Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Characterizing regional heterogeneity and optimal warning timing of brucellosis in China's five highest-incidence provinces: A comparative evaluation of time-series models with multi- source environmental predictors

Yawen Qin
Tianjin Medical University
Qianran Gao
Tianjin Medical University
Haoran Liu
Tianjin Medical University
Hanjing FAN
Tianjin Medical University
Qingqing Wang
Tianjin Medical University
Wenyi Zhang
China Medical University
Changping Li
Tianjin Medical University
Qiulan Chen
Chinese Center For Disease Control and Prevention
Zhuang Cui
zhuangcui@tmu.edu.cn

Tianjin Medical University

Research Article

Keywords: Brucellosis, Time-series modeling, LSTM, SARIMAX, Regional heterogeneity, Warning timing, Environmental drivers

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11203108/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Brucellosis is a severe zoonotic disease with pronounced seasonality and regional heterogeneity in high-incidence areas of China. While the seasonality of brucellosis is well established, the extent to which environmental drivers and optimal warning timing vary across high-incidence regions remains poorly quantified. Reliable tools are needed to characterize this regional heterogeneity and inform region-specific prevention strategies.

Methods

We collected monthly brucellosis incidence and 17 environmental variables from 2014 to 2024 across five high-incidence provinces: Inner Mongolia, Xinjiang, Shanxi, Heilongjiang, and Hebei. A three-step procedure—cross-correlation analysis, multicollinearity diagnostics, and stepwise regression—was used to select exogenous predictors. We then compared four time-series models: seasonal autoregressive integrated moving average (SARIMA), SARIMA with exogenous variables (SARIMAX), long short-term memory (LSTM), and LSTM with exogenous variables (LSTMX).

Results

All five provinces showed a unimodal seasonal pattern with peaks between April and July. However, the environmental drivers and optimal lag periods varied substantially by region, ranging from 1 to 6 months. In capturing region-specific temporal patterns, LSTM achieved the highest accuracy in Shanxi (R^2=0.925), Hebei (R^2=0.876), and Xinjiang (R^2=0.829), outperforming SARIMA and SARIMAX. LSTMX performed best in Inner Mongolia (R^2=0.759) and Heilongjiang (R^2=0.772) but showed weaker performance than LSTM in Shanxi and Hebei. The optimal environmental predictors and lag periods differed substantially across provinces: soil moisture at lag 4 in Inner Mongolia; ozone at lag 6, soil moisture at lag 1, NO_2 at lag 1, and precipitation at lag 2 in Xinjiang; ozone at lag 6 and wind speed at lag 1 in Shanxi; NDVI at lag 2, PM_{10} at lag 6, and wind speed at lag 6 in Heilongjiang; and ozone at lag 5, SO_2 at lag 1, and wind speed at lag 4 in Hebei. These lag periods directly determine the maximum warning lead time achievable in each region, ranging from 4 months in Inner Mongolia to 6 months in Xinjiang, Shanxi, and Heilongjiang, and 5 months in Hebei. Overall, adding exogenous variables did not consistently improve performance across provinces, indicating that the value of environmental predictors is region-dependent.

Conclusions

Our findings demonstrate that while LSTM-based models effectively capture region-specific temporal patterns of brucellosis, the added value of environmental predictors is region-dependent. These results support the development of tailored early warning systems that account for regional heterogeneity in environmental drivers and lag periods, rather than a one-size-fits-all forecasting approach.

1.Background

Brucellosis is a zoonotic bacterial infection caused by *Brucella* species[1], mainly infecting domestic animals such as cattle, sheep, and goats. The species most commonly associated with human disease are *B. melitensis*, *B. abortus*, *B. suis*, and *B. canis*[1,2]. Humans acquire infection through direct contact with infected animals, consumption of contaminated animal products (especially unpasteurized milk or cheese), or inhalation of infectious aerosols. Clinical presentation is nonspecific, including fever, sweats, fatigue, headache, and arthralgia, making it difficult to distinguish from other febrile illnesses[1,2]. Symptoms may persist for weeks to months, with complications such as endocarditis and neurological disorders[3]. Chronic brucellosis is defined as symptom persistence for more than twelve months[4,5].

Brucellosis is among the most widespread zoonotic diseases worldwide. Recent estimates suggest that annual global incidence may reach 1.6–2.1 million cases, indicating that the disease burden has likely been underestimated[6]. The disease remains persistently endemic across Asia and Africa, with the highest incidence rates reported in Syria, Kyrgyzstan, Mongolia, Iran, Algeria, and Kenya. Asia bears the heaviest burden, particularly in East, Central, and West Asian countries. Over the past decade, the epidemiological range has

expanded markedly, with affected countries increasing from 53 to at least 97[7]. Expansion of animal industries, urbanization, and inadequate hygienic measures in animal husbandry and food handling contribute to this sustained public health hazard[1].

Brucellosis remains an urgent public health challenge in China. Nationally, human brucellosis incidence increased overall between 2005 and 2021[8], declined during 2015–2018, and rebounded sharply after 2019[9]. Marked heterogeneity was observed across regions. Inner Mongolia bore the highest burden, with 87,961 cumulative cases and an average annual incidence of 71.88 per 100,000 population, ranking first nationwide for ten consecutive years. Xinjiang reported 27,845 cases, with the fastest growth among highest-incidence regions (average annual increase of 16.76%). Shanxi documented 21,932 cases (annual increase of 9.61%), Hebei reported 19,508 cases (annual increase of 4.63%), and Heilongjiang reported 17,637 cases (annual decrease of 0.95%). Nationwide, reported cases peaked in 2021 at more than 71,000 and remained above 69,000 through 2023[9].

During the past decade, machine learning techniques have been widely used in public health research to identify relationships between disease outcomes and environmental variables[10]. Time-series approaches, including autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) models, have also been applied to infectious disease forecasting in China[11,12]. A study on brucellosis in Xinjiang found that SARIMA outperformed nonlinear autoregressive neural network (NARNN) models, suggesting that seasonal time-series models perform well when brucellosis shows a clear periodic pattern, although the lack of environmental and policy data limited further explanation[13]. To address this limitation, the SARIMA model with exogenous variables (SARIMAX) allows time-variant environmental factors to be incorporated alongside temporal autocorrelation. SARIMAX has been successfully used to model hemorrhagic fever with renal syndrome in China[12], and a study on hepatitis C prediction showed that incorporating Baidu search index variables reduced forecasting errors compared with SARIMA alone[14].

Recently, long short-term memory (LSTM), a recurrent neural network (RNN)-based time-series model, has shown strong forecasting ability, including for COVID-19[15,16]. LSTM retains information from previous time steps, making it suitable for temporally correlated data. Its gated structure addresses the vanishing/exploding gradient problems of traditional RNNs[17,18]. In a tuberculosis study in Liaoning, Yang et al. found that a multivariate multi-step LSTM model reduced four error measures by 12.92–14.81% compared with ARIMA[19].

For brucellosis, existing prediction studies have mainly relied on traditional time-series or neural network models using historical incidence data alone. Given that the seasonality of brucellosis is well established, the key scientific question is not whether models can reproduce seasonal patterns, but rather: (1) to what extent do environmental drivers and their optimal lag periods vary across high-incidence regions; and (2) whether incorporating region-specific environmental predictors improves the characterization of local transmission dynamics. The performance of LSTM-based models for capturing such region-specific patterns in high-risk areas of China has not been systematically explored, nor has the added value of LSTM with exogenous variables (LSTMX), which incorporates screened multi-source environmental predictors. This gap is significant because brucellosis transmission is closely linked to livestock exposure, seasonal human activity, ecological conditions, and environmental variation. been systematically explored, nor has the added value of LSTM with exogenous variables (LSTMX), which incorporates screened multi-source environmental predictors. This gap is significant because brucellosis transmission is closely linked to livestock exposure, seasonal human activity, ecological conditions, and environmental variation.

Therefore, this study aims to evaluate four time-series approaches (SARIMA, SARIMAX, LSTM, and LSTMX) for characterizing region-specific temporal patterns of brucellosis in five highest-incidence provinces (Inner Mongolia, Xinjiang, Shanxi, Heilongjiang, and Hebei) using data from 2014 to 2024, and to identify region-specific environmental drivers, optimal lag periods, and warning timing that can inform tailored prevention strategies.

2.Materials and methods

2.1 Data of brucellosis

The time series of the monthly incidence of brucellosis in the top five provinces with the highest national brucellosis incidence in China, including Inner Mongolia Autonomous Region, Xinjiang Uygur Autonomous Region, Shanxi Province, Heilongjiang Province and Hebei Province, from January 2014 to December 2024 was extracted from the China Information System for Disease Control and

Prevention. All cases were diagnosed according to the criteria established by the National Health Commission of China. There were no missing values in this data set.

2.2 Meteorological data

All monthly meteorological and environmental datasets (2014–2024) for the five studied provinces were collected from authoritative databases. Mean temperature, maximum temperature, minimum temperature, precipitation, and Fractional Vegetation Cover (FVC) were obtained from the National Tibetan Plateau Data Center. Solar radiation, soil moisture, and wind speed were derived from the TerraClimate dataset. Normalized Difference Vegetation Index (NDVI) was acquired from the NASA MOD13A3 dataset. $PM_{2.5}$, PM_{10}, ozone, nitrogen dioxide, carbon monoxide, and sulfur dioxide were extracted from the ChinaHighAirPollutants (CHAP) dataset. Relative humidity and atmospheric pressure were obtained from a high-resolution meteorological forcing dataset published in *Scientific Data*. Detailed sources are provided in Table S1.

2.3 Variable Selection

A three-step procedure was established to screen exogenous meteorological variables. First, cross-correlation function (CCF) analysis with 12-order seasonal differencing identified optimal lagged relationships between each meteorological factor and monthly brucellosis incidence; variables with statistically significant positive lags were retained, and the lag period with the strongest correlation was selected. Second, variance inflation factor (VIF) testing was applied to the lagged variables with a threshold of VIF < 5, and an iterative stepwise elimination approach sequentially removed variables with excessive VIF values and the weakest correlation with brucellosis incidence. Third, forward stepwise regression with the Akaike Information Criterion (AIC) was employed on the VIF-retained candidate set, using log-transformed monthly brucellosis incidence as the dependent variable; the variable combination that minimized AIC and achieved optimal model fitting was reserved as the final exogenous predictors for multivariate models.

2.4 Modeling of SARIMA and SARIMAX

The ARIMA model is a widely applied time-series forecasting method with favorable prediction accuracy[20]. The SARIMA model is an extended version of the ARIMA model that takes seasonal fluctuations and seasonal trends into account. By incorporating seasonal autoregressive (AR) terms, differencing terms, and moving average (MA) terms, the SARIMA model can accurately forecast time-series data under the assumption of stable temporal trends[21].

The SARIMA model is denoted as SARIMA(p,d,q)(P,D,Q)s, where p, d, q are non-seasonal orders (autoregressive, differencing, moving average), P, D, Q are seasonal orders, and s is the seasonal period (s = 12 in this study). The ARIMAX model integrates exogenous variables to improve forecasting performance compared with conventional ARIMA[22]. In this study, we constructed SARIMAX models based on the SARIMA framework to quantify associations between meteorological factors and brucellosis incidence. By introducing exogenous regressors, these models capture inherent seasonal fluctuations while quantifying the driving effects of multi-source environmental variables (e.g., atmospheric pressure and precipitation) on disease risk.

2.5 Constructing the LSTM model

LSTM is a deep learning-based time-series model that uses three gating mechanisms (input, forget, and output gates) to capture long-term dependencies while avoiding vanishing/exploding gradients, making it suitable for monthly brucellosis incidence prediction. LSTM and LSTMX models were built using Python and TensorFlow, with LSTMX incorporating lagged meteorological factors identified through CCF, VIF, and stepwise regression in R. Log transformation was applied to incidence data to reduce skewness, and Min-Max normalization scaled all variables to [0,1]. To ensure reproducibility, a fixed random seed was set, GPU disabled, and TensorFlow deterministic mode activated. The last 12 months of data per province served as the independent test set, with the remaining data for training, and a 3-fold time-series cross-validation (chronological order) was used for hyperparameter optimization. Input samples were constructed using a sliding window, and a grid search optimized window length, number of LSTM neurons, and feature combinations, with the optimal model for each province selected based on minimum Root Mean Square Error under cross-validation. A single-layer LSTM architecture (input layer, LSTM hidden layer, Dropout, fully connected output layer) was adopted, using the Adam optimizer with MSE loss, while an early stopping mechanism monitored training loss and restored optimal weights, which together with Dropout mitigated overfitting and improved generalization.

2.6 Evaluation and comparison of models

Four evaluation metrics widely used in public health time-series forecasting—Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R^2)—were adopted to quantify model performance

on both training and test datasets. Smaller RMSE, MAE, and MAPE values indicate higher prediction accuracy, while R^2 closer to 1 reflects stronger explanatory power for data variation and better predictive performance.

2.7 Data analysis

In this study, R version 3.4.1 was used for variable selection and the development of SARIMA and SARIMAX models. Python was utilized to construct the LSTM and LSTMX models. The statistical significance level was set at $\alpha = 0.05$.

3.Results

3.1 Statistical description

This study analyzed the five highest-incidence provinces in China from 2014 to 2024: Inner Mongolia (144,908 cases), Xinjiang (63,229), Shanxi (52,832), Heilongjiang (45,238), and Hebei (44,898). These five provinces accounted for the vast majority of national cases, representing the core endemic areas. Annual average cases varied significantly: Inner Mongolia (1097.79 ± 556.23), Xinjiang (479.01 ± 288.28), Shanxi (400.24 ± 205.75), Heilongjiang (342.71 ± 156.25), and Hebei (340.14 ± 159.97). The data showed positive-skewed distributions with discrete high-incidence years.

Temporal trends are shown in Fig. 1 and Figures S1–S5. All provinces exhibited prominent seasonal patterns, with annual peaks between April and July and troughs from November to February. However, the magnitude, timing, and environmental correlates of these seasonal peaks varied substantially across provinces, underscoring the need for region-specific characterization rather than a uniform forecasting approach. Inter-provincial heterogeneity was evident: Inner Mongolia maintained the highest incidence with persistent seasonal oscillations but no clear monotonic trend; Xinjiang showed moderate fluctuations on a stable baseline; Shanxi displayed an initially high incidence followed by gradual decline; Heilongjiang exhibited small but consistent seasonal fluctuations; and Hebei had the lowest incidence with mild, stable seasonal variations. These divergent temporal trends reflect differences in livestock industry scale, vaccination and quarantine measures, and socioeconomic conditions across regions.

Meteorological variables are summarized in Table S2. CCF lag correlation analysis revealed that multiple meteorological factors had significant time-lag correlations with brucellosis cases, with lag effects varying by province. This regional heterogeneity underscores the need to consider province-specific predictors and lag effects when constructing models to improve accuracy and applicability. Notably, the optimal lag periods varied from 1 to 6 months across provinces, directly determining the maximum warning lead time achievable in each region. A summary of regional heterogeneity across the five provinces—including annual average cases, temporal trends, selected environmental variables and their lag periods, optimal models, and the effect of adding environmental variables—is presented in Table 1.

Table 1
Summary of regional heterogeneity in brucellosis across five high-incidence provinces of China

Province	Annual average cases	Temporal trend	Selected environmental variables (lag)	Best model	Effect of adding environmental variables
Inner Mongolia	1097.79	Persistent high incidence, no monotonic trend	Soil moisture(4)	LSTMX	Significant improvement (R^2 0.141→0.759)
Xinjiang	479.01	Moderate fluctuations on a stable baseline	Ozone(6), soil moisture(1), NO_2(1), precipitation(2)	LSTMX	Slight improvement (0.829→0.839)
Shanxi	400.24	Initially high, followed by gradual decline	Ozone(6), wind speed(1)	LSTM	Significant deterioration (0.925→0.512)
Heilongjiang	342.71	Small but consistent seasonal fluctuations	NDVI(2), PM_{10}(6), wind speed(6)	LSTMX	Improvement (0.718→0.772)
Hebei	340.14	Lowest and stable	Ozone(5), SO_2(1), wind speed(4)	LSTM	Deterioration (0.876→0.815)

3.2 Results of Variable Screening

This study applied CCF lag correlation analysis, VIF multicollinearity diagnostics, and AIC stepwise regression to screen environmental factors with significant correlation, no multicollinearity, and optimal AIC values. Cross-correlation heatmaps (Figure S6-S10) visually present the temporal correlation patterns across the five provinces, and Table S3 summarizes the maximum correlation coefficients, optimal lag orders, and VIF values. The correlation magnitude and lag patterns varied substantially across regions.

After systematic screening, the final exogenous covariates were determined: for Inner Mongolia, soil moisture at lag 4; for Xinjiang, ozone at lag 6, soil moisture at lag 1, NO_2 at lag 1, and precipitation at lag 2; for Shanxi, ozone at lag 6 and wind velocity at lag 1; for Heilongjiang, NDVI at lag 2, PM_{10} at lag 6, and wind velocity at lag 6; for Hebei, ozone at lag 5, SO_2 at lag 1, and wind velocity at lag 4. These lagged variables were used as exogenous inputs for SARIMAX and LSTMX models. Notably, the optimal lag periods varied from 1 to 6 months across provinces, directly determining the maximum warning lead time achievable in each region.

Figure S10-S15 display the time series of monthly brucellosis cases and selected lagged indicators (standardized by z-score). Distinct regional relationships were observed: in Inner Mongolia, cases were positively correlated with lagged soil moisture; in Xinjiang, negatively correlated with ozone at lag 6 and positively with soil moisture, NO_2, and precipitation; in Shanxi, ozone showed negative lagged associations while wind velocity showed positive correlation; in Heilongjiang, NDVI, PM_{10}, and wind velocity all showed positive lagged trends; in Hebei, lagged ozone was negatively correlated while SO_2 and wind velocity showed positive correlations. These temporal trends were consistent with the CCF correlation coefficients.

3.3 Performance of SARIMA and SARIMAX Models in Predicting Brucellosis Incidence

Various SARIMA and SARIMAX models were established for brucellosis forecasting in five provinces. All models adopted a 12-month seasonal cycle. Performance metrics (AIC, R^2, RMSE, MAE, MAPE) during the fitting and forecasting stages are presented in Table 2, and the optimal model structures are presented in Table S4.

At the forecasting stage, model performance varied significantly across provinces. In Inner Mongolia, SARIMAX showed better accuracy than SARIMA, with R^2 increasing from 0.141 to 0.622, and RMSE, MAE, and MAPE decreasing from 307.437 to 203.875, 269.121 to 174.367, and 22.779% to 14.393%, respectively. In contrast, SARIMAX performed worse than SARIMA in Xinjiang, Shanxi, and Heilongjiang. In Xinjiang, R^2 decreased from 0.744 to 0.562, while RMSE, MAE, and MAPE increased. In Shanxi and Heilongjiang, SARIMAX also showed lower R^2 (0.446 vs. 0.577; 0.503 vs. 0.732) and higher errors. For Hebei, SARIMAX showed marginal improvement in R^2 (0.864 vs. 0.851) and RMSE (45.490 vs. 47.548), but slightly higher MAE (38.189 vs. 40.821) and MAPE (15.374% vs. 13.614%).

Table 2
Model performance metrics for brucellosis forecasting in five Chinese provinces

Province	Model	Model fitting					Model forecasting			
		AIC	R2	RMSE	MAE	MAPE(%)	R2	RMSE	MAE	MAPE(%)
Inner Mongolia	SARIMA	-100.94	0.941	140.633	90.604	8.802	0.141	307.437	269.121	22.779
	SARIMAX	-107.71	0.939	143.525	93.674	8.777	0.622	203.875	174.367	14.393
	LSTM	-	0.884	203.373	151.323	15.677	0.187	299.085	258.654	22.479
	LSTMX	-	0.885	203.116	157.682	14.740	0.759	162.804	123.090	11.216
Xinjiang	SARIMA	1194.09	0.916	81.268	57.441	15.963	0.744	129.965	112.728	22.473
	SARIMAX	1342.15	0.913	82.612	63.179	17.074	0.562	169.969	142.345	29.805
	LSTM	-	0.824	112.351	82.499	21.888	0.829	106.226	86.373	13.214
	LSTMX	-	0.829	110.971	77.058	21.377	0.839	103.028	83.737	14.847
Shanxi	SARIMA	1104.38	0.917	50.946	38.312	12.173	0.577	91.187	82.549	27.539
	SARIMAX	1224.67	0.923	48.902	38.409	10.924	0.446	104.327	91.365	31.897
	LSTM	-	0.862	54.035	43.166	13.254	0.925	38.384	29.548	9.377
	LSTMX	-	0.893	47.575	38.265	12.418	0.512	97.919	82.352	21.341
Heilongjiang	SARIMA	1111.13	0.890	48.693	34.421	12.356	0.732	43.932	36.311	15.280
	SARIMAX	1253.11	0.868	53.285	41.671	15.037	0.503	59.850	54.944	27.032
	LSTM	-	0.812	64.328	47.751	15.993	0.718	45.085	36.746	18.757
	LSTMX	-	0.834	60.501	45.368	15.147	0.772	40.489	29.910	12.925
Hebei	SARIMA	1111.39	0.886	49.353	37.510	13.858	0.851	47.548	40.821	13.614
	SARIMAX	1249.76	0.864	53.846	40.709	14.041	0.864	45.490	38.189	15.374
	LSTM	-	0.839	50.143	38.830	14.249	0.876	43.347	33.145	12.494
	LSTMX	-	0.854	47.753	37.141	13.279	0.815	52.994	38.336	13.616

Note: SARIMAX/LSTMX incorporate screened lagged environmental predictors:

Inner Mongolia: SARIMAX (4-month lag soil moisture), LSTMX (4-month lag soil moisture); Xinjiang: SARIMAX (6-month lag O_3, 1-month lag soil moisture, 1-month lag NO_2, 2-month lag precipitation), LSTMX (1-month lag NO_2, 2-month lag precipitation); Shanxi: SARIMAX (6-month lag O_3, 1-month lag wind speed), LSTMX (6-month lag O_3); Heilongjiang: SARIMAX (2-month lag NDVI, 6-month lag PM_{10}, 6-month lag wind speed), LSTMX (2-month lag NDVI, 6-month lag PM_{10}); Hebei: SARIMAX (5-month lag O_3, 1-month lag SO_2, 4-month lag wind speed), LSTMX (5-month lag O_3).

3.4 Prediction Performance of LSTM and LSTMX Models for Brucellosis Incidence

A global grid search optimized the sliding window length, number of LSTM hidden layer neurons, and meteorological feature combinations for each province based on minimum RMSE under cross-validation. Performance metrics (R^2, RMSE, MAE, MAPE) for LSTM and LSTMX during fitting and forecasting are summarized in Table 2.

Notable variations in exogenous variables were identified across provinces. For Inner Mongolia, LSTMX incorporated soil moisture (lag 4). In Xinjiang, it included NO_2 (lag 1) and precipitation (lag 2). Shanxi's LSTMX used O_3 (lag 6). Heilongjiang's LSTMX integrated NDVI (lag 2) and PM_{10} (lag 6). Hebei's LSTMX adopted O_3 (lag 5). Optimal sliding window length, neuron number, and batch size also differed by province, reflecting regional variations in brucellosis patterns and heterogeneous meteorological effects.

Marked provincial differences were observed in the ability of models to capture region-specific temporal patterns. In Shanxi, LSTM achieved the best performance among all models (R^2=0.925), significantly higher than SARIMA (R^2=0.577) and SARIMAX (R^2=0.446), with the lowest errors (RMSE = 38.384, MAE = 29.548, MAPE = 9.377%). In Hebei, LSTM (R^2=0.876) also outperformed SARIMA (R^2=0.851) and SARIMAX (R^2=0.864). In Xinjiang, both LSTM (R^2=0.829) and LSTMX (R^2=0.839) outperformed SARIMAX (R^2=0.562). In Inner Mongolia, LSTMX performed best (R^2=0.759), higher than SARIMA (R^2=0.141) and SARIMAX (R^2=0.622), with the lowest errors (RMSE = 162.804, MAE = 123.090, MAPE = 11.216%). Conversely, LSTMX underperformed relative to LSTM in Shanxi (R^2 dropped from 0.925 to 0.512) and Hebei (R^2 dropped from 0.876 to 0.815). In Heilongjiang, LSTMX slightly outperformed LSTM (R^2: 0.772 vs. 0.718), with lower errors (RMSE: 40.489 vs. 45.085; MAE: 29.910 vs. 36.746; MAPE: 12.925% vs. 18.757%). Notably, the optimal environmental predictors and lag periods differed substantially across provinces (soil moisture at lag 4 for Inner Mongolia; ozone at lag 6, soil moisture at lag 1, NO_2 at lag 1, and precipitation at lag 2 for Xinjiang; ozone at lag 6 and wind speed at lag 1 for Shanxi; NDVI at lag 2, PM_{10} at lag 6, and wind speed at lag 6 for Heilongjiang; ozone at lag 5, SO_2 at lag 1, and wind speed at lag 4 for Hebei), indicating that region-specific environmental drivers, rather than a universal set of predictors, determine model performance.

3.5 The Comparisons of SARIMA,SARIMAX,LSTM and LSTMX Estimations

As shown in Fig. 2, the fitting and prediction results are separated by a vertical dashed line. The grey, orange, red, green, and purple lines represent actual cases, SARIMA, SARIMAX, LSTM, and LSTMX estimates, respectively. Table S5 presents the monthly predicted values for each province.

The main findings are as follows: (1) LSTM and LSTMX use a fixed 12-month sliding window, so their fitting curves start later (around 2015) than SARIMA/SARIMAX (2014). (2) LSTM and LSTMX produce smoother estimates with smaller local fluctuations and better capture long-term seasonal trends, while SARIMA/SARIMAX are more sensitive to short-term monthly fluctuations. All five provinces exhibit a consistent unimodal pattern: cases increase from January to May, peak between May and July, then decline from August to December, with no secondary peak. (3) SARIMA and SARIMAX tend to overestimate peak incidence, whereas LSTM and LSTMX provide more conservative and accurate peak estimates. During the low-incidence period (September–December), all four models capture low-level characteristics well, but SARIMA/SARIMAX show slight under- or overestimation in certain months, while LSTM/LSTMX yield more stable estimates with smaller biases. The multivariate LSTMX generally achieves better fitting accuracy than univariate models (SARIMA and LSTM) in regions with strong environmental drivers (e.g., Inner Mongolia, Heilongjiang), but not in regions where environmental signals are weak or complex (e.g., Shanxi, Hebei). This finding underscores that the value of environmental predictors is region-dependent. Combining peak timing (May–July) with environmental lag periods (1–6 months), the warning window differs by province: Xinjiang, Shanxi, and Heilongjiang can issue warnings as early as November of the previous year; Inner Mongolia in January; Hebei in December of the previous year.

4.Discussion

4.1 Spatiotemporal distribution of human brucellosis

Brucellosis is a severe zoonotic disease with pronounced seasonal fluctuations and regional clustering, remaining a priority for prevention and control in China[23]. The five highest-incidence provinces—Inner Mongolia, Xinjiang, Shanxi, Heilongjiang, and Hebei—collectively determine the national epidemic trend. Between 2014 and 2024, these five provinces accounted for the vast majority of reported cases nationwide, with Inner Mongolia ranking first (annual average: 1097.79 cases). This distribution pattern is consistent with existing studies identifying Inner Mongolia, Shanxi, Heilongjiang, and Hebei as highest-incidence provinces (2005–2010)[24], with Inner Mongolia and adjacent areas as main high-risk clusters[25].

All five provinces exhibited a consistent unimodal seasonal pattern, with peak incidence between April and July and a trough in winter. This aligns with brucellosis transmission dynamics: warmer spring/summer temperatures coincide with livestock birthing and increased human outdoor activity, jointly driving epidemic spread. However, the key finding of this study is not the seasonality itself—which is well established—but rather the substantial regional heterogeneity in environmental drivers, lag periods, and optimal warning timing that underlies this shared seasonal pattern. The divergent temporal trends—persistent high incidence in Inner Mongolia, rapid growth in Xinjiang, decline in Shanxi, and stable low levels in Heilongjiang and Hebei—reflect differences in livestock industry scale, vaccination and quarantine measures, and socioeconomic conditions across regions. Environmental factors in Xinjiang show both short-lag and cumulative effects (rainfall lag: ~0.42 months; cumulative effect: ~1.95 months)[26]. In northeastern China, low

temperatures have a longer lag effect (peak lag: 2 months) while high temperatures have a more immediate effect (peak lag: 0.6 months)[27]. Similar seasonality has been observed in Henan, where the annual incidence growth rate reached 34.9% after 2018[28].

4.2 Environmental drivers and region-specific lag effects

This study employed a three-step screening procedure (CCF lag analysis, VIF multicollinearity elimination, and AIC stepwise regression) to identify key predictors for each province, revealing significant regional heterogeneity in driving factors and lag periods. In Inner Mongolia, soil moisture at a lag of 4 months was the only selected exogenous variable, as it indirectly influences Brucella survival in the environment; a meta-analysis confirmed that seroprevalence in dairy cows in traditional northern farming areas exceeds 10%[29]. This reflects the region's grassland-based livestock system, where a single strong environmental driver dominates transmission dynamics. In Xinjiang, selected variables included O_3 at lag 6 months (negatively correlated), soil moisture at lag 1 month, NO_2, and precipitation at lag 2 months, with extreme climatic conditions increasing transmission risk[26]. The greater number of selected variables reflects the region's diverse ecological conditions and complex, multi-factor transmission dynamics. Results from Shanxi, Hebei, and Heilongjiang further underscored regional specificity, consistent with a Shaanxi study confirming that climatic conditions, livestock farming intensity, and socioeconomic status drive incidence in different subregions[30].In Shanxi and Hebei, the weaker performance of LSTMX relative to LSTM may reflect weak environmental signals or the dominance of unmeasured factors such as livestock transportation and vaccination coverage, leading to overfitting when environmental variables are added. In Heilongjiang, the importance of NDVI and PM_{10} may relate to vegetation cover influencing grazing activity and aerosol transmission.

NDVI and FVC were positively correlated with incidence in most regions, suggesting that higher vegetation cover increases exposure risk. Temperature promoted Brucella survival and reproduction, while precipitation was negatively correlated (possibly due to increased livestock confinement on rainy days). Wind speed was positively correlated, potentially facilitating aerosol spread. Atmospheric pressure, relative humidity, ozone, NO_2, and PM_{10} also showed significant regional associations, indicating that brucellosis epidemiology results from interactions among multiple meteorological, ecological, and air pollution factors. Optimal lag periods varied between 1 and 6 months across regions, highlighting that early warning models must incorporate region-specific lagged variables to ensure accuracy.

4.3 Model performance comparison and applicability analysis

The SARIMA model demonstrated stable fitting performance across all five provinces and effectively captured seasonal fluctuations, consistent with previous brucellosis prediction studies[28,30]. However, its inability to incorporate exogenous variables limits predictive accuracy. The SARIMAX model showed heterogeneous performance across regions: accuracy improved significantly in Inner Mongolia and Hebei but decreased in Xinjiang, Shanxi, and Heilongjiang. This suggests that the benefit of exogenous variables depends on the strength of local environmental drivers. In regions with a single strong driver, such as Inner Mongolia, model performance improved; in regions with complex or weak drivers, incorporating multiple exogenous variables led to overfitting due to limited sample size[31].

The LSTM model showed clear advantages in capturing region-specific temporal patterns in most provinces, particularly in Shanxi and Hebei where univariate LSTM achieved the highest accuracy. The LSTMX model performed best in Inner Mongolia and Heilongjiang. This advantage stems from the LSTM gating mechanism, which captures nonlinear dynamics and long-term dependencies in time series data[31,32]. Previous studies have confirmed that LSTM outperforms SARIMA[20] and that convolutional long short-term memory (ConvLSTM) outperforms traditional machine learning methods[30].Compared with SARIMA and SARIMAX, LSTM models produce smoother fitted curves, smaller peak prediction deviations, and stronger ability to capture long-term nonlinear patterns, overcoming limitations of linear models[31,32].

However, it is important to note that these performance metrics reflect the ability to reproduce historical seasonal patterns rather than to predict unpredictable events such as outbreaks caused by the introduction of infected livestock. Brucellosis, like most zoonoses, can experience outbreaks triggered by unexpected inputs (e.g., importation of infected herds), which are inherently difficult to predict. The value of our models therefore lies not in predicting such random events, but in characterizing region-specific baseline patterns and identifying optimal timing for routine prevention and control interventions. Introducing exogenous variables did not always improve performance. In Shanxi and Hebei, LSTMX underperformed relative to LSTM[33]. Possible reasons include: (1) environmental factors may have complex interactions or threshold effects not fully captured; (2) correlations between some meteorological variables and brucellosis may be non-causal; (3) multicollinearity screening may not fully address overfitting in time series forecasting[34]. A comparative study on hand, foot, and mouth disease also confirmed that LSTM outperformed SARIMA models[35].

From a public health perspective, the conservative peak estimation of LSTM models has greater practical value[30]. LSTM and other machine learning methods are more suitable for accurate long-term trend characterization, while SARIMA's high sensitivity may be more valuable for short-term real-time early warning.

In summary, we propose an adaptive strategy for brucellosis characterization: SARIMA for basic surveillance prioritizing stability and interpretability; LSTM for high-precision early warning; LSTMX or SARIMAX for regions with strong environmental drivers; and avoiding direct cross-regional application of a single model.

4.4 Public health and One Health implications for prevention and control

Machine learning methods are widely used for modeling nonlinear systems in public health, providing precise predictive models[36–38]. While the seasonality of brucellosis is well known, our study provides region-specific quantification of optimal timing for prevention and control interventions. Based on the optimal lag periods identified in this study, we recommend the following region-specific warning schedules: Inner Mongolia should initiate soil moisture monitoring in January; Xinjiang, Shanxi, and Heilongjiang should initiate ozone, NDVI, and PM_{10} monitoring in November of the previous year; and Hebei should initiate ozone monitoring in December of the previous year. This region-specific timing is more precise than a uniform national strategy based solely on seasonal patterns[39]. The region-specific environmental factors identified—soil moisture in Inner Mongolia, ozone and precipitation in Xinjiang, ozone and wind speed in Shanxi, NDVI and PM_{10} in Heilongjiang, and ozone and sulfur dioxide in Hebei—can serve as early warning indicators. Deep learning models such as LSTM can achieve monthly single-step predictions, supporting rational allocation of medical resources and targeted control measures.

Under the One Health framework, human, animal, and environmental health are integrally linked[40]. Brucellosis prevention and control require a comprehensive, cross-sectoral strategy, as animal epidemics are inseparably linked to sustained human transmission[41]. The environmental drivers identified in this study provide a scientific basis for cross-sectoral joint prevention and control. Through multisectoral collaboration among health, livestock, and meteorological authorities, monitoring environmental changes and implementing early intervention measures can effectively reduce the occurrence and spread of brucellosis.

4.5 Limitations and future directions

This study has several limitations. First, only 120 months of data were used for training (January 2014 to December 2023) and 12 months for testing (January to December 2024). This limited sample size may have constrained LSTM performance, as insufficient data can lead to overfitting, increased prediction error, and model instability[42]. Second, only environmental factors were incorporated; other important drivers—such as livestock farming scale (e.g., sheep and cattle inventory), immunization and quarantine measures, population mobility, and socioeconomic conditions (e.g., Gross Domestic Product (GDP), per capita disposable income, mutton price and production)—were not included, although these are significantly correlated with brucellosis incidence[43].In particular, outbreaks caused by the introduction of infected livestock, which are inherently stochastic and difficult to predict, cannot be captured by our models. This limitation underscores that our findings should be interpreted as characterizing region-specific baseline patterns and optimal warning timing, rather than predicting unpredictable outbreak events. Third, this study employed a basic LSTM architecture without exploring hybrid approaches combining attention mechanisms or convolutional neural networks (e.g., convolutional neural network-long short-term memory (CNN-LSTM), ConvLSTM) or SARIMA-LSTM combinations. Existing research shows that CNN-LSTM hybrid models outperform single LSTM models in mitigating overfitting and improving accuracy[44], ConvLSTM outperforms traditional LSTM[34], and combined ARIMA–Elman neural networks outperform single ARIMA models[45].

Future directions include: integrating multisource heterogeneous data (livestock epidemics, socioeconomic activities, ecological data) to build more interpretable and generalizable prediction models; developing hybrid deep learning architectures with attention or convolutional mechanisms to improve accuracy and interpretability; conducting long-term risk projections under different climate change scenarios; and developing a real-time early warning system for dynamic epidemic surveillance and visualization to translate modeling results into practical prevention and control applications.

Conclusion

In summary, this study systematically compared four time-series models (SARIMA, SARIMAX, LSTM, LSTMX) for characterizing region-specific temporal patterns of brucellosis. While the seasonality of brucellosis is well established, our study demonstrates substantial regional heterogeneity in environmental drivers, optimal lag periods, and warning timing. LSTM significantly outperformed SARIMA and

SARIMAX in capturing region-specific patterns in most provinces. However, incorporating exogenous meteorological variables did not consistently improve performance, and variable selection must balance regional specificity, biological plausibility, and model complexity. Importantly, our models are designed to characterize baseline seasonal patterns and optimal warning timing, not to predict unpredictable outbreak events such as those caused by the introduction of infected livestock. Further analysis revealed pronounced regional heterogeneity in the effects of meteorological-environmental factors on brucellosis transmission. These findings support the development of localized early warning systems for brucellosis in high-risk areas of China and provide scientific evidence for understanding region-specific transmission mechanisms.

Declarations

Ethics approval and consent to participate

This study used de-identified aggregated surveillance data from the China Information System for Disease Control and Prevention. No individual-level data were accessed. Ethical approval was not required as the study used publicly available, de-identified surveillance data.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Y.Q. and Q.G. contributed equally to this work. Y.Q., Q.G., and Z.C. designed the study. Y.Q., Q.G., H.L., H.F., and Q.W. collected and analyzed the data. W.Z., C.L., and Q.C. provided critical revisions. Z.C. and Q.C. supervised the study. All authors read and approved the final manuscript.

Acknowledgements

We express our gratitude to the China Information System for Disease Control and Prevention for providing the brucellosis incidence data.

Data Availability

The brucellosis incidence data are available from the China Information System for Disease Control and Prevention upon reasonable request. Environmental datasets are available from the sources listed in the manuscript: the National Tibetan Plateau Data Center, the TerraClimate dataset, the NASA MOD13A3 dataset, the ChinaHighAirPollutants (CHAP) dataset, and a high-resolution meteorological forcing dataset published in Scientific Data. Detailed sources are provided in Table S1.

References

1. Organization WH, Brucellosis. Accessed Apr 29, 2026, https://www.who.int/news-room/fact-sheets/detail/brucellosis
2. Prevention CfDCa. Clinical Overview of Brucellosis. Accessed Apr 29. 2026, https://www.cdc.gov/brucellosis/hcp/clinical-overview/index.html

3. Dean AS, Crump L, Greter H, Schelling E, Zinsstag J. Global burden of human brucellosis: a systematic review of disease frequency. PLoS Negl Trop Dis. 2012;6(10):e1865. 10.1371/journal.pntd.0001865.
4. Young EJ. An overview of human brucellosis. Clin Infect Dis Aug. 1995;21(2):283–9. 10.1093/clinids/21.2.283. quiz 290.
5. Spink WW. What is chronic brucellosis? Ann Intern Med. Aug 1951;35(2):358–74. 10.7326/0003-4819-35-2-358.
6. Laine CG, Johnson VE, Scott HM, Arenas-Gamboa AM. Global Estimate of Human Brucellosis Incidence. Emerg Infect Dis Sep. 2023;29(9):1789–97. 10.3201/eid2909.230052.
7. Liu Z, Gao L, Wang M, Yuan M, Li Z. Long ignored but making a comeback: a worldwide epidemiological evolution of human brucellosis. Emerg Microbes Infect Dec. 2024;13(1):2290839. 10.1080/22221751.2023.2290839.
8. Zhang M, Chen X, Bu Q, et al. Spatiotemporal dynamics and influencing factors of human brucellosis in Mainland China from 2005–2021. BMC Infect Dis Jan. 2024;11(1):76. 10.1186/s12879-023-08858-w.
9. Liu Z, Shi Y, Xue C, Yuan M, Li Z, Zheng C. Epidemiological and Spatiotemporal Clustering Analysis of Human Brucellosis - China, 2019–2023. China CDC Wkly Jan. 2025;24(4):130–6. 10.46234/ccdcw2025.020.
10. Carvajal TM, Viacrusis KM, Hernandez LFT, Ho HT, Amalin DM, Watanabe K. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. BMC Infect Dis Apr. 2018;17(1):183. 10.1186/s12879-018-3066-0.
11. Wang YW, Shen ZZ, Jiang Y. Comparison of autoregressive integrated moving average model and generalised regression neural network model for prediction of haemorrhagic fever with renal syndrome in China: a time-series study. BMJ Open Jun. 2019;16(6):e025773. 10.1136/bmjopen-2018-025773.
12. Chen Y, Hou W, Dong J. Time series analyses based on the joint lagged effect analysis of pollution and meteorological factors of hemorrhagic fever with renal syndrome and the construction of prediction model. PLoS Negl Trop Dis Jul. 2023;17(7):e0010806. 10.1371/journal.pntd.0010806.
13. Zheng Y, Zhang L, Wang C, et al. Predictive analysis of the number of human brucellosis cases in Xinjiang, China. Sci Rep Jun. 2021;1(1):11513. 10.1038/s41598-021-91176-5.
14. Ning H, Lan Y, Huang Q, et al. Enhancing public health surveillance: SARIMAX model incorporating Baidu search index for HCV prediction in China. BMC Med Res Methodol Apr. 2025;23(1):108. 10.1186/s12874-025-02562-w.
15. Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals Jun. 2020;135:109864. 10.1016/j.chaos.2020.109864.
16. Yu Y, Si X, Hu C, Zhang J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Comput Jul. 2019;31(7):1235–70. 10.1162/neco_a_01199.
17. Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber J. LSTM: A Search Space Odyssey. IEEE Trans Neural Netw Learn Syst Oct. 2017;28(10):2222–32. 10.1109/tnnls.2016.2582924.
18. Qin C, Chen L, Cai Z, Liu M, Jin L. Long short-term memory with activation on gradient. Neural Netw Jul. 2023;164:135–45. 10.1016/j.neunet.2023.04.026.
19. Yang E, Zhang H, Guo X, Zang Z, Liu Z, Liu Y. A multivariate multi-step LSTM forecasting model for tuberculosis incidence with model explanation in Liaoning Province, China. BMC Infect Dis May. 2022;23(1):490. 10.1186/s12879-022-07462-8.
20. Lu L, Yang T, Chen Z, Ge Q, Yang J, Sen G. Prediction analysis of human brucellosis cases in Ili Kazakh Autonomous Prefecture Xinjiang China based on time series. Sci Rep Jan. 2025;7(1):1232. 10.1038/s41598-024-80513-z.
21. Peng R, Wang Y, Zhai J, et al. Driving effect of multiplex factors on human brucellosis in high incidence region, implication for brucellosis based on one health concept. One Health Dec. 2022;15:100449. 10.1016/j.onehlt.2022.100449.
22. Zhao C, Yang Y, Wu S, et al. Search trends and prediction of human brucellosis using Baidu index data from 2011 to 2018 in China. Sci Rep Apr. 2020;3(1):5896. 10.1038/s41598-020-62517-7.
23. Zhou S, Qin H, Shi Q, Li S, Chen J, Chen Q. Changing Patterns of Epidemiological Characteristics and Spatial-Temporal Clusters of Human Brucellosis Based on County Level - China, 2011–2023. China CDC Wkly Apr. 2025;4(14):453–9. 10.46234/ccdcw2025.075.
24. Zhong Z, Yu S, Wang X, et al. Human brucellosis in the People's Republic of China during 2005–2010. Int J Infect Dis May. 2013;17(5):e289–92. 10.1016/j.ijid.2012.12.030.
25. Yang H, Chen Q, Li Y, Mu D, Zhang Y, Yin W, Epidemic Characteristics. High-Risk Areas and Space-Time Clusters of Human Brucellosis - China, 2020–2021. China CDC Wkly Jan. 2023;6(1):17–22. 10.46234/ccdcw2023.004.

26. Wei C, Zhang L, Shang B, Wang X, Zhao J. Interactions between brucellosis and environmental factors: spatiotemporal epidemiology from Xinjiang, China. BMC Public Health Nov. 2025;17(1):4004. 10.1186/s12889-025-25415-z.
27. Meng J, Yewu Z, Qing G, Xuemei S. Spatiotemporal distribution of human brucellosis and lagged effects of meteorological factors in northeastern China, 2014–2023. Disease Surveillance. 2026;2026–01–30(3):300–5. 10.3784/jbjc.202511070732.
28. Wang Y, Liang Y, Xue C, et al. Use of a Seasonal Autoregressive Fractionally Integrated Moving Average Model for the Time Series Analysis of Human Brucellosis. Zoonoses Public Health Sep. 2025;72(6):534–43. 10.1111/zph.13229.
29. Ran X, Cheng J, Wang M, et al. Brucellosis seroprevalence in dairy cattle in China during 2008–2018: A systematic review and meta-analysis. Acta Trop Jan. 2019;189:117–23. 10.1016/j.actatropica.2018.10.002.
30. Shen L, Jiang C, Weng F, et al. Spatiotemporal risk of human brucellosis under intensification of livestock keeping based on machine learning techniques in Shaanxi, China. Epidemiol Infect Oct. 2024;24:152:e132. 10.1017/s0950268824001018.
31. Yunita A, Pratama MI, Almuzakki MZ, et al. Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models. MethodsX Dec. 2025;15:103462. 10.1016/j.mex.2025.103462.
32. Waqas M, Humphries UW. A critical review of RNN and LSTM variants in hydrological time series predictions. MethodsX Dec. 2024;13:102946. 10.1016/j.mex.2024.102946.
33. Patharkar A, Cai F, Al-Hindawi F, Wu T. Predictive modeling of biomedical temporal data in healthcare applications: review and future directions. Front Physiol. 2024;15:1386760. 10.3389/fphys.2024.1386760.
34. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med May-Jun. 2004;66(3):411–21. 10.1097/01.psy.0000127692.23278.a9.
35. Wang Y, Xu C, Zhang S, et al. Development and evaluation of a deep learning approach for modeling seasonality and trends in hand-foot-mouth disease incidence in mainland China. Sci Rep May. 2019;29(1):8046. 10.1038/s41598-019-44469-9.
36. dos Santos BS, Steiner MTA, Fenerich AT, Lima RHP. Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to. 2018. *Computers & Industrial Engineering*. 2019/12/01/ 2019;138:106120. https://doi.org/10.1016/j.cie.2019.106120
37. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. J Glob Health Dec. 2018;8(2):020303. 10.7189/jogh.08.020303.
38. Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning: 2015 and beyond. Lancet Psychiatry Jan. 2016;3(1):13–5. 10.1016/s2215-0366(15)00549-0.
39. Liu Z, Wang M, Wang Y, Yuan M, Li Z. Retrospective Analysis of the Epidemiological Evolution of Brucellosis in Animals - China, 1951–1989 and 1996–2021. China CDC Wkly Nov. 2024;1(44):1159–70. 10.46234/ccdcw2024.235.
40. Rossetti CA, Arellano-Reynoso B, Gotuzzo EH, Oñate AA, Ponsart C. Editorial: Research advances toward One Health in brucellosis. Front Microbiol. 2026;17:1815921. 10.3389/fmicb.2026.1815921.
41. Bardenstein S, Grupel D, Blum SE, Motro Y, Moran-Gilad J. Public and animal health risks associated with spillover of Brucella melitensis into dairy farms. Microb Genom Apr. 2023;9(4). 10.1099/mgen.0.001014.
42. Horenko I. On a Scalable Entropic Breaching of the Overfitting Barrier for Small Data Problems in Machine Learning. Neural Comput Aug. 2020;32(8):1563–79. 10.1162/neco_a_01296.
43. Liu Z, Wang M, Shi Y, et al. Epidemiological Evolution Profile of Human Brucellosis and Socioeconomic Factor Correlation Analysis - Southern and Northern Areas, China, 1950–2021. China CDC Wkly Apr. 2025;4(14):460–6. 10.46234/ccdcw2025.076.
44. Susandri S, Zamsuri A, Nasution N, Efendi Y, Alwan HB. The Mitigating Overfitting in Sentiment Analysis Insights from CNN-LSTM Hybrid Models. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*. %03/%10. 2025;24(2):297–308. 10.30812/matrik.v24i2.4742.
45. Zhai M, Li W, Tie P, et al. Research on the predictive effect of a combined model of ARIMA and neural networks on human brucellosis in Shanxi Province, China: a time series predictive analysis. BMC Infect Dis Mar. 2021;19(1):280. 10.1186/s12879-021-05973-4.

Figures

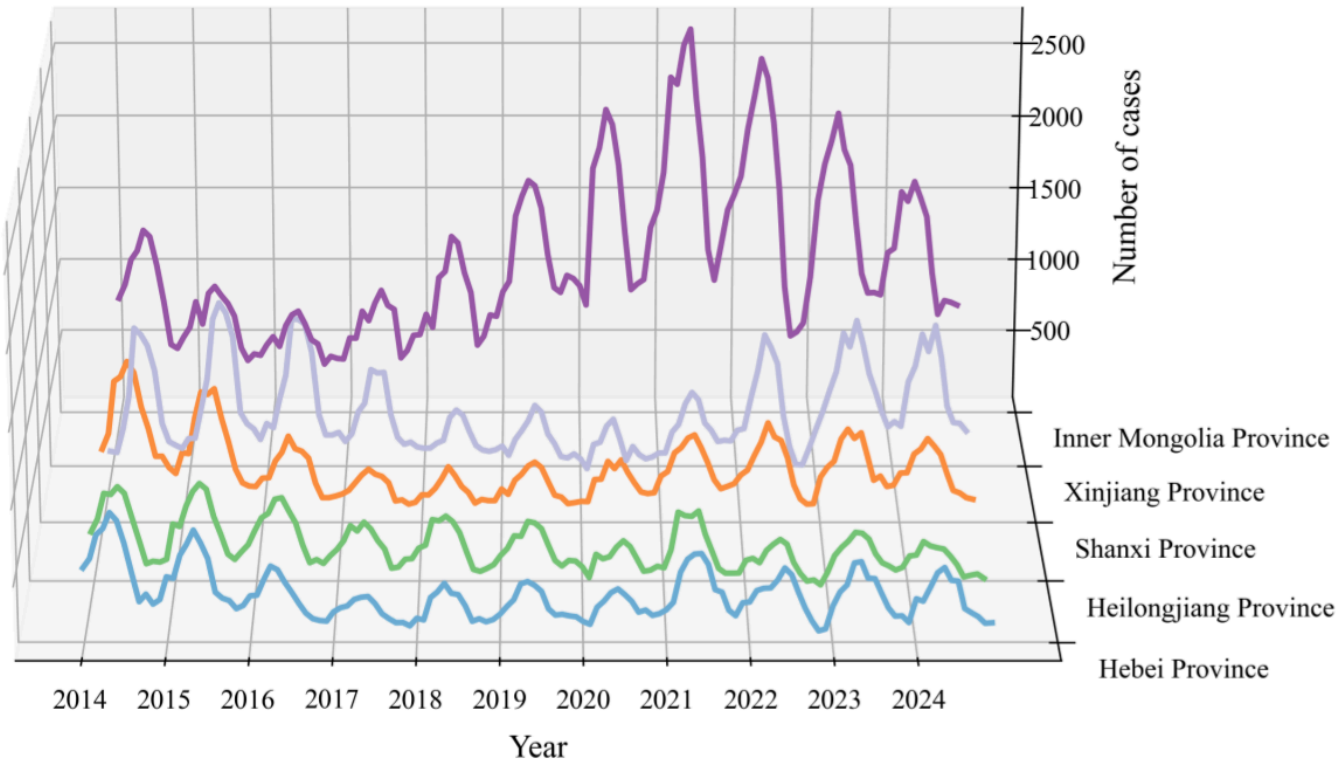

Figure 1

Annual trends of brucellosis cases in five provinces of China, 2014-2024.

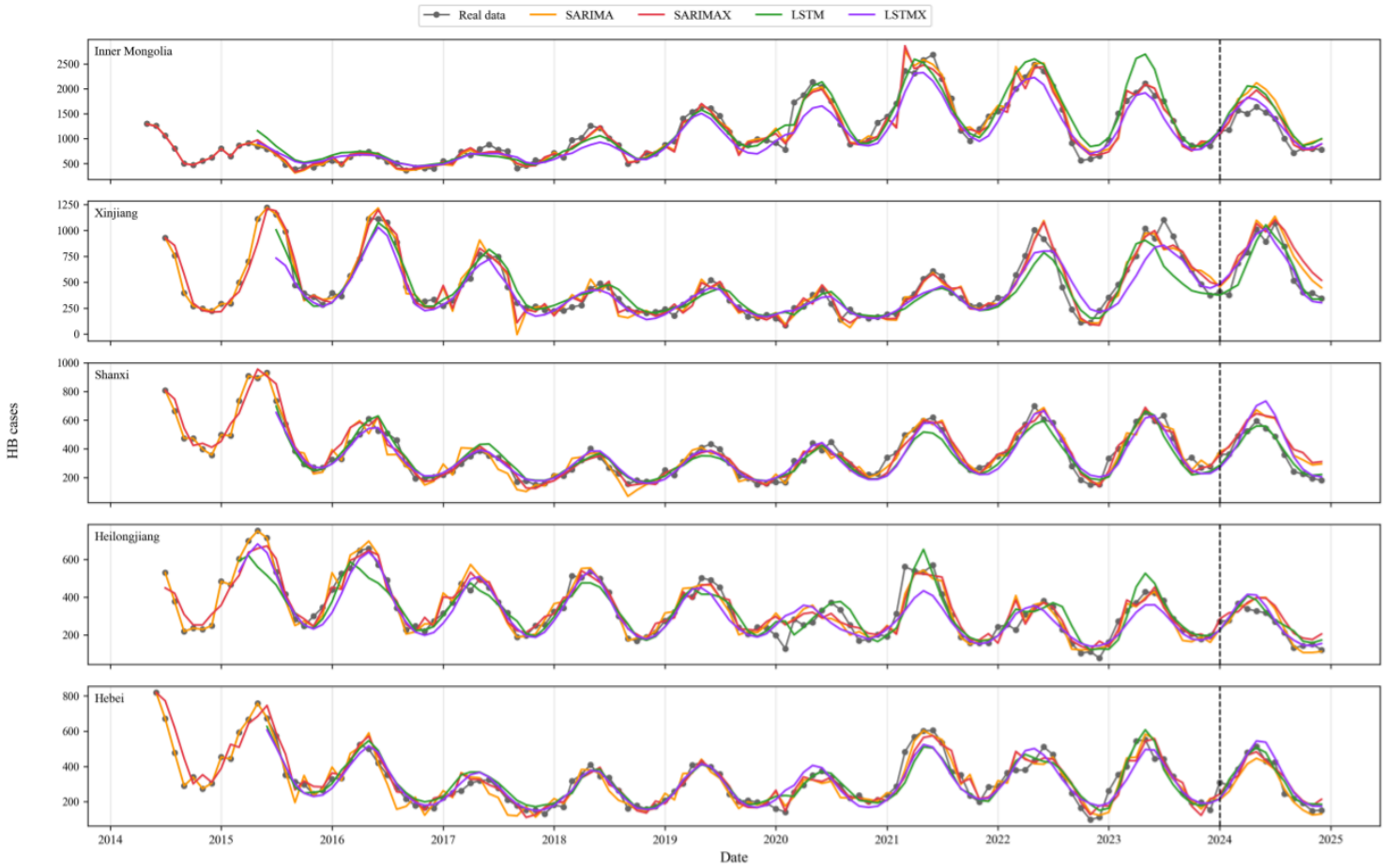

Figure 2

Fitting and prediction curves of four models for brucellosis in five Chinese provinces

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx