CSIAM Trans. Life Sci.
doi: 10.4208/csiam-ls.SO-2025-0028

Vol. **x**, No. x, pp. 1-22
xx 2026

Discovering Mechanistic Correlations Among Respiratory Diseases and Air Quality via Dynamic Modelling Combined with Deep Learning and Symbolic Regression

Mengqi He[1,2], Dingding Yan[3], Sha He[4], Jianhong Wu[5] and Sanyi Tang[6,*]

[1] *The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China.*
[2] *School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an 710061, China.*
[3] *School of Mathematics, Northwest University, Xi'an 710127, China.*
[4] *School of Mathematics and Statistics, Shaanxi Normal University, Xi'an 710119, China.*
[5] *Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, ON, M3J 1P3, Canada.*
[6] *School of Mathematics and Statistics, Shanxi University, Taiyuan 030006, China.*

Received 30 October 2025; Accepted 19 January 2026

Abstract. Air pollution and disease transmission constitute a complex feedback cycle. However, the mechanistic relationship between the variation of air pollution (measured by air quality index (AQI)) and the disease transmission dynamics remains incompletely understood. Here we develop a framework to explore this relationship and we illustrate this framework by inferring the disease transmission rate and inflow rate of air pollutants (IRAP) from AQI and disease incidence data. The coupled system of disease transmission dynamics model and AQI dynamics model is integrated into physics-informed neural networks to embed the information retrieved from the coupled system to the loss function of the neural networks using automatic differentiation, the outcome of this data-driven network for transmission rate and IRAP is then converted into analytic forms of transmission rate and IRAP depending on incidence and AQI, using correlation analyses and symbolic regressions. Based on data from Shaanxi Province of China, this framework is used to establish a linear positive correlation between transmission rate and AQI, and between IRAP and disease incidence. We observe frequent fluctuations in transmission rate and IRAP and show that the mechanistic model with nonlinear analytic forms for transmission rate and IRAP learned through symbolic regression has robust predictive capabilities.

AMS subject classifications: 92B05, 92B20, 92D30

Key words: Air pollution, disease transmission, neural networks, automatic differentiation, symbolic regression.

*Corresponding author. *Email addresses:* sytang@sxu.edu.cn, sytang@snnu.edu.cn (S. Tang)

1 Introduction

Comprehending inducing factors of infectious disease transmission is crucial for devising and implementing effective intervention measures to mitigate the burden of diseases anyway in the world [16]. This is particularly so for countries such as China with rapidly developing economy confronting a series of public health challenges arising from environmental changes [18, 46]. From 2010 to 2016, northern China suffered from severe smog, resulting in a clear upward trend in the air quality index during this period. Despite air quality in China has improved in recent years [6], fog and haze events are still common during the winter months in northern China, which seriously threatens human health. Therefore, quantifying the relationship between air pollution and disease transmission emerges as an important component of interdisciplinary research in tracking the driving factors of infectious disease spread [9, 29, 49].

There is increasing epidemiological evidence to show that air pollutants can have various adverse effects on the human health [17,34,37]. Particulate matter, sulfur dioxide, nitrogen oxides, carbon monoxide and volatile organic compounds in the air are shown to be among major factors causing respiratory diseases and cardiovascular diseases [12, 25,41,45]. Studies show that both short-term and long-term exposure to air pollution can reduce lung function, leading to increases in emergency visits for asthma and in the risk of chronic obstructive pulmonary disease [4, 10, 11, 32]. However, impacts of different types of air pollutants on various respiratory diseases vary [24]. Although numerous studies have confirmed that respiratory diseases are affected by air pollution, accurately quantifying the association between respiratory infection risk and AQI remain unclear and fall within the scope of this study.

Many data-driven methods have been constructed to better understand the seasonal fluctuations in air pollution and their implications for human health [26,28,30]. Dairi *et al.* [8] developed an integrated multiple directed attention deep learning architecture (called IMDA-VAE) based on the traditional variational autoencoder (VAE) and the attention mechanism, and used air pollution data from four US states to evaluate the predictive performance of their proposed method. Their results indicate that the proposed IMDA-VAE model exhibited satisfactory effectiveness in forecasting different pollutants in different locations. To examine the effect of air pollution on patient's hospital visits for respiratory diseases (and particularly acute respiratory infections (ARI)), Ravindra *et al.* [36] used eight machine learning algorithms (Random Forest, K-Nearest Neighbors regression, Linear regression, LASSO regression, Decision Tree Regressor, Support Vector Regression, X.G. Boost and Deep Neural Network) to analyze daily air pollutants and outpatient visits for ARI. To estimate daily hospital admissions due to air pollution, Araujo *et al.* [2] employed artificial neural networks to assess hospital admissions for respiratory diseases caused by particulate matter and meteorological variables of Campinas and São Paulo cities, Brazil. By collecting data on influenza-like visit rates, and linking the air pollution indicators and open data on temperature and humidity environmental factors, Kristiani *et al.* [21] implemented a deep ensemble technology to predict the rate

of emergency visits to influenza-like illness patients. Note that these data-driven methods do not incorporate prior disease transmission mechanisms, making it difficult to gain insights into the mechanism through which air pollution affects disease spread.

There are also several dynamic model-based methods developed to investigate the relationship between respiratory diseases and air pollution [5, 15, 42]. In order to measure the impact of air pollution on that respiratory infection risk, Tang *et al.* [40] developed an ordinary differential equation model with AQI-dependent incidence and AQI-based behaviour change interventions to depict the AQI trend and respiratory infection dynamics. Considering that both the removal of air pollutants and the transmission of diseases can be influenced by seasonal changes and stochastic perturbations, He *et al.* [13] introduced a coupled periodic stochastic differential equation model to study the dynamics of infectious respiratory disease transmission in an environment with air pollution. These studies indicate that the periodicity of air pollutants drives the periodicity of changes in the number of infected individuals. Zhou *et al.* [51] developed a reaction-diffusion model to determine the influences of spatial heterogeneity and mobility of infected individuals on the spread of respiratory infectious diseases affected by air pollution, their results show that spatial heterogeneity can facilitate the spread of respiratory infectious diseases affected by air pollution. In addition, with a new tuberculosis (TB) transmission dynamics model incorporating contaminated environments, Wang and Cai [43] investigated the effects of long-term exposure to ambient particulate air pollution on the TB transmission dynamics in Jiangsu, China, and claimed that PM_{10} is closely related to the incidence of TB. In these studies, various time-dependent rate functions are assumed to be in prescribed forms in order to simulate the dynamics of air pollution and disease transmission. These prescribed forms of time-dependent rate functions, reflecting the complex relationship between respiratory infections and air quality, have yet to be rigorously derived, informed and validated from the data. In addition, to integrate the dynamic changes in air pollution with the dynamic transmission of viral diseases requires a framework and technique to mechanistically couple the two dynamic processes in different scales, evaluated through multi-source data.

To address this intractable problem of deriving analytic coupling mechanisms from multi-source data, we propose here to leverage physics-informed neural networks (PINNs) [35] that integrate observational data, multi-scale models, and deep learning. Our approach compels training and learning neural networks to simultaneously adhere to the disease spread dynamics and AQI variations during the learning process, and meanwhile avoids assuming the specific rate functions a-priori. We illustrate and test this proposed method by simulating the trends of influenza-like illness and AQI in the city of Xi'an (Shaanxi province, China) from 2010 to 2016, to infer the temporal evolution patterns of dynamic parameters reflecting infection risk and air pollutants influx intensity. Subsequently, the precise analytical formulations that can characterize the complex relationship between respiratory infection risk and air pollution discovered by correlation analysis and symbolic regression. Finally, converting the dynamic parameters represented by neural networks to mathematical expressions results in a bidirectionally

coupled multi-scale system, which can accurately capture and predict the evolutionary trends of influenza-like illness (ILI) cases and AQI.

In this study, by coupling the deep neural network with the transmission dynamics of ILI cases and the evolution dynamics of air pollution, the bidirectional driving of the model and data has been achieved, and the interpretability of the unknown mechanism learning has been improved by reconstructing the function of the unknown mechanism. The proposed method utilizes the ability of deep learning to perform collaborative learning on multi-source data, providing new insights for constructing dynamical models of complex multi-scale coupled systems and reconstructing unknown mechanisms within these models.

2 Data and methodology

2.1 The data

We retrieved data on respiratory infections from the surveillance system at the Shaanxi Center for Disease Control and Prevention. The data consists of reports of daily cases who seek medical attention with influenza-like illness (symptoms typically include a body temperature exceeding 38°C, chills, dry cough, sore throat, loss of appetite, body aches, and nausea) in Xi'an from 15th November 2010 to 14th November 2016 [40], as shown in Fig. 1(A). Air pollution, a complex amalgamation of various pollutants, is communicated to the public by governmental agencies using the air quality index. AQI is a method to reflect and evaluate air quality. It is a comprehensive indicator that simplifies the concentrations of six routinely monitored air pollutants (including $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3 and CO) into a standardized index value, which is suitable for characterizing the air quality status and degree of air pollution in a city. The higher the index, the more serious the air pollution. According to the definition by Ministry of Environmental Protection of the People's Republic of China, AQI can be calculated using the following formula [22]:

$$IAQI_p = \frac{IAQI_{high} - IAQI_{low}}{BP_{high} - BP_{low}}(C_p - BP_{low}) + IAQI_{low}, \tag{2.1}$$

$$AQI = \max\{IAQI_1, IAQI_2, \cdots, IAQI_6\}. \tag{2.2}$$

First, the individual air quality index (IAQI) is calculated for each pollutant. Then, the maximum IAQI value from the six pollutants is selected as the AQI value. Where $IAQI_p$ is the index for pollutant p; C_p is the concentration of pollutant p; BP_{low} is the breakpoint concentration that is less than or equal to C_p; BP_{high} is the breakpoint concentration that is great than or equal to C_p; $IAQI_{high}$ is defined as the sub index amount relating to BP_{high}; $IAQI_{low}$ is defined as the sub index amount relating to BP_{low}. In this study, the data on AQI are obtained from the web of `www.tianqihoubao.com`, and the data on $PM_{2.5}$ are obtained from 11 monitoring stations established by the Xi'an City Environmental

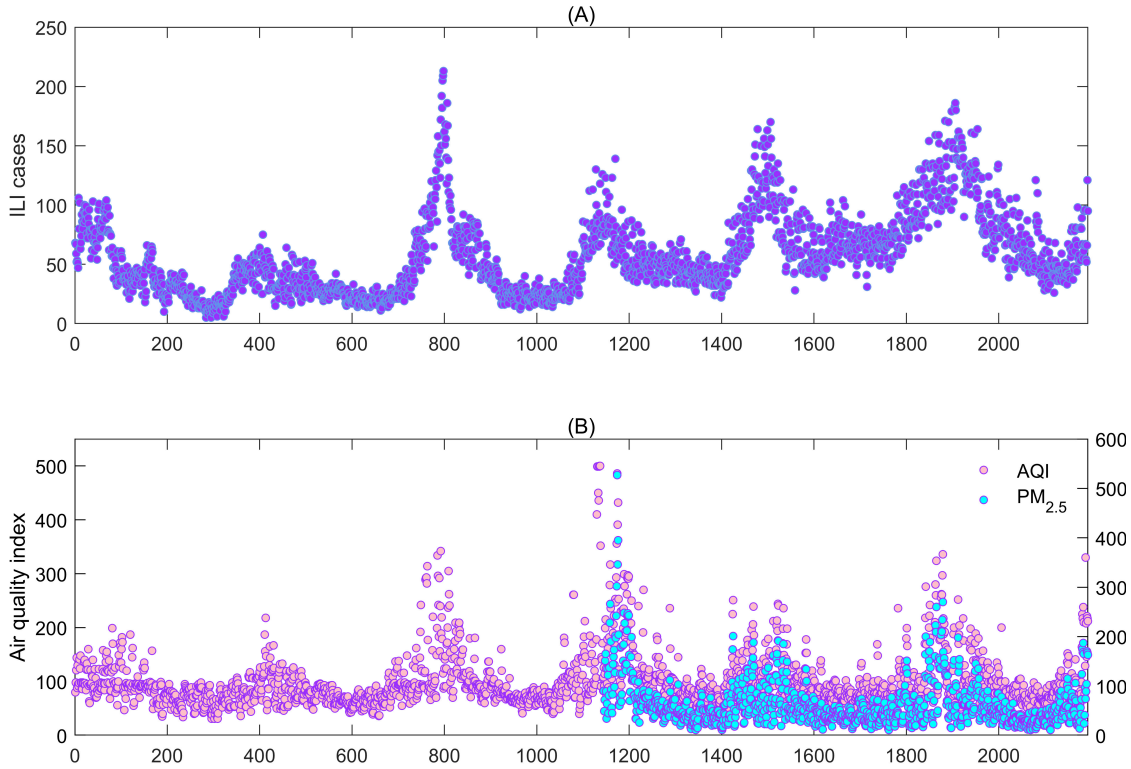

Figure 1: The data. (A) The newly reported ILI cases of sentinel surveillance from seven hospitals of Shaanxi province from November 15, 2010 to November 14, 2016; (B) The AQI for Xi'an, Shaanxi Province from November 15, 2010 to November 14, 2016, represented by pink solid dots; and $PM_{2.5}$ for Xi'an, Shaanxi Province from January 1, 2014 to November 14, 2016, represented by blue solid dots.

Monitoring Station, which are visualized in Fig. 1(B). Note that from November 15, 2010, to November 14, 2016, only partial $PM_{2.5}$ data for Xi'an was available, whereas complete AQI data for this period were accessible. As shown in Fig. 1(B), we observe that the trends of AQI and $PM_{2.5}$ are generally consistent in the available dataset. Therefore, in this study, AQI, as a comprehensive indicator, is employed to investigate the impact of air quality on ILI cases. Moreover, in this study, we divided the data into two parts: the observation values from November 15, 2010 to September 14, 2014 were used as training data for model identification; while those from September 15, 2014 to November 14, 2016 constituted the test data used to assess the predictive performance of the model.

Considering that the collected ILI cases and AQI data contain outliers, missing values, and other quality issues, we performed the following preprocessing steps on the raw data to improve data quality and enhance the efficiency of the machine learning model. Given that the collected sample size was sufficiently large and the proportion of outliers was relatively low, the outliers in the dataset were directly removed. For missing values in the data, the method we adopt is to replace the missing values with the average of the two days before and after the data.

2.2 Respiratory infection dynamics with air pollution

We leverage a multi-scale model that integrates the dynamics of disease transmission and the dynamics of air quality to investigate the relationship between air pollution and respiratory diseases. The corresponding system is governed by the following ordinary differential equations (ODEs) [13,40]:

$$\begin{cases} \dfrac{dS(t)}{dt} = -\lambda(t) + \gamma I(t), \\ \dfrac{dI(t)}{dt} = \lambda(t) - \gamma I(t), \\ \dfrac{dF(t)}{dt} = c\big(S(t), I(t), F(t)\big) - \mu(t)F(t). \end{cases} \tag{2.3}$$

In model (2.3), the first two equations describe the transmission dynamics of respiratory infectious diseases, which are driven by the structurally simple susceptible-infected-susceptible (SIS) compartmental model. This model divides the population into two states: susceptible (S) and infected (I) classes. In the SIS model, susceptible individuals become infected through effective contact with infected individuals. Once infected, individuals recover from the infectious state but do not gain immunity, returning to the susceptible class instead. Indeed, as we study the respiratory diseases, it is clear and appropriate to use a simple SIS model to describe the dynamic process of such epidemics. Here, the incidence rate $\lambda(t)$ given by the following equation:

$$\lambda(t) = \frac{\beta\big(F(t)\big)S(t)I(t)}{N}, \tag{2.4}$$

and constant N represents the total population. γ is the recovery rate of infections. The variable $F(t)$ measures the air quality. Furthermore, based on the evolution trend of the observation data of AQI, we assume the clearance rate of air pollutants $\mu(t)$ as a time-dependent function which can be given as follows [40]:

$$\mu(t) = \mu_0 + (\mu_1 + \mu_2 t)\sin(\omega t + \psi), \tag{2.5}$$

where $\omega = 2\pi/365$ is a periodic parameter assuming that there are 365 days in a year. The offset parameter and phase parameter are denoted by μ_0 and ψ, respectively. The amplitude is controlled by μ_1 and a small amplitude increasing rate μ_2. Previous studies have demonstrated a strong association between air pollution and respiratory infectious diseases [39,44], however, the underlying coupling mechanisms remain poorly understood. Motivated by this empirical evidence, we introduce two abstract functions, the transmission rate $\beta(F(t))$ and the inflow rate of air pollutants $c(S(t),I(t),F(t))$, in model (2.3) to capture the interactive evolution mechanism between disease transmission and air pollution. Notably, in traditional mechanism-based models, the transmission rate is usually assumed to be a specific function of the underlying variables (with parameters to be estimated from the data) to characterize the influence of air pollution on disease, and the

inflow rate of air pollutants is defined as a piecewise constant function to describe the fluctuation of air pollution [14, 40]. However, the impact of human and air pollutants themselves on the intensity of pollution control measures may not be accurately captured through a simple piecewise constant function. Therefore, in this study, we do not assume any specific form of the inflow rate of air pollutants as a function of (S,I,F); and we also do not make any assumption for transmission rate as a function of F in order to use data fitting and machine learning to discover the mechanistic effect of air pollution on disease transmission. We emphasize that without any prior assumption about the transmission rate and inflow rate of air pollutants, discovering their analytical expressions of the underlying variables based on available data to quantify the complex relationship between air pollution and respiratory diseases poses a significant challenge, to be tackled in the next subsection based on deep learning.

2.3 Coupling system of deep neural networks and multi-scale dynamic model

It is widely known that deep neural networks (DNNs) are universal approximators of continuous functions [33]. Therefore, in this study, we use three individual DNNs: one for the surrogate of each state variable (S,I,F) in model (2.3), which takes the time t as input, and two for the approximation of transmission rate and inflow rate of air pollutants, which take the state variables F and (S,I,F) as inputs respectively. In addition to the standard components such as the fully connected layers, our network incorporates two extra layers designed to facilitate the training process, described as follows:

- Input-scaling layer. Since the ILI case and AQI data analyzed in this study have a large temporal domain, the input time t could vary across several orders of magnitude, we first apply the max-min normalization method to scale t, i.e.

$$\tilde{t}=\frac{t-t_{\min}}{t_{\max}-t_{\min}}$$

  such that $\tilde{t}\sim\mathcal{O}(1)$. Here, $t_{\min}$ and $t_{\max}$ denotes the minimum and maximum values of the time t, respectively.

- Output-scaling layer. To enable flexible adjustment of the output range of the neural network, similar to the input scaling layer, we added an another scaling layer to transform the output of the last fully connected layer $\tilde{c},\tilde{\beta}$ to c,β, i.e. $c=c_0+(c_1-c_0)\tilde{c}$, $\beta=\beta_0+(\beta_1-c\beta_0)\tilde{\beta}$. Here, the output-scaling layer can be regarded as a hard constraint to satisfy $c\in[c_0,\,c_1],\beta\in[\beta_0,\,\beta_1]$.

Then, we integrate the multi-scale model coupling the transmission dynamics and the dynamics of air pollution with neural networks through physics-informed deep learning [19].

PINNs, originally proposed by Raissi *et al.* [35], have proven to be powerful tools to solve differential equations as well as infer unknown quantities from data and physical law. The core of this method utilizes DNN as surrogate models for the quantity of

interest, constructs a loss function with respect to the differential equation describing the underlying mechanisms, and optimizing the parameters of DNN to minimize this loss function, thereby addressing the forward and inverse problems of differential equations [27, 35]. In our study, S, I, F, β and c are approximated with DNNs, denoted by $S_\theta, I_\theta, F_\theta, \beta_\varphi$ and c_ϕ, respectively, where (θ,ϕ,φ) is a parameter set composed of network weights and biases. A schematic view of the workflow is illustrated in Fig. 2. Let $\mathcal{T}_d$ denote the set of t on which data of ILI and AQI are available and $\mathcal{T}_e$ denote the set of t on which residuals of the equation are computed. Then, the total loss is defined as a function of both (θ,φ,ϕ) and $(\gamma,\mu_0,\mu_1,\mu_2,\psi)$

$$\mathcal{L}(\theta,\varphi,\phi,\gamma,\mu_0,\mu_1,\mu_2,\psi) = \mathcal{L}^{data}(\theta,\varphi) + \mathcal{L}^{ode}(\theta,\varphi,\phi,\gamma,\mu_0,\mu_1,\mu_2,\psi), \tag{2.6}$$

where

$$\begin{aligned}
&\mathcal{L}^{data}(\theta,\varphi)\\
&= \frac{1}{\mathcal{T}_d}\sum_{t\in\mathcal{T}_d}\left|\frac{\beta_\varphi(F_\theta)S_\theta(t)I_\theta(t)}{N} - I^{data}_{\text{new}}(t)\right|^2 + \frac{1}{\mathcal{T}_d}\sum_{t\in\mathcal{T}_d}\left|F_\theta(t) - F^{data}(t)\right|^2,
\end{aligned} \tag{2.7}$$

$$\begin{aligned}
&\mathcal{L}^{ode}(\theta,\varphi,\phi,\gamma,\mu_0,\mu_1,\mu_2,\psi)\\
&= \frac{1}{\mathcal{T}_e}\sum_{t\in\mathcal{T}_e}\left|\frac{dS_\theta(t)}{dt} + \frac{\beta_\varphi(F_\theta)S_\theta(t)I_\theta(t)}{N} - \gamma I_\theta(t)\right|^2\\
&\quad + \frac{1}{\mathcal{T}_e}\sum_{t\in\mathcal{T}_e}\left|\frac{dI_\theta(t)}{dt} - \frac{\beta_\varphi(F_\theta)S_\theta(t)I_\theta(t)}{N} + \gamma I_\theta(t)\right|^2\\
&\quad + \frac{1}{\mathcal{T}_e}\sum_{t\in\mathcal{T}_e}\left|\frac{dF_\theta(t)}{dt} - c_\phi(S_\theta,I_\theta,F_\theta) + \big(\mu_0 + (\mu_1+\mu_2 t)\sin(\omega t+\psi)\big)F_\theta(t)\right|^2.
\end{aligned} \tag{2.8}$$

Here I^{data}_{new} and F^{data} represent the observational data of ILI cases and AQI, respectively. We see that the loss function (2.6) contains two terms: $\mathcal{L}^{data}$ measures the mismatch between the output of the neural network and the observed data at the training points, while $\mathcal{L}^{ode}$ guides neural networks to adhere to the evolutionary rules of transmission dynamics and AQI dynamics during the learning process. We employ automatic differentiation (AD) to analytically compute the derivative in $\mathcal{L}^{ode}$ (see more details of AD in [3]). Finally, gradient-descent based methods, such as Adam [20], can be employed to minimize the loss function (2.6) to simultaneously infer the neural network parameters as well as the unknown parameters of model (2.3).

To approximate the solution of the multi-scale system, we use a single hidden layer neural network with 128 neurons and tanh activation function to represent the state variables of model (2.3). In addition, the dynamic parameters are parameterized by separate networks of single hidden layers with 64 neurons and tanh activation function. In the training process, we fixed the learning rate at 0.001 and trained the neural network using

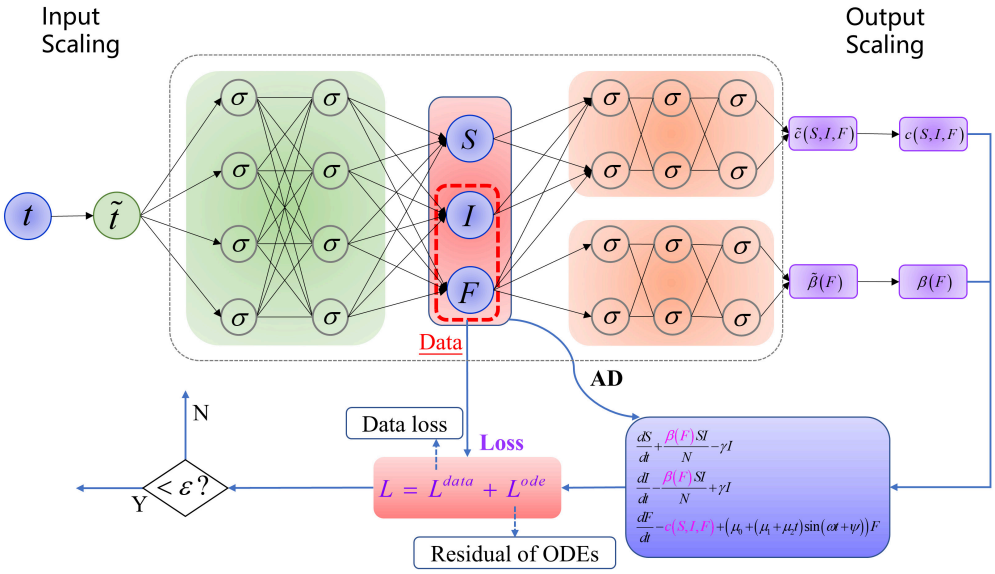

Figure 2: A schematic diagram of physics-informed neural networks, which integrate the dynamics of disease transmission and the dynamics of air quality. Different neural networks are used to represent the state variables (green shaded area) and dynamic parameters (pink shaded area) of model (2.3). The input-scaling layer and output-scaling layer are used to scale the network input and outputs. The symbol "AD" represents the automatic differentiation operator. The box marked "Loss" comprises two parts: the mismatch between available data and the outputs of neural networks, and the residual of ODEs.

the full batch of data. In particular, the algorithm is implemented using the TensorFlow learning framework base on Python [1], which is a widely popular and well-documented open-source software library designed for automatic differentiation and deep learning computations.

2.4 Maximal information coefficient

The maximum information coefficient (MIC) is a statistical method for capturing the correlation between two variables, was first proposed by Reshef *et al.* [38], capable of uncovering hidden linear or nonlinear relationships within the studied targets. The basic idea underlying MIC is that a grid can be drawn on the scatter plot of the two variables to partition and encapsulate this relationship and to get the maximal information entropy simultaneously if certain relationship exists between two variables. The MIC approach is summarized below.

For two discrete variables with n observations, $x=\{x_i|i=1,2,\ldots,n\}$ and $y=\{y_i|i=1,2,\ldots,n\}$, a finite set $D=\{(x_i,y_i)|i=1,2,\ldots,n\}$ of ordered pairs can be obtained. Then, the MIC is computed as follows [38,47]:

$$\mathrm{MIC}(D)=\max_{xy<B(n)}\{M(D)_{x,y}\},$$

where $B(n)$ represents the upper limit of xy grids, and $M(D)$ is the characteristic matrix of D, which defined by the following formula:

$$M(D)_{x,y}=\frac{I^*(D,x,y)}{\log\min\{x,y\}},$$

where I^* denotes the mutual information of the two variables in D. The MIC values are limited to the range $[0,1]$. If two target variables exhibit a correlated relationship, the MIC value is close to 1, otherwise it tends towards 0. In this study, we leveraged a natural measure of linearity and nonlinearity based on MIC, that is, if $\text{MIC}-\rho^2$ is close to zero, then represents a linear relationship and $\text{MIC}-\rho^2>0.2$ for nonlinear relationships, where ρ denotes the Pearson product-moment correlation coefficient, and the complete mathematical formulation of the $\text{MIC}-\rho^2$ metric is presented as follows [23,38]:

$$\text{MIC}-\rho^2=\max_{xy<B(n)}\left\{\frac{I^*(D,x,y)}{\log\min\{x,y\}}\right\}-\left[\frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x\sigma_y}\right]^2,$$

where σ_x and σ_y are the standard deviations of x and y respectively, and μ_x and μ_y correspond to the means of x and y respectively. In our case, due to the incomplete clarity regarding the interrelationships among the output results of PINNs, the MIC is utilized as an assessment metric to analyze the potential correlations among the output results of PINNs. This will lay the foundation for the determination of analytical expressions corresponding to various rate functions in subsequent steps.

2.5 Symbolic regression

In order to enhance the transparency of inference results in neural networks characterized by black-box attributes, symbolic regression emerges as a favorable option. Symbolic regression, a powerful methodology within the field of machine learning, is designed to discover a mathematical expression or equation that provides the optimal fit for a provided dataset. Diverging from traditional regression techniques (such as linear regression or polynomial regression), symbolic regression endeavors to reveal the intrinsic mathematical relationship between input variables and the target variable without making assumptions about the form of the equation in advance. Instead, initial expressions are formed by randomly combining mathematical building blocks such as simple mathematical operators $(+,-,\times,\div)$, unary analytic functions $(\sin,\cos,\exp,\ldots)$, constants, and state variables. Subsequently, for a given data set, symbolic regression is used to reveal the hidden functional relationships. Through the operation of evolutionary algorithms and symbolic structures, the generated mathematical expressions are continuously optimized to accurately capture patterns, dependencies, and potential regularities within the data.

For symbolic regression, we leverage the open-source software PySR [7], which has a configurable Python interface constructed upon the efficient Julia backend `SymbolicRegression.jl`. The core algorithm of PySR integrates tree search and regularized evolution. In our study, we use PySR to distill knowledge from the time series

of β_φ and c_ϕ inferred by PINNs to obtain interpretable symbolic expressions. This will lead to further understanding of the interrelationship between air pollution and respiratory diseases.

We summarize the execution process of this method in a flowchart, as shown in Fig. 3. First, we develop a mathematical model that couples respiratory infection transmission and air pollution evolution dynamics. Then, we replace the transmission rate and air pollutants inflow rate with neural networks, resulting in a mathematical model that integrates ODEs and neural networks. Next, we utilize the PINNs framework to infer the transmission and inflow rates based on ILI cases and AQI data. After that, we convert the transmission rate and inflow rate represented by neural networks into analytical forms through correlation analysis and symbolic regression, obtaining a partially learned mechanistic model. Finally, we use the partially learned mechanistic model to predict the development trends of ILI cases and AQI.

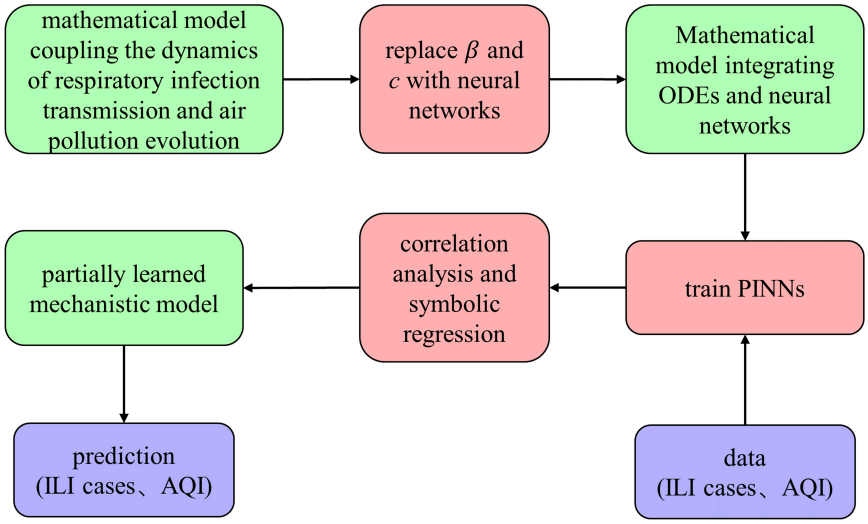

Figure 3: Flowchart of the method. The green cells represent the model at three different stages: the original mathematical model coupling the dynamics of respiratory infection transmission and air pollution evolution, the mathematical model integrating ODEs and neural networks and the partially learned mechanistic model. The red cells represent the steps of the method: embedding a neural network within the model, training PINNs and implementing correlation analysis and symbolic regression.

3 Results

3.1 Data fitting

In this part, based on the data related to the AQI and ILI case numbers during November 15, 2010 to September 14, 2014, we inferred the parameters associated with air pollution and epidemiology of respiratory infection in the multi-scale model (2.3) through

training PINNs. We present the PINNs results on data fitting for AQI and ILI case numbers in Fig. 4 (solid curves), and visualize the inferred time-dependent transmission rate and inflow rate of air pollutants in Fig. 4 (dashed curves). In addition, the estimation results of other parameters are listed in Table 1.

Table 1: Parameter definitions and estimation for model (2.3).

Paras	Definitions	Estimated values	Sources
β	Transmission rate	See text and Fig. 4(A)	Estimated
c	Clearance rate of air pollutants	See text and Fig. 4(B)	Estimated
γ	Recovery rate of infected individuals	0.180	Estimated
μ_0	Parameter in the clearance rate of pollutants	0.151	Estimated
μ_1	Parameter in the clearance rate of pollutants	0.011	Estimated
μ_2	Parameter in the clearance rate of pollutants	3.733×10^{-5}	Estimated
ψ	Parameter in the clearance rate of pollutants	3.799	Estimated

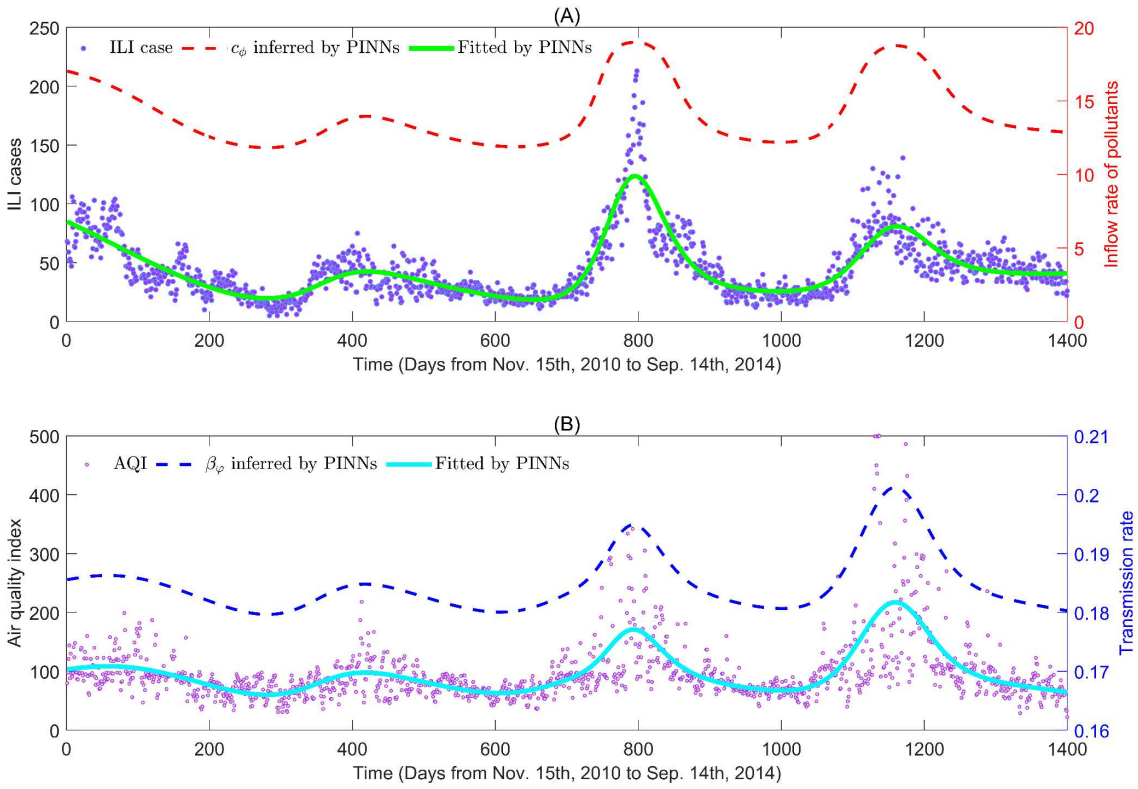

Figure 4: Data fitting and inference of the dynamic parameters by PINNs algorithm. The scatters represent the available observations of ILI cases and AQI, respectively. The solid curves represent the best fitting results of ILI cases and AQI by the PINNs algorithm. The estimated transmission rate and inflow rate of air pollutants are marked by dashed curves.

As we can observe from Fig. 4, the PINNs algorithm can not only fit ILI case numbers very well, but also accurately simulate the development trend of the AQI. Notably, even without any prior knowledge of the mathematical forms of disease transmission rate and inflow rate of air pollutants, the proposed method can autonomously extract the temporal variations of these rate functions from the available observational data. It can be observed that the inferred transmission rates of diseases and inflow rate of air pollutants exhibit highly intricate evolutionary patterns. Comparing the estimated time-dependent inflow rate and the time series of ILI cases, it can be observed from Fig. 4(A) that the peaks in the inflow rate of air pollutants always correspond to the peaks of ILI cases, which suggests that ILI case gradually increases during the ascending phase of the inflow rate of air pollutants. Furthermore, it can be observed that the temporal evolution trends of the transmission rate and AQI exhibit a high degree of consistency by combining the inferred transmission rates and the time series of AQI from Fig. 4(B), which indicates that the changes of AQI may, to some extent, influence the transmission rate of diseases.

In summary, we observe that, on one hand, the rapid descending of the inflow rate of air pollutants leads to the decreasing of the concentration of air pollutants, which lowers the AQI and results in the decreasing of the transmission rate, thereby leading to a reduction in ILI cases. On the other hand, we observe that the human environmental protection awareness may decrease as ILI cases decline to a lower level, causing a gradual rise in the inflow rate of air pollutants. This, consequently, leads to the increasing of the concentrations air pollutant, resulting in an elevated AQI and an upturn in the transmission rate, ultimately inducing the resurgence of ILI cases. This establishes a feedback loop between the levels of air pollution and ILI, which shows that the recurrent fluctuations in AQI serve as driving factors for the surges and subsequent resurgences of ILI cases.

3.2 Correlation analysis

It is worth noting that the transmission rate and inflow rate of air pollutants inferred by PINNs are merely two time series without particular forms, making it challenging to clearly elucidate the mechanism of interaction between air pollution and the spread of respiratory diseases. To further reveal the relationship between air pollution and respiratory diseases, it becomes imperative to formulate the appropriate functions for the transmission rate and inflow rate of air pollutants. Based on the high consistency in the change trends between the inflow rate of air pollutants and ILI cases in Fig. 4(A) and the transmission rate and AQI in Fig. 4(B), we infer that there may be some relationship between transmission rate and AQI, as well as between inflow rate of air pollutants and incidence rate of ILI case. However, the exact nature of these relationships remain incompletely understood. Therefore, in our study, the maximal information coefficient (MIC) is employed to analyze the correlation between these variables. We computed the MIC values corresponding to the transmission rate β_φ, the inflow rate of air pollutants c_ϕ, the simulated value ($F(t)$) of AQI, and the incidence rate $\lambda(t)$ resulting from PINNs algorithm and presented in Fig. 5(A). The calculated MIC values reveal a significant correla-

tion between the transmission rate and AQI, as well as a notable association between the inflow rate of air pollutants and the incidence rate. According to the linearity and nonlinearity analysis, the transmission rate has a significantly linear relationship with AQI (MIC$-\rho^2$ is near 0, pvalue less than 0.01), while the incidence rate has also significant linear relationships with the inflow rate of air pollutants (MIC$-\rho^2$ is near 0, pvalue less than 0.01).

3.3 Regression analysis

Based on the correlation analysis in the previous section, we have identified the interrelationships between the transmission rate and inflow rate of air pollutants and the state variables in the multiscale model (2.3), respectively. Next, we utilized the symbolic regression package PySR to discover closed-form expressions for these two functions β_φ and c_ϕ. To achieve this, we use the simulated values ($F(t)$) of AQI as the input variable for describing β_φ, and the incidence rate ($\lambda(t)$) as the input variable for characterizing c_ϕ, where $F(t)$ and $\lambda(t)$ are all inferred by PINNs. The binary operators are chosen to addition, subtraction, multiplication and division, the unary operators are exponential, sin,cos and reciprocal, and the complexities of unary operators are set to 3,3,3,3, respectively.

In the initial stage of the symbolic regression simulation, the algorithm randomly generates a set of expression binary trees, serving as candidate functions, based on defined input variables and operators. For each candidate expression, the fitness is evaluated by comparing the predicted outputs with the target values. Fitness measures how well the candidate function approximates the data, and here the mean square error is chosen as the fitness function. In our experiments, β_φ and c_ϕ experience 100 iterations of simulation to select the best candidate functions $\beta_{sym}=0.00014F(t)+0.17141$ and $c_{sym}=0.0904\lambda(t)+9.9753$, where $\lambda(t)=\beta_{sym}S(t)I(t)/N$. The expression binary tree corresponding to the best candidate function is shown in Figs. 5(B) and 5(C), where the leaves of the tree represent input variables or constants, and the internal nodes represent mathematical operations or functions. Furthermore, we present the fitting results of symbolic regression in Figs. 5(D) and 5(E), which indicate that the best candidate functions determined by symbolic regression can accurately simulate the temporal evolution pattern of the transmission rate and inflow rate of air pollutants inferred by PINNs. In addition, the selected best candidate functions suggest a linear positive correlation between the transmission rate and the variable $F(t)$, while the inflow rate of air pollutants exhibits a linear positive correlation with incidence rate $\lambda(t)$, further validating the conclusion of the correlation analysis. It also suggests that air pollution affects the evolution trend of respiratory diseases by influencing the transmission rate of the disease.

According to the above analysis, we utilized symbolic regression to get the best functions of transmission rate and inflow rate of air pollutants from the time series inferred by the PINNs algorithm, which enabled us to accurately quantify the complex relationship between respiratory disease and air pollution. Further, it is worth noting all functions

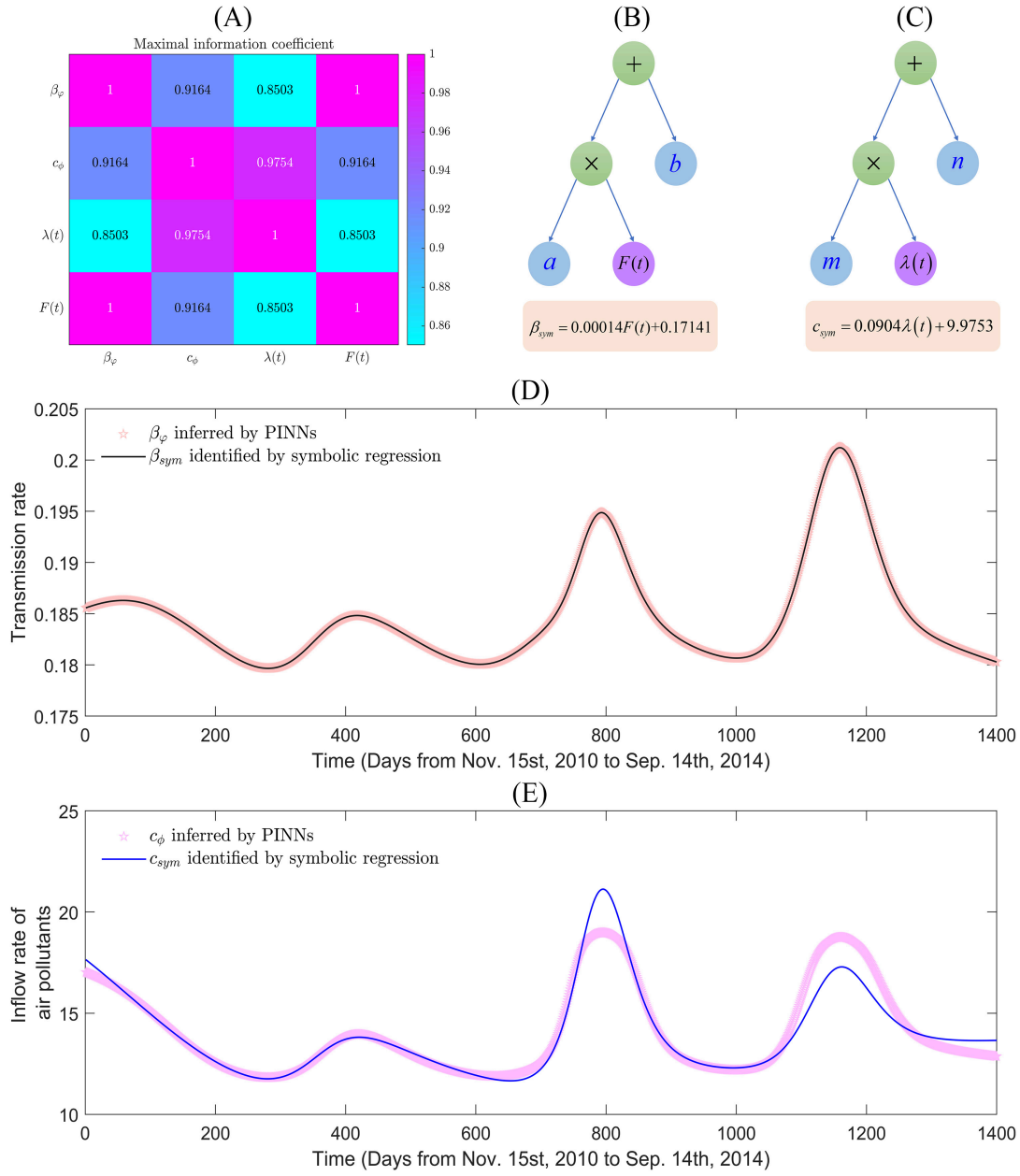

Figure 5: (A) MIC values among the transmission rate, the inflow rate of air pollutants, the simulated value of AQI and the incidence rate learned by PINNs. (B) The expression tree with depth 2 and size (total number of nodes) 5 represents the best analytical expressions of β_φ inferred by PINNs. (C) The expression tree with depth 2 and size (total number of nodes) 5 represents the best analytical expressions of c_ϕ inferred by PINNs. (D) and (E) The inference and fitting results of the transmission rate and inflow rate of air pollutants, where the pentagrams represent the neural network inference results, β_φ and c_ϕ, and the solid curves represent the fitting results, β_{sym} and c_{sym}, based on symbolic regression.

identified by symbolic regression have intuitive realistic meanings, thereby enhancing the interpretability of the inference results by deep learning. Therefore, our approach can aid in improving understanding how air pollution influences the spread of respiratory disease.

3.4 Validation and prediction

To validate the inference results of PINNs algorithm and symbolic regression, we firstly substitute the time-independent parameters $(\gamma, \mu_0, \mu_1, \mu_2, \psi)$ estimated by PINNs and the unknown rate functions (β_{sym} and c_{sym}) learned from symbolic regression back into the multi-scale model (2.3), which can be yielded a fully mechanistic model (FMM). Then, We utilize the built-in Python method `scipy.integrate.odeint` to solve this model within the time interval from November 15, 2010 to September 14, 2014. In the light pink regions of Figs. 6(A) and 6(E), we present the solutions of the fully mechanistic model corresponding to ILI cases and AQI, respectively, where we can observe that the simulated curves can match the data well. This indicates that combining the training results of the PINNs algorithm and the estimation results of symbolic regression can effectively assist the multi-scale model (2.3) to simulate the development process of disease and air quality and reveal their underlying relationship.

Subsequently, based on the inference results from PINNs and symbolic regression, we extrapolated the fully mechanistic model from September 16, 2014 to November 14, 2016 to examine the predictability of our model. The predicted curves of model (2.3)

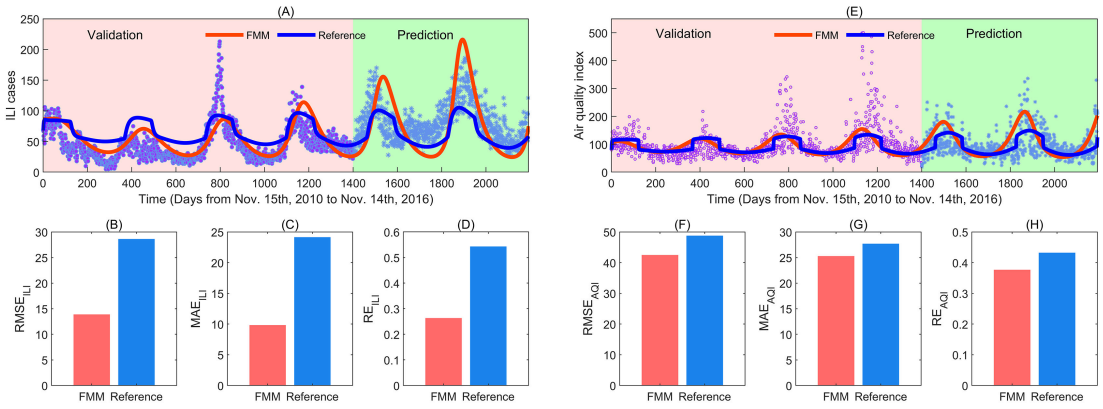

Figure 6: Comparison of the performance between the fully mechanistic model determined via PINNs and symbolic regression and the model established in the reference [40] (referred to as the reference model). (A) and (E) The fitting and predictive results of ILI cases and AQI by the FMM and the reference model, where the red curve represents the simulated results of the FMM and the blue curve represents the simulated results of the reference model. The solid dots and blue stars are the data for training and prediction, respectively. (B)-(D) The simulation performance metrics for ILI cases by FMM and the reference model. (F)-(G) The simulation performance metrics for AQI by FMM and the reference model.

corresponding to ILI cases and AQI are respectively depicted in the light green regions in Figs. 6(A) and 6(E). By comparing with the data that did not participate in the training during the inference stage of the PINNs algorithm, we find that the multi-scale model (2.3) calibrated by PINNs and symbolic regression can accurately capture the future evolution trends of ILI cases and AQI.

To further evaluate the performance of the fully mechanistic model reconstructed using PINNs and symbolic regression, we compared it with the model reported in the reference [40], and use three classic metrics – root mean square error (RMSE), mean absolute error (MAE), and relative error (RE) – to quantify the model's performance. These metrics can be calculated according to the following equations:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2},\quad \mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|,\quad \mathrm{RE}=\frac{\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}\hat{y}_i^2}},$$

where n is the total number of samples, with y_i and $\hat{y}_i$ denoting the observed and predicted values of the i-th sample, respectively. As shown in Figs. 6(B)-6(D) and 6(F)-6(H), the FMM reconstructed using PINNs and symbolic regression exhibits lower RMSE, MAE, and RE. This result indicates that compared to the model in the reference [40], the FMM derived in this study has better performance.

In light of the above simulation analysis results, we find that utilizing symbolic regression to convert the transmission rate and inflow rate of air pollutants represented by neural networks into analytical mathematical expressions, which not only assigns clear practical meanings to these rate functions, but also makes the learning results of DNNs more interpretable. Furthermore, by incorporating the rate functions determined through symbolic regression into model (2.3), a multiscale model of bidirectional coupling between air pollution and disease transmission can be derived. This enables us to gain a more comprehensive understanding and prediction of the dynamic processes of air pollution and disease transmission, potentially informing the development of effective public health policies and implementation of air pollution control strategies.

4 Discussion and conclusion

It is known that air pollution is one of the greatest environmental risks to public health, particularly, air pollution plays a crucial role in respiratory disease transmission and pathology. However, accurately quantifying the impact of air pollution on respiratory infection risk in the population remains challenging. In this study, to investigate the complicated relationship between respiratory disease and air pollution, we integrate neural networks and the multi-scale model coupling the transmission dynamics of respiratory disease and AQI evolution dynamics based on the principle of PINNs. In particular, we employed three independent DNNs: one to create a surrogate model for the multi-scale model, and the other two to respectively represent the time-varying transmission

rate and inflow rate of air pollutants. With the developed algorithm, we simulated the evolution trend of ILI cases and AQI data in Xi'an during 2010-2016, by simultaneously inferring the unknown constant parameters and dynamics parameters.

It is noteworthy that traditional mechanistic models typically assume various rate functions to characterize the relationship between air pollution and respiratory diseases [13,40]. This traditional approach needs to postulate complicated rate functions in order to make the involved mechanism models more consistent with the actual situation. This means that a large number of parameters must be introduced into the mechanism models, making the computation both cumbersome and dependent of the assumed form of the rate functions. In contrast, the PINNs algorithm enable us to directly capture the temporal evolution patterns of these rate functions purely from available data, eliminating the need to specify the exact mathematical forms of rate functions beforehand. This flexibility means we make fewer assumptions about the mechanistic models, offering a more efficient simulation and understanding of the intricate relationship between air pollution and respiratory diseases.

Since the information on the dynamics of disease transmission and air quality embedded in the neural network, our algorithm framework can simultaneously fit both ILI case numbers and AQI data well despite the fact that our baseline multi-scale model structure (2.3) is relatively simple and is purely guided by the underlying epidemiological process (SIR-compartmental model). Importantly, our data-driven integrative framework can extract the trends in transmission rate of disease and inflow rate of air pollutants from observational data (see Fig. 4 for detail), with high consistency in the dynamics of AQI and the transmission rate inferred by PINNs, as well as in the development trend of the inflow rate of air pollutants and ILI cases. Our results show how air quality affects the progression of the disease by altering the transmission rate. Recall that the transmission rate and inflow rate of air pollutants learning from PINNs inference are abstract time series, which alone makes it difficult for us to completely understand the relationship between air pollution and respiratory diseases. To resolve this issue, we analyzed the interrelationships among the output results of PINNs based on the maximum information coefficient (MIC) after simulating the training data using PINNs; and then we employed symbolic regression to discover analytical expressions that can accurately quantify the impact mechanism of air pollution on respiratory diseases. The study results reveal a linear positive correlation between the transmission rate of the disease and air quality, and the inflow rate of air pollutants exhibits linear positive correlation with the incidence rate. This implies that the transmission rate of respiratory disease increased progressively with increasing AQI, in line with observations of the existing study [31].

A key highlight of this study lies in the use of symbolic regression to regress the transmission rate and inflow rate of air pollutants inferred by PINNs back into symbolic forms, resulting in two interpretability functions ($\beta(F)$ and $c(S,I,F)$), revealing the following important practical implications: on one hand, the more severe the air pollution index, the higher the infection rate, exhibiting a positive linear relationship (i.e. $\beta(F(t))=0.00014F(t)+0.17141$). The data depicted in Fig. 4 further substantiate this find-

ing, demonstrating that severe pollution induces an increase in ILI case numbers. On the other hand, the high incidence rate $\lambda(t)$ of ILI implies a higher inflow rate of air pollutants $(c(S(t),I(t),F(t))=0.0904\lambda(t)+9.9753, \lambda(t)=\beta(F(t))S(t)I(t)/N)$, which shows the positive feedback cycle between air pollution and disease incidence rate.

This study introduced a novel hybrid algorithmic framework that integrates artificial intelligence, correlation analysis based on MIC, and symbolic regression. The workflow of this framework essentially involves three sequential steps. Firstly, the physics-informed deep learning framework, serving as a surrogate model, integrate the multi-scale model, neural networks and observational data. Subsequently, MIC is employed to identify various potential correlations among the outputs results of PINNs. Finally, symbolic regression is utilized to discover the exact expressions for unknown mechanisms within the multi-scale dynamical model. Particularly, the multi-scale model embedded in the neural network not only retains the interpretable characteristics of the dynamical model but also integrates the data learning capabilities of the neural network. Furthermore, our research findings indicate that a combination of deep learning, correlation analysis, and symbolic regression with available ILI case and AQI data can provide insights into the dynamic impact of air quality on respiratory infection. Although the data used in this study are specific to a particular region, the methodology can be applied to other regions as long as relevant data is available.

Our study also has some limitations. First, the framework we proposed can only infer the correlation between air pollution and respiratory infections, which may overlook the impact of unobserved confounders, as well as the observed confounders such as temperature and relative humidity on the relationship between air pollution and public health. To further elucidate the effects of these confounders, it is necessary to examine the causal relationships among air pollutants, climate factors, and the daily reported ILI cases. Therefore, future research could conduct more in-depth studies using causal analysis methods such as cross-mapping techniques [48] and intervened reservoir computing [50]. Second, we use the simplest SIS compartment model to describe the evolution of respiratory infectious diseases, which may ignore the important factors such as behavioral responses to epidemics (e.g. health-seeking), vaccination, and seasonal environmental changes on the development of ILI cases. Therefore, developing high-dimensional, multi-scale models that incorporate these factors will be a key challenge for our future work. Despite these limitations, the framework we developed provides a research paradigm for leveraging deep learning to discover the complex coupling between air pollution and respiratory diseases. We hope that this study serves as a starting point for more comprehensive analyses and that the approach combining deep learning with disease transmission dynamics and the evolution of air pollution can be more broadly applied.

Acknowledgments

S. Tang was funded by the National Key R & D Program of China (No. 2025YFA1016502) and by the National Natural Science Foundation of China (No. 92470107). J. Wu's re-

search has been supported by the York Research Chair Program, and by the Mathematics for Public Health Initiative. S. He was funded by the National Natural Science Foundation of China (No. 12571539).

M. He and D. Yan contributed equally to this work.

References

[1] M. Abadi et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, arXiv:1603.04467, 2016.

[2] L. N. Araujo, J. T. Belotti, T. A. Alves, Y. de Souza Tadano, and H. Siqueira, *Ensemble method based on Artificial Neural Networks to estimate air pollution health risks*, Environ. Model. Softw., 123:104567, 2020.

[3] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, *Automatic differentiation in machine learning: A survey*, J. Mach. Learn. Res., 18(153):1–43, 2018.

[4] A. J. Burbank and D. B. Peden, *Assessing the impact of air pollution on childhood asthma morbidity: How, when, and what to do*, Curr. Opin. Allergy Clin. Immunol., 18(2):124–131, 2018.

[5] Y. Cai, S. Zhao, Y. Niu, Z. Peng, K. Wang, D. He, and W. Wang, *Modelling the effects of the contaminated environments on tuberculosis in Jiangsu, China*, J. Theor. Biol., 508:110453, 2021.

[6] J. Cheng, F. Li, L. Liu, H. Jiao, and L. Cui, *Spatiotemporal variation air quality index characteristics in China's major cities during 2014-2020*, Water Air Soil Pollut., 234(5):292, 2023.

[7] M. Cranmer, *Interpretable machine learning for science with PySR and SymbolicRegression.jl*, arXiv:2305.01582, 2023.

[8] A. Dairi, F. Harrou, S. Khadraoui, and Y. Sun, *Integrated multiple directed attention-based deep learning for improved air pollution forecasting*, IEEE Trans. Instrum. Meas., 70:3520815, 2021.

[9] J. L. Domingo and J. Rovira, *Effects of air pollutants on the transmission and severity of respiratory viral infections*, Environ. Res., 187:109650, 2020.

[10] I. Galan, A. Tobias, J. R. Banegas, and E. Aranguez, *Short-term effects of air pollution on daily asthma emergency room admissions*, Eur. Respir. J., 22(5):802–808, 2003.

[11] E. Garshick, *Effects of short- and long-term exposures to ambient air pollution on COPD*, Eur. Respir. J., 44(3):558–561, 2014.

[12] W. J. Guan, X. Y. Zheng, K. F. Chung, and N. S. Zhong, *Impact of air pollution on the burden of chronic respiratory diseases in China: Time for urgent action*, Lancet, 388(10054):1939–1951, 2016.

[13] S. He, S. Y. Tang, Y. L. Cai, W. M. Wang, and L. B. Rong, *A stochastic epidemic model coupled with seasonal air pollution: Analysis and data fitting*, Stoch. Environ. Res. Risk Assess., 34:2245–2257, 2020.

[14] S. He, S. Y. Tang, Y. N. Xiao, and R. A. Cheke, *Stochastic modelling of air pollution impacts on respiratory infection risk*, Bull. Math. Biol., 80:3127–3153, 2018.

[15] S. He, S. Y. Tang, Q. M. Zhang, L. B. Rong, and R. A. Cheke, *Modelling optimal control of air pollution to reduce respiratory diseases*, Appl. Math. Comput., 458:128223, 2023.

[16] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak, *Global trends in emerging infectious diseases*, Nature, 451(7181):990–993, 2008.

[17] M. Kampa and E. Castanas, *Human health effects of air pollution*, Environ. Pollut., 151(2):362–367, 2008.

[18] H. D. Kan, R. J. Chen, and S. L. Tong, *Ambient air pollution, climate change, and population health in China*, Environ. Int., 42:10–19, 2012.

[19] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, *Physics-informed machine learning*, Nat. Rev. Phys., 3(6):422–440, 2021.

[20] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv:1412.6980, 2014.
[21] E. Kristiani, Y.-A. Chen, C.-T. Yang, C.-Y. Huang, Y.-T. Tsan, and W.-C. Chan, *Using deep ensemble for influenza-like illness consultation rate prediction*, Future Gener. Comput. Syst., 117:369–386, 2021.
[22] H. Li, S. You, H. Zhang, W. Zheng, X. Zheng, J. Jia, T. Ye, and L. Zou, *Modelling of AQI related to building space heating energy demand based on big data analytics*, Appl. Energy, 203:57–71, 2017.
[23] W. Li, H. Fang, G. Qin, X. Tan, Z. Huang, F. Zeng, H. Du, and S. Li, *Concentration estimation of dissolved oxygen in Pearl River Basin using input variable selection and machine learning techniques*, Sci. Total Environ., 731:139099, 2020.
[24] C. J. Liang, P. Y. Lin, Y. C. Chen, and J. J. Liang, *Effects of regional air pollutants on respiratory diseases in the basin metropolitan area of central Taiwan*, Sustain. Environ. Res., 33(1):1, 2023.
[25] W. M. Liang, H. Y. Wei, and H. W. Kuo, *Association between daily mortality from respiratory and cardiovascular diseases and air pollution in Taiwan*, Environ. Res., 109(1):51–58, 2009.
[26] C. Y. Lin, Y. S. Chang, and S. Abimannan, *Ensemble multifeatured deep learning models for air quality forecasting*, Atmos. Pollut. Res., 12(5):101045, 2021.
[27] L. Lu, X. H. Meng, Z. P. Mao, and G. E. Karniadakis, *DeepXDE: A deep learning library for solving differential equations*, SIAM Rev., 63(1):208–228, 2021.
[28] J. Ma, J. C. P. Cheng, C. Lin, Y. Tan, and J. Zhang, *Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques*, Atmos. Environ., 214:116885, 2019.
[29] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, *Environmental and health impacts of air pollution: A review*, Front. Public Health, 8:505570, 2020.
[30] A. I. Middya and S. Roy, *Pollutant specific optimal deep learning and statistical model building for air quality forecasting*, Environ. Pollut., 301:118972, 2022.
[31] Q. C. Pan, Z. H. Tang, Y. S. Yu, M. Xi, and G. Q. Zang, *Haze and influenza A virus: Coincidence or causation?*, Am. J. Infect. Control, 44(8):959–960, 2016.
[32] J. Park, H.-J. Kim, C.-H. Lee, C. H. Lee, and H. W. Lee, *Impact of long-term exposure to ambient air pollution on the incidence of chronic obstructive pulmonary disease: A systematic review and meta-analysis*, Environ. Res., 194:110703, 2021.
[33] A. Pinkus, *Approximation theory of the MLP model in neural networks*, Acta Numer., 8:143–195, 1999.
[34] C. A. Pope III, *Epidemiology of fine particulate air pollution and human health: Biologic mechanisms and who's at risk?*, Environ. Health Perspect., 108:713–723, 2000.
[35] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, J. Comput. Phys., 378:686–707, 2019.
[36] K. Ravindra, S. S. Bahadur, V. Katoch, S. Bhardwaj, M. Kaur-Sidhu, M. Gupta, and S. Mor, *Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections*, Sci. Total Environ., 858:159509, 2023.
[37] K. Ravindra, N. Chanana, and S. Mor, *Exposure to air pollutants and risk of congenital anomalies: A systematic review and metaanalysis*, Sci. Total Environ., 765:142772, 2021.
[38] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, *Detecting novel associations in large data sets*, Science, 334(6062):1518–1524, 2011.
[39] W. Su, X. Wu, X. Geng, X. Zhao, Q. Liu, and T. Liu, *The short-term effects of air pollutants on influenza-like illness in Jinan, China*, BMC Public Health, 19(1):1319, 2019.

[40] S. Tang, Q. Yan, W. Shi, X. Wang, X. Sun, P. Yu, J. Wu, and Y. Xiao, *Measuring the impact of air pollution on respiratory infection risk in China*, Environ. Pollut., 232:477–486, 2018.
[41] H. M. Tran et al., *The impact of air pollution on respiratory diseases in an era of climate change: A review of the current evidence*, Sci. Total Environ., 898:166340, 2023.
[42] T. D. Tuong, *Epidemic SIS model in air-polluted environment*, J. Appl. Math. Comput., 64(1):53–69, 2020.
[43] X. M. Wang and Y. L. Cai, *The influence of ambient air pollution on the transmission of tuberculosis in Jiangsu, China*, Infect. Dis. Model., 8(2):390–402, 2023.
[44] C. M. Wong et al., *Modification by influenza on health effects of air pollution in Hong Kong*, Environ. Health Perspect., 117(2):248, 2008.
[45] Y. J. Xia et al., *Associations of outdoor fine particulate air pollution and cardiovascular disease: Results from the prospective urban and rural epidemiology study in China (PURE-China)*, Environ. Int., 174:107829, 2023.
[46] P. Xu, Y. F. Chen, and X. J. Ye, *Haze, air pollution, and health in China*, Lancet, 382(9910):2067, 2013.
[47] D. Yang and H. M. Liu, *Maximal information coefficient applied to differentially expressed genes identification: A feasibility study*, Technol. Health Care, 27:249–262, 2019.
[48] X. Ying, S. Y. Leng, H. F. Ma, Q. Nie, Y. C. Lai, and W. Lin, *Continuity scaling: A rigorous framework for detecting and quantifying causality accurately*, Research (Wash D C), 2022:9870149, 2022.
[49] Z. B. Zhang, T. Xue, and X. Y. Jin, *Effects of meteorological conditions and air pollution on COVID-19 transmission: Evidence from 219 Chinese cities*, Sci. Total Environ., 741:140244, 2020.
[50] J. T. Zhao, Z. X. Gan, R. X. Huang, C. Guan, J. F. Shi, and S. Y. Leng, *Detecting dynamical causality via intervened reservoir computing*, Commun. Phys., 7(1):232, 2024.
[51] Q. Zhou, X. N. Li, J. Hu, and Q. M. Zhang, *Dynamics and optimal control for a spatial heterogeneity model describing respiratory infectious diseases affected by air pollution*, Math. Comput. Simulation, 220:276–295, 2024.