A Domain-Specific Bayesian Deep-learning Approach for Air Pollution Forecast

Yang Han*, Jacqueline C.K. Lam*, *Member, IEEE,* Victor O.K. Li, *Fellow, IEEE,* and Qi Zhang

Abstract—Given that poor air quality has obvious negative health impacts, predicting air pollution concentration is crucial and beneficial for public health. Motivated by recent advancements in deep-learning time series prediction, this study proposes a domain-specific Bayesian deep-learning model for long-term air pollution forecast in China and the United Kingdom, Our proposed model carries three novelties: First, we integrate a domain-specific knowledge taking into account the strong statistical relationship between $PM_{2.5}$ and PM_{10} as a regularization term; Second, we include an attention layer capable of capturing the influential historical feature, the recursive temporal correlation of air quality data, in our pollution prediction; Third, results generated from different multi-step forecast strategies have been combined based on corresponding uncertainty measures to improve our models performance. Using Beijing and London as our case studies, our results have shown that the Bayesian deep-learning model outperforms the baseline models. In particular, the incorporation of domain-specific knowledge into the Bayesian deep-learning model reduces prediction errors whilst the integration of Bayesian techniques allows the fusing of different forecast strategies to improve prediction accuracy. Feature selection can be performed and additional influential domain-specific features can be added in future to further improve our deep-learning models prediction accuracy and interpretability.

Index Terms—air pollution forecast, Bayesian deep-learning, prediction uncertainty, domain-specific knowledge, prediction fusion

1 Introduction

Rapid socio-economic development and urbanization have resulted in serious air pollution across many large cities in the world. Since poor air quality has clear negative impacts on physical and mental health [1], [2], accurately monitoring and predicting the concentration of air pollutants have become increasingly important to ensure that citizens can receive real-time health alerts and advice, and that government can make timely decisions such as tightening control of certain air pollutants on par with World Health Organization standards [3]. Among all air pollutants, particulate matters (PM), including $PM_{2.5}$ (PM with a diameter of less than 2.5 micrometers) and PM_{10} (PM with a diameter of less than 10 micrometers), have strongest negative impacts on public health [4]. $PM_{2.5}$ and PM_{10} may originate from similar emission sources; high statistical correlation between these two pollutants is often observed in empirical studies [5], [6]. Due to the similarities and differences in socio-economic structures and air pollution characteristics, we have selected Beijing, China, and London, the UK as our case study. We aim to predict the hourly $PM_{2.5}$ and PM_{10} concentration of the next 48 hours.

Urban air pollution forecast remains a challenge. Traditionally, physical models are used to simulate the air pollutant diffusion process based on the theories in atmospheric science [7]. Utilizing urban big data, recent advances in data-driven models [8], [9], especially deep-learning models [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], have made it possible to accurately predict air pollution levels based on the statistical patterns derived from historical data. However, data-driven approaches may suffer from the problems of limited and missing/noisy urban data. To tackle these challenges, prediction uncertainty, which takes into account the model and input uncertainty, will greatly enhance prediction reliability and interpretability of our model. Moreover, domain knowledge can be incorporated to further improve the prediction accuracy through an indirect supervision during the model training process. However, existing data-driven models for air pollution forecast often fail to give uncertainty estimation of the forecasts and to utilize domain knowledge. Therefore, this work presents a domain-specific Bayesian deep-learning approach to air pollution forecast. Our work addresses the aforementioned challenges in using urban big data technologies for air pollution forecast by providing uncertainty measures and integrating domain knowledge. The main contributions of our work are as follows:

- The first deep-learning model developed for air pollution forecast, taking into account of uncertainty measures.
- The first Bayesian deep-learning model that has incorporated a domain-specific knowledge into the model training procedure. The statistical relationship between pollutants is used to regularize our air pollution forecast.
- We put forward two multi-step air pollution forecast strategies, namely, one-time prediction and recursive prediction strategy, and provide fused results based on their uncertainty measures.
- Our Bayesian deep-learning model and domain-specific Bayesian deep-learning model both outperform the state-of-the-art deep-learning model and some existing statistical models in air pollution fore-

- *The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong.*
  E-mail: {yhan, jcklam, vli, zhangqi}@eee.hku.hk

* *Corresponding authors*

cast for both Beijing and London.

In relation to this work, we published a workshop paper, which aims to estimate the levels of air pollution without policy regulations using a Bayesian LSTM model [3]. We also archived a manuscript that focuses on the fine-grained air pollution forecast in the next hour by utilizing a CNN model [22]. However, our work is significantly different from these two pieces of work in the following ways. First, our current model focuses on providing long-term hourly air pollution forecast for individual monitoring stations in the next 48 hours, using Beijing, China and London, the UK as our case study. A prediction uncertainty measure is calculated for each predicted air pollution value. Second, the strong statistical relationship between $\mathrm{PM}_{2.5}$ and PM_{10}, a domain-specific knowledge of air-pollution modelling, has been incorporated to improve the models performance. Third, we investigate different air pollution forecast strategies and fuse the predicted values based on their corresponding uncertainty measures.

The rest of the paper is organized as follows. In Section 2, we review related works on air quality modelling and deep-learning methods. In Section 3, we discuss our collected data and the proposed Bayesian deep-learning framework in details. Experimental results are presented in Section 4 and are discussed in Section 5. Our work is concluded in Section 6.

2 Related Work

2.1 Urban Air Quality Modelling

Existing urban air quality models can be generally categorized into two approaches, namely, physical modelling (theory-driven) and machine learning (data-driven) [23]. On the one hand, physical models have been developed to simulate the air pollutant diffusion process and provide prediction at different scales including the city and street level. Typical models include ADMS, WRF-Chem, and CMAQ, which are based on atmospheric dispersion modelling, weather forecast modelling coupled with chemistry, and chemical transport modelling with weather data, respectively [7], [24]. However, these methods are often constrained by high computational cost, complex chemical processes modelling, and uncertainties in emission inventories [25], [26]. On the other hand, air pollution modelling can be based on analyzing historical data to establish the statistical patterns of air pollution and its relationship with other urban proxy variables such as meteorology and traffic, using linear models like Autoregressive Integrated Moving Average (ARIMA), or machine learning models such as Support Vector Regression (SVR) and Artificial Neural Network (ANN) which are capable of non-linear modelling in high dimensional space [7]. Recent advances in data-driven models have shown promising results in air pollution estimation and forecast based on urban big data [8], [9], [13], [17]. In particular, deep-learning methods such as Recurrent Neural Network (RNN) model and its variant Long Short-Term Memory (LSTM) and Gate Recurrent Unit (GRU) models have achieved state-of-the-art performance in many time series prediction tasks, and have been applied in air pollution forecast in some recent works such as [10], [11], [12], [14], [15], [19], [21]. Moreover, the spatial structures of monitoring station data are taken into account by convolutional neural network [16] or graph convolutional neural network [20], and the importance of various urban dynamics in predicting air quality are differentiated by neural attention network [18], [27]. More recently, state-of-the-art models for long-term air pollution forecast have utilized deep model fusion and ensemble strategies to account for the spatio-temporal characteristics of air pollution and weather data [13], [28]. However, long-term air pollution forecast remains a challenge since air pollution and other proxy data are often limited and missing/noisy. As such, appropriate tackling of missing/noisy data and accounting of the uncertainty of air quality modelling and urban big data, and the incorporation of domain-specific knowledge, can improve performance of more long-term (two-day) station-based air pollution forecast for Beijing and London.

2.2 Bayesian Deep-learning and Domain Knowledge-Based Deep-learning

Deep neural networks are able to approximate arbitrary functions from a large amount of data points, but are often criticized due to their "black box" challenge, given the limited interpretability and high uncertainty. Bayesian deep-learning combines probabilistic modelling with deep-learning, thus reducing model overfitting due to data sparsity and noise, while providing uncertainty quantification for the prediction [29]. Several approaches have been investigated for Bayesian deep-learning. On the one hand, stochastic processes can be adopted to model the distributions over functions, and in particular Gaussian processes (GPs) are shown to be equivalent to infinitely wide neural networks [30], [31]. This connection has motivated researchers to further combine Bayesian methods and neural networks more effectively by governing the mapping functions between network layers via GPs [32], equipping GPs with deep kernels [33], [34], or modelling the empirical distributions of functions through neural processes [35]. On the other hand, Bayesian methods are incorporated into deep neural networks by investigating the distributions over network weights. However, applying Bayesian inference in deep neural networks tends to be computationally intractable as the number of parameters is largely huge and complex non-linear relationships often exist. A number of methods are proposed to address these issues by using approximation techniques to estimate the posterior distribution of the network weight parameters [36]. For example, stochastic dropouts are used to generate random samples from the prediction distribution [37], [38]. This method does not require any changes in the existing network structure except for the additional dropout layers, but fails to model the network parameters as random variables directly. Alternatively, variational inference method is integrated into normal back-propagation in neural networks, in order to learn the posterior distribution on the weights of a neural network after observing the data, based on methods such as variational auto-encoder [39], expectation propagation [40], stochastic back-propagation [41], and Bayes by Backprop [42], [43].

Moreover, though deep-learning techniques have the potential to discover domain-specific knowledge from a large number of examples without prior information [44], domain-specific guidance can still be useful for enhancing the performance and interpretability of deep-learning models [45], [46], across a number of tasks such as machine translation [47], object recognition [48], and health risk prediction [49]. In general, to integrate domain-specific knowledge into the deep-learning framework, two approaches can be adopted. The first approach performs knowledge fusion by combing domain-specific features with the outputs of deep-learning models, in order to provide enhanced prediction [48]. Under such case, domain-specific knowledge is not involved during the model training process. The second approach incorporates domain-specific knowledge into the training procedure indirectly. From a Bayesian point of view, it can be encoded into the network weight parameters by properly specifying the prior distribution [50] or imposing constraints on the posterior distribution [51]. However, in many cases, establishing domain-specific priors or posteriors of the network weights could be a difficult task. Hence, it is more straightforward to regularize the predictions. More specifically, domain-specific knowledge can be utilized as an additional regularization term in the loss function to ensure that the predicted values are consistent with prior information, e.g., on the basis of a set of first-order logic rules [52] or physical laws described by partial differential equations [53].

3 The Proposed Approach

In this study, we aim to predict the hourly air pollution concentration of each of the monitoring stations in Beijing, China and London, the UK, respectively, in the next 48 hours. There are 35 air quality monitoring stations and 18 weather stations in Beijing, and 24 air quality monitoring stations and 26 weather stations in London. The target air pollutants to be predicted include $PM_{2.5}$ and PM_{10}.

The proposed approach consists of four components, as shown in Figure 1, namely, data collection, data pre-processing, model training, and model prediction.

3.1 Data Collection

We collected hourly air quality and meteorology data from 1 January 2017 to 31 May 2018, including station-level air pollution concentrations, station-level weather observations, and grid-level 48-hour weather forecasts at 10km x 10km resolution. All data was obtained from KDD Cup of Fresh Air website [54], except for the station-level meteorology data in London which was collected from the Met Office, the UK [55]. A summary of the data is shown in Table 1.

3.2 Data Pre-processing

Wind speed is decomposed into South-North and West-East components, based on wind direction. Each input is a vector representing the historical data at an air pollution monitoring station, including $PM_{2.5}$ concentration, PM_{10} concentration, and weather conditions. To reflect the temporal trends of air pollution, the hour, the day of the week, and the month have been included as an input vector. Moreover, to reflect station-fixed effects, station ID is included in the input vector. The corresponding output is a vector representing the $PM_{2.5}$ and PM_{10} values in the next 48 hours. To improve data quality, missing values in the air quality and weather time-series are imputed. Two imputation methods have been adopted to fill in the missing values. The first method utilizes the values of the adjacent observations to perform a spatio-temporal nearest neighbor imputation. First, for each station, forward linear temporal interpolation of the observed hourly values is used to fill in the missing hourly values less or equal to M hours in time series. Second, for each hour, the spatial interpolation (inverse distance weighting) of the observed values of N nearest stations is used. For the remaining missing values, the mean values of the same hour of the closest day where data is available are used. The second method exploits the inherent relationships of air pollutants and weather data. A multivariate iterative imputer is constructed, such that the missing values of each feature can be estimated based on the values of all other features [56]. After imputation has been conducted, the best imputation method is selected to fill in the missing values across our Beijing and London datasets. Finally, the meteorological conditions of air quality monitoring stations in Beijing and London, including temperature, pressure, humidity, wind speed (South-North), and wind speed (West- East), are derived as the inverse-distance-weighted values of three nearest weather observations.

3.3 Model Training

The pre-processed data is fed into the Bayesian deep-learning model for training. A Bayesian deep-learning model with network structure f and parameters θ is denoted as f_θ. Historical data at hour t consists of the features x_t including air quality and other covariates, while forecast data at hour t consists of the weather features z_t only. The model input consists of the observations over the past L hours (including current hour t): $X_{t-L+1:t} = \{x_{t-L+1}, \cdots, x_t\}$, the weather forecasts over the next H hours: $Z_{t+1:t+H} = \{z_{t+1}, \cdots, z_{t+H}\}$, and the corresponding actual observations: $Y_{t+1:t+H} = \{y_{t+1}, \cdots, y_{t+H}\}$, i.e., air pollution concentrations in the next H hours. Embedding layers are used to map the categorical features, including the time trends and station IDs, to vectors of continuous values [13]. The Bayesian model f_θ aims to find the optimal posterior distribution of the network weight parameters θ, given the observed pairs $(X_{t-L+1:t}, Z_{t+1:t+H})$ and $Y_{t+1:t+H}$. In this study, we focus on Bayesian RNN, which is capable of modelling time series data [43]. In the model, each weight parameter is a random variable with a prior, and the weight at each time step has the same distribution (see the Bayesian RNNs in Figure 1 for an illustration). By assigning a distribution instead of a fixed value to the parameters, the model reduces overfitting, addresses better the noisy and missing input data, and provides an uncertainty account for each output. More specifically, as $Y_{t+1:t+H}$ is a vector consisting of predictions in multiple steps, two commonly used multi-step forecast strategies are adopted [57], namely, one-time prediction and recursive prediction. One-time prediction aims to make multiple predictions in one single step. It consists of two parts, with each part processing different input data (see Figure 2a). One part focuses on the historical data including air quality and weather conditions, while

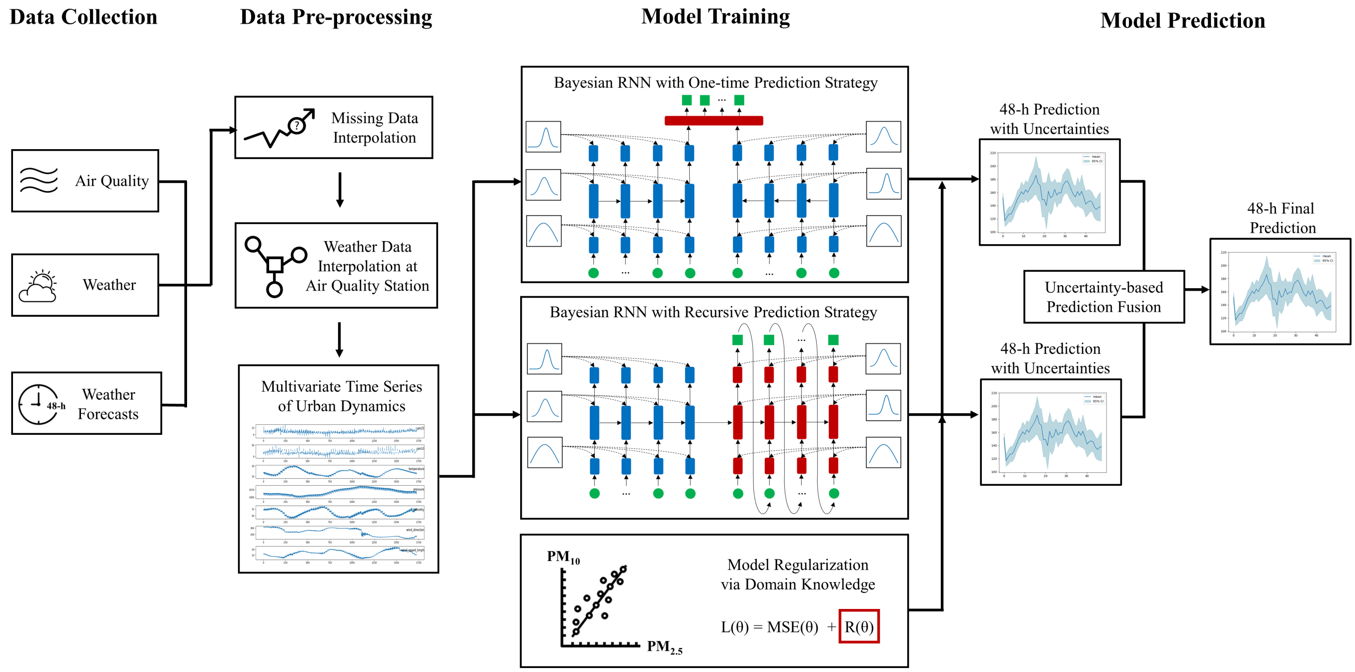

Data Collection
Data Pre-processing
Model Training
Model Prediction
Air Quality
Weather
48-h
Weather Forecasts
Missing Data Interpolation
Weather Data Interpolation at Air Quality Station
Multivariate Time Series of Urban Dynamics
Bayesian RNN with One-time Prediction Strategy
Bayesian RNN with Recursive Prediction Strategy
PM$_{10}$
PM$_{2.5}$
Model Regularization via Domain Knowledge
L(θ) = MSE(θ) + R(θ)
48-h Prediction with Uncertainties
48-h Prediction with Uncertainties
Uncertainty-based Prediction Fusion
48-h Final Prediction

Fig. 1. Overall framework of our proposed domain-specific Bayesian deep-learning network

TABLE 1
Dataset summary

Study Period: 31 January 2017 – 31 May 2018				
Urban dynamic	**Quantity (Beijing)**	**Quantity (London)**	**Granularity**	**Items**
Air quality	35 stations	24 stations	1 hour	PM$_{2.5}$, PM$_{10}$
Meteorology	18 stations	26 stations	1 hour	temperature, pressure, humidity, wind speed, wind direction
48-hour weather forecasts	651 grids	861 grids	1 hour	temperature, pressure, humidity, wind speed, wind direction

the other part utilizes the weather forecast data. The final hidden state of the first part (h_t^{hist}) and the second part (h_t^{fst}) are concatenated as H_t, and a dense layer is used to make final predictions $Y_{t+1:t+H}$. Conceptually, the network structure for one-time prediction is shown as follows:

$$h_t^{\text{hist}} = \text{Bayesian-RNN}(x_t, h_{t-1}^{\text{hist}}),$$

$$h_t^{\text{fst}} = \text{Bayesian-RNN}(z_{t+H}, h_{t-1}^{\text{fst}}),$$

$$H_t = \text{Concatenate}(h_t^{\text{hist}}, h_t^{\text{fst}}),$$

$$Y_{t+1:t+H} = \text{Bayesian-LINEAR}(H_t)$$

The second one, recurisve prediction, aims to make predictions recursively. It uses current single-step prediction as the input for the next single-step prediction. It consists of two parts, namely, an encoder and a decoder (see Figure 2b). The encoder encodes historical information including air quality data and weather data into a final hidden state h_t^{enc}. The decoder recursively predicts air quality in the next step based on current step information and the context (C), while utilizing weather forecast data at each step. The context is provided by an attention layer in the decoder, such that the decoder can capture the most influential parts of any input information that determines the air pollution forecast [27]. Finally, predictions of the 48 individual steps are combined for the final prediction $Y_{t+1:t+H}$. Conceptually, the network structure for recursive prediction is shown as follows:

$$h_t^{\text{enc}} = \text{Bayesian-RNN-Encoder}(x_t, h_{t-1}^{\text{enc}}),$$

$$h_t^{\text{dec}} = h_t^{\text{enc}}, C = h_t^{\text{enc}},$$

$$h_{t+1}^{\text{dec}} = \text{Bayesian-RNN-Decoder}(y_t, z_{t+1}, h_t^{\text{dec}}, C),$$

$$y_{t+1} = \text{Bayesian-LINEAR}(h_{t+1}^{\text{dec}}),$$

$$Y_{t+1:t+H} = \text{Concatenate}(y_{t+1}, \cdots, y_{t+H})$$

Since RNNs with simple recurrent units often suffer from gradients vanishing or exploding problems, more sophisticated recurrent units, including LSTM and GRU are proposed. LSTM and GRU have comparable performance [58], but GRU has fewer parameters and takes less time to train. For air quality time-series forecast, GRU performs better than LSTM [59]. Therefore, in this study, we use GRU as the recurrent unit in the proposed RNN models. To apply Bayesian inference for deep-learning, the statistical inference problem can be further transformed to an optimization problem. More specifically, given the training data set

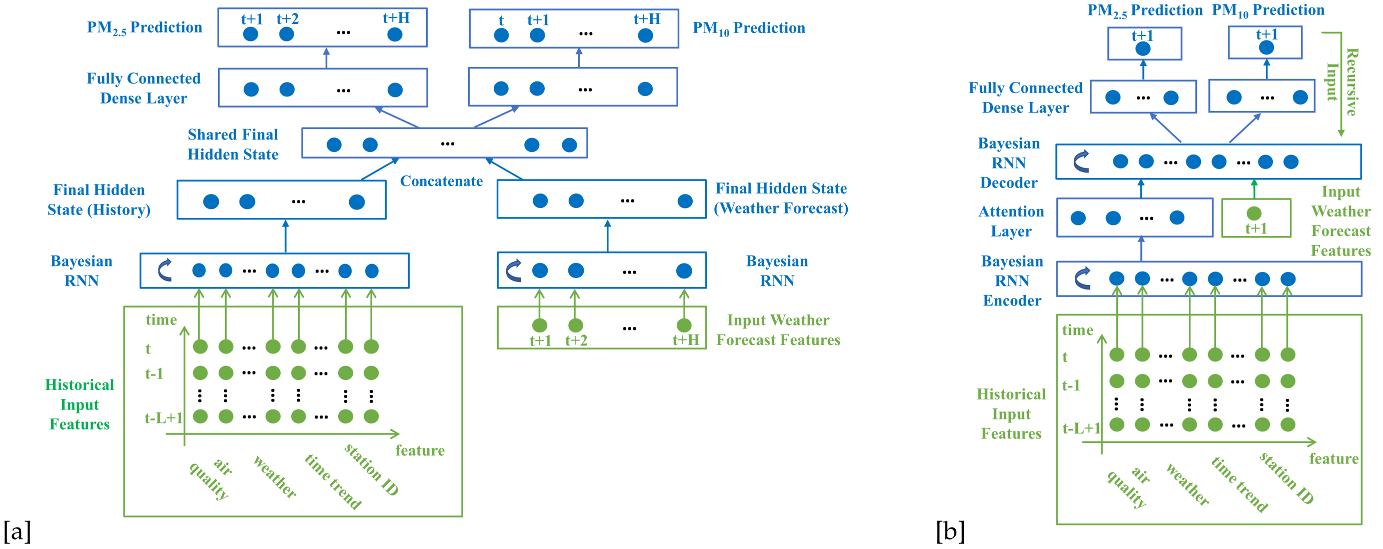


Fig. 2. Proposed Bayesian deep-learning network structures: (a) for one-time prediction strategy; (b) for recursive prediction strategy

$D = \{(X_{t-L+1:t}, Z_{t+1:t+H}), Y_{t+1:t+H}\}$, the true posterior distribution $p(\theta|D)$ is approximated by a variational distribution $q(\theta|\phi)$, and the distance between the true posterior distribution and the variational distribution is minimized to find the optimal posterior distribution. This distance is often measured by Kullback-Leibler (KL) divergence, and it can be further approximated as follows [42]:

$$\mathcal{L}(D,\varphi) = -\underbrace{\mathrm{E}_{q(\theta|\varphi)}[\log p(D|\theta)]}_{\text{Log likelihood cost}} + \underbrace{\mathrm{KL}[q(\theta|\varphi)||p(\theta)]}_{\text{KL cost}}$$

This cost function can be taken as a trade-off to balance the simplicity of the prior $p(\theta)$ and the complexity of the data D [31]. Furthermore, the log likelihood $\log p(D|\theta)$ of the forecast model with a Gaussian noise assumption can be formulated based on Mean Squared Error (MSE) [60] as follows:

$$\log p(Y_{t+1:t+H}|f_\theta(X_{x-L-1:t}, Z_{t+1:t+H})) \propto \frac{1}{H}\sum_{t+1}^{t+H} -\frac{1}{2\zeta^2}||y_i - \tilde{y}_i||^2 - \log\zeta$$

where $\tilde{y}_i$ is the predicted value given θ and ζ is the noise parameter. Hence, MSE loss is denoted as: $\mathrm{MSE}(\theta) = \frac{1}{H}\sum_{i=t+1}^{t+H}||y_i - \tilde{y}_i||^2$. In this work, since we aim to predict both PM$_{2.5}$ and PM$_{10}$ pollution concentrations, the MSE loss of these two target variables are calculated. Moreover, to integrate domain-specific knowledge with regards to the strong statistical relationship between PM$_{2.5}$ and PM$_{10}$ pollution [61], we also include a regularization term into the total loss, based on Pearson correlation coefficient between PM$_{2.5}$ and PM$_{10}$:

$$R(\theta) = \frac{\sum_{i=1}^{H}(\tilde{y}_i^{\mathrm{PM}_{2.5}} - \bar{y}_i^{\mathrm{PM}_{2.5}})(\tilde{y}_i^{\mathrm{PM}_{10}} - \bar{y}_i^{\mathrm{PM}_{10}})}{\sqrt{\sum_{i=1}^{H}(\tilde{y}_i^{\mathrm{PM}_{2.5}} - \bar{y}_i^{\mathrm{PM}_{2.5}})^2}\sqrt{\sum_{i=1}^{H}(\tilde{y}_i^{\mathrm{PM}_{10}} - \bar{y}_i^{\mathrm{PM}_{10}})^2}}$$

where $\bar{y}_i^{\mathrm{PM}_{2.5}}$ and $\bar{y}_i^{\mathrm{PM}_{10}}$ are the sample mean values. Finally, the total loss can be obtained as a weighted sum:

$$L(\theta) = \lambda_1\mathrm{MSE}_{\mathrm{PM}_{2.5}}(\theta) + \lambda_1\mathrm{MSE}_{\mathrm{PM}_{10}}(\theta) + \lambda_3 R(\theta)$$

where $\lambda_1, \lambda_2 \in (0,1)$ and $\lambda_3 < 0$ are hyper-parameters that represent the weights of different objectives. In order to ensure that the average total MSE is consistent, we set the sum of λ_1 and λ_2 to 1. λ_3 is a negative value as we expect a higher correlation with PM$_{2.5}$ or PM$_{10}$ will generate a lower total loss. Although these hyper-parameters are given fixed weights during training, we can fine-tune them based on the validation set, with more sophisticated extensions, such as, a ”hyper-parameter free” approach that learns the relative weights based on data uncertainty [60] or a more efficient approach that identifies the most important hyper-parameters and their interactions [62]. During the model training, the loss function can calibrate the predicted air pollution forecasts based on the ground truth data, while also imposing penalties on the learning procedure when the results of any predicted PM$_{2.5}$ and PM$_{10}$ values contradict the domain-specific knowledge that the values of these two pollutants are strongly correlated. To train our proposed models, we follow the work done by [43], and use a mixture Gaussian distribution as the prior and a diagonal Gaussian distribution as the variational posterior. Bayes by Backprop is adopted to update the weight parameters of the network while minimizing the loss and KL complexity (see Algorithm 1).

3.4 Model Prediction

After model training, we can perform air pollution forecast based on the fitted network model f_θ. One of the significant advantages of using Bayesian techniques is the ability of providing uncertainty estimation. Prediction uncertainty gives the level of confidence on the forecast values and improves the interpretability of the results. The distribution of the forecasts $Y^*_{t+1:t+H}$ for a new input $(X^*_{t-L+1:t}, Z^*_{t+1:t+H})$ can be calculated by marginalizing out the posterior distri-

Algorithm 1 Model Training via Bayes by Backprop

```
Require: training data $D$, network structure $f$, batch size $B$, and learning rate $\alpha$
repeat
    Sample a mini-batch of size $B$ from the training data $D$
    Sample $\epsilon \sim \mathcal{N}(0, I)$
    Set network parameters $\theta = \mu + \sigma\epsilon$, where $\mu$ and $\sigma$ are the mean and standard deviation, respectively
    Compute the gradients of domain-specific knowledge regularized loss with respect to $\theta$ using normal back-propagation: $g_\theta^L$
    Compute the gradients of $F(\mu, \sigma, \theta) = \log \mathcal{N}(\mu, \sigma^2) - \log p(\theta)$ with respect to $\mu, \sigma, \theta$: $g_\mu^F, g_\sigma^F, g_\theta^F$
    Update $\mu = \mu - \alpha \frac{g_\theta^L + g_\theta^F + g_\mu^F}{B}$
    Update $\sigma = \sigma - \alpha \frac{g_\theta^L \epsilon + g_\theta^F \epsilon + g_\sigma^F}{B}$
until stopping criteria is met
return fitted network model $f_\theta$
```

bution of θ:

$$\begin{aligned} p(Y_{t+1:t+H}^* | X_{t-L+1:t}^*, Z_{t+1:t+H}^*) &\\ = \int p(Y_{t+1:t+H}^* | f_\theta(X_{t-L+1:t}^*, Z_{t+1:t+H}^*)) p(\theta|D) d\theta &\\ = \mathrm{E}_{p(\theta|D)}[p(Y_{t+1:t+H}^* | f_\theta(X_{t-L+1:t}^*, Z_{t+1:t+H}^*))] & \end{aligned}$$

This can be seen as averaging forecasts from an infinite number of models. In practice, the variance of the forecast distribution is often used to quantify the prediction uncertainty, and it can be further decomposed into two terms using the law of total variance [63], [64], namely, model uncertainty and data uncertainty. Model uncertainty refers to the uncertainty introduced by the model parameters θ. Similar to the ideas proposed in [37], [38] that use Monte Carlo dropout as an approximation to model uncertainty, we use the sample variance of the values predicted by the network model f with different weight parameters $\{\theta_1, \theta_2, \cdots, \theta_T\}$ as a measure of model uncertainty. The weight parameters θ_i is randomly drawn from the posterior distribution. This is repeated for T times. Then, given a new input, the model uncertainty is calculated as follows:

$$\begin{aligned} \text{Model Uncertainty} = \frac{1}{T} \sum_{i=1}^{T} (f_{\theta_i}(X_{t-L+1:t}^*, Z_{t+1:t+H}^*) &\\ - \frac{1}{T} \sum_{i=1}^{T} f_{\theta_i}(X_{t-L+1:t}^*, Z_{t+1:t+H}^*))^2 & \end{aligned}$$

Data uncertainty refers to the irreducible noise inherent in the data, which could be estimated by the residual sum of squares on the independent validation dataset [64], assuming that the irreducible noise in the air quality data is homogeneous. Given a validation set of size V, the data uncertainty is calculated as follows:

$$\begin{aligned} &\text{Data Uncertainty} = \\ &\frac{1}{V} \sum_{v=1}^{V} (Y_{t+1:t+H}^v - \frac{1}{T} \sum_{i=1}^{T} f_{\theta_i}(X_{t-L+1:t}^v, Z_{t+1:t+H}^v))^2 \end{aligned}$$

In summary, given a new input, Algorithm 2 shows how the proposed models make predictions with uncertainty measures.

Algorithm 2 Prediction with Uncertainty Measure

```
Require: input $X_{t-L+1:t}^*$ and $Z_{t+1:t+H}^*$, fitted network model $f_\theta$ with parameters $\theta = \{\mathcal{N}(\mu_j, \sigma_j^2)\}_{j=1}^{j=W}$, sample size $T$, and an independent validation set of size $V$
for $i = 1$ to $i = T$ do
    for $j = 1$ to $j = W$ do
        Sample $w_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$
    end for
    Let $\theta_i = \{w_j\}_{j=1}^{j=W}$
end for
Compute final prediction $Y_{t+1:t+H}^*$:
    $$\frac{1}{T} \sum_{i=1}^{T} f_{\theta_i}(X_{t-L+1:t}^*, Z_{t+1:t+H}^*)$$
Compute model uncertainty $\eta_1$:
    $$\frac{1}{T} \sum_{i=1}^{T} (f_{\theta_i}(X_{t-L+1:t}^*, Z_{t+1:t+H}^*) - Y_{t+1:t+H}^*$$
Compute data uncertainty $\eta_2$:
    $$\frac{1}{V} \sum_{v=1}^{V} (Y_{t+1:t+H}^v - \frac{1}{T} \sum_{i=1}^{T} f_{\theta_i}(X_{t-L+1:t}^v, Z_{t+1:t+H}^v))^2$$
Compute prediction uncertainty $\eta$: $\eta_1 + \eta_2$
return $Y_{t+1:t+H}^*$ and $\eta$
```

Furthermore, predictions generated from different forecast strategies could exhibit different characteristics [57]. In general, recursive prediction could capture temporal dependencies among prediction steps, but may suffer from error accumulations; and one-time prediction tends to perform better at the later prediction steps as error propagation is less significant, but may perform worse at the initial steps. To combine the potential strengths of the two multi-step forecast strategies, a hybrid Bayesian RNN model is therefore proposed to fuse the results derived from these two strategies by utilizing the uncertainty measures. In particular, prediction uncertainties are calculated for each hour and each pollutant, and the predicted values are weighted according to their uncertainty measures, in order to generate the fused outputs:

$$\begin{aligned} Y_{t+1:t+H}^{\text{Hybrid}} = \rho(\eta^{\text{MP}}, \eta^{\text{RP}}) \circ Y_{t+1:t+H}^{\text{MP}} &\\ + (1 - \rho(\eta^{\text{MP}}, \eta^{\text{RP}})) \circ Y_{t+1:t+H}^{\text{RP}} & \end{aligned}$$

where $\circ$ denotes element-wise product, and ρ is the weighting function. We adopt two weighting strategies. The first one aims to provide an uncertainty-averaged output:

$$\rho(\eta^{\text{MP}}, \eta^{\text{RP}}) = \frac{\eta^{\text{MP}}}{\eta^{\text{MP}} + \eta^{\text{RP}}}$$

The second one aims to select the predicted values with the lowest uncertainties:

$$\rho(\eta^{\text{MP}}, \eta^{\text{RP}}) = \begin{cases} 1, & \text{if } \eta^{\text{MP}} < \eta^{\text{RP}} \\ 0, & \text{otherwise} \end{cases}$$

3.5 Baseline Selection

To compare the relative performance of our proposed models with the existing approaches to air pollution forecast, we include both the state-of-the-art deep-learning models and the time-series models (see Section **??** below). To evaluate the relative performance of Bayesian method and domain-specific knowledge regularization to deep-learning, our baseline deep-learning models adopt the same network structure and parameters as the proposed Bayesian deep-learning models. Given that Lasso regression model, which accounts for the feature selection and regularization during model fitting, has achieved a better performance in air pollution forecast when compared to other regression models such as ARIMA [13], we include it as one of our baseline models. State-of-the-art methods for air pollution forecast have taken into account of the spatial and temporal dimensions by means of fusion techniques [13], [28], . In this study, our proposed models have focused more on the temporal dimensions of air pollution forecast, in particular, how one-time and recursive time-series forecast strategies vary in the long-term (2-day) forecast performance. In addition, we directly compare the performance reported in [28] since the same datasets are used. The baseline and proposed models are defined as follows.

Our proposed Bayesian deep-learning models:

- **BayesAir (OP)**: Two Bayesian RNN models with GRU unit are used to predict the air pollution values in the next 48 hours in one shot, based on the concatenated hidden states from the two Bayesian RNNs. The two Bayesian RNNs take the historical air quality data and the weather forecast data as the inputs, respectively.
- **BayesAir (RP)**: One sequence to sequence (seq2seq) Bayesian RNN model with GRU unit and an attention layer, utilizing weather forecast data as the inputs for the Bayesian seq2seq decoder.
- **DBayesAir (OP)**: BayesAir (OP) model with a regularization term in the loss function, based on the domain-specific knowledge about the strong statistical relationship between $PM_{2.5}$ and PM_{10} pollution.
- **DBayesAir (RP)**: BayesAir (RP) model with a regularization term in the loss function, based on the domain-specific knowledge about the strong statistical relationship between $PM_{2.5}$ and PM_{10} pollution.

State-of-the-art baseline models:

- **LassoAir (OP)**: 48 Lasso regression models are used to predict the pollutant values in the next 48 hours, respectively, taking the historical air quality data and the weather forecast data as the inputs.
- **LassoAir (RP)**: One Lasso regression model is used to predict the air pollutant values in the next hour, taking the air quality data in the previous hour and the weather forecast data in the next hour as the inputs. This is repeated for the next 48 hours.
- **TradAir (OP)**: Two RNN models with GRU unit [58] are used to predict the air pollution values in the next 48 hours in one shot, based on the concatenated hidden states from the two RNNs. The two RNNs take the historical air quality data and the weather forecast data as the inputs, respectively.
- **TradAir (RP)**: One seq2seq RNN model with GRU unit and an attention layer, utilizing weather forecast data as the inputs for the seq2seq decoder [27].

3.6 Experimental Settings

We use five random 80/20 splits of data dated from January 2017 to April 2018 as the training data and the validation data. We select "last month data" from May 2018 as the independent testing data (i.e., the testing data was not used during training and validation). We use data of the past 72 hours, that are combined with the weather forecast of the next 48 hours, to predict air pollution concentrations of the next 48 hours. To measure the relative error rate and compare model performance between the two cities, Symmetric Mean Absolute Percentage Error (SMAPE) is used as the metric for model evaluation.

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|A_t - F_t|}{(A_t + F_t)/2}$$

where F_t is the forecast value, A_t is the actual value, and n is the number of samples. For each hour, we averaged the SMAPE of each pollutant at each station. The model training and evaluation procedures are listed as follows. First, for each of the prediction strategies, traditional statistical models (LassoAir) and traditional deep-learning models (TradAir) are selected as baselines. For all deep-learning models (TradAir, BayesAir, and DBayesAir), the same settings are used, including the learning rate (0.001), the batch size (64), the number of recurrent layers (3), and the number of hidden units (256). For Bayesian deep-learning models, the network weight priors are selected according to the settings in [43]. In order to obtain a reasonable prediction distribution from the Bayesian deep-learning models, we set the number of simulations to 100 [38]. Then, we fine-tune the models to select the best hyper-parameters, and choose the models with the lowest SMAPE of the validation set as our final models. Early stop will be adopted if the SMAPE on the validation set started increasing. More specifically, the setting of hyper-parameters is listed as follows. For the traditional and the Bayesian deep learning models (TradAir and BayesAir), two hyper-parameters are fine-tuned, namely, λ_1 and λ_2, with the aim to account for the relative weights of the prediction errors in the loss function for $PM_{2.5}$ and PM_{10}. Three pairs of λ_1, λ_2 are tested, including, (0.4, 0.6), (0.5, 0.5), and (0.6, 0.4). For the Bayesian deep learning model with domain-specific knowledge regularization (DBayesAir), in addition to λ_1 and λ_2, λ_3 is used to account for the relative weights of the domain-specific constraint term. When fine-tuning the domain-specific Bayesian deep learning model, in order to capture the impact of λ_3, we set λ_1 and λ_2 to the best λ_1 and λ_2 obtained from the corresponding Bayesian deep learning model. λ_3 is set as -0.01, -0.1, or -1.0. Finally, we evaluate the models on the test set. This is repeated on the five data splits, and the mean and the standard deviation of our model performance are reported.

4 Results

The SMAPEs of different prediction strategies are shown in Table 2 and Table 3. The SMAPEs of different fusion strategies are shown in Table 4. Improvements in deep-learning

TABLE 2
SMAPE (%) of models using different prediction strategies (Beijing, China)

Prediction strategy	One-time prediction			Recursive prediction		
Model / Period	**1-24h**	**25-48h**	**Overall**	**1-24h**	**25-48h**	**Overall**
LassoAir	57.4 (1.2)	60.6 (1.1)	59.0 (1.1)	54.2 (1.0)	55.9 (1.0)	55.1 (1.0)
TradAir	50.9 (1.4)	53.4 (0.6)	52.1 (1.0)	47.1 (1.6)	55.7 (3.5)	51.4 (2.6)
BayesAir	50.4 (0.5)	53.0 (0.2)	51.7 (0.4)	46.2 (0.6)	54.0 (1.2)	50.1 (0.9)
DBayesAir	**50.1** (0.6)	**52.9** (0.3)	**51.5** (0.4)	**45.8** (0.7)	**53.2** (0.6)	**49.5** (0.6)
Notes						
1. For each column, lowest error is in boldface. 2. For each column, standard deviation is shown in parentheses.						

TABLE 3
SMAPE (%) of models using different prediction strategies (London, the UK)

Prediction strategy	One-time prediction			Recursive prediction		
Model / Period	**1-24h**	**25-48h**	**Overall**	**1-24h**	**25-48h**	**Overall**
LassoAir	**35.0** (0.4)	44.0 (0.7)	39.5 (0.6)	42.0 (0.3)	42.8 (0.3)	42.4 (0.3)
TradAir	40.4 (3.6)	41.0 (3.7)	40.7 (3.6)	40.6 (2.4)	43.0 (3.0)	41.8 (2.7)
BayesAir	38.0 (1.3)	38.7 (1.9)	38.4 (1.6)	35.4 (0.8)	43.1 (1.5)	39.2 (1.2)
DBayesAir	37.5 (1.6)	**38.6** (1.0)	**38.1** (1.3)	**33.6** (0.7)	**39.7** (1.3)	**36.6** (1.0)
Notes						
1. For each column, lowest error is in boldface. 2. For each column, standard deviation is shown in parentheses.						

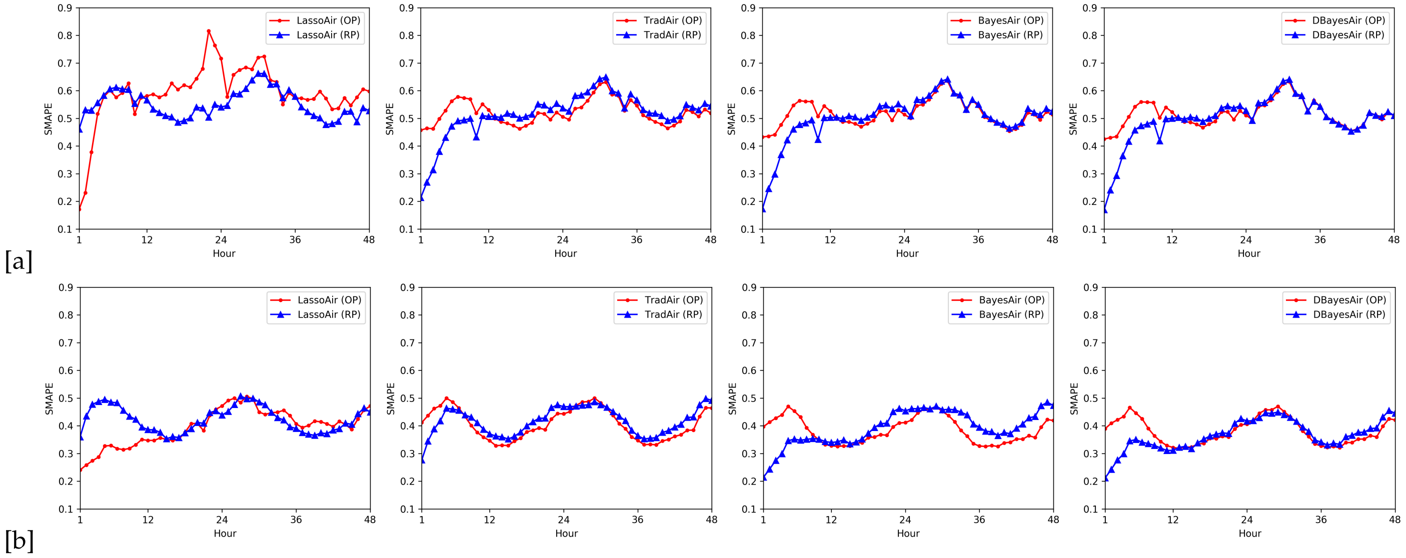


Fig. 3. 48-h SMAPE trends for models with one-time prediction (OP) and recursive prediction (RP) strategies: (a) for Beijing; (b) for London

performance using the Bayesian method and the domain-specific knowledge are shown in Table 5. The following paragraphs will compare our proposed models with the state-of-the-art baseline models and [28]. Implications of the results will be highlighted in Section 5.

First, for prediction errors across the first day and the second day, we find that the Bayesian models can outperform their traditional counterparts across most of the time periods, using either one-time prediction strategy or recursive prediction strategy. Compared to the baseline deep-learning models, incorporating the Bayesian method can reduce the prediction errors by a maximum of 2.5% and 6.2% for Beijing and London, respectively (see Table 5). More importantly, results have shown that incorporating domain-specific knowledge into the training of Bayesian models can further improve the models' performance. Among all models listed in Table 2 and Table 5, domain-specific Bayesian deep-learning models can achieve the lowest overall SMAPEs on average. As compared to Bayesian models, adding domain-specific knowledge can further reduce the prediction errors by a maximum of 1.2% and 6.2% for Beijing and London, respectively (see Table 5).

Second, for all models shown in Table 2 and Table 3, the prediction errors of individual hours can be found in Figure 3. Results indicate that recursive prediction strategy significantly outperforms one-time prediction strategy for

TABLE 4
SMAPE (%) of models using different fusion strategies

City	Beijing, China			London, the UK		
Model (Fusion Strategy) / Period	**1-24h**	**25-48h**	**Overall**	**1-24h**	**25-48h**	**Overall**
LassoAir (Average-based prediction fusion)	50.9 (0.5)	53.2 (0.4)	52.0 (0.5)	35.3 (0.2)	40.6 (0.3)	37.9 (0.3)
TradAir (Average-based prediction fusion)	47.4 (1.1)	52.8 (0.7)	50.1 (0.9)	39.0 (2.4)	40.8 (2.7)	39.9 (2.6)
BayesAir (Average-uncertainty-based prediction fusion)	48.7 (0.5)	53.1 (0.5)	50.9 (0.5)	35.4 (1.1)	40.0 (1.7)	37.7 (1.4)
BayesAir (Lowest-uncertainty-based prediction fusion)	46.1 (0.7)	53.1 (0.6)	49.6 (0.7)	34.9 (1.0)	40.5 (1.4)	37.7 (1.2)
DBayesAir (Average-uncertainty-based prediction fusion)	48.4 (0.3)	52.8 (0.3)	50.6 (0.3)	34.5 (0.8)	**38.6** (0.8)	36.6 (0.8)
DBayesAir (Lowest-uncertainty-based prediction fusion)	**45.6** (0.6)	**52.7** (0.5)	**49.2** (0.6)	**33.5** (0.7)	38.7 (0.7)	**36.1** (0.7)
Notes						
1. For each column, lowest error is in boldface. 2. For each column, standard deviation is shown in parentheses.						

TABLE 5
Relative improvement of overall SMAPE (%) with the Bayesian method and the domain-specific knowledge

City	Beijing, China			London, the UK		
Model / Prediction Strategy	**OP**	**RP**	**Fusion**	**OP**	**RP**	**Fusion**
TradAir	52.1	51.4	50.1	40.7	41.8	39.9
BayesAir	51.7 (0.8%)	50.1 (2.5%)	49.6 (1.0%)	38.4 (5.7%)	39.2 (6.2%)	37.7 (5.5%)
DBayesAir	**51.5** (1.2%)	**49.5** (3.7%)	**49.2** (1.8%)	**38.1** (6.4%)	**36.6** (12.4%)	**36.1** (9.5%)
Notes						
1. Average overall SMAPEs in Table 2, 3, and 4 are used. Fusion strategy is the lowest-uncertainty-based prediction fusion. 2. For each column, lowest error is in boldface. 3. For each column, relative improvement compared to the baseline (TradAir) is shown in parentheses.						

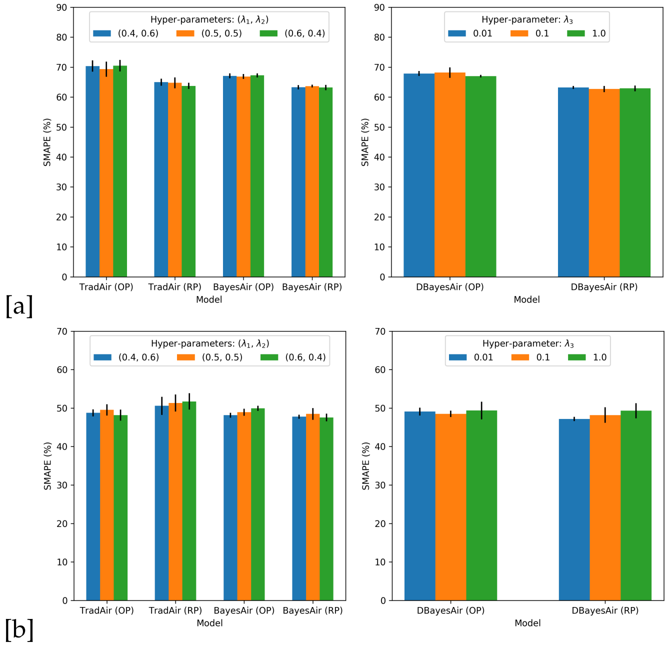


Fig. 4. The influence of different hyper-parameters on the validation set: (a) for Beijing; (b) for London

initial-hours prediction (from the 1st hour to the 12th hour), and one-time prediction strategy generally performs better for later-hours prediction (from the 24th hour to the 48th hour). This trend is more significant in the case of London than that of Beijing.

Moreover, we compare the hybrid models where predictions are fused by different prediction strategies (see Table 4). For the Bayesian hybrid models, we report the results derived from the two fusing strategies listed in Section 3.3. For the traditional hybrid models, prediction results are simply averaged since no weighting measures could be adopted. Results show that hybrid models can generally achieve lower SMAPEs compared to the corresponding single models without prediction fusion. Also, domain-specific knowledge regularization can improve the performance of hybrid Bayesian deep-learning models. Among all hybrid models, domain-specific Bayesian deep-learning models can achieve the lowest overall SMAPEs. The lowest overall SMAPEs of hybrid models are 49.2% and 36.1% for Beijing and London, respectively. The corresponding error reductions compared to the traditional hybrid deep-learning baseline models are 1.8% and 9.5% for Beijing and London, respectively.

Further, we compare our proposed methods with state-of-the-art models in [28] since the same datasets have been used. As shown in Table 1 in [28], for Beijing and London, the average overall SMAPE during the testing period is 39.2% and 43.2%, respectively for the 1-48-hr forecast and the 24-48-hr forecast. As shown in Table 4, the best performance for the 1-48-hr forecast is 49.2% and 36.1% for Beijing and London, respectively, whilst the best performance for the 24-48-hr forecast is 52.7% and 38.6% for Beijing and London, respectively.

Finally, we compare the influence of different hyper-parameters during the model training. The mean and the standard deviation of the model performance (based on the validation set) across different hyper-parameters are shown in Figure 4. On average, for the same hyper-parameters, the proposed models that adopt the recursive prediction strategy tend to have lower error rates (consistent with the results from the test set). Regarding the relative weights of $PM_{2.5}$ and PM_{10} (i.e., λ_1 and λ_2), no consistent patterns can be found, suggesting that the characteristics of $PM_{2.5}$ and PM_{10} pollution concentrations could be different across different training/validation splits. In terms of the domain-specific knowledge constraint (i.e., λ_3), a low or moderate λ_3 (0.01 or 0.1) can better improve the performance of the proposed models most of time. However, for a one-time prediction strategy based on the Beijing data, a high λ_3 (1.0) can lead to a better performance, suggesting that the effect of such regularization could be dependent on data quality and prediction strategy.

5 Discussion

We have proposed a domain-specific Bayesian deep-learning model for long-term (2-day) air pollution forecast, using Beijing and London as the case studies. Although Bayesian deep-learning has the potential to process limited and noisy data [36], such model has not been tailored to long-term air pollution forecast. The novelties of our proposed Bayesian deep-learning approach include the following: First, we integrate a domain-specific knowledge taking into account the strong statistical relationship between $PM_{2.5}$ and PM_{10} for regularization; Second, we include an attention layer capable of capturing an influential historical feature, the recursive temporal pattern of air quality data for pollution prediction; Third, deviating from state-of-the-art fusion strategies [13], [28], we utilize model and input prediction uncertainties generated from different forecast strategies, to provide uncertainty-based prediction fusion. Experimental results show that our proposed methods have achieved better results when compared to the traditional time-series and deep-learning baseline models (i.e., LassoAir and TradAir). The following paragraphs will highlight how theses novelties are linked to the improvements of long-term air pollution forecast as shown in Section **??**, and what can be done to further improve our proposed approach.

First, domain-specific knowledge taking into account of the temporal-spatial nature of air pollution data can improve further the performance of our model over long-term forecast. More specifically, the strong correlation between $PM_{2.5}$ and PM_{10} values is used as a regularization term in the loss function. Such regularization procedure can produce more accurate predictions over the second day (see Table 2 and 3). In addition, to capture the periodic patterns of air pollution, we incorporate the temporal trend (such as peak/non-peak hours or weekday/weekend) into our deep-learning models, which serve an attention layer to better capture the recurrent temporal patterns (recurrent daily, weekly, monthly, and seasonal patterns). We also include the station IDs to capture the spatial characteristics, though more urban morphology information such as building height and density are yet to be integrated. Overall, though the Bayesian deep-learning models in theory should automatically capture better any underlying domain-specific knowledge based on the training data, such as the correlation between $PM_{2.5}$ and PM_{10} values or the recurrent temporal patterns of $PM_{2.5}$ and PM_{10} concentrations, our tailored-made domain-specific learning procedure generally performs better in practice. This suggests that incorporating any influential, high-saliency domain-specific knowledge to our model, such as the high temporal- or high spatial-correlation feature of the air pollution data, can further improve our models prediction accuracy.

Second, different forecast strategies can be exploited to further improve our model's performance over long-term forecast. For any air pollution forecast of the next 48 hours, recursive prediction strategy tends to perform better over the initial hours (the first 12 hours), while one-time prediction strategy tends to improve gradually over the later hours (the last 24 hours), due to the following reasons. First, recursive prediction can better capture the temporal correlation across each individual hour, thus resulting in more accurate forecast during the initial hours. Second, one-time prediction can achieve lower error rates during the later hours, as predictions are less likely affected by error accumulation. Although the error trends of the two forecast strategies are consistent across all deep-learning models, the Lasso regression models (see Table 2 and Table 3) have exhibited a different trend. This may be explained by their difference in model structures: First, for one-time prediction, the temporal correlation has been ignored because 48 Lasso models have been trained to predict air qualities of the next 48 hours individually; Second, for recursive prediction, the Lasso model only uses the previous-hour data for predicting the next-hour air pollution values. Moreover, it is noted that some Lasso regression models (LassoAir using one-time prediction strategy) can even outperform the deep-learning models for prediction in the short term (in particular, the first 12-h) in London. This finding is consistent with some previous studies in air quality modelling. For example, a European study finds that machine learning models cannot add benefits to the performance of air quality prediction, especially when (1) the non-linear relationships between air pollution and other predictors are not significant and (2) the variation of air pollution concentrations is low [65]. The results of Lasso regression models have two implications for deep learning-based air pollution forecast, especially for short-term forecast in London. First, as compared to our proposed deep learning models, one advantage of Lasso regression is that the most important input variables are selected during the training process (i.e., some regression coefficients can be zero). Therefore, instead of using all the data as the input, feature analysis and selection could be performed when training the deep learning models. Second, Lasso regularization can be linked to Bayesian regression models with Laplace priors [66]. Therefore, instead of using Gaussian priors, more informative priors that better characterize the air quality data could be examined in the Bayesian deep learning models. Furthermore, for our proposed deep-learning models based on one-time and recursive strategies, each performs better in air pollution forecast over the longer-term, given that the temporal correlation of the predictions across different periods have been taken into account in the network structures.

Furthermore, uncertainty-based prediction fusion can enable more accurate forecast in the short and long term (see Table 4 and 5). More specifically, motivated by the characteristics of one-time prediction strategy and recursive prediction strategy, hybrid models are also used so that the hourly predictions derived from these two strategies can automatically complement each other based on their associated uncertainty measures. As shown in Table 5, we have found that hybrid models generally perform better than models based on single-prediction strategy; while our Bayesian hybrid models (BayesAir and DBayesAir) have lower error rates when compared to the baseline hybrid models (TradAir and LassoAir). This suggests that different forecast strategies can complement each other over long-term prediction, and uncertainty-based prediction fusion strategies tend to give a higher accuracy. Further, the lowest uncertainty-based strategy performs better across the Beijing and the London dataset in general, while the average uncertainty-based strategy performs slightly better on the second day prediction across the London dataset. To fully capitalize on the strengths of each forecasting strategy across different periods, it would be important to select the most trusted prediction (prediction with the lowest uncertainty). The uncertainty-based strategy can be adopted if the data are more predictable and the second day prediction is of more interests.

In general, as shown in Table 5, air pollution forecast for London achieves a higher accuracy as compared to Beijing, suggesting that the London data is more regular and predictable than the Beijing data, due to the following reasons. First, more missing values were observed in the Beijing data, and errors due to missing data imputation could lead to lower data quality and result in lower prediction accuracy. More sophisticated missing data recovery capturing the temporal data correlation characteristics can be used in future [67]. Similar to the iterative back-propagation method proposed by [11], an iterative Bayesian back-propagation approach can be used to integrate missing data recovery strategies with our model to recover long-term missing air pollution data and improve the quality of our data input and model performance over longer-term pollution forecast. Second, during the test period, some sudden changes in air quality due to sandstorms in Beijing have been observed. This may lead to lower prediction accuracy because the ad-hoc air pollution patterns may not be learned automatically during the training process.

Our study has demonstrated the feasibility of using a domain-specific Bayesian model for more accurate long-term (48-hr) air pollution forecast. When compared with the state-of-the-art models, our proposed approach may achieve an even better performance by adding certain spatial factors (such as nearby air pollution concentrations, weather conditions, street canyon effects, etc.). Based on the separate feature importance analysis conducted by [28] (see Table 2 in [28]), the SMAPE is respectively 39.5% and 44.4%, with the model taking into account (1) both the spatial and the temporal features, or (2) only the temporal features. This implies that our proposed deep-learning models can be further improved if important spatio-temporal features of air quality dynamics are taken into account in future modelling.

One major limitation of our work is the lack of important spatial or temporal features that determine the occurrence of air pollution. Although the strong statistical relationship between $PM_{2.5}$ and PM_{10} has been incorporated into our models as one type of domain-specific knowledge, more critical features that contribute to air pollution in the urban areas should be taken in account in our future study. Based on salient score analysis and Granger causality test, previous literature suggest that traffic speed/density, street canyon effects, local air quality and weather measurements from adjacent stations, and regional environmental conditions, are important factors for air quality modelling [9], [11], [68]. These critical temporal or spatial features of air pollution, when combined with feature selection and neural attention modelling, can be added to our domain-specific air pollution forecast model to further improve the performance of our existing model. Moreover, we can use the saliency scores to determine which factors should be taken into account for air pollution forecast, and how domain-specific knowledge can help our deep-learning model attend to the most influential part instead of the full dataset. This can further improve the interpretability of our proposed model. Another limitation of our work is about the distribution assumption of the network weights in Bayesian deep-learning models. Previous Bayesian deep-learning studies often take the assumption of a Gaussian distribution for network weights and prediction outputs, but this may not hold true across all types of data especially the environmental data [69]. Future work may investigate other distribution assumptions.

6 Conclusion

Providing air pollution forecasts with uncertainty measures and domain-specific knowledge-integration has been largely overlooked in previous data-driven deep-learning approaches. This study investigates the air pollution forecast problem through a Bayesian deep-learning approach with domain-specific knowledge. Using Beijing, China and London, the UK as case studies, experimental results show that on average, incorporating Bayesian methods and domain-specific knowledge can reduce the prediction errors by a maximum of 3.7% and 12.4% for Beijing and London, respectively. Moreover, hybrid Bayesian models are able to achieve the lowest prediction errors and the best ones can improve the traditional hybrid baselines by 1.8% to 9.5% for Beijing and London, respectively. Our results highlight the importance of including domain-specific knowledge and suggest that introducing Bayesian techniques not only improves the accuracy of traditional deep-learning models, but also allows fusing of different forecast strategies to provide more accurate results. In future, our proposed model will include more influential factors for air pollution forecast, such as the spatio-temporal data from nearby monitoring stations. We will also integrate more influential domain-specific knowledge (e.g., by incorporating building height and density that capture the spatial effects of air pollution nearby the roadside stations) and evaluate the relative performance of air pollution forecast.

Acknowledgments

We gratefully acknowledge the provision of air quality and weather data provided by ACM SIGKDD Cup 2018, as well as the London weather data provided by the Met Office, the

UK. This research is supported in part by the Theme-based Research Scheme of the Research Grants Council of Hong Kong, under Grant No. T41-709/17-N.

References

[1] D. Y. Pui, S.-C. Chen, and Z. Zuo, "$\mathrm{PM}_{2.5}$ in China: Measurements, sources, visibility and health effects, and mitigation," *Particuology*, vol. 13, pp. 1–26, 2014.

[2] Y. Li, D. Guan, S. Tao, X. Wang, and K. He, "A review of air pollution impact on subjective well-being: Survey versus visual psychophysics," *Journal of Cleaner Production*, vol. 184, pp. 959–968, 2018.

[3] Y. Han, J. C. K. Lam, and V. O. K. Li, "A Bayesian LSTM model to evaluate the effects of air pollution control regulations in China," in *2018 IEEE Big Data Workshop on Big Data and AI for Air Quality Estimation, Forecasting, and Health Advice*. IEEE, 2018, pp. 4465–4468.

[4] D. W. Dockery and C. A. Pope, "Acute respiratory effects of particulate air pollution," *Annual Review of Public Health*, vol. 15, no. 1, pp. 107–132, 1994.

[5] J. Duan, Y. Chen, W. Fang, and Z. Su, "Characteristics and relationship of PM, PM_{10}, $\mathrm{PM}_{2.5}$ concentration in a polluted city in Northern China," *Procedia Engineering*, vol. 102, pp. 1150–1155, 2015.

[6] X. Yang, L. Jiang, W. Zhao, Q. Xiong, W. Zhao, and X. Yan, "Comparison of ground-based $\mathrm{PM}_{2.5}$ and PM_{10} concentrations in China, India, and the US," *International Journal of Environmental Research and Public Health*, vol. 15, no. 7, p. 1382, 2018.

[7] L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: An overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, p. 780, 2018.

[8] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2013, pp. 1436–1444.

[9] J. Y. Zhu, C. Sun, and V. O. K. Li, "An extended spatio-temporal Granger causality model for air quality estimation with heterogeneous urban big data," *IEEE Transactions on Big Data*, vol. 3, no. 3, pp. 307–319, 2017.

[10] B. T. Ong, K. Sugiura, and K. Zettsu, "Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $\mathrm{PM}_{2.5}$," *Neural Computing and Applications*, vol. 27, no. 6, pp. 1553–1566, 2016.

[11] V. O. K. Li, J. C. K. Lam, Y. Chen, and J. Gu, "Deep learning model to estimate air pollution using M-BP to fill in missing proxy urban data," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[12] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation," *Environmental Pollution*, vol. 231, pp. 997–1004, 2017.

[13] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep distributed fusion network for air quality prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 965–973.

[14] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting air quality time series using deep learning," *Journal of the Air & Waste Management Association*, vol. 68, no. 8, pp. 866–886, 2018.

[15] B. Wang, Z. Yan, J. Lu, G. Zhang, and T. Li, "Deep multi-task learning for air quality prediction," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 93–103.

[16] C.-J. Huang and P.-H. Kuo, "A deep CNN-LSTM model for particulate matter ($\mathrm{PM}_{2.5}$) forecasting in smart cities," *Sensors*, vol. 18, no. 7, p. 2220, 2018.

[17] Z. Qi, T. Wang, G. Song, W. Hu, X. Li, and Z. Zhang, "Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2285–2297, 2018.

[18] W. Cheng, Y. Shen, Y. Zhu, and L. Huang, "A neural attention model for urban air quality inference: Learning the weights of monitoring stations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[19] Y. Zhou, F.-J. Chang, L.-C. Chang, I.-F. Kao, and Y.-S. Wang, "Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts," *Journal of Cleaner Production*, vol. 209, pp. 134–145, 2019.

[20] Y. Qi, Q. Li, H. Karimian, and D. Liu, "A hybrid model for spatiotemporal forecasting of $\mathrm{PM}_{2.5}$ based on graph convolutional neural network and long short-term memory," *Science of The Total Environment*, 2019.

[21] W. Tong, L. Li, X. Zhou, A. Hamilton, and K. Zhang, "Deep learning $\mathrm{PM}_{2.5}$ concentrations with bidirectional LSTM RNN," *Air Quality, Atmosphere & Health*, pp. 1–13, 2019.

[22] Q. Zhang, V. O. K. Li, J. C. K. Lam, and Y. Han, "Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution forecast."

[23] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, p. 195, 2019.

[24] J. Wang, B. Zhao, S. Wang, F. Yang, J. Xing, L. Morawska, A. Ding, M. Kulmala, V.-M. Kerminen, J. Kujansuu *et al.*, "Particulate matter pollution over China and the effects of control policies," *Science of The Total Environment*, vol. 584, pp. 426–447, 2017.

[25] X.-H. Liu, Y. Zhang, S.-H. Cheng, J. Xing, Q. Zhang, D. G. Streets, C. Jang, W.-X. Wang, and J.-M. Hao, "Understanding of regional air pollution over China using CMAQ, part I performance evaluation and seasonal variation," *Atmospheric Environment*, vol. 44, no. 20, pp. 2415–2426, 2010.

[26] X. Li, Y. Qiao, J. Zhu, L. Shi, and Y. Wang, "The APEC Blue endeavor: Causal effects of air pollution regulation on air quality in China," *Journal of Cleaner Production*, vol. 168, pp. 1381–1388, 2017.

[27] B. Liu, S. Yan, J. Li, G. Qu, Y. Li, J. Lang, and R. Gu, "An attention-based air quality forecasting method," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 728–733.

[28] Z. Luo, J. Huang, K. Hu, X. Li, and P. Zhang, "Accuair: Winning solution to air quality prediction for KDD Cup 2018," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1842–1850.

[29] H. Wang and D.-Y. Yeung, "Towards Bayesian deep learning: A framework and some existing methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3395–3408, 2016.

[30] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[31] J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, "Deep neural networks as Gaussian processes," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=B1EA-M-0Z

[32] A. Damianou and N. Lawrence, "Deep Gaussian processes," in *Artificial Intelligence and Statistics*, 2013, pp. 207–215.

[33] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Artificial Intelligence and Statistics*, 2016, pp. 370–378.

[34] M. Al-Shedivat, A. G. Wilson, Y. Saatchi, Z. Hu, and E. P. Xing, "Learning scalable deep kernels with recurrent structure," *Journal of Machine Learning Research*, vol. 18, no. 82, pp. 1–37, 2017.

[35] M. Garnelo, D. Rosenbaum, C. J. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. J. Rezende, and S. Eslami, "Conditional neural processes," *arXiv preprint arXiv:1807.01613*, 2018.

[36] Y. Gal, "Uncertainty in deep learning," *University of Cambridge*, 2016.

[37] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 1019–1027.

[38] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2016, pp. 1050–1059.

[39] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[40] J. M. Hernández-Lobato and R. Adams, "Probabilistic backpropagation for scalable learning of Bayesian neural networks," in *International Conference on Machine Learning*, 2015, pp. 1861–1869.

[41] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning*, 2014, pp. 1278–1286.

[42] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*, 2015, pp. 1613–1622.

[43] M. Fortunato, C. Blundell, and O. Vinyals, “Bayesian recurrent neural networks,” *arXiv preprint arXiv:1704.02798*, 2017.

[44] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, p. 436, 2015.

[45] Ç. Gülçehre and Y. Bengio, “Knowledge matters: Importance of prior information for optimization,” *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 226–257, 2016.

[46] Z. Hu, Z. Yang, R. Salakhutdinov, and E. Xing, “Deep neural networks with massive learned knowledge,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1670–1679.

[47] J. Zhang, Y. Liu, H. Luan, J. Xu, and M. Sun, “Prior knowledge integration for neural machine translation using posterior regularization,” *arXiv preprint arXiv:1811.01100*, 2018.

[48] X. Ding, Y. Luo, Q. Li, Y. Cheng, G. Cai, R. Munnoch, D. Xue, Q. Yu, X. Zheng, and B. Wang, “Prior knowledge-based deep learning method for indoor object recognition and application,” *Systems Science & Control Engineering*, vol. 6, no. 1, pp. 249–257, 2018.

[49] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, “Risk prediction on electronic health records with prior medical knowledge,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1910–1919.

[50] A. Atanov, A. Ashukha, K. Struminsky, D. Vetrov, and M. Welling, “The deep weight prior,” in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=ByGuynAct7

[51] C. Du, C. Du, L. Huang, and H. He, “Reconstructing perceived images from human brain activities with Bayesian deep multiview learning,” *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[52] M. Diligenti, S. Roychowdhury, and M. Gori, “Integrating prior knowledge into deep learning,” in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 920–923.

[53] M. Raissi, P. Perdikaris, and G. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[54] “KDD CUP of Fresh Air,” https://biendata.com/competition/kdd_2018/, 2018.

[55] Met Office, “MIDAS: UK hourly weather observation data,” https://catalogue.ceda.ac.uk/uuid/916ac4bbc46f7685ae9a5e10451bae7c, 2006.

[56] S. v. Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in R,” *Journal of Statistical Software*, pp. 1–68, 2010.

[57] I. Fox, L. Ang, M. Jaiswal, R. Pop-Busui, and J. Wiens, “Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1387–1395.

[58] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.

[59] V. Athira, P. Geetha, R. Vinayakumar, and K. Soman, “DeepAirNet: Applying recurrent networks for air quality prediction,” *Procedia Computer Science*, vol. 132, pp. 1394–1403, 2018.

[60] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[61] J.-F. Wang, M.-G. Hu, C.-D. Xu, G. Christakos, and Y. Zhao, “Estimation of citywide air pollution in Beijing,” *PloS One*, vol. 8, no. 1, p. e53400, 2013.

[62] H. Hoos and K. Leyton-Brown, “An efficient approach for assessing hyperparameter importance,” in *International Conference on Machine Learning*, 2014, pp. 754–762.

[63] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5574–5584.

[64] L. Zhu and N. Laptev, “Deep and confident prediction for time series at Uber,” in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017, pp. 103–110.

[65] J. Chen, K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, A. van Donkelaar, U. A. Hvidtfeldt, K. Katsouyanni *et al.*, “A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide,” *Environment International*, vol. 130, p. 104934, 2019.

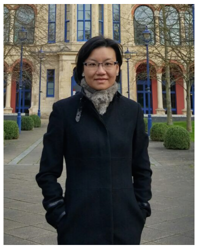
[66] T. Park and G. Casella, “The Bayesian Lasso,” *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[67] Y. Yu, J. James, V. O. K. Li, and J. C. K. Lam, “Low-rank singular value thresholding for recovering missing air quality data,” in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 508–513.

[68] J. Y. Zhu, C. Zhang, H. Zhang, S. Zhi, V. O. K. Li, J. Han, and Y. Zheng, “pg-causality: Identifying spatiotemporal causal pathways for air pollutants with urban big data,” *IEEE Transactions on Big Data*, vol. 4, no. 4, pp. 571–585, 2017.

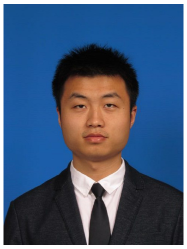
[69] T. Vandal, E. Kodra, J. Dy, S. Ganguly, R. Nemani, and A. R. Ganguly, “Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2377–2386.

Yang Han Yang Han received his MSc degree in Computer Science with distinction from the University of Hong Kong (HKU) in 2014. He obtained his bachelor’s degree from the Department of Information Systems, Beihang University, Beijing, China, in 2013. Currently he is undertaking PhD in the Department of Electrical & Electronic Engineering, HKU. His recent work focuses on spatio-temporal data analysis and its applications in environmental pollution and policy studies in China. He has published on Environmental Science and Policy.



Jacqueline C.K. Lam Jacqueline C.K. Lam is Associate Professor in the Department of Electrical and Electronic Engineering, the University of Hong Kong and Co-Director of the HKU-Cambridge Clean Energy and Environment Research Platform, and of the HKU-Cambridge AI-WiSe Research Platform. She was the Hughes Hall Visiting Fellow before taking up the Visiting Senior Research Fellow and Associate Researcher in Energy Policy Research Group, Judge Business School, the University of Cambridge. Her research studies clean energy and environment using interdisciplinary approaches, with a special focus on China and the UK. Her recent research focuses on the use of big data and machine learning techniques to study personalized air pollution monitoring and health management. Jacqueline has received three times the research grants awarded by the Research Grants Council, HKSAR Government, from 2011-2017. The funded amount totalled USD 7.8M in PI or Co-PI capacity. Her recent research study with Mr. Yang Han and Prof. Victor OK Li, on $PM_{2.5}$ pollution and environmental inequality in Hong Kong, has been published in Environmental Science and Policy, and widely covered by more than 30 local and overseas newspapers and television media. She is currently a visiting fellow at MIT CEEPR.

Victor O.K. Li Victor O.K. Li received SB, SM, EE and ScD degrees in Electrical Engineering and Computer Science from MIT. Prof. Li is Chair of Information Engineering and Cheng Yu-Tung Professor in Sustainable Development at the Department of Electrical & Electronic Engineering (EEE) at the University of Hong Kong. He is the Director of the HKU-Cambridge Clean Energy and Environment Research Platform, and of the HKU-Cambridge AI to Advance Well-being and Society Research Platform, which are interdisciplinary collaborations with Cambridge University. He was the Head of EEE, Assoc. Dean (Research) of Engineering and Managing Director of Versitech Ltd. He serves on the board of Sunevision Holdings Ltd., listed on the Hong Kong Stock Exchange and co-founded Fano Labs Ltd., an AI company with his PhD student. Previously, he was Professor of Electrical Engineering at the University of Southern California (USC), Los Angeles, California, USA, and Director of the USC Communication Sciences Institute. His research interests include big data, AI, optimization techniques, and interdisciplinary clean energy and environment studies. In Jan 2018, he was awarded a USD 6.3M RGC Theme-based Research Project to develop deep-learning techniques for personalized and smart air pollution monitoring and health management. Sought by government, industry, and academic organizations, he has lectured and consulted extensively internationally. He has received numerous awards, including the PRC Ministry of Education Changjiang Chair Professorship at Tsinghua University, the UK Royal Academy of Engineering Senior Visiting Fellowship in Communications, the Croucher Foundation Senior Research Fellowship, and the Order of the Bronze Bauhinia Star, Government of the HKSAR. He is a Fellow of the Hong Kong Academy of Engineering Sciences, the IEEE, the IAE, and the HKIE.



Qi Zhang Qi Zhang received the BE degree in electronic engineering and BEc degree in economics from Tsinghua University, Beijing, China, in 2017. He is working towards the PhD degree in the Department of Electrical & Electronic Engineering, the University of Hong Kong. He is a holder of the Hong Kong PhD Fellowship. His research interests include deep-learning and its applications on air pollution, spatio-temporal data analysis, and urban computing.