

Short-Term Forecasting of Monthly Syphilis Cases Using a Stability and Recency-Aware Selective Ensemble

Shanhua Zhang

Jiangsu Vocational College of Electronic and Information

Xinxin Shen

xinxin.shen1997@outlook.com

Jinhu County Center for Disease Control and Prevention

Article

Keywords: Syphilis, Time series forecasting, Selective ensemble, Rolling-origin validation, Particle swarm optimization, Disease surveillance

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11268766/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Short-Term Forecasting of Monthly Syphilis Cases Using a Stability and Recency-Aware Selective Ensemble

Shanhua Zhang 1, 2, and Xinxin Shen 3*

1Department of Digital Equipment, Jiangsu Vocational College of Electronic and Information, Huai'an 223003, China; 201203@jsei.edu.cn

2School of Traffic and Transportation, Shijiazhuang Tiedao University, Shijiazhuang 050000, China

3Jinhu County Center for Disease Control and Prevention, Jinhu 211600, China; xinxin.shen1997@outlook.com

* Correspondence:

Xinxin Shen

xinxin.shen1997@outlook.com

Abstract: Accurate short-term forecasting of syphilis cases is important for disease surveillance and public health planning. This study developed a stability and recency-aware selective ensemble (SRSE) framework using monthly syphilis cases in Jiangsu Province from 2021–2024 for model development and 2025 data for independent testing. Nine statistical forecasting models were evaluated by expanding-window rolling-origin validation. SRSE integrated recency-weighted errors, cross-origin stability, and candidate-model screening, with key hyperparameters optimized using Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Sparrow Search Algorithm (SSA). PSO-SRSE achieved the lowest test errors, with MAE, RMSE, MAPE, and sMAPE of 109.39, 132.35, 3.89%, and 3.87%, respectively, outperforming the unoptimized SRSE, equal-weighted ensemble, and all individual models. These findings suggest that PSO-SRSE may improve point forecasting for short monthly syphilis surveillance series.

Keywords: Syphilis; Time series forecasting; Selective ensemble; Rolling-origin validation; Particle swarm optimization; Disease surveillance

1 Introduction

Syphilis is a chronic sexually transmitted disease caused by *Treponema pallidum* infection. It can be transmitted via sexual contact, blood-borne routes, and mother-to-child vertical transmission, and remains a major global public-health concern to this day [1]. Studies on the global burden of disease have indicated that sexually transmitted infections continue to impose a considerable disease burden, with marked disparities across regions [2]. In China, reported syphilis cases have remained at a persistently high level over a prolonged period. Based on surveillance data from 31 provinces in mainland China between 2004 and 2019, a total of 5,527,399 syphilis cases were reported during this period, exhibiting an overall upward trend alongside significant regional heterogeneity [3]. Therefore, timely identification of the temporal patterns of syphilis case counts and the conduct of short-term forecasting are of great importance for understanding epidemic trends, rational allocation of health resources, and the formulation of evidence based prevention and control strategies.

Time series approaches have been widely adopted to forecast syphilis incidence trends. Zhang et al. constructed ARIMA and ARIMAX models using syphilis surveillance data across China from 2005 to 2012, verifying the feasibility of conventional time series models for

predicting syphilis trends and seasonal fluctuations[4]. For Jiangsu Province, further comparative investigations between ARIMA and Bayesian structural time series models revealed divergent forecasting performances across independent test periods[5]. Moreover, ARIMA, neural networks, long short term memory (LSTM), and various hybrid time series models have been applied to predict syphilis incidence in China in recent years, where different methodologies exhibit distinct advantages under varying forecast horizons and data conditions[6,7]. Existing studies suggest that the relative superiority of individual forecasting models is data dependent. Hence, comparing candidate models with divergent structural assumptions and further synthesizing their forecasting outputs may yield better adaptability than relying on a single standalone model.

Given that the training dataset in this study only contains 48 monthly observations, statistically parsimonious statistical models with alternative time series assumptions are selected to build the candidate model library, so as to mitigate overfitting risks of complex models under limited sample conditions. Mean, Naive and Seasonal Naive serve as fundamental forecasting benchmarks, and large scale forecasting competitions have highlighted the reference value of simple baselines for model comparison[8]. Simple exponential smoothing (SES) characterizes local level shifts of time series via exponentially decaying weights[9]; the Holt model further incorporates trend components[10], while damped Holt introduces a damping mechanism to mitigate risks arising from excessive extrapolation of linear trends over long horizons[11]. Exponential smoothing state space (ETS) models describe level, trend and seasonal patterns within a unified state space framework[12]. The ARIMA model captures trends and serial correlations embedded in time series data through autoregressive, differencing and moving average terms[13]. In addition, Fourier harmonic terms can describe smooth periodic variations with few parameters, and can be combined with autoregressive error structures to establish harmonic regression models[14]. Conventional statistical models and their combinations remain vital benchmark comparators for influenza forecasting and health care demand prediction tasks[15,16]. On this basis, nine models, namely Mean, Naive, Seasonal Naive, SES, Holt, Damped Holt, ETS, ARIMA and Fourier-AR, are adopted as candidate forecasting models to cover diverse time series features including level, trend, seasonal periodicity and short term autocorrelation. This setup of candidate models is consistent with previous methodological frameworks.

Rigorous evaluation of time series models requires strict preservation of temporal ordering between training inputs and future observations. Rolling origin validation generates out of sample forecasts iteratively from multiple temporal origins. It enables more comprehensive assessment of model performance across different time periods compared with a single train-test split, and alleviates evaluation bias induced by look ahead information[17,18]. On the other hand, forecast combination integrates information from diverse models and mitigates model selection risks stemming from dependence on a single “best performing” model[19]. Infectious disease forecasting literature has indicated that ensemble forecasting generally yields robust outcomes; nevertheless, performance gains brought by additional models tend to diminish when low quality or highly correlated models are incorporated into ensembles[20]. Meanwhile, model predictive performance may vary over time, which provides statistical justification for time varying weighting schemes for candidate model combination[21]. For disease surveillance time series affected by shifting transmission dynamics, reporting practices and population activity patterns, model screening and weighting merely based on average forecasting errors across all historical

stages may fail to reflect recent predictive capability. Ignoring error volatility across rolling windows may further assign high ensemble weights to models with satisfactory average accuracy yet insufficient stability.

Motivated by the above considerations, this study proposes the SRSE framework. First, expanding window rolling origin validation is implemented to evaluate out of sample performance of the nine candidate models. An exponential time decay mechanism is introduced such that validation results closer to the end of the training period receive higher weights. Candidate models are then screened according to recency weighted forecasting errors, and error fluctuations across rolling-origin windows are incorporated into the composite loss function to jointly optimize predictive accuracy and stability. Non-negative ensemble weights are subsequently derived from the composite loss. PSO is further employed for joint optimization of hyperparameters, including the number of retained models, time decay coefficient, stability penalty coefficient and weight adjustment parameter. Comparative analyses are performed against GA and SSA[22-24]. PSO, GA and SSA adopt distinct swarm or evolutionary search mechanisms and are suitable for multi-parameter optimization problems. Data spanning January 2021 to December 2024 are utilized for model fitting, rolling origin validation and hyperparameter tuning, while data from January 2021 to December 2025 are reserved as a fully independent test set to assess out of sample performance of the final model. This work explores a selective ensemble forecasting approach tailored for disease surveillance time series with limited sample size.

The main contributions of this paper are summarized as follows:

First, a SRSE framework is proposed for short disease surveillance time series. It jointly incorporates recency weighted validation error, cross origin prediction stability, and candidate model screening to reduce the influence of outdated or unstable forecasts on the final ensemble.

Second, a leakage free expanding window rolling-origin strategy is used for model evaluation and calibration. All model selection, stability assessment, and ensemble weighting are based solely on training-period out of sample forecasts, while the 2025 observations are reserved as an independent test set.

Third, a joint hyperparameter optimization scheme is developed for SRSE, with PSO, GA, and SSA compared under identical settings. The optimization performance, repeated-run stability, and sensitivity to the search space are further evaluated to assess the robustness of the resulting ensemble.

This study proposes an optimized SRSE model, and its modeling procedure is illustrated in Figure 1.

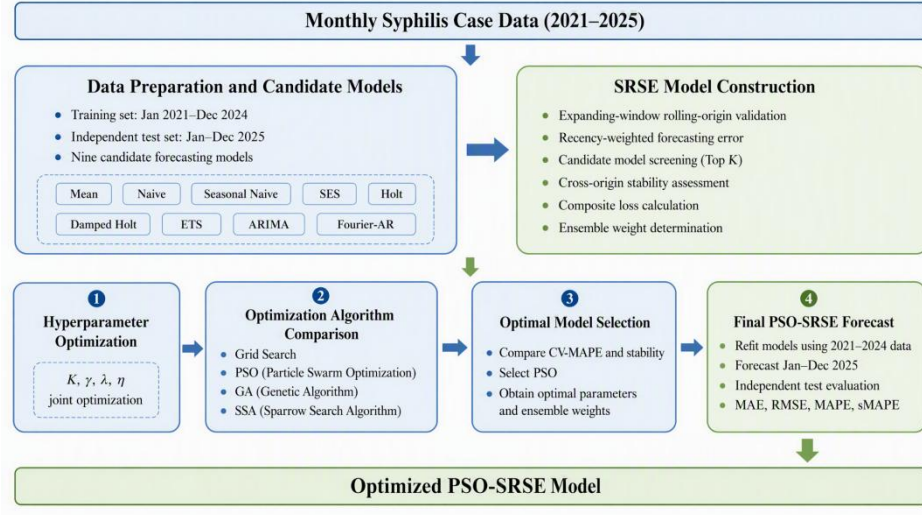


Figure 1. Modeling procedure

2 Candidate forecasting models

2.1 Data partitioning and research design

This study incorporated syphilis case report data from January 2021 to December 2025 in Jiangsu Province, comprising a total of 60 monthly observations. The temporal trend of case counts is illustrated in Figure 2. To evaluate the out-of-sample forecasting performance of the models on future unseen time periods, the dataset was chronologically partitioned into a training set and an independent test set. Specifically, data from January 2021 to December 2024, encompassing 48 monthly observations, were designated as the training set for model fitting, hyperparameter selection, and combination weight determination; while data from January to December 2025, comprising 12 monthly observations, were reserved as the independent test set for the final evaluation of each model's forecasting performance.

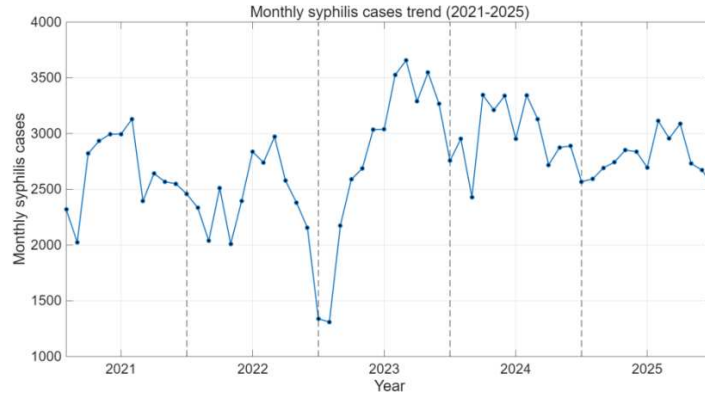


Figure 2. Monthly trend of reported syphilis cases, 2021–2025

2.2 Rolling-origin time series validation

Design ideas for training samples: To evaluate the out-of-sample forecasting performance of different candidate models within the training period, this study employed a rolling-origin expanding-window validation approach based on the training data from 2021 to 2024. Let the r -th rolling forecast origin be denoted as T_r , and the data used for model fitting in this round be $\{y_1, y_2, \dots, y_{T_r}\}$. After model fitting, an h -step ahead forecast was generated from the corresponding forecast origin, denoted as $\hat{y}_{m,r,T_r+h}$, where m indexes the candidate model, r indexes the rolling forecast origin, and h denotes the forecast horizon. The fitting samples and

prediction intervals corresponding to each rolling window are illustrated in Figure 3.



Figure 3. Sample setting of rolling origin expanding window verification

Since this study focuses on evaluating the multi-step ahead forecasting performance of each model over the next 12 months, each rolling forecast origin that satisfies the sample condition was used to generate consecutive 12-month ahead forecasts, i.e., $h=1,2,\dots,12$. The fitting sample for all rolling windows starts from January 2021. Considering the annual cyclical characteristics and the limited sample conditions, the first rolling window employed 24 months of data from January 2021 to December 2022 for model fitting and generated forecasts for January to December 2023. Thereafter, the forecast origin was moved forward month by month, with the fitting window expanding by one month each time, and the sample size gradually increasing from 24 to 36 months. The final rolling window used data from January 2021 to December 2023 for fitting and generated forecasts for January to December 2024.

Design ideas for test samples: The dataset spanning from January 2021 to December 2024 is adopted for model fitting, and predictions are generated for the period from January to December 2025.

For each rolling forecast origin and forecast horizon, the prediction error between the model forecast and the actual observation was calculated to evaluate the out-of-sample forecasting performance of each candidate model. The prediction error of the m -th model at the r -th rolling forecast origin for the h -th month ahead is defined as:

$$e_{m,r,h} = \frac{|y_{T_r+h} - \hat{y}_{m,r,T_r+h}|}{y_{T_r+h}} 100\% \quad (1)$$

where y_{T_r+h} denotes the actual number of cases in the h -th future month, $\hat{y}_{m,r,T_r+h}$ denotes the forecast value for the h -th future month produced by model m at the r -th rolling forecasting origin, and $e_{m,r,h}$ denotes the corresponding relative error. Averaging $e_{m,r,h}$ over the next 12 months yields:

$$E_{m,r} = \frac{1}{12} \sum_{h=1}^{12} e_{m,r,h} \quad (2)$$

$E_{m,r}$ denotes the MAPE of model m for continuous 12-month forecasts starting from the r -th rolling forecasting origin. This design can repeatedly generate prediction results for multiple samples under limited sample conditions, thereby reducing the randomness introduced by model selection that relies solely on a single train-test split.

2.3 Principles of Candidate Models

The training set in this study comprises only 48 monthly observations. Due to the limited sample size, complex machine learning or deep learning models may face a high risk of overfitting, which could compromise their out-of-sample forecasting performance. Therefore, this study prioritized the construction of a candidate model pool comprising statistical forecasting models with relatively few parameters, suitability for short time series, and distinct modeling assumptions. The candidate models include the Mean model, the Naive model, the Seasonal Naive model, Simple Exponential Smoothing (SES), Holt's linear trend model, Damped Holt's model, the Exponential Smoothing state space model (ETS), the Autoregressive Integrated Moving Average model (ARIMA), and a harmonic regression model with Fourier terms for characterizing periodic patterns combined with an autoregressive error structure (Fourier-AR).

(1) Mean model

The Mean model assumes that future case numbers fluctuate around the long-term average level of the historical observed series. At the r -th rolling forecasting origin T_r , all training data available up to this time point, $\{y_1, y_2, \dots, y_{T_r}\}$, are used to compute the sample mean, which is taken as the future predicted value. Therefore, the 12-step forecasts of the Mean model yield a single constant value. This model captures no trend, seasonal, or autocorrelation structures, and it mainly serves as a baseline forecasting model with low complexity and low variance.

(2) Naive model

The Naive model assumes that future case numbers remain at the most recent observed level. At the rolling forecasting origin T_r , all future forecast horizons take the value of the last observed case count. The Naive model is equivalent to the one-step and multi-step forecasts of the drift-free random walk model. It is used to verify whether the performance of complex models can truly outperform the simplest assumption that “the future equals the present”.

(3) Seasonal Naive model

Considering that monthly infectious-disease data may exhibit a 12-month seasonal cycle, this study sets the seasonal period $m=12$. The Seasonal Naive model predicts future corresponding months using the observed values from the same month of the previous year. This model does not estimate any parameters and can serve as a benchmark for examining whether the annual seasonal pattern possesses stable repeatability.

(4) Simple Exponential Smoothing (SES)

SES model assumes that a time series consists mainly of a time-varying local level, without explicit trend and seasonal components. The multi-step forecasts of SES are constant extensions of the final estimated level. In this study, the smoothing parameter α is not manually specified; instead, it is estimated via numerical optimization within the training window to minimize the model fitting error. The initial level is determined by the model-internal estimation method based on the training series.

(5) Holt's linear trend model

The Holt model further incorporates a linear trend term based on the local level of SES. Its state equations are given by:

$$\begin{aligned}\sigma_t &= \alpha y_t + (1 - \alpha)(\sigma_{t-1} + b_{t-1}) \\ b_t &= \beta(\sigma_t - \sigma_{t-1}) + (1 - \beta)b_{t-1}\end{aligned}\tag{3}$$

where σ_t is the local level; b_t is the local trend; α is the level smoothing parameter; β is the trend smoothing parameter. This study adopts the additive linear-trend formulation without seasonal terms. The parameters α and β , together with the initial level and initial trend, are estimated by numerical optimization within the current training window.

(6) Damped Holt model

Considering that the unlimited extrapolation of linear trends into the future may lead to over-extrapolation for long term forecasts, this study further constructs the damped Holt model. Where ϕ is the damping parameter. For $0 < \phi < 1$, the contribution of the trend gradually decays as the forecast horizon increases. When $\phi = 1$, the model reduces to the standard Holt's linear trend model.

The level smoothing parameter, trend smoothing parameter and damping parameter of the model are all estimated using data within the current training window. No seasonal terms are incorporated in this model.

(7) Exponential Smoothing state space model (ETS)

The ETS model adopts a state space representation consisting of three components: Error, Trend, and Seasonal. In this study, $m=12$ is set for the monthly case-count data. An additive seasonal structure is employed to avoid excessive fluctuations induced by multiplicative seasonal terms caused by magnitude changes in case counts.

Level update:

$$\rho_t = \alpha(y_t - s_{t-12}) + (1 - \alpha)(\rho_{t-1} + \mu_{t-1}) \quad (4)$$

Trend update:

$$\mu_t = \beta(\rho_t - \rho_{t-1}) + (1 - \beta)\mu_{t-1} \quad (5)$$

Seasonal update:

$$s_t = \gamma_s(y_t - \rho_t) + (1 - \gamma_s)s_{t-12} \quad (6)$$

where ρ_t is the level component; μ_t is the trend component; s_t is the 12-month seasonal component; α, β , and γ_s are the corresponding smoothing parameters.

(8) ARIMA(p,q,d) model

The autoregressive integrated moving average (ARIMA) model is adopted to characterize the linear autocorrelation structure within the monthly syphilis case-count series. Its general form is denoted as ARIMA(p,d,q), where p is the autoregressive order, d is the differencing order, and q is the moving-average order. where B denotes the back-shift operator, $(1 - B)^d$ represents the d -th order differencing, $\tau(B)$ and $\theta(B)$ are the autoregressive polynomial and moving-average polynomial, respectively, and ε_t is the random error term. The optimal parameters p, q and d are selected according to the model comparison results over the training period.

(9) Fourier-AR model

Considering that fixed 12-month seasonal factors may be overly rigid, this study further constructs a Fourier-AR model. Sine and cosine functions are adopted to characterize smooth periodic variations, while an autoregressive error structure is used to describe the remaining temporal autocorrelation. For monthly data, the seasonal period is set as $m=12$. Where β_0 denotes the intercept; β_1 is the linear time-trend coefficient; a_1 and b_1 represent the Fourier harmonic coefficients; and ϕ is the residual AR(1) coefficient.

The multi-step forecasting equations for the nine models are shown in Table 1.

Table 1. Multi-step forecasting equations of the nine models

Model	Multi- step forecasting formula $\hat{y}_{t+h}$	Description
-------	---	-------------

Mean	$\hat{y}_{t+h} = \bar{y}_t$	Historical sample mean $\bar{y}_t = \frac{1}{t} \sum_{i=1}^t y_i$
Naive	$\hat{y}_{t+h} = y_t$	Last observed value
Seasonal Naive	$\hat{y}_{t+h} = y_{t+h-12}$	Value from the same period of the previous year; seasonal period $m=12$
SES	$\hat{y}_{t+h} = \ell_t$	Local level, $\ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1}$ $\alpha \in (0,1)$
Holt	$\hat{y}_{t+h} = \sigma_t + hb_t$	Additive linear trend, where σ_t is the level and b_t is the trend
Damped Holt	$\hat{y}_{t+h} = \sigma_t + \left(\sum_{j=1}^h \phi^j \right) b_t$	Damped trend, $0 < \phi < 1$
ETS	$\hat{y}_{t+h} = \rho_t + h\mu_t + s_{t+h-12k}$	Level, trend, and seasonal updates; $k = [(h-1)/12] + 1$
ARIMA(p,q,d)	$\tau(B)(1-B)^d y_{t+h} = c + \theta(B)\varepsilon_{t+h}$	AR coefficients $\tau_1, \dots, \tau_p$; MA coefficients $\theta_1, \dots, \theta_q$; differencing order d
Fourier-AR	$y_{t+h} = \beta_0 + \beta_1(t+h) + a_1 \sin\left(\frac{2\pi(t+h)}{12}\right) + b_1 \cos\left(\frac{2\pi(t+h)}{12}\right) + u_{t+h}$	Linear trend + first-order Fourier harmonics + AR(1) error correction $u_{t+h} = \phi u_{t+h-1} + \varepsilon_{t+h}$

When $t=T_r$, the multi-step forecasting formulas presented in the table above can be applied to generate out-of-sample predictions within the rolling-origin validation procedure. In this case, the model is fitted using the current training window and produces forecasts for the subsequent months beyond that window. When t is set to December 2024, i.e., the final forecast origin, the same formulas are used to generate the final forecasts for January to December 2025.

3 Stability- and recency-aware selective ensemble(SRSE)

3.1 Recency-weighting mechanism of forecasting performance

Disease-surveillance time series can be affected by changes in disease transmission status, reporting practices, and population activity patterns. Consequently, a model's predictive performance in earlier periods may not adequately reflect its forecasting capability for recent periods. To mitigate the influence of early validation outcomes on current model evaluation, this study introduces an exponential time-decay mechanism. Validation errors from different rolling forecasting origins are weighted, such that rolling-validation results temporally closer to the final training time point are assigned higher weights.

Suppose that rolling-origin validation contains R forecasting origins, where $R=13$ in this study. The time weight corresponding to the r -th rolling forecasting origin is re-defined as:

$$q_r = \gamma^{R-r} \quad (7)$$

where γ denotes the time-decay coefficient with $0 < \gamma \leq 1$. The weight q_r is further normalized:

$$\tilde{q}_r = \frac{q_r}{\sum_{j=1}^R q_j} \quad (8)$$

where $\tilde{q}$ is referred to as the rolling weight, and $\sum_{r=1}^R \tilde{q}_r = 1$. When $\gamma < 1$, rolling-validation results temporally closer to the final training time point are assigned higher

weights. When $\gamma = 1$, equal weights are allocated to all rolling forecasting origins. On this basis, the recency-weighted average forecasting error of model m is defined as:

$$\bar{E}_m = \sum_{r=1}^R \tilde{q}_r E_{m,r} \quad (9)$$

The selection of the time-decay coefficient γ is entirely based on the rolling-validation results within the training period, without using information from the 2025 independent test set, so as to prevent test-data leakage. The detailed method will be discussed in the subsequent section.

3.2 Candidate-model performance screening

For short term time series scenarios, incorporating all candidate models with poor forecasting performance or insufficient stability into the model ensemble may increase the prediction variance and reduce the robustness of the combined model. Therefore, prior to determining the final ensemble weights, this study first ranks the candidate models based on the recency-weighted forecasting errors derived from rolling-validation over the training period, according to:

$$\bar{E}_1 \leq \bar{E}_2 \leq \dots \leq \bar{E}_M \quad (10)$$

Only the top K models with better forecasting performance are retained for the subsequent ensemble stage. The retained-model number K is also determined solely based on the internal validation results within the training period, without utilizing information from the 2025 independent test set. This screening step aims to reduce the extra noise introduced by weak-performing models in the ensemble, so as to improve the stability of the combined forecasts.

3.3 Stability-aware loss function

Selecting models merely based on average forecasting errors may overlook model stability across different prediction origins. If a model yields low average error yet exhibits extremely large prediction biases in several time windows, its reliability for practical disease-early-warning applications remains limited. Therefore, this study further calculates the dispersion degree of forecasting errors for each model during rolling-validation. Let the forecasting errors of model m at each rolling origin be $E_{m,1}, E_{m,2}, \dots, E_{m,R}$. Its error-stability index is defined as:

$$S_m = SD(E_{m,1}, E_{m,2}, \dots, E_{m,R}) \quad (11)$$

where $SD(\cdot)$ denotes the standard deviation of forecasting errors for model m across different rolling prediction origins. Further, a composite loss function that simultaneously considers forecasting accuracy and forecasting stability is constructed:

$$L_m = \bar{E}_m + \lambda S_m \quad (12)$$

where $\bar{E}_m$ reflects the average forecasting accuracy of the model, and S_m measures the performance fluctuation of the model across different rolling forecasting windows. λ denotes the stability-penalty coefficient, and its value characterizes the influence of performance fluctuation on the composite loss function. A smaller L_m indicates that the model not only achieves low overall forecasting error but also exhibits more stable performance over different time periods. The determination method of λ will be discussed in subsequent sections.

3.4 Determination of Weights and Prediction for the Ensemble Model

After completing candidate-model performance screening, non-negative ensemble weights are determined based on the composite loss of the selected models. For the m -th candidate model entering the final ensemble, its ensemble weight is defined as:

$$w_m = \frac{L_m^{-\eta}}{\sum_{j=1}^K L_j^{-\eta}} \quad (13)$$

where w_m denotes the model weight determined according to recent forecasting performance and error stability, satisfying $w_m \geq 0, \sum_{m=1}^K w_m = 1$. η is the weight-adjustment parameter, which is used to tune the weight discrepancy among different models. When η is small, the weights of individual models become closer. As η increases, models with smaller rolling-validation loss are assigned higher weights. The determination method of η will be discussed in subsequent sections.

Finally, for the forecasting horizon h , the prediction result of the ensemble model is given by:

$$\hat{Y}_{t+h}^{SRSE} = \sum_{m=1}^K w_m \hat{Y}_{m,t+h} \quad (14)$$

where t denotes the starting point for final-horizon prediction, and $\hat{Y}_{m,t+h}$ is the predicted value for the h -th month obtained from the m -th candidate model.

3.5 Rules for Hyperparameter Selection

The SRSE model involves four main hyperparameters, namely the retained-model number K , the time-decay coefficient γ , the stability-penalty coefficient λ , and the ensemble-weight adjustment parameter η . All hyperparameters are selected solely based on the rolling time-series validation results within the training data from January 2021 to December 2024, and the actual case data from the 2025 independent test set are never used during parameter optimization. After determining the optimal parameters, the candidate models are re-fitted using the complete 48-month training data from 2021 to 2024, and the final forecasts for January–December 2025 are generated according to the model-screening rules and ensemble weights determined during the training period.

(1) Grid-search algorithm

In this study, joint grid-search is adopted to simultaneously optimize the hyperparameters of the SRSE model, instead of optimizing each parameter sequentially and independently. The retained-model number is set as $K \in \{2, 3, 4, 5, 6\}$, the time-decay coefficient as $\gamma \in \{0.8, 0.9, 1.0\}$, the stability-penalty coefficient as $\lambda \in \{0, 0.5, 1\}$, and the weight-adjustment parameter as $\eta \in \{0.5, 1, 2\}$, yielding a total of 135 parameter combinations. For each parameter combination, out-of-sample MAPE is calculated via leak-free rolling time-series validation. The parameter combination with the minimum validation error is selected for the final model, and the optimal parameters are determined by $\theta^* = \arg \min_{\theta} CV - MAPE(\theta)$, where $\theta = (K, \gamma, \lambda, \eta)$ denotes an arbitrary parameter combination, θ^* represents the optimal parameter combination, and the calculation of $CV - MAPE(\theta)$ is given in Eq. (16).

(2) Metaheuristic algorithms

Apart from joint grid-search, metaheuristic optimization algorithms are further adopted to search the hyperparameters of the SRSE model. These metaheuristic algorithms share the same parameter ranges as grid-search: $K \in \{2, 3, \dots, 6\}$, $\gamma \in [0.8, 1]$, $\lambda \in [0, 1]$, and $\eta \in [0.5, 2]$. The retained-model number K is an integer variable, and continuous values generated by the algorithm are rounded to integers; γ , λ and η are optimized as continuous variables. Compared with grid-search limited to discrete candidate values, metaheuristic optimization algorithms can explore potential parameter combinations with higher search resolution within given parameter bounds. The out-of-sample CV-MAPE obtained from rolling-origin cross-validation is taken as the fitness function. The absolute percentage error for a single forecasting point is defined as:

$$e_{r,h}(\theta) = \frac{|y_{T_r+h} - \hat{y}_{r,h}^{SRSE}(\theta)|}{y_{T_r+h}} \quad (15)$$

$\hat{y}_{r,h}^{SRSE}(\theta)$ denotes the predicted value of the SRSE model for the h -th month under the r -th rolling-forecast origin. Furthermore, the prediction errors satisfying the strict rolling-validation conditions are averaged to obtain:

$$Fitness(K, \gamma, \lambda, \eta) = \min_{K, \gamma, \lambda, \eta} CV - MAPE(\theta) = \frac{100\%}{N} \sum_{(r,h) \in \mathcal{V}} \frac{|y_{T_r+h} - \hat{y}_{r,h}^{SRSE}(\theta)|}{y_{T_r+h}} \quad (16)$$

where $\theta = (K, \gamma, \lambda, \eta)$; $\mathcal{V}$ denotes all valid rolling-validation predictions; N is the total number of prediction points involved in validation.

3.6 Prediction performance evaluation

To comprehensively evaluate the forecasting performance of each model on the independent test set of the year 2025, four metrics are adopted in this study: mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (sMAPE). The calculation formulas for these metrics are given as follows:

$$\begin{aligned} MAE &= \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \\ RMSE &= \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \\ MAPE &= \frac{1}{n} \sum_{t=1}^n \frac{|y_t - \hat{y}_t|}{|y_t|} \times 100\% \\ sMAPE &= \frac{1}{n} \sum_{t=1}^n \frac{2|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \times 100\% \end{aligned} \quad (17)$$

MAE and RMSE reflect the absolute deviations between predicted and actual values, where RMSE imposes a higher penalty on large prediction errors. MAPE and sMAPE evaluate forecasting performance from the perspective of relative errors. Smaller values of these metrics indicate better model forecasting performance. Considering that different evaluation metrics have different emphases, this study does not judge model superiority based on a single metric. Instead, MAE, RMSE, MAPE, and sMAPE of each candidate model and the ensemble model on the independent test set are comprehensively compared.

4 Result

4.1 Forecasting performance of candidate prediction models

Nine candidate models are adopted to forecast the reported syphilis case counts in Jiangsu Province from January to December 2025. The implementation-related parameters for each model are presented in Table 2.

Table 2. Implementation parameters of the nine candidate models

Model	Trend	Seasonality	Seasonal Period	Key Parameter Handling	Forecast Horizon
Mean	None	None	—	Historical sample mean	12
Naive	None	None	—	Last observed value	12
Seasonal	None	Fixed	12	Value from the same	12

Naive		seasonal		period of last year	
SES	None	None	—	α automatically estimated on training set	12
Holt	Additive linear	None	—	α, β automatically estimated	12
Damped Holt	Damped trend	None	—	α, β, ϕ automatically estimated	12
ETS	Additive trend	Additive	12	Smoothing parameters automatically estimated	12
ARIMA	AR(1)	None	—	ARIMA(1,0,0) with intercept	12
Fourier-AR	Linear trend	Fourier periodic	12	Fourier coefficients+AR(1) error	12

The forecasting results of the nine candidate models are obtained, and MAE, RMSE, MAPE and sMAPE are calculated. The forecasting performance and ranking of each candidate model are illustrated in Figure 4.

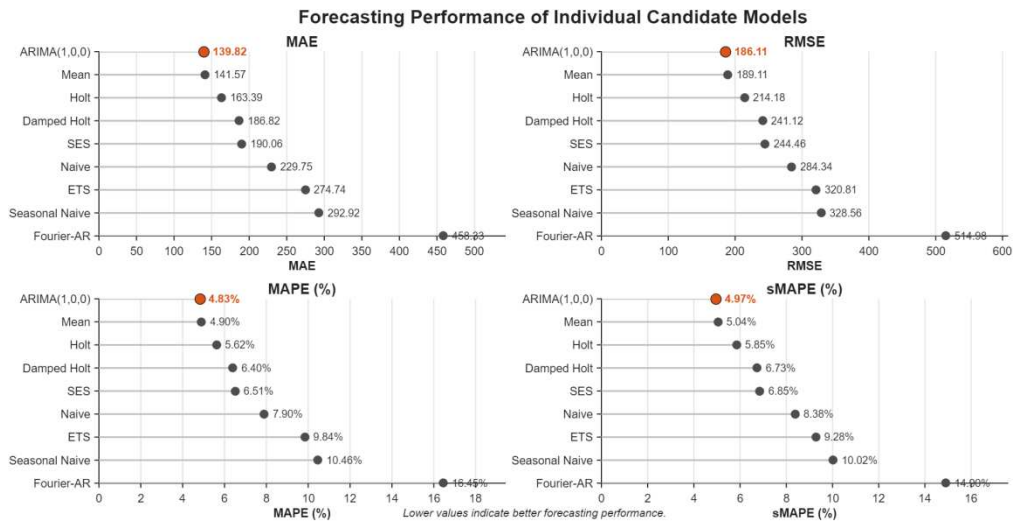


Figure 4. Forecasting results and comprehensive ranking of different methods

According to Figure 4, the ARIMA(1,0,0) model achieves the best performance among the nine single-candidate models, with MAE, RMSE, MAPE and sMAPE of 139.82, 186.11, 4.83% and 4.97%, respectively. The Mean model ranks second, with corresponding metrics of 141.57, 189.11, 4.90% and 5.04%. In contrast, the Fourier-AR model yields relatively high forecasting errors.

4.2 Grid-search based hyperparameter determination for the SRSE model

According to the preset parameter ranges, a joint grid search is performed over 135 hyperparameter combinations. The minimum CV-MAPE of rolling-origin validation on the training set is adopted as the selection criterion for the optimal parameter combination. The parameter values corresponding to the minimum CV-MAPE obtained from the search are:

$K=6, \gamma=0.9, \lambda=1, \eta=0.5$. Based on the optimal parameter settings, the recency-weighted error, forecasting stability and composite loss of each candidate model are calculated. The top-6 ranked models are screened for the final ensemble. The composite loss and combination weights of each selected model are presented in Table 3.

Table 3. Comprehensive evaluation and final combination weights of the screened base models

Model	Recency-weighted Error	Stability	Composite Loss	Final Weight
ARIMA(1,0,0)	0.1102	0.0816	0.1918	19.14%
Mean	0.1291	0.0866	0.2156	18.05%
SES	0.1471	0.1237	0.2707	16.11%
Naive	0.1496	0.1264	0.2760	15.95%
ETS	0.1546	0.1205	0.2751	15.98%
Damped Holt	0.1771	0.1454	0.3226	14.76%

The six finally-selected models are ARIMA(1,0,0), Mean, SES, Naive, ETS, and Damped Holt, with combination weights of 19.14%, 18.05%, 16.11%, 15.95%, 15.98%, and 14.76%, respectively. Among them, ARIMA(1,0,0) achieves the highest combination weight owing to its lowest recency-weighted error and composite loss. Based on the above combination weights, the prediction outputs of the six candidate models are weighted to obtain the SRSE forecasting results for monthly reported syphilis cases in Jiangsu Province in 2025, as shown in Figure 5.

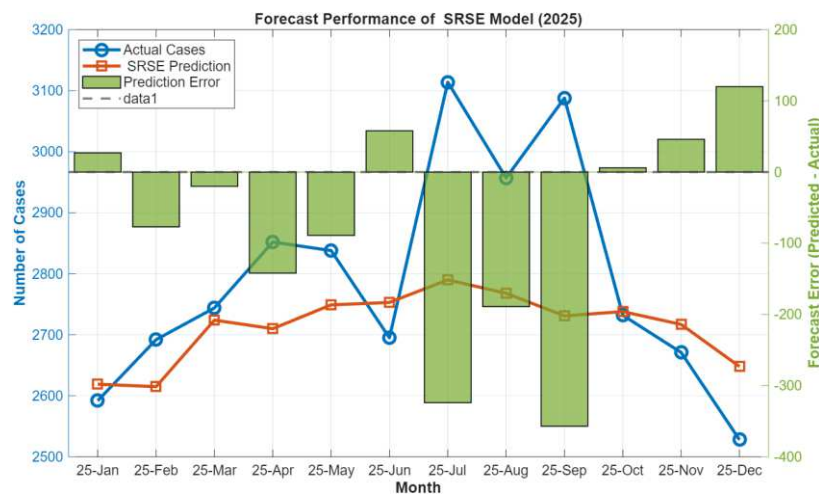


Figure 5. Prediction results and forecast errors of the SRSE model for the year 2025

Sensitivity analysis: Given the small sample size in this study, sensitivity analysis of key hyperparameters is performed to evaluate whether the prediction performance of the SRSE model is overly dependent on specific hyperparameter configurations. Adopting the single-factor control-variable method, the effects of the number of retained candidate models K , time-decay coefficient γ , stability penalty coefficient λ , and weight-adjustment parameter η on CV-MAPE are examined. While one hyperparameter is assessed, the remaining parameters are kept constant, and the rolling-origin validation procedure is repeated. Relatively stable validation errors within a reasonable parameter variation range indicate favorable model robustness with respect to the corresponding hyperparameter. Detailed settings are given below:

- (1) Fix $\gamma=0.9, \lambda=1, \eta=0.5$, vary $K=2, 3, 4, 5, 6$;
- (2) Fix $K=6, \gamma=0.9, \eta=0.5$, vary $\lambda=0, 0.5, 1$
- (3) Fix $K=6, \lambda=1, \eta=0.5$, vary $\gamma=0.8, 0.9, 1$

(4) Fix $K=6$, $\gamma=0.9$, $\lambda=1$, vary $\eta=0.5, 1, 2$

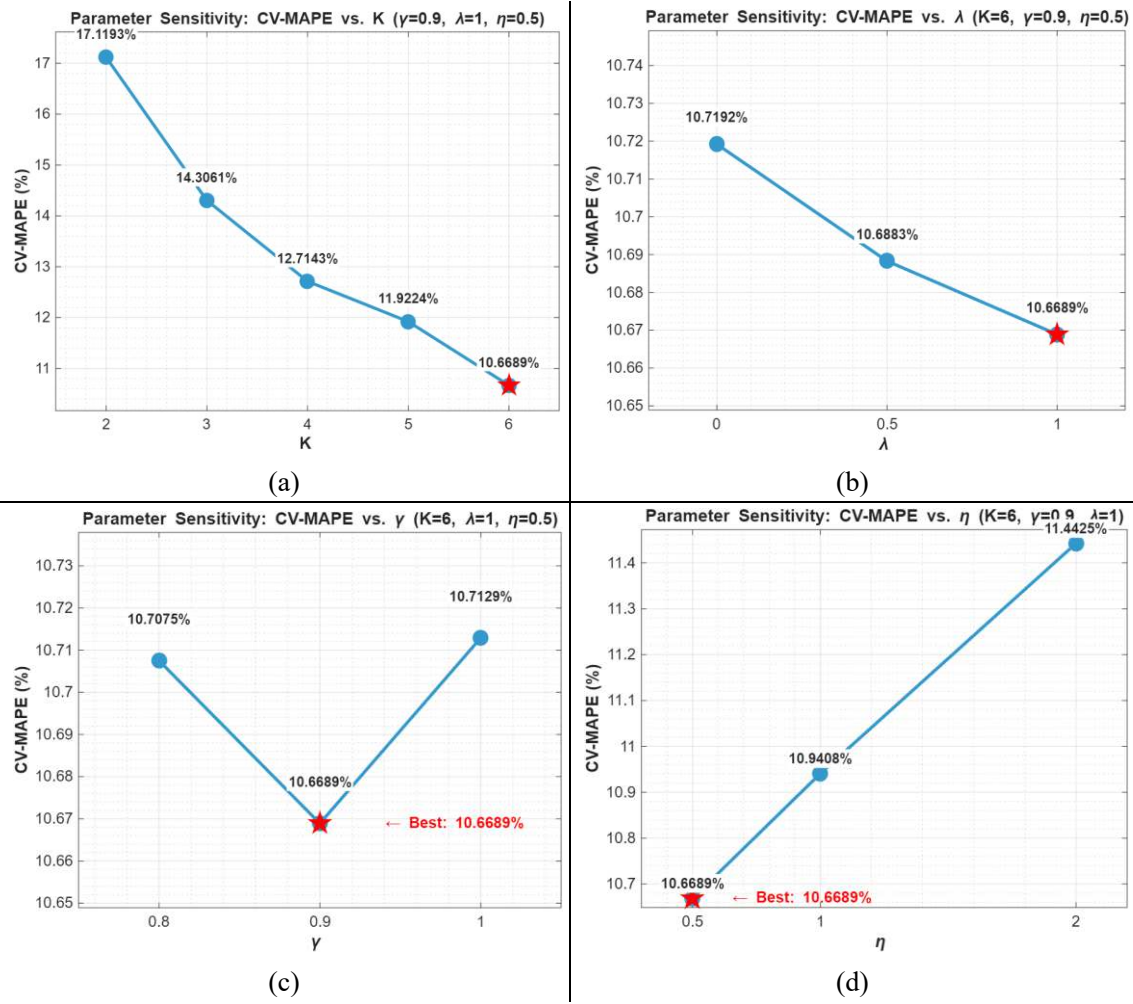


Figure 6. Sensitivity-analysis results of hyperparameters K , λ , γ , and η on CV-MAPE

From the four groups of results, parameter sensitivities are not identical. Overall, the model is most sensitive to K and exhibits considerable sensitivity to η , whereas it is relatively insensitive to variations in λ and γ . This indicates that the final parameter combination not only yields the lowest CV-MAPE, but also achieves favorable stability of partial parameters around their optimal values.

4.3 Hyperparameter determination of the SRSE model by intelligent optimization algorithms

To further improve the flexibility of hyperparameter search for the SRSE model, PSO, GA, and SSA are adopted for hyperparameter optimization. All algorithms are run under identical software-hardware environments and the same objective function, with the maximum number of iterations set to 100, the population size set to 30, and the search-space dimension equal to 4. The key parameter configurations of the three algorithms are presented in Table 4.

Table 4. Key parameter settings for five models

PSO	GA	SSA
Initial inertia weight: 0.9	Crossover probability: 0.8	Discoverer ratio: 20%
Final inertia weight: 0.4	Mutation probability: 0.1	Alert ratio: 10%
Inertial weight formula:	Elite reservation quantity: 2	Security threshold: 0.8
$\omega_t = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \frac{t}{T-1}$	Selection method: Tournament selection	

Cognitive learning factor $c_1=1.7$ Social learning factor $c_2=1.7$	Tournament size: 3 encoding method: Real-valued encoding	
common parameters Maximum number of iterations: 100 population size: 30 Search space dimension: 4		

To evaluate the repeat-run stability of stochastic intelligent optimization algorithms, each algorithm is run independently 30 times, and the optimal CV-MAPE obtained in each run is recorded. On this basis, statistical metrics including Best-CV-MAPE, Mean-CV-MAPE, standard deviation (SD), and Median are calculated. The results are presented in Table 5.

Table 5. Comparison of Stability for Three Stochastic Algorithms over 30 Runs

Method	Best CV-MAPE	Mean CV-MAPE	SD	Median
PSO	10.621904%	10.621904%	≈ 0	10.621904%
GA	10.621905%	10.621933%	0.000027	10.621922%
SSA	10.621954%	10.645796%	0.039391	10.622654%

Within the initial parameter search range, the best CV-MAPE values obtained by PSO, GA, and SSA are 10.621904%, 10.621905%, and 10.621954%, respectively, indicating that the optimal results of the three algorithms are very close. The SD of PSO across 30 independent runs is close to zero, GA also exhibits small fluctuations, whereas SSA yields a relatively higher SD. This demonstrates that PSO achieves favorable repeat-run stability under the current experimental settings.

The optimal parameters identified by PSO, GA and SSA are $\alpha = 6, \lambda = 1.000, \eta = 0.500$, with only the γ value differing: γ equals 0.9977744, 0.9977743, and 0.9977648 correspondingly. The optimal CV-MAPE values show minor discrepancies, which implies that all three algorithms are applicable for parameter searching in this study. Considering both the optimal CV-MAPE and the stability over 30 independent runs, the PSO algorithm is adopted for hyperparameter optimization of the SRSE model in subsequent analyses. Six candidate models are screened by PSO-SRSE for final ensemble, and the model composition and combination weights are shown in Figure 7.

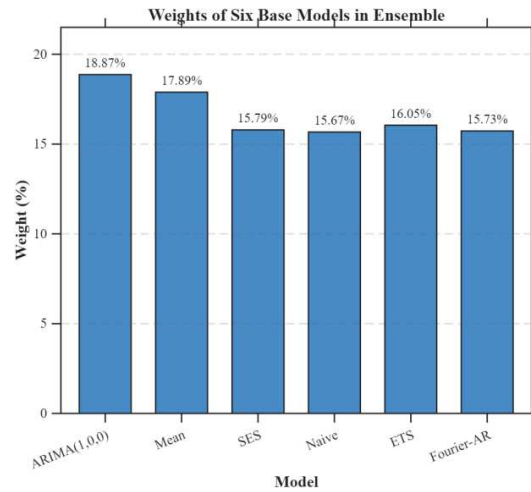


Figure 7. Weight distribution of six base models in the integrated framework

The prediction results of the PSO-SRSE model for the reported syphilis cases in Jiangsu Province from January to December 2025 are calculated according to the corresponding combination weights, and the results are shown in Figure 8.

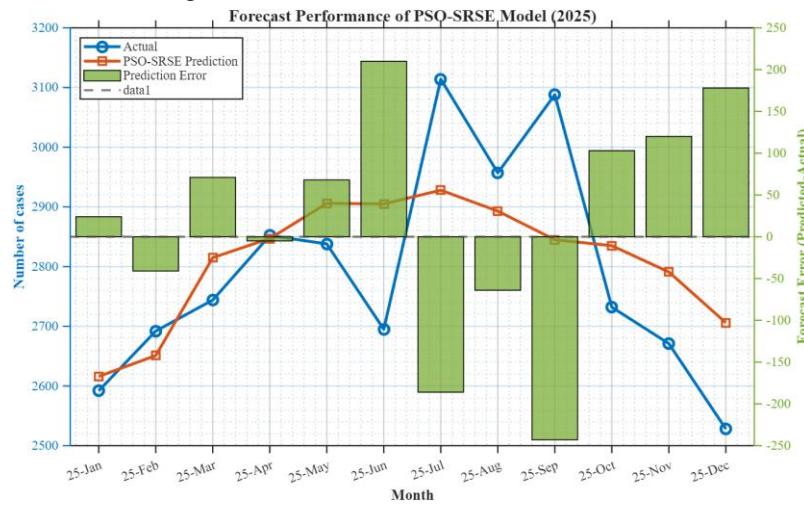


Figure 8. Prediction results and prediction errors of the PSO-SRSE model for 2025

4.4 Comparison of prediction performance among different models

To further evaluate the prediction performance of the proposed ensemble approach, the prediction results of the PSO-SRSE and SRSE models are compared with those of nine single candidate models on the 2025 independent test set. Meanwhile, an equal-weighted ensemble model consisting of the nine single candidate models ($w_m = 1/9$) is further constructed as a control group to verify the effectiveness of combination weights determined by the performance-based screening and weighting strategy. The prediction results are presented in Figure 9

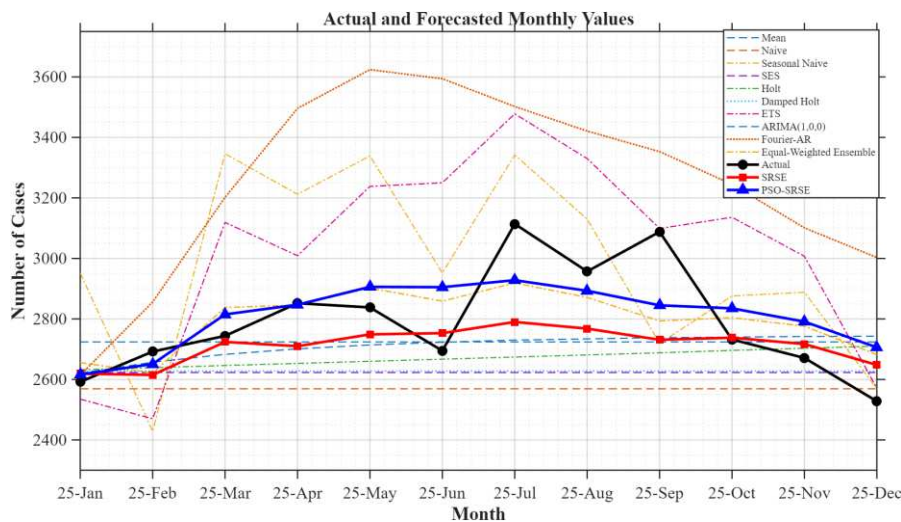


Figure9. Comparison of monthly actual values and forecasted values from different models

The prediction results of each model are substituted into evaluation formulas for calculating MAE, RMSE, MAPE and sMAPE, as illustrated in Figure 10.

Forecasting model	MAE		RMSE		MAPE		sMAPE		Rank
ARIMA(1,0,0)	139.82	4	186.11	4	4.83%	4	4.97%	4	12
Mean	141.57	5	189.11	5	4.90%	5	5.04%	5	11
Holt	163.39	6	214.18	6	5.62%	6	5.85%	6	10
Damped-Holt	186.82	7	241.12	7	6.40%	7	6.73%	7	9
SES	190.06	8	244.46	8	6.51%	8	6.85%	8	8
Naive	229.75	9	284.34	9	7.90%	9	8.38%	9	7
ETS	274.74	10	320.81	10	9.84%	10	9.28%	10	6
Seasonal Naive	292.92	11	328.56	11	10.46%	11	10.02%	11	5
Fourier-AR	458.33	12	514.98	12	16.45%	12	14.90%	12	4
Equal-Weighted Ensemble	114.03	2	135.71	2	4.03%	2	4.05%	2	3
RSE	121.12	3	163.99	3	4.16%	3	4.29%	3	2
PSO-SRSE	109.39	1	132.35	1	3.89%	1	3.87%	1	1

Figure 10. Comparison of evaluation metrics and ranks for different forecasting models

On the 2025 independent test set, the MAE, RMSE, MAPE and sMAPE of the PSO-SRSE model are 109.39, 132.35, 3.89% and 3.87%, respectively. All four metrics are the lowest among all compared models. The corresponding metrics of the equal-weighted ensemble model are 114.03, 135.71, 4.03% and 4.05%, while those of the SRSE model are 121.12, 163.99, 4.16% and 4.29%. Among the single candidate models, ARIMA(1,0,0) achieves the best performance with four metrics of 139.82, 186.11, 4.83% and 4.97%. The results reveal that PSO-SRSE achieves favorable overall prediction performance within the test dataset of this study, and its point-forecast errors are lower than those of the unoptimized SRSE, equal-weighted ensemble and all single candidate models.

4 Conclusion

This study constructs a SRSE for short-term disease surveillance time series. PSO is employed to jointly optimize the model retention number, time-decay coefficient, stability penalty coefficient, and weight-adjustment parameter. Model construction is implemented using monthly reported syphilis case data of Jiangsu Province from 2021 to 2024, and out-of-sample independent evaluation is performed on the 2025 dataset. The results show that the MAE, RMSE, MAPE and sMAPE of PSO-SRSE are 109.39, 132.35, 3.89% and 3.87%, respectively. Within the test period of this study, these metrics are lower than those of SRSE, each single candidate model, and the equal-weighted ensemble of all candidate models, indicating that the proposed method possesses certain application potential for short-term monthly case point-forecasting.

Nevertheless, this study has several limitations. Since the analysis is conducted merely on data from a single region with limited time-series length, external validation should be carried out across more regions, different diseases and longer time series in future work. Further assessments regarding different forecasting horizons and prediction uncertainty are also required.

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.

Author Contributions

“Conceptualization, S.Z. and X.S.; methodology, S.Z.; software, S.Z.; validation, S.Z. and X.S.; formal analysis, S.Z.; investigation, X.S.; resources, X.S.; data curation, X.S.; writing—original

draft preparation, S.Z.; writing—review and editing, S.Z.; visualization, X.S.; supervision, X.S.; project administration, S.Z.; funding acquisition, S.Z. All authors have read and agreed to the published version of the manuscript.”

Funding

This work was funded by Huai'an Basic Research Plan (Joint Special Project) Project (Grant no. HABL202515); the 2026 "Green-Blue Project" Program for Outstanding Young Backbone Teachers of Jiangsu Universities; Youth Fund Project of Jiangsu Vocational College of Electronics and Information (Grant no. JSEIZYB202403).

References

1. Peeling RW, Mabey D, Chen XS, Garcia PJ. **Syphilis**. *Lancet*. 2023;402(10398):336-346. doi:10.1016/S0140-6736(22)02348-0.
2. Zheng Y, Yu Q, Lin Y, et al. **Global burden and trends of sexually transmitted infections from 1990 to 2019: an observational trend study**. *Lancet Infect Dis*. 2022;22(4):541-551. doi:10.1016/S1473-3099(21)00448-5.
3. Li J, Yang Y, Huang B, Zeng J. **Epidemiological characteristics of syphilis in mainland China, 2004 to 2019**. *J Int Med Res*. 2024;52(6):03000605241258465. doi:10.1177/03000605241258465.
4. Zhang X, Zhang T, Pei J, Liu Y, Li X, Medrano-Gracia P. **Time Series Modelling of Syphilis Incidence in China from 2005 to 2012**. *PLoS One*. 2016;11(2):e0149401. doi:10.1371/journal.pone.0149401.
5. Zhang F, Li Y, Li X, Zhang B, Xue C, Wang Y. **Comparison of ARIMA and Bayesian Structural Time Series Models for Predicting the Trend of Syphilis Epidemic in Jiangsu Province**. *Infect Drug Resist*. 2024;17:5745-5754. doi:10.2147/IDR.S462998.
6. Zhu Z, Zhu X, Zhan Y, Gu L, Chen L, Li X. **Development and comparison of predictive models for sexually transmitted diseases—AIDS, gonorrhea, and syphilis in China, 2011–2021**. *Front Public Health*. 2022;10:966813. doi:10.3389/fpubh.2022.966813.
7. Wang ZD, Yang CX, Zhang SK, Wang YB, Xu Z, Feng ZJ. **Analysis and forecasting of syphilis trends in mainland China based on hybrid time series models**. *Epidemiol Infect*. 2024;152:e93. doi:10.1017/S0950268824000694.
8. Makridakis S, Hibon M. **The M3-Competition: results, conclusions and implications**. *Int J Forecast*. 2000;16(4):451-476. doi:10.1016/S0169-2070(00)00057-1.
9. Muth JF. **Optimal properties of exponentially weighted forecasts**. *J Am Stat Assoc*. 1960;55(290):299-306. doi:10.1080/01621459.1960.10482064.
10. Holt CC. **Forecasting seasonals and trends by exponentially weighted moving averages**. *Int J Forecast*. 2004;20(1):5-10. doi:10.1016/j.ijforecast.2003.09.015.
11. Gardner ES Jr, McKenzie E. **Forecasting trends in time series**. *Manage Sci*. 1985;31(10):1237-1246. doi:10.1287/mnsc.31.10.1237.
12. Hyndman RJ, Koehler AB, Snyder RD, Grose S. **A state space framework for automatic forecasting using exponential smoothing methods**. *Int J Forecast*. 2002;18(3):439-454. doi:10.1016/S0169-2070(01)00110-8.
13. Hyndman RJ, Khandakar Y. **Automatic time series forecasting: the forecast package for R**. *J Stat Softw*. 2008;27(3):1-22. doi:10.18637/jss.v027.i03.

14. Young PC, Pedregal DJ, Tych W. **Dynamic harmonic regression.** *J Forecast.* 1999;18(6):369-394.
doi:10.1002/(SICI)1099-131X(199911)18:6<369::AID-FOR748>3.0.CO;2-K.
15. Lu J, Meyer S. **Forecasting Flu Activity in the United States: Benchmarking an Endemic-Epidemic Beta Model.** *Int J Environ Res Public Health.* 2020;17(4):1381.
doi:10.3390/ijerph17041381.
16. Monks T, Harper A, Allen M, Collins L, Mayne A. **Forecasting the daily demand for emergency medical ambulances in England and Wales: a benchmark model and external validation.** *BMC Med Inform Decis Mak.* 2023;23:117. doi:10.1186/s12911-023-02218-z.
17. Tashman LJ. **Out-of-sample tests of forecasting accuracy: an analysis and review.** *Int J Forecast.* 2000;16(4):437-450. doi:10.1016/S0169-2070(00)00065-0.
18. Hewamalage H, Ackermann K, Bergmeir C. **Forecast evaluation for data scientists: common pitfalls and best practices.** *Data Min Knowl Discov.* 2023;37(2):788-832.
doi:10.1007/s10618-022-00894-5.
19. Wang X, Hyndman RJ, Li F, Kang Y. **Forecast combinations: an over 50-year review.** *Int J Forecast.* 2023;39(4):1518-1547. doi:10.1016/j.ijforecast.2022.11.005.
20. Fox SJ, Kim M, Meyers LA, Reich NG, Ray EL. **Optimizing Disease Outbreak Forecast Ensembles.** *Emerg Infect Dis.* 2024;30(9):1967-1969. doi:10.3201/eid3009.240026.
21. Sun Y, Hong Y, Wang S, Zhang X. **Penalized time-varying model averaging.** *J Econom.* 2023;235(2):1355-1377. doi:10.1016/j.jeconom.2022.09.007.
22. Poli R, Kennedy J, Blackwell T. **Particle swarm optimization.** *Swarm Intell.* 2007;1(1):33-57. doi:10.1007/s11721-007-0002-0.
23. Srinivas M, Patnaik LM. **Genetic algorithms: a survey.** *Computer.* 1994;27(6):17-26.
doi:10.1109/2.294849.
24. Xue J, Shen B. **A novel swarm intelligence optimization approach: sparrow search algorithm.** *Syst Sci Control Eng.* 2020;8(1):22-34. doi:10.1080/21642583.2019.1708830.