Smart Agricultural Technology 12 (2025) 101381

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

An innovative two-stage, zero-inflated, hybrid count time series model for predicting rice yellow stem borer using weather parameters in hotspot locations of India

Bojjireddygari Nanda Kumar Reddy [a,1], Santosha Rathod [b,i,1,*], Yerram Sridhar [b,*], Supriya Kallakuri [a], Pramit Pandit [c], Bellamkonda Jyostna [j], Seetalam Malathi [d], R Shravan Kumar [d], Sanjay Sharma [e], K Karthikeyan [f], NRG Varma [g], Sitesh Chatterjee [h], Ayyagari Phani Padmakumari [b], Nethi Somasekhar [b], Sailaja Banda [b], Chintalapati Padmavathi [b], Chitra Shanker [b], Ponnuraj Jeyakumar [b], Raman Meenakshi Sundaram [b]

[a] College of Agriculture, Rajendranagar, Professor Jayashankar Telangana Agricultural University, Hyderabad 500 030, India
[b] ICAR-Indian Institute of Rice Research (IIRR), Rajendranagar, Hyderabad 500 030, India
[c] Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur 741 252, India
[d] Regional Agricultural Research Station, Warangal 506 007, Telangana, India
[e] Indira Gandhi Krishi Vishwavidyalaya, Chhattisgarh, 492 012, India
[f] Regional Agricultural Research Station, Kerala Agricultural University, Pattambi, Kerala 679 306, India
[g] Institute of Rice Research, Professor Jayashankar Telangana Agricultural University, Rajendranagar, Hyderabad 500 030, India
[h] Rice Research Station, Chinsurah, Hooghly, West Bengal 712 102, India
[i] ICAR-National Institute of Abiotic Stress Management, Baramati, Maharastra 413 115, India
[j] Agricultural College, Bapatla, Acharya N G Ranga Agricultural University, Andhra Pradesh 522 101, India

ARTICLE INFO

Keywords:
Yellow stem borer
Count time series
Zero inflated models
Two stage models
INGARCH-ANN
ZIPAR-ANN
ZINBAR-ANN

ABSTRACT

The Yellow Stem Borer (YSB) (*Scirpophaga incertulas* Walker) is a major pest in rice agroecosystems, and timely prediction of its occurrence is crucial for effective management. This study considered data from 2013 to 2023 from five YSB hotspot regions in India, namely Warangal, Rajendranagar, Pattambi, Raipur, and Chinsurah to develop a reliable forewarning model for predicting YSB populations using weather parameters. Daily YSB catches were recorded using light traps with 200 W incandescent bulbs, and various weather variables were also considered. Stepwise regression identified key weather parameters influencing YSB population density, including minimum and maximum temperatures, evening and morning relative humidity, sunshine hours, and rainfall. The study utilized weekly cumulative YSB populations and average climatological data as inputs to several count time series models, including the Integer-valued Generalized Autoregressive Conditional Heteroscedastic (INGARCH) model, Zero-Inflated Poisson Autoregressive (ZIPAR) model, Zero-Inflated Negative Binomial Autoregressive (ZINBAR) model, and the Artificial Neural Network (ANN), a machine learning model. Additionally, innovative two-stage hybrid models viz., INGARCH-ANN, ZIPAR-ANN, and ZINBAR-ANN were developed and evaluated. Classical count time series models, such as INGARCH, underperformed when a high proportion of zeros were observed due to the absence of YSB in certain Standard Meteorological Weeks (SMWs). Zero-inflated models were found to be better suited for such cases. Classical models showed significant residual patterns, indicating the need for model correction. To address this, hybrid models were constructed to normalize the residuals and enhance forecasting accuracy. Among all the models, the two-stage ZINBAR-ANN model outperformed the others, showing the lowest Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) across most locations in both training and testing datasets for both rainy and post-rainy seasons. This innovative two-stage

* Corresponding authors.
E-mail addresses: Santosha.Rathod@icar.org.in (S. Rathod), Yerram.Sridhar@icar.org.in (Y. Sridhar).
[1] Equal Contribution

https://doi.org/10.1016/j.atech.2025.101381
Received 30 January 2025; Received in revised form 21 July 2025; Accepted 25 August 2025
Available online 26 August 2025

zero-inflated hybrid modeling approach offers an effective early warning system for YSB population prediction, supporting sustainable and localized pest management strategies, and helping to mitigate rice yield losses.

1. Introduction

Rice is the most essential and widely consumed food crop in East and Southeast Asia. The importance of rice stretches beyond its nutritional value encompassing social, economic, and environmental benefits. India is among the world's largest rice producing countries; however, substantial yield loss due to the damage by insect pests such as the stem borers, plant hoppers, gall midge, etc. remains a major area of concern [1]. Post green revolution yellow stem borer (YSB) (*Scirpophaga incertulas* Walker) had emerged as one of the important pests throughout the India [1] inflicting around 20 % and 80 % yield losses in early and late planted crops respectively. The YSB larvae bore into the central stem leading to the production of dead tillers at the vegetative stage, popularly known as 'dead heart' and chaffy ear heads called as 'white ears' at reproductive stage (Fig. 1). Continuous flooding and application of high doses of nitrogenous fertilisers are known to be favourable for the population build-up of the stem borers [2]. In addition to natural reproductive potential of insect pests, abiotic factors play a major role in determining their abundance in a crop ecosystem, so, developing of effective statistical model-based early warning system to predict the growth of YSB population is crucial for designing and executing a proactive, site-specific pest control and management strategy.

Count time series modelling is widely used to analyse discrete count data that exhibit autocorrelation, where the observations are usually assumed to follow Poisson or negative binomial distributions [3]. Crop pest modelling is a significant area of research in count time series modelling, where the focus is on daily or weekly counts of insects or pests that exhibit autocorrelation. While count time series models and ML methods have been successfully applied in various fields, their application in modelling and predicting YSB populations is relatively new and innovative. Traditional count time series models have been applied in various fields such as stock exchange data [4,5], monthly claims count of workers in manufacturing industry [6], monthly strike count time series [7], Campylobacterosis infections count time series [8–10] influenza activity prediction using Poisson-INGARCH model and dengue incidents prediction in Jakarta [11], as well as network traffic count time series [12]. In agriculture, crop pest prediction has been reviewed in [13], which explored both regression- and ML-based approaches. Hybrid time series and ML models have also been developed for forecasting crop yields [14]. ML models have been employed in diverse agricultural applications, including banana yield forecasting [15], rice blast disease forecasting [16,17], rice pest prediction [18], early blight severity prediction in tomato crops [19], sugarcane borer disease prediction [20], rice yellow stem borer (YSB) forecasting in the Cauvery command area of Karnataka, India [21], and YSB population prediction using long short-term memory (LSTM) models [22].

The accurate prediction of YSB populations based on climatological parameters is crucial for the implementation of effective and preventive crop protection measures. However, previous attempts on forecasting insect pest populations relied mainly on multiple regression analysis and classical time series models. These methods have limitations while dealing with count data that follows Poisson and negative binomial distributions. Attempt to normalize this type of data does not always lead to accurate prediction models [23,24]. Moreover, in a dataset where a high proportion of zero counts are observed, even a traditional count model may underestimate the variance of the count data, leading to incorrect inference and predictions. Despite the generalised linear model (INGARCH) being better suited for count data, their ability to handle excess zeros in comparison with that of expected number of counts under a Poisson or negative binomial distribution is questionable [25]. These zeros can arise due to various reasons such as the presence of structural zeros (i.e., certain events cannot occur in certain time periods) or excess zeros (i.e., some events have a low probability of occurrence) [26]. To model such phenomena in an effective way, zero inflated models came into picture, where the probability of obtaining a zero count is modelled separately from the probability of obtaining non-zero counts. This is done by incorporating two components into the model: a binary component that models the probability of zero counts and a count component that models the distribution of non-zero counts. The binary component is usually modelled using logistic or a related model, which estimates the probability of a zero count. The count component is typically modelled using a Poisson or negative binomial model, which estimates the distribution of non-zero counts. Zero-inflated models have applications in various fields, including epidemiology, ecology, finance, and social sciences [27–31].

Crop pest modelling is a significant application in this field, where daily or weekly pest counts serve as dependent variable and corresponding weather variables such as temperature, rainfall, relative humidity, sunshine hours etc. as exogenous variables. In dealing with complex zero-inflated datasets, a parametric model may not be sufficient to adequately capture the population dynamics. ML models such as ANN is useful in situations when it is data-driven and devoid of any stringent model assumptions. Moreover, if the residuals of a fitted linear model reveal significant autocorrelation pattern, sequential implementation or hybridization of two models are likely to result superior forecast than its component models [32,33].

This study develops a reliable statistical model for predicting YSB populations by utilizing count time series and machine learning approaches based on climatic input parameters that directly influence the life cycle of YSB. It marks the first attempt to introduce a two-stage modeling framework that integrates a zero-inflated model with an artificial neural network (ANN), using weather variables for insect pest modeling in agriculture, thereby extending the application of ML techniques in forecasting pest populations. Furthermore, this work attempts to combine zero inflated models such as zero inflated Poisson autoregressive (ZIPAR) model and zero inflated Negative Binomial

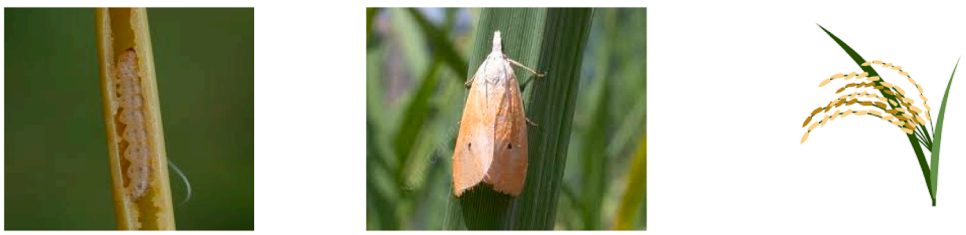

Fig. 1. (a) larva YSB (b) adult YSB (c) symptoms of YSB infected rice.

autoregressive (ZINBAR) model along with ANN for YSB population prediction.

The methodological framework begins with basic descriptive statistics, correlation and stepwise regression analysis to explore the causal relationships between YSB populations and weather variables. Advanced computational methods, such as INGARCH, ZIPAR, ZINBAR, ANN, INGARCH-ANN, ZIPAR-ANN and ZINBAR-ANN are developed to model and forecast YSB populations in hot spot regions of India.

2. Materials and methods

2.1. Data collection

Light trap data on YSB populations from five hotspot locations in India (Warangal, Rajendra Nagar, Pattambi, Chinsurah, and Raipur) (Fig. 2) were utilized for modeling. The data were generated under the All India Coordinated Research Project on Rice (AICRPR) entomology program across years. YSB moths were trapped using Robinson type light traps fitted with 200 W incandescent bulbs, which were illuminated daily from 6:00 pm to 6:00 am. Moths were collected each

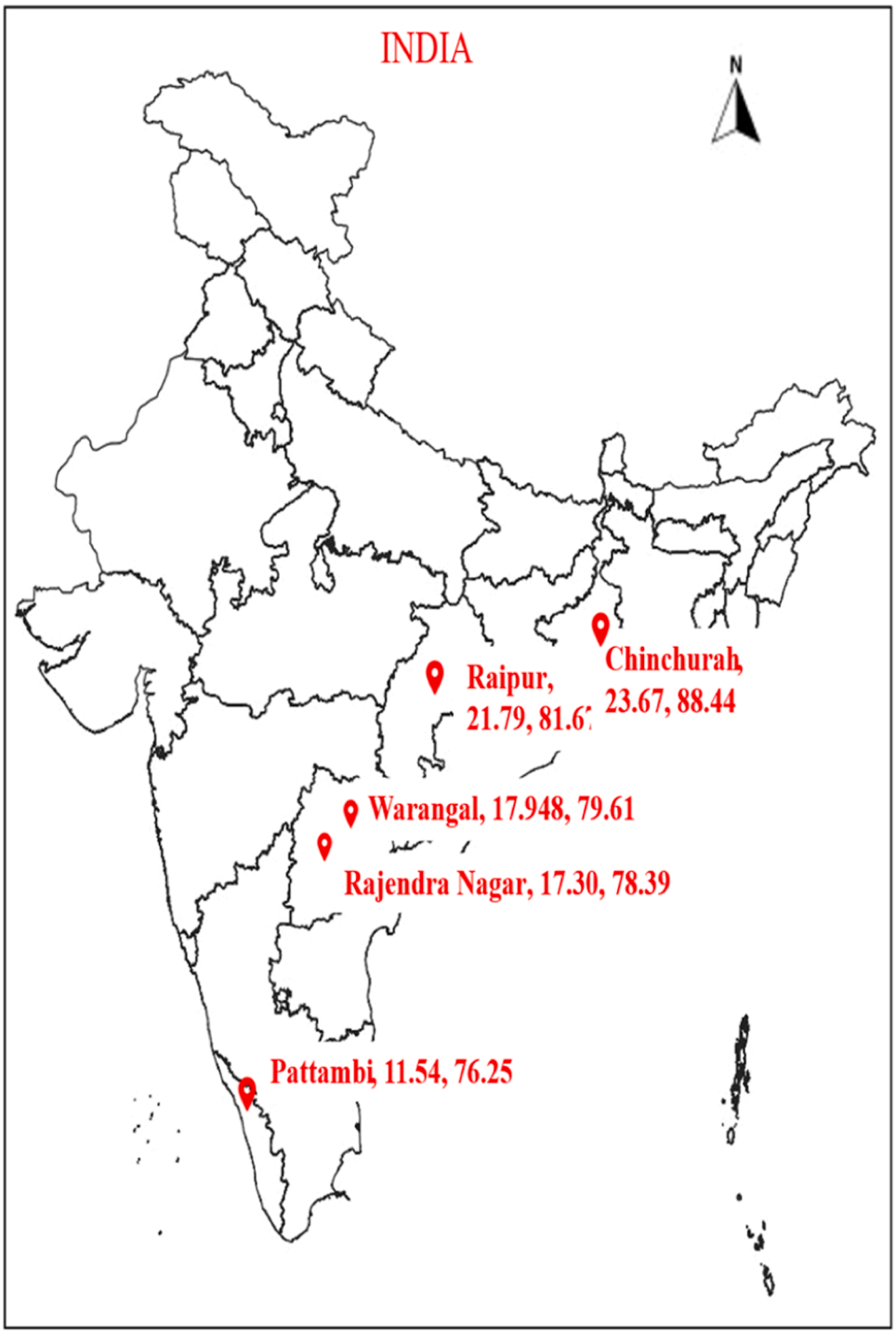

Fig. 2. Study area of YSB pest population.

morning and counted manually. Corresponding daily weather data on MAXT, MINT, RF, MRH, ERH, and SSH were obtained from automatic weather stations at the respective locations. Standard Meteorological Week (SMW) wise cumulative YSB moth catches and weekly mean weather parameters during the past 11 years (2013–2023) were considered for modeling.

As the YSB occurs in both the rainy (*Kharif*) and post-rainy (*Rabi*) season crop of rice with discernible population peaks, the data were apportioned into two sets namely, data set 1 and data set 2 for rainy and post rainy seasons respectively. For data set 1, the first 525 observations were used as the training data set for model building and the last 7 weeks' observations were used as testing data set for validation purposes. Similarly, for data set 2, the first 555 observations and the last 10 weeks' observations were used as the training and the validation (testing) data set, respectively.

2.2. Statistical models

To understand the nature of the data descriptive statistics viz., mean, standard error (SE), skewness, kurtosis, minimum, maximum, and coefficient of variations (CV) were estimated. Graphical depiction of data with time series plots was done. Stepwise correlation analysis preceded by Pearson's correlation analysis was performed for understanding the relationship between the YSB population and exogenous weather variables. The time series plots, INGARCH, ANN and two stage models were developed in R software [34].

2.2.1. INGARCH (Integer valued generalized autoregressive conditional heteroscedastic) model

INGARCH model is a special case of generalised linear model (GLM), where the conditional distribution of dependent variable assumed to follow popular discrete distributions like Poisson, negative binomial, generalised Poisson and double Poisson distributions [35]. Let us denote the count time series by $\{Y_t : t \in N\}$ and time varying r-dimensional covariate vector say $\{X_t : t \in N\}$ *i.e.* $X_t = (X_{t,1} \ldots, X_{t,r})^T$. The conditional mean becomes $E\left(\frac{Y_t}{F_{t-1}}\right) = \lambda_t$ and F_t is historical data.

The generalised model form is expressed as follows:

$$g(\lambda_t) = \beta_o + \sum_{k=1}^{p} \alpha_k \tilde{g}\left(Y_{t-i_k}\right) + \sum_{l=1}^{q} \beta_1 g\left(\lambda_{t-jl}\right) + \eta^T \quad (1)$$

Case 1: Consider the situation where g and $\tilde{g}$ are equal to identity i.e., $g(x) = \tilde{g}(x) = x$, further, Y_t follows (Poisson) INGARCH (p, q) model with $p>1$ and $q \leq 0$ if

a) Y_t conditioned on $Y_{t-1}, Y_{t-2}, \ldots$, is poison distributed

b) The conditional mean $\lambda_t = E\left(\frac{Y_t}{Y_{t-1}, Y_{t-2}, \ldots}\right)$ satisfies

$$\lambda_t = \beta_o + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j} \text{ with } \beta_o > 0 \text{ and } \alpha_1, .., \alpha_p, .., \beta_1, .., \beta_q \geq 0 \quad (2)$$

Assuming further that $Y_t|Y_{t-1}$ is Poisson distributed, then we obtain an INGARCH model of order p and q, abbreviated as INGARCH (p, q) model. If $q = 0$, the model can be referred as the INARCH(p) model.

Case 2: The Negative Binomial (NB) distribution allows for a conditional variance to be larger than the mean λ_t which is often referred to as over-dispersion parameter. It is assumed that $Y_t|F_{t-1} \sim$NB (λ_t, $\varnothing$). When $\varnothing \to \infty$. The Poisson distribution is a limiting case of the Negative binomial distribution by the assumption;

$$\frac{\boldsymbol{Y_t}}{\boldsymbol{Y_{t-1}}, \boldsymbol{Y_{t-2}}, \ldots} \sim \boldsymbol{B}\left(\boldsymbol{n}, \boldsymbol{\beta} + \boldsymbol{\alpha}\frac{\boldsymbol{Y_{t-1}}}{\boldsymbol{n}}\right) \quad (3)$$

Estimation through INGARCH model using conditional likelihood estimation, especially on the asymptotic properties, are given by [36, 37]. The INGARCHX model is an extended version of the INGARCH model, where future values of a variable depend on its past values and past values of exogenous variables.

2.2.2. Zero inflated Poisson Autoregressive (ZIPAR) model

Poisson regression is used to predict a dependent variable that consists of count data given one or more independent variables. The zero inflated Poisson autoregressive (ZIPAR) model is expressed as follows [38]:

$$pr(Y_i = j) = \pi + (1 - \pi)exp(-\mu), \text{ if } j = 0 \quad (4)$$

The Poisson distribution is described as follows

$$(1 - \pi)\frac{\mu^y exp(-\mu)}{y_i}, \text{ if } j > 0 \quad (5)$$

Where, y_i is the logistic link function defined below. The Poisson component can include an exposure time t and a set of k regressor variable. the expression relating these quantities is

$$\mu_i = \exp(ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}) \quad (6)$$

Often, $x_1 = 1$, in which case β_1 is called the intercept, the regression coefficients β_2, β_3,…, β_k are unknown parameters that are estimated from a set of data and their estimates are symbolised as b_1, b_2 …. b_k. This logistic link function π_i is given by

$$\pi_i = \frac{\lambda_i}{1 + \lambda_i} \quad (7)$$

Where, $\lambda_i = \exp\ (ln\ (t_i) + y_1 z_{1i} + y_2 z_{2i} + \ldots + y_m z_{mi})$

The logistic component includes time t and a set of m regressor variables.

2.2.3. Zero inflated negative binomial autoregressive (ZINBAR) model

The zero inflated negative binomial regression is used for count data that exhibit over dispersion and excess zeros. The data distribution combines the negative binomial distribution and the logit distribution [39,40] The possible values of y are the non-negative integers: 0, 1, 2, …

$$Pr\ (y_i = j) = \begin{cases} \pi_i + (1 - \pi_i)g(y_i = 0) \text{ if } j = 0 \\ (1 - \pi_i)g(y_i) \text{ if } j > 0 \end{cases} \quad (8)$$

Where, π_i is the logistic link function defined below and $g(y_i)$ is the negative binomial distribution given by

$$g(y_i) = Pr(Y = y_i|\mu_i,\ \alpha)\ = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\ \Gamma(y_i + 1)}\left(\frac{1}{1 + \alpha\mu_i}\right)\alpha^{-1}\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)y_i \quad (9)$$

The negative binomial component can include an exposure time t and a set of k regressor variable. The expression related to these quantities is

$$\mu_i = \exp\left(ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \phi_1 Y_{i-1} + \phi_2 Y_{i-2} + \ldots\ + \phi_p Y_{i-p}\right) \quad (10)$$

Often, $x_1 = 1$, in which case β_1 is called the intercept, the regression coefficients β_2, β_3,…, β_k are unknown parameters that are estimated from a set of data and their estimates are symbolised as b_1, b_2 …. b_k.

2.2.4. Artificial neural network model (ANN)

ANN is mostly used model among the machine learning techniques. ANNs are non-linear, nonparametric and self-adaptive approaches as opposed to the model-based non-linear methods. Neural networks are composed of layers of neurons where each layer receives input from the

previous layer and passes the output to the next layer.

The general expression for the final output Y_t of a multi-layer feed forward autoregressive neural network is expressed as follows:

$$Y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} Y_{t-i}\right) + \varepsilon_t \quad (11)$$

Where, $\alpha_j (j = 0, 1, 2, \ldots, q)$ and $\beta_{ij} (i = 0, 1, 2, \ldots, p, j = 0, 1, 2, \ldots, q)$ are the model parameters, also called as the synopsis weights, p is the number of input nodes, q is the number of hidden nodes, and g is the activation function. Training part in ANN minimises the error function between actual and predicted values. The error function of autoregressive ANN is expressed as follows:

$$E = \frac{1}{N-p} \sum_{i=t}^{n} (e_t)^2$$
$$= \frac{1}{N-p} \sum_{t=p+1}^{N} \left[Y_t - \left\{ \alpha_0 + \sum_{j=1}^{q} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} Y_{t-i}\right) \right\} \right]^2 \quad (12)$$

Where, N is the total number of error terms.

2.2.5. Proposed two stage modelling

The rationale for selecting the Zero-Inflated Negative Binomial Autoregressive (ZINBAR) model lies in its ability to handle over dispersed count data with excess zeros, which are common in pest and disease time series [25,39]. Classical models such as Poisson-INGARCH assume equi-dispersion and are often inadequate in capturing both the excess zeros and the temporal dependency structure [4]. Furthermore, while ZINBAR captures the statistical structure of zero-inflated, auto-correlated count data, it may fall short in modeling non-linear patterns and complex interactions [40]. This is where Artificial Neural Networks (ANNs) are beneficial, as they excel in capturing hidden non-linear dependencies and adapt to irregularities in input-output relationships. The hybrid ZINBAR-ANN framework integrates the strengths of both paradigms: the probabilistic rigor of ZINBAR for zero-inflation and auto-correlation, and the flexibility of ANN for non-linearity. This hybridization has demonstrated improved predictive accuracy in modeling time series with structural zeros. The proposed two stage modelling in this work considers the time series *Yt* as a combination of both auto-correlated original time series and significant residuals of the model. This approach follows the Zhang's [41] hybrid approach, accordingly the relationship between auto-correlated count time series and significant residuals were considered. In this work, the auto-correlated count time series were modelled using INGARCH, ZIPAR and ZINBAR models (Stage-I) and significant residuals were modelled using ANN model (Stage II).

The proposed methodology consists of two steps. Firstly, an INGARCH, ZIPAR and ZINBAR models were employed to model the count time series data. In the second step, if the residuals obtained from INGARCH, ZIPAR and ZINBAR models were found (Stage II) to be significant by Box pierce test and nonlinear by the BDS (Brock-Dechert-Scheinkman) test, then they were modelled and predicted using the ANN model. Finally, the forecasted values from stage 1 and stage 2 components were combined to generate aggregated forecasted values. The schematic representation of proposed methodology is depicted in Fig. 3.

$$\widehat{Y}_t = \widehat{S_1} + \widehat{S_2} \quad (13)$$

Where, $\widehat{S_1}$ and $\widehat{S_2}$ represent the predicted count time series (stage I) and predicted significant residual components (stage II), respectively.

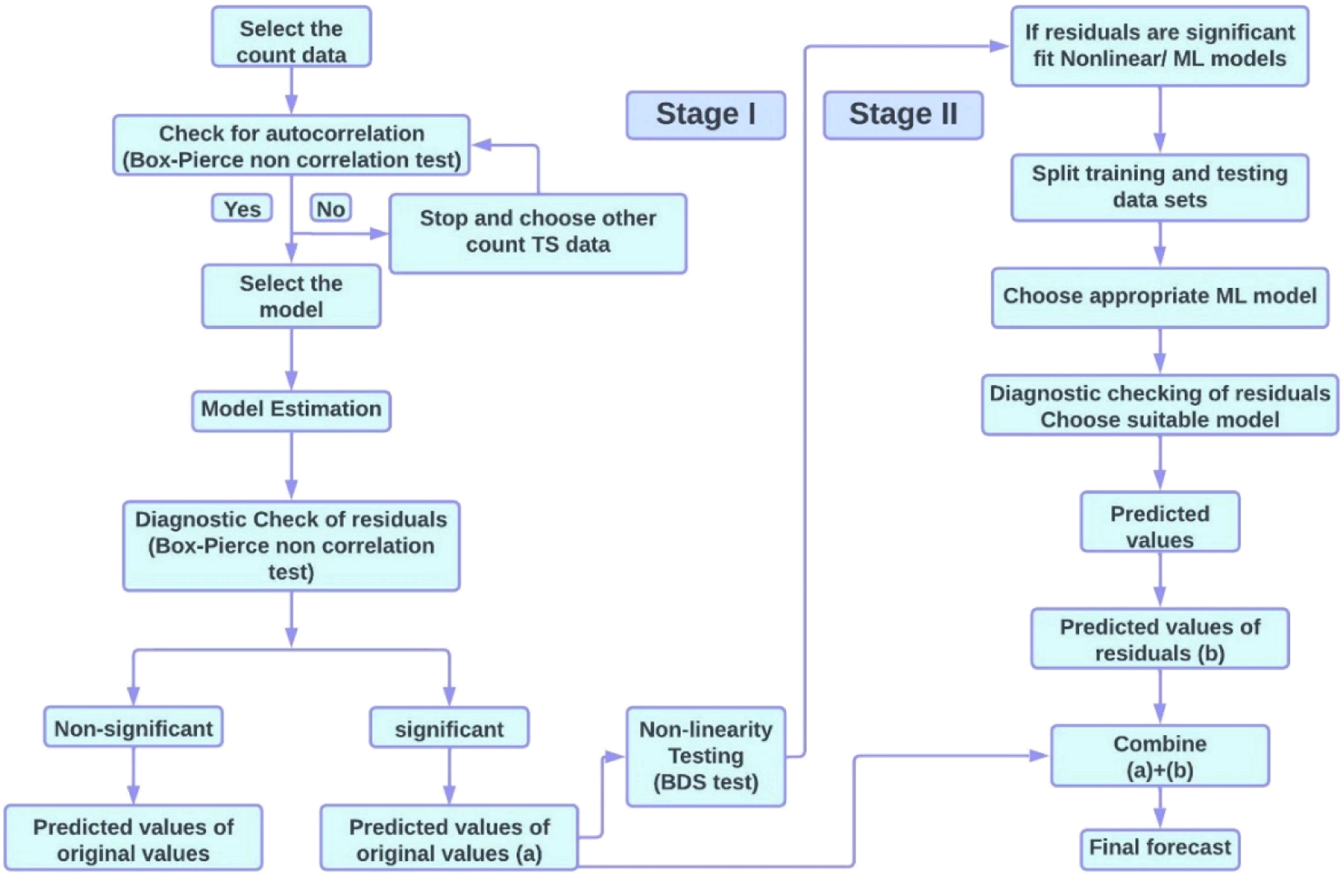

Fig. 3. Schematic representation of proposed two stage methodology.

2.3. Comparison criteria

2.3.1. MSE and RMSE

The mean squared error (MSE) and root mean square error (RMSE) are the two criteria used to measure model accuracy in this study. MSE measures error in statistical models by calculating the average squared difference between observed and predicted values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{14}$$

Where, n is number of data points, Y_i and $\widehat{Y}_i$ are the observed and predicted values, respectively. RMSE is the square root of the mean of the square of all the errors.

2.3.2. MAE

Mean Absolute Error (MAE) measures the average absolute difference between the predicted and actual values. It is a scale-dependent metric but is widely used due to its simplicity and interpretability.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y}_t| \tag{15}$$

Where, y_t is actual observed value at time t, $\widehat{y}_t$ is predicted value at time t and n is the total number of observations

2.3.3. MASE

Mean Absolute Scaled Error (MASE) is a scale-independent metric proposed by [42] for evaluating forecast accuracy. It is calculated by dividing the MAE of the forecasting model by the MAE of a naïve one-step lag model.

$$MASE = \frac{\frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|} \tag{16}$$

In this formula, the numerator represents the Mean Absolute Error (MAE) of the proposed forecasting model, reflecting the average absolute difference between the predicted and observed values. The denominator corresponds to the MAE of the naïve one-step lag model, which assumes that the forecast at time t is equal to the observed value at time t-1 i.e., $\widehat{y}_t = y_{t-1}$. This naïve model serves as a basic benchmark, and by scaling the model's MAE against it, MASE provides a standardized measure of forecast accuracy. A MASE value less than one indicates that the model outperforms the naïve approach, while a value greater than one suggests inferior performance.

2.4. Diebold–Mariano test

The Diebold–Mariano (DM) test is used to determine whether the two forecasts are significantly different or not [43]. Let e_i and r_i be the residuals for the two forecasts;

$$e_i = y_i - f_i \tag{17}$$

$$r_i = y_i - g_i \tag{18}$$

Let d_i be defined as $d_i = e_i^2 - r_i^2$ or $d_i = |e_i| - |r_i|$ (19)

The time series d_i is called the loss-differential**.** Clearly, the first of these formulas are related to the MSE metric and the second is related to the MAE metric. We now define:

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i,\ \ \mu = E[d_i] \tag{20}$$

$$\text{For } n > k \geq 1,\ \text{define } \widehat{\gamma}_k = \frac{1}{n}\sum_{i=k+1}^{n}(d_i - \overline{d})(d_{i-k} - \overline{d}) \tag{21}$$

γ_k is the autocovariance at lag k. For $h \geq 1$, Diebold-Mariano statistic is defined as follows:

$$DM = \frac{\overline{d}}{\sqrt{\left[\widehat{\gamma}_0 + 2\sum_{k=1}^{h-1}\gamma_k\right]/n}} \tag{22}$$

It is generally sufficient to use the value $h = n^{1/3} + 1$. Under the assumption that $\mu = 0$ (the null hypothesis), DM follows a standard normal distribution i.e., $DM \sim N(0, 1)$. Thus, there is a significant difference between the forecasts if, $DM \geq Z_{crit}$, where Z_{crit} is the two-tailed critical value for the standard normal distribution. The key assumption for using the DM test is that the loss differential time series d_i is stationary.

3. Results

The time series plots of weekly counts of YSB light trap catches of five study sites during the study period were plotted in Fig. SF1. The time series plots showed that at all examined locations, the YSB incidence was higher between the 35th to 45th standard meteorological weeks (SMWs), except at the Pattambi and Chinsurah centres, where it showed between the 20th to 30th SMWs. At the Rajendra Nagar centre, it showed two peaks, between 1st to 10th and the 35th to 45th SMWs. Summary statistics of the yellow stem borer population and exogenous weather variables were presented in supplementary Tables ST1 and ST2, respectively. In most of the centres, the YSB count and the weather variables were highly skewed and leptokurtic in nature. The coefficient of variation of YSB were of highly heterogeneous in nature.

The Pearson correlation coefficients between YSB populations and the weather variables are presented in Supplementary Table ST3. YSB population was having a significantly low positive correlation with that of sunshine hours (SSH) at the Warangal, Rajendra Nagar, and Pattambi centres. Similarly, a low significant correlation was found between YSB populations and maximum temperature (MAXT) at the Raipur centre. Additionally, a low significant correlation was observed between YSB populations and MAXT, minimum temperature (MINT), and evening relative humidity (ERH) at other centres. However, at Warangal, the correlation between YSB populations and both MAXT and MINT was significantly negative. In Raipur, the correlation between YSB populations and ERH as well as rainfall (RF) was also significantly negative. At the Pattambi centre, a significant negative correlation was found between YSB populations and both MINT and RF. Supplementary Table ST3 provides a clear representation of the relationships between YSB populations and the various weather variables.

To identify the climatological factors influencing the incidence of the YSB population, a stepwise regression analysis was conducted, and the results are shown in Supplementary Table ST4. The MINT, SSH, and RF at Warangal; SSH and RF at Rajendra Nagar; RF, MINT, MAXT, and ERH at Pattambi; ERH at Raipur; and MAXT, MINT, and RF at Chinsurah were found to significantly contribute to the YSB population. However, the R^2 values for the fitted regression models at all five centres were low, indicating a poor fit. This may be attributed to the presence of non-linear and highly heterogeneous relationships among the variables.

3.1. Results of INGARCH models

Box-Pierce non-correlation test indicated that the data under consideration were auto-correlated ($p<0.0001$) in nature. As a next step, the INGARCH model with exogenous climatological variables were fitted and the model summaries for the five centres for data set 1 were presented in Table 1 (for data set 2, refer to Supplementary Table ST5). Even though the coefficients of the lagged observations were found to be significant, effect of no climatological parameters were significant. Moreover, diagnostic checking of residuals by the Box-Pierce non-correlation test revealed that the residuals were auto-correlated ($p< 0.0001$) in nature.

Table 1
Parameter estimates of INGARCH models for the study centres for data set 1.

	Parameters	Estimate	Std. error	Z value	Pr(>|z|)	Box-Pierce non-correlation test for residuals
Warangal	Intercept	5.00	14.1	0.38	0.70	
	beta_4	0.5	0.11	4.62	<0.001	
	alpha_4	0.2	0.12	2.02	0.04	
	MAXT	<0.0001	0.42	0.00	0.99	χ^2 = 132.07
	MINT	<0.0001	0.50	0.00	1.00	(p< 0.0001)
	RF	0.06	0.08	0.68	0.49	
	MRH	0.0001	0.19	0.00	0.99	
	ERH	0.02	0.18	0.10	0.91	
	SSH	0.12	0.08	1.42	0.15	
Rajendra Nagar	Intercept	1.08	0.2	0.49	0.62	
	beta_2	0.13	0.96	1.37	0.16	
	alpha_2	0.46	0.34	1.36	0.17	χ^2 = 0.003 (p = 0.92)
	MAXT	<0.0001	0.67	0.005	1	
	MINT	<0.0001	0.78	0.002	1	
	RF	0.10	0.15	0.63	0.52	
	MRH	0.64	0.25	0.03	0.97	
	ERH	<0.0001	0.21	0.00	1	
	SSH	0.02	0.08	0.29	0.77	
Pattambi	Intercept	205	653	0.314	0.753	
	beta_2	0.51	0.188	2.731	0.006	
	alpha_2	<0.0001	0.175	0.051	0.959	χ^2 = 119.6 (p<0.001)
	MAXT	<0.0001	0.113	0.000	1	
	MINT	0.198	8.26	0.002	0.998	
	RF	0.172	0.46	0.370	0.711	
	MRH	<0.0001	5.7	0.001	0.999	
	ERH	0.064	2.47	0.026	0.979	
	SSH	0.001	0.008	0.23	0.63	
Raipur	Intercept	1.83	2.31	0.79	0.42	
	beta_2	0.29	0.67	4.36	<0.001	
	alpha_2	0.65	0.8	8.13	<0.001	
	MAXT	<0.0001	0.2	0.005	1	χ^2 = 126.18 (p<0.001)
	MINT	<0.0001	0.27	0.003	1	
	RF	<0.0001	0.56	0.01	0.99	
	MRH	<0.0001	0.6	0.002	0.99	
	ERH	0.23	0.8	0.28	0.77	
Chinsurah	Intercept	<0.0001	2.66	0.00	1.00	
	beta_2	0.43	0.13	3.23	0.001	
	alpha_2	0.18	0.14	1.28	0.20	
	MAXT	0.2	1.9	0.11	0.92	χ^2 = 150 (p<0.001)
	MINT	2.7	1.82	1.51	0.13	
	RF	<0.0001	0.27	0.007	1.00	
	MRH	<0.0001	0.22	0.005	1.03	
	ERH	<0.0001	0.54	0.001	0.3	

3.2. Results of ZIPAR models

The ZIPAR models with 5 lags were chosen based on the lowest AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values. The parameter estimates of ZIPAR models for all five centres for data set 1were represented in Table 2 (for data set 2, refer to Supplementary Table ST6). Most of the parameter estimates are non-significant and residuals shows significance as the probability of Box-Pierce non-correlation was less than 0.05, indicates residuals under consideration are significant for all the centres except for Rajendra Nagar where residual probabilities are non-significant as probability is more than 0.05 for both data set 1 and data set 2.

3.3. Results of ZINBAR models

The ZINBAR models with 5 lags were chosen on the basis of the lowest AIC and BIC values. The parameter estimates of the ZINBAR models for all the five centres for data set 1 were represented in Table 3 (for data set 2, refer to Supplementary Table ST7). Majority of the parameter estimates are non-significant, while the residuals are significant, as indicated by the Box-Pierce test, with non-correlation probabilities below 0.05. This means the residuals are significant for all centers except Rajendra Nagar, where the residual probabilities exceed 0.05, making them non-significant for both data set 1 and data set 2.

3.4. Results of ANN models

The sigmoidal and linear activation functions were used in the input to hidden layer and in hidden to output layer, respectively. The weather variables namely MAXT, MINT, MRH, ERH, SSH and rainfall, were also used as exogenous variables in input layer. The suitable candidate models were chosen based on lowest MSE and RMSE values. Table 4 showed the specifications of the selected ANN models (for data set 2, refer to Supplementary Table ST8). After fitting of the models, the diagnostic checking of the residuals was carried out by Box- Pierce non-correlation test and found that the residuals were non-correlated in nature.

3.5. Two stage modelling and forecasting

The first step in the two stage model building process was to obtain the predicted and the residual values from the INGARCH, ZIPAR and ZINBAR models. In the next step, the presence of autocorrelation in the residuals of fitted models was tested along with nonlinearity. After confirmation of the presence of autocorrelation and non-linear structure in the residual series of all the study locations except Rajendra Nagar,

Table 2
Parameter estimates of the ZIPAR models for the study centers for data set 1.

	Parameter	P (Y>0)				P(Y=0)				Box-Pierce non correlation test for residuals
		Estimate	Std. Error	Z value	Pr(>|z|)	Estimate	Std. Error	Z value	Pr(>|z|)	
Warangal	Intercept	2.70	0.01	163.64	<0.0001	0.33	0.22	1.50	0.134	χ^2 = 9.08
	ysb_lag1	0.01	0.0002	42.71	<0.0001	-0.60	0.11	-5.32	<0.0001	p = 0.03
	ysb_lag2	0.002	0.0003	7.57	<0.0001	-0.01	0.03	-0.43	0.664	
	ysb_lag3	0.003	0.0003	8.45	<0.0001	0.01	0.03	0.40	0.693	
	ysb_lag4	0.002	0.0003	7.64	<0.0001	0.003	0.02	-0.17	0.862	
	ysb_lag5	0.001	0.0003	2.57	0.01	0.01	0.01	0.54	0.591	
Rajendra Nagar	Intercept	3.01	0.01	210.84	< 0.001	0.65	0.18	3.63	0.0002	χ^2 = 0.002
	ysb_lag1	0.004	<0.001	54.82	< 0.001	-0.37	0.06	-5.79	<0.001	p = 0.91
	ysb_lag2	0.002	<0.001	23.28	< 0.001	0.05	0.02	2.83	0.004	
	ysb_lag3	<0.001	<0.001	2.84	0.004	-0.08	0.03	-2.65	0.008	
	ysb_lag4	0.001	<0.001	12.66	< 0.001	0.02	0.01	2.05	0.04	
	ysb_lag5	0.0017	<0.001	16.75	< 0.001	-0.01	0.01	-1.45	0.14	
Pattambi	Intercept	4.65	0.066	699.1	<0.001	-1.45	0.001	-8.74	<0.001	χ^2 = 9.4
	ysb_lag1	<0.0001	<0.0001	510.46	<0.001	-0.001	0.001	-0.8	0.42	p= 0.02
	ysb_lag2	<0.0001	<0.0001	193.87	<0.001	-0.004	0.003	-1.19	0.23	
	ysb_lag3	<0.0001	<0.0001	-65.25	<0.001	0.002	0.002	1.11	0.26	
	ysb_lag4	<0.0001	<0.0001	150.27	<0.001	-0.007	0.003	-0.02	0.97	
	ysb_lag5	<0.0001	<0.0001	89.16	<0.001	-0.004	0.003	-1.44	0.15	
Raipur	Intercept	4.65	0.066	699.1	<0.001	-1.45	0.001	-8.74	<0.001	χ^2 = 1.54
	ysb_lag1	<0.0001	<0.0001	510.46	<0.001	-0.001	0.001	-0.8	0.42	p = 0.21
	ysb_lag2	<0.0001	<0.0001	193.87	<0.001	-0.004	0.003	-1.19	0.23	
	ysb_lag3	<0.0001	<0.0001	-65.25	<0.001	0.002	0.002	1.11	0.26	
	ysb_lag4	<0.0001	<0.0001	150.27	<0.001	-0.007	0.003	-0.02	0.97	
	ysb_lag5	<0.0001	<0.0001	89.16	<0.001	-0.004	0.003	-1.44	0.15	
Chinsurah	Intercept	4.5	0.52	869.4	<0.001	-2.416	0.534	-4.53	<0.001	χ^2= 4.6
	ysb_lag1	0.13	<0.0001	270.4	<0.001	-0.090	0.031	-2.92	0.003	p = 0.03
	ysb_lag2	1.57	<0.0001	20.9	<0.001	0.020	0.007	2.79	0.005	
	ysb_lag3	<0.0001	<0.0001	3.4	0.0597	-0.003	0.003	-0.94	0.34	
	ysb_lag4	<0.0001	<0.0001	12.3	<0.001	0.004	0.002	1.68	0.092	
	ysb_lag5	<0.0001	<0.0001	40.5	<0.001	0.002	0.002	1.11	0.26	

the ANN model was used for modelling and forecasting of the INGARCH, ZIPAR and ZINBAR residuals for the rest four centres. The residuals predicted from the ANN were then combined with the predicted values obtained from the INGARCH, ZIPAR, and ZINBAR models. The results of the BDS test of the fitted model residuals for datasets 1 and 2 are presented in Supplementary Tables ST9 and ST10, respectively

3.5.1. *Results of INGARCH-ANN models*

After validating autocorrelation by Box-Pierce test (Table 1 and Supplementary Table ST5) and nonlinearity by the BDS test (Supplementary Table ST9 and ST10), the same residuals were modelled using ANN model along with exogenous weather varaibles. Further, the predicted residuals were combined with the forecasts obtained from original INGARCH model. This modeling procedure is called as INGARCH-ANN two stage count time series methodology. Supplementary Tables ST11 and ST12, showed the specifications of the selected ANN models for INGARCH residuals for data set 1 and 2, respectively. After fitting of the model, the diagnostic checking of the residuals by Box-Pierce test showed that residuals were non-correlated in nature.

3.5.2. *Results of ZIPAR-ANN models*

Similar to the INGARCH-ANN two stage count time series methodology, the significant residuals of ZIPAR (Table 2 and Supplementary Table ST6) were modelled using ANN model along with exogenous weather varaibles as the residuals were nonlinear (Supplementary Table ST9 and ST10). Further, the forecasted residuals were combined with the forecasts obtained from original ZIPAR model. This modeling procedure is called as ZIPAR-ANN two stage count time series methodology. Supplementary Tables ST13 and ST14, showed the specifications of the selected ANN models for ZIPAR residuals for data set 1 and 2, respectively. Diagnostic checking of the residuals of ZIPAR-ANN methodology were also found to be non-correlated.

3.5.3. *Results of ZINBAR-ANN models*

Similarly, in ZINBAR-ANN two stage count time series methodology, once the autocorrelation (Table 3 and Supplementary Table ST8) and nonlinearity of ZINBAR residuals (Supplementary Tables ST9 and ST10) were confirmed, the residuals were modelled using ANN model along with exogenous weather varaibles. Further the fitted residuals were combined with the predicted values obtained from original ZINBAR model. Supplementary Tables ST15 and ST16, showed the specifications of the selected ANN models for ZINBAR residuals for data set 1 and 2, respectively. The diagnostic checking of the residuals also confirmed the non-existence of autocorrelation structure.

4. Discussion

Prior to developing the proposed forewarning models, the cause-and-effect relationship between Yellow Stem Borer (YSB) populations and weather variables were analysed. Stepwise regression analysis revealed that minimum temperature, maximum temperature, rainfall, sunshine hours, and morning and evening relative humidity significantly influence the occurrence of YSB populations. It was observed that different weather variables have varying effects on the YSB population across different centres. A detailed estimation of the parameters and their significance is provided in Supplementary Tables ST4. These findings align with the results reported by [44].

The comparative assessment of different models employed models in terms of MSE and RMSE values for data set 1 were presented in Table 5 for both training and testing sets (for data set 2, refer to Supplementary Tables ST17). In the INGARCH and ANN models, weather variables were directly incorporated as exogenous variables. For ZIPAR and ZINBAR models, weather data were also treated as exogenous variables, but with an additional step; significant residuals of the ZIPAR and ZINBAR models were fitted as exogenous variables. When building the ANN models for these residuals, the weather variables were used as

Table 3
Parameter estimates of the ZINBAR models for the study centers for data set 1.

	Parameter	P (Y>0)				P(Y=0)				Box-Pierce non correlation test for residuals
		Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)	
Warangal	Intercept	2.26	0.09	26.02	<0.0001	0.90	0.32	2.83	<0.0001	$\chi^2 = 8.08$
	ysb_lag1	0.02	0.002	7.72	<0.0001	-0.83	0.30	-2.75	0.01	p = 0.05
	ysb_lag2	0.003	0.002	1.41	0.159	-0.71	0.42	-1.67	0.10	
	ysb_lag3	0.003	0.002	1.19	0.236	0.00	0.06	-0.07	0.95	
	ysb_lag4	0.002	0.002	0.75	0.452	0.07	0.08	0.92	0.36	
	ysb_lag5	0.002	0.002	1.23	0.22	-0.02	0.03	-0.80	0.43	
	Log(theta)	0.053	0.018	2.94	0.003					
Rajendra Nagar	Intercept	2.34	0.09	23.6	< 0.001	1.4	0.27	5.03	<0.001	$\chi^2 = 0.085$
	ysb_lag1	0.02	0.2	8.11	<0.001	-1.07	0.36	-2.90	0.003	p= 0.92
	ysb_lag2	0.001	0.02	1.09	0.27	-0.46	0.20	-2.27	0.02	
	ysb_lag3	0.0014	0.01	0.09	0.92	-0.03	0.07	-0.5	0.61	
	ysb_lag4	0.0015	0.002	0.69	0.49	-0.05	0.06	-0.82	0.41	
	ysb_lag5	0.0012	0.001	1.33	0.18	-0.009	0.03	-0.25	0.79	
	Log(theta)	2.34	0.09	23.61	< 0.001	1.4	0.27	5.03	<0.001	
Pattambi	(Intercept)	2.9	0.096	40.4	<0.001	-0.10	0.40	-0.25	0.8	$\chi^2 = 9.07$
	ysb_lag1	0.001	<0.0001	7.62	<0.001	-1.48	0.52	-2.87	0.004	p = 0.02
	ysb_lag2	<0.0001	<0.0001	1.65	0.09	0.002	0.03	0.07	0.94	
	ysb_lag3	<0.0001	<0.0001	-0.17	0.86	0.007	0.07	0.11	0.91	
	ysb_lag4	<0.0001	<0.0001	2.24	0.02	-0.001	0.06	-0.01	0.98	
	ysb_lag5	<0.0001	<0.0001	0.5	0.57	-0.041	0.06	-0.64	0.52	
	Log(theta)	-0.9	0.64	-13.6	<0.001					
Raipur	Intercept	2.5	0.071	35.20	<0.001	-7.7	0.42	-4.27	<0.001	$\chi^2 = 3.91$
	ysb_lag1	0.015	0.0023	10.17	<0.001	-34.7	0.032	-1.07	1.03	p = 0.04
	ysb_lag2	0.001	0.002	0.63	0.53	1.4	0.045	2.04	0.56	
	ysb_lag3	0.0015	0.001	0.90	0.36	-1.0	0.02	-0.49	0.45	
	ysb_lag4	0.0013	0.002	0.38	0.7	2.5	0.01	0.35	1.05	
	ysb_lag5	0.003	0.001	2.34	0.01	-1.3	0.02	-0.05	0.24	
	Log(theta)	0.402	0.07	5.71	<0.001					
Chinsurah	(Intercept)	4.02	0.06	61.95	<0.001	-9.48	141.30	-0.07	0.94	$\chi^2 = 4.92$
	ysb_lag1	0.003	0.0004	9.98	<0.001	-4.81	143.25	-0.03	0.97	p = 0.03
	ysb_lag2	-0.0007	0.0003	-2.25	0.024	0.11	51.13	0.00	0.998	
	ysb_lag3	0.0001	0.0003	0.44	0.66	-0.12	7.07	-0.02	0.98	
	ysb_lag4	0.0005	0.0004	1.49	0.13	0.32	17.33	0.005	0.95	
	ysb_lag5	0.0007	0.0003	2.47	0.014	0.01	0.97	0.01	0.99	
	Log(theta)	0.32	0.06	4.96	<0.001					

Table 4
ANN model specifications for the study centres for data set 1.

Specifications	Warangal	Rajendra Nagar	Pattambi	Raipur	Chinsurah
Input lag	4	5	4	1	1
Output variable/ dependent variable	1	1	1	1	1
Hidden nodes	5	10	5	5	5
Hidden layers	1	1	1	1	1
Exogenous variables	6	6	6	5	5
Model	4:5S:1L	5:10S:1L	3:6S:1L	1:5S:1L	1:5S:1L
Total number of parameters	61	131	61	41	41
Network type	Feed Forward	Feed Forward	Feed Forward	Feed Forward	Feed Forward
Activation function (I: H)	Sigmoidal	Sigmoidal	Sigmoidal	Sigmoidal	Sigmoidal
Activation function(H: O)	Identity	Identity	Identity	Identity	Identity
Box-Pierce non-correlation test for residuals	χ^2 = 0.03 p = 0.85	χ^2 = 0.08 (p = 0.07)	χ^2 = 0.02 (p = 0.871)	χ^2 = 0.26 (p = 0.60)	χ^2 = 3.25 (p = 0.07)

independent variables in the input layer, while the residuals were considered as the dependent variables. This adjustment was made to simplify the model-building process by indirectly accounting for weather information in the modelling phase.

Among the models studied, the ZINBAR-ANN model outperformed the INGARCH, ZIPAR, ZINBAR, INGARCH-ANN, and ZIPAR-ANN models in both training and testing datasets, as indicated by its lowest MSE and RMSE values in both data set 1 and 2. The performance hierarchy of these models can be represented as follows: ZINBAR-ANN > ZIPAR-ANN > INGARCH-ANN > ANN > ZINBAR > ZIPAR > INGARCH, across all study locations except the Rajendra Nagar centre in training and testing population in both the data sets. For the Rajendra Nagar centre, hybrid models were not developed because the residuals of the INGARCH, ZIPAR, and ZINBAR models were non-significant. The performance hierarchy of the models for YSB count data at Rajendra Nagar was as follows; ANN > ZINBAR > ZIPAR > INGARCH for both the data sets. Centre-wise forecasts of different models for both data set 1 and 2 were provided in Supplementary Tables ST18-ST27, respectively. Actual vs. fitted plots for Data Set 1 are shown in Figs. 4a–4e, whereas for Data Set 2, the actual vs. fitted plots are provided in Supplementary Figure SF2.

The error reduction achieved by the proposed ZINBAR-ANN model compared to classical models is shown in supplementary Tables ST29-ST33 for both training and testing splits of the datasets. The ZINBAR-ANN model consistently outperforms both the INGARCH and ZINBAR models, with significant error reductions across various datasets and locations. In Dataset 1, for Warangal, it reduces errors by 68.3 % during training and 60.7 % during testing compared to the ZINBAR model, and by 73.7 % and 65.7 % compared to the INGARCH model. Similar

Table 5
Comparison criteria for different models for YSB populations in training and testing sets for data set 1.

	Data Split	Criteria	INGARCH	ZIPAR	ZINBAR	ANN	INGARCH-ANN	ZIPAR-ANN	ZINBAR-ANN
Warangal	Training Set	MSE	1090.92	890.23	756.96	339.69	152.96	112.79	75.75
		RMSE	33.03	29.80	27.51	18.43	12.37	10.62	8.70
		MAE	23.02	18.81	16.38	10.82	7.12	5.43	4.06
		MASE	1.56	1.29	1.12	0.74	0.49	0.38	0.28
	Testing Set	MSE	1103.17	1057.40	839.36	429.75	280.86	218.11	129.97
		RMSE	33.21	32.52	28.97	20.73	16.76	14.77	11.40
		MAE	29.43	30.86	26.14	17.00	13.86	11.57	8.00
		MASE	1.40	1.47	1.24	0.81	0.66	0.55	0.38
Rajendra Nagar	Training Set	MSE	4205.90	3469.80	2117.30	340.50			
		RMSE	64.85	58.90	46.01	18.45			
		MAE	33.86	25.80	19.03	10.17			
		MASE	1.39	1.10	0.81	0.43			
	Testing Set	MSE	7364.73	5903.67	2220.57	1102.14			
		RMSE	85.82	76.84	47.12	33.20			
		MAE	73.29	65.57	46.00	22.14			
		MASE	2.36	2.12	1.48	0.71			
Pattambi	Training Set	MSE	248,930.42	140,362.22	98,544.57	72,631.99	36,138.79	20,735.13	15,915.50
		RMSE	498.93	374.65	313.92	269.50	190.10	144.00	126.16
		MAE	219.20	143.23	116.19	103.58	71.71	43.64	41.45
		MASE	1.50	0.99	0.80	0.72	0.50	0.30	0.28
	Testing Set	MSE	299,906.78	194,479.43	171,990.57	119,360.71	59,570.43	40,866.29	19,093.60
		RMSE	547.64	441.00	414.72	345.49	244.07	202.15	138.18
		MAE	470.29	372.86	346.29	303.00	213.57	196.29	123.29
		MASE	0.91	0.72	0.67	0.59	0.41	0.38	0.24
Raipur	Training Set	MSE	885.20	634.56	501.23	310.67	272.52	86.28	56.80
		RMSE	29.75	25.19	22.39	17.63	16.51	9.29	7.54
		MAE	18.57	15.42	14.05	11.51	9.72	5.82	4.66
		MASE	0.74	0.63	0.58	0.45	0.40	0.24	0.19
	Testing Set	MSE	1569.00	1108.14	891.14	553.57	402.71	308.57	108.29
		RMSE	39.61	33.29	29.85	23.53	20.07	17.57	10.41
		MAE	35.57	30.43	26.86	15.57	17.86	14.57	8.29
		MASE	0.50	0.43	0.38	0.22	0.25	0.21	0.12
Chinsurah	Training Set	MSE	2380.55	2065.50	1810.82	1415.25	810.03	585.32	379.02
		RMSE	48.79	45.45	42.55	37.62	28.46	24.20	19.47
		MAE	9.89	17.60	15.59	14.34	9.89	8.88	8.19
		MASE	0.13	0.24	0.21	0.20	0.13	0.12	0.11
	Testing Set	MSE	2999.18	2308.33	1946.41	1627.14	1418.00	1024.40	554.79
		RMSE	54.76	48.05	44.12	40.34	37.70	32.00	23.55
		MAE	52.86	44.43	39.71	36.00	33.71	29.14	20.57
		MASE	1.97	1.66	1.48	1.34	1.26	1.09	0.77

improvements are seen in Pattambi, Raipur, and Chinsurah, with reductions ranging from 46.9 % to 75.8 %. For Dataset 2, the ZINBAR-ANN model achieves error reductions of up to 75.6 % in training and 67.8 % in testing compared to the INGARCH model, with consistent results across locations. At the Rajendra Nagar Centre, the ANN model also showed substantial error reductions compared to the INGARCH and ZIPAR models, especially in Dataset 1, where it reduced errors by 71.1 % in training and 60.0 % in testing. However, the ANN model showed smaller improvements compared to the ZINBAR model, particularly during testing, with reductions of 23.0 % for Dataset 1 and 29.5 % for Dataset 2. These results emphasize the strong performance of the proposed two stage ZINBAR-ANN model, especially compared to classical models.

Along with the MSE and RMSE, the study also incorporates two additional measures, MAE and MASE. The MAE provides a direct measure of the average magnitude of forecasting errors, expressed in the same units as the original data, making it easy to interpret and compare across models. In the present analysis, the proposed ZINBAR-ANN hybrid model consistently provided the lowest MAE values across all locations and data splits, both in training and testing phases. For example, in the testing set for Warangal, the MAE drops significantly from 29.43 (INGARCH) to 8.00 (ZINBAR-ANN), while in Pattambi, the reduction is even more substantial from 470.29 to 123.29 for data set 1. Similar trends are observed in other locations such as Raipur and Chinsurah, demonstrating the hybrid model's effectiveness in reducing forecasting error. The same pattern holds for dataset 2 as well.

The MASE metric further enhances interpretability by scaling the error against that of a naïve forecast model, which assumes that the forecast at time t equals the observed value at time t-1. A MASE value below 1 indicates that the proposed model outperforms the naïve benchmark. Across all study locations, the ZINBAR-ANN model consistently records MASE values well below 1, mostly below 0.5, highlighting its robust predictive performance. For instance, in Warangal (testing), the MASE decreases from 1.40 (INGARCH) to 0.38 (ZINBAR-ANN); in Raipur, it improves from 0.50 to 0.12; and in Pattambi, from 0.91 to 0.24 for data set 1. Similar results were observed for dataset 2. These results confirm that the ZINBAR-ANN model not only reduces absolute forecasting error but also generalizes better across different datasets and locations. Hence, the inclusion of MAE and MASE provides a more comprehensive and scalable evaluation, reinforcing the superiority of the hybrid model in predicting complex zero-inflated, over dispersed pest count data.

However, these comparison criteria only show the differences between the observed and predicted values of the models. Therefore, the Diebold–Mariano (DM) test was employed to assess the statistical significance of the differences in model performance. While absolute metrics quantify the magnitude of forecast errors, they do not indicate whether one model performs significantly better than another in a statistical sense. The DM test addresses this limitation by comparing the forecast errors of two models over time, testing the null hypothesis that both models have equal predictive accuracy. The results for both Dataset 1 and Dataset 2 (Supplementary Tables ST34 and ST35) clearly

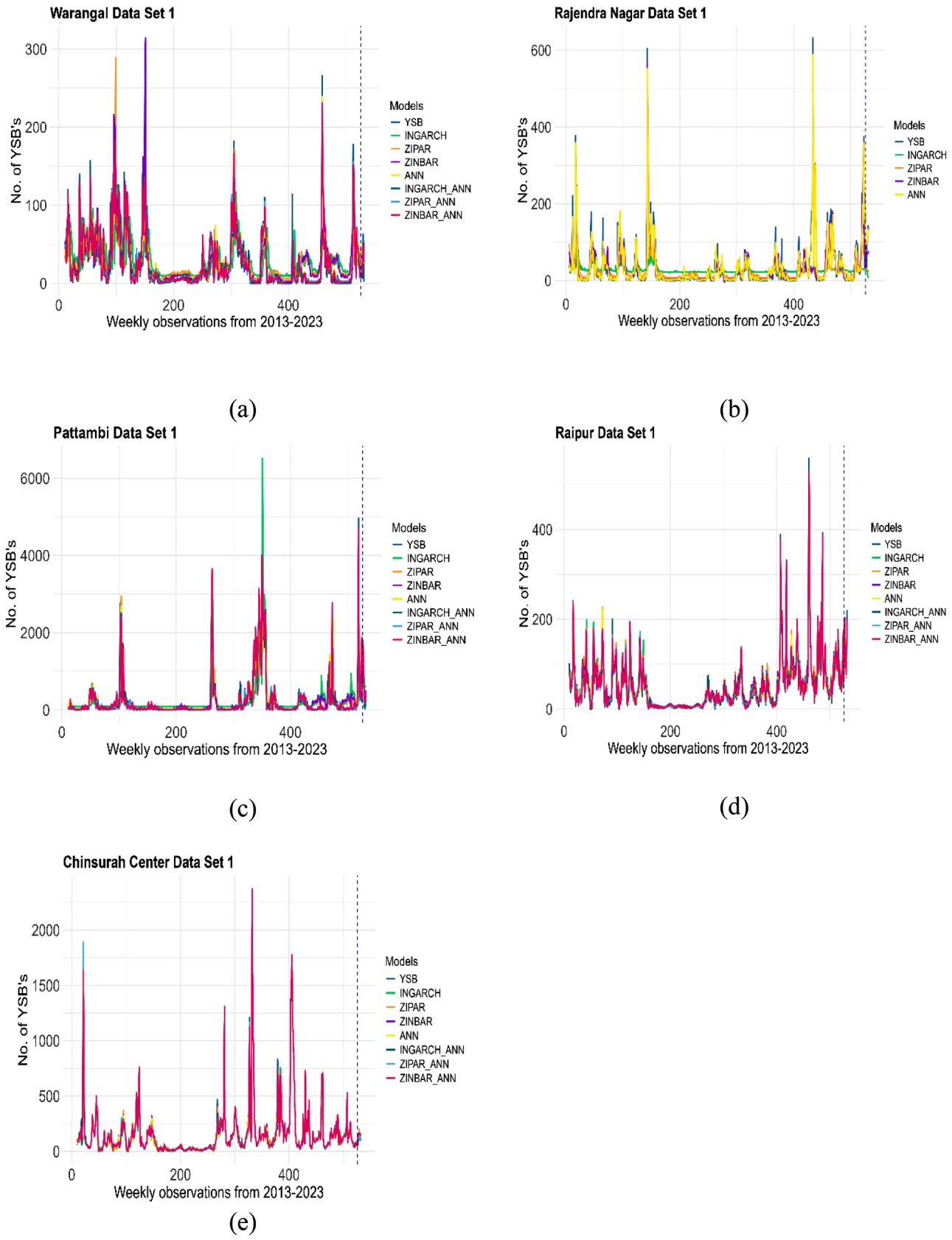

Fig. 4. Actual vs. fitted plots of YSB populations for data set 1 depicting a) Warangal, b) Rajendra Nagar, c) Pattambi, d) Raipur and e) Chinsurah.

demonstrate that the ZINBAR-ANN hybrid model (M7) significantly outperforms not only the individual models (INGARCH, ZIPAR, ZINBAR, ANN) but also the other hybrid models (INGARCH-ANN, ZIPAR-ANN) across almost all centres and model pairings. For instance, in Warangal (Dataset 1), the DM statistic for M1 vs. M7 is 7.56 ($p < 0.001$), and in Raipur, it is 5.24 ($p < 0.01$), indicating strong evidence against the null hypothesis. Similarly, in Dataset 2, the M1 vs. M7 comparisons show highly significant values across centres, such as 6.62 in Warangal and 5.79 in Raipur (both $p < 0.0001$). Even when comparing M6 (ZIPAR-ANN) with M7 (ZINBAR-ANN), significant DM statistics are observed (for example, 3.65 in Pattambi and 3.84 in Raipur), indicating that ZINBAR-ANN consistently provides superior forecasts even among hybrid competitors. An exception is seen in Rajendra Nagar, where the ANN model (M4) occasionally outperforms others, aligning with earlier findings. Overall, the consistently significant DM test results underscore the robustness and statistical superiority of the proposed ZINBAR-ANN model in forecasting complex, overdispersed, and nonlinear count time series data.

Several studies have reported that hybrid two-stage models and machine learning models perform better than classical models. However, these studies are not specifically related to count time series models. For instance, [35] modeled and forecasted rice gall midge populations using count and machine learning models and found that machine learning approaches outperformed the classical count time series INGARCH model. Similarly, hybrid models demonstrated superior performance in forecasting rainfall [45], predicting rice yield [46], and forecasting temperature [47,48]. In addition, two-stage hybrid models have shown effectiveness in credit risk assessment [49], short-term wind direction forecasting [50], and measuring the sustainable performance of the Indian retail chain. Furthermore, diagnostic checks on the residuals obtained from the INGARCH, ZIPAR, ZINBAR, ANN, INGARCH-ANN, and ZIPAR-ANN models showed that they were non-autocorrelated and non-random, indicating that the models under consideration were adequate. The proposed two stage ZINBAR-ANN approach was effective in modeling over-dispersed count data, addressing challenges like excess zeros, which commonly observed in light trap count data and results in over dispersion, which simpler models like Poisson often struggle to handle. The ANN component excels at capturing non-linear relationships, learning from the residual patterns left after fitting the ZINBAR model, allowing it to detect more complex and heterogeneous patterns that a simple count model may overlook.

YSB is a serious insect pest of rice inflicting significant crop loss across the Indian sub- continent. It is a monophagous insect with an interesting reproductive biology. The female moths start laying eggs next day of its first appearance in the season, for 1–2 days. The eggs hatch in 7–8 days [51,52] and the neonate larvae enter inside the rice plant within 35 min and remain inside until the adult emergence. This very behaviour makes YSB management challenging as it is protected from the natural enemies and to chemical insecticides. Hence, the tactics aimed at YSB management, especially the insecticides need to be delivered during this narrow window of about 10-days, targeting the egg stage and neonate larvae and thus making the 'timing' a critical factor. In this scenario, the two-stage ZINBAR-ANN model because of its high precision, as evidenced in this study will be of immense value in forecasting YSB population based on the weather data. An efficient forecasting model of YSB population in rice crop ecosystem will be a precursor for developing an area wide forewarning system to disseminate advisories for the timely application of management tactics to avoid yield losses by the YSB damage in rice crop.

5. Conclusion

In the current study, two-stage zero-inflated count time series models were developed to predict the occurrence of the Yellow Stem Borer (YSB) population using weather variables. Weather variables such as maximum and minimum temperature, morning and evening relative humidity, rainfall, and sunshine hours were found to affect the occurrence of the rice YSB population. The classical count time series INGARCH model fails to perform well when a significant proportion of zeros occur in the population due to the non-occurrence of YSB in certain Standard Meteorological Weeks (SMWs). In such cases, zero-inflated models were found to be suitable alternatives.

The traditional count time series models, namely INGARCH, ZIPAR, and ZINBAR, exhibited significant residuals after model fitting, indicating that these models did not provide a satisfactory fit. To improve forecasting accuracy, two-stage models were developed, where the significant residuals were fitted using an Artificial Neural Network (ANN) model. The proposed zero-inflated methodology was applied to both kharif (dataset 1) and rabi (dataset 2) seasons, and it outperformed classical models in both training and validation datasets. Additionally, the percentage error reduction achieved by the ZINBAR-ANN strategy, compared to classical and other models, indicated a substantial improvement in forecasting accuracy. The methodology proposed in this study aids in predicting the YSB population, enabling the proactive implementation of preventive and curative pest management strategies. This approach can, in turn, assist in planning strategies to reduce rice yield loss due to the YSB pest. In the future, prediction models for different crop pests can be developed using the proposed models and tested with various combinations of count time series and machine learning models.

Ethical statement

This statement is to certify that all authors have seen and approved the manuscript being submitted, have contributed significantly to the work, attest to the validity and legitimacy of the data and its interpretation, and agree to its submission to the Smart Agricultural Technology.

We attest that the article is the Authors' original work, has not received prior publication and is not under consideration for publication elsewhere. We adhere to the statement of ethical publishing as appears in the Smart Agricultural Technology.

On behalf of all Co-Authors, the corresponding Author shall bear full responsibility for the submission. Any changes to the list of authors, including changes in order, additions or removals will require the submission of a new author agreement form approved and signed by all the original and added submitting authors.

All authors are requested to disclose any actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations within three years of beginning the submitted work that could inappropriately influence, or be perceived to influence, their work. If there are no conflicts of interest, the COI should read: "The authors report no relationships that could be construed as a conflict of interest".

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

CRediT authorship contribution statement

Bojjireddygari Nanda Kumar Reddy: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Santosha Rathod:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yerram Sridhar:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Investigation, Data curation, Conceptualization. **Supriya Kallakuri:** Writing – review & editing, Methodology, Conceptualization. **Pramit Pandit:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis. **Bellamkonda Jyostna:** Writing – review

& editing, Software, Methodology, Formal analysis. **Seetalam Malathi:** Writing – review & editing, Validation, Data curation. **R Shravan Kumar:** Writing – review & editing, Validation, Data curation. **Sanjay Sharma:** Writing – review & editing, Validation, Data curation. **K Karthikeyan:** Writing – review & editing, Validation, Data curation. **NRG Varma:** Writing – review & editing, Validation, Data curation. **Sitesh Chatterjee:** Writing – review & editing, Validation, Data curation. **Ayyagari Phani Padmakumari:** Writing – review & editing, Validation, Resources, Data curation. **Nethi Somasekhar:** Writing – review & editing, Validation, Data curation. **Sailaja Banda:** Writing – review & editing, Software, Methodology. **Chintalapati Padmavathi:** Writing – review & editing, Validation, Data curation. **Chitra Shanker:** Writing – review & editing, Validation, Data curation. **Ponnuraj Jeyakumar:** Writing – review & editing, Visualization, Validation. **Raman Meenakshi Sundaram:** Writing – review & editing, Visualization, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Professor Jayashankar Telangana Agricultural University, Hyderabad; All India Coordinated Research Project on Rice(AICRPR) and Indian Council of Agricultural Research-Indian Institute of Rice Research (ICAR-IIRR), Hyderabad, for providing the necessary support to carry out this work.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.atech.2025.101381.

Data availability

Data will be made available on request.

References

[1] G. Katti, C. Padmavathi, R.M. Kumar, Integrated pest management in rice-based cropping systems in India. Integrated Pest Management in Diverse Cropping Systems, Apple Academic Press, 2023, pp. 111–135.
[2] R.A. Balikai, V. Madhurima, S. Desai, Non-chemical approaches for the management of insect pests in agri-horti crops and storage, J. Eco-friendly Agric. 15 (2020) 95–111.
[3] T. Liboschik, K. Fokianos, R. Fried, tscount: an R package for analysis of count time series following generalized linear models, J. Stat. Softw. 82 (2017) 1–51.
[4] K. Fokianos, A. Rahbek, D. Tjøstheim, Poisson autoregression, J. Am. Stat. Assoc. 104 (2009) 1430–1439.
[5] F. Zhu, Modeling count time series with Markov processes based on binomial thinning, J. Time Ser. Anal. 33 (2012) 150–162.
[6] Weib, Monthly claims count of workers in manufacturing industry, J. Risk Insur. 76 (2009) 663–684.
[7] Weib, Monthly strike count time series: an application of the Poisson model, Appl. Econ. 42 (2010) 3277–3286.
[8] F. Zhu, Z. Wang, Campylobacterosis infections count time series modeling, Stat. Modelling. 10 (2010) 79–94.
[9] T. Liboschik, L. Hufnagel, A. Ziegler, Count time series of campylobacteriosis infections in Germany, BMC Public Health 14 (2014) 239.
[10] W. Yang, M. Qiu, Time series prediction of influenza activity using Poisson-INGARCH model with Google Trends, Epidemiology & Infect. 147 (2019) e123.
[11] E.P.Basuki Tanawi, S. Yulianto, Prediction of dengue incidents in Jakarta using time series analysis, Int. J. Environ. Res. Public Heal. 18 (2021) 1332.
[12] M. Kim, Network traffic prediction based on INGARCH model, Wirel. Netw. 26 (2020) 6189–6202.
[13] Y.H. Kim, S.J. Yoo, Y.H. Gu, J.H. Lim, D. Han, S.W. Baik, Crop pests prediction method using regression and machine learning technology: survey, IERI Procedia 6 (2014) 52–56.
[14] W. Alam, K. Sinha, R.R. Kumar, M. Ray, S. Rathod, K.N. Singh, P. Arya, Hybrid linear time series approach for long-term forecasting of crop yield, Indian J. Agric. Sci. 88 (2018) 1275–1279.
[15] S. Rathod, G.C. Mishra, K.N. Singh, Hybrid time series models for forecasting banana production in Karnataka State, India, J. Indian Soc. Agric. Statist. 71 (2017) 193–200.
[16] K. Sriwanna, Weather-based rice blast disease forecasting, Comput. Electron. Agric. 193 (2022) 106685.
[17] V. Amaratunga, L. Wickramasinghe, A. Perera, J. Jayasinghe, U. Rathnayake, Artificial neural network to estimate the paddy yield prediction using climatic data, Math. Probl. Eng. 2020 (2020) 1–11.
[18] C.C. Ma, Y. Liang, X. Lyu, Weather analysis to predict rice pest using neural network and DS evidential theory, in: International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2019, pp. 277–283.
[19] S.Mondal Paul, A. Ghosh, Prediction of early blight severity in tomato using machine learning models, J. Plant Pathol. 101 (2019) 801–809.
[20] T. Huang, R. Yang, W. Huang, Y. Huang, X. Qiao, Detecting sugarcane borer diseases using support vector machine, Inf. Process. Agric. 5 (2018) 74–82.
[21] N.R. Prasannakumar, S. Chander, V.L. Kumar, Development of weather-based rice yellow stem borer prediction model for the Cauvery command rice areas, Karnataka, India, Cogent Food Agric. 1 (2015) 995281.
[22] K.G. Bapatla, B.G. Gadratagi, N.B. Patil, G.P.P. Govindharaj, L.N. Thalluri, B. B. Panda, Predictive modelling of yellow stem borer population in rice using light trap: A comparative study of MLP and LSTM networks, Ann. Appl. Biol (2024).
[23] R.B. O'Hara, D.J. Kotze, Do not log-transform count data, Methods Ecol. Evol. 1 (2010) 118–122.
[24] A.P. St-Pierre, V. Shikon, D.C. Schneider, Count data in biology—Data transformation or model reformation? Ecol. Evol. 8 (2018) 3077–3085.
[25] A.H. Lee, K. Wang, J.A. Scott, K.K. Yau, G.J. McLachlan, Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros, Stat. Methods Med. Res. 15 (2006) 47–61.
[26] H.E. Hua, T.A. Wan, W.A. Wenjuan, P. Crits-Christoph, Structural zeroes and zero-inflated models, Shanghai. Arch. Psychiatry 26 (2014) 236–242.
[27] L.E. Nieto-Barajas, D. Bandyopadhyay, A zero-inflated spatial gamma process model with applications to disease mapping, J. Agric. Biol. Environ. Stat. 18 (2013) 137–158.
[28] S. Islam, C.E. Haque, S. Hossain, K. Rochon, Role of container type, behavioural and ecological factors in Aedes pupal production in Dhaka, Bangladesh: an application of zero-inflated negative binomial model, Acta Trop. 193 (2019) 50–59.
[29] Jia, Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data, SIAM. J. Appl. Math. 80 (2020) 1336–1355.
[30] L.P. Fávero, J.F. Hair Jr, R.D.F. Souza, M. Albergaria, T.V. Brugni, Zero-inflated generalized linear mixed models: a better way to understand data relationships, Mathematics 9 (2021) 1100.
[31] K.I. Kang, K. Kang, C. Kim, Risk factors influencing cyberbullying perpetration among middle school students in Korea: analysis using the zero-inflated negative binomial regression model, Int. J. Environ. Res. Public Heal. 18 (2021) 2224.
[32] Y. Hao, C. Tian, A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting, Appl. Energy 238 (2019) 368–383.
[33] J. Liu, S. Zhang, H. Fan, A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network, Expert. Syst. Appl. 195 (2022) 116624.
[34] R Core Team, R: A language and Environment for Statistical Computing (Version 3.5.1) [Computer Software], R Foundation for Statistical Computing, 2018.
[35] S. Rathod, S. Yerram, P. Arya, G. Katti, J. Rani, A.P. Padmakumari, N. Somasekhar, C. Padmavathi, G. Ondrasek, S. Amudan, S. Malathi, N.M. Rao, K. Karthikeyan, N. Mandawi, P. Muthuraman, R.M. Sundaram, Climate-based modeling and prediction of Rice Gall Midge populations using count time series and machine learning approaches, Agronomy 12 (2022).
[36] Heinen, Modelling time series count data: an autoregressive conditional Poisson model, SSRN. 1117187 (2003).
[37] K. Fokianos, Some recent progress in count time series, Statistics. 45 (2011) 49–58.
[38] K. Tawiah, W.A. Iddrisu, K.S. Asosega, Zero-Inflated Time series modelling of Covid-19 deaths in Ghana, J. Environ. Public Heal. (2021) 1–9.
[39] J. Kim, S. Park, H. Lee, A hybrid model combining the negative binomial and logit distributions for count data analysis, J. Stat. Comput. Simul. 91 (2021) 2456–2471.
[40] F.L. Pinheiro, M.P. Andrade, G. Kreimann, Comparison of zero-inflated models for overdispersed count data, Stat. Methods Appt. 30 (2021) 111–128.
[41] G.P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing. 50 (2003) 159–175.
[42] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (4) (2006) 679–688.
[43] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, J. Bus. Econ. Stat. 13 (3) (1995) 253–263.
[44] B.N.K. Reddy, S. Rathod, S. Kallakuri, Y. Sridhar, M. Admala, S. Malathi, P. Pandit, B. Jyostna, Modelling the relationship between weather variables and rice yellow stem borer population: A count data modelling approach, Int. J. Environ. Clim. Chang. 12 (2022) 3623–3632.
[45] K.N.Singh Saha, M. Ray, S. Rathod, A hybrid spatio-temporal modelling: an application to space-time rainfall forecasting, Theor. Appl. Climatol. 142 (2020) 1271–1282.
[46] S. Rathod, A. Saha, R. Patil, G. Ondrasek, C. Gireesh, M.S. Anantha, D.V.K.N. Rao, N. Bandumula, P. Senguttuvel, A.K. Swarnaraj, S.N. Meera, A. Waris, P. Jeyakumar, B. Parmar, P. Muthuraman, R.M. Sundaram, Two-stage

spatiotemporal time series modelling approach for rice yield prediction & advanced agroecosystem management, Agronomy 11 (2021) 2502.
[47] K.N.Singh Saha, M. Ray, S. Rathod, M. Dhyani, Fuzzy rule-based weighted space-time autoregressive moving average models for temperature forecasting, Theor. Appl. Climatol. 150 (2022) 1321–1335.
[48] M.Liu Rao, M. Goh, J. Wen, 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers, Appl. Soft. Comput. 95 (2020) 106570.
[49] Z. Tang, G. Zhao, T. Ouyang, Two-phase deep learning model for short-term wind direction forecasting, Renew. Energy 173 (2021) 1005–1016.
[50] N. Pachar, J.D. Darbari, K. Govindan, P.C. Jha, Sustainable performance measurement of Indian retail chain using two-stage network DEA, Ann. Oper. Res. 310 (2022) 505–527.
[51] D. Panigrahi, S. Rajamani, Studies on the biology and reproductive behaviour of Scirpophaga incertulas Walker, Oryza, 45 (1) (2008) 137–141.
[52] G. Nayak, A. Prabhuraj, S. Hurali, S.G. Hanchinal, M. Bheemanna, B.G. Koppalkar, J.M. Nidagundi, Studies on reproductive biology of yellow stem borer, scirpophaga incertulas Walker in the changing climate scenario, Int. J. Environ. Clim. Chang. 13 (11) (2023) 2354–2361.