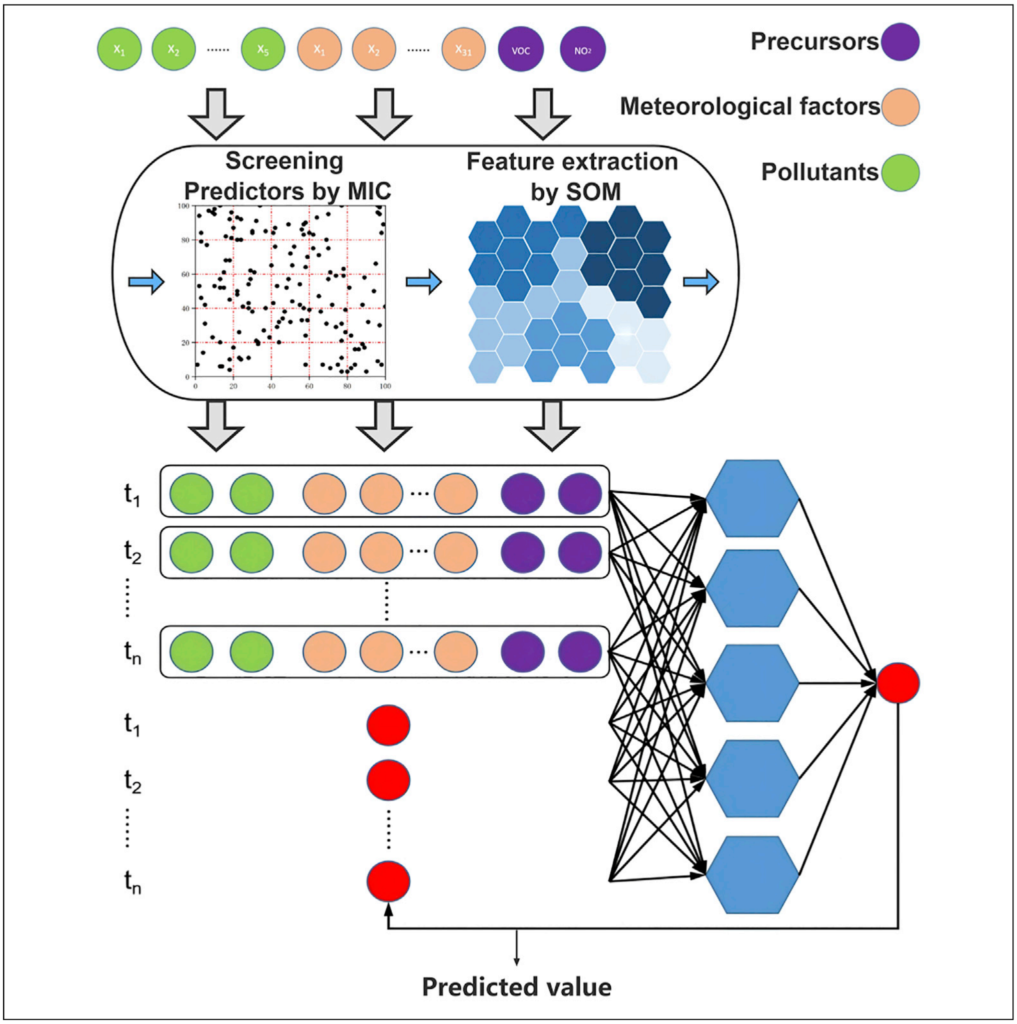
iScience

CellPress
OPEN ACCESS

Article

Prediction of ground-level ozone by SOM-NARX hybrid neural network based on the correlation of predictors



Qinqing Xiong,
Wenju Wang,
Mingya Wang,
Chunhui Zhang,
Xuechun Zhang,
Chun Chen,
Mingshi Wang

mingshiwang@hpu.edu.cn

Highlights

For the first time, the SOM-NARX hybrid neural network is proposed to predict ozone

Feature extraction using the SOM network for the first time

Predictors are screened by MIC to improve prediction accuracy

The prediction accuracy of SOM-NARX is better than other dynamic hybrid models

Xiong et al., iScience *25*, 105658
December 22, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.105658

iScience

CellPress
OPEN ACCESS

Article

Prediction of ground-level ozone by SOM-NARX hybrid neural network based on the correlation of predictors

Qinqing Xiong,[1,3] Wenju Wang,[1,3] Mingya Wang,[1] Chunhui Zhang,[1] Xuechun Zhang,[1] Chun Chen,[2] and Mingshi Wang[1,4,*]

SUMMARY

Current approaches to ozone prediction using hybrid neural networks are numerous but not perfect. Decomposition algorithms ignore the correlation between predictors and ozone, and feature extraction methods rarely select appropriate predictors in terms of correlation, especially for VOCs. Therefore, this study proposes a hybrid neural network model SOM-NARX based on the correlation of predictors. The model is based on MIC to filter predictors, using SOM to make predictors as feature sequences and using NARX networks to make predictions. Data from the JCDZURI site were used for training, testing, and validation. The results show that the correlation of the predictors, classification numbers of SOM, neuron numbers, and delay steps can affect prediction accuracy. Model comparison shows that the SOM-NARX model has 13.82, 10.60, 6.58% and 12.05, 9.44, 68.14% RMSE, MAE, and MAEP in winter and summer, which is smaller than CNN-LSTM, CNN-BiLSTM, CNN-GRU, SOM-LSTM, SOM-BiLSTM, and SOM-GRU.

INTRODUCTION

Particulate pollution in China has significantly decreased since the Air Pollution Prevention and Control Action Plan was put into effect there in 2013. But the presence of ground-level ozone (O_3) pollution grows. From 2015 to 2019, the annual average $PM_{2.5}$ and PM_{10} concentrations in 337 Chinese cities decreased by 24 $\mu g/m^3$, whereas the annual average O_3-M8h value increased from 134 $\mu g/m^3$ to 148 $\mu g/m^3$,[1] exceeding the WHO air quality standard (100 $\mu g/m^3$) by 0.48 times.[2] Studies have shown that for every 10 $\mu g/m^3$ increase in O_3-M8h, the total risk of death in the population increases by 0.31%.[3,4]Ozone pollution has become the second factor affecting China air quality and threatens the health of Chinese residents.

Ozone pollution in China has obvious seasonal and regional differences. High ozone pollution occurs mainly from May to September, while ozone concentrations are lowest from December to January, mainly due to the high temperature, high humidity, and light in summer, which promote the rate of photochemical reactions and accelerate ozone accumulation.[5] This seasonal difference also varies by region, for example, ozone pollution in the Pearl River Delta urban agglomeration occurs in May and October respectively, but the seasonal difference is not significant, whereas in the central-eastern region, where the seasonal difference is significant, the high pollution period is concentrated in June to September, and the pollution level is also higher than in other regions.[6–8] The most serious ozone pollution areas are the "2 + 26 city cluster" around Beijing, Tianjin and Hebei, with annual average O3-M8h values between 158 $\mu g/m^3$ and 199 $\mu g/m^3$ from 2015 to 2021. Except for 2015, all other years have reached the "light pollution" level of the China Ambient Air Quality Standard.[9] In the severe ozone pollution situation, it is important to develop effective control measures. An accurate and reliable ozone warning system is an important part of pollution control, not only to provide a scientific reference for control strategies but also to remind residents to avoid long-term exposure to pollution.

Currently, ozone prediction methods can be generally classified into three categories: air quality models, statistical methods, and neural network models. Representative air quality models include a third-generation Air Quality Modeling System (Model-3/CMAQ), Weather Research and Forecasting model coupled to Chemistry (WRF-Chem), and Nested Air Quality Prediction Modeling System (NAQPMS), whose main

[1]College of Resource and Environment, Henan Polytechnic University, Jiaozuo 454003, China

[2]Henan Key Laboratory for Environmental Monitoring Technology, Zhengzhou 450004, China

[3]These authors contributed equally

[4]Lead contact

*Correspondence: mingshiwang@hpu.edu.cn

https://doi.org/10.1016/j.isci.2022.105658

principles are based on the complex physicochemical reactions of precursors under different meteorological conditions, However, uncertainties in physicochemical parameters and emission inventories may lead to some bias in the predicted values.[10–12] The statistical method, such as Autoregressive Integrated Moving Average model (ARIMA),[13] hidden Markov model (HMM) and generalized linear models (GLMs), Multiple Linear Regression (MLR),[14] Partial Least Squares (PLS), as well as Principal Component Regression (PCR),[15] is favored by many scholars because of its low operational requirements and easy implementation. However, the statistical methods assume a linear relationship between the dependent and independent variables,[16] inconsistent with the complex nonlinear relationships between the ozone and precursors (VOC and NOx), as well as meteorological factors.[17]

Artificial neural networks have received considerable attention because of their strong nonlinear mapping ability, especially in solving the nonlinear problems of pollutants with irreplaceable advantages.[18–20] The mechanism of ozone generation is very complex, and a single model cannot efficiently fit the dynamic relationship between the ozone and influencing factors, so many studies combine the advantages of multiple algorithms to construct hybrid prediction models. At present, the strategies for constructing hybrid models are mainly divided into feature extraction and decomposition algorithms. For example, wavelet decomposition (WD) combined with gated recurrent unit (GRU) neural network and support vector regression (SVR) model to predict O_3-M8h[21]; convolutional neural network CNN combined with long short-term memory neural network (LSTM) to predict PM2.5.[22] However, studies using these two methods to predict ozone have not been perfect. Decomposition algorithm improves the mapping ability by decomposing the time series into more stable and regular subcolumns, but ignores the correlation between the predictors and ozone, destroying the potential relationship between them.[23] Relatively, feature extraction methods are superior to decomposition algorithms. It can extract the temporal and spatial characteristics of predictors and maintain their temporal and spatial correlation with ozone. But the existing studies mostly use specified environmental pollutant factors (CO, $PM_{2.5}$, and NOx, and so forth) and meteorological factors (temperature, light, and wind speed, and so forth) as predictors. Appropriate predictors are rarely selected from the perspective of correlation, and VOCs, which are important precursors of ozone, are rarely used to the extent that the mapping relationships established are incomplete.[24]

In order to make up for the deficiency of neural networks in predicting ozone, this study proposes a hybrid neural network model (SOM-NARX) that combines a self-organizing mapping neural network (SOM) with nonlinear autoregressive models with exogenous inputs (NARX). Although the model uses a feature extraction strategy, it differs from the commonly used feature extraction methods in that for the first time the clustering capability of the SOM network is used to transform similar predictors into feature factors for dimensionality reduction. In addition, the effect of correlation on model accuracy was investigated based on the mutual information coefficient (MIC) of predictors and ozone. Finally, the prediction performance of CNN-LSTM, CNN-BiLSTM, CNN-GRU, SOM-LSTM, SOM-BiLSTM, SOM-GRU and SOM-NARX is investigated by comparing.

RESULTS

In this section, the effect of the correlation between predictors and ozone on the SOM-NARX model is first analyzed, and then the relationship between different parameters and prediction accuracy is discussed. Finally, the prediction performance of the SOM-NARX network is analyzed by model comparison.

Correlation of predictors

MIC is used to quantify the correlation between the factors and ozone, with a larger MIC indicating a stronger correlation between the factors and ozone. By calculating the MIC of the factors, the effect of correlation on the accuracy of the model is investigated and the optimal predictor is screened. Predictors with MIC >0, MIC >0.05, MIC >0.10, MIC >0.15, and MIC >0.20 were experimented with separately to analyze the prediction accuracy of SOM-NARX for different MICs. The remaining parameters of the model were used as initial parameters.

The MIC calculation results of the predictors are shown in Table S1. NO_2, temperature, light, air pressure, wind, and ozone are strongly correlated; they are the main factors affecting the generation and transmission of ozone. VOCs are weakly correlated with ozone. This phenomenon may be related to the high randomness of VOC time series caused by the combination of irregular emission levels and complex atmospheric diffusion conditions. In addition, previous studies have shown that relative humidity can reduce the ground-light

Table 1. Error results under different MIC conditions

MIC	RMSE (μg/m^3)	MAE(μg/m^3)	MAPE (%)
MIC>0	17.59	14.22	94.55%
MIC>0.5	16.75	13.80	91.65%
MIC>0.1	16.34	13.45	90.78%
MIC>0.15	15.60	12.19	76.02%
MIC>0.2	15.93	13.34	92.45%

intensity and inhibit the rate of photochemical reactions.[25] However, in this study, the MIC of relative humidity was only 0.9, showing a weak correlation with ozone. The main reason is that Jiaozuo City is located between North China and the Huanghuai River Basin. It is dry in summer and less rainy in winter. The relative humidity is very low throughout the year, and the inhibition of the photochemical rate is very weak.

The experimental results are shown in Table 1. The table shows that with the increase in MIC, the predictors of low correlation gradually decrease, and RMSE, MAE, and MAPE decrease accordingly. When MIC>0.15, the three errors reach the minimum value. This result suggests that irrelevant variables can directly affect the prediction accuracy of the SOM-NARX model. However, the error tends to increase as the MIC continues to increase. This phenomenon may be due to the extremely few predictors, so the mapping relationship constructed by the model tends to be more single, and the fitting ability decreases accordingly, thereby reducing the prediction accuracy. This result shows that the prediction accuracy of SOM-NARX is mainly affected by the correlation and number of predictors. Therefore, variables with MIC>0.15 were selected as the best predictors in this study, including NO_2, air temperature, surface temperature, potential evaporation, evaporation, sea level pressure, ground pressure, UV intensity, total sunlight intensity, north wind speed, east wind speed, and CO. In addition, although the correlation between VOCs and ozone is not strong, VOCs are important precursors of ozone and directly determine ozone generation. Therefore, using VOCs as the best predictor.

Optimal parameters and data processing

In order to investigate the effect of parameters on prediction accuracy, the optimal parameters of the SOM-NARX model were screened. Test experiments were conducted for different parameters separately, and the experiments were repeated 10 times for each group of parameters, and the mean value was used as the final result. RMSE, MAE, and MAPE were used to evaluate the experimental results. The key parameters of the SOM network are the number of categories, and the key parameters of the NARX network are the delay step and the number of neurons.

Parameter selection

The experimental results of the parameters are shown in Figure 1 The figure shows that when the number of categories is less than 45, RMSE, MAE, and MAPE decrease with the increase in the number of categories; the three errors increase and fluctuate greatly with more than 55 categories. The finding shows that within 45 categories, the number of categories is inversely proportional to the prediction accuracy of SOM-NARX; with more than 55 categories, the error has no correlation with the number of categories; the number of categories is between 45 and 55, and the prediction accuracy is the highest. In addition, the RMSE, MAE, and MAPE are not remarkably different in categories 45 and categories 55, However, the RMSE and MAE are the smallest in categories 45, whereas only MAPE is the smallest in categories 55, so the optimum parameter for the number of categories is 45.

For the number of neurons, if the number of neurons is less than 9, then RMSE, MAE, and MAPE gradually decrease with this number; if it is more than 9, then it tends to be stable, indicating that the fitting ability of the NARX network increases with the increase in the number of neurons. However, the excessive number of neurons will increase the complexity of the network and reduce the operating efficiency of the model. Therefore, the optimal parameter for the number of neurons is 9.

The delay step is a key parameter that determines the structure of NARX and is highly correlated with the temporal correlation between the ozone and predictors. So, it is necessary to investigate not only the effect of the delay step on the prediction accuracy but also the relationship between time dependence and

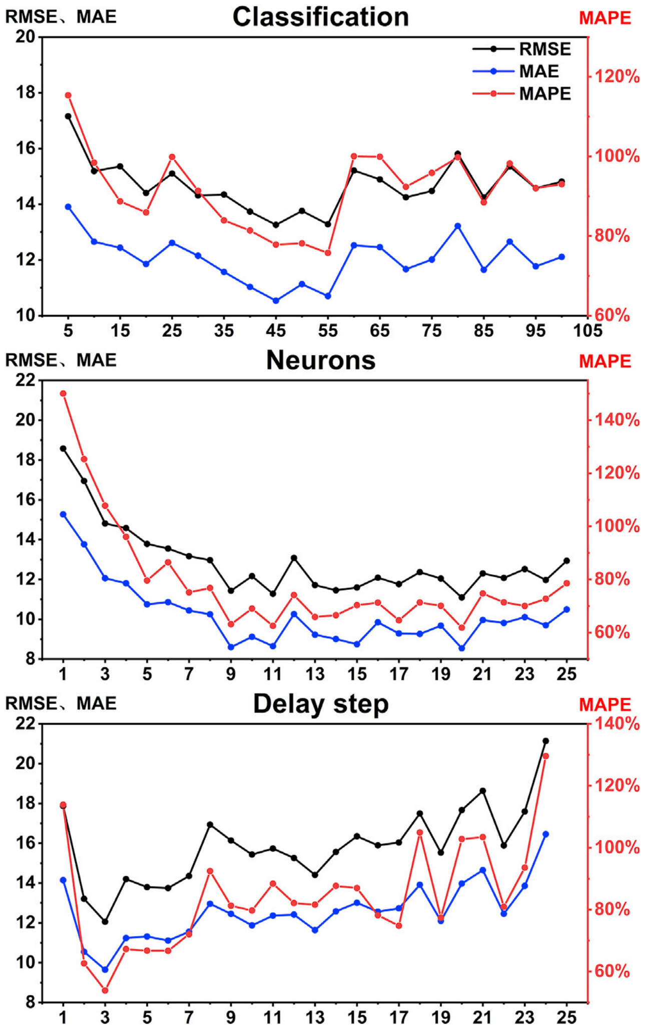


Figure 1. Error in the number of SOM classifications, NARX neurons, and delay step

model accuracy. Since this article is concerned with the strength of correlations between ozone and predictors, rather than positive and negative correlations. Therefore, the absolute value of the correlation coefficient is used to indicate the strength of the correlation coefficient, as shown in Table 2, where the color from green to red indicates the correlation coefficient from small to large. Combine Figure 1 and Table 2, the RMSE in the delay steps 1–12 shows an upward trend, and the correlation coefficients of all predictors except VOCs show a downward trend. The RMSE and correlation trends were opposite, indicating that the temporal correlation between predictors and ozone was positively correlated with the accuracy of the model. When the delay step is 1, the RMSE, MAE, and MAPE are very large, and then drop significantly, and when the delay step is 3, they reach the minimum, and then these errors begin to increase. The main reason is that the correlation coefficients are close to the peak value in delay steps 1–3, and the peak value is also in this interval. However, extremely few delay steps cause the fitting ability of the NARX network to become insufficient and increase the error. As the delay step increases, the fitting ability gradually increases, and the factor restricting the prediction accuracy gradually changes from fitting ability to correlation. The three smallest errors (with a delay step size of 3) are the result of the combined effect of correlation and fitting ability. Thus, the optimal parameter for the delay step is 3.

Standardization and normalization

Before making predictions, the data should be processed to ensure that the different factors are in the same dimension. In this study, experiments were conducted using normalization and standardization,

Table 2. Cross-correlation of predictors with ozone

Daley	NO2	VOCs	Pressure	Temperature	Wind	Light	Evaporation	Potential evaporation	CO	RMSE
−12	0.11	0.07	0.49	0.36	0.03	0.18	0.07	0.07	0.02	12.42
−11	0.14	0.07	0.49	0.39	0.04	0.09	0.02	0.01	0.02	12.37
−10	0.18	0.07	0.49	0.42	0.05	0.01	0.05	0.09	0.01	11.87
−9	0.23	0.06	0.49	0.47	0.07	0.12	0.12	0.19	0	12.45
−8	0.28	0.05	0.49	0.52	0.1	0.24	0.21	0.29	0	12.96
−7	0.33	0.05	0.49	0.57	0.13	0.36	0.29	0.4	0	11.55
−6	0.39	0.05	0.49	0.63	0.16	0.47	0.38	0.5	0	11.11
−5	0.45	0.05	0.5	0.68	0.19	0.56	0.45	0.6	0.01	11.32
−4	0.5	0.04	0.5	0.72	0.22	0.63	0.51	0.67	0.01	11.24
−3	0.55	0.04	0.51	0.76	0.24	0.66	0.55	0.73	0.02	9.65
−2	0.6	0.03	0.53	0.75	0.26	0.64	0.55	0.72	0.03	10.55
−1	0.63	0.03	0.54	0.75	0.27	0.58	0.52	0.69	0.04	14.15
0	0.64	0.03	0.54	0.72	0.28	0.48	0.45	0.61	0.04	

respectively, so that the best data processing method could be screened. The experimental results are shown in Table 3. By normalizing and standardizing the eigenfactors, the MAPE can be reduced by 15.04 and 24.44%. The finding shows that standardization or normalization can significantly improve the prediction accuracy of the SOM-NARX model, and the prediction error is the smallest after standardization. To further investigate the reasons for this discrepancy, the processed data were analyzed. The result shows that the maximum value of VOCs data is 2842.62 $\mu g/m^3$, the minimum value is 1.69 $\mu g/m^3$, whereas the average value is only 114.76 $\mu g/m^3$, and the frequency of high concentrations of VOCs is very low. By comparing the two data processing methods (see as Figure 2), we found that 80% of the VOCs data are distributed between −0.323 and 0.274 after standardization, and the distribution range of all data is not remarkably different. However, 80% of the normalized data are distributed in the range of 0.0026–0.016, which is not of the same order of magnitude as other predictors. The difference in data magnitude increases the training difficulty of the NARX network, thereby increasing the systematic error. Thus, standardization is better than normalization.

DISCUSSION

Self-organizing mapping neural network-nonlinear autoregressive models with exogenous inputs performance evaluation

This section investigates the seasonal differences and multiscale prediction accuracy of the SOM-NARX and NARX networks. In addition, the stability and applicability of the model are analyzed by comparing SOM-NARX with several commonly used hybrid neural networks. The experimental data were used from September 3 to 15, 2020 and compared with the data from January 3 to 15, 2022. The No. 1 and No. 2 data are used as pre-experiments, not as prediction results.

Improvement of nonlinear autoregressive models with exogenous inputs prediction performance by self-organizing mapping neural network

In the ozone predictions of NARX and SOM-NARX in different seasons (as shown in Table 4 and Figure S3), the predicted values are in good agreement with the observed values. However, the fit between the predicted value and the observed value in January 2022 is very high, the fit in September 2020 is poor, and the RMSE, MAE, and MAPE in January 2022 are generally smaller than those in September 2020. This finding shows the evident seasonal differences in the prediction accuracy of the two models, and the prediction performance in winter is better than that in summer, similar to Wang Hongwei's research.[26] The authors suggest that this seasonal difference may be related to seasonal differences in ozone.

From the comparison of the two models, the RMSE, MAE, and MAEP of SOM-NARX in different seasons are smaller than NARX, and the mean error of the two seasons is also the smallest for SOM-NARX. The result shows that the SOM network can improve the prediction performance of the NARX network. On the other

Table 3. Standardization and normalized prediction errors

Number of categories	RMSE (μg/m^3)	MAE(μg/m^3)	MAPE (%)
standardization	14.21	11.38	70.84%
normalization	14.58	12.09	80.24%
Raw data	16.70	13.67	95.28%

hand, the RMSE, MAE, and MAEP of the NARX model in winter and summer differed by 3.26%, 2.26%, and 67.87%, respectively, whereas the difference in SOM-NARX was 1.77%, 1.15%, and 61.56%, and the seasonal difference of SOM-NARX was significantly smaller than that of NARX. This finding indicates that after the dimensionality reduction of the SOM network, the seasonal difference of the NARX network can be effectively improved, and the applicability of the NARX network in different seasons can be improved.

In the comparison of multistep prediction, absolute error (AE; the absolute value of the difference between the observed value and the predicted value, i.e., |Oi-Pi|) is the main reference for measuring deviations in the size of different forecasting steps. The results are shown in Figure 3, where the AE of the two models showed a trend of initially increasing and then become stable with the increase in the step size. It continued to increase in the 0–3 steps, and the AE tended to be stable after three steps. It shows that the short-term prediction accuracy of the NARX and SOM-NARX models is better than the long-term prediction. The difference is that the error distribution of SOM-NARX in each step is more concentrated than NARX, and the MAE is also smaller than NARX. It shows that in multistep prediction, the prediction accuracy of SOM-NARX is higher than that of the NARX network.

Model comparison

To further evaluate the advantages of the SOM-NARX model in ozone prediction. Two commonly used feature extraction networks (i.e., SOM and CNN) were compared with four dynamic neural networks (NARX, LSTM, BiLSTM, GRU) combined into eight hybrid neural networks to discuss the performance differences between the SOM-NARX network and other models. 10 experiments were performed for each model, and the average value was used as the experimental result. The optimal parameters were determined by Trial-and-Error testing before the experiment. The prediction errors are shown in Table 4.

The two feature extraction networks are compared, and the results showed that the average error of the SOM network in the two seasons reaches 31.65%, 23.65%, and 56.47%, and the CNN is only 19.72%, 15.71%, and 67.63%. The summer error of SOM is much larger than that of CNN, whereas the error in winter and the R2 are smaller than the CNN. Except for SOM-NARX, the seasonal difference in SOM is significantly stronger than that of CNN. This finding shows that the stability and applicability of CNN in different seasons are better than the SOM network in terms of feature extraction. The reason may be that the feature sequences classified by SOM only maintain the correlation between similar predictors, and LSTM, BiLSTM, and GRU can only deal with temporal correlation. When the two types of models are combined, the interactions between different predictors cannot be effectively fitted, resulting in large errors. Especially in summer, ozone precursors are more closely correlated with temperature, light, wind speed, and other factors, widening the error and strengthening seasonal differences.

The prediction performance of dynamic neural networks indicates that when the four networks are combined with CNN and SOM, the prediction errors from small to large are NARX, LSTM, GRU, and BiLSTM. Although the RMSE and MAE of the CNN-NARX network in winter are larger than those of CNN-LSTM, and the correlation coefficient is smaller than that of CNN-LSTM. They are basically at the same level, without evident difference, and the error in summer is much smaller than the other dynamic neural network, indicating that the prediction accuracy of the NARX network is better than other dynamic networks. In addition, the prediction errors of NARX in different seasons are not remarkably different, whereas the seasonal differences of other dynamic networks are more evident, indicating that NARX is better than LSTM, BiLSTM, and GRU in fitting the interaction relationship between different predictors. The prediction error of the SOM-NARX model in different seasons is much smaller than that of other hybrid models, and the correlation coefficient R2 between the predicted value and the observed value is the largest, indicating that the NARX network can effectively make up for the shortcomings of the SOM network in feature extraction.

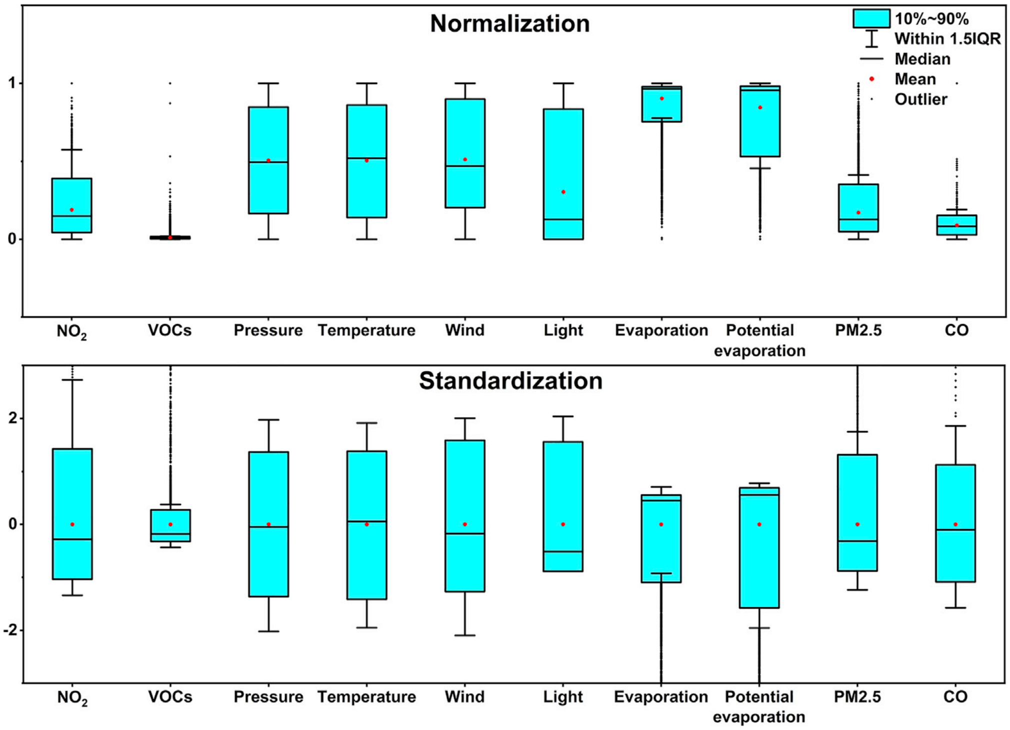


Figure 2. Boxplot of data distribution after standardization and normalization

Conclusions

This study proposes a new hybrid neural network model (SOM-NARX) for the prediction of ground-level ozone concentration. The model uses the maximum information coefficient to screen out environmental factors and meteorological factors that are highly related to ozone as predictors; it fully considers the correlation between predictors. Similar predictors are clustered as feature factors by the SOM network to avoid over-fitting problems caused by repeated variables. Finally, the NARX network is used for training and prediction. Experimental results show that the model can effectively predict near-ground ozone concentration, outperforming competitors. The following summarizes several key findings of this study:

The accuracy of SOM-NARX prediction is affected by the correlation and quantity of predictors. The optimal predictors selected were air temperature, surface temperature, potential evaporation, potential evaporation, sea level pressure, surface pressure, UV intensity, net sunshine intensity, total sunshine intensity, wind speed and direction, and CO.

Table 4. RMSE, MAE, MAPE of the prediction results of 8 mixed models

	September-20			January-22			
Model	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	MAPE (%)	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	MAPE (%)	R^2
cnn-lstm	23.18	18.02	11.25%	15.64	13.39	118.42%	0.8978
cnn-bilstm	23.01	17.92	11.43%	20.50	18.03	148.16%	0.8832
cnn-gru	22.36	17.70	11.13%	17.24	14.80	123.21%	0.8979
cnn-narx	18.39	14.92	9.54%	16.22	13.82	107.90%	0.8742
som-lstm	50.70	42.46	25.04%	15.99	13.10	94.88%	0.7865
som-bilstm	49.98	42.35	24.96%	16.44	13.64	105.01%	0.7925
som-gru	52.49	44.38	26.22%	16.15	13.22	100.86%	0.7730
som-narx	13.82	10.60	6.58%	12.05	9.44	68.14%	0.9161
narx	16.22	12.45	7.58%	12.96	10.19	75.45%	0.9056

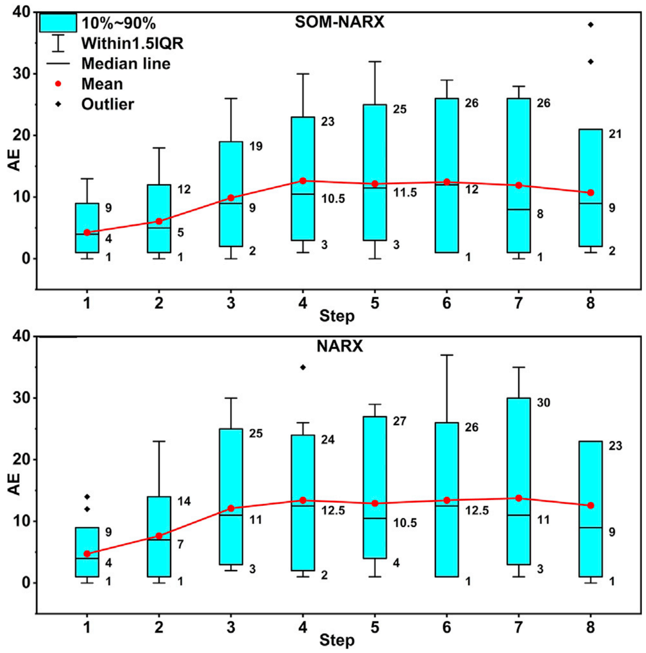


Figure 3. Accuracy boxplots of SOM-NARX and NARX networks at different prediction steps

Excessively high or low number of SOM classification numbers, neuron numbers, and delay steps will increase the prediction error of SOM-NARX. When the three parameters are 45, 9, and 3, the model accuracy is the highest. Standardization or normalization of the eigenfactors can effectively reduce the prediction error by 24.44 and 15.04% respectively, and the model error is the smallest after standardization.

The RMSE, MAE, and MAEP of the NARX model in winter and summer differed by 3.26%, 2.26%, and 67.87%, respectively, whereas the difference in SOM-NARX was 1.77%, 1.15%, and 61.56%. The SOM

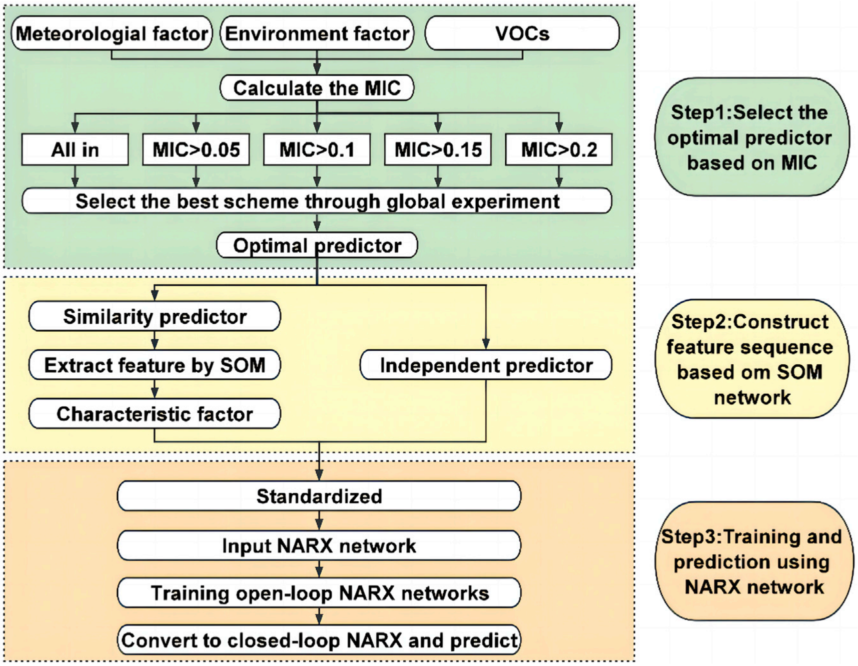

Meteorologial factor
Environment factor
VOCs
Calculate the MIC
All in
MIC>0.05
MIC>0.1
MIC>0.15
MIC>0.2
Select the best scheme through global experiment
Optimal predictor
Similarity predictor
Extract feature by SOM
Characteristic factor
Independent predictor
Standardized
Input NARX network
Training open-loop NARX networks
Convert to closed-loop NARX and predict
Step1:Select the optimal predictor based on MIC
Step2:Construct feature sequence based om SOM network
Step3:Training and prediction using NARX network

Figure 4. SOM-NARX network architecture

network can reduce the seasonal differences of the NARX network. In multistep forecasting, the SOM-NARX network has higher prediction accuracy than the NARX network.

Through model comparison, although the stability and applicability of CNN in different seasons are better than that of the SOM network, NARX is better than LSTM, BiLSTM, and GRU in fitting the interaction relationship between different predictors, effectively compensating for the shortcomings of SOM network in feature extraction. The RMSE, MAE, and MAEP of the SOM-NARX model in winter and summer were 13.82, 10.60, 6.58% and 12.05, 9.44, 68.14%, respectively, with smaller prediction errors than CNN-LSTM, CNN-BiLSTM, CNN-GRU, SOM-LSTM, SOM-BiLSTM, and SOM-GRU.

Limitations of the study

The SOM-NARX network model based on the correlation of predictors is not only suitable for ozone prediction, but also for the early warning of other pollutants and prediction of carbon emissions, providing a scientific reference for pollution prevention and control. A limitation of this study is that the dataset is obtained from a 13-month time series at one site and does not consider the spatial correlation of predictors with ozone. Longer time horizons and broader target areas can help obtain more accurate forecasts and deeper insights. Inter-regional interactions are also the next focus of this study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Research area and data
  - Pre-processing of data
  - Maximum information coefficient (MIC)
  - Self-organizing mapping neural network (SOM)
  - Nonlinear AutoRegressive models with exogenous inputs (NARX)
  - Evaluation indicators
  - Construction of SOM-NARX model

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105658.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (NO.41977284).

AUTHOR CONTRIBUTIONS

Q. X. and W. W.: investigation, conceptualization, writing, formal analysis, data curation, and editing; M. W. and M. W.: conceptualization, methodology, supervision, project administration, and review; C. C.: investigation, experimental analysis, data curation, and review; X. Z. and C. Z.: experimental analysis, data curation, investigation, and experimental analysis.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 12, 2022
Revised: October 22, 2022
Accepted: November 21, 2022
Published: December 22, 2022

REFERENCES

1. MEEPRC (2021). China Ecological Environment Status Bulletin, p. 2020.

2. Lancet, T. (2006). WHO's global air-quality guidelines, Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, *368* (World Health Organization (WHO)), p. 1302.

3. Shin, H.H., Parajuli, R.P., Gogna, P., Maquiling, A., and Dehghani, P. (2021). Pollutant-sex specific differences in respiratory hospitalization and mortality risk attributable to short-term exposure to ambient air pollution. Sci. Total Environ. *755*, 143135. https://doi.org/10.1016/j.scitotenv.2020.143135.

4. Ruan, F.F., Liu, J.X., Chen, Z.W., and Zeng, X.G. (2022). Meta-analysis of the impact of different ozone metrics on total mortality in China. Environ. Sci. J. Integr. Environ. Res. *43*, 37–45. https://doi.org/10.13227/j.hjkx.202105118.

5. Li, Z., Yue, M., Heng, B.C., Liu, Y., Zhang, P., and Zhou, Y. (2022). Factors influencing ozone production and characteristics of ozone pollution in China. Int. J. Oral Sci. *14*, 54–56. https://doi.org/10.16317/j.cnki.12-1377/x.2022.05.020.

6. Lin, W., and Guo, X. (2022). Spatial and temporal distribution characteristics of ozone in Urban agglomerations in China. China Environ. Sci. *42*, 2481–2494. https://doi.org/10.19674/j.cnki.issn1000-6923.20220207.002.

7. Jiang, H., Zheng, X., Hao, J., Dai, A., and Jiang, N. (2022). Spatial and temporal distribution characteristics of ozone concentration in China from 2014 to 2020. J. Green Sci. Tech. *24*, 159–162. (in Chinese). https://doi.org/10.16663/j.cnki.lskj.2022.10.031.

8. Liu, Y. (2021). Research on the spatiotemporal distribution characteristics and influencing factors of ozone pollution in China. CNKI. https://doi.org/10.26991/d.cnki.gdllu.2021.001779.

9. Song, X, Yan, L, Liu, W, He, J, Wang, Y, Huang, T, Li, Y, Chen, M, Meng, J and Hou, Z, Spatiotemporal distribution characteristics of co-pollution of PM2.5 and ozone over BTH with surrounding area for years of 2015-2021. Environ. Sci. J. Integr. Environ. Res. (In Chinese). 1-17. 10.13227/j.hjkx.202205089

10. Cuchiara, G.C., Li, X., Carvalho, J., and Rappenglück, B. (2014). Intercomparison of planetary boundary layer parameterization and its impacts on surface ozone concentration in the WRF/Chem model for a case study in Houston/Texas. Atmos. Environ. X. *96*, 175–185. https://doi.org/10.1016/j.atmosenv.2014.07.013.

11. Thomas, A., Huff, A.K., Hu, X., and Zhang, F. (2019). Quantifying uncertainties of ground-level ozone within WRF-chem simulations in the mid-atlantic region of the United States as a response to variability. J. Adv. Model. Earth Syst. *11*, 1100–1116. https://doi.org/10.1029/2018MS001457.

12. Zhong, M., Saikawa, E., Liu, Y., Naik, V., Horowitz, L.W., Takigawa, M., Zhao, Y., Lin, N.H., and Stone, E.A. (2016). Air quality modeling with WRF-Chem v3.5 in East Asia: sensitivity to emissions and evaluation of simulated air quality. Geosci. Model Dev. (GMD) *9*, 1201–1218. https://doi.org/10.5194/gmd-9-1201-2016.

13. Li, Y.R., Han, T.T., Wang, J.X., Quan, W.J., He, D., Jiao, R.G., Wu, J., Guo, H., and Ma, Z.Q. (2021). Application of arima model for mid- and long-term forecasting of ozone concentration. Environ. Sci. J. Integr. Environ. Res. *42*, 3118–3126. https://doi.org/10.13227/j.hjkx.202011237.

14. Sun, W., Zhang, H., and Palazoglu, A. (2013). Prediction of 8 h-average ozone concentration using a supervised hidden Markov model combined with generalized linear models. Atmos. Environ. X. *81*, 199–208. https://doi.org/10.1016/j.atmosenv.2013.09.014.

15. Lengyel, A., Héberger, K., Paksy, L., Bánhidi, O., and Rajkó, R. (2004). Prediction of ozone concentration in ambient air using multivariate methods. Chemosphere *57*, 889–896. https://doi.org/10.1016/j.chemosphere.2004.07.043.

16. Arsić, M., Mihajlović, I., Nikolić, D., Živković, Ž., and Panić, M. (2020). Prediction of ozone concentration in Ambient air using multilinear regression and the artificial neural networks methods. Ozone: Sci. Eng. *42*, 79–88. https://doi.org/10.1080/01919512.2019.1598844.

17. Jia, M., Zhao, T., Cheng, X., Gong, S., Zhang, X., Tang, L., Liu, D., Wu, X., Wang, L., and Chen, Y. (2017). Inverse relations of PM2.5 and O3 in air compound pollution between cold and hot seasons over an urban area of east China. Atmosphere *8*, 59. https://doi.org/10.3390/atmos8030059.

18. Yi, J., and Prybutok, V.R. (1996). A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environ. Pollut. *92*, 349–357. https://doi.org/10.1016/0269-7491(95)00078-X.

19. Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., and Jung, J. (2020). Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance. Neural Network. *121*, 396–408. https://doi.org/10.1016/j.neunet.2019.09.033.

20. Feng, R., Zheng, H.J., Zhang, A.R., Huang, C., Gao, H., and Ma, Y.C. (2019). Unveiling tropospheric ozone by the traditional atmospheric model and machine learning, and their comparison:A case study in hangzhou, China. Environ. Pollut. *252*, 366–378. https://doi.org/10.1016/j.envpol.2019.05.101.

21. Cheng, Y., He, L.Y., and Huang, X.F. (2021). Development of a high-performance machine learning model to predict ground ozone pollution in typical cities of China. J. Environ. Manag. *299*, 113670. https://doi.org/10.1016/j.jenvman.2021.113670.

22. Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., and Chi, T. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Sci. Total Environ. *654*, 1091–1099. https://doi.org/10.1016/j.scitotenv.2018.11.086.

23. Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., and Bi, J. (2020). Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: a machine learning approach. Environ. Int. *142*, 105823. https://doi.org/10.1016/j.envint.2020.105823.

24. Cabaneros, S.M., Calautit, J.K., and Hughes, B.R. (2019). A review of artificial neural network models for ambient air pollution prediction. Environ. Model. Software *119*, 285–304. https://doi.org/10.1016/j.envsoft.2019.06.014.

25. Chen, Z., Li, R., Chen, D., Zhuang, Y., Gao, B., Yang, L., and Li, M. (2020). Understanding the causal influence of major meteorological factors on ground ozone concentrations across China. J. Clean. Prod. *242*, 118498. https://doi.org/10.1016/j.jclepro.2019.118498.

26. Wang, H.W., Li, X.B., Wang, D., Zhao, J., He, H.d., and Peng, Z.R. (2020). Regional prediction of ground-level ozone using a hybrid sequence-to-sequence deep learning approach. J. Clean. Prod. *253*, 119841. https://doi.org/10.1016/j.jclepro.2019.119841.

27. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., and Sabeti, P.C. (2011). Detecting novel associations in large data sets. Science *334*, 1518–1524. https://doi.org/10.1126/science.1205438.

28. Kohonen, T. (1987). Adaptive, associative, and self-organizing functions in neural computing. Appl. Opt. *26*, 4910–4918. https://doi.org/10.1364/AO.26.004910.

29. Lin, T., Horne, B.G., Tino, P., and Giles, C.L. (1996). Learning long-term dependencies in NARX recurrent neural networks. IEEE Trans. Neural Network. *7*, 1329–1338. https://doi.org/10.1109/72.548162.

30. Bai, Y., Zeng, B., Li, C., and Zhang, J. (2019). An ensemble long short-term memory neural network for hourly PM2.5 concentration forecasting. Chemosphere *222*, 286–294. https://doi.org/10.1016/j.chemosphere.2019.01.121.

31. Catalano, M., Galatioto, F., Bell, M., Namdeo, A., and Bergantino, A.S. (2016). Improving the prediction of air pollution peak episodes generated by urban transport networks. Environ. Sci. Pol. *60*, 69–83. https://doi.org/10.1016/j.envsci.2016.03.008.

32. Mishra, D., and Goyal, P. (2016). Neuro-Fuzzy Approach to Forecasting Ozone Episodes over the Urban Area of Delhi, India. Environ. Tech. Innov. *5*, 83–94. https://doi.org/10.1016/j.eti.2016.01.001.

33. Fernando, H.J.S., Mammarella, M.C., Grandoni, G., Fedele, P., Di Marco, R.,

Dimitrova, R., and Hyde, P. (2012). Forecasting PM10 in metropolitan areas: efficacy of neural networks. Environ. Pollut. *163*, 62–67. https://doi.org/10.1016/j.envpol.2011.12.018.

34. Arhami, M., Kamali, N., and Rajabi, M.M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. Environ. Sci. Pollut. Res. Int. *20*, 4777–4789. https://doi.org/10.1007/s11356-012-1451-6.

35. Kumar, N., Middey, A., and Rao, P.S. (2017). Prediction and examination of seasonal variation of ozone with meteorological parameter through artificial neural network at NEERI, Nagpur, India. Urban Clim. *20*, 148–167. https://doi.org/10.1016/j.uclim.2017.04.003.

36. Liu, J., Lu, K., and Liu, X. (2020). The analysis and countermeasures of ozone pollution of the ambient air in Jiaozuo city. J. Jiaozuo Univ. *34*, 92–94. https://doi.org/10.16214/j.cnki.cn41-1276/g4.2020.04.026.

37. Carter, W. (2010). Updated maximum incremental reactivity scale and hydrocarbon bin reactivities for regulatory applications. California Air Resources Board Contract *1*, 7–339. https://doi.org/10.1126/science.1205438.

38. Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., and Pak, C. (2020). Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: a case study of Beijing, China. Sci. Total Environ. *699*, 133561. https://doi.org/10.1016/j.scitotenv.2019.07.367.

39. Fan, R., Meng, D., and Xu, D. (2014). Survey of research process on statistical correlation analysis. Mathematical Modeling and Its Applications *3*, 1–12.

40. Tan, S., Zhang, X., Li, Q., and Ai, C. (2018). Information push model-building based on maximum mutual information coefficient. J. Jilin Univ. (Eng. Technol. Ed.) *48*, 558–563. https://doi.org/10.1126/science.1205438.

41. Jolliffe, I.T. (2002). Principal component analysis. J. Marketing Res. *87*, 513.

42. Boulesteix, A.L., and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings Bioinf. *8*, 32–44. https://doi.org/10.1093/bib/bbl016.

43. Liu, Y., Weisberg, R.H., and Mooers, C.N.K. (2006). Performance evaluation of the self-organizing map for feature extraction. J. Geophys. Res. *111*, C05018. https://doi.org/10.1029/2005JC003117.

44. Chang, L.C., Shen, H.Y., and Chang, F.J. (2014). Regional flood inundation nowcast using hybrid SOM and dynamic neural networks. J. Hydrol. X. *519*, 476–489. https://doi.org/10.1016/j.jhydrol.2014.07.036.

45. Pak, U., Kim, C., Ryu, U., Sok, K., and Pak, S. (2018). A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. Air Qual. Atmos. Health *11*, 883–895. https://doi.org/10.1007/s11869-018-0585-1.

46. Ma, J., Li, Z., Cheng, J.C.P., Ding, Y., Lin, C., and Xu, Z. (2020). Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network. Sci. Total Environ. *705*, 135771. https://doi.org/10.1016/j.scitotenv.2019.135771.

47. Leontaritis, I.J., and Billings, S.A. (1985). Input-output parametric models for non-linear systems Part I: deterministic non-linear systems. Int. J. Control *41*, 303–328. https://doi.org/10.1080/0020718508961129.

48. Menezes, J.M.P., and Barreto, G.A. (2008). Long-term time series prediction with the NARX network: an empirical evaluation. Neurocomputing *71*, 3335–3343. https://doi.org/10.1016/j.neucom.2008.01.030.

49. Narendra, K.S., and Parthasarathy, K. (1991). Learning automata approach to hierarchical multiobjective analysis. IEEE Trans. Syst. Man Cybern. *21*, 263–272.

50. Graupe, D. (2016). Deep Learning Neural Networks: Design and Case Studies (World Scientific Publishing Company).

51. Marquardt, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. *11*, 431–441.

52. Hagan, M., and Menhaj, M. (1994). Training feedforward networks with the marquardt algorithm. IEEE Trans. Neural Network. *5*, 989–993. https://doi.org/10.1109/72.329697.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Pollutants data (national control sites)	Qingyue Data	https://data.epmap.org/page/index
Meteorological data	NASA GES DISC MERRA	https://disc.gsfc.nasa.gov/
VOCs data	Demonstration area VOCs detection station	https://doi.org/10.57760/sciencedb.06259; https://www.scidb.cn/s/Y3YFri
Software and algorithms		
IBM SPSS	IBM	https://www.ibm.com/analytics/spss-statistics-software
Origin 2018	OriginLab	https://www.originlab.com/
MATLAB 2020b	MathWorks	https://ww2.mathworks.cn/?s_tid=gn_logo
Maximal Information coefficient	Reshef et al.[27]	https://doi.org/10.57760/sciencedb.06273; https://www.scidb.cn/s/JFBJry
Self-Organizing Mapping Neural Network	Kohonen et al.[28]	https://doi.org/10.57760/sciencedb.06273; https://www.scidb.cn/s/JFBJry
Nonlinear AutoRegressive models with exogenous inputs	Tsungnan et al.[29]	https://doi.org/10.57760/sciencedb.06273; https://www.scidb.cn/s/JFBJry

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Mingshi Wang (mingshiwang@hpu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. A detailed description of the data can be found in the supplementary document "Research area and data".
- All original code has been deposited at https://www.scidb.cn/s/JFBJry and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Research area and data

Jiaozuo City is one of the "2 + 26 cities" in China with serious ozone pollution, and the annual average value of O_3-M8h in 2020 is ranked second to last among 337 cities,[1] especially in Jiaozuo City Demonstration Zone of Urban-Rural Integration (Abbreviation JCDZURI), where ozone pollution is the most serious. JCDZURI is located in the south of Jiaozuo City, which is a complex functional area with urban-rural integration, industrial integration and integrated development. It contains one national high-tech industrial development zone, one two-star industrial cluster in Henan Province, a national demonstration base for new industrialization industry, a national science and technology business incubator, and the second batch of national pilot area for science and technology service industry.

The data set includes hourly data of 6 pollutants (national control sites), 31 meteorological factors, and 118 VOCs from the Sanqing website (https://imee.3clear.com) NASA GES DISC MERRA (https://disc.gsfc.nasa.gov/), and VOCs fixed detection stations in the demonstration area. In this study, the datasets are divided into three types: training set, test set and validation set. The training set is used to train the network to learn the features in the dataset. In using machine learning to predict air pollutants, many scholars tend to use one year of data as the training set,[30–33] because the maximum period of temporal variation of pollutants is one year and one year of data covers all temporal characteristics of pollutants,[34,35] such as annual characteristics, seasonal characteristics, daily characteristics, weekend effects, etc. Therefore, the full year 2021 data, containing 8760 hourly data from 1 January to 31 December, was selected as the training set for this study. The test set and validation set are used to tune the model parameters and evaluate the performance of the model respectively. The study showed that the annual variation of ozone in Jiaozuo City gradually increased from January and remained at a high level from May to September.[36] In order to investigate the difference in model accuracy between high and low values, the test and validation sets were taken for 15 consecutive days in January 2022 and September 2020, respectively.

Pre-processing of data

The acquired training data were missing 170 hourly data for pollutants and 254 for VOCs. This study interpolated the missing data using linear interpolation to reduce the impact of non-continuous data on temporal correlation. On the other hand, different parameters have different scales, and large values disproportionately mask the impact of smaller inputs.[24] As a result, the data is preprocessed using min-max normalization or standardization. Calculations as Equations 1 and 2.

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad \text{(Equation 1)}$$

$$x' = \frac{x - \bar{x}}{\sigma} \qquad \text{(Equation 2)}$$

where x* is the normalized value, dimensionless; x is the data at a moment of the factor; xmin and xmax represent, respectively, the minimum and maximum values in the dataset made up of x; x' is the standardized value, dimensionless; $\bar{x}$ is the mean; and σ is the standard deviation.

Numerous VOC species contribute differently to ozone, and the ozone formation potential (OFP) based on the maximum incremental reactivity (MIR) can assess the ability of VOCs to produce ozone. In this study, to establish a mapping relationship between VOCs and ozone, OFP was used as a predictor of the model, OFP is calculated as Equations 3 and 4.

$$OFP_i = MIR_i \times [VOC_i] \qquad \text{(Equation 3)}$$

$$OFP = \sum OFP_i \qquad \text{(Equation 4)}$$

where VOC_i is the observed concentration of VOC species i; MIR_i is the maximum incremental reactivity coefficient for i; The coefficient comes from the MIR updated by Carter in 2010.[37]

Maximum information coefficient (MIC)

Predictors are critical to the model's prediction performance, and too many irrelevant variables or missing key variables will affect the prediction accuracy.[24] Predictor screening is mainly divided into trial-and-error[26] and analytical methods,[38] although the trial-and-error method is simple to operate, the amount of arithmetic is very large and does not reflect the relationship between predictors and accuracy, while the analytical method based on factor correlation is better than the trial-and-error method. However, Pearson and Spearman correlation coefficients are only sensitive to linear relationships and cannot effectively capture the nonlinear relationships between both meteorological factors and precursors and ozone. Although mutual information (MI) has good performance in analyzing nonlinear relationships between variables, the probability density functions of the variables are unknown and the mutual information is difficult to estimate.[39,40] In contrast, MIC is applicable to any functional relationship, whether linear or nonlinear, and the outliers of the variables have less impact on the results. Therefore, this study used the maximum information coefficient (MIC) to screen out factors with some correlation with ozone as predictors.

Reshef proposed the maximum information coefficient (MIC) to analyze the nonlinear correlation of big data.[27] MIC is calculated by mutual information and grid division. Mutual information is an important indicator for determining the degree of correlation between variables, and it is defined as Equations 5–8:

$$\mathrm{MI}(A,B) = \sum_{a\in A}\sum_{b\in B} P(a,b)\log_2\frac{P(a,b)}{P(a)P(b)} \qquad \text{(Equation 5)}$$

$$\mathrm{MI}^*(D,x,y) = \max \mathrm{MI}\left(\frac{D}{G}\right) \qquad \text{(Equation 6)}$$

$$M(D)_{x,y} = \frac{\mathrm{MI}^*(D,x,y)}{\log_2 \min\{x,y\}} \qquad \text{(Equation 7)}$$

$$\mathrm{MIC}(D) = \max_{xy<B(n)}\left\{M(D)_{x,y}\right\} \qquad \text{(Equation 8)}$$

Where $A = \{a_i, i = 1, 2, \cdots, n\}$; $B = \{b_i, i = 1, 2, \cdots, n\}$; n denotes the number of samples; The joint probability density of A and B is p(a, b); The marginal probability densities of A and B are denoted by p(a) and p(b), respectively; MIC is the maximum information coefficient; D/G denotes that data D is divided using G; $M(D)_{x,\ y}$ is the maximum normalized MI value obtained by dividing a feature matrix into different divisions; B(n) is the upper limit of grid division x×y, which is generally defined as $\omega(1) \le B(n) \le O(n^{1-\varepsilon})$, $0<\varepsilon<1$.

Self-organizing mapping neural network (SOM)

The mechanism of ozone generation is complex, and the effects of similar variables (such as net sunshine intensity and UV light intensity, ground temperature and air temperature, etc.) on ozone generation are still unclear. If similar variables are excluded, the prediction accuracy may be reduced due to the lack of key factors. Therefore, this research adopts the method of feature extraction to achieve the purpose of dimensionality reduction. Traditional feature extraction methods such as Principal Component Analysis (PCA),[41] Partial Least Squares (PLS),[42] etc. mostly rely on the correlation between variables or specific functional relationships, and will be limited to varying degrees.[43] The self-organizing map network can cluster high-dimensional input vectors without any prior knowledge of statistical distribution to form a visual one-dimensional or multi-dimensional topology map and maintain the relationship between variables.[44]

The SOM network is an unsupervised neural network proposed by Kohonen based on the phenomenon of "lateral inhibition" between biological neurons,[28] as shown in Figure S1. It includes an input layer and a clustering layer consisting of nodes in a two-dimensional graph. Associated with each node is a weight vector with the same dimensions as the input vector (variable, e.g., ground temperature) and a location in map space. The training method of SOM is different from other neural networks. Different from other neural networks, the training network of SOM adopts the competitive learning method. Different from other neural networks, the training network of SOM adopts the competitive learning method. The winning neuron is determined according to the minimum Euclidean distance d_j between the weight vector ω(j) and the input vector x(t). Winning neurons and in-neighbor neurons can update weights. Calculated as Equations 9–13:

$$d_j = x(t) - \omega(j) = \sqrt{\sum_{i=n}\left(x(t)_i - \omega(j)_i\right)^2} \qquad \text{(Equation 9)}$$

$$r(t+1) = \mathrm{INT}\left((r(t)-1)\times\left(1-\frac{t}{T}\right)\right)+1 \qquad \text{(Equation 10)}$$

$$k_j = e^{\left(-\frac{\omega(j)-\omega(j)^{*2}}{2r^2 t}\right)} \qquad \text{(Equation 11)}$$

$$\eta(t+1) = \eta(t) - \frac{\eta(0)}{T} \qquad \text{(Equation 12)}$$

$$\omega(j)_{new} = \omega(j)_{old} + \eta(t)k_j\left(x(t) - \omega(j)_{old}\right) \qquad \text{(Equation 13)}$$

where n is the dimension of the input vector and weight vector, ω(j) represents the j_{th} competition layer neuron, x(t) represents the t_{th} input layer neuron; k_j is the neighborhood function, generally using Gaussian function or bubble The function, which represents the update coefficient of the weight vector in the neighborhood of the j_{th} winning neuron, the smaller the Euclidean distance between the neighborhood neuron and the winning neuron, the larger the value of k_j; r(t) is the winner when the t_{th} vector is input Neighborhood radius of the neuron, INT is the rounding function, T is the iteration period; ω(j) denotes the j_{th} winning neuron's weight vector; The winning neuron neighborhood's weight vector is ω(j)*; t is the learning efficiency function, which gradually decreases with the iterative training, and the value range is (0, 1).

Nonlinear AutoRegressive models with exogenous inputs (NARX)

In pollutant prediction, dynamic neural network RNN (such as LSTM, GRU, etc.) with memory function has been widely used due to its irreplaceable advantages in dealing with time series problems, but it is still insufficient in solving function approximation, so many Scholars build hybrid prediction models based on LSTM, such as CNN-LSTM hybrid model,[45] spatial transfer bidirectional long short-term memory network,[46] Attention-based Seq2Seq learning for temporal modeling,[26] etc. Another dynamic neural network, NARX, also has good performance in prediction. It adds a delay and feedback mechanism on the basis of a multi-layer perceptron, which makes it have the ability to memorize historical data and can take into account both time series and function approximation[29,47,48]

The NARX network is a two-layer feedforward network, the hidden layer is mainly in the form of full connection, and the sigmoid is used as the activation function. The output y_t is influenced by the input ($x_{t,}$ x_{t-1}, x_{t-2}, x_{t-3}, ⋯ x_{t-n}) and (y_{t-1}, y_{t-2}, y_{t-3}, ⋯ y_{t-n}), where n is also called the delay step. The Equation 14 is the NARX model's defining equation:

$$y(t) = f\left(y(t-1), y(t-2), \cdots, y(t-n\_y), x(t-1), x(t-2), \cdots, x(t-n_x)\right) \quad \text{(Equation 14)}$$

where x is the input variable, y is the output variable, and n_y and n_x are the delay steps of the output and input variables, respectively.

The NARX network is divided into open-loop mode and closed-loop mode, as shown in Figure S2. The open-loop mode is a network without feedback, and the target and predictor are combined into the input layer; the closed-loop mode is a network with a feedback mechanism, and the output value is fed back to the input layer of the feedforward network instead of the target value. Usually, the network is trained in open-loop mode to ensure the accuracy of the input; prediction is made in closed-loop mode to achieve dynamic prediction.[49]

Given that the network in open-loop mode is a feedforward neural network, training can be done using the static back-propagation technique to increase training effectiveness. The mean square error is used to adjust the network's weights and biases during training (MSE). The training algorithm adopts Levenberg-Marquardt, which trains very fast. Although it is inefficient to train large networks with many weights, it has good performance in fitting nonlinear regression.[50–52]

The reasons for using the NARX network can be found in the supplementary document " Nonlinear AutoRegressive models with exogenous inputs (NARX)".

Evaluation indicators

This study employs three statistical indicators to assess model performance: mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE), with lower values representing better model performance, as defined as Equations 15–17:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \quad \text{(Equation 15)}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|O_i - P_i|}{O_i} \quad \text{(Equation 16)}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - p_i)^2} \qquad \text{(Equation 17)}$$

where O_i is the value of the observed data, P_i is the predicted value, and n is the length of the data.

Construction of SOM-NARX model

The SOM-NARX model uses the SOM network to cluster the time series of similar predictors, thereby transforming them into corresponding feature factor sequences. Finally, the eigenfactors are fed into the NARX network together with other predictors for training and prediction. Before the SOM-NARX model predicts, it is necessary to screen out the factors highly correlated with ozone as predictors based on MIC. The construction of the model is divided into three parts, as shown in Figure 4. Screening predictors based on MIC; 2. Constructing characteristic factors based on SOM network; 3. Using the NARX network to train and predict ozone. Please refer to the Supporting information for the detailed process.

Step 1. Screening predictors based on MIC

Mutual information (MI) coefficient estimated probability density from histogram, assuming that D = {(ai, bi), i = 1, 2, …, n} is a finite set of ordered pairs, grid division G (an x×y grid) divides the range of variable A into x segments and the range of B into y segments. Mutual information MI (A, B) is calculated inside each of the obtained grid divisions. There are many grid division methods for the same x×y, and the maximum value of MI (A, B) in different division methods is taken as the value of G. mutual information value. We refer to Reshef's research[27] and set the number of networks to be $B(n) = n^{0.6,}$ where n is the data volume of 8760 in 2021.

Step 2. Constructing characteristic factors based on SOM network

The 8760 hourly data of similar factors are divided into m categories, and the cluster labels are used as eigenvalues, and the new time series formed is called the eigenfactor sequence. The eigenfactors and other predictors make up the input variables for subsequent models. The number of classifications "m" is the size of the output layer of the SOM network, which determines the topology of the network mapping. We set the data structure of m to be a one-dimensional column vector. because a one-dimensional column vector can maintain the continuity and temporal correlation of the eigenfactor sequence. The optimal number of classifications is determined through global experiments.

Step 3. Using the NARX network to train and predict ozone

The prediction module uses the standard NARX network. Trained in open-loop mode, the input layer consists of the predictor x1, the feature factor x2, and the target variable y. The root mean square error (MSE) is used as the training basis. The training is finished if the MSE does not decrease for six consecutive tests. The closed-loop mode is used for prediction. The output y'(t) at time t and the predictor x1(t+1) and characteristic factor x2(t+1) at time t+1 constitute input variables. The variables are mapped by the trained NARX network and the output value y'(t+1) at time t+1, and y'(t+1) is the predicted value. For example, to predict the ozone concentration for 8 consecutive hours from 12:00 to 19:00 on January 1, 2022, it is necessary to input the variables at n times before 12:00 into the closed-loop NARX network to obtain the first predicted value at 12:00. The predicted value at 12:00 is used as the target variable in the input layer to predict the next moment at 13:00. Continuous prediction is achieved by circular mapping. The optimal parameters of the preprocessing of the input variables, the delay step and the number of neurons in the hidden layer are gradually determined through global experiments.