Modeling Earth Systems and Environment (2023) 9:4493–4505
https://doi.org/10.1007/s40808-023-01738-x

ORIGINAL ARTICLE

Relative humidity prediction with covariates and error correction based on SARIMA-EG-ECM model

Jiajun Guo[1] · Liang Zhang[1] · Ruqiang Guo[1]

Received: 21 October 2022 / Accepted: 7 February 2023 / Published online: 14 March 2023

Abstract

RH is a physical quantity measuring atmospheric water vapor content. Predicting RH is of great importance in weather, climate, industrial production, crops, human health, and disease transmission, since it is helpful in making critical decisions. In this paper, the effects of covariates and error correction on relative humidity (RH) prediction have been studied, and a hybrid model based on seasonal autoregressive integrated moving average (SARIMA) model, cointegration (EG), and error correction model (ECM) named SARIMA-EG-ECM (SEE) has been proposed. The prediction model was performed in the meteorological observations of Hailun Agricultural Ecology Experimental Station, China. Based on the SARIMA model, the meteorological variables that interact with RH were used as covariates to perform EG tests. A cointegration model has been constructed. It revealed that RH had a cointegration relationship with air temperature (TEMP), dew point temperature (DEWP), precipitation (PRCP), atmospheric pressure (ATMO), sea-level pressure (SLP), and 40 cm soil temperature (40ST), which revealed the long-term equilibrium relationship between series. An ECM was established which indicated that the current fluctuations of DEWP, ATMO, and SLP have a significant impact on the current fluctuations of RH. The established ECM describes the short-term fluctuation relationship between the series. With the increase of the forecast horizon from 6 to 12 months, the prediction performance of the SEE model decreased slightly. A comparative study has also been introduced, indicating that the SEE performs superior to SARIMA and Long Short-Term Memory (LSTM) network.

Keywords Relative humidity · Cointegration model · Error correction model · Multiplicative seasonality model

Introduction

As the strongest greenhouse gas in the atmosphere and the main component of the water cycle, the greenhouse contribution of water vapor is several times that of carbon dioxide (Jones et al. 2007). It in the lower troposphere is the main source of precipitation for all weather systems. Water vapor absorbs radiation through the formation and evolution of clouds, and affects the changes of other variables in the climate system (Nian et al. 2018). Relative Humidity (RH) refers to the ratio of the maximum amount of water vapor in the atmosphere to the amount of water vapor that the air can contain at a certain temperature (Xie et al. 2011)). It is a physical quantity measuring atmospheric water vapor content. Predicting RH is of great importance in weather, climate, industrial production, crops, human health, and disease transmission, since it is helpful in making critical decisions. RH plays a vital role in driving electricity demand during the warm months (June–September) (Xie et al. 2018). Negative temperature and high RH are important conditions in the prediction of aircraft icing area (Ivanova 2009). The study of Duan et al. (2019) demonstrated that the encountering high and low RH, the daily allergic rhinitis outpatients increased. Humans are more susceptible to respiratory novel coronavirus (COVID-19) when the RH decreases (Mangla et al. 2021). In crops, RH is crucial in regulating root hydraulic characteristics (Calvo-Polanco et al. 2017). Models for dust storm predicting may be improved by utilizing RH and wind speed as main drivers for dust generation and transport (Csavina et al. 2014). Kwon et al. (2019) uses public weather forecast information about temperature, RH,

✉ Liang Zhang
zhanglsd@126.com

Jiajun Guo
0161121866@mail.imu.edu.cn

Ruqiang Guo
guoruqiang227@nwafu.edu.cn

[1] College of Science, Northwest A and F University, Yangling, Shaanxi 712100, China

dew point, and sky coverage as a training set in the naive Bayes classifier classification of hourly resolution for global horizontal irradiance prediction.

Quite a few methods have been utilized to predict RH. Yu (2009) used correlation analysis with the index station and RH reference value for predicting precipitation with RH. In addition, Lu and Viljanen (2009) used external input nonlinear autoregressive (NNARX) model and genetic algorithm to establish a neural network to achieve the purpose of prediction. Practice of Kuzugudenli (2018) has proved that the artificial neural network method had greater predictive power than the model developed with multiple linear regression. However, Tkacz (2001) has found that artificial neural networks are not able to improve on an autoregressive model. Although the regression model, the correlation analysis, and back propagation (BP) neural network method have their own advantages, such non-parametric methods have a great dependence on the choice of variables (Li et al. 2019b). In recent years, Long Short-Term Memory (LSTM) network has performed well in predicting meteorological variables with dynamic characteristics, such as temperature (TEMP), RH, and precipitation (PRCP) due to its special network (Gao et al. 2021; Hutapea et al. 2020; Casallas et al. 2021).

Since RH is a time series recorded at intervals of time, there may be a certain trend and periodicity between the series. Autoregressive moving average method (ARIMA) is one of the commonly used prediction methods in parametric methods (Eymen and Köylü 2019; Rathod et al. 2017; Fernández-González et al. 2016). For dealing with seasonal time series, such as RH, seasonal autoregressive integrated moving average (SARIMA) had a great effect for forecasting as shown by (Valipour 2015; Bas Cerdá et al. 2017; Fang and Lahdelma 2016; Qiu et al. 2021; Murthy et al. 2018; Cong et al. 2019; Shad et al. 2022).

To overcome the problems of non-stationary of the time series, Engle and Granger (1987) provides the cointegration theory. If there is a cointegration relationship between non-stationary time series, there will be no pseudo-regression problem. Cointegration theory does not require all sequences to be stable, only their regression residual sequence is stable. The cointegration model performs well in measuring the long-term equilibrium relationship of the series (Granger and Swanson 2010; Zhang et al. 2015; Abdi et al. 2022). While the error correction model (ECM) as a complementary model performs well in explaining the short-term fluctuation relationship of the series as indicated in (Li et al. 2013; Ma et al. 2015; Abdi et al. 2022). Meanwhile, some researchers started to introduce the cointegration theory into the meteorological field and also found many valuable results. Statistical analysis was performed on water level, temperature, and humidity using cointegrated vector autoregression models by Appiah (2017). Htet (2017) proposed the Airline Error Correction Model (AECM), and forecast CO using traffic, precipitation, and air temperature as extrinsic variables. A novel multi-step forecasting method of hourly PM2.5 concentration is proposed with ECM using for correcting the prediction error according to studies of Yin et al. (2021). However, it is relative rare in the relationship for RH. The purpose of this study was to combine SARIMA with cointegration theory to form the SARIMA-EG-ECM (SEE) model, and to use the SEE model to predict RH at the Agricultural Ecological Experimental Station of the Chinese Academy of Science. This paper utilizes the cointegration model on the basis of SARIMA to establish a dynamic model with air temperature, dew point temperature, precipitation, and other meteorological variables as covariates, and use the ECM model to discuss the effect of the current fluctuation of covariates on the fluctuation of RH. This paper will verify the performance of the SEE model by comparing SARIMA model (including the multiplicative seasonality model and the additive seasonality model), LSTM model, and SEE model.

Materials and methods

SARIMA model

The SARIMA method can be used to model series with seasonal effects and periodic fluctuations. According to the difficulty of extracting seasonal effects, it is divided into additive seasonality model and multiplicative seasonality model (Danhui 2019).

Additive seasonality model

In the additive seasonality model, the seasonal change S_t, the trend T_t, and the immediate I_t in the time series are in the additive relationship shown in the formula (1), namely

$$x_t = S_t + T_t + I_t. \tag{1}$$

The series can be smoothed by the trend difference and the seasonal difference, and the smoothed series can be fitted by ARIMA model. The structure of the additive season model is

$$\nabla_D \nabla^d x_t = \frac{\Theta(B)}{\Phi(B)} \varepsilon_t, \tag{2}$$

where D is the step size of the seasonal period, d is the order of the difference, and $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ is the q-order AR coefficient polynomial. $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the MA coefficient polynomial of order p. $\{\varepsilon_t\}$ is a white noise series, and $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \sigma_\varepsilon^2$.

Multiplicative seasonality model

Usually, the long-term trend effect, seasonal effect, and random fluctuation of time series are not easy to be separate like the previous subsection because of the complex interaction between them. At this time, the additive seasonality model cannot fully extract their interaction. The multiplicative seasonality model needs to be adopted. The construction principle of the multiplicative seasonal model is shown in Fig. 1. In fact, due to the multiplicative relationship between the short-term correlation of the series and the seasonal effect, the multiplicative seasonality model is the product of ARMA(p, q) and ARMA(P, Q), denoted as ARIMA$(p, d, q) \times (P, D, Q)_S$. The structure is

$$\nabla^d \nabla_S^D x_t = \frac{\Theta(B)\Theta_S(B)}{\Phi(B)\Phi_S(B)}\varepsilon_t, \tag{3}$$

where $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ is the non-seasonal q-order AR coefficient polynomial. $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the non-seasonal MA coefficient polynomial of order p. $\Theta_S(B) = 1 - \theta_1 B^S - \cdots - \theta_Q B^{QS}$ is the seasonal Q-order AR coefficient polynomial. $\Phi_S(B) = 1 - \phi_1 B^S - \cdots - \phi_P B^{PS}$ is the seasonal MA coefficient polynomial of order P.

Cointegration model

Some series themselves change unevenly, but there are close long-term equilibrium relationships between the series. Cointegration model can measure whether there are long-term equilibrium relationships between the series effectively. Assuming that series of independent variables are $\{x_1\}, \{x_2\}, \cdots, \{x_n\}$. And the series of response variable is $\{y_t\}$. We can construct a regression model

$$y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{it}. \tag{4}$$

If the residual series $\{\varepsilon_t\}$ in the regression model is stationary, it is said that there is a cointegration relationship between the series of response variable $\{y_t\}$ and the series of independent variables $\{x_1\}, \{x_2\}, \cdots, \{x_n\}$.

Error correction model

As a supplementary model of the cointegration model, ECM was originally proposed by Hendry and Anderson (1977), which can explain the short-term fluctuation relationship of the series.

If there is a cointegration relationship among the series of response variable $\{y_t\}$ and the series of independent variables $\{x_1\}, \{x_2\}, \cdots, \{x_n\}$, that is

$$\begin{aligned} y_t &= \beta x_t + \varepsilon_t, \\ \varepsilon_t &= y_t - \beta x_t \sim I(0). \end{aligned} \tag{5}$$

According to Eq. (5), there is

$$y_t - y_{t-1} = \beta x_t - y_{t-1} + \varepsilon_t. \tag{6}$$

Combine Eq. (6) with $y_{t-1} = \beta x_{t-1} + \varepsilon_{t-1}$, there is

$$y_t - y_{t-1} = \beta x_t - \beta x_{t-1} + \varepsilon_{t-1} + \varepsilon_t. \tag{7}$$

Let the least square estimate of β be $\hat{\beta}$. Then, $\hat{\varepsilon}_{t-1} = y_{t-1} - b\hat{e}ta_{t-1}$ stands for the error from the previous period, denoted as ECM_{t-1}. Equation (7) can be written as

$$\nabla y_t = \beta \nabla x_t - ECM_{t-1} + \varepsilon_t. \tag{8}$$

According to Eq. (8), there are three main types of short-term fluctuations that will influence the current fluctuations (∇y_t) of the response series. They are:

1. ∇x_t: Current fluctuation of the input series;
2. ECM_{t-1}: Error from the previous period;
3. ε_t: Random fluctuations in the current period.

In summary, the structure of the model is

$$\nabla y_t = \beta_0 \nabla x_t + \beta_1 ECM_{t-1} + \varepsilon_t. \tag{9}$$

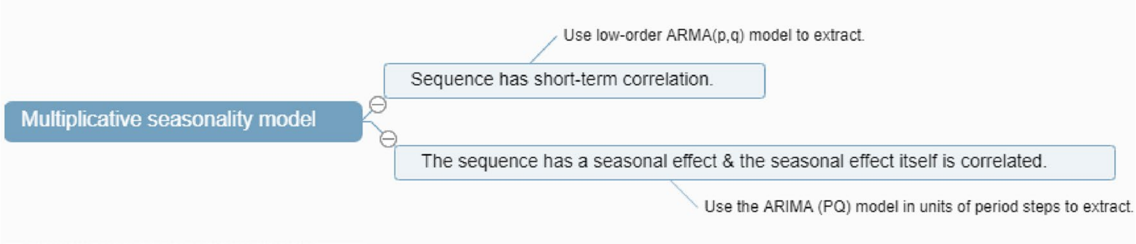

Use low-order ARMA(p,q) model to extract.
Sequence has short-term correlation.
Multiplicative seasonality model
The sequence has a seasonal effect & the seasonal effect itself is correlated.
Use the ARIMA (PQ) model in units of period steps to extract.

Fig. 1 Construction principle of the multiplicative seasonality model

Among them, $\beta_1(\beta_1 < 0)$ is the coefficient of error correction, indicating the extent to which the error correction term can correct the current fluctuation.

Long short-term memory networks

Hochreiter and Schmidhuber (1997) proposed the LSTM, which is an improved Recurrent Neutral Network (RNN) model. The LSTM unit consists of input gate i_t, forgetting gate f_t, and output gate o_t. The Forget Gate f_t controls how much information is forgotten by the internal state $c_t - 1$ at the previous moment, the input gate i_t controls how much information is saved by the candidate state c_t at the current moment, and the output gate o_t controls how much information is output by the internal state c_t at the current moment to the external state h_t. LSTM structure is shown in Fig. 2, where '×' and '+' represent the multiplication and addition operations of the matrix, respectively. σ and tanh are activation functions. The mathematical definitions are as follows:

$$\begin{aligned}
f_t &= \sigma\left(W_f[h_{t-1}, x_t] + b_f\right),\\
i_t &= \sigma\left(W_i[h_{t-1}, x_t] + b_i\right),\\
\widetilde{C}_t &= \tan h(W_c[h_{t-1}, x_t] + b_c),\\
C_t &= f_t \times C_{t-1} + i_t \times \widetilde{C}_t,\\
\sigma_t &= \sigma\left(W_0[h_{t-1}, x_t] + b_0\right),\\
h_t &= \sigma_t \times \tan\sigma(C_t),
\end{aligned} \tag{10}$$

where W_f, W_i, W_0 is the weight parameter, and b_i, b_f, b_0 is the deviation parameter. The mathematical formula mentioned above is for a unit. The work of an LSTM network has layers, and each layer has several units. In this paper, the network configuration is characterized by the following parameters: batch size=1, epochs=3000, and neurons=6.

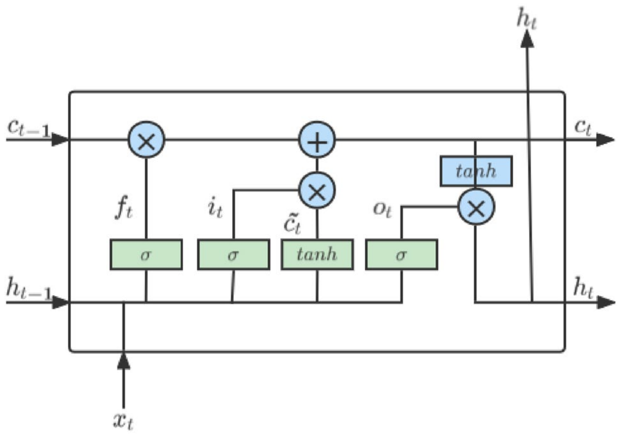


Fig. 2 The unit structure of LSTM network

SARIMA-EG-ECM hybrid model

Previous studies have used the dynamic ARIMA model with covariates (Li et al. 2021). Our research focuses on the seasonality of RH. Therefore, based on the SARIMA model, the SEE model is established. The methodology used for the determination of the SEE model includes three steps. First, a cointegration test is performed on RH and other meteorological variables to consider the long-term equilibrium relationship among the series. Second, a cointegration model based on the SARIMA model is fitted using the meteorological variables that have a cointegration relationship with RH. Third, an ECM is established as a supplement to the cointegration model to explore the impact of the current fluctuation of meteorological variables on the current fluctuation of RH, so as to describe the short-term fluctuation relationship among the series. Figure 3 describes the procedure.

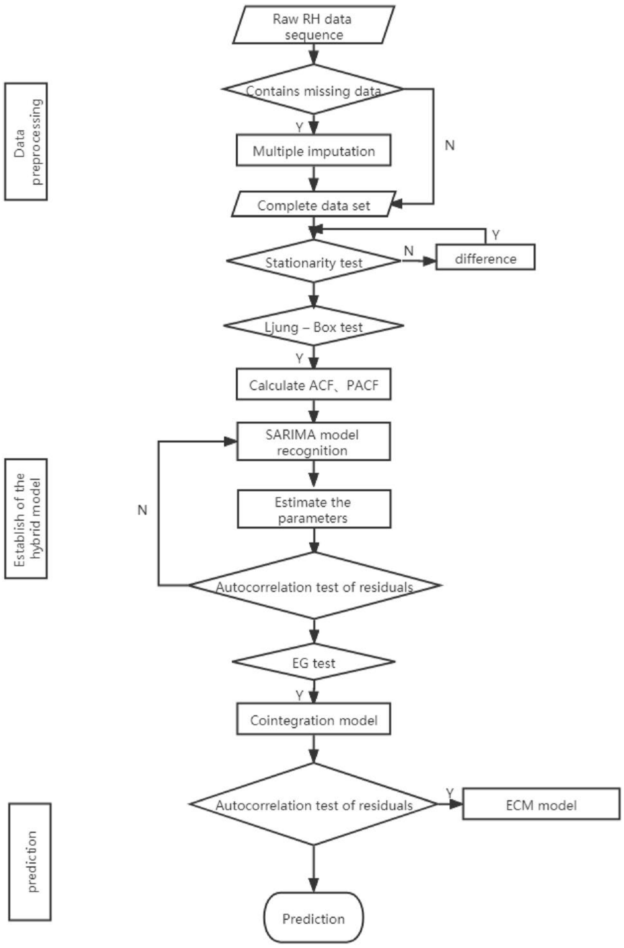


Fig. 3 Modeling flowchart of the hybrid model

Forecast accuracy measures

The choice of the fitted model should be considered from two aspects: on the one hand, the likelihood function is maximized, and on the other hand, the number of unknown parameters in the model is minimized. The larger the likelihood function value, the better the model fitting effect. The more unknown parameters in the model, the more independent variables, the more flexible the model changes, and the higher the accuracy of model fitting. However, only measuring the pros and cons of the model by fitting accuracy will result in an increasing number of unknown parameters in the model and an increase in unknown risks. Correspondingly, the model becomes more and more complex, and the estimation of parameter becomes more and more difficult (Danhui 2019). Therefore, when selecting a fitting model, it is necessary to choose a comprehensive optimal configuration of fitting accuracy and the number of unknown parameters.

1. Akaike Information Criterion

Akaike Information Criterion (AIC) was proposed by Japanese statistician Akaike (Akaike 1973). AIC is a weighted function of fitting accuracy and the number of parameters: the calculation method is as follows:

$$AIC = 2N_1 - 2ln\,(L_1), \tag{11}$$

where N_1 is the number of model parameters and L_1 is the maximum likelihood function of the model.

2. Bayesian Information Criterion

Although, the AIC criterion provides an effective criterion for the choice of fitting model. When faced with a complex model containing multiple independent variables, the information provided by the fitting error in the AIC criterion will be amplified by the sample size and the number of parameters. The penalty factor of the number is always 2, which has nothing to do with the sample size. Therefore, when the sample size is large, the fitting model selected using the AIC criterion contains more unknown parameters than the real model, and does not converge to the real model. Bayesian Information Criterion (SBC) was proposed by Schwarz (1978) based on Bayes theory. The penalty weight for the number of unknown parameters was changed from a constant 2 to the logarithmic function of the sample size $ln(n)$, which made up for the deficiency of the AIC criterion in the case of large sample size. The calculation method of SBC is

$$SBC = ln\,(n)\,N_2 - 2ln\,(L_2), \tag{12}$$

where N_2 is the number of parameters in the model and L_2 is the value of maximum-likelihood function.

When selecting the fitting model in this paper, using the AIC criterion and the SBC criterion helps us find the relative optimum fitting model within a limited range of orders. The model that minimizes the AIC or SBC function is the relatively optimal model.

In addition to AIC criterion and SBC criterion, the performance of the hybrid model can be evaluated by various statistical metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Residual Standard Error (RSE), and Coefficient of Determination (R^2).

MSE and RMSE are used in detecting the deviation between the predicted value of the model and the true value. When their value equals 0, the model used for prediction is the optimal model, and accordingly, the larger the error, the larger the value.

RSE describes the average offset between the target and the real regression line, which is used in estimating the standard deviation of the residual. Values of RSE close to 0 represent the optimal performances.

R^2 is the proportion that reflects the total variance of the dependent variable that can be explained by the independent variable through the regression relationship. R^2 provides a method to evaluate the performance of the same model on different data. Its value ranges from 0 to 1, with 0 that indicates the optimal performances.

The formulas can be defined as follows:

$$\begin{aligned} RMSE &= \sqrt{\frac{\sum_{i=1}^{n}(f_i - y_i)^2}{n}};\\ MSE &= \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2;\\ RSE &= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(f_i - y_i)^2};\\ R^2 &= 1 - \frac{\sum_{i=1}^{n}(f_i - y_i)^2}{\sum_{i=1}^{n}(f_i - y_i)^2}, \end{aligned} \tag{13}$$

where n is the total number of time series data, f_i is the predicted value of the i-th data, and y_i is the measured value of the i-th data.

Results

Data preparation

The meteorological data used in this paper are derived from the meteorological data set observed by Hailun Agricultural Ecology Experimental Station, China Academic of Science, and a total of data set take the month as the scale to collate

Table 1 Variance information of the multiple imputation procedure

Parameter	Variance			Relative increase	Fraction missing	Relative efficiency
	Between	Within	Total			
RH mean	0.001792	0.894575	0.896725	0.002403	0.002400	0.999520

Table 2 Parameter estimates (H_0 : $parameter = \theta_0$)

Parameter	Estimate	Std error	Minimum	Maximum	θ_0	t	Pr> t
RH mean	69.413333	0.946955	69.341667	69.441667	0	73.30	< 0.0001

Table 3 Phillips–Perron unit root test of the original series

Type	Lag	ρ	$Pr < \rho$	τ	$Pr < \tau$
Zero mean	0	− 1.3630	0.4133	− 0.80	0.3669
	1	− 1.3035	0.4222	− 0.78	0.3751
	2	− 0.9889	0.4730	− 0.68	0.4219
Single mean	0	− 65.5623	0.0010	− 6.80	< 0.0001
	1	− 72.1335	0.0010	− 7.00	< 0.0001
	2	− 66.6754	0.0010	− 6.84	< 0.0001
Trend	0	− 65.7102	0.0004	− 6.78	< 0.0001
	1	− 72.3065	0.0004	− 6.98	< 0.0001
	2	− 66.8148	0.0004	− 6.81	< 0.0001

10-year meteorological data published from 2009 to 2018 including 17 meteorological variables (Li et al. 2019a). Average RH for 108 months from January, 2009 to December, 2017 in Hailun Agricultural Ecology Experimental Station were selected to establish predictive models. The 12-month data from January, 2018 to December, 2018 are utilized in evaluating. In this study, we used the SEE model to predict RH for 6 and 12 months. It is found that due to voltage instability or other unknown reasons, part of the data contained missing values. To ensure the effectiveness of data analysis and prediction, we require imputing the 17 time series of meteorological variables in the data set with the multiple imputation method.

Take the RH time series as an example: as shown in Table 1, the number of multiple imputation is 5, and the relative efficiency is as high as 0.999520. P value of the parameter estimation for imputing the vacancies is less than 0.05, which passed the hypothesis test, as shown in Table 2.

Model fitting and prediction based on multiplicative seasonality model

The Phillips–Perron unit root test is performed on the RH time series, and it can be seen from Table 3 that the autoregressive process of the drift-free term of the series is non-stationary.

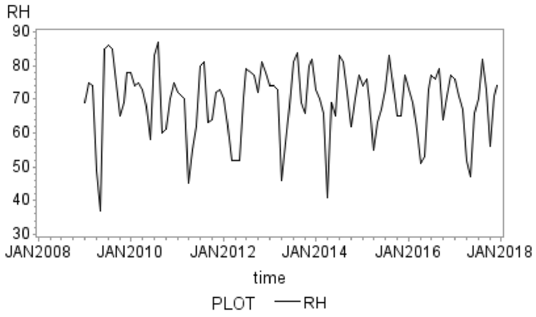


Fig. 4 Timing diagram of RH

Table 4 White noise test of the original series

lag	χ^2	Df	$Pr > \chi^2$
6	21.13	6	0.0017
12	36.47	12	0.0003
18	39.27	18	0.0026

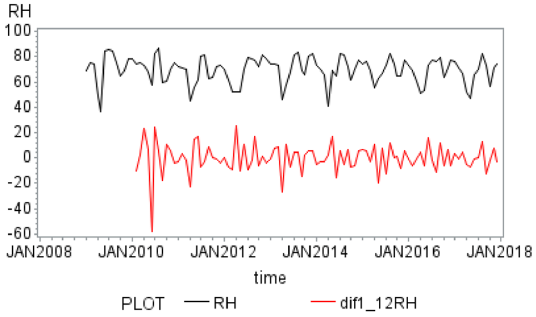


Fig. 5 Timing diagram of RH after first-order 12-step difference

We drew a timing diagram of the RH as Fig. 4. It can be seen that the series has a periodic effect with a year as a period. We performed the Phillips–Perron unit root test. As shown in Table 4, not all P values are less than 0.05, which indicates that the series after the difference is a non-white noise series. It contains relevant information worthy of being extracted. Therefore, we performed first-order 12-step difference on the original series to extract the information contained in the original series. The red part in Fig. 5 illustrates

Table 5 Phillips–Perron unit root test of the series after first-order 12-step difference

type	lag	ρ	$Pr < \rho$	τ	$Pr < \tau$
Zero mean	0	– 132.038	0.0001	– 14.88	< 0.0001
	1	– 125.235	0.0001	– 15.53	< 0.0001
	2	113.008	0.0001	– 17.79	< 0.0001
Single mean	0	– 132.043	0.0001	– 14.80	< 0.0001
	1	– 125.239	0.0001	– 15.44	< 0.0001
	2	– 113.005	0.0001	– 17.68	< 0.0001
Trend	0	– 132.045	0.0001	– 14.72	< 0.0001
	1	– 125.241	0.0001	– 15.35	< 0.0001
	2	– 113.002	0.0001	– 17.57	< 0.0001

Table 6 Conditional least-squares estimation of parameters of multiplicative seasonality model

Parameter	Estimate	Std error	t value	Approximate Pr> t	Lag
MA1,1	0.80061	0.06206	12.90	< 0.0001	1
MA2,1	0.50315	0.09231	5.45	< 0.0001	12

differential series of RH. To judge whether the differential series is stationary, we performed the Phillips–Perron unit root test. P values less than 0.05 in Table 5 show that the differential series is significantly stationary.

Figure 6 illustrates the trend and correlation analysis for RH after the first-order 12-step difference. Combined with the characteristics of the autocorrelation coefficient and partial autocorrelation coefficient mentioned in Fig. 6, the multiplicative seasonality model $(\mathrm{ARIMA}(0,1,1)\times(0,1,1)_{12})$ with $AIC = 669.8517$ and $SBC = 674.9594$ performs better than additive seasonality model $(\mathrm{ARIMA}(0,1,1)\times(0,1,0)_{12})$ with $AIC = 683.5248$ and $SBC = 688.6325$. We use the least square method to estimate the parameters. From Table 6, we can see that the parameters are significant and pass the test with P values less than 0.05. The model is

$$\nabla\nabla_{12}x_t = (1-0.80061 * B) * (1-0.50315 * B^{12})\varepsilon_t, \\ Var(\varepsilon_t) = 66.17914. \tag{14}$$

From Table 7, it can be seen that the residual series has passed the test with P values larger than 0.05. This confirms that the model has fully extracted the seasonal effect and short-term correlation of the series, and fits well.

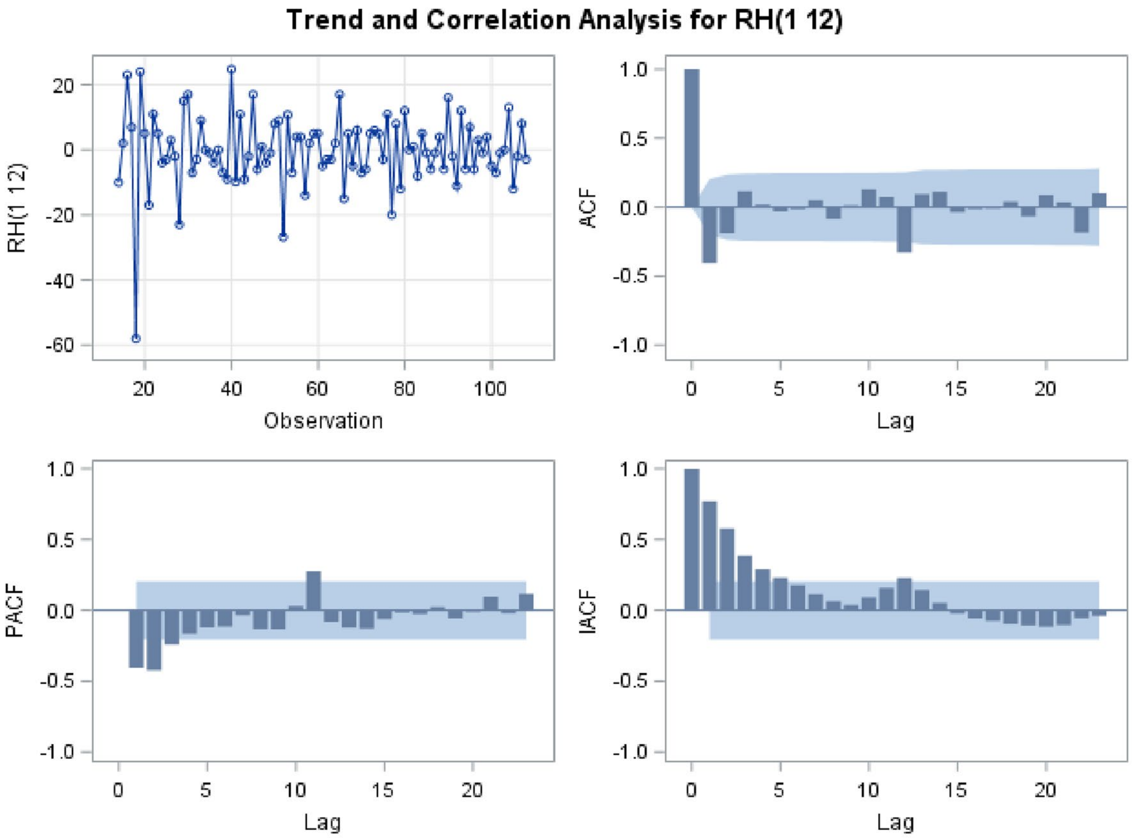
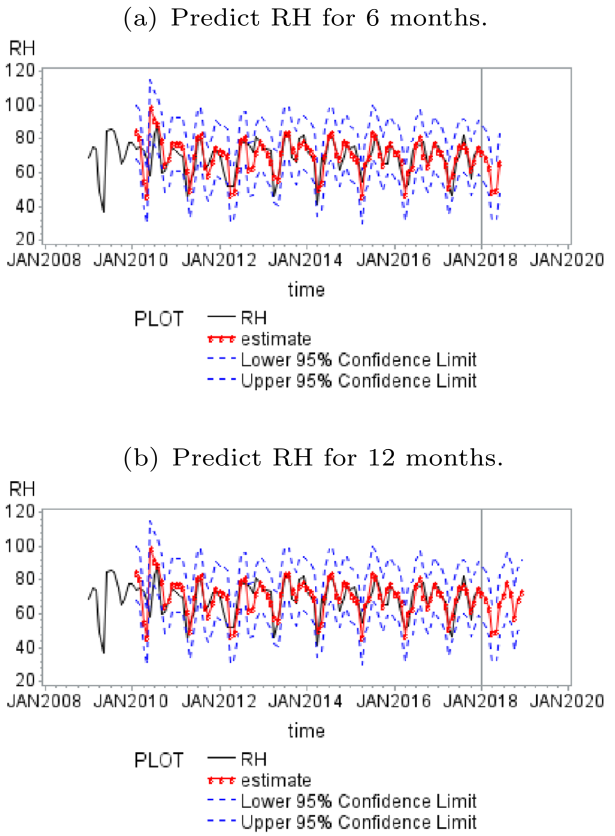


Fig. 6 Autocorrelations and partial autocorrelations of RH after first-order 12-step difference

Table 7 Residual autocorrelation test of multiplicative seasonality model

Lag	χ^2	Df	$Pr > (\chi)^2$
6	2.34	4	0.6738
12	3.66	10	0.9615
18	4.39	16	0.9980
24	7.52	22	0.9982


(a) Predict RH for 6 months.

(b) Predict RH for 12 months.

Fig. 7 RH prediction map based on multiplicative seasonality model

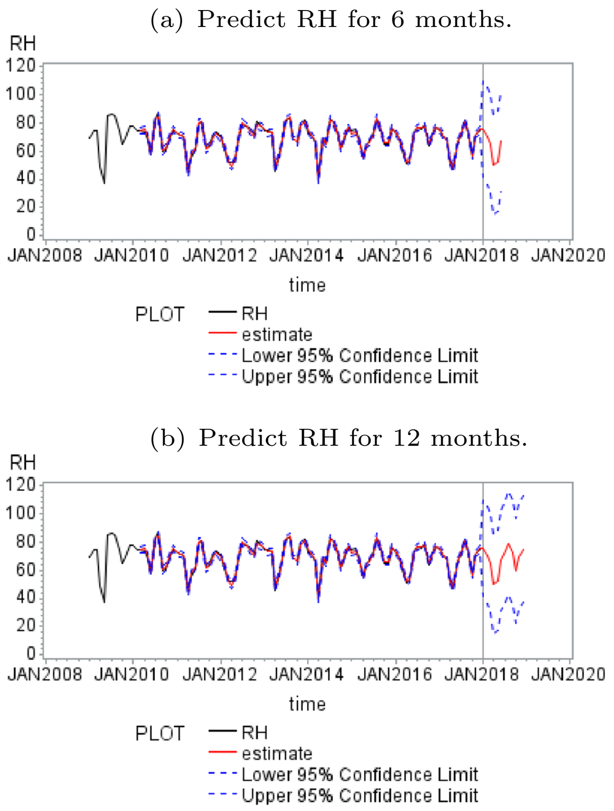
According to the multiplicative seasonality model, we made prediction (Fig. 7), where the black asterisk is the actual measured value of RH, the red line is the predicted value of RH, and the blue line is the 95% confidence upper and lower limit of the series.

RH forecast based on SEE model

In fact, changes in RH are not only affected by the changes in the series itself, but also by other meteorological conditions, such as temperature, precipitation, etc. Therefore, we also took into account changes in other meteorological conditions to obtain more accurate results. The cointegration test was performed on other meteorological condition series and RH series in the processed data set, and the results are shown in Table 8. It shows that RH has a cointegration relationship with TEMP, dew point temperature (DEWP), PRCP, atmospheric pressure (ATMO), sea-level pressure (SLP), and 40 cm soil temperature (40ST), which reveals the long-term equilibrium relationship between sequences. These meteorological variables play a significant role in the model of predictions. To fit the interaction between RH and other meteorological conditions, a dynamic regression model can be constructed

Table 8 Parameter estimation results of cointegration model

Parameter	Estimate	Standard error	*t* value	$Pr > \|t\|$	Variable
MA1,1	0.74194	0.07800	9.51	< 0.0001	RH
MA2,1	0.66551	0.08486	7.84	< 0.0001	RH
NUM1	− 3.60887	0.10131	− 35.62	< 0.0001	TEMP
NUM2	3.40734	0.07133	47.77	< 0.0001	DEWP
NUM3	0.0047355	0.0028287	1.67	0.0978	PRCP
NUM4	3.05656	0.73004	4.19	< 0.0001	ATMO
NUM5	− 3.08589	0.70871	− 4.35	< 0.0001	SLP
NUM6	0.20930	0.04896	4.28	< 0.0001	40ST


(a) Predict RH for 6 months.

(b) Predict RH for 12 months.

Fig. 8 RH forecast map based on SEE model

$$\begin{aligned}\nabla_{12}y_{RHt} = {} & -3.60887 * x_{TEMPt} + 3.40734 * x_{DEWPt} \\ & + 0.0047355 * x_{PRCPt} + 3.05656 * x_{ATMOt} \\ & - 3.08589 * x_{SLPt} + 0.20930 * x_{40STt} \\ & + (1 - 0.74194B)(1 - 0.66551B^{12})\varepsilon_t, \\ & Var(\varepsilon_t) = 1.459924.\end{aligned} \tag{15}$$

The prediction results of the SEE model are illustrated in Fig. 8.

We used the differential series of meteorological conditions that have a cointegration relationship with the RH and previous error series to construct an ECM model

$$\begin{aligned}\nabla\nabla_{12}y_{RHt} = {} & -3.57290\nabla\nabla_{12}x_{TEMPt} + 3.47467\nabla\nabla_{12}x_{DEWPt} \\ & + 0.00383\nabla\nabla_{12}x_{PRCPt} + 2.04517\nabla\nabla_{12}x_{ATMOt} \\ & - 2.08302\nabla\nabla_{12}x_{SLPt} + 0.09009\nabla\nabla_{12}x_{40STt} \\ & - 0.00011056ECM_{t-1} + \varepsilon_t,\end{aligned} \tag{16}$$

where $\nabla\nabla_{12}x_{TEMPt}$ is the first-order 12-step difference series of air temperature, $\nabla\nabla_{12}x_{DEWPt}$ is the first-order 12-step difference series of dew point temperature, $\nabla\nabla_{12}x_{PRCPt}$is the first-order 12-step difference series of precipitation, $\nabla\nabla_{12}x_{ATMOt}$ is the first-order 12-step difference series of atmospheric pressure, $\nabla\nabla_{12}x_{SLPt}$is the first-order 12-step difference series of sea-level presssure, $\nabla\nabla_{12}x_{40STt}$ is the first-order 12-step difference series of 40 cm soil temperature, $ECM_{t-1}$is the previous error series, and ε_t is the residual sequence of regression.

The results of the analysis of variance shown in Table 9 indicates that the equation was significantly linearly correlated, and the value of R^2 was 0.8889. Parameter estimations of ECM model shown in Table 10 indicate that the current fluctuations of TEMP, DEWP, ATMO, and SLP have a significant impact on the current fluctuations of RH, and the adjustment range of RH fluctuations is large. Their adjustments are, respectively, −3.57290, 3.47467, 2.04517, −2.08302 for a unit, which explains the short-term volatility relationship between the series. While, PRCP, 40ST, and previous period errors have no significant impact on current fluctuations, and the adjustment range of current fluctuations of RH is not large. Their adjustments are, respectively, 0.00383, 0.09009, and −0.00011056 for a unit.

Table 9 The results of the analysis

F value	*Pr* > *F*	RMSE	R^2	Df
701.21	< 0.0001	1.55281	0.9812	7

Table 10 Parameter estimation of ECM model

Variable	Parameter estimation	Standard error	*t* value	*Pr* > *t*
$\nabla\nabla_{12}x_{TEMP}$	− 3.57290	0.08791	− 40.64	< 0.0001
$\nabla\nabla_{12}x_{DEWP}$	3.47467	0.07534	46.12	< 0.0001
$\nabla\nabla_{12}x_{PRCP}$	0.00383	0.00286	1.34	0.1843
$\nabla\nabla_{12}x_{ATMO}$	2.04517	0.58081	3.52	0.0007
$\nabla\nabla_{12}x_{SLP}$	− 2.08302	0.56312	− 3.70	0.0004
$\nabla\nabla_{12}x_{40ST}$	0.09009	0.07096	1.27	0.2076
ECM_{t-1}	− 0.00011056	0.00078976	− 0.14	0.8890

The forecast evaluations of the SEE model, the additive seasonality model, the multiplicative seasonality model, and the LSTM model are shown in Table 11. For simple, we denote the additive seasonality model as ASM and the multiplicative seasonality model as MSM in the table. The model with minimum AIC, SBC, RMSE, RSE, and maximum R^2 is the optimal model. The optimal results have been boldly marked in Table 11. These results are discussed in the following section.

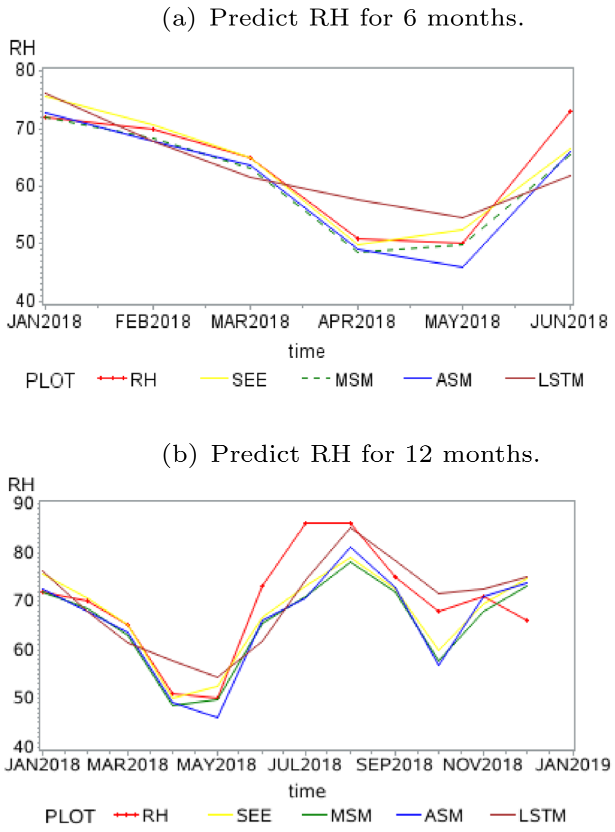
Discussion

This section discusses the modeling results. Figure 9 shows prediction of SEE model, the additive seasonality model, the multiplicative seasonality model, and the LSTM model for 6 months and 12 months. The radar charts shown in Fig. 10 summarize the performance of several methods in predicting RH for different periods and prepare for further discussion.

First, the SEE model performs better than the other models with minimal RMSE, RSE, and maximal R^2 (RMSE=3.1776, RSE=0.1111, R^2=0.8889, AIC=309.9675, and SBC=330.3138 for 6-month predicting;

Table 11 Performance metrics of four models

	ASM	MSM	SEE	LSTM
6 months				
AIC	683.5248	669.8517	309.9675	
SBC	688.6325	674.9594	330.3138	
RMSE	3.5226	3.4104	3.1776	6.096
RSE	0.1365	0.1279	0.1111	0.4087
R^2	0.8635	0.8721	0.8889	0.5913
12 months				
AIC	683.5248	669.8517	309.9675	
SBC	688.6325	674.9594	330.3138	
RMSE	6.5807	6.6724	5.946	6.2406
RSE	0.3841	0.3949	0.3136	0.3454
R^2	0.6159	0.6051	0.6864	0.6546



(a) Predict RH for 6 months.

(b) Predict RH for 12 months.

Fig. 9 Predicted and measured RH

RMSE=5.946, RSE=0.3136, R^2=0.6864, AIC=309.9675, and SBC=330.3138 for 12-month predicting). Compared with the multiplicative seasonality model, the SEE model performs better in fitting and predicting RH, resulting in 53.73% reduction in AIC, 51.06% reduction in SBC, 6.83% reduction in RMSE, 13.14% reduction in RSE, and 1.93% increase in R^2 for 6-month predicting; 10.89% reduction in RMSE, 20.59% reduction in RSE, and 13.44% increase in R^2 for 12-month predicting. The comparation between SEE model and additive seasonality model indicates that the SEE model results in 54.65% reduction in AIC, 52.03% reduction in SBC, 9.79% reduction in RMSE, 18.61% reduction in RSE, and 2.94% increase in R^2 for 6-month predicting; 9.64% reduction in RMSE, 18.35% reduction in RSE, and 11.45% increase in R^2 for 12-month predicting. AS for the LSTM model, the SEE model results in 47.87% reduction in RMSE, 72.82% reduction in RSE, and 50.33% increase in R^2 for 6-month predicting; 4.72% reduction in RMSE, 9.21% reduction in RSE, and 4.86% increase in R^2 for 12-month predicting. Moreover, the study confirms that when the prediction horizon is 6 months, the SARIMA model performs better than the artificial intelligence method with smaller MSE, RSE, and larger R^2, which is consistent with the results in the research of Aghelpour et al. (2021).

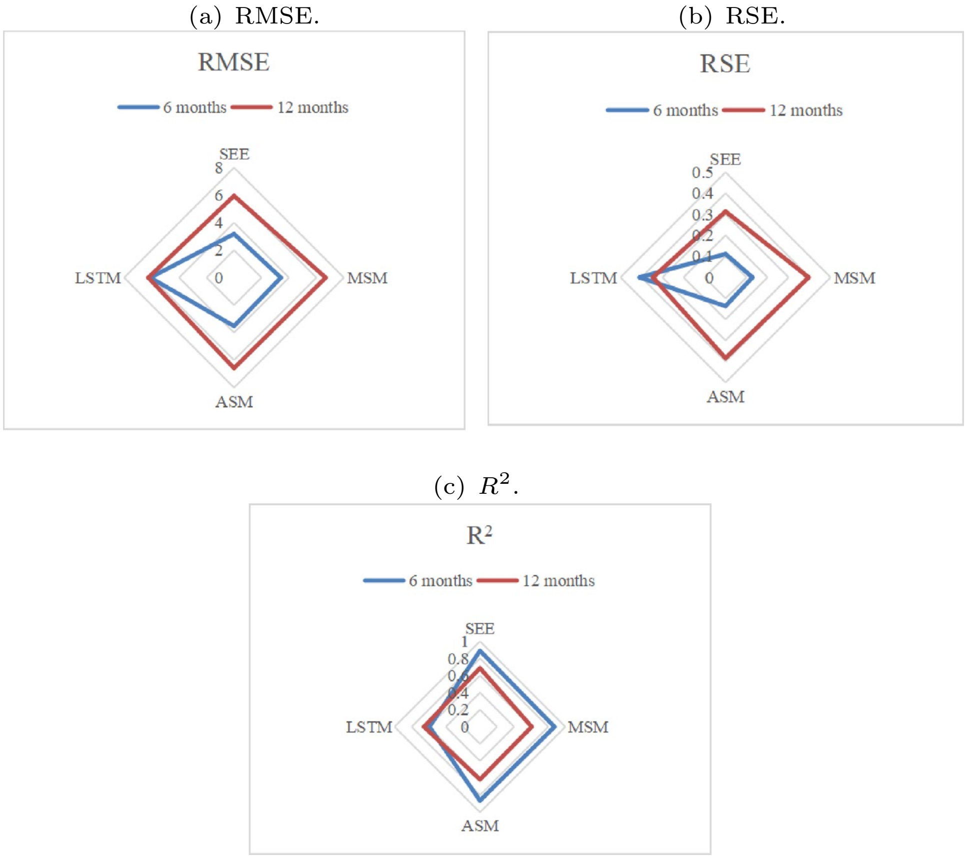
Second, incorporating EG theory and ECM into SARIMA model is able to increase the forecasting accuracy. SEE introduces EG theory and ECM based on SARIMA model. We perform cointegration tests on RH and other meteorological conditions, as shown in Table 8, and establish a cointegration model as indicated in formula (15). It shows that RH has a cointegration relationship with TEMP, DEWP, PRCP, ATMO, SLP, and 40ST, which reveals the long-term equilibrium relationship among series; Table 10 indicates that the current fluctuations of TEMP, DEWP, ATMO, and SLP have a significant impact on the current fluctuations of RH. Their adjustments are, respectively, −3.57290, 3.47467, 2.04517, −2.08302 for a unit, which explains the short-term volatility relationship between the series. In contrast, the performance of the SEE model is better than the SARIMA model including the multiplicative seasonality model and the additive seasonality model according to the value of AIC, SBC, RMSE, RSE, and R^2. The time series modeling of Li et al. (2021) also reveals that adding covariates can improve the prediction performance of ARIMA model.

Third, increasing the prediction horizon from 6 months to 12 months results in a decrease in the accuracy of the SEE model. According to Table 11, it can be calculated that the increase in the prediction horizon results in 22.78 %reduction in R^2 for 6-month predicting. Nevertheless, the SEE model still performs better than the other models in predicting RH.

Fourth, Fig. 9 illustrates that RH will decrease from January to April and from September to October. It will increase from May to August and from November to December. The RH will reach the minimum in April and the maximum in August. Studies of Shad et al. (2022) have similar discussions. The spread of respiratory diseases, such as COVID-19, is enhanced when the RH decreases (Mangla et al. 2021). Therefore, the prediction method proposed in this paper is helpful to prepare for the transmission and prevention of diseases that may occur in the future.

Conclusions

This paper proposes an SARIMA-EG-ECM model suitable for RH prediction. The accuracy of the model is evaluated by various statistical metrics. Monthly predictions of the RH in Hailun Agricultural Ecological Experimental Station in the next 6 months and 12 months have been carried out. It demonstrates that the SEE model performs better than the multiplicative seasonality, the additive seasonality model, and the LSTM model with minimal RMSE, RSE, and maximal


(a) RMSE.

(b) RSE.

(c) R^2.

Fig. 10 Comparation based on individual metrics

R^2. The SEE model has the following characteristics: it takes full account of seasonality of the time series; it shows that RH has a cointegration relationship with TEMP, DEWP, PRCP, ATMO, SLP, and 40ST, which reveals the long-term equilibrium relationship among series; it indicated that the current fluctuations of TEMP, DEWP, ATMO, and SLP have a significant impact on the current fluctuations of RH. Their adjustments are respectively: −3.57290, 3.47467, 2.04517, −2.08302 for a unit, which explains the short-term volatility relationship between the series.

The accuracy of the SEE model decreased slightly when the prediction horizon was increased from 6 to 12 months. Nevertheless, the SEE model still performs better than the other models in predicting RH. We can observe from the prediction results of the SEE model that there will be a decrease in the RH from January to April and from September to October in the next year. There will be an increase in the RH from May to August and from November to December. The RH will reach the minimum in April and reach the maximum in August. The results will help to evaluate the applicability of SEE model in predicting RH in the future development of this study.

Acknowledgements We would like to thank the referees for their careful reading and helpful comments. This work was supported in part by Chinese Universities Scientific Fund (No. 2,452,022,364), NSF of Shaanxi Province (No. 2023-JC-YB-083), and NSF of China (No. A010704).

Data availability statements The meteorological datasets used during the current study were observed by Hailun Agroecosystem Experimental Station, China Academic of Science (2009–2018) at the Science Data Bank.

Declarations

Conflict of interest The authors declare that they have no conflict of interest. Informed consent was obtained from all individual participants or organizations included in the study.

Ethical statement I certify that this manuscript is original and has not been published and will not be submitted elsewhere for publication while being considered by "Modeling Earth Systems and Environment" *Stochastic Environmental Research and Risk Assessment*'. And the study is not split up into several parts to increase the quantity of submissions and submitted to various journals or to one journal over time. No data have been fabricated or manipulated (including images) to support our conclusions. And authors whose names appear on the submission have contributed sufficiently to the scientific work and therefore share collective responsibility and accountability for the results.

References

Abdi AH, Warsame AA, Sheik-Ali IA (2022) Modelling the impacts of climate change on cereal crop production in east africa: evidence from heterogeneous panel cointegration analysis. Environmental Science and Pollution Research pp 1–12. https://doi.org/10.1007/s11356-022-24773-0

Aghelpour P, Singh VP, Varshavian V (2021) Time series prediction of seasonal precipitation in iran, using data-driven models: a comparison under different climatic conditions. Arabian J Geosci 14(7):1–14. https://doi.org/10.1007/s12517-021-06910-0

Akaike H (1973) Information theory as an extension of the maximum likelihood. Intersympon Inform Theory 1:610–624. https://doi.org/10.1007/978-1-4612-0919-5_38

Appiah K (2017) Statistical analysis of water level, temperature and humidity using cointegrated vector autoregression (var) models. PhD thesis, University of Ghana

Bas Cerdá MdC, Ortiz Moragón J, Ballesteros Pascual L et al (2017) Evaluation of a multiple linear regression model and sarima model in forecasting 7be air concentrations. Chemosphere 177:326–333. https://doi.org/10.1016/j.chemosphere.2017.03.029

Calvo-Polanco M, Ibort P, Molina S et al (2017) Ethylene sensitivity and relative air humidity regulate root hydraulic properties in tomato plants. Planta 246(5):987–997. https://doi.org/10.1007/s00425-017-2746-0

Casallas A, Ferro C, Celis N, et al (2021) Long short-term memory artificial neural network approach to forecast meteorology and pm2. 5 local variables in bogotá, colombia. Model Earth Syst Environm pp 1–14. https://doi.org/10.1007/s40808-021-01274-6

Cong J, Ren M, Xie S et al (2019) Predicting seasonal influenza based on sarima model, in mainland china from 2005 to 2018. Int J Environm Res Public Health 16(23):4760. https://doi.org/10.3390/ijerph16234760

Csavina J, Field J, Félix O et al (2014) Effect of wind speed and relative humidity on atmospheric dust concentrations in semi-arid climates. Sci Total Environm 487:82–90. https://doi.org/10.1016/j.scitotenv.2014.03.138

Danhui Y (2019) Apply Time Series Analysis. China Renmin University Press, Beijing

Duan J, Wang X, Zhao D et al (2019) Risk effects of high and low relative humidity on allergic rhinitis: Time series study. Environm Res 173:373–378. https://doi.org/10.1016/j.envres.2019.03.040

Engle RF, Granger CW (1987) Co-integration and error correction: representation, estimation, and testing. Econometrica pp 251–276. https://doi.org/10.2307/1913236

Eymen A, Köylü Ü (2019) Seasonal trend analysis and arima modeling of relative humidity and wind speed time series around yamula dam. Meteorol Atmospheric Phys 131(3):601–612. https://doi.org/10.1007/s00703-018-0591-8

Fang T, Lahdelma R (2016) Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system. Appl Energy 179:544–552. https://doi.org/10.1016/j.apenergy.2016.06.133

Fernández-González M, Ramos-Valcárcel D, Aira MJ et al (2016) Prediction of biological sensors appearance with arima models as a tool for integrated pest management protocols. Ann Agricul Environm Med 10.5604/12321966.1196868

Gao W, Gao J, Yang L et al (2021) A novel modeling strategy of weighted mean temperature in china using rnn and lstm. Remote Sensing 13(15):3004. https://doi.org/10.3390/rs13153004

Granger CWJ, Swanson N (2010) Future developments in the study of cointegrated variables*. Oxford Bull Econom Statist 58(3):537–553

Hendry DF, Anderson GJ (1977) Testing dynamic specification in small simultaneous systems: An application to a model of building society behavior in the united kingdom. Front Quantit Econom pp 361–383

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Htet MS (2017) Airline error correction model and its application to forecast the california carbon monoxide, precipitation, and air temperature. PhD thesis, Southeast Missouri State University

Hutapea MI, Pratiwi YY, Sarkis IM, et al (2020) Prediction of relative humidity based on long short-term memory network. In: AIP Conference Proceedings, AIP Publishing LLC, p 060003, https://doi.org/10.1063/5.0003171

Ivanova A (2009) An experience of the humidity forecasts verification and assessment of their applicability in forecasting of the aircraft icing zones. Russian Meteorol Hydrol 34(6):354–363. https://doi.org/10.3103/S106837390906003X

Jones P, Trenberth K, Ambenje P, et al (2007) Observations: surface and atmospheric climate change. Climate change pp 235–336

Kuzugudenli E (2018) Relative humidity modeling with artificial neural networks. Appl Ecol Environm Res 16(4), 5227–5235. https://doi.org/10.15666/aeer/1604_52275235

Kwon Y, Kwasinski A, Kwasinski A (2019) Solar irradiance forecast using naïve bayes classifier based on publicly available weather forecasting variables. Energies 12(8):1529. https://doi.org/10.3390/en12081529

Li F, Wang Z, Liu G (2013) Towards an error correction model for dam monitoring data analysis based on cointegration theory. Structural Safety 43:12–20. https://doi.org/10.1016/j.strusafe.2013.02.005

Li M, Hu B, Han X, et al (2019a) 2009-2018 hailun agricultural ecological experimental station meteorological data set of chinese academy of sciences. China Scientific Data

Li YR, Han TT, Wang JX, et al (2021) Application of arima model for mid-and long-term forecasting of ozone concentration. Huan Jing ke Xue= Huanjing Kexue 42(7):3118–3126. https://doi.org/10.13227/j.hjkx.202011237

Li Z, Zou H, Qi B (2019b) Application of arima and lstm in relative humidity prediction. In: 2019 IEEE 19th International Conference on Communication Technology (ICCT), IEEE, pp 1544–1549, https://doi.org/10.1109/ICCT46805.2019.8947142

Lu T, Viljanen M (2009) Prediction of indoor temperature and relative humidity using neural network models: model comparison. Neural Comput Appl 18(4):345–357. https://doi.org/10.1007/s00521-008-0185-3

Ma T, Zhou Z, Abdulhai B (2015) Nonlinear multivariate time-space threshold vector error correction model for short term traffic state prediction. Transport Res Part B Methodol 76:27–47. https://doi.org/10.1016/j.trb.2015.02.008

Mangla S, Pathak AK, Arshad M et al (2021) Impact of environmental indicators on the covid-19 pandemic in delhi, india. Pathogens. https://doi.org/10.3390/pathogens10081003

Murthy KN, Saravana R, Kumar KV (2018) Modeling and forecasting rainfall patterns of southwest monsoons in north-east india as a sarima process. Meteorol Atmos Phys 130(1):99–106. https://doi.org/10.1007/s00703-017-0504-2

Nian D, Deng Q, Fu Z (2018) Research progress of relative humidity and its changing annual cycle. Adv Earth Sci

Qiu H, Zhao H, Xiang H et al (2021) Forecasting the incidence of mumps in chongqing based on a sarima model. BMC Public Health 21(1):1–12. https://doi.org/10.1186/s12889-021-10383-x

Rathod S, Singh K, Arya P et al (2017) Forecasting maize yield using arima-genetic algorithm approach. Outlook Agric 46(4):265–271. https://doi.org/10.1177/0030727017744933

Schwarz GE (1978) Estimating the dimension of a model. Ann Stat. https://doi.org/10.1007/978-1-4612-0919-5_38

Shad M, Sharma Y, Singh A (2022) Forecasting of monthly relative humidity in delhi, india, using sarima and ann models. Model Earth Syst Environm. https://doi.org/10.1007/s40808-022-01385-8

Tkacz G (2001) Neural network forecasting of canadian gdp growth. Int J Forecast 17(1):57–69. https://doi.org/10.1016/S0169-2070(00)00063-7

Valipour M (2015) Long-term runoff study using sarima and arima models in the united states. Meteorol Appl 22(3):592–598. https://doi.org/10.1002/met.1491

Xie B, Zhang Q, Ying Y (2011) Trends in precipitable water and relative humidity in china: 1979–2005. J Appl Meteorol Climatol 50(10):1985–1994. https://doi.org/10.1175/2011JAMC2446.1

Xie J, Chen Y, Hong T et al (2018) Relative humidity for load forecasting models. IEEE Transact Smart Grid 9(1):191–198. https://doi.org/10.1109/TSG.2016.2547964

Yin S, Liu H, Duan Z (2021) Hourly pm2. 5 concentration multi-step forecasting method based on extreme learning machine, boosting algorithm and error correction model. Digital Signal Process 118:103–221. https://doi.org/10.1016/j.dsp.2021.103221

Yu X (2009) Indication of relative humidity of ecmwf in precipitation forecast in hainan prefecture. Qinghai Meteorol 3:17–20

Zhang J, Zhao Y, Xiao W (2015) Multi-resolution cointegration prediction for runoff and sediment load. Water Resour Manag 29(10):3601–3613. https://doi.org/10.1007/s11269-015-1018-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.