Effectiveness of Cascading Time Series Models Based on Meteorological Factors in Improving Health Risk Prediction

Shuopeng Jia  
China Pharmaceutical University

Weibin She  
Dongguan Kanghua Hospital

Zhipeng Pi  
China Pharmaceutical University

Buying Niu  
China Pharmaceutical University

Jinhua Zhang  
Meteorological Bureau of Panyu District

Xihan Lin  
China Pharmaceutical University

Mingjun Xu  
China Pharmaceutical University

Weiya She  
Meteorological Bureau of Panyu District

Jun Liao (liaojun@cpu.edu.cn)  
China Pharmaceutical University  https://orcid.org/0000-0003-0617-5840

Research Article

Keywords: LRTIs, health risk, meteorological factors, time series, forecast

DOI: https://doi.org/10.21203/rs.3.rs-544703/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Effectiveness of cascading time series models based on meteorological factors in improving health risk prediction

Shuopeng Jia, MS\textsuperscript{1*}, Weibin She, MD\textsuperscript{2*}, Zhipeng Pi, BS\textsuperscript{3}, Buying Niu, BS\textsuperscript{4}, Jinhua Zhang, MS\textsuperscript{5}, Xihan Lin, MS\textsuperscript{1}, Mingjun Xu, MS\textsuperscript{1}, Weiya She, BS\textsuperscript{5}, Jun Liao, PhD\textsuperscript{4*}

\textsuperscript{1} School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, China
\textsuperscript{2} Dongguan Kanghua Hospital, Dongguan, China
\textsuperscript{3} School of Pharmacy, China Pharmaceutical University, Nanjing, China
\textsuperscript{4} School of Science, China Pharmaceutical University, Nanjing, China
\textsuperscript{5} Meteorological Bureau of Panyu District, Guangzhou, China

\textsuperscript{*} Shuopeng Jia and Weibin She contributed equally to this work.
\textsuperscript{*} Corresponding author: Jun Liao

Tel: 8613952040425; Fax: +862586185122; Email: liaojun@cpu.edu.cn

Address:
\textsuperscript{135} China Pharmaceutical University, #639 Longmian Avenue, Jiangning District, Nanjing 211198, Jiangsu province, China
\textsuperscript{2} Dongguan Kanghua Hospital, #1000 Dongguan Avenue, Dongguan 523080, Guangdong province, China
\textsuperscript{4} Meteorological Bureau of Panyu District, #5 Landscape Avenue, Hengjiang village, Shatou Street, Panyu district, Guangzhou 511400, Guangdong province, China
Abstract

Meteorological factors, which are cyclical and regular, have an essential impact on human health. Accurate health risk prediction based on meteorological factors is essential for optimal resource allocation in healthcare units. However, due to the non-stationary and nonlinear of the original hospitalization sequence, traditional methods are less robust to predict it. This study aims to investigate hospital admission prediction models using time series pre-processing algorithms and deep learning approach based on Meteorological factors. Using the hospital electronic medical record data of Panyu Central Hospital and meteorological data of Panyu district from 2003-2019, 46,089 eligible patients with lower respiratory tract infections (LRTIs) and four meteorological factors were identified to build and evaluate the prediction models. A novel hybrid model, Cascade GAM-CEEMDAN-LSTM Model (CGCLM) was established in combination with Generalized Additive Model (GAM), Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Long-Short Term Memory Networks (LSTM) for predicting daily admissions of patients with LRTIs. The experimental results show that the CGCLM multistep method proposed in this paper outperforms the single LSTM model in the prediction of health risk time series at different time window sizes. Moreover, our results also indicate that the CGCLM has the best prediction performance when the time window is set to 61 days (RMSE=1.12, MAE=0.87, $R^2=0.93$). Adequate extraction of exposure-response relationships between meteorological factors and diseases and suitable handling of sequence pre-processing have an important role in time series prediction. This hybrid climate-based model to LRTIs disease prediction can also be extended to the time series prediction of other epidemic disease.

Key Words: LRTIs, health risk, meteorological factors, time series, forecast
1. Introduction

Climate change causes or aggravates a wide range of exposures with multiple impacts on health, both direct and indirect (Linares, et al., 2020). The 2020 report of The Lancet Countdown on health and climate change states that no country, rich or poor, is exempt from the effects of climate change on health (Watts, et al., 2020). LRTIs are a broad description of a group of respiratory diseases that comprise acute bronchitis, pneumonia, and exacerbations of chronic lung disease (Woodhead, et al., 2011) and it has become the fourth most common cause of death according to the World Health Organization (WHO) Global Health Estimates 2019 (WHO, 2020). In the context of continuing global climate change, identifying factors influencing the occurrence of LRTIs could contribute to the prediction of future outbreaks and facilitate the development of transmission prevention measures.

In the case of LRTIs, meteorological factors not only affect the activity of influenza viruses and make their activities seasonal (Price, et al., 2019), but also affect the immune system and alter the susceptibility of the body (Mäkinen, et al., 2009). As a result, the incidence of LRTIs also presents a certain rule. A study in Shaanxi Province, China, using stepwise multiple linear regression found that LRTIs hospitalization rate was significantly correlated with air temperature, relative humidity and atmospheric pressure (Liu, et al., 2016). Besides, a study conducted in Finland using a generalized additive model found that low temperature and low relative humidity associated with increased incidence of LRTIs (Mäkinen, et al., 2009). Although these statistical regression models can assess the relationship between meteorological factors and the incidence of LRTIs, they generally do not provide sufficiently accurate incidence forecasts for medical management.

Traditional time series forecasting models are usually statistical models, such as the classic
Autoregressive Integrated Moving Average (ARIMA). ARIMA has been widely applied in the field of medical time series forecasting (Zhang, et al., 2015). Although ARIMA models are quite flexible, their major limitation is that the nonlinear effects of various complex factors cannot be captured by ARIMA (Zhang, 2003). As machine learning algorithms evolved, it proved capable of capturing non-linear hidden features in time series. One study used machine learning to predict cardiovascular admissions, and they have achieved excellent results by defining a threshold of 85% that translates into a classification problem (Qiu, et al., 2020). GU et al (Gu, et al., 2019) used the optimized three-layer LSTM model to predict the incidence of Hand-foot-mouth disease (HFMD) in 14 administrative regions of Guangxi, China, and confirmed that LSTM has a good predictive performance. Navares et al (Navares, et al., 2020) proposed an algorithm based on environmental indicators to predict the number of daily hospital admissions due to respiratory and circulatory diseases using LSTM and Convolutional Neural Network (CNN) networks, providing a new way to solve the problem of hospital admission prediction.

In this paper, we collected 46,089 cases of LRTIs in Panyu Central Hospital from January 1, 2003 to July 15, 2019. Meteorological data including temperature, atmospheric pressure (AP), relative humidity and wind velocity were also collected from Panyu District, Guangzhou during the same period. In this paper a multi-step cascade prediction method (CGCLM) combined with GAM, CEEMADAN, and LSTM was established to predict short-term LRTIs admissions by incorporating the local exposure-response relationship between weather and disease. The main contribution of this paper is to validate the feasibility and effectiveness of the CGCLM approach. Also, the method has been shown to be effective in improving the accuracy of short-term hospital admissions prediction.

2. Materials and methods
2.1 Data collection

Data on the daily number of hospitalized patients with lower respiratory tract infection were obtained from Panyu Central Hospital. This data includes the number of inpatients with the first diagnosis of LRTIs (International Statistical Classification of Diseases and Health Related Problems-10: J12.0-J22.0) in Panyu Central Hospital on 1 January 2003 and 15 July 2019. Daily meteorological data including mean temperature (°C), relative humidity (%), AP (hPa) and wind velocity (m/s) for the same time were collected from Guangzhou Meteorological Bureau. A total of 6,040 days and 46,089 cases were included in the study. All patients' private information was desensitized.

2.2 Data processing

During the study period, the proportion of missing values of each meteorological variable was the temperature of 0.05% (3/6040), relative humidity 0.05% (3/6040) and atmospheric pressure 0.116% (7/6040) respectively. We used the average value method to fill the missing values of meteorological conditions. The time-series features were extracted from date, including year, month (month of year), half of month, three-part of month, day (day of month), and DOW (day of week).

2.3 Methodology

An overview diagram of the experimental design is shown in Figure1, with the following implementation steps.

(1) The daily number of patients admitted for LRTIs in Panyu Central Hospital from January 1, 2003 to July 15, 2019 and the local meteorological parameters in the same period were obtained as modeling data.

(2) The meteorological data are interpolated with vacancy values and the exposure-response
relationship model is built using GAM to obtain the maximum relative risk (RR) and the corresponding time lag for each meteorological variable. According to the different maximum lag days of the four meteorological factors, a moving average is processed as the input of the LSTM in step (6) respectively.

(3) CEEMDAN was applied to decompose the original admission time series to obtain 9 Intrinsic Mode Functions (IMF) components and 1 trend component.

(4) The sample entropy values of each IMF component were calculated separately, while the subsequences with similar sample entropy values were merged and recombined to obtain five new subsequences with significant differences in complexity.

(5) To improve the prediction accuracy of the model, the new subseries and meteorological parameters are normalized before the data are fed into the model by the following formula:

\[ X'(i) = \frac{(X(i) - \min\{X(i)\})}{(\max\{X(i)\} - \min\{X(i)\})} \]  

(1)

Where \( X(i) \) is the original data; \( X'(i) \) is the normalized data.

(6) LSTM neural network prediction models of corresponding parameter space were established for each new sub-sequence by adding meteorological factors, and the predicted values of each model were output.

(7) The five predicted values obtained in step (4) are back-normalized and their arithmetic sums are calculated to obtain the predicted daily admissions for LRTIs.
Figure 1 The overview of the research framework, which can be divided into three parts. The first part includes data used in the experiment, data preprocessing and feature extraction. The second part is the decomposed time series features and the use of sample entropy algorithm to do feature clustering. The third part is the independent prediction of each feature using LSTM as a prediction model, and the final output of the combined prediction results.
2.4 Generalized Additive Model (GAM)

GAM was first proposed by Hastie and Tibshirani (Hastie and Tibshirani, 1987) as an extension of the generalised linear model (GLM), which has significant advantages in dealing with complex nonlinear relationships between variables in time series analysis. GAM can be used to capture the exposure-response relationship between meteorological factors and LRTIs in Guangzhou, and to obtain the lagged characteristics of meteorological factors. Considering that the number of hospital admissions for LRTIs obeys a Poisson distribution, a univariate model using a logarithmic function (log) was constructed. The model expression is as follows:

\[
Y_t \sim \text{Poisson}(\mu_t):
\]
\[
\log(\mu_t) = \beta X_t + ns(time, df = 6) + DOW + \alpha
\]

Where \( t \) is the day of the observation; \( Y_t \) is the number of hospital admissions observed on day \( t \); \( \log \) is the logarithmic transformation of \( \mu_t \); \( X_t \) represents the individual variables, while \( \beta \) represents their coefficients; \( DOW \) is the day of the week and \( \alpha \) is the intercept and time was introduced together as confounders. The regression coefficients (\( \beta \)), standard error (SE) and quad distance (IQR) for each meteorological factor were estimated according to the model, and the RR and 95% confidence intervals (95% CI) were calculated. RR and 95% CI were calculated as follows:

\[
RR = \exp(\beta \times IQR)
\]

\[
95\% CI = \exp[(\beta \pm 1.96SE) \times IQR]
\]

2.5 Complete Ensemble Empirical Mode Decomposition with Adaptive Noise

CEEMDAN is a noise-assisted data analysis method that addresses the shortcomings of Empirical Mode Decomposition (EMD) and Ensemble Empirical Mode Decomposition (EEMD). EMD is a signal processing method based on instantaneous frequency proposed by Norden E. Huang (Huang, et al., 1998).
This method can decompose the signal into different IMF according to the time-scale characteristics of the data itself, breaking the limitation of traditional data methods that need to set the basis function in advance.

The data is treated as an IMF, and each IMF needs to satisfy two conditions: I. the number of local extreme points and the number of cross-zero points are equal or differ by 1 over the whole time range; II. for any zero points, the mean envelope of the local maximum and local minimum values are 0, respectively. For the original time series s(t), the EMD decomposition process is as follows:

1. Search for all extreme points of the original signal sequence.
2. The upper envelope $x_{\text{max}}(t)$ and lower envelope $x_{\text{min}}(t)$ of the signal are obtained by cubic spline interpolation and their mean values $n(t)$ are calculated.

$$n(t) = \frac{x_{\text{max}}(t) + x_{\text{min}}(t)}{2} \quad (5)$$

3. Calculate the difference between $s(t)$ and $n(t)$

$$h(t) = s(t) - n(t) \quad (6)$$

4. Test $h(t)$: (a) If $h(t)$ satisfies the two rules of judgment for IMFs, then $h(t)$ is defined as the $i$th IMF. (b) If $h(t)$ does not satisfy these two rules, then $h(t)$ is not an IMF, so the iterative process continues with $h(t)$ instead of $s(t)$.

5. Repeat steps 1 to 4 until the residual term $r(t)$ becomes a monotonic function or the number of extremes is less than or equal to 1, and no more IMFs can be extracted.

6. Eventually, the original signal sequence is decomposed into multiple IMFs and a and one trend item, which can be expressed as:

$$s_t = \sum_{i=1}^{n} \text{imfi}(t) + r_n(t) \quad (7)$$

Where $s_t$ is the original signal, imfi is the $i$th IMF component, $r_n(t)$ is the trend term.
In order to solve the modal aliasing problem in EMD decomposition, Wu et al. (Wu and Huang, 2009) proposed the EEMD method by introducing auxiliary noise with uniform frequency distribution in the EMD decomposition. The steps of EEMD decomposition are as follows: add different auxiliary white noise with uniform frequency distribution to the target signal each time, then use EMD to decompose the signal with additional white noise, and repeat the above process N times, finally integrate the average of the IMFs and trend terms obtained from the decomposition to obtain the final decomposition result of the original signal. This method avoids the appearance of modal mixing. However, in practice, the white noise added by EEMD is averaged over several times, but it is not offset completely. By adding adaptive white noise at each decomposition stage, the CEEMDAN method achieves a reconstruction error of almost zero with fewer averaging times (Torres, et al., 2011). Thus, the CEEMDAN method overcomes the modal mixing of EMD and also solves the incompleteness of the EEMD decomposition and the computational inefficiency caused by the need to reduce the reconstruction error by increasing the averaging times. The calculation process of CEEMDAN is listed in the supplementary material.

2.6 Sample Entropy

Sample Entropy (SE) is a metric for determining the complexity of time series proposed in 2000 by Richman et al. (Richman and Moorman, 2000). SE is more relatively consistent than Approximately Entropy (AE) and has been used successfully in the analysis of the complexity of biological signal sequences.

The estimated value of the sample entropy is calculated by the following formula:

\[
SampEn(m, r, N) = -\ln \frac{B^{m+1}(t)}{B^m(t)}
\]

(8)

Where the range of values for each IMF component sequence is \( \{x(i)|1 < i < N\} \), \( m \) is the mode
dimension and \( r \) is the similarity tolerance. In this study, the maximum epoch length (m) is set to 2 and the default tolerance (r) is given as 0.2.

### 2.7 Long Short-Term Memory Networks

LSTM, as a special Recursive Neural Network (RNN), incorporates sequence correlation, where the current state during the learning process will contain all historical information from previous time series, has the ability to learn long-term dependencies and solves the problem of RNN gradient explosion (Gers, et al., 2000). The LSTM neural network structure consists of a series of loops of connected subnetworks (i.e., memory modules), each containing one or more self-linked cells (cells), and control information. A system of three thresholds for mobile input gates \( i_t \), output gates \( o_t \) and forgetting gates \( f_t \). When the input value \( x_t \) is input to the LSTM cell in equation (9), the function \( f_t \) determines the information to be discarded and retained in the cell layer. In equation (10,11), the input value \( x_t \) goes through \( \sigma \) and \( \tanh \) respectively to determine the value to be renewed and generate new candidate values for renewal. In equation (12) the cell layer \( \tilde{C}_t \) is updated using \( f_t, i_t, \) and \( C \). In Equation (13), the cell layer’s information is used and \( o_t \) is the output. In Equation (14) the updated unit state \( \tilde{C}_t \) goes through the \( \tanh \) function along with the output gate \( o_t \) for the final output \( h_t \). In an LSTM network, the steps of implementation can be summarized in the following equations:

\[
f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)
\]

\[
i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)
\]

\[
C = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)
\]

\[
\tilde{C}_t = f_t \ast C_{t-1} + i_t \ast C
\]

\[
o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)
\]
\[ h_t = o_t \ast \tanh(\bar{C}_t) \]  

(14)

Where \( x_t \) and \( h_t \) are input vector and output vector, \( f_t, o_t, h_t \) stand for forgetting gate, input gate and output gate respectively; \( C_{t-1} \) and \( \bar{C}_t \) are the last moment and current unit state; \( h_{t-1} \) and \( h_t \) are the last moment and current hidden layer unit output; \( \sigma \) and \( \tanh \) stand for sigmoid activation and tangent functions, \( W \) and \( b \) are weight matrices and deviation vectors.

The LSTM network models involved in this research are optimized with common values. The parameters of the models are shown in Table 1.

### Table 1 Details of the parameters of the model.

| Model | Symbol | Meaning                     | Value   | Reason            |
|-------|--------|-----------------------------|---------|-------------------|
| LSTM  | \( n_i \) | Number of input layer nodes | -       | Number of feature inputs |
|       | \( n_h \) | Number of hidden layer nodes | 4       | Common value [2, 4, 8, 10,…] |
|       | \( n_o \) | Number of output layer nodes | 1       | Time series regression |
|       | \( \mu \) | Fixed learning rate         | 0.005   | Common value [0.001, 0.005, 0.01, 0.05,…] |
|       | \( T \)  | Size of batch               | 32      | Common value [8, 16, 32, 50, 100,…] |
|       | \( E_p \) | Epochs of training          | 100     | Converged         |

2.8 Sliding time window

In implementing the prediction process, this paper designs a sliding time window regarding Jacinta et al. (Jacinta, et al., 2016), as shown in Figure 2. The strength of this mode is that the data in the current window is being updated constantly, and deleting some of the old data and adding new data. Thus, the sliding time window pattern has the property of updating the data in real-time dynamically. The prediction process implements a segmented prediction approach to ensure the generalizability and usefulness of the model. The process consists of three parts: the training part is used to train the model and update the model parameters, with a data set length of 365*3 days; the validation part is used to adjust the hyper-parameters to obtain the best settings of the model parameters; and the last part is the
test part, which uses the best model to predict the data. The validation and test windows are set to the same size.

**Figure 2** The composition of sliding time window and the distribution of training set, validation set and testing set. The length of time for N were 14 days, 31 days, 61 days and 91 days.

### 2.9 Assessment criteria

The predictions were compared to the actual data and the errors were calculated separately for the single LSTM model and the CGCLM hybrid model. We used three measures: Root Mean Squared Error (RMSE), which it has traditionally been widely used in forecasting evaluation (Armstrong and Collopy, 1992); Mean Absolute Error (MAE), because MAE is less sensitive to outliers in predicted values than RMSE (Hyndman and Koehler, 2006) which can better reflect the actual situation of prediction error. R-square, for the reason it is as a measure of fit or model variability in health-forecasting-related studies is very common in the literature (Soyiri and Reidpath, 2012).

RMSE, MAE and $R^2$ can be expressed as:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]  
(15)

\[
MAE(X,h) = \frac{1}{n} \sum_{i=1}^{n} |h(x_i) - y_i|
\]  
(16)

\[
R^2 = 1 - \frac{\sum (y^{(i)} - \hat{y}^{(i)})^2}{(\bar{y} - y^{(i)})^2}
\]  
(17)
The GAM model was built using the "mgcv" package from R software version 3.6.2. The CEEEMDAN algorithm, SE algorithm and all deep learning models are implemented on python version 3.6.

3. Results

3.1 Data description

A total of 46089 patients with LRTIs were incorporated into the study from January 2003 to July 2019. In terms of age distribution, the majority of patients were less than 18 years old (26,261 cases, 57.0%), while male patients (28,177 cases, 61.1%) were more than female patients (17,912 cases, 38.9%) in terms of gender.

Table 2 Statistical characteristics of LRTIs patients admitted to Panyu Central Hospital, Guangzhou from January 2003 to July 2019.

| Daily hospital admission | Number | Percentage |
|------------------------|--------|------------|
| **Age**                |        |            |
| 18<                    | 26261  | 57.0%      |
| 18-65                  | 9196   | 20.0%      |
| >65                    | 10632  | 23.1%      |
| **Gender**             |        |            |
| Male                   | 28177  | 61.1%      |
| Female                 | 17912  | 38.9%      |
| **Total**              | 46089  | 100%       |

The daily number of LRTIs cases and the distribution of four meteorological variables in Panyu are shown in Table 3. The average daily number of hospital admissions in Panyu was 7.64. The average daily temperature (T), relative humidity, AP and wind velocity were 23.11°C, 74.81%, 1011.33 hPa and 1.90 m/s, respectively.

Table 3 Statistics on the daily number of LRTIs patients admitted to Panyu Central Hospital and daily
air temperature, daily relative humidity, daily air atmospheric pressure and daily wind velocity in Panyu District from January 2003 to July 2019.

| Variables          | Mean     | SD       | Min | 25% | 50% | 75% | Max | IQR |
|--------------------|----------|----------|-----|-----|-----|-----|-----|-----|
| Number             | 7.64     | 4.298    | 0   | 4.0 | 7.0 | 10.0| 33.0| 6.0 |
| Temperature        | 23.108   | 6.1386   | 4.3 | 18.600 | 24.500 | 28.200 | 34.1 | 9.6 |
| Relative humidity  | 74.805   | 12.529   | 21.0| 68.000 | 76.000 | 84.000 | 99.0 | 16  |
| AP                 | 1011.33  | 6.8247   | 990.| 1006.10 | 1011.20 | 1016.50 | 1035 | 10.4 |
| Wind velocity      | 1.900    | 0.8867   | 0   | 1.300 | 1.700 | 2.300 | 7.7 | 1.0 |

3.2 Relationship results of the risk exposure response

Figure 3 shows four forest plots of the effect of temperature, relative humidity, atmospheric pressure, and wind velocity on the admission RR and 95% CI of LRTIs. The results demonstrated that temperature, relative humidity, atmospheric pressure and wind velocity were correlated with LRTIs admission, with the maximum RR of lag5 (RR= 1.015, 95%CI, 1.001 to 1.03), lag0 (RR=1.07, 95%CI, 1.058 to 1.83), lag0 (RR=0.978, 95%CI, 0.964 to 0.991) and lag3 (RR=1.026, 95%CI, 1.014 to 1.039), respectively.
Figure 3 Forest plots of the effects of temperature, relative humidity, atmospheric pressure, and wind velocity on admission RR and 95% CI of LRTIs. Lag0 lag1, lag2 lag3, lag4, lag5 represent the zeroth day, the first day, the second day, the third day, the fourth day, the fifth day of lag effect.

3.3 CEEMDAN decomposition results of the original admission sequence

According to the CEEMDAN decomposition process described above, the daily admission sequence of LRTIs was decomposed. The results are shown in Figure 4, 9 IMFs and 1 residual term are arranged from top to bottom, in which the horizontal axis represents time and the vertical axis represents the frequency of each IMF. The frequency from IMF1-IMF9 to the residual term gradually decreases, and the change mode is simpler than the original sequence. The bar chart depicts the Pearson correlation.
coefficients of the nine IMF series and one residual term with the original series.

**Figure 4** The CEEMDAN decomposition consequence of LRTIs admission and the correlation coefficients of each sub-sequence.

### 3.4 Reconstruction of decomposed sub-sequences via SE

In order to reduce the complexity of LSTM prediction modeling appropriately and avoid model overfitting to predict the number of daily admissions effectively, the sample entropy theory was used to evaluate the complexity of each IMF component obtained from CEEMDAN decomposition, and the SE values of each remaining component were calculated separately (Xie, et al., 2019). As shown in Figure 5 the IMFs were reorganized according to the sample entropy values, (IMF1), (IMF2), (IMF3), (IMF4, IMF5, IMF6) and (IMF7, IMF8, IMF9, Residual) can be reconstructed to NEW1, NEW2, NEW3, NEW4 and NEW5, which is shown in right tendency chart.
Figure 5 The sample entropy result of each sub-sequence of LRTIs admission and reconstruction consequence.

3.5 Forecast results of hospital admissions

RMSE and $R^2$ values of five NEW IMF predicted values corresponding to four time Windows with different sizes are shown in Figure S1.

As the first part of the experiment, we adopt a single LSTM model for prediction, and the prediction results of different time windows are shown in Table 4. The RMSE values of 14-day, 31-day, 61-day and 91-day are 1.92, 1.86, 1.73 and 1.60 respectively, the MAE values are 1.53, 1.47, 1.35 and 1.23, and the $R^2$ values are 0.79, 0.80, 0.83 and 0.85 respectively. The second part of the experiment aims to reduce the errors produced by a single LSTM module and improve the accuracy of the predictions. CEEMDAN decomposition is introduced to fully extract the temporal features. Table 3 shows the comparison between the predicted value of single LSTM and CGCLM method. It can be clearly seen that the CGCLM method, which includes the CEEMDAN module, helps to better extract time-series features to make more accurate predictions, with the 14-day predicted MSE value decreasing from 1.92 to 1.25, the MAE value decreasing from 1.53 to 0.98, and the coefficient of determination $R^2$ improving from 0.79 to 0.91. The 31-day predicted RMSE decreased
from 1.86 to 1.15, the MAE decreased from 1.47 to 1.90, and the $R^2$ increased from 0.80 to 0.92.

The 61-day predicted RMSE value dropped from 1.73 to 1.12, the MAE value reduced from 1.35 to 0.87, and the $R^2$ rose from 0.83 to 0.93. The 91-day predicted RMSE value fell from 1.60 to 1.30, the MAE value declined from 1.23 to 0.99, and the $R^2$ improved from 0.85 to 0.90. Overall, model prediction performance is best when the prediction window size is 61 days using the CGCLM method, and the trend plot of predicted versus true values is shown in Figure 6. In addition, we calculated the RMSE, MAE, and $R^2$ separately for each year, also showing the excellent performance of the CGCLM method as shown in Table S1.

| Metrics | CGCLM | LSTM |
|---------|-------|------|
|         | 14days | 31days | 61days | 91days | 14days | 31days | 61days | 91days |
| RMSE    | 1.25   | 1.15   | 1.12   | 1.30   | 1.92   | 1.86   | 1.73   | 1.60   |
| MAE     | 0.98   | 0.90   | 0.87   | 0.99   | 1.53   | 1.47   | 1.35   | 1.23   |
| $R^2$   | 0.91   | 0.92   | 0.93   | 0.90   | 0.79   | 0.80   | 0.83   | 0.85   |

Table 4 Year-by-year comparison of the performance of CGCLM and single LSTM models for different window sizes.
Figure 6 Waterfall plot of true values (green curve), CGCLM model predicted values (orange curve) and single LSTM model predicted values (purple curve) for different windows.

To further evaluate the performance of two models, the predicted value distribution of the CGCLM was compared with the single LSTM model. Figure S2 shows the distribution of prediction results of two different models and different predicted time size. Obviously, the eight models had different prediction results for Panyu Center Hospital LRTIs daily admission time series. In order to compare the results, we set three calibration lines at the maximum, average and minimum respectively. It can be clearly seen that the predictions of CGCLM for maximum, mean and minimum are closer to the true values. However, the single LSTM model generally predicts a large value of maximum value, a small value of minimum value, and slightly larger deviation of the average value. Therefore, from the perspective of predicted value distribution, CGCLM method also has better prediction performance.
4. Discussion

The purpose of this study is to propose a time series prediction model to achieve health risk prediction through meteorological factors in advance. We constructed a multistep cascade model CGCLM using a dataset totaling 4,6089 patients with LRTIs and explored the predictive effects of different lengths of time windows. The results showed that the CGCLM approach integrating GAM, CEEMDAN and LSTM worked better than the single LSTM model. CGCLM shows the best performance when the time window is set to 61 days.

Discovering potential relationships and patterns from healthcare-related data to help hospitals and public health departments make decisions has garnered a lot of attention. For example, in 2009, Google predicted the outbreak of Influenza A (H1N1) with the help of big data technology from users' search behaviors, and issued an early warning ahead of the US Centers for Disease Control and Prevention (Davidson, et al., 2015). A study constructing a predictive model of influenza-like case rates based on users' posts on Twitter has yielded relatively accurate results (Lampos, et al., 2010). In recent years, a large number of studies have focused on the impact of environmental variables on diseases. Some studies have paid attention to the use of predictive models, such as machine learning and neural network for admission prediction (Kassomenos, et al., 2008). Other studies have concentrated on exploring the relationship between air quality and meteorological data and the occurrence of pneumonia through the use of statistical methods such as Generalized Linear Model (Glick, et al., 2018), Generalized Additivity Model (Li, et al., 2018), or autoregressive integrated moving averages (Ruchiraset and Tantrakarnapa, 2018). This difference in approach explains the diversity between predictive and explanatory models. The former usually provides high precision but low interpretability, while the latter approach often provides high interpretability but
In this study, we presented a multi-step health risk prediction method (CGCLM) based on meteorological factors and time series cascade models, which fully considered the explanatory and predictive properties of the model and confirmed its excellent performance. To reduce the effect of noise on prediction, CEEMDAN was used in conjunction with LSTM to predict the time series of LRTIs admissions. Since CEEMDAN is a Fourier transform-based signal decomposition method, it processes any non-linear and non-stationary signal adaptively. The original time series is decomposed into multiple subseries at different frequencies by CEEMDAN processing, and these subseries are predicted separately as input data for the LSTM model. Finally, all predictions are reconstructed to obtain the final results. After analyzing the results, we found that the CGCLM method achieved good prediction performance, but there was some variation in the performance of the LSTM model leading to fluctuations in the prediction accuracy from year to year. This may be due to the fact that in certain populations, socioeconomic factors may be the primary driver of disease admissions instead of meteorological one (Sahni, et al., 2017; Trachtenberg, et al., 2014).

The lag effect of meteorological conditions has been extensively demonstrated in epidemiological studies and shows that there are regional differences in this effect. Based on previous research, Krishan et al. adopted the representative lags regard as predictors to predict the peak demand day of respiratory disease visits (Khatri, et al., 2017). Ramadona, A.L., et al set the meteorological variables with a time lag of 4 months for the exposure-response relationship study and predicted the number of dengue cases and potential outbreak risk based on the time lag results (Ramadona, et al., 2016). In our study, the GAM model was used to determine the risk relationships between several meteorological conditions and LRTIs admissions. Compared with the
traditional regression model, GAM provides a flexible and effective technique for modeling nonlinear time-series in studies of the health effects of environmental factors.

In this work, we observed that the effect of relative humidity on the admission of LRTIs was at a high level, suggesting that it may be a key factor influencing LRTIs. This finding is consistent with previous studies, a meta-analysis by Gao et al. demonstrated that environmental relative humidity generally plays an important role in the incidence and prevalence of climate-sensitive diseases, such as respiratory diseases (Jinghong, et al., 2014). Existing studies explain the mechanism of the association between relative humidity and LRTIs, one of the study reports that relative relative humidity is positively correlated with RSV activity, with more RSV infection as relative humidity increases, ultimately accompanying the risk of LRTIs (Meerhoff, et al., 2009).

Enveloped viruses such as influenza viruses require 20-30% relative humidity to grow and they can only survive in high relative humidity (50-70%) environments (Lowen, et al., 2007). What is more, temperature and wind velocity also showed some correlation. Temperature changes may impair sensor cell function and antiviral responses, thereby affecting the body's immune function (Iwasaki, et al., 2017). Moreover, temperature also affects people's direct exposure to the spread of the virus, e.g. people's physical activity is increasing during the warmer months (Suminski, et al., 2008). Wind velocity has been found to increase the activity of respiratory viruses (Darniot, et al., 2018), possibly explained by the fact that it promotes the circulation and distribution of air pollutants, such as virus-carrying particles, to speed up their transport. Besides, we also found that atmospheric pressure was negatively associated with the risk of hospital admission for LRTIs. Shaman and Kohn stated that atmospheric pressure affects the spread and survival of viruses (Shaman and Kohn, 2009). However, Liu et al. found a positive correlation between atmospheric pressure and the number of
hospitalizations for LRTIs (Liu, et al., 2016).

The WHO reports that effective health service delivery requires a number of key resources, of which comprehensive health information is a critical component (WHO, 2020). Given the increasing public demand for coverage and quality of health care services, health care providers are often under increased pressure to respond to the demands associated with high disease prevalence (Sharifi and Saberi, 2014). As frontline health care services and providers do not have sufficient information or resources to meet the "above normal" demand for health care services. Capacity planning decisions are particularly important for the medical profession, as they not only involve the management of highly specialized and expensive resources (i.e., nurses, doctors and advanced medical equipment) but also can be life-or-death in critical situations (Hans EW, 2012). Health forecasting allows both individuals and service providers to anticipate situations so that necessary steps can be taken to manage peaks or extreme events (Soyiiri and Reidpath, 2013). Therefore, our research model promises to provide health care organizations with accurate forecasts of health care demand, thereby optimizing resource allocation and improving the service quality.

Although the experimental results are satisfactory, we also observe that there are still large errors in the prediction of the model at many local extremum points. Second, the study was limited by the amount of data available. Also, the limitation of variable availability makes pollutants such as PM$_{2.5}$, PM$_{10}$ and NO$_2$ excluded from this study, which has been proved to have a direct impact on the respiratory system.

**Conclusions**

In this study, a multi-step health risk time series prediction method (CGCLM) is established by combining with GAM model, CEEMDAN signal decomposition algorithm and LSTM. This
method is applied to forecast the daily hospital admission of LRTIs. The main steps are the following: according to the actual meteorological conditions of the study site, the exposure-response relationship is established by using GAM model; then, the CEEMDAN method is used to decompose the unsteady admission time series into several smooth IMF components; finally, the LSTM neural network is used to predict the IMF component. By error analysis (RMSE, MAE and $R^2$), CGCLM method shows a lower prediction error than single LSTM model. The realization of accurate admission prediction not only provides support for the establishment of a prediction model for the number of medical patients and the decision-making of public health departments but also provides a reference for people to choose the time of medical treatment, which is of great research significance.

**Ethics approval**

This study was originally approved by the Clinical Research Ethics Committee of the Panyu center hospital with code [2020]25.

**Consent to participate**

This is a retrospective study therefore no patients actually participated in this study. The information collected in this study includes the date of admission, gender, age and ICD number, and does not include sensitive information such as the patient's name. The patients' personal information was adequately protected.

**Consent to publication**

All authors consent to publish this article in Environmental Science and Pollution Research.

**Funding**
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China [grant number 81773806, 81874331]; Double-Class University project [grant numbers CPU2018GY19].

**Acknowledgment**

The author(s) would like to thank the support provided by High Performance Computing Center, China Pharmaceutical University.

**Competing interests**

The authors declare that they have no competing interests.

**Authors' contributions**

- **Shuopeng Jia**: Methodology, Software, Writing - Original Draft
- **Weibin She**: Conceptualization, Resources, Data Curation
- **Zhipeng Pi**: Software, Writing - Review & Editing
- **Buying Niu**: Software
- **Jinhua Zhang**: Resources, Investigation
- **Xihan Lin**: Writing - Review & Editing
- **Mingjun Xu**: Data Curation
- **Weiya She**: Resources
- **Jun Liao**: Conceptualization, Supervision, Funding

**Availability of data and materials**

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

**References**

Armstrong, J.S. and Collopy, F.J.I.J.o.F. Error measures for generalizing about forecasting methods:
Empirical comparisons. 1992;8(1):69-80.

Darniot, M., Pitoiset, C., Millière, L., Aho-Glélé, L.S., Florentin, E., Bour, J.B., et al. Different meteorological parameters influence metapneumovirus and respiratory syncytial virus activity. Journal of Clinical Virology 2018:S1386653218301252.

Davidson, M.W., Haim, D.A. and Radin, J.M.J.S.R. Using Networks to Combine "Big Data" and Traditional Surveillance to Improve Influenza Predictions. 2015;5:8154.

Gers, F.A., Schmidhuber, J. and Cummins, F. Learning to Forget: Continual Prediction with LSTM. Neural Computation 2000;12(10):2451-2471.

Glick, A.F., Tomopoulos, S., Fierman, A.H., Elixhauser, A. and Trasande, L. Association Between Outdoor Air Pollution Levels and Inpatient Outcomes in Pediatric Pneumonia Hospitalizations, 2007 to 2008. Academic Pediatrics 2018.

Gu, J., Liang, L., Song, H., Kong, Y., Ma, R., Hou, Y., et al. A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China. Sci Rep 2019;9(1):17928.

Hans EW, V.H.M.a.H.P. A framework for healthcare planning and control. Handbook of Healthcare System Scheduling 2012.

Hastie, T. and Tibshirani, R. Generalized Additive Models: Some Applications. Publications of the American Statistical Association 1987;82(398):371-386.

Huang, N., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. A 1998;454:679-699.

Hytndman, R.J. and Koehler, A.B. Another look at measures of forecast accuracy. International Journal of Forecasting 2006;22(4):679-688.

Iwasaki, A., Foxman, E.F. and Molony, R.D. Early local immune defences in the respiratory tract. Nat Rev Immunol 2017;17(1):7-20.

Jacinta, Chan, Phooi, M'ng, Mohammadali and One, M.J.P. Forecasting East Asian Indices Futures via a Novel Hybrid of Wavelet-PCA Denoising and Artificial Neural Network Models. 2016.

Jinghong, G., Yunzong, S., Yaogui, L., Liping, L. and Tang, J.W.J.P.O. Impact of Ambient Humidity on Child Health: A Systematic Review. 2014;9(12):e112508.

Kassomenos, P., Papaloukas, C., Petrakis, M. and Karakitsios, S. Assessment and prediction of short term hospital admissions: the case of Athens, Greece. Atmospheric Environment 2008;42(30):7078-7086.

Khatri, K.L., Tamil, L.I.J.o.B. and Informatics, H. Early Detection of Peak Demand Days of Chronic Respiratory Diseases Emergency Department Visits Using Artificial Neural Networks. 2017:1-1.

Lampos, V., De Bie, T. and Cristianini, N. Flu Detector - Tracking Epidemics on Twitter. In: Balcázar, J.L., et al., editors, Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 599-602.

Li, D., Wang, J.B., Zhang, Z.Y., Shen, P. and Chen, K. Effects of air pollution on hospital visits for pneumonia in children: a two-year analysis from China. Environmental Science & Pollution Research International 2018;25(11):1-9.

Linare, C., Martínez, G.S., Kendrovski, V. and Diaz, J. A new integrative perspective on early warning systems for health in the context of climate change. Environmental Research 2020;187:109623.

Liu, Y., Liu, J., Chen, F., Shamsi, B.H., Wang, Q., Jiao, F., et al. Impact of meteorological factors on lower respiratory tract infections in children. J Int Med Res 2016;44(1):30-41.

Lowen, A.C., Mubareka, S., Steel, J. and Palese, P. Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathog 2007;3(10):1470-1476.
Cold temperature and low humidity are associated with increased occurrence of respiratory tract infections. *Respir Med* 2009;103(3):456-462.

Meerhoff, T.J., Paget, J.W., Kimpen, J.L. and Schellevis, F.O.J.P.I.D.J. Variation of Respiratory Syncytial Virus and the Relation With Meteorological Factors in Different Winter Seasons. 2009;28(10):860.

Navares, R., Computing, J.L.A.J.N. and Applications. Deep learning architecture to predict daily hospital admissions. 2020;1-10.

Price, R.H.M., Graham, C. and Ramalingam, S. Association between viral seasonality and meteorological factors. *Sci Rep* 2019;9(1):929.

Qiu, H., Luo, L., Su, Z., Zhou, L., Wang, L. and Chen, Y. Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. *BMC Med Inform Decis Mak* 2020;20(1):83.

Ramadona, A.L., Lazuardi, L., Hii, Y.L., Holmner, Å., Kusnanto, H. and Rocklöv, J. Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data. *PLoS One* 2016;11(3):e0152688-e0152688.

Richman, J.S. and Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *American journal of physiology. Heart and circulatory physiology* 2000;278(6):H2039-2049.

Ruchiraset, A. and Tantrakarnapa, K. Time series modeling of pneumonia admissions and its association with air pollution and climate variables in Chiang Mai Province, Thailand. *Springer Open Choice* 2018;25(33).

Sahni, S., Talwar, A., Khanjo, S. and Talwar, A. Socioeconomic status and its relationship to chronic respiratory disease. *Adv Respir Med* 2017;85(2):97-108.

Sainani and Kristin, L. Explanatory Versus Predictive Modeling. *Pm & R* 2014;6(9):841-844.

Shaman, J. and Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc Natl Acad Sci U S A* 2009;106(9):3243-3248.

Sharifi, S. and Saberi, K. CAPACITY PLANNING IN HOSPITAL MANAGEMENT: AN OVERVIEW. 2014;4.

Soyiri, I.N. and Reidpath, D.D. Evolving forecasting classifications and applications in health forecasting. *Int J Gen Med* 2012;5:381-389.

Soyiri, I.N. and Reidpath, D.D. An overview of health forecasting. *Environmental Health & Preventive Medicine* 2013;18(1):1-9.

Suminski, R.R., Poston, W.C., Market, P., Hyder, M. and Sara, P.A. Meteorological conditions are associated with physical activities performed in open-air settings. *International Journal of Biometeorology* 2008;52(3):189-197.

Torres, M.E., Colominas, M.A., Schlotthauer, G. and Flandrin, P. A complete ensemble empirical mode decomposition with adaptive noise. In, 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2011. p. 4144-4147.

Trachtenberg, A.J., Dik, N., Chateau, D. and Katz, A. Inequities in ambulatory care and the relationship between socioeconomic status and respiratory hospitalizations: a population-based study of a canadian city. *Ann Fam Med* 2014;12(5):402-407.

Watts, N., Amann, M., Arnell, N., Ayeb-Karlsson, S., Beagley, J., Belesova, K., et al. The 2020 report of The Lancet Countdown on health and climate change: responding to converging crises. *The Lancet* 2020.

WHO. 2019 Global Health Estimates :The top 10 causes of death. 2020.
Woodhead, M., Blasi, F., Ewig, S., Garau, J., Huchon, G., Ieven, M., et al. Guidelines for the management of adult lower respiratory tract infections—summary. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases 2011;17 Suppl 6:1-24.

Wu, Z. and Huang, N. Ensemble Empirical Mode Decomposition: a Noise-Assisted Data Analysis Method. Advances in Adaptive Data Analysis 2009;1:1-41.

Xie, G., Liang, T., Mi, D. and Wang, J. PM10 Concentration Forecasting Based on CEEMDAN, SE and LSTM Neural Network. In, 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2). 2019. p. 2398-2402.

Zhang, G.P. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 2003;50:159-175.

Zhang, X., Pang, Y., Cui, M., Stallones, L. and Xiang, H. Forecasting mortality of road traffic injuries in China using seasonal autoregressive integrated moving average model. Ann Epidemiol 2015;25(2):101-106.
The overview of the research framework, which can be divided into three parts. The first part includes data used in the experiment, data preprocessing and feature extraction. The second part is the decomposed time series features and the use of sample entropy algorithm to do feature clustering. The
The third part is the independent prediction of each feature using LSTM as a prediction model, and the final output of the combined prediction results.

| 2003.01.01 | 365x3 days | N days | N days |
|------------|------------|--------|--------|
| Training   | Validation | Testing|

Figure 2

The composition of sliding time window and the distribution of training set, validation set and testing set. The length of time for N were 14 days, 31 days, 61 days and 91 days.
**Figure 3**

Forest plots of the effects of temperature, relative humidity, atmospheric pressure, and wind velocity on admission RR and 95% CI of LRTIs. Lag0, lag1, lag2, lag3, lag4, lag5 represent the zeroth day, the first day, the second day, the third day, the fourth day, the fifth day of lag effect.
Figure 4

The CEEMDAN decomposition consequence of LRTIs admission and the correlation coefficients of each sub-sequence.

Figure 5

The sample entropy result of each sub-sequence of LRTIs admission and reconstruction consequence.
Figure 6

Waterfall plot of true values (green curve), CGCLM model predicted values (orange curve) and single LSTM model predicted values (purple curve) for different windows.

Supplementary Files

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

- SupplementaryMaterial.docx