IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 13, 2020, accepted May 5, 2020, date of publication May 11, 2020, date of current version May 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993610

Enhancing the Data Learning With Physical Knowledge in Fine-Grained Air Pollution Inference

RUI MA[1], NING LIU[1], XIANGXIANG XU[1], (Student Member, IEEE), YUE WANG[1], HAE YOUNG NOH[2], PEI ZHANG[3], AND LIN ZHANG[4]

[1]Department of Electronic Engineering, Tsinghua University, Beijing 10084, China
[2]Department of Civil and Environmental Engineering, Stanford University, Stanford, CA 94305, USA
[3]Department of Electrical and Computer Engineering, Carnegie Mellon University, Moffett Field, CA 94035, USA
[4]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China

Corresponding author: Rui Ma (mr15@mails.tsinghua.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC0212100, and in part by the Shenzhen Science and Technology Program under Grant KQTD20170810150821146.

ABSTRACT Fine-grained air pollution monitoring has attracted increasing attention worldwide. Even with an increasing amount of both static and mobile sensing systems, an inference algorithm is still essential to achieve a comprehensive understanding of the urban atmospheric environment. Conventional physical model-based methods are unable to involve all the influencing factors with limited prior knowledge, and data-driven methods lacking physical interpretation may result in bad generalization ability. This paper presents a multi-task learning scheme, which combines the physical model and the data-driven model with both merits. It enhances the data learning of a neural network with the aid of prior knowledge on atmospheric dispersion, and also controls the impact of the knowledge with a tunable weighting coefficient. Evaluations over a real-world deployment in Foshan, China show that, with the resolution of 500m×500m×15min, the proposed method outperforms the state-of-the-art ones with 7.9% error reduction and 6.2% correlation increase. Benefited from the physical knowledge, the neural network obtains stable performance with lower variance, as well as higher robustness against negative background conditions.

INDEX TERMS Air pollution inference, data-driven method, multitask learning, physical model.

I. INTRODUCTION

Air pollution has been among the top global risks. According to the World Health Organization (WHO), air pollution causes cardiovascular and respiratory diseases, leading to about 7 million deaths a year [1]. In order to prevent people from its damage, air quality monitoring stations have been deployed for routine environmental monitoring in many countries. However, human influences on the environment continue to grow and the resulting risks are continuously generating diseases and injuries [1]. Existing static official stations are only capable of obtaining the background pollution level, but fail in capturing the dynamic street-level pollution patterns resulted from human's urban activities, which brings difficulties in effective environment governing and policy making [2].

In recent years, sensor networks that consist of calibrated low-cost gas sensors are proposed to achieve fine-grained air pollution monitoring by a series of previous work [3]–[7]. These sensors can be deployed more densely in the objective area due to their lower unit price and compact size. Some researchers equip air quality sensors on vehicles that can freely move in the city, therefore these systems obtain greater granularity and higher possibility to capture detailed pollution variations. However, even with more sensing nodes, the observations from these systems are not able to cover the entire spatial-temporal space. An inference algorithm that recovers the entire pollution map from partial observations is still essential to achieve a comprehensive understanding of the urban atmospheric environment.

Previous algorithms for air pollution inference are mainly based on the physical model or data-driven: a) Physical model-based methods are conventional in civil and environmental engineering, of which the most typical one is called pollutant dispersion model [8]–[11]. Several inference algorithms are proposed base on it to fit specific scenarios with different assumptions [12]–[14]. However, on one hand, physical model-based methods require abundant prior

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Maaz Rehan.

knowledge like initial states and pollution sources, which is unavailable in real circumstances. On the other hand, physical models are unable to involve all the influencing factors and apply to all scenarios because empirical parameters of specific scenarios can hardly be generalized. b) Data-driven methods are proposed as the amount of data samples grows. Pollution field can be modeled as Gaussian-Markov random field [15], [16], Gaussian process [5] and random walk motion [6]. Various type of neural networks, especially those with deep learning structures, are also applied in static sensing [17]–[19]. Nevertheless, these data-driven methods are always seen as black boxes and the selection of these models is barely backed up with physical interpretation, which results in high generalization error. Besides, sampling bias in time and space, as well as in background scenarios, also lead to bad generalization ability.

Above all, there are two major challenges in fine-grained air pollution inference. *(i) Physical knowledge is beneficial but cannot be fully applied in fine-grained scenarios:* we cannot extensively obtain the required prior knowledge with fine resolution, and it has a high computational cost with a large amount of data. *(ii) Sampling bias from mobile sensing may lead to bad performance:* biased training samples may lead to bad performance in data-driven methods, and may also result in insufficient data for physical model estimation. Hybrid algorithms have the potential to utilize the advantages from both physical models and data-driven models. However, existing combination approaches including parallel, serial, and mixture [20], are still unable to address the above challenges.

In this paper, we present a multi-task learning scheme that combines the physical dispersion model and data-driven model quantitively, evolved from our previous work [21]. Under the scheme, these two models are seen as two parallel tasks and the whole model is trained by minimizing the weighted sum of the losses from these two tasks. In fine-grained air pollution inference, the relative impact from the selected physical model and the observed data can be adjusted according to our confidence in this physical model, therefore it relaxes the requirement on prior parameters and addresses the challenge *(i)*. Besides, we feed these two tasks with different data during multi-task learning, therefore the physical model works over the entire spatiotemporal field and the data-driven model works over the observed areas. It reduces the negative impact from sampling bias to a minimum extent, which addresses the challenge *(ii)*. Evaluations over a real-world deployment in Nanhai district of Foshan, China show that, the proposed method obtains 7.9% error reduction with lower variance and higher robustness against negative background conditions.

Our contributions can be summarized as below:

- A physically-based pollution dispersion model is reformulated, and a quantitative method is proposed to evaluate how well an inferred pollution map fits with this physical model.
- A multi-task learning method is proposed for fine-grained air pollution inference, which utilizes both knowledge from physical model and data-driven model.
- The algorithm is evaluated over a real-world air pollution sensing system. The effect of model integration is assessed and the performance of this hybrid method is compared with existing ones.

The rest of this paper is organized as follows. Section II lists related work. Section III elaborates on the reformulation of the physically-based pollution dispersion model. In Section IV, we present the multi-task learning scheme and introduce the details about how we use this hybrid algorithm for air pollution inference. Section V evaluates our method and compares it with existing works. Finally, this paper is concluded in Section VI.

II. RELATED WORK

This section introduces previous methods for air pollution inference, which recovers the entire pollution map from partial observation. These methods contain physical model-based algorithms and data-driven algorithms. Besides, hybrid algorithms that attempt to combine these two are also listed in this section.

A. PHYSICAL MODEL-BASED ALGORITHMS

Physical model-based approaches are widely used in the civil and environmental field. Sophisticated models like Community Multiscale Air Quality (CMAQ) and Weather Research and Forecasting model coupled with Chemistry (WRF-Chem) are widely used in country-level or city-level air pollution inference, with data from official air quality and meteorological stations [22], [23]. These approaches strictly follow the empirically validated physical and chemical laws, however, they suffer from a great deal of computational resources and are in great demand for information on the weather as well as the pollution source.

Models that only consider the physical dispersion process are also studied for long [8]–[11]. Several inference algorithms are proposed to fit specific scenarios with different assumptions [12]–[14]. While applying the pollutant dispersion model, similar challenges like great demand for fine-grained knowledge on weather and pollution source also exist. Assumptions like spatial homogeneity and temporal constancy in neighborhoods can alleviate these challenges, but sacrifice the model precision. Besides, empirical parameters from coarse-grained studies can also bring errors when they are applied into detail pollution patterns. Therefore, in fine-grained air pollution recovery, we cannot fully trust the physical dispersion model with limited background knowledge.

B. DATA-DRIVEN ALGORITHMS

With the development of electrochemical and optical sensors as well as wireless communication techniques, air quality sensor networks including AirSense [24], [25], BlueAer [6],

AirCloud [5], Gotcha [3], [4] and so on, are allowed to monitor the pollution field at fine spatiotemporal scales with both static and mobile deployment. As the number of data samples and the resolution of pollution map increase, the computational complexity of air pollution inference becomes higher and data-driven methods are proposed.

Among the data-driven algorithms, a typical kind of method utilizes auxiliary information like weather conditions, population density, traffic intensity, and point of interest, to infer the missing samples of pollution monitoring. For example, land-use regression (LUR) models a mapping relation from these influencing factors to corresponding pollution concentration at individual locations [26]. It is commonly used with static deployment and mobile sensing under routine trajectories in OpenSense [27], [28]. This kind of method obtains satisfied performance but not convincing enough for increasing the pollution map resolution under cross-validation with only static or regular samples. Besides, it also requires huge manpower and material costs to obtain fine-grained auxiliary information, especially those time-variant ones.

Another kind of method is interpolation. Previous work includes inverse distance weighting (IDW), nearest-neighbor interpolation, and Kriging interpolation [29]. These interpolation methods infer unobserved samples using its nearby observed ones, which are relatively easy to implement, but their accuracy at a specific location is greatly influenced by its nearby sampling conditions. Therefore, they suffer from irregular sampling and accidental errors, which happen frequently under mobile sensing, resulting in bad inference performance.

Customized data-driven algorithms for mobile sensing are proposed along with mobile systems. BlueAer team models the particle motion as random walk therefore proposed a probabilistic concentration estimation method (PCEM) [30]. AirCloud team uses Gaussian Process Regression (GPR), which models the pollution map as a multivariate Gaussian distribution and considers related features including GPS coordinates, location-related humidity, temperature as well as point of interest (POI) [5]. Besides, neural networks are also applied in air pollution inference with different forms. For example, U-Air team establishes a deep network with a spatiotemporal correlated structure, which models the dependency in the neighborhoods of the pollution map [18], [19]. Networks with recurrent and convolutional structures are also used in a series of recent works [31], [32].

However, these data-driven methods including neural networks may suffer from low robustness and uncertain performance. On one hand, the data amount from air quality sensing systems are not sufficient to be large-scale. It is hard to determine the proper capacity of a data-driven model, which is able to both describe the fine-grained pollution variations and avoid the over-fitting problem under limited data amounts. On the other hand, sampling bias happens frequently in mobile sensing from both time and space dimension, which also lead to bad generation ability.

C. HYBRID ALGORITHMS

To address the above challenges from physical models and data-driven models, efforts have been done to combine these two models and make a balance between them in recent years.

A deep autoencoder method is proposed by providing its inner convolutional long short-term memory structure with physical interpretations [33]. PCEM and GPR can also be seen as algorithms with basic physical assumptions. Nevertheless, these three methods apply physical knowledge qualitatively but not quantitatively in data-driven models, and it is still unknown how the physical knowledge and data-driven models couple with each other.

A physics guided and adaptive approach (PGA) is proposed to adaptively estimate the pollution level with the physical dispersion model or an artificial neural network under a particle filter structure [34], but it needs to calculate both the physical model and the neural network at the same resolution thus it still suffers from high computational complexity. A Gaussian plume model is applied on the basis of the neural network [35], however, this physical model only works when obtaining exact knowledge on the pollution sources. A physics-informed CoKriging is also proposed, which uses Monte Carlo simulations on the stochastic physical model therefore enables the Kriging interpolation without Gaussian assumptions [36]. However, in our scenario, it is hard to apply the Monte Carlo simulation with such a number of uncertain physical parameters.

In order to address the above challenges in hybrid methods, we propose a multi-task learning scheme to combine the data-driven model and physical model. This scheme enhances the learning process of neural network with a convective-diffusion model, and allows us to deal with varying levels of knowledge on auxiliary information as well as different confidence in the prior physical model. The basic form of this multi-task scheme is initially presented in our previous work [21], which considers the physical task with constant parameters. In this paper, we further specify the application of auxiliary information and launch a more exhaustive evaluation of this evolved algorithm. Details are discussed in the following sections.

III. MEASURING HOW WELL THE RECOVERED POLLUTION MAP FITS WITH PRIOR PHYSICAL MODEL

This section introduces how we measure the fitting degree of our inferred pollution map to our prior physical knowledge about air pollution dispersion. Firstly, we introduce a classic pollution dispersion model. Then we make discreteness approximations over this model, and then break the problem over the whole pollution map into subproblems within small neighborhoods. These reformulations allow us to evaluate how well a recovered pollution map accords with the physical model in these small neighborhoods, which also offers a way to construct the physical model-based task in our proposed multi-task learning scheme.

A. ORIGINAL PHYSICAL DISPERSION MODEL

Without loss of generality, we choose a classic formulation in atmospheric theory, which describes the atmospheric dispersion process using a convective-diffusion equation [8], [37],

$$\frac{\partial C}{\partial t} = \nabla \cdot (\mathbf{K} \nabla C) - \nabla \cdot (\boldsymbol{v} C) + S. \tag{1}$$

In this equation, C [kg/m^3] is pollutant concentration matrix, t[s] is the time, $\boldsymbol{v}$ [ms] is wind velocity vector, S [kg/m^3 s] is the source term indicating locations and the emission rates of sources, and $\mathbf{K}$ [m^2/s] is a diagonal matrix representing the diffusion coefficient, with its entries as turbulent eddy diffusivities. The pollution concentration change over time is the combined result from three components: diffusion, convection, and the source emission or destruction.

However, there are a number of limitations when we apply this physical model in different real-world scenarios. 1) It requires a preset initial pollution distribution with the same resolution we expect, which is unavailable to be extensively observed. 2) The real-world pollution field is nearly infinite, but we could only consider a finite-size area at one time, which requires ideal settings on the boundary conditions. 3) The number of uncertain parameters is always too large to be learned with enough data, thus assumptions like even distribution are made in $\mathbf{K}$, v, S. However, these assumptions also limit the expression ability of the model, especially in fine-grained air pollution modeling which reveals more detailed variation.

B. DISCRETENESS APPROXIMATION

The original physical model (1) can be reformulated with sequential spatial and temporal discretizations (details are expanded in Appendix).

Here we define the pollution concentration at a grid point of the discrete pollution field:

$$C_{[i,j,k]} = C(i\Delta x, j\Delta y, k\Delta t),$$

where Δx, Δy, Δt are the discretization intervals on x, y, t and i, j, k are their corresponding values. Here, x, y denote the two dimensions of the space (omitting the vertical variation) and t denotes time. For succinctness, we further denote the 3-dimensional coordinate of a grid point as a 3-dimensional vector $\boldsymbol{\theta} = [i, j, k]$, and $C_{\boldsymbol{\theta}}$ represents the concentration $C_{[i,j,k]}$. Before expatiating the discretized physical model, we define the *physically-related neighborhood* of $\boldsymbol{\theta}$:

$$\mathcal{T}(\boldsymbol{\theta}) = \big\{[i, j, k-1], [i-1, j, k-1], [i+1, j, k-1], \\ [i, j+1, k-1], [i, j-1, k-1]\big\}.$$

Then the reformulated physical dispersion model, can be presented as below, which describes the relationship between the concentration at $\boldsymbol{\theta}$ and the concentrations in its physically-related neighborhood $\mathcal{T}(\boldsymbol{\theta})$:

$$\widetilde{C}_{\boldsymbol{\theta}} = \boldsymbol{A}_{\boldsymbol{\theta}} \cdot \Big[C_\tau, \tau \in \mathcal{T}(\boldsymbol{\theta})\Big]^T + S_{\boldsymbol{\theta}}, \tag{2}$$

where $\widetilde{C}_{\boldsymbol{\theta}}$ denotes the physically inferred concentration at $\boldsymbol{\theta}$, and τ is the coordinate of the samples in physically-related neighborhood, then $[C_\tau, \tau \in \mathcal{T}(\boldsymbol{\theta})]$ represents a 5-dimensional vector composed of the pollution concentrations at each sample inside $\mathcal{T}(\boldsymbol{\theta})$, while $\boldsymbol{A}_{\boldsymbol{\theta}}$ represents the coefficient vector with each element corresponding to the concentration of each neighbor.

C. DEFINING THE FITTING DEGREE IN SMALL NEIGHBORHOODS

According to the above expressions, we are able to measure how well a pollution map fits with the dispersion model at $\boldsymbol{\theta}$. For an objective pollution map C, the fitting degree at $\boldsymbol{\theta}$ can be defined as

$$\mathcal{S}_{\boldsymbol{\theta}} = -\mathcal{L}\Big(C_{\boldsymbol{\theta}}, \widetilde{C}_{\boldsymbol{\theta}}\Big), \tag{3}$$

where $C_{\boldsymbol{\theta}}$ is the concentration of objective map at $\boldsymbol{\theta}$, $\mathcal{L}(\cdot, \cdot)$ could be any form of loss functions that measures the distance between these two variables. Based on this metric, the fitting degree of the whole pollution map can be calculated by going through each coordinate over it. Consequently, we can approach the physical model by maximizing the fitting degree and further construct an additional learning task to enhance conventional neural networks.

IV. A MULTI-TASK LEARNING SCHEME FOR ENHANCING NEURAL NETWORKS WITH PHYSICAL KNOWLEDGE

This section introduces our multi-task learning scheme for air pollution inference. An overview is given at first. Then we separately elaborate on our two tasks corresponding to the artificial neural network and the physical dispersion model. Finally, we explain how we implement our inference algorithm under these two tasks in detail.

A. OVERVIEW

The objective of air pollution inference is to obtain the whole pollution map based on partial observations, which is also, to predict the pollution concentration at unobserved time and positions. As we discuss previously, we aim to quantitively combine the physical model and the data-driven model, which could utilize the dispersion model with an acceptable computational complexity and also reduce the impact of sampling bias.

In order to accomplish the objective with the above requirements, we establish a multi-task learning scheme as shown in FIGURE 1. Under this scheme, Task I is data-driven, which learns from the observed dataset, while Task II is physical model-based, which learns from the pollution dispersion model. During the learning process, Task I and Task II are trained simultaneously with a preset weight. The weight can be adjusted according to our confidence level in our selected physical model. Besides, these two tasks can be performed over different sets of data samples. Task I learns from the observed dataset, while Task II works on an artificial dataset which is evenly distributed over the pollution map with a

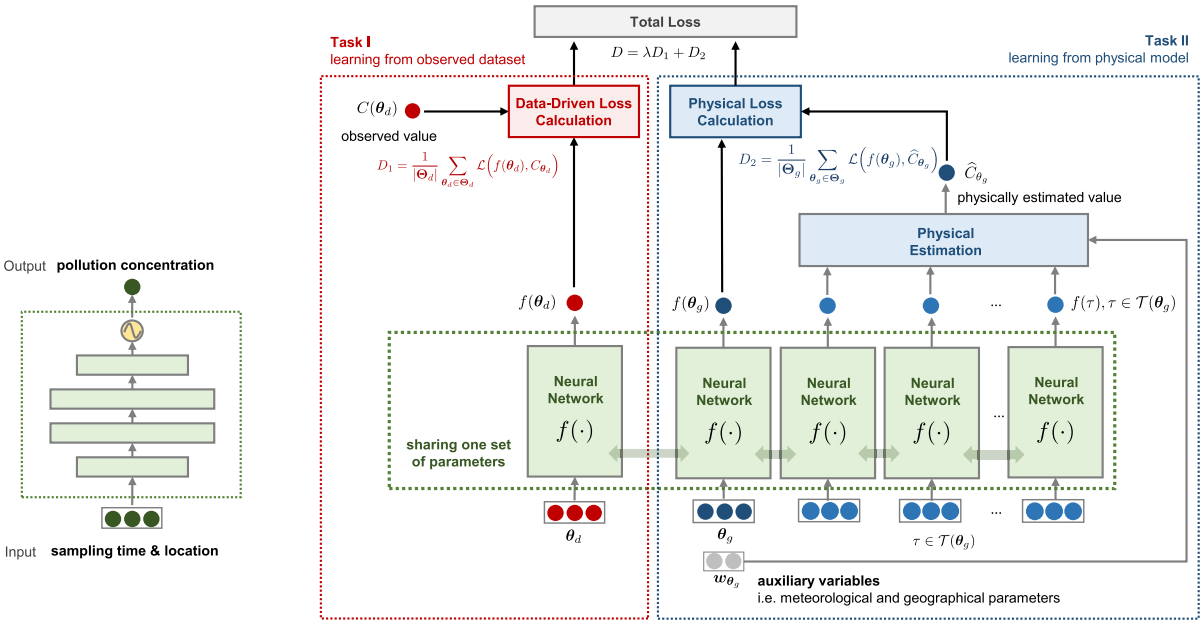


FIGURE 1. Sketch of the multi-task inference algorithm.

coarser resolution. Therefore, it addresses the problems of both the computational complexity and the sampling bias.

Inside these two tasks, a *Neural Network Module* is used as a basic module in multiple places, which aims to learn a mapping $f(\cdot)$ from the spatiotemporal coordinate to its corresponding concentration. Accordingly, the pollution concentration at $\boldsymbol{\theta}$ ($\boldsymbol{\theta}$ is a three-dimensional vector with two axes of space and one axes of time) is estimated as $f(\boldsymbol{\theta})$. This network is used multiple times with the same parameters in our learning scheme. In the training phase, the parameters are optimized under both tasks of fitting the observed dataset and fitting the physical model. Later in the inference phase, the concentration of the whole pollution field can be obtained with any resolution by inputting the coordinate of every grid point into this network. The details of the dual tasks are elaborated below.

B. TASK I: LEARNING FROM OBSERVED DATASET

Task I aims to fit the observed dataset. When inputting the coordinates of observed samples into the Neural Network Module, it learns to output their corresponding pollution concentration values. Here, we define coordinates that have been observed as $\boldsymbol{\theta}_\mathrm{d}$, and the observed value at $\boldsymbol{\theta}_\mathrm{d}$ as $C_{\boldsymbol{\theta}_\mathrm{d}}$. Then the loss of Task I, also the fitting error from the observed dataset, can be defined as:

$$D_1 = \frac{1}{|\boldsymbol{\Theta}_\mathrm{d}|} \sum_{\boldsymbol{\theta}_\mathrm{d} \in \boldsymbol{\Theta}_\mathrm{d}} \mathcal{L}\left(f(\boldsymbol{\theta}_\mathrm{d}), C_{\boldsymbol{\theta}_\mathrm{d}}\right), \tag{4}$$

where $\boldsymbol{\Theta}_\mathrm{d}$ is the observed dataset, $|\boldsymbol{\Theta}_\mathrm{d}|$ is the observed sample amount. $\mathcal{L}(\cdot,\cdot)$ is the loss function selected according to actual requirement.

This task is achieved by training on the observed dataset. Measurements from fixed air quality monitoring stations and mobile air quality sensors usually include GPS locations, sampling time, and observed concentrations. For each data sample, its GPS location and its sampling time determine its coordinate vector $\boldsymbol{\theta}_\mathrm{d}$, meanwhile its corresponding observed value determines $C_{\boldsymbol{\theta}_\mathrm{d}}$. Then, the coordinate vector $\boldsymbol{\theta}_\mathrm{d}$ and its corresponding concentration value $C_{\boldsymbol{\theta}_\mathrm{d}}$ constitute an input-output pair of the training set. The spatial-temporal coordinates of all the observed samples constitute the set $\boldsymbol{\Theta}_\mathrm{d}$.

On the other hand, the Neural Network Module with only Task I also works by directly minimizing D_1, which can be seen as a purely data-driven method. The performance of this method is presented as a baseline in Section V.

C. TASK II: LEARNING FROM PHYSICAL MODEL

Task II aims to fit the physical model. The deformation of the pollution dispersion model in Section III provides an approach to measure how well an inferred pollution map fits with the physical model. Specifically, as the right side of FIGURE 1 shows, the Neural Network Module in Task I is replicated six times and laid out in parallel. Suppose the coordinates for Task II are denoted as $\boldsymbol{\theta}_\mathrm{g}$. These modules are first fed with coordinates of a sample $\boldsymbol{\theta}_\mathrm{g}$ and its physically-related neighbors $\mathcal{T}(\boldsymbol{\theta}_\mathrm{g})$, and output the inferred concentration values at these coordinates. Then it is able to evaluate its fitting degree of this physically-related neighborhood from our prior physical knowledge using Equation (2) and (3).

In consideration of our imprecise knowledge about the parameters, we adopt an extra neural network as the *Coefficient Inference Module*, which supports the physical estimation of Equation (2). According to our prior physical

knowledge, the elements of coefficient $\boldsymbol{A}_{\boldsymbol{\theta}}$ in Equation (2) can be inferred from the value of wind speed as well as the diffusion coefficient, and $S_{\boldsymbol{\theta}}$ is the pollution source. Even though we are unable of achieving these parameters with fine, measurements on meteorology and geography can augment our inference in a data-driven way. Therefore, the Coefficient Inference Module f_A, f_S are fed with meteorological and geographical information $\boldsymbol{w}_{\boldsymbol{\theta}_g}$, including weather measurements, POIs, etc, to infer the coefficients $\boldsymbol{A}$ and S at $\boldsymbol{\theta}_g$.

Then the physical estimation based on these coefficients can be deployed over the neighborhood of the input sample, with all its concentration values inferred by our Neural Network Module:

$$\begin{aligned}\widehat{C}_{\boldsymbol{\theta}_g} &= \mathcal{F}(\boldsymbol{\theta}_g, \boldsymbol{w}_{\boldsymbol{\theta}_g}, \mathcal{T}(\boldsymbol{\theta}_g)) \\ &= f_A(\boldsymbol{\theta}_g, \boldsymbol{w}_{\boldsymbol{\theta}_g}) \cdot \left[f(\tau), \tau \in \mathcal{T}(\boldsymbol{\theta}_g)\right]^T + f_S(\boldsymbol{\theta}_g, \boldsymbol{w}_{\boldsymbol{\theta}_g}), \quad (5)\end{aligned}$$

where $\mathcal{F}$ represents the whole function of Task II, and $\widehat{C}_{\boldsymbol{\theta}_g}$ denotes the estimated concentration at $\boldsymbol{\theta}_g$ using above physically-based method. $[f(\tau), \tau \in \mathcal{T}(\boldsymbol{\theta}_g)]$ represents a 5-dimensional vector composed of the outputs of $f(\cdot)$ with each coordinate inside $\mathcal{T}(\boldsymbol{\theta}_g)$ as input.

Further, by comparing this physically estimated value with the direct output of Neural Network Module with $\boldsymbol{\theta}_g$, we can define the loss of Task II as:

$$D_2 = \frac{1}{|\boldsymbol{\Theta}_g|} \sum_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g} \mathcal{L}\left(f(\boldsymbol{\theta}_g), \widehat{C}_{\boldsymbol{\theta}_g}\right), \quad (6)$$

where $\boldsymbol{\Theta}_g$ denotes a grid of samples over the total space where the physical constraint is deployed, and $|\boldsymbol{\Theta}_g|$ is the sample amount of $\boldsymbol{\Theta}_g$. The loss function $\mathcal{L}(\cdot, \cdot)$ is selected according to actual requirement.

This task is achieved by training over an artificial dataset, constituted with a grid of samples over the targeted space. The grid size, which is also the discretization step, can be adjusted according to the resolution of our knowledge on meteorology and geography, as well as the computational complexity we can afford. The input of a sample at the discrete coordinate $\boldsymbol{\theta}_g$ is formed by $\boldsymbol{\theta}_g$ along with the coordinates of its physically-related neighbors $\mathcal{T}(\boldsymbol{\theta}_g)$ as well as the auxiliary information $\boldsymbol{w}_{\boldsymbol{\theta}_g}$. Then the training dataset is constituted with tuples at every discrete grid in the spatial-temporal space.

D. INFERENCE UNDER DUAL TASKS

Under our multi-task learning scheme, the above two tasks can be achieved at the same time by combining their losses:

$$D = \lambda D_1 + D_2, \quad (7)$$

where the regularization coefficient λ determines the relative impact of these two tasks. Therefore, the impact of our prior physical knowledge can be controlled. It can be adapted to different monitoring conditions in different application scenarios. For example, if the dataset is reliable with small sampling bias and noise, λ can be set to a large value. If we are confident in our physical model, λ can be set to a smaller value.

During the model training phase, parameters in all the Neural Network Modules are optimized simultaneously under two tasks, while parameters in the *Coefficient Inference Module* for physical estimation exclusively work in Task II. By minimizing the total loss D, parameters in both the Neural Network Module and the Coefficient Inference Modules are trained.

Above all, the proposed multi-task learning method provides a way to improve the data learning process with prior physical knowledge. In air pollution inference, the effect of the pollutant dispersion model is tunable by adjusting the spatial-temporal discretization intervals, the loss regularization coefficient λ, as well as the input of meteorological and geographical factors. It is also worth noticing that the data-driven task and the physical model-based task are fed with different data samples, which resolves the sampling bias problem to some extent. The data-driven task is trained over the observed dataset $\boldsymbol{\Theta}_d$, and the physical model-based task is trained over an artificial dataset $\boldsymbol{\Theta}_g$. Data sample $\boldsymbol{\theta}_g$ in the artificial dataset need not to be observed, which is, $\boldsymbol{\theta}_g$ need not belong to the sampled dataset $\boldsymbol{\Theta}_d$. In practical operation, a sample from $\boldsymbol{\Theta}_d$ is randomly paired up with a sample from $\boldsymbol{\Theta}_g$, and passed into our hybrid algorithm as a whole.

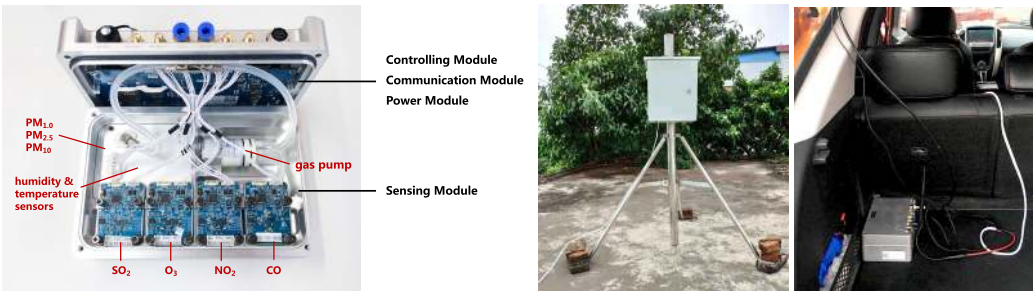
V. EVALUATIONS

The basic form of the multi-task algorithm has been evaluated over simulations as well as a 9-day real-world deployment in Tianjin in our previous work [21]. In order to further evaluate our evolved algorithm proposed in this paper, experiments are deployed over another real-world sensing system in the city of Foshan in China. Our algorithm is thoroughly evaluated over a two-week frequent data collection, and the results are compared with existing methods.

A. SYSTEM DEPLOYMENT

Our system adopts a centralized network, which consists of distributed sensing units and a cloud server. With either mobile or static configuration, these sensing units keep collecting air quality data and transmit it to the cloud server for further process. Each sensing unit contains four modules: sensing module, controlling module, communication module, and power module, as shown in FIGURE 2. The sensing module includes 7 types of gas sensors (SO_2, O_3, NO_2, CO, $PM_{1.0}$, $PM_{2.5}$, PM_{10}), a humidity sensor and a temperature sensor. The other three modules work together to support the sensing module. The observed air pollution data are then labeled with their recording time and GPS location and sent to the cloud server through the communication module.

Our system is deployed over an urban area in the Nanhai district in the city of Foshan, China, which is inside the Guangzhou Province. The area covers 104.77km^2, from 113.100°E to 113.200°E and from 23.000°N to 23.092°N. According to its local conditions, we install sensing units on 10 statical sites including security booths and building roofs, and on 8 environmental cruisers. The sensing units are deployed inside the cruisers, but can get access to the



(a) Hardware Design of Our Sensing Unit

(b) Static and Mobile Configurations

FIGURE 2. Real-world system deployment.

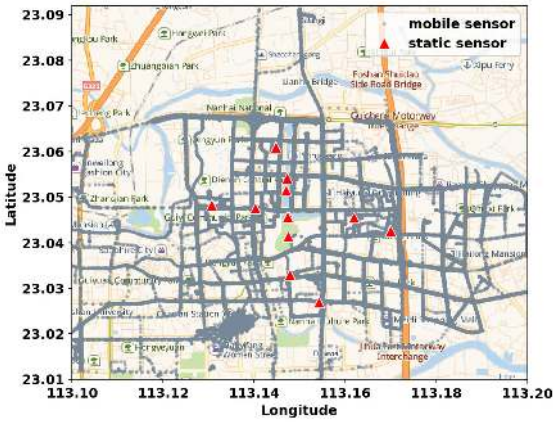


FIGURE 3. Spatial sampling distribution.

outside air through a tube. Experiments are run from Oct 8th to Oct 15th, and from Nov 19th to Nov 25th. During the experimental periods, these environmental cruisers move around the blocks in our experiment area. FIGURE 3 presents their trajectories during these two weeks, which is also the spatial distribution of our sampling data.

B. EXPERIMENT SETUP

We introduce the experiment setups of our evaluation, including the datasets, the performance metrics, the model training details, and the baseline methods.

1) DATASET DESCRIPTION

Our dataset contains three parts: the air pollution records from our distributed sensing system, the meteorological parameters from official weather stations, as well as the POI information.

The air pollution data are recorded by the distributed sensing units and gathered at the cloud server. The air pollution is sampled every 3 seconds and it is labeled with sampling time and GPS location. The data in cloud server is formatted as:{device id, timestamp, latitude, longitude, CO concentration [mg/m^3], NO_2 concentration [μg/m^3], SO_2 concentration [μg/m^3], O_3 concentration [μg/m^3], $PM_{1.0}$ concentration [μg/m^3], $PM_{2.5}$ concentration [μg/m^3], PM_{10} concentration [μg/m^3]}. Among these pollutants, we only focus on $PM_{2.5}$ in our algorithm evaluation, which is seen as the health indicator according to WHO. Therefore, the longitude, the latitude, and the sampling time constitute the 3-dimensional coordinates $\boldsymbol{\theta}$ of the samples, while the $PM_{2.5}$ constitutes the corresponding concentrations. Moreover, all the sensors are calibrated in the lab before deployment, therefore we assume the accuracy of these sensors during these two weeks.

The meteorological parameters are obtained from an open-source API, called Dark Sky.[1] Among all the information items, we select the top five related meteorological parameters, including the icon (the summary of weather conditions, including clear, rainy, snowy, windy, foggy, and cloudy), temperature, relative humidity, wind speed, and wind direction.

The POI information is collected through the web API of Amap.[2] We divide all the POIs into 6 categories that may impact the surrounding pollution level: catering service, car service, natural scenery, factory, traffic hinge, and others (including the office buildings and the residential buildings).

2) PERFORMANCE METRICS

To evaluate the performance of our proposed algorithm, we use cross-validation among the sensing units. In each round, samples from one mobile sensor are seen as the testing set, while samples from other 7 mobile sensors and all the static sensors are seen as the training set. After 8 rounds of evaluation, the validation results are averaged over the rounds to represent the algorithm performance.

Specifically, we use following three metrics to measure the validation results:

- Root Mean Squared Error

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\boldsymbol{\Theta}_t|}\sum_{\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t}\left(f(\boldsymbol{\theta}_t) - C_{\boldsymbol{\theta}_t}\right)^2},$$

- Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{|\boldsymbol{\Theta}_t|}\sum_{\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t}\left|f(\boldsymbol{\theta}_t) - C_{\boldsymbol{\theta}_t}\right|,$$

[1]Dark Sky: https://darksky.net/
[2]Amap: https://lbs.amap.com

- Pearson Correlation

$$\text{CORR} = \frac{\sum_{\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t} (f(\boldsymbol{\theta}_t) - \overline{f(\boldsymbol{\theta}_t)})(C_{\boldsymbol{\theta}_t} - \overline{C_{\boldsymbol{\theta}_t}})}{\sqrt{\sum_{\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t} (f(\boldsymbol{\theta}_t) - \overline{f(\boldsymbol{\theta}_t)})^2} \sqrt{\sum_{\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t} (C_{\boldsymbol{\theta}_t} - \overline{C_{\boldsymbol{\theta}_t}})^2}},$$

where $\boldsymbol{\Theta}_t$ represents the testing set, $|\boldsymbol{\Theta}_t|$ represents the sample amount of the testing set. The $\overline{f(\boldsymbol{\theta}_t)}$ denotes the mean value of $f(\boldsymbol{\theta}_t)$, and $\overline{C_{\boldsymbol{\theta}_t}}$ denotes the mean value of $C_{\boldsymbol{\theta}_t}$.

3) MODEL TRAINING

- *Preprocessing:* For consistency, we consider the pollution map as a discrete field with 3 dimensions, so that our method can be compared with existing ones. We divided our experimental area into 500m×500m spatial grids and set the temporal interval as 15min, thus the pollution map for a day is of size $20 \times 20 \times 96$.
- *Model Settings:* The physical model in Task II is considered with a resolution coarser than the pollution map, which is $10 \times 10 \times 24$. Therefore, the amount of physical model computation becomes $1/16$ of what it should be. In the pure neural network $f(\cdot)$ of Task I, we adopt a 4-layer fully connected structure (3-8-16-32-16-1). Then inside the physical estimation in Task II, we use 2-layer fully connected networks for both the coefficient inference module $f_A(\cdot)$ (5-8-5) and $f_S(\cdot)$ (6-4-1). For both Task I and Task II, we apply the mean-squared-error loss in the proposed task loss $\mathcal{L}(\cdot, \cdot)$, and each whole model works exclusively for each day.
- *Hyper-Parameter Selection:* The hyper-parameters in our proposed multi-task learning scheme is selected by cross-validation. FIGURE 4 shows the performance of our algorithm under different values of the hyper-parameters. The regularization coefficient λ, which is also the loss weights ratio between Task I and Task II, determines the relative impact of these two tasks. As shown in the left subfigure, the error of the hybrid algorithm decreases as λ increases from 10^{-2} to 1, and gradually increases as λ continues to grow. It indicates that, under the proposed scheme, Task I plays a dominant role in pollution inference, while Task II can further enhance the performance of data learning. Here, we set the regularization coefficient as 1, under which value the hybrid algorithm achieves the lowest error. This might because both the loss of Task I and the loss of Task II are based on the inferred pollution concentration, thus share the same order of magnitude. Besides, the learning rate during our model training is set as 10^{-3}, which achieves good performance with the highest efficiency.

4) BASELINE METHODS

We compare the performance of our algorithm with the following state-of-art methods that can be used in mobile sensing:

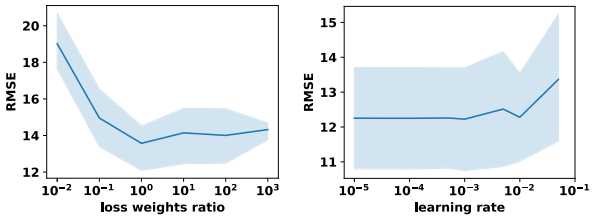


FIGURE 4. Effect of hyper-parameters on RMSE.

- *Inverse distance weighting interpolation (IDW)* [38] infers unobserved pollution concentration with its nearest observed spatial neighbors, which is a typical method in spatial interpolation.
- *Land use regression (LUR)* [26] combines the monitoring of air pollution and the development of stochastic models using predictor variables. Here, we consider the predictor variables including meteorological parameters and POIs that we mentioned above.
- *Gaussian process regression (GPR)* [5] is adopted by the AirCloud team, which models the air pollution field as a Gaussian Process. The probability distribution is learned from observed samples and then used to predict unobserved ones. Kriging interpolation, which is popular in the conventional environmental domain, is also based on a simplified form of Gaussian Process Regression.
- *Artificial neural network (ANN)* uses fully connected layers to map the relation between geo features and corresponding pollutant concentrations. This is a basic kind of neural networks, which can also be seen as the inference only under Task I.
- *Convolutional long short-term memory (ConvLSTM)* [31], [32] is another kind of neural network with a recurrent structure that considers the temporal dependencies in physical neighborhoods, and with convolutional structure that considers the spatial dependencies. It is usually used in static sensing with a large amount of sensing nodes.

In addition, we also evaluate the impact of different parts in our algorithm by comparing the performance of the following models:

- *Task I*: It represents the Neural Network Module under only the loss of Task I, here we adopt the same network structure in the baseline method ANN.
- *Task I + Task II*: It represents the Neural Network Module under both the loss of Task I and the loss of Task II, without the input of auxiliary information. Therefore, the Coefficient Inference Module degrades into coefficient variables A and S, which are constant over time and space, and their values are estimated during model training.
- *Task I + Task II + Info*: It is the whole model that we proposed in this paper, with auxiliary information including weather and POI.

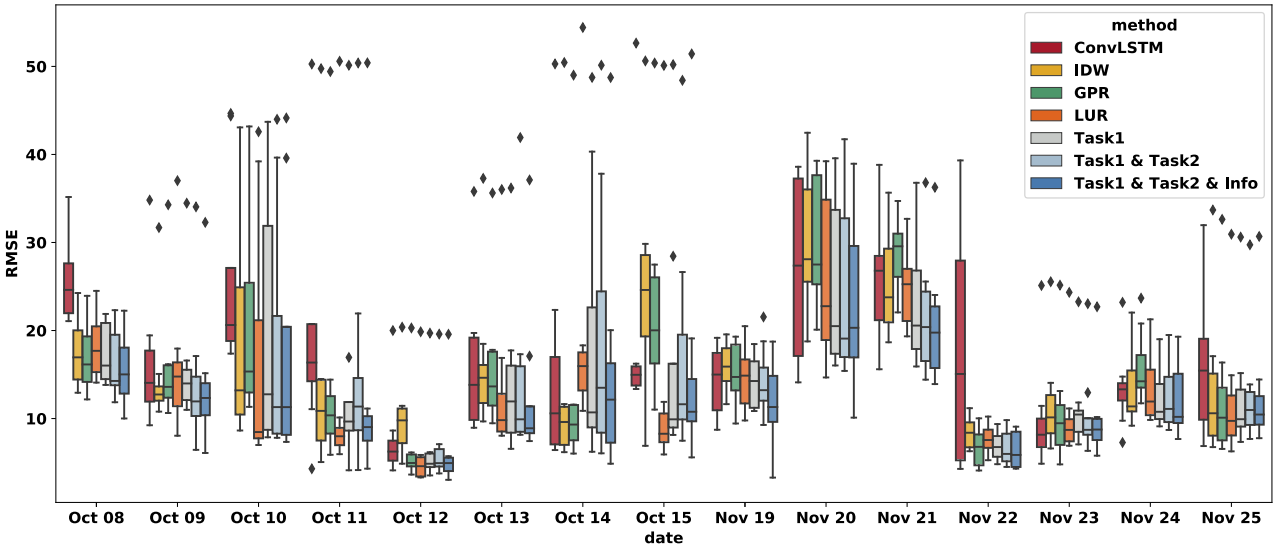


FIGURE 5. Daily performance on RMSE of different algorithms.

TABLE 1. Overall performance of different algorithms.

Method	RMSE	MAE	CORR
IDW	16.842	10.740	0.559
GPR	16.618	10.901	0.489
LUR	15.201	10.921	0.554
ConvLSTM	17.947	12.915	0.410
Task I (also ANN)	15.409	10.367	0.581
Task I + Task II	15.148	10.047	0.580
Task I + Task II + Info	**13.999**	**9.502**	**0.617**

It is worth noticing that we do not evaluate the individual performance of Task II, which is the Neural Network Module under only the loss of Task II. Because it can output any kind of pollution map as long as it satisfies the physical rules. The optimization under no observations is meaningless and does not converge.

C. ALGORITHM PERFORMANCE

The algorithm performances are first presented from an overall perspective and then discussed under different background conditions.

1) OVERALL PERFORMANCE

The average performance over two weeks are shown in TABLE 1. Overall, our method outperforms all the other baselines in various indicators. Among the baselines, GPR and IDW obtain a similar performance. Although ConvLSTM aims to capture the spatiotemporal dependencies in the pollution field, it fails to deal with sparse and irregular data from mobile sensing in its original form, therefore performs worst. The LUR is more cost-effective because it performs a few better than IDW, GPR, and ConvLSTM with lower computation cost. When comparing Task I, Task I + Task II and Task I + Task II + Info, the impact of the physical structure in Task II and the impact of auxiliary information are verified by the progressive error reduction and correlation increase, which also proves the effectiveness of our proposed method.

The daily performance on RMSE of these methods are shown as boxplots in FIGURE 5. The RMSEs on each day almost follows the same trend as the average. However, the performance of ConvLSTM shows great variations over different days, meanwhile the performances of regression methods IDW, GPR, LUR, and pure ANN are relatively stable. It indicates that the training of ConvLSTM may be significantly affected by the sampling condition, and the distributions of samples from mobile sensing are fairly irregular and sparse. Modeling the mapping relation from geo features to concentrations, instead of the mapping relation among concentrations at different locations, can be supported with more training samples.

During the days that all these methods get high RMSE like Oct 10, Oct 13, Oct 14, Nov 20, etc., our whole model (Task I + Task II + Info) not only achieves lower mean error than the pure ANN (Task I), but also reduces the variation of the errors from different folds. This proves that the utilization of the physical models, as well as the construction of virtual samples, help address the challenge of biased sampling and meanwhile capture the spatiotemporal dependencies in pollution maps effectively.

2) PERFORMANCE UNDER DIFFERENT BACKGROUND CONDITIONS

Since the performances of these algorithms show significant differences over various days, we further classify these days by their weather types and background pollution level, thus evaluate the algorithm robustness under different background conditions. As shown in FIGURE 6, the performances of

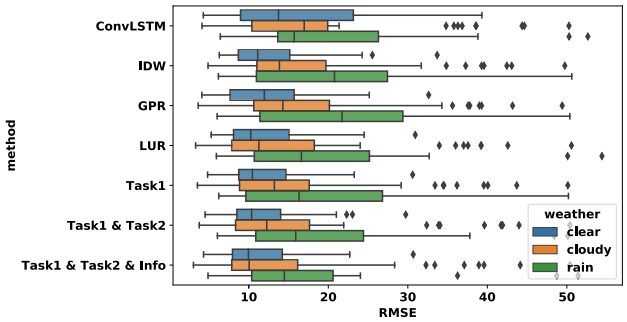
these methods are affected by two considered influencing factors. The errors of these methods under different weather types are shown in the top subfigure. As shown, all these algorithms have remarkably different RMSE under different weather types, they always perform best in clear days and worst in rainy days. When it is clear or cloudy, the differences of RMSE between our method and other baselines are not obvious. However, when it is rainy, our method has both significantly lower median and fewer variations on RMSE, which illustrates its robustness against negative impact from background conditions, benefited from the prior physical knowledge.

The errors of these methods under different background pollution levels are shown in the bottom subfigure. Here, the background pollution level is determined by the mean pollution concentration of the entire day, where "excellent" is from 0 to $35\mu g/m^3$, "good" is from $35\mu g/m^3$ to $75\mu g/m^3$, and "light pollution" is from $75\mu g/m^3$ to $115\mu g/m^3$. Among these methods, the proposed one obtains the least median errors under nearly all the background pollution levels. Besides, all these algorithms perform worst when the background level is "light pollution", however, most of them have better performance under "good" than the performance under "excellent". While looking into the weather type of these days, we find that there are a higher proportion of rainy days with an "excellent" pollution level than that with a "good" pollution level. It may indicate that the performances of these algorithms are easier to be affected by the weather type than the background pollution level, but more data are needed to confirm this point.

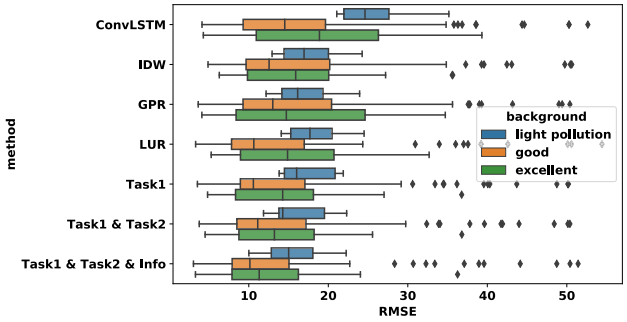
VI. CONCLUSION

This paper presents a novel method for fine-grained air pollution inference, which enhances the data learning process with physical knowledge using a multi-task learning scheme. Specifically, the data-driven model and the physical dispersion model are seen as two parallel tasks, and further these two tasks are trained by minimizing a weighted sum of their losses. By involving this reformulated physical model with a weighting coefficient, we can control its impact on our air pollution inference. On the other hand, the data-driven model works on the observed dataset and meanwhile the physical model works on a virtual dataset that covers the entire space, therefore it releases the negative impact of sampling bias from mobile sensing.

Our method is evaluated over a real-world deployment with both static and mobile air quality sensors in Foshan, China. With the resolution of 500m × 500m × 15min, the proposed method obtains the least error and highest correlation among existing algorithms, which is 7.9% less root-mean-squared error than the second place. Detailed evaluations also show that, benefited from the physical knowledge, our algorithm obtains stable performance with low variance and high robustness against negative conditions including bad weather and high-polluted environment.



(a) Weather Types



(b) Background Pollution Level

FIGURE 6. **Performance under different conditions.**

APPENDIX
MATHEMATICAL DETAILS IN DISCRETENESS APPROXIMATION

In this appendix, we expand the details of the discreteness approximation for the physical dispersion model.

First, assume that we only consider the near-ground air pollution, and the vertical variations of pollution field can be omitted. Here, we use x, y to denote the 2 dimensions of the space. Then $\boldsymbol{v}$ is with form (v_x, v_y), $\mathbf{K}$ is with form $\mathrm{diag}(K_x, K_y)$. After spatial and temporal discretization successively, the convective-diffusion equation can be rewritten as:

$$\frac{\partial C}{\partial t} + v_x\frac{\partial C}{\partial x} + v_y\frac{\partial C}{\partial y} + \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial x}\right)C = K_x\frac{\partial^2 C}{\partial x^2} + \frac{\partial K_x}{\partial x}\cdot\frac{\partial C}{\partial x} + K_y\frac{\partial^2 C}{\partial y^2} + \frac{\partial K_y}{\partial y}\cdot\frac{\partial C}{\partial y} + S. \quad (8)$$

Later, we express the discrete pollution map in terms of a three-dimensional matrix C with size $M \times N \times T$, and also assume the discretization intervals of x and y are $\Delta x = \Delta y = l$, the discretization interval of t is Δt. The pollutant concentration at the discrete spatial-temporal coordinate $[i, j, k]$ (corresponding to x-axis, y-axis of space, and t-axis of time) can be estimated by a group of surrounding samples. The mathematical expression of this relationship is:

$$\begin{aligned} C[i,j,k] = {} & A_1[i,j,k]\cdot C[i,j,k-1] \\ & + A_2[i,j,k]\cdot C[i-1,j,k-1] \\ & + A_3[i,j,k]\cdot C[i+1,y,k-1] \end{aligned}$$

$$+A_4[i,j,k]\cdot C[i,j-1,k-1]$$
$$+A_5[i,j,k]\cdot C[i,j+1,k-1]$$
$$+S[i,j,k], \tag{9}$$

where $S[i,j,k]$ [kg/m^3] is the source emission concentration, and $A_1[i,j,k],\cdots,A_5[i,j,k]$ are the coefficients matrixes corresponding to the concentrations of physically-related neighbors, which can be calculated from the value of v_x, v_y, K_x, K_y in this physically-related neighborhood:

$$A_1[i,j,k] = 1-\Big\{\frac{2}{l^2}\big(K_x[i,j]+K_y[i,j]\big) - \frac{1}{2l}\big(v_x[i+1,j,k-1]-v_x[i-1,j,k-1] + v_y[i,j+1,k-1]-v_y[i,j-1,k-1]\big)\Big\}\Delta t,$$

$$A_2[i,j,k] = \Big\{\frac{1}{4l^2}\big(4K_x[i,j]+K_x[i+1,j] - K_x[i-1,j]\big)-\frac{1}{2l}v_x[i,j,k-1]\Big\}\Delta t,$$

$$A_3[i,j,k] = \Big\{\frac{1}{4l^2}\big(4K_x[i,j]-K_x[i+1,j] + K_x[i-1,j]\big)-\frac{1}{2l}v_x[i,j,k-1]\Big\}\Delta t,$$

$$A_4[i,j,k] = \Big\{\frac{1}{4l^2}\big(4K_y[i,j]+K_y[i,j+1] - K_y[i,j-1]\big)-\frac{1}{2l}v_y[i,j,k-1]\Big\}\Delta t,$$

$$A_5[i,j,k] = \Big\{\frac{1}{4l^2}\big(4K_y[i,j]-K_y[i,j+1] + K_y[i,j-1]\big)-\frac{1}{2l}v_y[i,j,k-1]\Big\}\Delta t.$$

Besides, it is worth noting that the physical model can be applied with different spatial-temporal discretization intervals Δx, Δy, Δt as needed. It can also be applied under different parametric assumptions. If v_x, v_y, K_x, K_y, S are assumed to be spatially-temporally homogeneous, the relationship between $C[i,j,k]$ and $C[i,j,k-1]$, $C[i-1,j,k-1]$, $C[i+1,y,k-1]$, $C[i,j-1,k-1]$, $C[i,j+1,k-1]$ is independent of time and space. Otherwise, the relationship at particular coordinates is determined by $\boldsymbol{v}$, $\mathbf{K}$ and S at $[i,j,k]$.

REFERENCES

[1] World Helth Organization. (2019). *Healthy Environments for Healthier Populations: Why Do They Matter, and What Can We Do?*. [Online]. Available: https://www.who.int/publications-detail/healthy-environments-for-healthier-populations-why-do-they-matter-and-what-can-we-do

[2] P. Kumar, L. Morawska, C. Martani, G. Biskos, M. Neophytou, S. Di Sabatino, M. Bell, L. Norford, and R. Britter, "The rise of low-cost sensing for managing air pollution in cities," *Environ. Int.*, vol. 75, pp. 199–205, Feb. 2015.

[3] X. Xu, P. Zhang, and L. Zhang, "Gotcha: A mobile urban sensing system," in *Proc. 12th ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, 2014, pp. 316–317.

[4] X. Xu, X. Chen, X. Liu, H. Y. Noh, P. Zhang, and L. Zhang, "Gotcha II: Deployment of a vehicle-based environmental sensing system: Poster abstract," in *Proc. 14th ACM Conf. Embedded Netw. Sensor Syst. (CD-ROM)*, Nov. 2016, pp. 376–377.

[5] Y. Cheng, X. Li, Z. Li, S. Jiang, Y. Li, J. Jia, and X. Jiang, "AirCloud: A cloud-based air-quality monitoring system for everyone," in *Proc. 12th ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, 2014, pp. 251–265.

[6] Y. Hu, G. Dai, J. Fan, Y. Wu, and H. Zhang, "BlueAer: A fine-grained urban PM2.5 3D monitoring system using mobile sensing," in *Proc. IEEE 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.

[7] H. Guo, G. Dai, J. Fan, Y. Wu, F. Shen, and Y. Hu, "A mobile sensing system for urban monitoring with adaptive resolution," *J. Sensors*, vol. 2016, May 2016, Art. no. 7901245.

[8] S. P. Arya, *Air Pollution Meteorology and Dispersion*, vol. 310. New York, NY, USA: Oxford Univ. Press, 1999.

[9] Z. Zlatev and I. Dimov, *Computational and Numerical Challenges in Environmental Modelling*, vol. 13. Amsterdam, The Netherlands: Elsevier, 2006.

[10] N. S. Holmes and L. Morawska, "A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available," *Atmos. Environ.*, vol. 40, no. 30, pp. 5902–5928, Sep. 2006.

[11] P. Zannetti, *Air Pollution Modeling: Theories, Computational Methods, and Available Software*. Cham, Switzerland: Springer, 2013.

[12] H. Zhang, G. Chen, J. Hu, S.-H. Chen, C. Wiedinmyer, M. Kleeman, and Q. Ying, "Evaluation of a seven-year air quality simulation using the weather research and forecasting (WRF)/Community multiscale air quality (CMAQ) models in the Eastern United States," *Sci. Total Environ.*, vols. 473–474, pp. 275–285, Mar. 2014.

[13] Z. Wang, T. Maeda, M. Hayashi, L.-F. Hsiao, and K.-Y. Liu, "A nested air quality prediction modeling system for urban and regional scales: Application for high-ozone episode in Taiwan," *Water, Air, Soil Pollut.*, vol. 130, nos. 1–4, pp. 391–396, Aug. 2001.

[14] P. E. Saide, G. R. Carmichael, S. N. Spak, L. Gallardo, A. E. Osses, M. A. Mena-Carrasco, and M. Pagowski, "Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–chem CO tracer model," *Atmos. Environ.*, vol. 45, no. 16, pp. 2769–2780, May 2011.

[15] M. Kanevski, "Advanced mapping of environmental data: Geostatistics," in *Machine Learning and Bayesian Maximum Entropy*. London, U.K.: Wiley, 2008.

[16] M. Sherman, *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Hoboken, NJ, USA: Wiley, 2011.

[17] M. A. Elangasinghe, N. Singhal, K. N. Dirks, J. A. Salmond, and S. Samarasinghe, "Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering," *Atmos. Environ.*, vol. 94, pp. 106–116, Sep. 2014.

[18] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2013, pp. 1436–1444.

[19] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 2267–2276.

[20] M. von Stosch, R. Oliveira, J. Peres, and S. Feyo de Azevedo, "Hybrid semi-parametric modeling in process systems engineering: Past, present and future," *Comput. Chem. Eng.*, vol. 60, pp. 86–101, Jan. 2014.

[21] R. Ma, X. Xu, Y. Wang, H. Y. Noh, P. Zhang, and L. Zhang, "Guiding the data learning process with physical model in air pollution inference," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 4475–4483.

[22] D. Byun and K. L. Schere, "Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (CMAQ) modeling system," *Appl. Mech. Rev.*, vol. 59, no. 2, p. 51, 2006.

[23] G. A. Grell, S. E. Peckham, R. Schmitz, S. A. McKeen, G. Frost, W. C. Skamarock, and B. Eder, "Fully coupled 'online' chemistry within the WRF model," *Atmos. Environ.*, vol. 39, no. 37, pp. 6957–6975, 2005.

[24] K. Aberer, S. Sathe, D. Chakraborty, A. Martinoli, G. Barrenetxea, B. Faltings, and L. Thiele, "OpenSense: Open community driven sensing of environment," in *Proc. ACM SIGSPATIAL Int. Workshop GeoStreaming (GIS)*, 2010, pp. 39–42.

[25] J. J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel, "Sensing the air we breathe–the opensense zurich dataset," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 1–3.

[26] G. Hoek, R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs, "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmos. Environ.*, vol. 42, no. 33, pp. 7561–7578, Oct. 2008.

[27] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervas. Mobile Comput.*, vol. 16, pp. 268–285, Jan. 2015.

[28] M. D. Mueller, D. Hasenfratz, O. Saukh, M. Fierz, and C. Hueglin, "Statistical modelling of particle number concentration in zurich at high spatio-temporal resolution utilizing data from a mobile sensor network," *Atmos. Environ.*, vol. 126, pp. 171–181, Feb. 2016.

[29] M. Wu, J. Huang, N. Liu, R. Ma, Y. Wang, and L. Zhang, "A hybrid air pollution reconstruction by adaptive interpolation method," in *Proc. 16th ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2018, pp. 408–409.

[30] Y. Hu, J. Fan, H. Zhang, X. Chen, and G. Dai, "An estimated method of urban $PM_{2.5}$ concentration distribution for a mobile sensing system," *Pervas. Mobile Comput.*, vol. 25, pp. 88–103, Jan. 2016.

[31] C. Wen, S. Liu, X. Yao, L. Peng, X. Li, Y. Hu, and T. Chi, "A novel spatiotemporal convolutional long short-term neural network for air pollution prediction," *Sci. Total Environ.*, vol. 654, pp. 1091–1099, Mar. 2019.

[32] R. Ma, X. Xu, H. Y. Noh, P. Zhang, and L. Zhang, "Generative model based fine-grained air pollution inference for mobile sensing systems," in *Proc. 16th ACM Conf. Embedded Networked Sensor Syst.*, Nov. 2018, pp. 426–427.

[33] R. Ma, N. Liu, X. Xu, Y. Wang, H. Y. Noh, P. Zhang, and L. Zhang, "A deep autoencoder model for pollution map recovery with mobile sensing networks," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput. UbiComp/ISWC*, 2019, pp. 577–583.

[34] X. Chen, X. Xu, X. Liu, S. Pan, J. He, H. Y. Noh, L. Zhang, and P. Zhang, "PGA: Physics guided and adaptive approach for mobile fine-grained air pollution estimation," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervas. Ubiquitous Comput. Wearable Comput. (UbiComp)*, 2018, pp. 1321–1330.

[35] Y. Yang, Z. Zheng, K. Bian, L. Song, and Z. Han, "Real-time profiling of fine-grained air quality index distribution using UAV sensing," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 186–198, Feb. 2018.

[36] X. Yang, D. Barajas-Solano, G. Tartakovsky, and A. M. Tartakovsky, "Physics-informed CoKriging: A Gaussian-process-regression-based multifidelity method for data-model convergence," *J. Comput. Phys.*, vol. 395, pp. 410–431, Oct. 2019.

[37] J. M. Stockie, "The mathematics of atmospheric dispersion modeling," *SIAM Rev.*, vol. 53, no. 2, pp. 349–372, Jan. 2011.

[38] M. Jerrett, R. T. Burnett, B. S. Beckerman, M. C. Turner, D. Krewski, G. Thurston, R. V. Martin, A. van Donkelaar, E. Hughes, and Y. Shi, "Spatial analysis of air pollution and mortality in california," *Amer. J. Respiratory Crit. Care Med.*, vol. 188, no. 5, pp. 593–599, 2013.

RUI MA received the B.Sc. degree in electronic engineering from Tsinghua University, Beijing, China, in 2015, where she is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. Her research interests include machine learning and cyber-physical systems, especially data analytics in environmental sensing.

NING LIU received the B.Sc. degree in electronic engineering from Tsinghua University, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His current research interests include machine learning and wireless sensor networks.

XIANGXIANG XU (Student Member, IEEE) received the B.Sc. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His research interests include information theory and machine learning, with applications in data analytics.

YUE WANG received the B.Sc. and Ph.D. degrees from the Electronic Engineering Department, Tsinghua University, in 1999 and 2005, respectively. He is currently an Associate Professor with Tsinghua University. His research interests include computer networks, data fusion, and complex networks.

HAE YOUNG NOH received the B.S. degree in mechanical and aerospace engineering from Cornell University, the M.S. degree in civil and environmental engineering and the second M.S. degree in electrical engineering from Stanford University, and the Ph.D. degree in civil and environmental engineering from Stanford University. She is currently an Associate Professor with the Department of Civil and Environmental Engineering, Stanford University. Her researches focus on indirect sensing and physics-guided data analytics to enable low-cost, and non-intrusive monitoring of cyber-physical-human systems. She is particularly interested in developing smart structures and systems to be self-, user-, and surrounding-aware to provide safe and resilient environments and improve users quality of life, while reducing maintenance and operational costs. The result of her work has been deployed in a number of real-world applications from trains, to the Amish community, to eldercare centers, to pig farms. Dr. Noh received a number of awards, including the Google Faculty Research Awards, in 2013 and 2016, the Deans Early Career Fellowship, in 2018, and the National Science Foundation CAREER award, in 2017.

PEI ZHANG received the bachelor's degree (Hons.) from the California Institute of Technology, in 2002, and the Ph.D. degree in electrical engineering from Princeton University, in 2008. While at Princeton University, he developed the ZebraNet system, which is used to track zebras in Kenya. It was the first deployed, wireless, ad-hoc, mobile sensor network. He is currently an Associate Research Professor with the ECE Departments, Carnegie Mellon University. His recent work includes SensorFly (focus on groups of autonomous miniature-helicopter based sensor nodes) and muscle activity recognition (MARS). Beyond research publications, his work has been featured in popular media including CNN, science channel, discovery channel, CBS news, CNET, popular science, BBC focus, etc. He is also a Co-Founder of the startup Vibradotech. Dr. Zhang is a member of the Department of Defense Computer Science Studies Panel. In addition, he received several awards including the NSF CAREER Award, the SenSys Test of Time Award, and the Google Faculty Award.

LIN ZHANG received the B.Sc., M.Sc., and Ph.D. degrees from Tsinghua University, Beijing, in 1998, 2001, and 2006, respectively. He was a Visiting Professor with the University of California at Berkeley, from 2011 to 2013. He has been teaching the courses in selected topics in communication networks and information theory to senior undergraduate and graduate students at Tsinghua University. He is currently a Professor with the Tsinghua–Berkeley Shenzhen Institute, Tsinghua University. His research interests include efficient protocols for sensor networks, statistical learning and data mining algorithms for sensory data processing, and information theory. Since 2006, he has been implementing wireless sensor networks in a wide range of application scenarios, including underground mine security, precision agriculture, industrial monitoring, and also the 2008 Beijing Olympic Stadium (the Bird's Nest) structural security surveillance Project and a metropolitan area sensing and operating network (MASON) in Shenzhen. Dr. Zhang received the IEEE/ACM SenSys 2010 Best Demo Awards, the IEEE/ACM IPSN 2014 Best Demo Awards, and the IEEE CASE 2013 Best Paper Awards, and the Excellent Teacher Awards from Tsinghua University, in 2004 and 2010.