atmosphere

MDPI

Article

A Gaussian Process Method with Uncertainty Quantification for Air Quality Monitoring

Peng Wang [1,*], Lyudmila Mihaylova [2], Rohit Chakraborty [3], Said Munir [3], Martin Mayfield [3], Khan Alam [4], Muhammad Fahim Khokhar [5], Zhengkai Zheng [6], Chengxi Jiang [7] and Hui Fang [8]

1 Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BH, UK
2 Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield S10 2TN, UK; l.s.mihaylova@sheffield.ac.uk
3 Department of Civil and Structural Engineering, The University of Sheffield, Sheffield S10 2TN, UK; rohit.chakraborty@sheffield.ac.uk (R.C.); smunir2@sheffield.ac.uk (S.M.); martin.mayfield@sheffield.ac.uk (M.M.)
4 Department of Physics, University of Peshawar, Peshawar 25120, Pakistan; khanalam@uop.edu.pk
5 Institute of Environmental Sciences and Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan; fahim.khokhar@iese.nust.edu.pk
6 Yueqing Xinshou Agricultural Development Co., Ltd., Yueqing 325604, China; zhengzhengkai@ls.zjcoon.cn
7 College of Electrical and Electronic Engineering, Wenzhou University, Wenzhou 325035, China; 20190408@wzu.edu.cn
8 College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China; hfang@zju.edu.cn
* Correspondence: p.wang@mmu.ac.uk

Citation: Wang, P.; Mihaylova, L.; Chakraborty, R.; Munir, S.; Mayfield, M.; Muhammad, K.A.; Khokhar, M.F.; Zheng, Z.; Jiang, C.; Fang, H. A Gaussian Process Method with Uncertainty Quantification for Air Quality Monitoring. *Atmosphere* **2021**, *12*, 1344. https://doi.org/10.3390/atmos12101344

Academic Editors: Fabio Galatioto, Prashant Kumar and Francis Pope

Received: 13 September 2021
Accepted: 2 October 2021
Published: 14 October 2021

Abstract: The monitoring and forecasting of particulate matter (e.g., $PM_{2.5}$) and gaseous pollutants (e.g., NO, NO_2, and SO_2) is of significant importance, as they have adverse impacts on human health. However, model performance can easily degrade due to data noises, environmental and other factors. This paper proposes a general solution to analyse how the noise level of measurements and hyperparameters of a Gaussian process model affect the prediction accuracy and uncertainty, with a comparative case study of atmospheric pollutant concentrations prediction in Sheffield, UK, and Peshawar, Pakistan. The Neumann series is exploited to approximate the matrix inverse involved in the Gaussian process approach. This enables us to derive a theoretical relationship between any independent variable (e.g., measurement noise level, hyperparameters of Gaussian process methods), and the uncertainty and accuracy prediction. In addition, it helps us to discover insights on how these independent variables affect the algorithm evidence lower bound. The theoretical results are verified by applying a Gaussian processes approach and its sparse variants to air quality data forecasting.

Keywords: Gaussian process; uncertainty quantification; air quality forecasting; low-cost sensors; sustainable development

1. Introduction

It is generally believed that urban areas provide better opportunities in terms of economic, political, and social facilities compared to rural areas. As a result, more and more people are migrating to urban areas. At present, more than fifty percent of people worldwide live in urban areas, and this percentage is increasing with time. This has led to several environmental issues in large cities, such as air pollution [1].

Landrigan reported that air pollution caused 6.4 million deaths worldwide in 2015 [2]. According to World Health Organization (WHO) statistical data, three million premature deaths were caused by air pollution worldwide in 2012 [3]. Air pollution has a strong link with dementia, causing 850,000 people to suffer from dementia in the UK [4]. Children growing up in residential houses near busy roads and junctions have a much higher risk of developing various respiratory diseases, including asthma, due to high levels of

air pollution [5]. Polluted air, especially air with high levels of NO, NO_2, and SO_2 and particulate matter ($PM_{2.5}$), is considered the most serious environmental risk to public health in urban areas [6]. Therefore, many national and international organisations are actively working on understanding the behaviour of various air pollutants [7]. This eventually leads to the development of air quality forecasting models so that people can be alerted in time [8].

Essentially, being like a time series, air quality data can be easily processed by models that are capable of time series data processing. For instance, Shen applies an autoregressive moving average (ARMA) model in $PM_{2.5}$ concentration prediction in a few Chinese cities [9]. Filtering techniques like Kalman filter are also applied to adjust data biases to improve air quality prediction accuracy [10]. These methods, though with good results reported, are limited by the requirement of a prior model before data processing. Machine learning methods, on the other hand, can learn a model from the data directly. This has enabled them to attract wide attention in recent decades in the field of air quality forecasting. For instance, Lin et al. propose the support vector regression with logarithm preprocessing procedure and immune algorithms (SVRLIA) method, which outperforms general regression neural networks (GRNN) [11] and BackPropagation neural networks (BPNN) [12] in Taiwan air quality forecasting [13].

Recently, inspired by the fact that large scale data are accumulated, deep learning models have been applied in air quality prediction [14]. Some work has added these deep learning models with the ability to quantify uncertainties introduced by inputs. For instance, Garriga-Alonso et al. endow a deep convolutional network with uncertainty quantification, by taking it as an equivalent of a Gaussian processes (GPs) model [15]. This is because GPs predictions are accompanied by confidence intervals, which are usually taken as a metric to measure prediction uncertainties. Applications of GPs in air quality forecasting can be found in [16,17]. However, the involvement of matrix inversion in GPs limits their application in large-scale datasets [18]. This has inspired research on improving the efficiency of GP models, and a series of efficient GP models have been published [19]. We also proposed an efficient GP model with application in air quality forecasting [17]. Despite the rich number of GP models published, there lacks work that investigates how noise level, hyperparameters, etc. affect the performance of GP models. It is necessary because air quality data vary due to seasonal variations and sensor degradations. A well-trained GP model may not work when fed with new data, simply due to measurement noise level change. By knowing how the variation of GPs performance can be attributed to noise level and hyperparameters, etc., we will still be able perform analysis when noise level or hyperparameters vary.

Aiming at this, a general solution is proposed in this paper. It provides insights on how a GP model's performance is related to measurement noise level and hyperparameters, etc. The main contribution of this work includes (1) a general method for analysing how noise level and hyperparameters of a GP model affect the prediction performance. The variation of the evidence lower bound (ELBO) and the upper bound of the marginal likelihood (UBML) with respect to the noise level and hyperparameters are also given. (2) Neumann series is exploited to approximate the matrix inversion involved in GPs. This helps construct an analytical relation between noise level, hyperparameters, etc., and model performance. (3) A comparative air quality forecasting study between Sheffield, UK, and Pershawar, Pakistan is given, demonstrating that the proposed solution is able to capture how noise level and hyperparameters affect GPs performance.

The remaining part of this paper is as follows. Section 2 provides the theoretical fundamentals involved in this paper; Section 3 elaborates the proposed uncertainty quantification solution. In Section 4, we provide a comparative study of air quality prediction in the same period between the British city Sheffield and Pakistani city Pershawar, and the paper is concluded in Section 5. Appendix A describes the data collection process in Peshawar, Pakistan, and in Sheffield, United Kingdom, and presents maps of the considered areas

of these cities. Appendix B gives the World Health Organisation (WHO) criteria for air pollutants. Appendix C gives the approximate derivatives of the GP kernel.

2. Background Knowledge

2.1. Gaussian Processes

Given a set of training data $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$ where $\mathbf{x}_i \in \mathcal{X}$ is the input and $y_i \in \mathbb{R}$ is the observation, we can determine a GP model $f(\cdot)$ to predict y_* for a new input $\mathbf{x}_*$. For instance, when the output is one-dimensional, the GP model is formulated as

$$f \sim \mathcal{GP}\big(\bar{f}(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big), \quad y = f(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $\bar{f} : \mathcal{X} \to \mathbb{R}$ is the mean function defined as

$$\bar{f}(\mathbf{x}) = \mathbb{E}\big[f(\mathbf{x})\big], \tag{2}$$

and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the kernel function [18] defined as

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\big[(f(\mathbf{x}) - \bar{f}(\mathbf{x}))(f(\mathbf{x}') - \bar{f}(\mathbf{x}'))\big], \tag{3}$$

where ε is the additive, independent, identically distributed Gaussian measurement noise with variance $\sigma^2 \neq 0$, and $\mathbb{E}$ denotes the mathematical expectation operation.

Given $\mathbf{x}_i$ a $D \times 1$ vector, the n inputs can be aggregated into a matrix $\mathbf{X}_{D\times n}$, or briefly $\mathbf{X}$ with the corresponding output vector $\mathbf{y}_{n\times 1}$, or $\mathbf{y}$. Similarly, the function values at the test inputs $\mathbf{X}_*$ with dimensions of $D \times N$ can be denoted as $\mathbf{f}_*$, and we next write the joint distribution of $\mathbf{y}$ and $\mathbf{f}_*$ as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{nn} + \sigma^2\mathbf{I} & \mathbf{K}_{nN} \\ \mathbf{K}_{Nn} & \mathbf{K}_{NN} \end{bmatrix}\right), \tag{4}$$

where $\mathbf{I}$ represents the identity matrix. $\mathbf{K}_{nn} + \sigma^2\mathbf{I}$ is the $n \times n$ prior covariance matrix of $\mathbf{y}$ with entry $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma^2\delta_{ij}$, where δ_{ij} is one iff $i = j$ and zero otherwise, and $\mathbf{x}_i$ and $\mathbf{x}_j$ are column vectors from $\mathbf{X}$. The matrix $\mathbf{K}_{NN}$ denotes the $N \times N$ prior covariance matrix of $\mathbf{f}_*$ with entry $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are column vectors from $\mathbf{X}_*$. The matrices $\mathbf{K}_{Nn}$ and $\mathbf{K}_{nN}$ satisfy $\mathbf{K}_{Nn} = \mathbf{K}_{nN}^{\mathrm{T}}$, and the entry of the $N \times n$ prior covariance matrix of $\mathbf{f}_*$ and $\mathbf{y}$ is $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_i$ is a column vector from $\mathbf{X}_*$ and $\mathbf{x}_j$ is a column vector from $\mathbf{X}$.

By deriving the conditional distribution of $\mathbf{f}_*$ from (4), where the prior mean is set to be zero for simplicity [20], we have the predictive posterior at new inputs $\mathbf{X}_*$ as

$$\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)\big), \tag{5}$$

where

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}\big[\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*\big] = \mathbf{K}_{Nn}\big[\mathbf{K}_{nn} + \sigma^2\mathbf{I}\big]^{-1}\mathbf{y}, \tag{6}$$

is the prediction at $\mathbf{X}_*$, and

$$\mathrm{cov}(\mathbf{f}_*) = \mathbf{K}_{NN} - \mathbf{K}_{Nn}\big[\mathbf{K}_{nn} + \sigma^2\mathbf{I}\big]^{-1}\mathbf{K}_{Nn}^{\mathrm{T}}, \tag{7}$$

denotes the covariance of $\mathbf{f}_*$.

The hyperparameter $\boldsymbol{\theta}$ incorporated in the mean and covariance functions underpin the predictive performance of GP models, and they are usually estimated by maximising the logarithm of the marginal likelihood

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\big(\mathbf{K}_{nn} + \sigma^2\mathbf{I}\big)^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_{nn} + \sigma^2\mathbf{I}| - \frac{n}{2}\log 2\pi. \tag{8}$$

2.2. Neumann Series Approximation

Given a matrix inverse $\mathbf{A}^{-1}$, it can be expanded as the following Neumann series [21]

$$\mathbf{A}^{-1} = \sum_{n=0}^{\infty} \left(\mathbf{X}^{-1}(\mathbf{X} - \mathbf{A})\right)^n \mathbf{X}^{-1}, \tag{9}$$

which holds if $\lim_{n\to\infty} \left(\mathbf{I} - \mathbf{X}^{-1}\mathbf{A}\right)^n = \mathbf{0}$ is satisfied. In our case, suppose

$$\mathbf{A} = \mathbf{K} + \sigma_n^2 \mathbf{I} \triangleq \mathbf{D}_A + \mathbf{E}_A, \tag{10}$$

where $\mathbf{D}_A$ is the main diagonal of $\mathbf{A}$ and $\mathbf{E}_A$ is the hollow. If we substitute $\mathbf{X}$ in Equation (9) by $\mathbf{D}_A$, we get

$$\mathbf{A}^{-1} = \sum_{n=0}^{\infty} \left(-\mathbf{D}^{-1}\mathbf{E}_A\right)^n \mathbf{D}_A^{-1}, \tag{11}$$

which is guaranteed to converge when $\lim_{n\to\infty} \left(-\mathbf{D}_A^{-1}\mathbf{E}_A\right)^n = \mathbf{0}$. We investigated the convergence condition in [17], where we proved that if $\mathbf{A}$ is diagonally dominant, then Neumann series can approximate $\mathbf{A}^{-1}$ both fast and accurate. In case $\mathbf{A}$ is not diagonally dominant, we also provided a way to convert it into a diagonally dominant matrix in [17], such that $\mathbf{A}^{-1}$ can still be approximated by Neumann series. When Neumann series given in (11) converges, we can then approximate $\mathbf{A}$ with only the first L terms. The L-term approximation is computed as follows:

$$\widetilde{\mathbf{A}}_L^{-1} = \sum_{n=0}^{L-1} \left(-\mathbf{D}_A^{-1}\mathbf{E}_A\right)^n \mathbf{D}_A^{-1}, \tag{12}$$

For instance, when $L = 1, 2, 3$, we have the approximations

$$\widetilde{\mathbf{A}}_L^{-1} = \begin{cases} \mathbf{D}_A^{-1}, & L = 1 \\ \mathbf{D}_A^{-1} - \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}, & L = 2 \\ \mathbf{D}_A^{-1} - \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1} + \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}. & L = 3 \end{cases} \tag{13}$$

3. Uncertainty Quantification in Gaussian Processes

3.1. Uncertainty in Measurements

It is intuitive that noisy measurements would result in less accurate predictions, just as a poor model would do. However, it is not direct from Equations (6) and (7). We will show in detail how the measurement noise would affect the prediction accuracy.

From Equations (6) and (7), we can see that the measurement noise ϵ affects the prediction and the covariance by adding a term $\sigma_n^2\mathbf{I}$ to the prior covariance $\mathbf{K}$ in comparison to the noisy free scenario [20]. From the way that they originated, we know that both $\mathbf{K}$ and $\sigma_n^2\mathbf{I}$ are symmetrical. Then, a matrix $\mathbf{P}$ exists such that

$$\mathbf{K} = \mathbf{P}^{-1}\mathbf{D}_\mathbf{K}\mathbf{P}, \tag{14}$$

where $\mathbf{D}_\mathbf{K}$ is a diagonal matrix with eigen values of $\mathbf{K}$ along the diagonal. As $\sigma_n^2\mathbf{I}$ a diagonal matrix itself, we have

$$\sigma_n^2\mathbf{I} = \mathbf{P}^{-1}\sigma_n^2\mathbf{I}\mathbf{P}. \tag{15}$$

Therefore, we have the partial derivative of Equation (6) with respect to σ_n^2 as

$$\frac{\partial \bar{\mathbf{f}}_*}{\partial \sigma_n^2} = \mathbf{K}_*\mathbf{P}(\mathbf{D}_\mathbf{K} + \sigma_n^2\mathbf{I})^{-2}\mathbf{P}^{-1}\mathbf{y}, \tag{16}$$

The element-wise form of Equation (16) can be therefore obtained as

$$\Big(\frac{\partial \bar{\mathbf{f}}_*}{\partial \sigma_n^2}\Big)_o = -\sum_{h=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{hj}p_{ij}k_{oh}\Lambda_j^{-1}y_i, \tag{17}$$

where $\Lambda_j = (\lambda_j + \sigma_n^2)^2$. p_{hj} and p_{ij} are the entries indexed by the j-th column, h-th and i-th row, respectively. k_{oh} is the o-th row and h-th column entry of $\mathbf{K}_*$. y_i is the i-th element of $\mathbf{y}$. $o = 1, \cdots, s$ denotes the o-th element of the partial derivation.

We can see that the sign of Equation (17) is determined by p_{hj} and p_{ij}. This is because we can actually transform $\mathbf{y}$ to either positive or negative with a linear transformation, which will not be an issue for the GPs model. When we impose no constraints on p_{hj} and p_{ij}, Equation (17) could be any real number, indicating that $\bar{\mathbf{f}}_*$ is multimodal with respect to σ_n^2, which means that one σ_n^2 can lead to different $\bar{\mathbf{f}}_*$, or equivalently, different σ_n^2 can lead to the same $\bar{\mathbf{f}}_*$. In such cases, it is difficult to investigate how σ_n^2 affects the prediction accuracy. In this paper, to facilitate the study of the monotonicity of $\bar{\mathbf{f}}_*$, we constrain p_{hj} and p_{ij} to satisfy

$$\Big(\frac{\partial \bar{\mathbf{f}}_*}{\partial \sigma_n^2}\Big)_o \begin{cases} > 0, & p_{hj}p_{ij} < 0, \\ < 0, & p_{hj}p_{ij} > 0, \\ = 0, & p_{hj}p_{ij} = 0. \end{cases} \tag{18}$$

Then, we can see that $\bar{\mathbf{f}}_*$ is monotonic. It means that changes of σ_n^2 can cause arbitrarily large/small predictions, whereas a robust method should bound the prediction errors regardless of how σ_n^2 varies.

Similarly, the partial derivative of Equation (7) with respect to σ_n^2 is

$$\frac{\partial \mathrm{cov}(\mathbf{f}_*)}{\partial \sigma_n^2} = (\mathbf{K}_*\mathbf{P})(\mathbf{D}_\mathbf{K} + \sigma_n^2\mathbf{I})^{-2}(\mathbf{K}_*\mathbf{P})^T = \sum_{i=1}^{n}\Lambda_i^{-1}\vec{\mathbf{p}}_i\vec{\mathbf{p}}_i^T, \tag{19}$$

where we denote the $m \times n$ dimension matrix $\mathbf{K}_*\mathbf{P}$ as

$$\mathbf{K}_*\mathbf{P} = [\vec{\mathbf{p}}_1, \vec{\mathbf{p}}_2, \cdots, \vec{\mathbf{p}}_n], \tag{20}$$

with $\vec{\mathbf{p}}_i$ a $m \times 1$ vector, and $i = 1, \cdots, n$.

As the uncertainty is indicated by the diagonal elements, we only show how these elements change with respect to σ_n^2. The diagonal elements are given as

$$\begin{aligned} \mathrm{diag}\Big(\sum_{i=1}^{n}\Lambda_i^{-1}\vec{\mathbf{p}}_i\vec{\mathbf{p}}_i^T\Big) &= \mathrm{diag}\Big(\sum_{i=1}^{n}\Lambda_i^{-1}p_{1i}^2, \sum_{i=1}^{n}\Lambda_i^{-1}p_{2i}^2, \cdots, \sum_{i=1}^{n}\Lambda_i^{-1}p_{mi}^2\Big) \\ &= \mathrm{diag}\Big(\Sigma_{11}, \Sigma_{22}, \cdots, \Sigma_{mm}\Big), \end{aligned} \tag{21}$$

with $\mathrm{diag}(\cdot)$ denoting the diagonal elements of a matrix. We see that $\Sigma_{jj} \geqslant 0$ stands for $j = 1, \cdots, m$, which implies that $\mathrm{cov}(\mathbf{f}_*)$ is non-decreasing as σ_n^2 increases. This means that the increase of measurement noise level would cause the non-deceasing of the prediction uncertainty.

3.2. Uncertainty in Hyperparameters

Another factor that affects the prediction of a GPs model is the hyperparameters. In Gaussian processes, the posterior, as shown in Equation (5), is used to do the prediction, while the marginal likelihood is used for hyperparameters selection [18]. The log marginal likelihood as shown in Equation (22) is usually optimised to determine the hyperparameter with a specified kernel function.

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K} + \sigma_n^2\mathbf{I}| - \frac{N}{2}\log 2\pi. \tag{22}$$

However, the log marginal likelihood could be non-convex with respect to the hyperparameters, which implies that the optimisation may not converge to the global maxima [22]. A common solution dealing with it is to sample multiple starting points from a prior distribution, then choose the best set of hyperparameters according to the optima of the log marginal likelihood. Let's assume $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_s\}$ being the hyperparameter set and θ_s denoting the s-th of them, then the derivative of $\log p(\mathbf{y}|\mathbf{X})$ with respect to θ_s is

$$\frac{\partial}{\partial \theta_s} \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{2}\mathrm{tr}\Big(\big(\boldsymbol{\alpha}\boldsymbol{\alpha}^T - (\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\big) \frac{\partial(\mathbf{K} + \sigma_n^2\mathbf{I})}{\partial \theta_s} \Big), \tag{23}$$

where $\boldsymbol{\alpha} = (\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}$, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. The derivative in Equation (23) is often multimodal and that is why a fare few initialisations are used when conducting convex optimisation. Chen et al. show that the optimisation process with various initialisations can result in different hyperparameters [22]. Nevertheless, the performance (prediction accuracy) with regard to the standardised root mean square error does not change much. However, the authors do not show how the variation of hyperparameters affects the prediction uncertainty [22].

An intuitive explanation to the fact of different hyperparameters resulting with similar predictions is that the prediction shown in Equation (6) is non-monotonic itself with respect to hyperparameters. To demonstrate this, a direct way is to see how the derivative of (6) with respect to any hyperparameter $\theta_s \in \boldsymbol{\theta}$ changes, and ultimately how it affects the prediction accuracy and uncertainty. The derivatives of $\bar{\mathbf{f}}_*$ and $\mathrm{cov}(\mathbf{f}_*)$ of θ_s are as below

$$\frac{\partial \bar{\mathbf{f}}_*}{\partial \theta_s} = \Big(\mathbf{K}_* \frac{\partial(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}}{\partial \theta_s} + \frac{\partial \mathbf{K}_*}{\partial \theta_s}(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1} \Big)\mathbf{y}. \tag{24}$$

We can see that Equations (24) and (25) are both involved with calculating $(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}$, which becomes enormously complex when the dimension increases. In this paper, we focus on investigating how hyperparameters affect the predictive accuracy and uncertainty in general. Therefore, we use the Neumann series to approximate the inverse [21].

$$\begin{aligned} \frac{\partial \mathrm{cov}(\mathbf{f}_*)}{\partial \theta_s} =& \frac{\partial \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*)}{\partial \theta_s} - \frac{\partial \mathbf{K}_*}{\partial \theta_s}(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{K}_*^T - \mathbf{K}_* \frac{\partial(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}}{\partial \theta_s}\mathbf{K}_*^T \\ & - \mathbf{K}_*(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\frac{\partial \mathbf{K}_*^T}{\partial \theta_s}. \end{aligned} \tag{25}$$

3.3. Derivatives Approximation with Neumann Series

The approximation accuracy and computationally complexity of Neumann series varies with L. This has been studied in [21,23], as well as in our previous work [17]. This paper aims at providing a way to quantify uncertainties involved in GPs. We therefore choose the 2-term approximation as an example to carry out the derivations. By substituting the 2-term approximation into Equations (24) and (25), we have

$$\frac{\partial \bar{\mathbf{f}}_*}{\partial \theta_s} \approx \Big(\mathbf{K}_* \frac{\partial\big(\mathbf{D}_A^{-1} - \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}\big)}{\partial \theta_s} + \frac{\partial \mathbf{K}_*}{\partial \theta_s}\big(\mathbf{D}_A^{-1} - \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}\big) \Big)\mathbf{y}, \tag{26}$$

$$\begin{aligned} \frac{\partial \mathrm{cov}(\mathbf{f}_*)}{\partial \theta_s} \approx& \frac{\partial \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*)}{\partial \theta_s} - \frac{\partial \mathbf{K}_*}{\partial \theta_s}\big(\mathbf{D}_A^{-1} - \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}\big)\mathbf{K}_*^T \\ & - \mathbf{K}_* \frac{\partial\big(\mathbf{D}_A^{-1} - \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}\big)}{\partial \theta_s}\mathbf{K}_*^T - \mathbf{K}_*\big(\mathbf{D}_A^{-1} - \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}\big)\frac{\partial \mathbf{K}_*^T}{\partial \theta_s}. \end{aligned} \tag{27}$$

Due to the simple structure of matrices $\mathbf{D}_A$ and $\mathbf{E}_A$, we can get the element-wise form of Equation (26) as

$$\Big(\frac{\partial \bar{\mathbf{f}}_*}{\partial \theta_s}\Big)_o = \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(k_{oj}\frac{\partial d_{ji}}{\partial \theta_s} + \frac{\partial k_{oj}}{\partial \theta_s}d_{ji}\Big)y_i. \tag{28}$$

Similarly, the element-wise form of Equation (27) is

$$\Big(\frac{\partial \mathrm{cov}(\mathbf{f}_*)}{\partial \theta_s}\Big)_{oo} = \frac{\partial \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*)_{oo}}{\partial \theta_s} - \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\frac{\partial k_{oj}}{\partial \theta_s} d_{ji} k_{oi} + k_{oj}\frac{\partial d_{ji}}{\partial \theta_s} k_{oi} - k_{oj} d_{ji}\frac{\partial k_{oi}}{\partial \theta_s}\Big), \tag{29}$$

where $o = 1, \cdots, m$ denotes the o-th output, d_{ji} is the j-th row and i-th column entry of $\mathbf{D}_A^{-1} - \mathbf{D}_A^{-1}\mathbf{E}_A\mathbf{D}_A^{-1}$, k_{oj} and k_{oi} are the o-th row, j-th and i-th entries of matrix $\mathbf{K}_*$, respectively. When the kernel function is determined, Equations (26)–(29) can be used for GPs uncertainty quantification.

3.4. Impacts of Noise Level and Hyperparameters on ELBO and UBML

The minimisation of $\mathrm{KL}\big(q(\mathbf{f}, \mathbf{u}) \| p(\mathbf{f}, \mathbf{u}|\mathbf{y})\big)$ is equivalent to maximise the ELBO [18,24] as shown in

$$\mathcal{L}_{\text{lower}} = -\frac{1}{2}\mathbf{y}^T\mathbf{G}_n^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{G}_n| - \frac{N}{2}\log(2\pi) - \frac{t}{2\sigma_n^2}, \tag{30}$$

where $\mathbf{G}_n = \mathbf{G}_{\mathbf{xx}} + \sigma_n^2\mathbf{I}$, and $t = \mathrm{Tr}(\mathbf{K}_{\mathbf{xx}} - \mathbf{G}_{\mathbf{xx}})$. Combining it with UBML, as shown in Equation (31), an interval can be given to quantify the uncertainty in marginal likelihood.

$$\mathcal{L}_{\text{upper}} = \frac{1}{2}\mathbf{y}^T\Big(\mathbf{G}_n + t\mathbf{I}\Big)^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{G}_n| - \frac{N}{2}\log(2\pi). \tag{31}$$

This paper, however, focuses on investigating how ELBO and UBML change according to σ_n^2 only. Because the investigation of how ELBO and UBML change with respect to kernel hyperparameters involves multiple Neumann series approximations, which makes the analysis less convincing. We shall leave it as an open problem for future study. The derivatives of Equations (30) and (31) with respect to σ_n^2 are as follows,

$$\frac{\partial \mathcal{L}_{\text{lower}}}{\partial \sigma_n^2} = \frac{1}{2}\Big[\sum_{i=1}^{n}(\lambda_i + \sigma_n^2)^{-2}\Big(\sum_{j=1}^{n} y_j v_{ji}\Big)^2 - \sum_{i=1}^{n}\frac{1}{\lambda_i + \sigma_n^2} + \frac{t}{\sigma_n^4}\Big], \tag{32}$$

$$\frac{\partial \mathcal{L}_{\text{upper}}}{\partial \sigma_n^2} = -\frac{1}{2}\Big[\sum_{i=1}^{n}(\lambda_i + \sigma_n^2 + t)^{-2}\Big(\sum_{j=1}^{n} y_j v_{ji}\Big)^2 + \sum_{i=1}^{n}\frac{1}{\lambda_i + \sigma_n^2}\Big]. \tag{33}$$

Figure 1 shows how σ_n^2 affects ELBO and UBML. We set σ_n^2 to increase from 0.1 to 200.0 with a step of 0.01. Both ELBO and UBML are recorded step by step. From the figure, we can see that when σ_n^2 is small ($\sigma_n^2 \in [0.1, 1.5]$), ELBO increases with different speeds, however, UBML fluctuates as the derivative of UBML jumps between positive and negative. When σ_n^2 is in [1.5, 3.0], ELBO still increases, but the speeds slow down significantly. In comparison, UBML keeps decreasing with reducing speeds. The decrements of UBML mean that when σ_n^2 increases, though ELBO could be increased still, but the maximum (which is the UBML) can decrease. When $\sigma_n^2 \in [3.0, 20.0]$, ELBO starts to decrease when $\sigma_n^2 \approx 3.2$, while UBML keeps decreasing. This means that as σ_n^2 increases, both ELBO and UBML decrease, which indicates that the model becomes less and less effective to explain the data. When σ_n^2 keeps increasing ($\sigma_n^2 \in [20.0, 200.0]$), the decreasing speeds of ELBO and UBML becomes similar and approaches zero. This means that UBML and ELBO both converge and together define an interval for the marginal likelihood, which however, can result in non-optimal hyperparameters. Our conclusion is that when σ_n^2 increases, UBML tends to decrease, which decreases the maximum that ELBO can reach. ELBO, on the other hand, is robust to the change of σ_n^2 (as it keeps increasing when σ_n^2 is below $\sim$3.2). However, when σ_n^2 exceeds a certain threshold, ELBO turns to decrease, indicating that the GPs model becomes less and less reliable. However, both ELBO and UBML converge, even when σ_n^2 becomes very significant, though we can no longer trust the model.

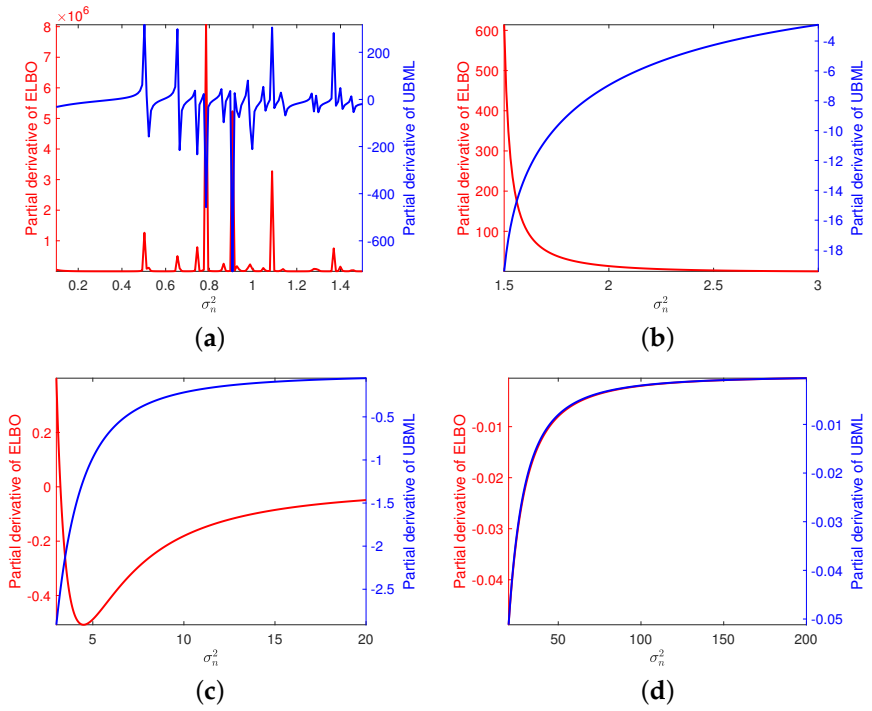

Figure 1. Impacts of σ_n^2 on ELBO and UBML: (**a**) $\sigma_n^2 \in [0.1, 1.5]$, (**b**) $\sigma_n^2 \in [1.5, 3.0]$, (**c**) $\sigma_n^2 \in [3.0, 20.0]$, (**d**) $\sigma_n^2 \in [20.0, 200.0]$.

4. Experiments and Analysis

To verify that the proposed solution can help to identify the impacts of σ_n^2 and $\boldsymbol{\theta}$ on the predition accuracy and uncertainty of GPs model and its sparse variants such as the fully independent training conditional (FITC) [25] and variational free energy (VFE) [24] models, we conduct various experiments to process air quality data collected from Sheffield, UK, and Pershawar, Pakistan (see Appendix A), during the time period of 24 June 2019–14 July 2019 for three weeks, which will be denoted as W1, W2, and W3 hereafter. The data were collected with digital sensors called AQMesh pod with a 15 min time interval. Though the sensor itself is able to measure the concentrations of quite a few atmospheric pollutants, here we only analyse the concentrations of NO, NO_2, SO_2, and $PM_{2.5}$. Figure 2 shows the raw data. We can see directly that the air quality of Sheffield is much better than Pershawar on average. Especially during daytime, concentrations of NO_2 and $PM_{2.5}$ in Pershawar exceed the WHO criteria (see Appendix B). Meanwhile, those in Sheffield are much lower than the criteria. Being a postindustrial city itself, Sheffield has improved air quality significantly. The experience can be spread to help cities like Pershwar to improve air quality.

4.1. Air Quality Prediction

Figures 3 and 4 show Sheffield and Pershawar forecasting results of GPs, FITC, and VFE, with 3σ confidence intervals (denoted as Conf in the figures) indicated by the shaded area. We can see that the GPs model reports the best results in general, in terms of absolute error between predicts and measurements (denoted as Meas in the figures). However, the performance of all the models varies from pollutant types to cities. This is actually one of the reasons why the investigation of how measurement noise level and hyperparameters affect prediction accuracy and uncertainty is necessary. To make the results more convincing, we normalise the data from both cities for uncertainty quantification studies.

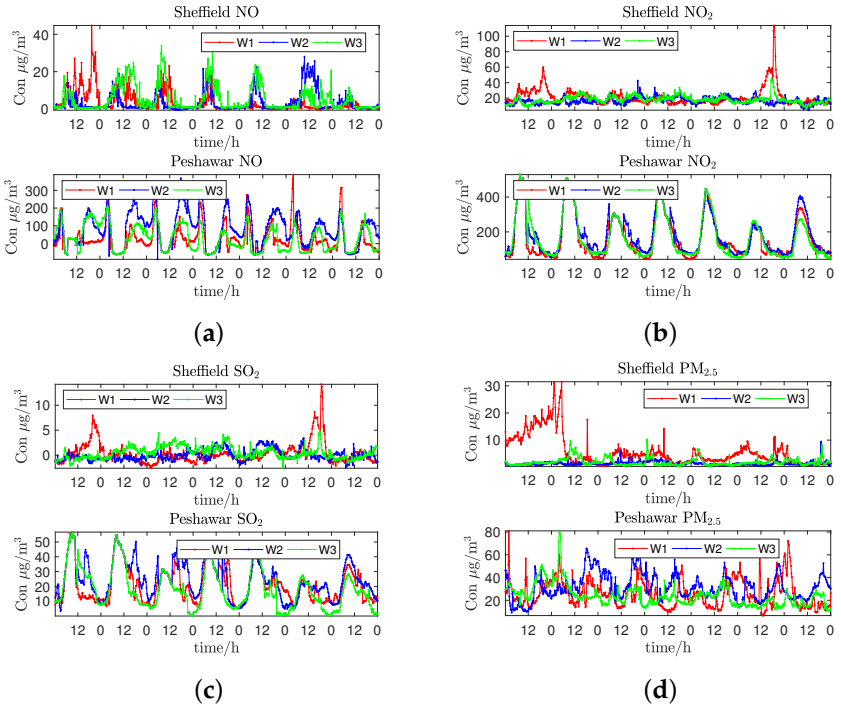

Figure 2. Concentration of pollutants recorded at the same time period in both Sheffield and Peshawar: (**a**) NO, (**b**) NO_2, (**c**) SO_2, (**d**) $PM_{2.5}$.

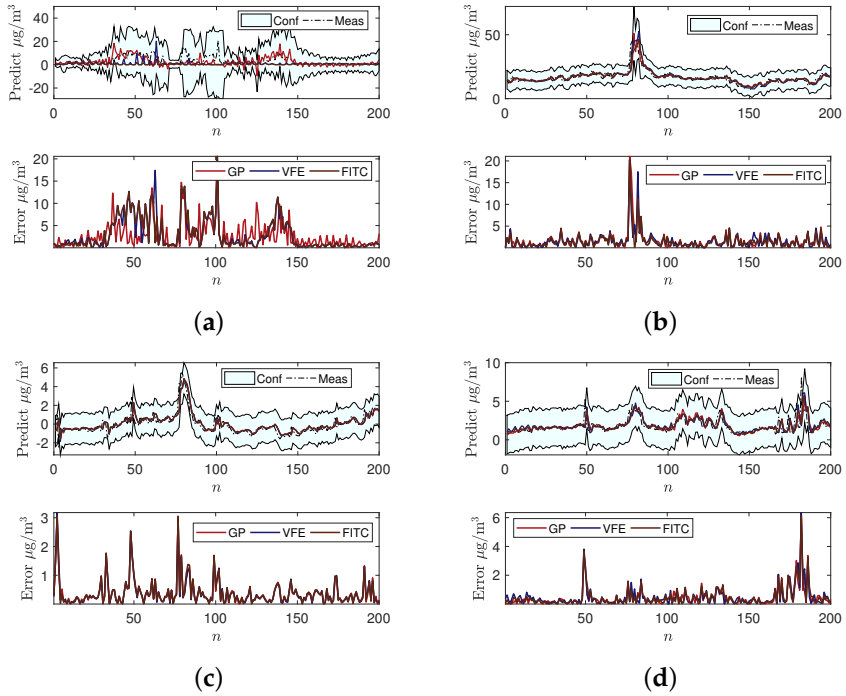

Figure 3. Prediction and absolute error of pollutants in Sheffield: (**a**) NO, (**b**) NO_2, (**c**) SO_2, (**d**) $PM_{2.5}$.

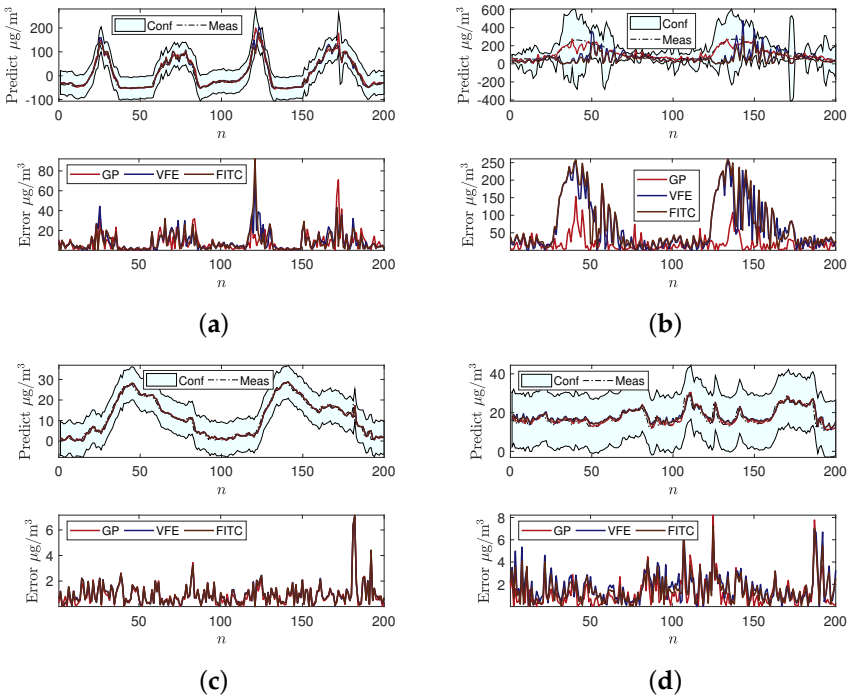

Figure 4. Prediction and absolute error of pollutants in Peshawar: (**a**) NO, (**b**) NO_2, (**c**) SO_2, (**d**) $PM_{2.5}$.

4.2. Impacts of Measurement Noise Level and Hyperparameters

To demonstrate how noise level σ_n^2 and hyperparameters affect prediction accuracy and uncertainty, three sets of experiments are conducted. This paper adopts the squared exponential (SE) kernel, with hyperparameters s_f and l. The analytical derivation can be found in Appendix C. The prediction accuracy is identified by the root mean square error (RMSE), as shown in Equation (34), while the uncertainty is identified by $\frac{1}{2}\sigma$ confidence bound. Configurations of the experiments are as follows.

Experiment 1: Impacts of σ_n^2 on prediction accuracy and uncertainty. Both s_f and l are fixed to be the optimised values. σ_n^2 varies from 0.1 through to 20.0. NO, NO_2, SO_2, and $PM_{2.5}$ data from both cities are processed. Six inducing points are applied to both FITC and VFE.

Experiment 2: Impacts of s_f on prediction accuracy and uncertainty. l is set to the optimised value. s_f varies from 0.1 through to 30.0. σ_n^2 is set to 0.5 and 1.5, respectively. NO data from both cities are processed. Six inducing points are applied to both FITC and VFE.

Experiment 3: Impacts of l on prediction accuracy and uncertainty. s_f is set to the optimised value. l varies from 0.1 through to 30.0. σ_n^2 is set to 0.5 and 1.5, respectively. NO data from both cities are processed. Six inducing points are applied to both FITC and VFE.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{Num}(y_i - \widehat{y}_i)^2}{Num}}, \tag{34}$$

where y_i is the ground truth value and $\widehat{y}_i$ represents predicted meant. Num is the sample number in testing set.

Figures 5 and 6 show the results from **Experiment 1**. To make the results more distinguishable, the horizontal axes of the figures are set to $\log(\sigma_n^2)$. We can see from Figure 5 that when σ_n^2 is small, GPs perform the best in general, while the performance of FITC and VFE varies. We can also observe that as σ_n^2 keeps increasing, the RMSE becomes very significant for all methods/pollutants. Similar results can be observed from Figure 6 as well. Both comply with our theoretical conclusions, despite the fact that the Neumann series is used to approximate the matrix inverse. We also notice that σ_n^2 has a more significant impact on Sheffield data as RMSE increases ealier after $\log(\sigma_n^2)$ reaches

zero. From Figure 6b,c, we also see that the uncertainty bounds of Sheffield data are greater after $\log(\sigma_n^2)$ reaches zero. We think the reason is that Sheffield data are generally less periodical than Pershawar data (see Figure 2), which influences the performance of the models.

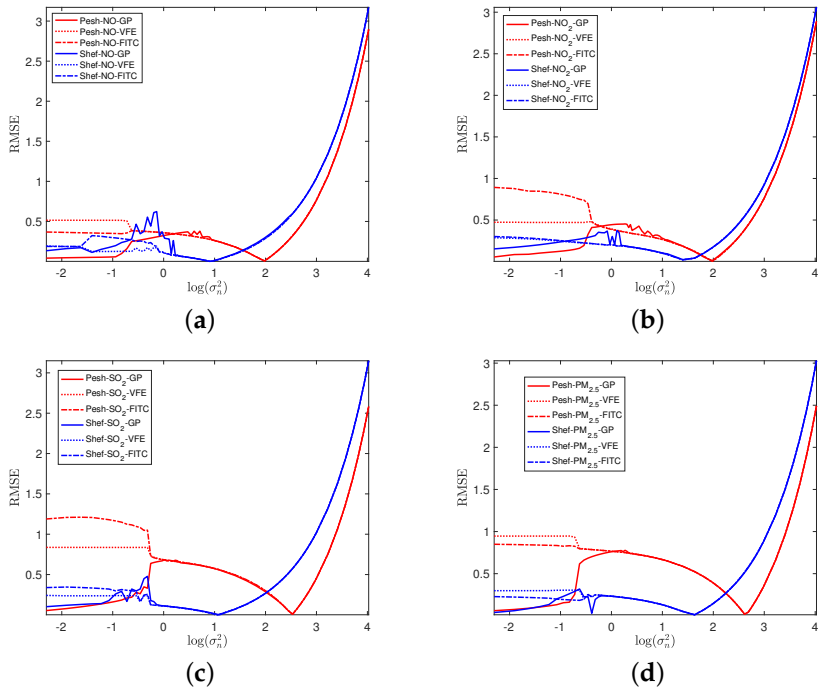

Figure 5. Relationship of σ_n^2 with four pollutants prediction RMSE: (**a**) NO, (**b**) NO_2, (**c**) SO_2, (**d**) $PM_{2.5}$.

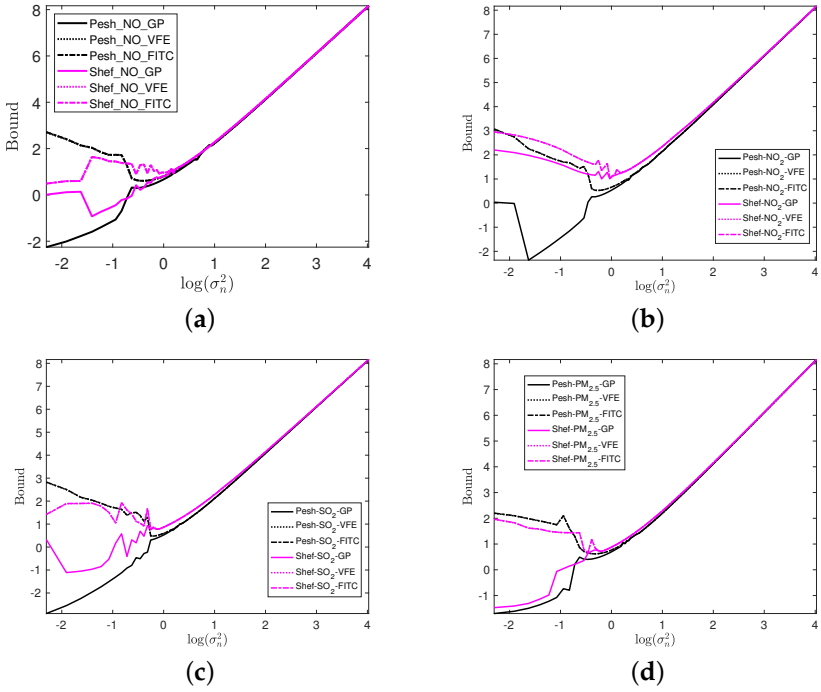

Figure 6. Relationship of σ_n^2 with pollutants prediction uncertainty bound: (**a**) NO, (**b**) NO_2, (**c**) SO_2, (**d**) $PM_{2.5}$.

4.3. Impacts of Noise Level on ELBO and UBML

Figure 7 shows the results from **Experiment 2**. According to our theoretical results, the impact of s_f on the uncertainty should become greater as s_f increases. This is verified by the results shown in Figure 7b,d. Our theoretical results also suggest that the variation of s_f would not affect the prediction accuracy. We can see from Figure 7a,c that when s_f is smaller, it does affect the prediction accuracy, but when it exceeds a certain value, the impacts become negligible. Considering the Neumann series approximation, we would say that the experimental results comply with the theoretical conclusion.

The results of **Experiment 3** are shown in Figure 8. We can see that when l is smaller, both RMSE and the uncertainty bounds change rapidly. While after it exceeds certain values, both converge. This again complies with our theoretical conclusions and simulation results. We should also notice from Figures 7 and 8 that the increment of s_f tends to increase the uncertainty, whereas the increment of l tends to decrease the uncertainty. Taking both into consideration, an optimised uncertainty bound can be obtained.

We also conduct an experiment to demonstrate how the noise level σ_n^2 affects the ELBO and UBML. In our experiment, we set σ_n^2 to vary from 0.5 to 4.5. The results are shown in Figure 9. To make the results distinguishable, we set the vertical axes to $\log(-ELBO/UBML)$. To make the logrithm work, we reverse the signs of both ELBO and UBML. This is the reason why ELBO is 'greater' than UBML in Figure 9. The full GPs model is trained by setting σ_n^2 to $\{1, 7, 13, 19, 25, 31, 37, 43, 49\}$ to obtain 9 sets of hyperparameters. For each set of them, we then set σ_n^2 to vary from 0.5 to 4.5. The darker the colour in Figure 9, the smaller σ_n^2 is for model training. We can see that generally, greater σ_n^2 can slow down the convergence speed of both ELBO and UBML, while training a model. When the model is trained, the increment of σ_n^2 can lower down UBML, which is the maximum that ELBO can reach. This implies that the increment of σ_n^2 can cause the failure of a sparse GPs model, as ELBO is deeply related to determine a sparse GPs model. Nevertheless, the experimental results again comply with our theoretical conclusions.

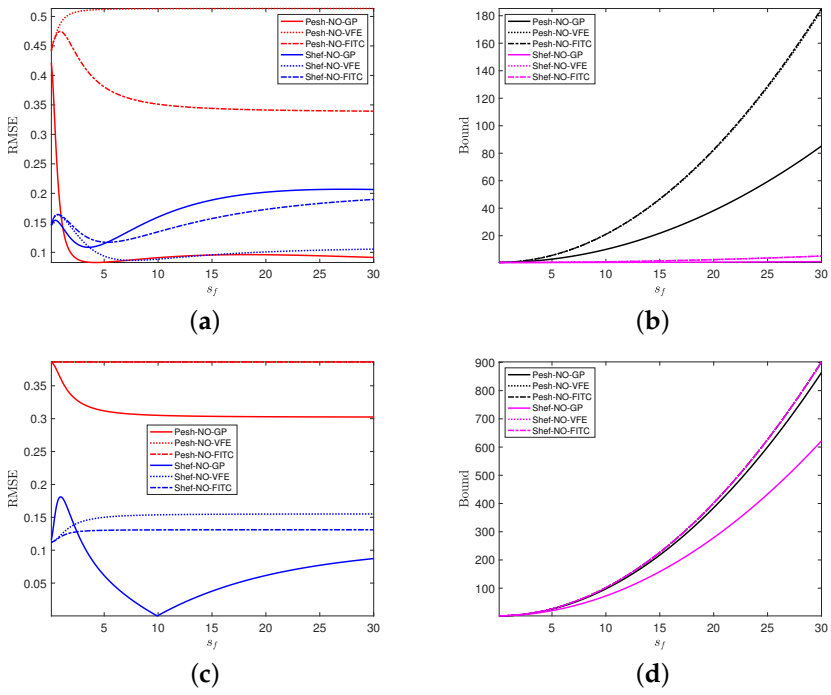

Figure 7. Relationship of s_f on NO prediction RMSE and uncertainty bound: (**a**) $\sigma_n^2 = 0.5$, (**b**) $\sigma_n^2 = 0.5$, (**c**) $\sigma_n^2 = 1.5$, (**d**) $\sigma_n^2 = 1.5$.

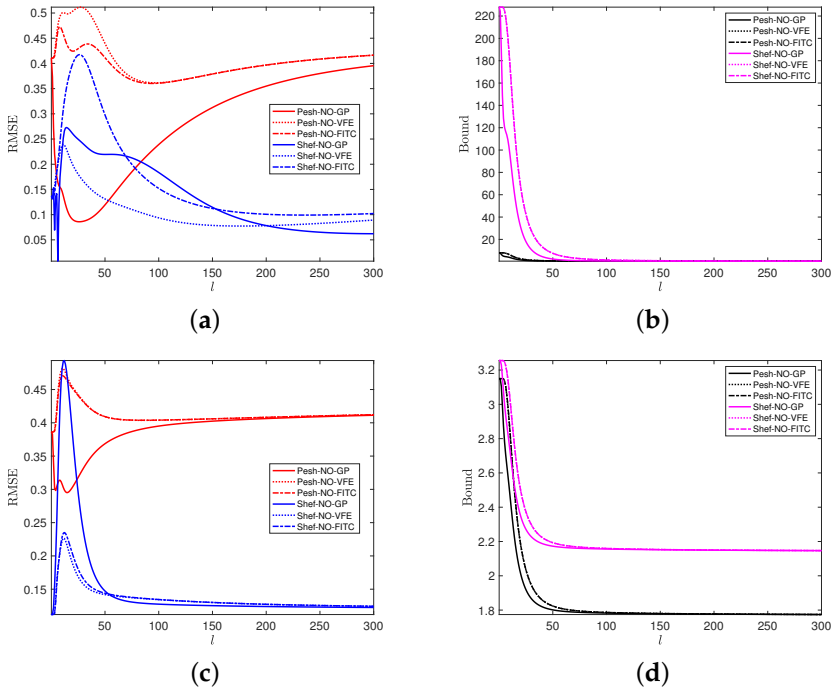

Figure 8. Relationship of l on NO prediction RMSE and uncertainty bound: (**a**) $\sigma_n^2 = 0.5$, (**b**) $\sigma_n^2 = 0.5$, (**c**) $\sigma_n^2 = 1.5$, (**d**) $\sigma_n^2 = 1.5$.

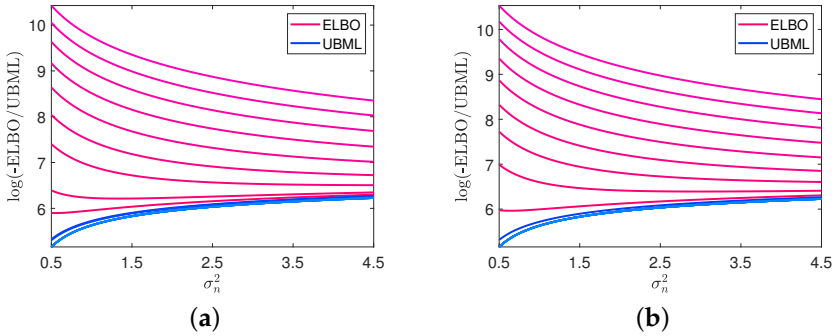

Figure 9. Effects of σ_n^2 on ELBO and UBML: (**a**) NO in Sheffield, (**b**) NO in Peshawar.

5. Conclusions

This paper proposes a general method to investigate how the performance variation of a Gaussian process model can be attributed to hyperparameters and measurement noises, etc. The method is demonstrated by applying it to process particulate matter (e.g., $PM_{2.5}$) and gaseous pollutants (e.g., NO, NO_2, and SO_2) from both Sheffield, UK, and Peshawar, Pakistan. Experimental results show that the proposed method provides insights on how measurement noises and hyperparameters, etc. affect the prediction performance of a Gaussian process. The results align with the analytical derivations, which is enabled by adopting Neuman series to approximate matrix inversions in Gaussian process models. The theoretical findings and experimental results combined demonstrate that the proposed method can generate air quality forecasting results. In the meantime, it provides a way to link uncertainties in measurements and hyperparameters, etc. with the forecasting results. This will help with forecasting performance analysis when measurement noise level or model hyperparameters vary, making the method more general.

Author Contributions: Conceptualization, P.W., L.M., M.M., R.C., S.M., K.A. and M.F.K.; methodology, P.W.; software, P.W.; validation, P.W., Z.Z., C.J. and H.F.; formal analysis, P.W., L.M.; investigation, P.W.; data curation, S.M., R.C., K.A. and M.F.K.; writing—original draft preparation, P.W., L.M., R.C., S.M., K.A. and M.F.K.; writing—review and editing, P.W. and L.M.; visualization, P.W., R.C.; supervision, L.M., M.M.; funding acquisition, L.M., P.W., M.M., S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the UK EPSRC through EP/T013265/1 project NSF-EPSRC:ShiRAS. Towards Safe and Reliable Autonomy in Sensor Driven Systems, a joint project with the USA National Science Foundation under Grant NSF ECCS 1903466. Other funders are NSFC (61703387) and the Global Challenges Research Funds (QR GCRF—Pump priming awards (Round 2), project entitled: "Collaborating with North Pakistan for monitoring and reducing the air pollution (X/160978)".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not Applicable.

Acknowledgments: We are grateful to UK EPSRC for funding this work through EP/T013265/1 project NSF-EPSRC:ShiRAS. Towards Safe and Reliable Autonomy in Sensor Driven Systems. This work was also supported by the USA National Science Foundation under Grant NSF ECCS 1903466. We also appreciate the support of NSFC (61703387). We are also grateful to the Global Challenges Research Funds (QR GCRF - Pump priming awards (Round 2), entitled: "Collaborating with North Pakistan for monitoring and reducing the air pollution (X/160978))". We also thank Urban FLows Observatory, the University of Sheffield for providing the air quality sensors for collecting air pollution data in Pakistan.

Conflicts of Interest: : The authors declare no conflict of interest.

Appendix A. Data Collection

Peshawar (34.015° N, 71.52° E) is a city located in Khyber Pakhtunkhwa, Pakistan, situated at an elevation of 340 m above sea level. Peshawar covers an area of 1257 km^2 and has a population of 1,218,773 making it the biggest city in Khyber Pakhtunkhwa. Peshawar is predominantly hot during summer (May–Mid July) with an average maximum temperature of 40 °C followed by monsoon and cold winter.

Local vehicular emission, fossil fuel energy plants and industrial processes are the significant sources of air pollution in Peshawar. Wind direction and wind speed also play a crucial role to observe transboundary pollution build-up. Furthermore, at this site, the distribution and dispersion of air pollution are further impacted by the nearby buildings, and its proximity to Grand Trunk Road, creating a built-up street canyon environment, generated primarily from nearby, increasing traffic pollution.

The air quality monitoring sensor (AQMS) was installed at the University of Peshawar's Physics Department Building (see Figure A1) at 6 m height from the ground surface level. It is described as an urban background site.

Sheffield (53°23′ N, 1°28′ W) is a geographically diverse city located in county South Yorkshire, UK, built on several hills thus situated at an elevation of 29–500 m above sea level. Sheffield covers a total area of 367.9 km^2 with a growing population of 582,506. Sheffield is claimed to be the "greenest city" in England by the local city council. Sheffield enjoys a temperate climate with July considered the hottest month, with an average maximum temperature of 20.8 °C.

The air pollution in the city is primarily due to both road transport and industry, and to a lesser extent, fossil fuel-run processes, such as energy supply and commercial or domestic heating systems (for example, wood burners).

The AQMS is installed at 2.5 m height from the elevated ground surface level at the playground of Hunter's Bar Infants School (see Figure A2), which lies in close proximity to a busy roundabout, and at the intersection of Ecclesall Road, Brocco Bank, Sharrow Vale Road and Junction Road; thus, traffic is the primary source of pollution. It is also described as an urban background site.

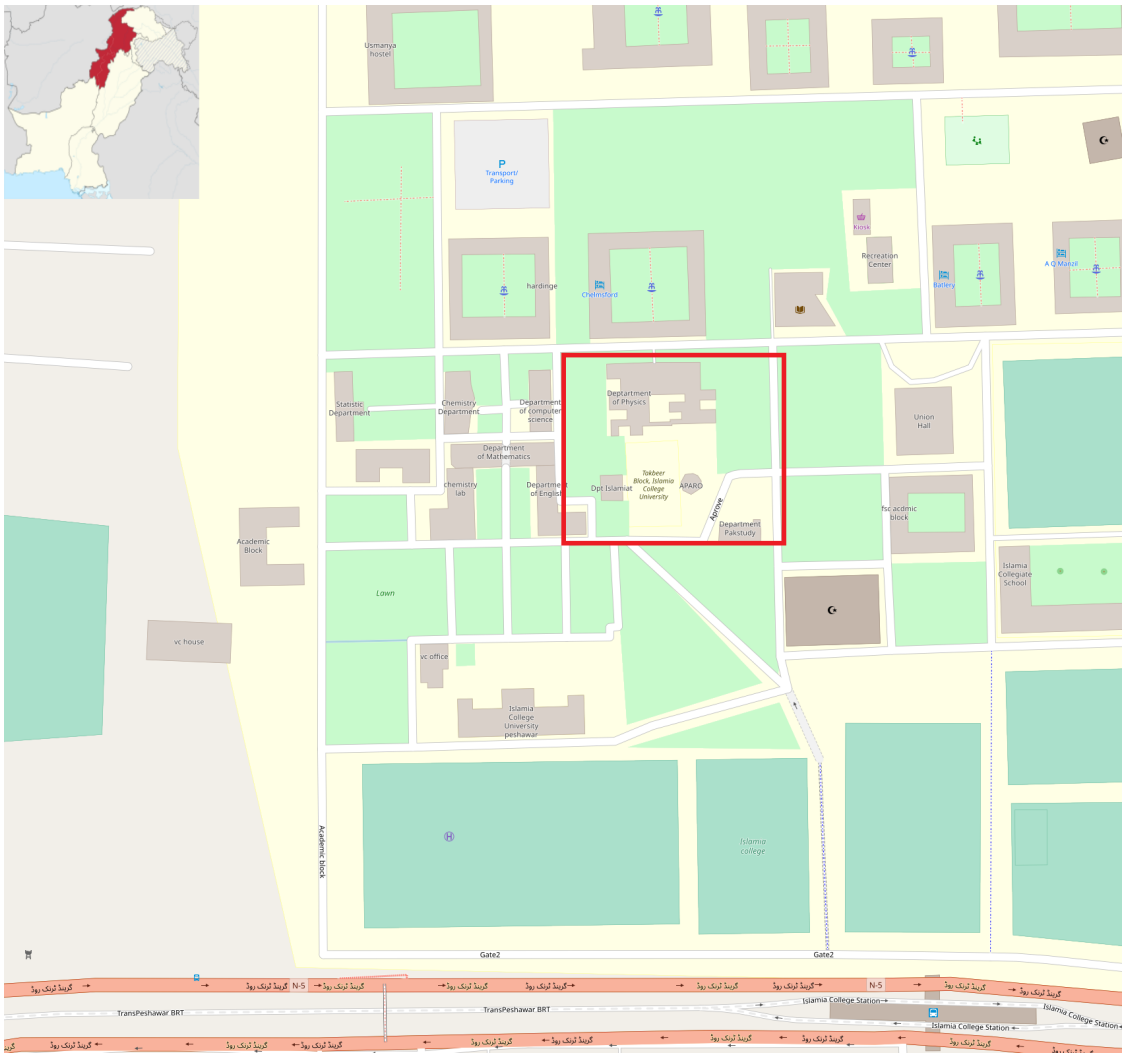

Figure A1. Peshawar study site © OpenStreetMap contributors.

In our case, the AQMSs are commercially low cost sensor nodes AQMesh. They have been deployed at the two sites in Peshawar and Sheffield. A "black box" post calibration is applied to the data by the manufacturer to eliminate the impact of humidity and temperature on the sensor and to eliminate cross sensitivity. The data are aggregated and sampled every 15 min. The data collected from these nodes are transferred to the cloud-based AQMesh database via standard GPRS communication integrated. The data are then accessed through the dedicated API.

Figure A2. Sheffield study site © OpenStreetMap contributors.

Appendix B. The WHO Concentration Criteria for Pollutants

All data from 'WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide' [26].

WHO NO_2

Table A1. WHO Nitrogen dioxide guidelines.

Nitrogen Dioxide	Annual Mean	1-h Mean
NO_2	40 $\bar{g}/m^3$	200 $\bar{g}/m^3$

WHO SO_2

Table A2. WHO sulfur dioxide guidelines.

Sulfur Dioxide	24-h Mean	10-min Mean
SO_2	20 $\bar{g}/m^3$	500 $\bar{g}/m^3$

WHO $PM_{2.5}$ and PM_{10}

Table A3. WHO particulate matter guidelines.

Particulate Matter	Annual Mean	24-h Mean
$PM_{2.5}$	10 $\bar{g}/m^3$	25 $\bar{g}/m^3$
PM_{10}	20 $\bar{g}/m^3$	50 $\bar{g}/m^3$

WHO O_3

Table A4. WHO Ozone guidelines.

Ozone	8-h Mean
O_3	100 $^{-}$g/m^3

Appendix C. Approximated Derivatives of SE Kernel

By specifying a kernel function, we can obtain analytical forms of Equations (28) and (29) immediately. In this paper, we adopt the widely used SE kernel shown in Equation (A1) as an example.

$$k_{SE}(x, x') = s_f^2 \exp\Big(-\frac{(x-x')^2}{2l^2}\Big). \tag{A1}$$

There are two hyperparameters, i.e., the signal variance s_f and length-scale l are involved. Equations (A2) and (A3) show the expectation (prediction mean) partial derivative (EPD) and covariance partial derivative (CPD) of s_f,

$$\begin{aligned}
&\Big(\frac{\partial \bar{\mathbf{f}}_*}{\partial \theta_s}\Big)_o\Big|_{\theta_s=s_f} \\
&=\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(k_{oj}\frac{\partial d_{ji}}{\partial s_f}+\frac{\partial k_{oj}}{\partial s_f}d_{ji}\Big)y_i \\
&=\sum_{i=1}^{n}\sum_{j=1}^{n}y_i\left\{\begin{array}{ll} 0, & j\neq i \\ 0, & j=i \end{array}\right\},
\end{aligned} \tag{A2}$$

$$\begin{aligned}
&\Big(\frac{\partial \mathrm{cov}(\mathbf{f}_*)}{\partial \theta_s}\Big)_{oo}\Big|_{\theta_s=s_f} \\
&=\frac{\partial \mathbf{K}(\mathbf{X}_*,\mathbf{X}_*)_{oo}}{\partial s_f}-\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\frac{\partial k_{oj}}{\partial s_f}d_{ji}k_{oi}+k_{oj}\frac{\partial d_{ji}}{\partial s_f}k_{oi}-k_{oj}d_{ji}\frac{\partial k_{oi}}{\partial s_f}\Big) \\
&=2s_f-\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\begin{array}{ll} 2s_f\exp(-\frac{(x_o-x_j)^2+(x_j-x_i)^2+(x_o-x_i)^2}{2l^2}), & j\neq i \\ -2s_f\exp(-\frac{(x_o-x_j)^2+(x_o-x_i)^2}{2l^2}), & j=i \end{array}\right\}.
\end{aligned} \tag{A3}$$

While the derivatives of l are given in Equations (A4) and (A5),

$$\begin{aligned}
&\Big(\frac{\partial \bar{\mathbf{f}}_*}{\partial \theta_s}\Big)_o\Big|_{\theta_s=l} \\
&=\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(k_{oj}\frac{\partial d_{ji}}{\partial l}+\frac{\partial k_{oj}}{\partial l}d_{ji}\Big)y_i \\
&=\sum_{i=1}^{n}\sum_{j=1}^{n}y_i\left\{\begin{array}{ll} -\exp(-\frac{(x_o-x_j)^2+(x_j-x_i)^2}{2l^2})\frac{(x_o-x_j)^2+(x_j-x_i)^2}{l^3}, & j\neq i \\ \exp(-\frac{(x_o-x_j)^2}{2l^2})\frac{(x_o-x_j)^2}{l^3}, & j=i \end{array}\right\},
\end{aligned} \tag{A4}$$

$$\begin{aligned}
&\Big(\frac{\partial \mathrm{cov}(\mathbf{f}_*)}{\partial \theta_s}\Big)_{oo}\Big|_{\theta_s=l} \\
&=\frac{\partial \mathbf{K}(\mathbf{X}_*,\mathbf{X}_*)_{oo}}{\partial l}-\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\frac{\partial k_{oj}}{\partial l}d_{ji}k_{oi}+k_{oj}\frac{\partial d_{ji}}{\partial l}k_{oi}-k_{oj}d_{ji}\frac{\partial k_{oi}}{\partial l}\Big) \\
&=\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\begin{array}{ll} \exp(-\frac{(x_o-x_j)^2+(x_j-x_i)^2+(x_o-x_i)^2}{2l^2}) & \\ *\ \frac{(x_o-x_j)^2+(x_j-x_i)^2-(x_o-x_i)^2}{l^3}s_f^2, & j\neq i \\ 0, & j=i \end{array}\right\}.
\end{aligned} \tag{A5}$$

References

1. WHO. *WHO Global Ambient Air Quality Database (Update 2018)*; World Health Organization: Geneva, Switzerland, 2018.
2. Landrigan, P.J. Air pollution and health. *Lancet Public Health* **2017**, *2*, e4–e5. [CrossRef]
3. WHO. *Health Effects of Particulate Matter: Policy Implications for Countries in Eastern Europe, Caucasus and Central Asia (2013)*; World Health Organization Regional Office for Europe: Copenhagen, Denmark, 2013.
4. Chen, H.; Kwong, J.C.; Copes, R.; Tu, K.; Villeneuve, P.J.; Van Donkelaar, A.; Hystad, P.; Martin, R.V.; Murray, B.J.; Jessiman, B.; et al. Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: A population-based cohort study. *Lancet* **2017**, *389*, 718–726. [CrossRef]
5. Khreis, H.; de Hoogh, K.; Nieuwenhuijsen, M.J. Full-chain health impact assessment of traffic-related air pollution and childhood asthma. *Environ. Int.* **2018**, *114*, 365–375. [CrossRef] [PubMed]
6. *Improving Air Quality in the Tackling Nitrogen Dioxide in Our Towns and Cities*; UK Overview Document; Department for Environment, Food & Rural Affairs and Department for Transport: London, UK, 2017.
7. Rai, A.C.; Kumar, P.; Pilla, F.; Skouloudis, A.N.; Di Sabatino, S.; Ratti, C.; Yasar, A.; Rickerby, D. End-user perspective of low-cost sensors for outdoor air pollution monitoring. *Sci. Total Environ.* **2017**, *607*, 691–705. [CrossRef] [PubMed]
8. Zheng, T.; Bergin, M.H.; Sutaria, R.; Tripathi, S.N.; Caldow, R.; Carlson, D.E. Gaussian process regression model for dynamically calibrating and surveilling a wireless low-cost particulate matter sensor network in Delhi. *Atmos. Meas. Tech.* **2019**, *12*, 5161–5181. [CrossRef]
9. Shen, J. $PM_{2.5}$ concentration prediction using times series based data mining. *City* **2012**, *2013*, 2014–2020.
10. Silibello, C.; D'Allura, A.; Finardi, S.; Bolignano, A.; Sozzi, R. Application of bias adjustment techniques to improve air quality forecasts. *Atmos. Pollut. Res.* **2015**, *6*, 928–938. [CrossRef]
11. Specht, D.F. A general regression neural network. *IEEE Trans. Neural Netw.* **1991**, *2*, 568–576. [CrossRef] [PubMed]
12. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
13. Lin, K.; Pai, P.; Yang, S. Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms. *Appl. Math. Comput.* **2011**, *217*, 5318–5327. [CrossRef]
14. Mao, Y.; Lee, S. Deep Convolutional Neural Network for Air Quality Prediction. *J. Phys. Conf. Ser.* **2019**, *1302*, 032046. [CrossRef]
15. Garriga-Alonso, A.; Rasmussen, C.E.; Aitchison, L. Deep convolutional networks as shallow gaussian processes. *arXiv* **2018**, arXiv:1808.05587.
16. Bai, L.; Wang, J.; Ma, X.; Lu, H. Air pollution forecasts: An overview. *Int. J. Environ. Res. Public Health* **2018**, *15*, 780. [CrossRef] [PubMed]
17. Wang, P.; Mihaylova, L.; Munir, S.; Chakraborty, R.; Wang, J.; Mayfield, M.; Alam, K.; Khokhar, M.F.; Coca, D. A computationally efficient symmetric diagonally dominant matrix projection-based Gaussian process approach. *Signal Process.* **2021**, *183*, 108034. [CrossRef]
18. Burt, D.R.; Rasmussen, C.E.; Van Der Wilk, M. Rates of Convergence for Sparse Variational Gaussian Process Regression. *arXiv* **2019**, arXiv:1903.03571.
19. Liu, H.; Ong, Y.S.; Shen, X.; Cai, J. When Gaussian process meets big data: A review of scalable GPs. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 4405–4423. [CrossRef] [PubMed]
20. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; Number 3; MIT Press: Cambridge, MA, USA, 2006.
21. Wu, M.; Yin, B.; Wang, G.; Dick, C.; Cavallaro, J.R.; Studer, C. Large-scale MIMO detection for 3GPP LTE: Algorithms and FPGA implementations. *IEEE J. Sel. Top. Signal Process.* **2014**, *8*, 916–929. [CrossRef]
22. Chen, Z.; Wang, B. How priors of initial hyperparameters affect Gaussian process regression models. *Neurocomputing* **2018**, *275*, 1702–1710. [CrossRef]
23. Zhu, D.; Li, B.; Liang, P. On the matrix inversion approximation based on Neumann series in massive MIMO systems. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 1763–1769.
24. Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In Proceedings of the Artificial Intelligence and Statistics, Clearwater Beach, FL, USA, 16–18 April 2009; pp. 567–574.
25. Snelson, E.; Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, 4–9 December 2006; pp. 1257–1264.
26. *WHO. Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulphur Dioxide. Global Update 2005*; World Health Organization: Geneva, Switzerland, 2006.