India. The sector holds social, economic, and physical value for approximately 33–35 lakh small and marginal farmers in India (Sarkar and Majumdar, 2016). In 2023–2024, West Bengal cultivated 6.46 lakh hectares of jute, producing 94.49 lakh bales (Directorate of Economics and Statistics, MoAFW, Govt. of India). Cooch Behar is one of the major districts in the state for jute fiber cultivation.

Jute is an essential cash crop for farmers, but challenges such as pest infestations and pricing issues hinder its growth. Oversupply in the market and the scarcity of high-yielding seeds restrict the farmers' ability to adopt advanced technologies, reducing their interest in high-yielding jute varieties (Hussain et al., 2002). Jute cultivation, typically carried out during the pre-kharif season, suffers significant physical and economic losses each year due to infestations by major insect pests like yellow mite (*Polyphagotarsonemus latus* Banks) and jute semilooper (*Anomis sabulifera* Guen) at various stages of crop growth. In West Bengal, the avoidable loss in fiber yield has been estimated to range from 31% to 34% (Rahman and Khan, 2012). These losses could be minimized through sustainable plant protection measures, such as integrated pest management, biological control, and mechanical methods.

Timely and accurate forecasting of pest incidence using mathematical, statistical, and simulation models can help minimize these losses by enabling farmers to implement appropriate pest management strategies. With the advancement of computing power, machine learning models are increasingly used for precise forecasting (Durgabai and Bhargavi, 2018). Numerous studies have shown that weather variables—particularly temperature, relative humidity, and rainfall—play a crucial role in the occurrence and survival of various insect pests on jute crops (Rahman and Khan, 2012; Suyal et al., 2018). It has also been found that pest incidence is correlated with both the current time period and lead times ranging from 1 to 4 weeks (Katke et al., 2009; Balikai and Venkatesh, 2019). Therefore, incorporating weather variables along with past pest incidence data provides a solid foundation for developing reliable pest prediction models. Several researchers have already developed weather-based models for predicting crop pests and diseases (Sarkar et al., 2023; Vaidheki et al., 2023). Therefore, in the present study, an attempt has been made to forecast the mean incidence of major pests of jute crop in Cooch Behar district of West Bengal using machine learning models and weather data.

Materials and methods

Description of data

The incidence data of major jute pests used in this study are secondary data obtained from the All India Network Project (AINP) on Jute and Allied Fibers, Uttar Banga Krishi Viswavidyalaya (UBKV) Centre, Pundibari, Cooch Behar, West Bengal, covering the period from 2013 to 2023. During each crop season, yellow mite and jute semilooper incidences are available at 25, 35, 45, 55, 65, and 75 days after sowing (DAS). The infestation level of the semilooper is measured as the percentage of infestation, while yellow mite incidence is quantified by counting its number per square centimeter on the second unfold leaf. The pest incidence data used in this study were collected from control fields, i.e., without pest control operations, from a widely cultivated variety of jute, JRO-524 recommended for the region. The same variety was consistently used across all years of the study (2013–2023) to maintain uniformity in pest incidence data collection and avoid variability arising from genetic differences among varieties. It was consistently observed across all years that pest incidence at 25 and 75 DAS was zero. Including these zero values in the analysis could lead to inconsistencies in model fitting and forecasting. Moreover, the study focuses primarily on pest incidences that exceed the economic threshold levels—approximately five mites per square centimeter for yellow mite and 10% infestation for jute semilooper, as reported in the literature. Therefore, data from 25 and 75 DAS were excluded, and the analysis was conducted using the remaining data points where significant pest incidence was observed. Similarly, the weather data used in this study are secondary data obtained from the records of Gramin Krishi Mausam Sewa (GKMS), Agrometeorological Field Unit (AMFU), Pundibari, UBKV. The dataset includes daily observations of rainfall (mm), maximum and minimum temperatures (MaxT and MinT in °C), and maximum and minimum relative humidity (MaxRH and MinRH in %) spanning the period from 2013 to 2023. These daily records were aggregated into weekly data based on the Standard Meteorological Weeks (SMWs) to correspond with the pest incidence survey dates. For each SMW, the rainfall values were summed, while the temperature and relative humidity variables were averaged to ensure consistency with the timing of pest observations.

Methodology

To check the presence of seasonality in pest incidence, time plots of the mean incidence of both pests are constructed along with the Webel–Ollech (WO) seasonality test. Pest weather relationship has been studied using the Pearson correlation coefficient between mean pest incidence and weather variables in the current week as well as at the 1- to 2-week lag. Various time series forecasting models, including Autoregressive Integrated Moving Average

(ARIMA), ARIMA with exogenous variables (ARIMAX), Seasonal ARIMA (SARIMA), and SARIMA with exogenous variables (SARIMAX), are implemented. In addition, machine learning models such as Random Forest, Support Vector Regression (SVR), and Time Delay Neural Network with exogenous variables (TDNNX) are used to predict the mean pest incidence based on weather variables. A total of 44 data points from 2013 to 2023 are available, out of which the initial 40 data points (2013–2022) were used for model development and the remaining 4 data points from 2023 were reserved for model validation, maintaining a 90:10 train–test ratio. The analysis has been carried out using R software.

Autoregressive Integrated Moving Average

Being one of the most prevalent time series models, ARIMA (Box and Jenkins, 1976) is suitable for short-term forecasting, and it is dependent on past values of the variable being forecast. The basic formulation of ARIMA (p, d, q) could be narrated as

$$\nabla^d y_t = \mu + \phi_1 \nabla^d y_{t-1} + \ldots + \phi_p \nabla^d y_{t-p} - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where y_t is the value of the dependent variable at time t; y_{t-1}, y_{t-2},…, y_{t-p} are values of the dependent variable at time lags $t-1$, $t-2$…, $t-p$, respectively; μ is the constant mean; φ_1, φ_2,…, φ_p are p autoregression (AR) coefficients to be estimated; θ_1, θ_2,…, θ_q are q moving average (MA) coefficients to be estimated; ε_t is the forecast error at time t, independently and normally distributed with zero mean and constant variance σ^2 $\forall$ $t = 1, 2,\ldots, T$; and d is the order of differencing.

ARIMA with exogenous variables

ARIMAX (Bierens, 1987) is an acronym for autoregressive integrated moving average with exogenous variables. It is a logical extension of the pure ARIMA model that incorporates independent variables that add explanatory value. Conceptually, it is a merging of ARIMA and the regression model. It can be expressed as

$$\nabla^d y_t = \mu + \phi_1 \nabla^d y_{t-1} + \ldots + \phi_p \nabla^d y_{t-p} - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} + \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \varepsilon_t$$

where, x_{1t}, x_{2t},…, x_{kt} are the values of k exogenous variables at time t.

Seasonal ARIMA

The SARIMA model (Box and Jenkins, 1976) is an extension of ARIMA that completely deals with the time series data consisting of seasonal components. In order to improve the performance of the conventional ARIMA model, seasonal data patterns are added to develop the SARIMA model (Box and Jenkins, 1976). The SARIMA model can be formulated as

$$\Phi_P(B^s)\phi_p(B)\nabla_s^D \nabla^d y_t = \mu + \Theta_Q(B^s)\theta_q(B)\varepsilon_t$$

where μ is the intercept or mean term, ε_t is the residual at time t follows $N(0, \sigma^2)$, B is the backward shift operator, and s denotes the number of periods per season. The polynomials $\phi_p(B)$ and $\theta_q(B)$ represent the non-seasonal autoregressive and moving average terms with orders p and q, respectively. Similarly, the seasonal autoregressive and moving average terms of order P and Q, respectively, are represented by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ polynomials, and also, the seasonal and non-seasonal differencing terms are represented by ∇_s^D and ∇^d respectively.

SARIMA with exogenous variables

The SARIMAX model (Box and Jenkins, 1976) is a rational extension of SARIMA that allows the incorporation of explanatory variables. It is an integration of regression and the SARIMA model. If only the SARIMA model is not sufficient to provide an acceptable efficiency, it is very obvious to look for other processes that have the potential to implant in past values of the dependent variables.

Random Forest

Random Forest (Breiman, 2001) has become widely used in machine learning. The method of Random Forest usually works in two steps. In the first step, Random Forest constructs n number of binary classification/regression trees using multiple bootstrap samples with replacement obtained from the original observations. The correct classification/regression is determined by the majority vote/average value of all the trees. The second step is to randomly select input variables (*mtry*) from a random subset of the features (exogenous variables) and to calculate the best split for the tree based only on these selected variables.

Support vector regression

The SVR is a nonlinear modeling procedure that utilizes the principle of structured risk minimization (Vapnik, 2000). The SVR model can be expressed as

$$y = w^T \varphi(x) + b$$

where $\varphi(\,.\,): R^n \rightarrow R^{n_h}$ is a nonlinear mapping function from the original input space into a higher-dimensional feature space, $w \in R^{n_h}$ is the weight vector, b is the bias term, and the superscript T denotes transpose.

Time Delay Neural Network with exogenous variables

A neural network consists of a set of connected cells called neurons or node. The neurons receive information from either

input cells or other neurons and perform some kind of transformation of the input and transmit the outcome to other neurons to output cells. The neural networks are built from layers of neurons connected so that one layer receives input from the preceding layer of neurons and passes the output on to the subsequent layer. In TDNNX (Hyndman and Athanasopoulos, 2021), all neurons are connected through weights. To design a TDNNX, (a) number of input, hidden, and output layer; (b) number of input, hidden, and output node; (c) activation function; (d) bias; and (e) exogenous variables are to specified. The TDNNX performs the following non-linear mapping between input and output.

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, x_{1t}, x_{2t}, \ldots, x_{qt}) + \varepsilon_t$$

where f is the function of the network structure and connection weights. Here, $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, the pth order lag of study variable y, and $x_{1t}, x_{2t}, \ldots, x_{qt}$, the q exogenous variables selected from Random Forest model, are network input nodes.

Forecast evaluation methods

Different forecasting models are evaluated using the criteria of root mean square error (RMSE) and root median square error (RMdSE), which are expressed as

$$RMSE = \sqrt{(1/n)\left[\sum_{t=1}^{n}(y_t - \hat{y}_t)^2\right]}, \text{ and}$$

$$RMdSE = \sqrt{Median(y_t - \hat{y}_t)^2}$$

where y_t and $\hat{y}_t$ are the actual and predicted values of pest incidence at time t, respectively. The RMSE was used as an evaluation criterion to measure the accuracy of the forecasted pest incidence, as described by Chai and Draxler (2014). The RMdSE was employed as an additional criterion to assess forecast performance, providing robustness against extreme values, as discussed in Hyndman and Koehler (2006).

Results and discussion

Descriptive statistics

The descriptive statistics of incidence of both pests and weather variables in the current week are presented in Table 1. It is evident from Table 1 that the variability in pest incidence is quite high for both pests since the coefficient of variation (CV) is found to be 134% and 102% for yellow mite and jute semilooper, respectively. Among the weather variables, rainfall in the current week shows considerably high CV.

Seasonal incidence

The seasonal plots of yellow mite and jute semilooper are presented in Figures 1 and 2, respectively. For both pests, peak incidence is observed on 55 DAS followed by 45 DAS. The results of the WO test in Table 2 indicate that seasonality is present for jute semilooper but is absent for yellow mite.

Correlation analysis

Results of Pearson's correlation analysis in Table 3 reveal that the mean incidence of yellow mite is significantly positively correlated with maximum temperature at a 2-week lag, whereas it is significantly negatively correlated with maximum RH at a 1-week lag, and minimum and maximum RH at a 2-week lag. Similarly, the mean incidence of jute semilooper is also significantly positively correlated with maximum temperature at a 2-week lag, whereas it is significantly negatively correlated with maximum and minimum RH at a 2-week lag. The weather variables that are significantly correlated with pest incidence are further subjected to multicollinearity analysis. The results of multicollinearity analysis in Table 4 reveal that maximum and minimum RH at a 2-week lag show moderate multicollinearity for yellow mite incidence, whereas for jute semilooper, minimum RH at a 2-week lag shows moderate

TABLE 1 Descriptive statistics of pest incidence and weather variables in the current week.

Statistics	Yellow mite (n/cm^2)	Jute semilooper (% infestation)	MaxT (°C)	MinT (°C)	MaxRH (%)	MinRH (%)	Rainfall (mm)
Mean	3.97	5.17	31.64	22.17	83.91	72.56	104.48
Median	2.27	4.23	31.57	22.25	83.94	72.93	57.85
Standard deviation	5.33	5.29	1.97	2.12	10.12	10.42	117.59
CV(%)	134	102	6	10	12	14	112
Minimum	0.00	0.00	28.29	13.00	46.43	37.86	0.00
Maximum	25.62	18.99	36.86	25.29	99.14	92.57	533.90

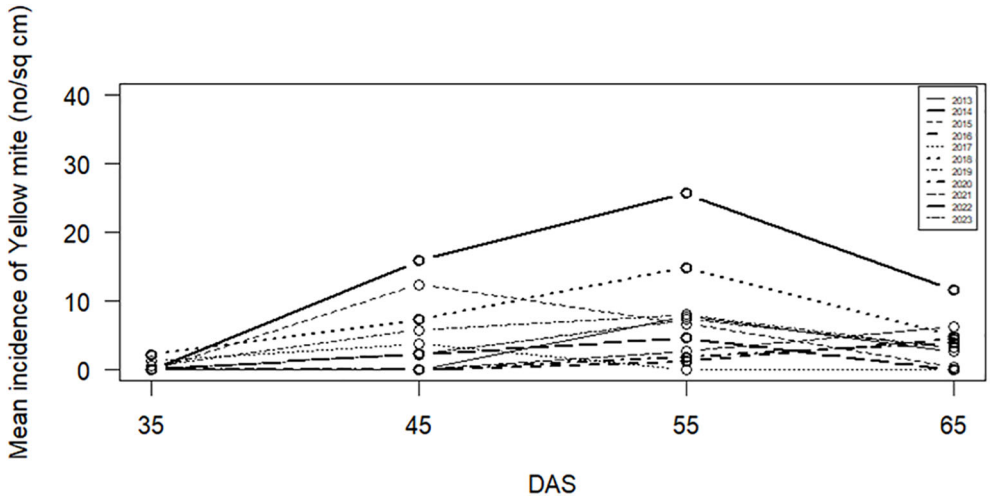

FIGURE 1
Seasonal plot of yellow mite incidence.

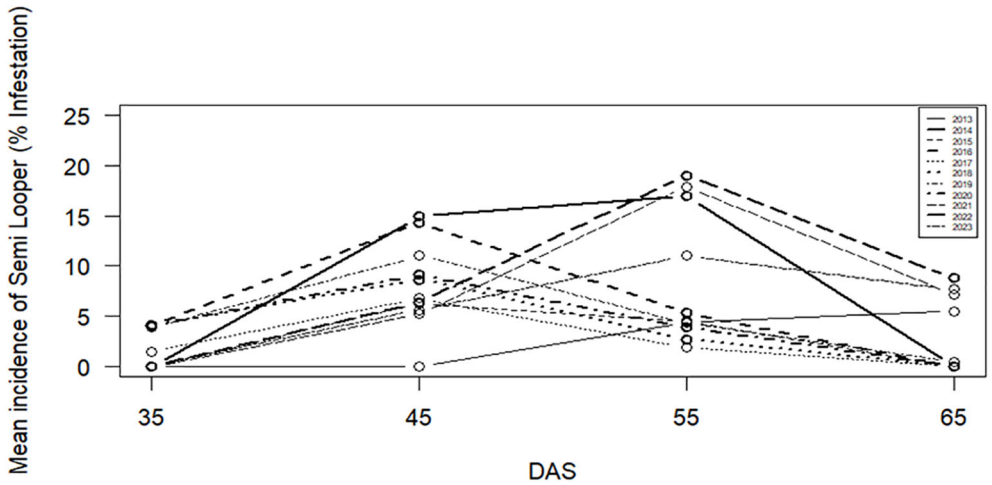

FIGURE 2
Seasonal plot of jute semilooper incidence.

multicollinearity. Therefore, maximum RH at a 1-week lag and maximum temperature at a 2-week lag are to be used as exogenous variables in time series models for yellow mite incidence, whereas for jute semilooper, the exogenous variables are maximum temperature and maximum RH at a 2-week lag.

TABLE 2 WO test to check seasonality.

	Test statistic	*p*-value		
Yellow mite	0.04	1	0.014	0.04
Jute semilooper	1	0.0001	0.0004	0.003*

*Significant at the 5% level of significance.

Fitting of different models for yellow mite and jute semilooper incidence

To check the presence of stationarity in the data series, the Augmented Dickey–Fuller (ADF) test and Phillips–Perron (PP) test have been applied, and the results are presented in Table 5. It is found that the both data series are stationary and, therefore, regular differencing is not required.

After confirming the stationarity of the time series data of yellow mite incidence, an ARIMA model is fitted and the parameters are presented in Table 6. The residuals of the fitted ARIMA model are found to be non-normal, as evident from Table 7; therefore, the original data series is transformed using square root transformation with the addition of 0.5 as few zero values are there. The transformed data exhibit seasonality as evident from the WO test, and therefore, a suitable SARIMA model is fitted

TABLE 3 Correlation between the mean incidence of pests with weather variables.

Weather variables	Yellow mite			Jute semilooper		
	Current week	One-week lag	Two-week lag	Current week	One-week lag	Two-week lag
MaxT	−0.16	0.26	0.34*	0.02	0.15	0.39**
MinT	−0.27	−0.16	−0.21	−0.02	−0.05	−0.08
MaxRH	−0.19	−0.40**	−0.54**	−0.15	−0.16	−0.37*
MinRH	−0.01	−0.27	−0.46**	−0.07	−0.15	−0.41**
Rainfall	−0.17	−0.21	−0.18	−0.05	−0.04	−0.24

*: Significant at 5% level of significance ; **Significant at the 1% level of significance.

TABLE 4 VIF values of significantly correlated weather variables.

Yellow mite		Jute semilooper	
Variables	VIF	Variables	VIF
MaxRH lag1	3.15	MaxT lag2	1.66
MaxT lag2	2.10	MaxRH lag2	5.54
MaxRH lag2	7.92^	MinRH lag2	6.34^
MinRH lag2	6.39^		

^Moderate multicollinearity.

based on the minimum AIC and BIC criteria, and this model is referred to as SARIMA-T. The parameters of the fitted SARIMA-T model are also presented in Table 6. Subsequently, MaxRH at 1 week and MaxT at 2 weeks are used as exogenous variables in the SARIMA-T model and, therefore, referred to as the SARIMAX-T model.

Similarly, the results of ADF and PP tests for jute semilooper incidence in Table 4 indicate that the data series is stationary, and the WO test in Table 2 indicates that the series is seasonal. Therefore, the SARIMA model is fitted to predict the incidence of jute semilooper and the parameters of the model are presented in Table 8. Since the residuals of the fitted SARIMA model depict non-normality as evident from Table 9, square root transformation is therefore applied on the original data with the addition of 0.5 to it. The WO test on the square root transformed data of semilooper incidence indicates non-seasonality, and accordingly, the ARIMA (2, 0, 1) model is fitted on this transformed data, and subsequently, the model is referred to as ARIMA-T. The parameters of the fitted ARIMA-T model are also presented in Table 8. In case of jute semilooper, MaxT and MaxRH at 2 weeks are found to be the most important exogenous variables.

TABLE 5 ADF and PP test for stationarity.

Yellow mite				Jute semilooper			
ADF test		PP test		ADF test		PP test	
Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value
−4.01	0.01**	−4.128	0.015*	−5.17	0.01**	−5.16	0.01**

*: Significant at 5% level of significance; **: Significant at 1% level of significance.

TABLE 6 Parameter estimates of the ARIMA (0, 0, 1), SARIMA-T (0, 0, 1) $(1,0,0)_4$, and SARIMAX-T (0,0,1) $(1,0,0)_4$ models for yellow mite incidence.

Model	Parameters	Estimate	SE	*p*-value
ARIMA (0, 0, 1)	C	3.98	1.20	0.001**
	MA1	0.71	0.11	0.000***
SARIMA-T (0, 0, 1) $(1,0,0)_4$	C	1.76	0.32	0.000***
	MA1	0.57	0.15	0.000***
	SAR1	0.27	0.16	0.09
SARIMAX-T (0,0,1) $(1,0,0)_4$	MA1	0.49	0.17	0.004**
	SAR1	0.33	0.15	0.035*
	MaxRH lag1	−0.03	0.01	0.029*
	MaxT lag2	0.13	0.04	0.000**

*: Significant at 5% level of significance; **: Significant at 1% level of significance; ***Significant at the 0.1% level of significance.

TABLE 7 Residual diagnostics test for time series models of yellow mite incidence.

Diagnostic test	ARIMA		SARIMA-T		SARIMAX-T	
	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value
Box–Ljung	6.31	0.61	2.98	0.94	3.57	0.89
Shapiro–Wilk	0.87	0.0003*	0.95	0.07	0.96	0.14
ARCH LM	19.29	0.01	14.15	0.08	7.65	0.47

*: Significant at 5% level of significance.

Therefore, ARIMAX (3,1,0) is found to be the best fitted model on the original time series data of jute semilooper incidence. The parameter estimates with standard error (SE) and *p*-values are presented in Table 8. However, the residuals of fitted ARIMAX model are found to be non-normal and therefore, ARIMAX model is fitted on the square root transformed data. This model is referred to as ARIMAX-T. The estimate of parameters, its SE, and the respective *p*-values are also presented in Table 8.

TABLE 8 Parameter estimates of the SARIMA (0, 0, 1) $(1,0,0)_4$, ARIMA-T (2, 0, 1), ARIMAX (3, 1, 0), and SARIMAX-T (0, 0, 1) $(1,0,0)_4$ models for jute semilooper incidence.

Model	Parameters	Estimate	SE	*p*-value
SARIMA (0, 0, 1) $(1,0,0)_4$	C	5.11	1.96	0.009**
	MA1	0.63	0.15	0.000***
	SAR1	0.49	0.14	0.000***
ARIMA-T (2, 0, 1)	C	2.06	0.11	0.000***
	AR1	−0.02	0.14	0.912
	AR2	−0.70	0.13	0.000***
	MA1	0.63	0.25	0.011*
ARIMAX (3, 1, 0)	AR1	−0.75	0.14	0.000***
	AR2	−0.92	0.09	0.000***
	AR3	−0.61	0.14	0.000***
	MaxT lag2	0.49	0.28	0.079
	MaxRH lag2	−0.19	0.05	0.000***
ARIMAX-T (0, 0, 1) $(1,0,0)_4$	MA1	0.87	0.12	0.000***
	SAR1	0.70	0.12	0.000***
	MaxT lag2	0.11	0.02	0.000***
	MaxRH lag2	−0.01	0.01	0.015*

*: Significant at 5% level of significance; **: Significant at 1% level of significance; ***: Significant at 0.1% level of significance.

Random Forest relies on two key algorithms: bagging and random feature selection. Bagging involves selecting a specified number of samples (*ntree*) from the dataset using simple random sampling with replacement (SRSWR) to construct multiple trees. Random feature selection determines the number of variables (*mtry*) to consider at each split. The hyperparameters of the fitted Random Forest model are provided in Table 10. The Random Forest model also evaluates the % increase in node impurity (IncNodePurity) to identify important variables. For predicting the mean incidence of yellow mite, rainfall and MaxT at a 2-week lag are identified as significant weather variables, as shown in Figure 3. Similarly, for predicting the jute semilooper incidence, rainfall and MaxT at a 2-week lag are also recognized as important weather variables, as shown in Figure 4. The key weather variables identified from the variance importance plot in the Random Forest model, rainfall and MaxT at a 2-week lag, are used as exogenous variables in both the SVR and TDNNX models.

For both pests, the SVR model is trained with the mean incidence as the dependent variable and rainfall and MaxT at a 2-

TABLE 10 Hyperparameters of the Random Forest model for both yellow mite and jute semilooper incidences.

Hyperparameters	Yellow mite	Jute semilooper
Type of Random Forest	Regression	Regression
Number of trees (*ntree*)	500	500
No. of variables tried at each split (*mtry*)	5	5
Mean of squared residuals	27.62	32.36

TABLE 9 Residual diagnostics test for time series models of jute semilooper incidence.

Diagnostic test	SARIMA		ARIMA-T		ARIMAX		ARIMAX-T	
	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic	*p*-value
Box–Ljung	5.60	0.69	2.76	0.95	5.22	0.73	6.97	0.54
Shapiro–Wilk	0.86	0.0001***	0.88	0.0006***	0.92	0.007**	0.92	0.008*
ARCH LM	2.72	0.95	1.92	0.98	4.34	0.82	4.94	0.76

*: Significant at 5% level of significance; **: Significant at 1% level of significance; ***: Significant at 0.1% level of significance.

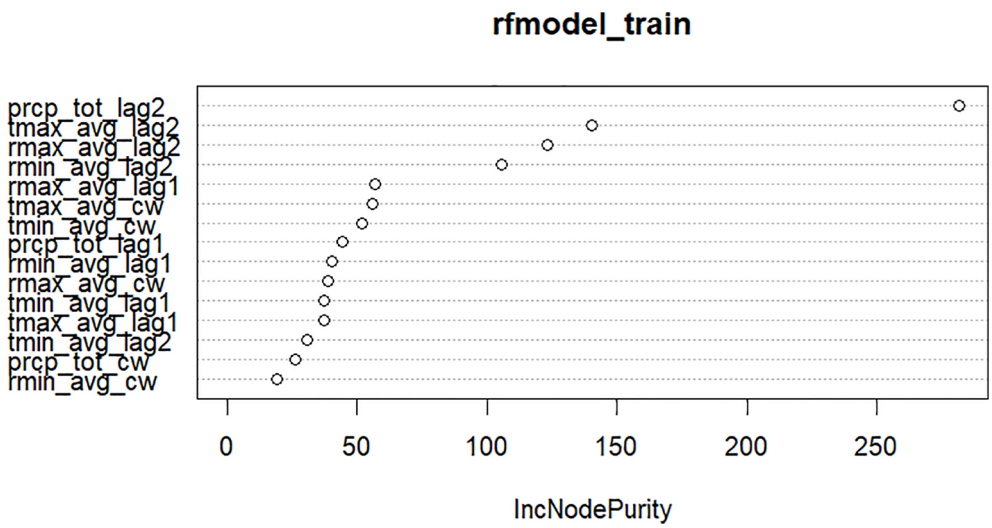

FIGURE 3
Variable importance plot of the Random Forest model for yellow mite incidence.

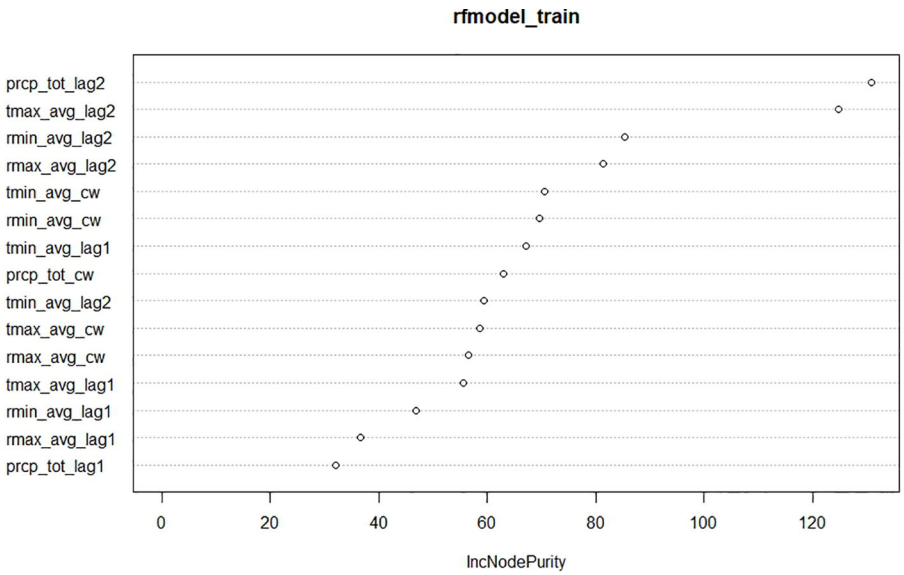

FIGURE 4
Variable importance plot of the Random Forest model for jute semilooper incidence.

week lag as the exogenous variables. The hyperparameters of the fitted SVR model are provided in Table 11.

The TDNNX model is fitted between the mean incidence of pest with its lagged values up to the significant order and exogenous weather variables selected from the Random Forest model. In the TDNNX (p,k) model, p and k indicate the number of nodes in the input and hidden layer, respectively. For both yellow mite and semilooper incidences, there are five nodes in the input layer that

TABLE 11 Hyperparameters of the SVR model for both yellow mite and jute semilooper incidences.

Pest	Type	Kernel	Cost (C)	Gamma	Epsilon (ε)	No. of support vectors
Yellow mite	eps-regression	Radial	1	0.5	0.1	39
Jute semilooper	eps-regression	Radial	1	0.5	0.1	37

TABLE 12 Hyperparameters of the TDNNX model for both yellow mite and jute semilooper incidences.

Number of input nodes (*p*)	Number of hidden nodes (*k*)	Exogenous variables	Activation function
5	3	Rainfall at a 2-week lag MaxT at a 2-week lag	Sigmoid (input layer) Linear (output layer)

correspond to lagged values of mean pest incidence up to the order of three and two exogenous weather variables. The hyperparameters of the fitted TDNNX model are presented in Table 12.

Model validation

The predictive abilities of different time series and machine learning models are compared using RMSE and RMdSE as the evaluation criteria for both training and testing dataset, and the

TABLE 13 Predictive abilities of different models for both yellow mite and jute semilooper incidences.

Pest	Model	RMSE		RMdSE	
		Training	Testing	Training	Testing
Yellow mite	ARIMA (0,0,1)	4.50	2.11	2.60	1.72
	SARIMA-T (0,0,1) (1,0,0) [4]	4.72	2.25	2.06	1.06
	SARIMAX-T (0,0,1)(1,0,0)$_4$	4.14	1.85	1.89	1.76
	SVR	4.52	1.33	2.14	0.98
	Random Forest	2.28	2.92	1.41	2.46
	TDNNX	1.39	0.60	0.85	0.39
Jute semilooper	SARIMA (0,0,1) (1,0,0)$_4$	4.23	1.01	1.95	0.87
	ARIMA-T (2,0,1)	3.43	2.91	1.18	2.83
	ARIMAX (3,1,0)	3.35	7.47	1.81	7.07
	SARIMAX-T (0,0,1) (1,0,0)$_4$	3.91	4.82	1.13	2.34
	SVR	4.86	3.10	2.42	2.40
	Random Forest	2.40	3.96	1.78	0.45
	TDNNX	0.82	2.49	0.30	0.91

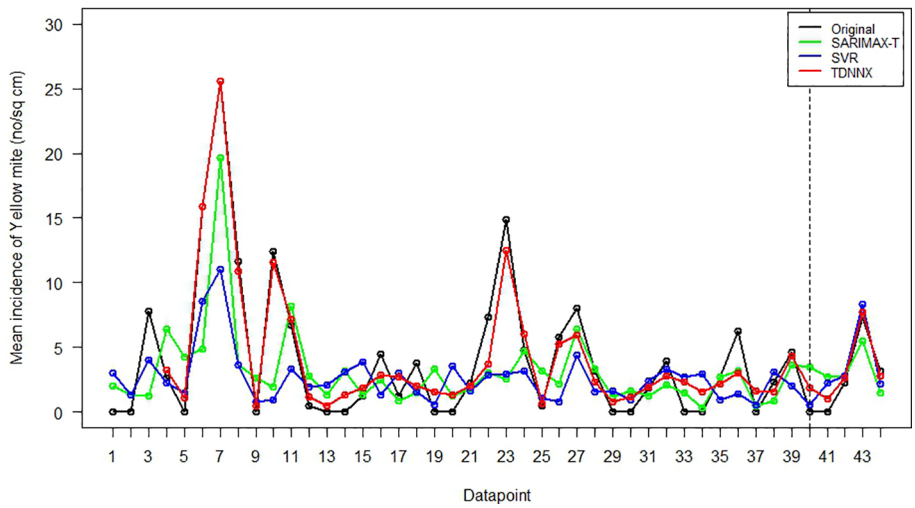

FIGURE 5
Plot of observed vs. fitted values by SARIMAX-T, SVR, and TDNNX models for yellow mite incidence.

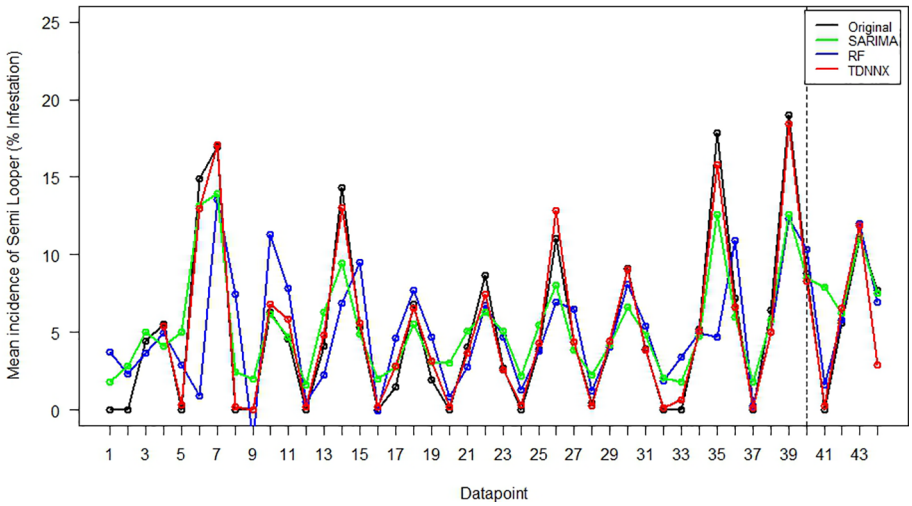

FIGURE 6
Plot of observed vs. fitted values of SARIMA, Random Forest, and TDNNX models for semilooper incidence.

TABLE 14 Out-of-sample forecast for the mean incidence of yellow mite and jute semilooper for the year 2024 using forecasted weather data.

DAS	Forecasted rainfall at lag2	Forecasted MaxT at lag2	Yellow mite mean incidence (n/cm^2)	Jute semilooper mean incidence (% infestation)
35	73.84	31.50	1.93	0.05
45	77.21	31.50	2.35	1.77
55	78.19	31.50	2.17	7.81
65	78.47	31.50	2.06	15.30

TABLE 15 Out-of-sample forecast for the mean incidence of yellow mite and jute semilooper for the year 2024 using average weather data from 2013 to 2023.

DAS	Average rainfall at lag2	Average MaxT at lag2	Yellow mite mean incidence (n/cm^2)	Jute semilooper mean incidence (% infestation)
35	77.5	31.1	1.91	0.11
45	51.9	31.6	2.42	1.41
55	88.5	31.7	2.12	8.12
65	97.4	31.6	2.05	12.60

results are presented in Table 13. For the mean incidence of yellow mite, TDNNX is found to be the best fitted model followed by SVR and SARIMA-T on the basis of RMSE and RMdSE values. Similarly, for jute semilooper, it is observed that the TDNNX model produces the least RMSE and RMdSE value for the training dataset followed by Random Forest, but for the testing dataset, SARIMA has the least RMSE values followed by Random Forest and TDNNX. However, the residuals of the fitted SARIMA model are not normally distributed, and therefore, model assumptions are violated. Thus, by considering both RMSE and RMdSE values, TDNNX may also be considered as the best fitted model for prediction of the mean incidence of jute semilooper followed by Random Forest and SARIMA. The plots of observed vs. fitted values for the mean incidence of yellow mite and jute semilooper across different models are shown in Figures 5 and 6, respectively.

Forecasting of pest incidence

The out-of-sample forecasts for pest incidence of yellow mite and jute semilooper are obtained using the TDNNX model for the year 2024 at 35, 45, 55, and 65 DAS. The forecasts are obtained using forecasted values of weather variables, i.e., rainfall and MaxT at a 2-week lag, from the ARIMA model, and the results are presented in Table 14. Additionally, out-of-sample forecasts of pest incidence are also made using the average values of weather

variables (rainfall and MaxT at a 2-week lag) for the period from 2013 to 2023, and the results are presented in Table 15.

Conclusion

The study found that the peak mean incidence of jute semilooper and yellow mite occurs at 45 and 55 DAS, respectively, with seasonality observed only in jute semilooper. Among the crop years studied, the peak mean incidence of yellow mite was recorded in 2014, while for jute semilooper, it occurred in 2022. The study also revealed that incidence of yellow mite has a significant positive correlation with maximum temperature at a 2-week lag while the correlation with maximum relative humidity at a 1- and 2-week lag is highly significant in a negative direction. This suggests that dry weather with high temperatures 2 weeks prior leads to higher yellow mite infestations at present. Consequently, if such weather conditions are observed, there is a likelihood of increased mite infestation in the following 2 weeks. This information can assist farmers in better preparing for pest emergence and in making informed decisions regarding pest control measures. Among the different time series and machine learning models, the TDNNX model was found to be the most accurate for predicting the mean incidence of both yellow mite and jute semilooper using weather variables.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

PB: Writing – review & editing, Conceptualization, Writing – original draft. SS: Investigation, Writing – original draft, Writing – review & editing. DG: Supervision, Writing – review & editing, Writing – original draft. TP: Visualization, Methodology, Writing – review & editing. MD: Software, Writing – review & editing. PS: Writing – review & editing, Visualization. SH: Writing – review & editing, Data curation. SK: Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Balikai, R. A., and Venkatesh, H. (2019). Weather based prediction models to forecast major pests of rabi sorghum in Karnataka, India. *J. Exp. Zool. India* 22, 131–134.

Bierens, H. J. (1987). ARMAX model specification testing, with an application to unemployment in the Netherlands. *J. Econom.* 35, 161–190. doi: 10.1016/0304-4076(87)90086-8

Box, G. E. P., and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control. Revised Edition* (San Francisco: Holden Day).

Breiman, L. (2001). Random forests. *Mach. Learn* 45, 5–32. doi: 10.1023/A:1010933404324

Chai, T., and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model. Dev.* 7, 1247–1250. doi: 10.5194/gmd-7-1247-2014

Durgabai, R. P. L., and Bhargavi, P. (2018). Pest management using machine learning algorithms: a review. *Int. J. Comp. Sci. Eng. Info. Tech. Res.* 8, 13–22.

Hussain, M., Rahman, M., Uddin, M. N., Taher, M. A., Nabi, M. N., and Mollah, A. F. (2002). Problems and solutions in jute cultivation faced by the farmers in a selected area of Bangladesh. *J. Biol. Sci.* 2 (9), 628–629. doi: 10.3923/jbs.2002.628.629

Hyndman, R. J., and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice (3rd ed.)* (Melbourne, Australia: OTexts).

Hyndman, R. J., and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *Int. J. Forecast* 22, 679–688. doi: 10.1016/j.ijforecast.2006.03.001

Katke, M. K., Balikai, R. A., and Venkatesh, H. (2009). Seasonal incidence of grape mealy bug, Maconellicoccus hirsutus (green) and its relation with weather parameters. *Pest Mgmt. Hortic. Eco.* 15, 9–16.

Rahman, S., and Khan, M. R. (2012). Incidence of pests in jute (Corchorus olitorius L.) ecosystem and pest–weather relationships in West Bengal, India. *Arch. Phyto. Pl Protec* 45, 591–607. doi: 10.1080/03235408.2011.588053

Sarkar, P., Basak, P., Panda, C. S., Gupta, D. S., and Ray, M. (2023). Prediction of peak pest population incidences in jute crop based on weather variables using statistical and machine learning models: A case study from West Bengal. *J. Agrometeorol.* 25, 305–311. doi: 10.54386/jam.v25i2.1915

Sarkar, S., and Majumdar, B. (2016). Present status of jute production and technological and social interventions needed for making jute agriculture sustainable and remunerative in West Bengal. *Indian J. Nat. Fib.* 3, 23–36.

Suyal, P., Gaur, N., Rukesh, P. K. N., and Devrani, A. (2018). Seasonal incidence of insect pests and their natural enemies on soybean crop. *J. Entomon. Zoo Stud.* 6, 1237–1240.

Vaidheki, M., Gupta, D. S., Basak, P., Debnath, M. K., Hembram, S., and Ajith, S. (2023). Prediction of potato late blight disease severity based on weather variables using statistical and machine learning models: A case study from West Bengal. *J. Agrometeorol.* 25, 583–588. doi: 10.54386/jam.v25i4.2272

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory. 2nd Edition* (Berlin: Springer). doi: 10.1007/978-1-4757-3264-1