© 2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/
The definitive publisher version is available online at http://doi.org/10.1016/j.scs.2019.101970

Urban Air Pollution Estimation Using Unscented Kalman Filtered Inverse Modeling With Scaled Monitoring Data

Abstract

The increasing rate of urbanization requires effective and reliable techniques for air quality monitoring and control. For this, The Air Pollution Model and Chemical Transport Model (TAPM-CTM) has been developed and used in Australia with emissions inventory data, synoptic data and terrain data used as its input parameters. Since large uncertainties exist in the emissions inventory (EI), further refinements and improvements are required for accurate air quality prediction. This study evaluates the performance of urban air quality forecasting, using TAPM-CTM, and improves accuracy of air pollution estimation by using a two-stage optimization technique to upgrade EI with validation from monitoring data. The first stage is based on statistical analysis for EI correction and the second stage is based on the unscented Kalman filter (UKF) to take into account the spatio-temporal distributions of air pollutant levels utilizing a Matérn covariance function. The predicted nitrogen monoxide (NO) and nitrogen dioxide (NO_2) concentrations with *a priori* emissions are first compared with observations at monitoring stations in the New South Wales (NSW). Ozone (O_3) is also considered since at the ground level it represents a major air pollutant affecting human health and the environment. In the second stage, with the improved EI, TAPM-CTM model errors are reduced further by using the UKF to calibrate EI. Results obtained show effectiveness of the proposed technique, which is promising for air quality inverse modeling, an important aspect of air pollution control in smart cities to achieve environmental sustainability.

Keywords: Urban air quality, Unscented Kalman Filter, TAPM-CTM, Emissions inventory, Nitrogen oxide, Nitrogen dioxide, Ozone

1. Introduction

The idea of Smart Cities (SCs) has emerged as a new trend for urbanization. In the context of environmental sustainability, one of the requirements for SCs is to monitor air quality, particularly in a metropolis. Air pollution sources located in modern cities include industrial enterprises, municipal services, motor transport, biogenic and geogenic emissions and many others, which cause adverse effects on the quality of the atmospheric environment and health of municipal residents. In many countries, the city authorities have taken steps to achieve better air quality by monitoring and using air quality simulation software to predict air quality under different scenarios. Air quality modeling is mainly based on the forward theory, whereby its gas concentration prediction is obtained from some physical or mathematical model using a given set of model parameters (and perhaps some other appropriate information, such as geometry, shape, size, contrast etc.). In [18], a forward model was used to simulate transport air pollutants and their dispersion in different seasons in the coastal metropolitan city of Chennai, for assessment of ground level NO_X concentrations. Air pollutant levels were estimated and predicted using the forward approach for Qatar and nine major European cities respectively by Shaban *et al.* [19] and Nanaki *et al.*[20]. Table 1 summarizes some relevant studies on air quality research, the methods used and forecasting models, published in the past decade. Of interest are such pollutants as ozone (O_3), carbon monoxide (CO), carbon dioxide (CO_2), nitrogen monoxide (NO), nitrogen dioxide (NO_2), benzene (C_6H_6), particulate matter (PM_{10} and $PM_{2.5}$), sulfur dioxide (SO_2), ammonia (NH_3) and formaldehyde (HCHO). Other physical parameters of concern are noise level, temperature, relative humidity (RH) and planetary boundary-layer (PBL). In terms of methodology, techniques used include Neural Networks (NNs), fuzzy logic, Variational Mode Decomposition (VMD), Sample Entropy (SE), Long Short-Term Memory (LSTM) neural network, Least Squares Support Vector Machine (LSSVM), Morgan-Mercer-Flodin (MMF), source-Receptor matrix (SRM), Ensemble Kalman Filter (EnKF), Genetic Algorithm (GA), Extended Fractional Kalman Filter (EFKF), Normalized Mean Bias (NMB), and Raman returns. The corresponding models of interest have been SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCIAMACHY), Lagrangian Particle Dispersion Model (LPDM), Nested Air Quality Prediction Modeling System (NAQPMS), Community Multiscale Air Quality (CMAQ), Weather

Table 1: **Summary of studies on forecasting air pollutants and air quality indexes using different models.**

Pollutants/ Parameters	**Methodology**	**Models used/ Instruments used**	**Authors**
O_3	NN	SCIMACHY	Sellitto *et al.* [1]
O_3	NN	Satellite UV measurements	Xu *et al.* [2]
Noise level, CO, NO_2, O_3, C_6H_6 and PM_{10}	Fuzzy	Monitoring station data	Silva & Mendes [3]
HCHO, PM_{10}, CO_2, CO, Temperature and RH	Fuzzy	Sensor data	Yuan *et al.* [4]
$PM_{2.5}$, SO_2, NO_2 and CO	NN	Monitoring station data	Liu *et al.* [5]
$PM_{2.5}$, PM_{10}, CO, O_3, SO_2 and NO_2	VMD, SE and LSTM NN	Monitoring station data	Wu & Lin [6]
$PM_{2.5}$	LSSVM	Monitoring station data	Sun & Sun [7]
PM_{10}	MMF	Monitoring station data	Ortolani & Vitale [8]
$PM_{2.5}$	SRM	LPDM	Yu *et al.* [9]
CO	EnKF	NAQPMS	Tang *et al.* [10]
PBL height	Back-trajectory	CMAQ and WRF analysis	Banks & Baldasano [11]
PM_{10} and SO_2	GA	CMAQ	Li *et al.* [12]
NO, NO_2 and O_3	EFKF	TAPM-CTM	Metia *et al.* [13]
NO_2	UKF	OMI and TAPM-CTM	Metia *et al.* [14]
NH_3 and $PM_{2.5}$	Bulk emissions into daily	CMAQ	Hsu *et al.* [15]
NH_3 and $PM_{2.5}$	NMB	CMAQ and CIMS	Bray *et al.* [16]
$PM_{2.5}$	Raman returns	CMAQ and AOD	Vladutescu *et al.* [17]

Research and Forecasting (WRF)and The Air Pollution Model with Chemical Transport Model (TAPM-CTM), and the corresponding instruments are satellite ultra violet (UV) measurements, monitoring station data, sensor data, Ozone Monitoring Instrument (OMI), Chemical Ionization Mass Spectrometry (CIMS), and Aerosol Optical Depth (AOD). Despite numerous methods have been proposed for air quality monitoring in cities, prediction of air pollutants profiles often depends on the location of the monitoring stations. Also, due to large uncertainties involved, improving performance of the air pollution estimation remains a difficult problem. In this work, a new method will be developed for achieving higher estimation accuracy for urban air pollution via optimization of a scaling factor and enhanced calibration with Matérn covariance-based unscented Kalman filtering.

In the present study, a two-stage optimization approach is proposed and implemented on emissions inventory (EI) for TAPM-CTM. Statistical analysis is used to improve EI, via a scaling factor for local suburbs with real data settings, and then a UKF based Matérn covariance function is used to smooth inventory data to ameliorate prediction accuracy. Finally, NO and NO_2 concentrations are simulated using TAPM-CTM with the recalibrated EI, and modeling results are then compared with monitoring station data.

This paper is organized as follows. After the introduction, Section 2 describes the modeling and estimation system framework. Section 3 presents the algorithm description. Results are presented in Section 4. Finally, a conclusion is drawn in Section 5.

2. System description

Emissions inventory plays an important role in air quality modeling. Figure 1 shows TAPM-CTM for air quality prediction, using synoptic, terrain and emissions data as the input.

2.1. Monitoring station data and TAPM-CTM

To provide the Australian state of New South Wales with up-to-date information about air quality, monitoring stations have been set up in Sydney and several regional areas. These stations monitor particulate matters (PM_{10}, $PM_{2.5}$), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2) and visibility. Meteorological parameters such as wind speed and direction, air temperature and humidity are also recorded. Data are obtained from the Office of Environment and Heritage (OEH) for this study. The modeling system consists of a TAPM prognostic meteorological model and a chemical transport model (CTM) which the Commonwealth Scientific and Industrial Research Organisation (CSIRO) originally developed for use with the Australian Air Quality Forecasting System.

2.2. Emissions inventory

The air emissions inventory is a detailed listing of pollutants discharged into the atmosphere by each source type during a given time period and at a specific location. These data are stored in a database maintained by

OEH. The inventory includes emissions from biogenic (i.e. natural and living), geogenic (i.e. natural non-living) and anthropogenic (i.e. human-made) sources. The Emissions Data Management System (EDMS) v2.0 has a number of features, including emissions charting according to a local government area (LGA) and emissions forecasting up to 2036 with air pollutant emissions given in $gm/m^2/sec$ and spatial resolution of $1km$ by $1km$.

2.3. Terrain and synoptic data

Terrain data are as input to TAPM-CTM based on the vertical and horizontal dimensions of land surface. Over a large area, terrain conditions can affect surface water flow and distribution, hence weather and climate patterns. The Bureau of Meteorology (BOM) is an Executive Agency of the Australian Government, responsible for providing weather services to Australia and surrounding areas. BOM has provided meteorological data to many organisations including OEH. As such, information of temperature, specific humidity, wind directions and wind speeds are made available as input to TAPM-CTM and other forecasting models.

2.4. Statistical inventory

Emissions inventory, initial conditions, lateral boundary conditions and meteorological fields are known as subject to large uncertainties in atmospheric CTMs [21]. The primary source of model errors in air quality forecasts is due to uncertainties in ad hoc inventory [22]. In [23], statistical modeling was used for EI while in [24], different methodologies were adopted, considering Gaussian distributions in statistical modeling. In this paper, errors between the model output and observations are also assumed to follow a Gaussian distribution in nature.

2.5. Inventory improvement with UKF

The UKF has an advantage over the EKF, since it does not use linearization for the prediction of state and error covariances [25, 26]. This facilitates more accurate computation of Kalman gain and error covariance matrices, which can ultimately lead to an enhanced estimation of random process signals. Since large uncertainties exist in EI, further improvements and refinements are required. This is the main reason the UKF is used to calibrate EI in this study.

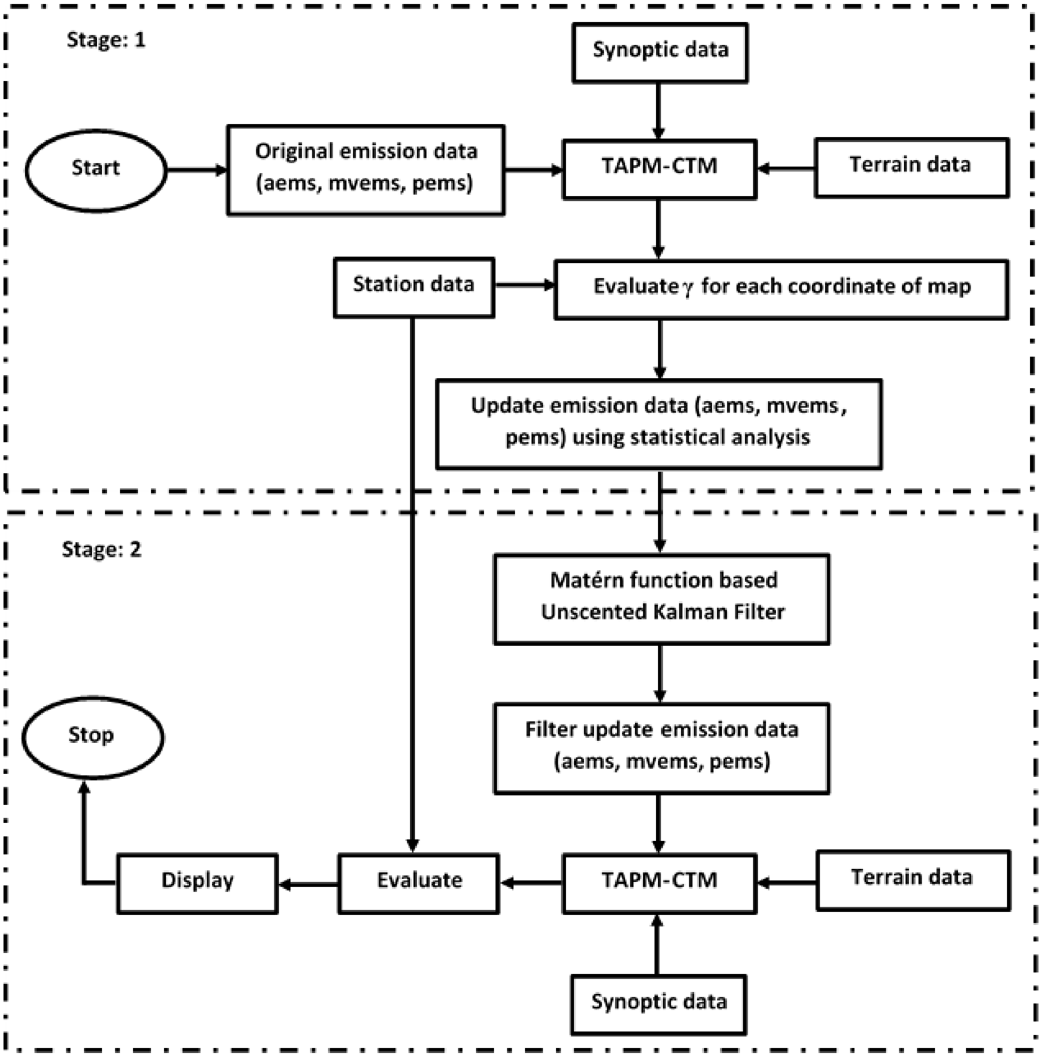

Stage: 1
Synoptic data
Start
Original emission data (aems, mvems, pems)
TAPM-CTM
Terrain data
Station data
Evaluate γ for each coordinate of map
Update emission data (aems, mvems, pems) using statistical analysis
Stage: 2
Matérn function based Unscented Kalman Filter
Stop
Filter update emission data (aems, mvems, pems)
Display
Evaluate
TAPM-CTM
Terrain data
Synoptic data

Figure 1: Flowchart of optimizing emissions inventory with inverse modeling.

3. Algorithm description

The observation vector comprises pollutant concentrations collected at the surface monitoring stations over the region of interest (here, the state of New South Wales). The relationship between observation vector y with noise and the state vector x can be described as follows under a forward model

$$y = H(x, b) + \epsilon, \tag{1}$$

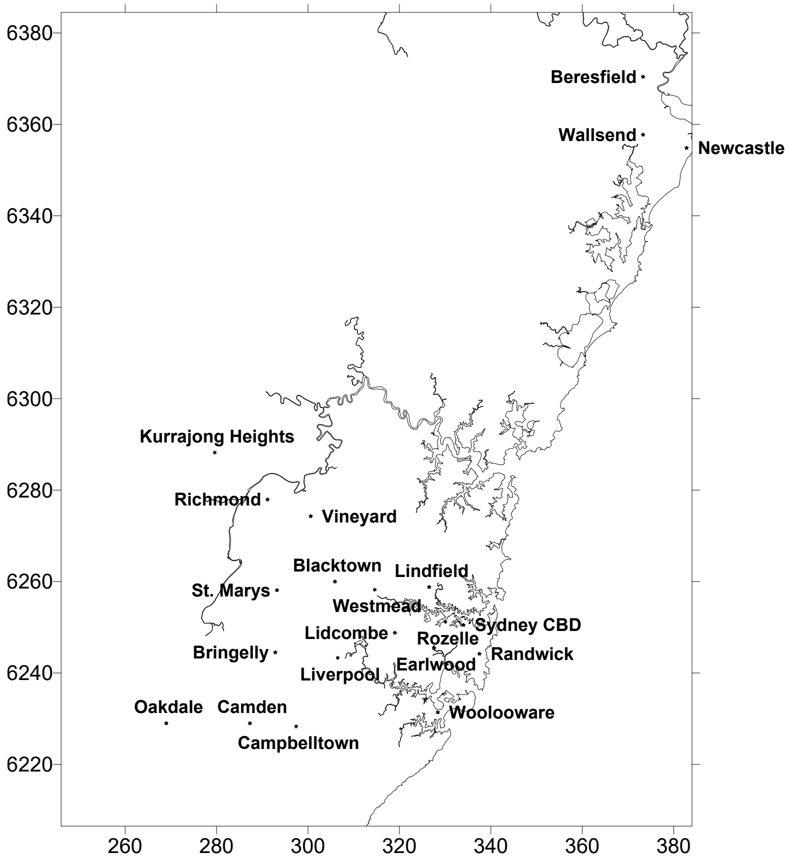


Figure 2: Locations of monitoring stations across the NSW.

where b is the model parameter vector, ϵ is the noise vector, and H stands for the forward model. The entries of the state vector x are the pollutant concentrations obtained at different grid levels across NSW, while vector b includes parameters describing terrain and synoptic data. To solve (eqn. 1) a two-stage inverse modeling framework is proposed as shown in Figure 1. Herein, Stage 1 is based on statistical analysis and Stage 2 is based on Matérn functioned based Unscented Kalman filtering of emissions inventory data.

3.1. Stage 1

A mean error is evaluated for each coordinate output air quality variable of the map as follows,

$$\mu_{i,j}^{error} = \frac{\sum_1^n \left(S_{i,j} - S'_{i,j}\right)}{n}, \tag{2}$$

where $S_{i,j}$ stands for monitoring station data at (i,j) coordinate, $S^{'}_{i,j}$ stands for forward model output from TAPM-CTM, n is the number of samples, and $\mu^{error}_{i,j}$ is calculated using (eqn. 2). The standard deviation is evaluated as

$$\sigma^{error}_{i,j} = \sqrt{\frac{\sum_1^n \left(S_{i,j} - S^{'}_{i,j} - \mu^{error}_{i,j}\right)^2}{n-1}}. \tag{3}$$

According to Poor & Verdu [27] and Bich [28], the estimated error following a Gaussian distribution, can be generated from $\sigma^{error}_{i,j}$ and $\mu^{error}_{i,j}$ as

$$f(X_{i,j}|\mu^{error}_{i,j}, (\sigma^{error}_{i,j})^2) = \frac{1}{\sqrt{2\pi(\sigma^{error}_{i,j})^2}} e^{-\frac{(X_{i,j} - \mu^{error}_{i,j})}{2(\sigma^{error}_{i,j})^2}}. \tag{4}$$

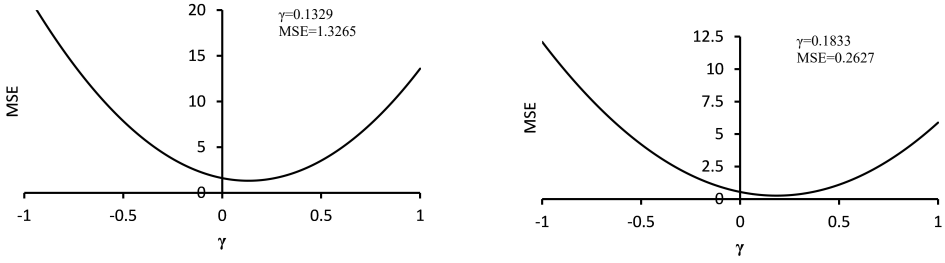
The updated EI is then given by

$$x^{update}_{i,j} = x_{i,j} \pm (\gamma \times X_{i,j}), \tag{5}$$

where "+" is for under estimation and "-" for over estimation, respectively. In this equation, $x_{i,j}$ is the original emissions data at (i,j) coordinate, $X_{i,j}$ is the error, $x^{update}_{i,j}$ is the update the same coordinate, and γ represents the scaling factor. This factor depends spatially on pollutant variables and is a function of emissions data. The model data $S^{'}_{i,j}$ is generated from TAPM-CTM with $x_{i,j}$ as EI input. The mean square error (MSE) can be calculated using $x^{update}_{i,j}$ at (i,j) coordinate as,

$$MSE = \frac{1}{n} \sum_1^n \left[S_{i,j} - (\gamma \times S^{'}_{i,j})\right]^2. \tag{6}$$

The database EDMS v2.0 is mainly used to export emissions data to the air quality modeling software. Since, EDMS v2.0 normally generates separate data, based on the type of emissions rather than the total summation of the grid data, it is not straight forward to use this system for air quality modeling. The scaling remains therefore a tedious process. For example, NO_X is generated due to biogenic, geogenic and anthropogenic emissions, in which biogenic and geogenic emissions contribute 3.1% while anthropogenic emissions contributes 96.9%. In terms of emissions type area sources contribute 61.9% of NO_X emissions, 1.2% is from point sources, while motor vehicles are responsible for the rest 36.9% of NO_X emissions contribution [29].



(a) (b)

Figure 3: MSE versus γ at Earlwood monitoring station (a) NO, and (b) NO_2.

The scaling process is optimized by using (eqn. 6), where factor γ is varied in the interval $-1.0 \leq \gamma \leq 1.0$ to minimize the mean square error (MSE). The scaling factor varies according to the location of air quality data collection for which Figure 2 shows coordinates of NSW monitoring stations. For example, the optimal value of γ at Earlwood monitoring station is found at 0.1329 and 0.1833 respectively for NO and NO_2, obtained by minimization of MSE (eqn. 6), as shown in Figure 3. Similarly, for other monitoring stations, the scaling factors are calculated from minimization of MSE (eqn. 6). The scaling factor at locations in between monitoring stations is averaged accordingly. These scaling factors are used in (eqn. 5) to update emissions inventory. The updated EI is defined as $x_{i,j}^{update}$. These data are further improved in the next stage.

3.2. Stage 2

The UKF is used to calibrate $x_{i,j}^{update}$ for further improvements and refinements. In [13, 30], Gaussian process models adopted in the state space can be solved by the classical Kalman filtering theory where the link between state-space model and Matérn covariance function is analytically obtained. Here, Matérn function based UKF is used to refine EI. As such, the covariance is of a function family [31].

$$\upsilon(S_{i,j}, S_{k,l}) = \frac{\sigma_\nu^2}{2^{\phi-1}\Gamma(\phi)}(\lambda d_{ij,kl})K_\phi(\lambda d_{ij,kl}), \lambda > 0, \phi \geq 0, \tag{7}$$

where $d_{ij,kl}$ is the distance between sites $S_{i,j}$ and $S_{k,l}$, $K_\phi(\cdot)$ is the modified Bessel function of order ϕ, $\Gamma(\phi)$ is a Gamma function, and σ_ν^2 is the variance. The corresponding 3^{rd} order system for air pollutant dynamics is described

Table 2: Summary statistics of predicted using different emissions inventory

Station	Pollutant	Original Emissions Inventory			Statistical Emissions Inventory			UKF Emissions Inventory		
		MSE	R^2	p-value	MSE	R^2	p-value	MSE	R^2	p-value
Earlwood	NO	3.6016	0.5366	4.97×10^{-2}	1.4895	0.7159	4.09×10^{-2}	0.3724	0.8488	2.05×10^{-2}
	NO_2	5.8868	0.5191	1.14×10^{-3}	0.8626	0.7735	1.13×10^{-5}	0.2013	0.8601	1.91×10^{-6}
	O_3	4.0274	0.5860	4.98×10^{-2}	2.1453	0.7583	1.43×10^{-5}	0.5245	0.8165	5.89×10^{-6}
Lindfield	NO	4.4517	0.5678	9.84×10^{-3}	1.1296	0.7704	6.21×10^{-4}	0.2824	0.8364	1.88×10^{-4}
	NO_2	6.5598	0.5219	5.50×10^{-6}	0.3426	0.7199	3.82×10^{-8}	0.0818	0.8657	1.48×10^{-9}
	O_3	4.5094	0.5161	6.69×10^{-1}	2.2201	0.7561	1.61×10^{-1}	0.5550	0.8531	7.25×10^{-2}
Richmond	NO	2.9669	0.6331	1.21×10^{-6}	0.0378	0.7855	3.44×10^{-7}	0.0094	0.8633	1.97×10^{-7}
	NO_2	5.7423	0.5343	5.75×10^{-20}	0.3309	0.7338	2.19×10^{-21}	0.0766	0.8324	1.17×10^{-23}
	O_3	7.6874	0.5592	4.42×10^{-8}	3.8256	0.7173	1.75×10^{-8}	0.9471	0.8123	7.00×10^{-9}

in [13]

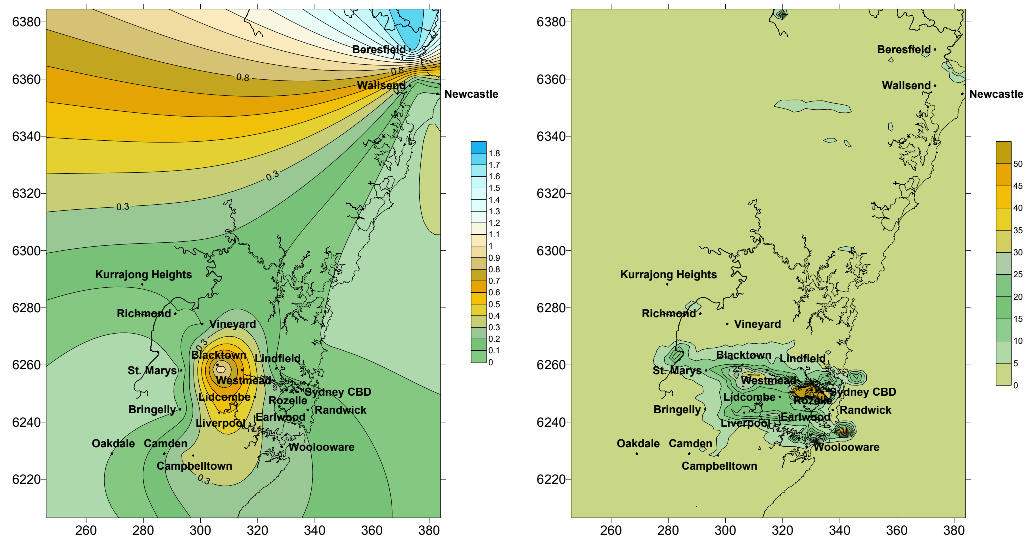
$$\begin{aligned}\frac{dx(t)}{dt} &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\lambda^3 & -3\lambda^2 & -3\lambda \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} q(t) \\ y(t) &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x(t) + r(t), \end{aligned} \tag{8}$$

where $x(t)$ is the state vector, $q(t)$ is the white noise of the process, λ is a coefficient depending on the correlation length and smoothness of the process, $y(t)$ is the model output, and $r(t)$ the measurement noise, respectively for n data points.

Unlike the EKF which uses the first-order approximation of the nonlinear system, the UKF represents a derivative-free alternative with lesser computational complexity [32]. An unscented transform (UT) is implemented to estimate power plant pollutant emissions [14]. Here, the UKF is represented by the following equations:

$$\begin{aligned} x_k &= f_k(x_{k-1}, u_{k-1}) + q_{k-1} \\ y_k &= h_k(x_k, u_k) + r_k, \end{aligned} \tag{9}$$

where $x_k \in \mathbb{R}^n$ is the state, $y_k \in \mathbb{R}^m$ is the measurement, $u_k \in \mathbb{R}^v$ is the input, $q_{k-1} \in \mathbb{R}^n$ is the Gaussian process noise $q_{k-1} \sim \mathcal{N}(0, Q_{k-1})$, $r_k \in \mathbb{R}^m$ is the Gaussian measurement noise $r_k \sim \mathcal{N}(0, R_k)$, and Q_{k-1} and R_k are covariances correspondingly. The design procedure of a generic UKF is summarized in Appendix.

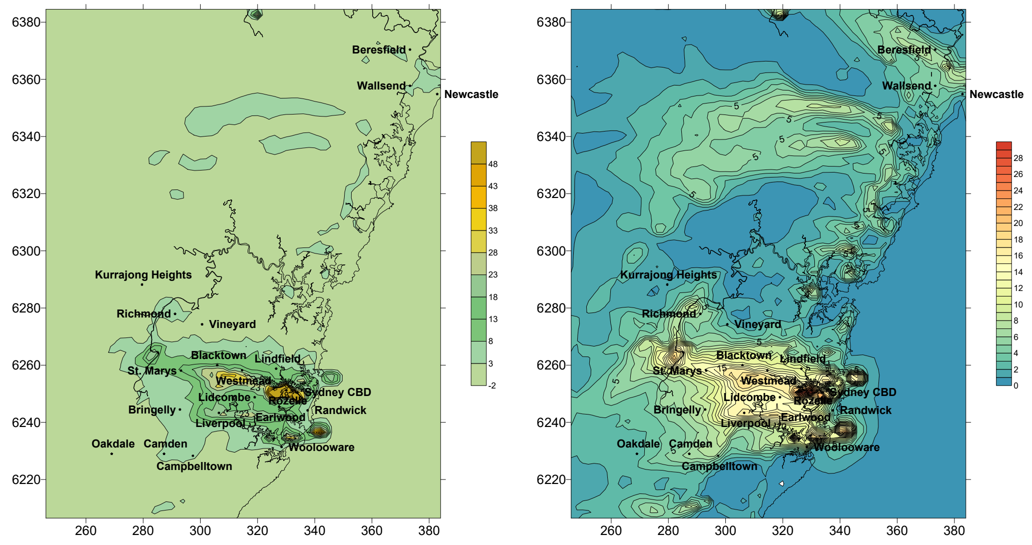


(a) (b)

Figure 4: NO concentration (pphm) distribution on January 8, 2008 at 7 A.M., (a) based on monitoring station data, and (b) based on original inventory.

4. Results and discussion

This section compares the station data with TAPM-CTM prediction data using original EI, statistical calibrated EI and UKF calibrated EI. Figures 4a and 4b show spatial distribution of nitrogen monoxide with monitoring data and original inventory, respectively. Local monitoring stations at Earlwood, Lindfield and Richmond are considered as experimental sites for this study. The choice of these suburbs reflected the variation in emission level, mainly due to the density of motor vehicle in a big city like Sydney. In terms of temporal distribution, there are diurnal variations in nitrogen oxides and ozone concentrations. In general, the O_3 concentration slowly increases after sunrise, reaching its maximum during midday, and then slowly decreases until the next morning. The diurnal cycles of NO and NO_2 are shaped like double waves with the morning peak being higher in magnitude than the evening one. Therefore, 7A.M. and 1 P.M. were considered for spatial plots of nitrogen oxides and ozone, respectively shown in Figure 5 and Figure 6, for



(a) (b)

Figure 5: Pollutant concentration (pphm) difference between original inventory and monitoring station data on January 8, 2008 at 7 A.M., (a) NO, and (b) NO_2.

the difference between original inventory and data collected at a monitoring station.

4.1. Comparison between monitoring station data and original emissions inventory

The performance of original EI is evaluated by comparing it with station data. Table 2 shows the comparison in terms of mean square error (MSE), coefficient of determination (R^2) and probability value (p-value). The MSE and p-value for NO are 13.6016 and 0.0497 respectively. They imply that there are significant differences between station data and TAPM-CTM output with original EI. To illustrate these differences are shown in Figure 5a and Figure 5b, respectively for NO and NO_2. These discrepancies are due to the overestimation of EI and the sparsity of monitoring stations over the region of interest. Indeed, there are only 33 monitoring stations active to monitor air quality across NSW mainly located in the city basin, the state has a 210 km × 270 km grid in EI. As such, 99.94% of its data are estimated

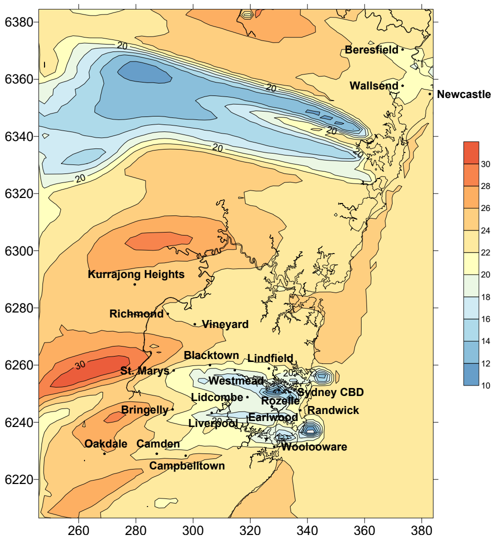
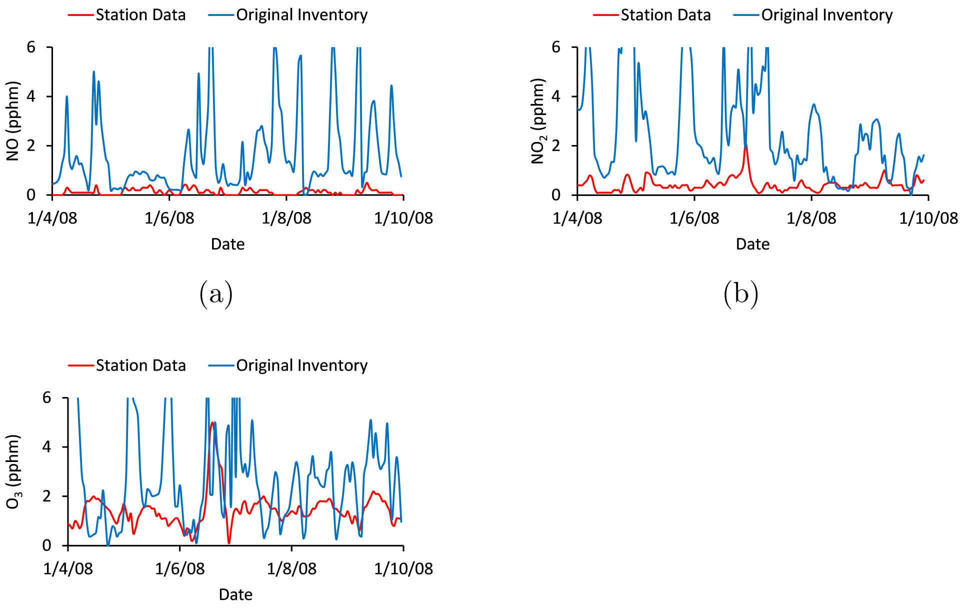


Figure 6: O_3 concentration (pphm) difference between original inventory and monitoring station data on January 8, 2008 at 1 P.M.

using Kriging interpolation, from the data collected at monitoring stations. Similarly, Figure 6 shows the difference between TAPM-CTM based on original EI and monitoring station data of O_3. Diurnal variations of NO, NO_2 and O_3 concentration are shown in Figure 7, indicating clearly the large discrepancy of the original inventory with the station data. These mismatches between original inventory and monitoring station data, apparently higher in municipal areas of the Sydney basin require a suitable method to improve accuracy of air pollution estimation [33]. The overestimation problem is due to activity data and emission factors. Emissions inventory includes emissions from biogenic (i.e. natural) and anthropogenic (i.e. human) derived sources. Biogenic sources cover bushfires, trees and windborne dust. Anthropogenic sources include all emissions activities by human such as quarries, service stations, house painting, mowing, heating, mining, oil refineries, power stations, steelworks, aircraft, railways, and other transport means. Activity data have been obtained from industry groups, government departments and other service providers. Uncertainties from these sources together explain why large discrepancies exist between original inventory and monitoring station data.



(a)

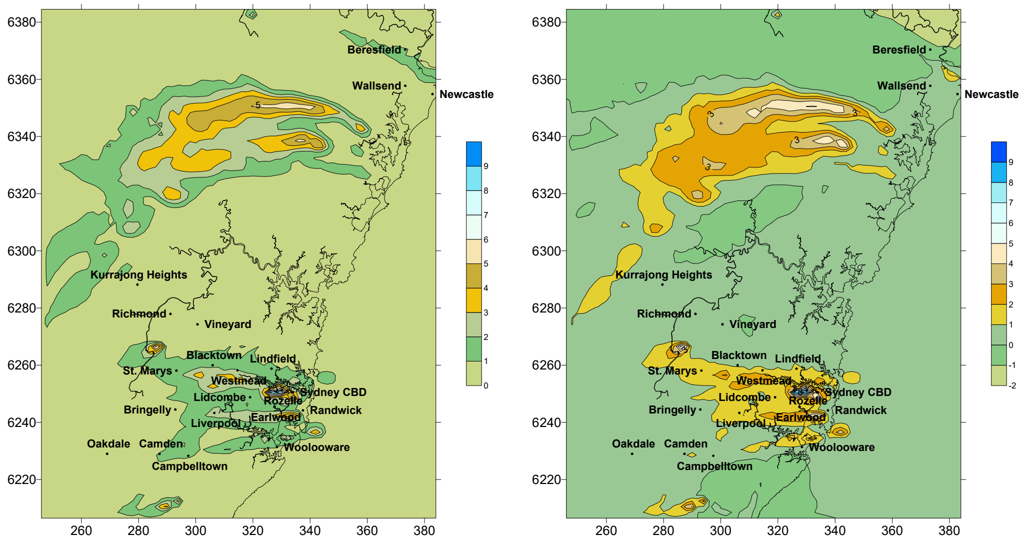
(b)

(c)

Figure 7: Pollutant concentration (pphm) level at Earlwood station from 4^{th}-10^{th} January, 2008, (a) NO, (b) NO_2, and (c) O_3.

4.2. Comparison between station data and statistical emissions inventory

Table 2 shows also detailed statistical analysis of station data, at Stage 1 in the flowchart of Figure1 with TAPM-CTM output based on statistical EI. At this stage, the value of MSE and p-value for NO are respectively 1.4895 and 0.0409, which still show differences between station data and TAPM-CTM output using statistical EI. Indeed, Figure 8a shows the nitrogen monoxide spatial distribution using TAPM-CTM based on statistical EI, while it differences between TAPM-CTM based on statistical EI and station data are plotted in Figure 8b. Similarly, the differences between TAPM-CTM based on statistical EI and station data are plotted in Figure 9a and Figure 9b for NO_2 and O_3 respectively. To show diurnal variations, Figures 10a, 10b and 10c depict respectively the NO, NO_2 and O_3 profiles as extracted originally from TAPM-CTM model, statistical EI, and the measurements at Earlwood station. These figures show that TAPM-CTM prediction output based on statistical EI is closer to monitoring data as compared to original inventory but still has some mismatch, especially for nitrogen dioxide and ozone.



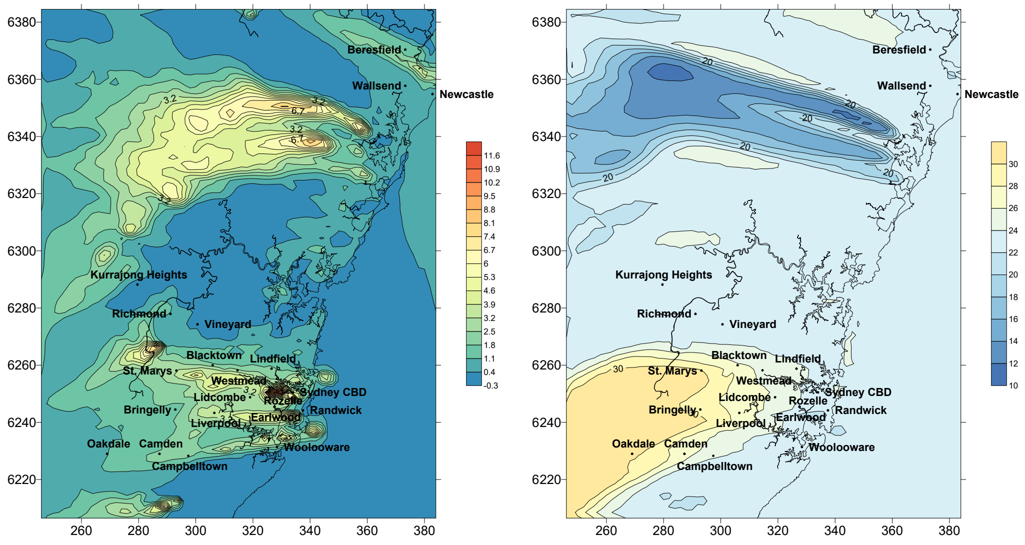
(a)

(b)

Figure 8: NO concentration (pphm) distribution on January 8, 2008 at 7 A.M., (a) based on statistical inventory, and (b) difference between statistical inventory and monitoring station data.

4.3. Comparison between station data and UKF emissions inventory

Values of determination of coefficient are given in Table 2, where R^2 for TAPM-CTM based on UKF EI is shown higher compared to statistical EI and original EI, which clearly demonstrate the advantage of our proposed approach to urban air pollution estimation. Using TAPM-CTM output based on UKF EI, Figure 11a shows the distribution of NO and its difference between TAPM-CTM output based on UKF EI and station data in Figure 11b. Similarly, Figures 12a and 12b show difference between TAPM-CTM output based on UKF EI and station data for NO_2 and O_3, respectively. As can be seen, TAPM-CTM output based on UKF EI has significantly improved the inverse modeling for urban air pollution as compared to monitoring station. Diurnally, Figures 13a, 13c and 13c show variations respectively of NO, NO_2 and O_3 as extracted from TAPM-CTM model with UKF EI, and the measured emissions profiles at Earlwood station. They indicate that TAPM-CTM prediction output based on UKF EI are quite coincident with



(a)

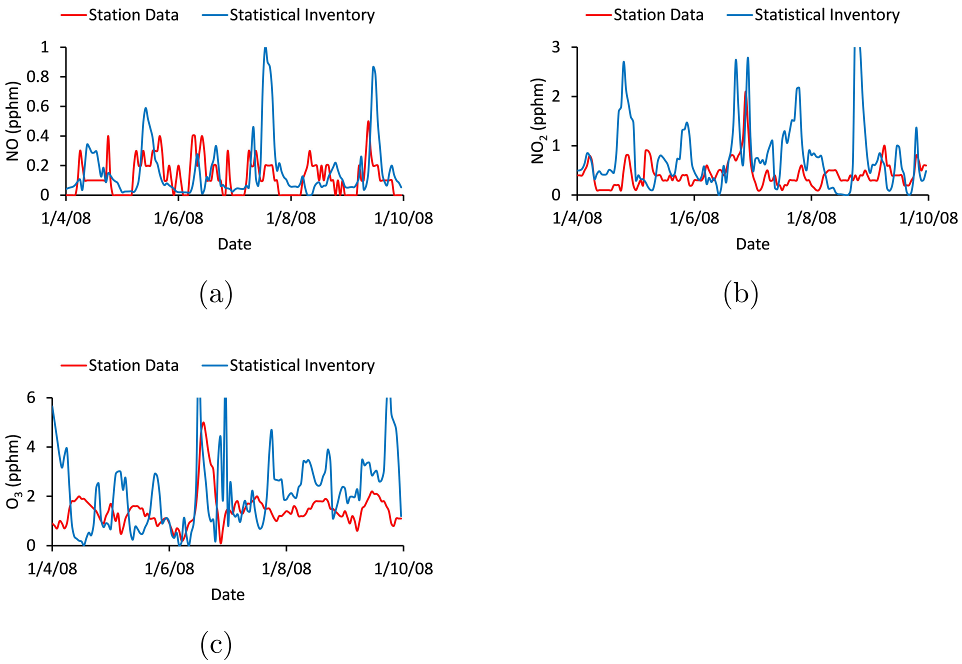
(b)

Figure 9: Pollutant concentration (pphm) difference between statistical inventory and monitoring station data on January 8, 2008, (a) NO_2 at 7 A.M., and (b) O_3 at 1 P.M.

monitoring data. Indeed, UKF EI in Stage 2 with Matérn function-based covariance and the selection of suitable scaling factor γ in Stage 1 for a lowest value of MSE, have resulted in a significant reduction of the difference between estimation and measurements. Notably, this is also reliant on the station location. Similarly to Figure 3, optimal values for the scaling factor were found respectively as 0.1261 for NO and 0.1972 for NO_2 at Lindfield, and as 0.1401 for NO and 0.1841 for NO_2 at Richmond. The results obtained have confirmed the merit of our method in terms of accuracy improvement of TAPM-CTM prediction output towards achieving sustainability for Sydney city.

4.4. Urban air quality monitoring and city sustainability

The fast growth of medium sized and mega cities as well as population increase have severely contributed to air pollution. In addition to problems associated with environmental changes, the rapid urbanization and economic



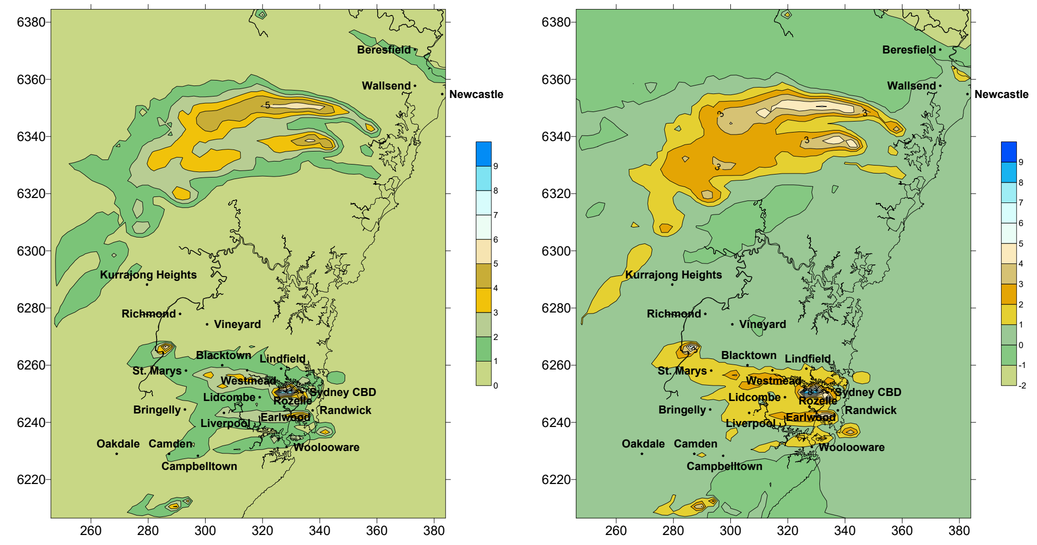
(a)

(b)

(c)

Figure 10: Pollutant concentration (pphm) level at Earlwood station from 4^{th}-10^{th} January, 2008, (a) NO, (b) NO_2, and (c) O_3.

development needs have imposed significant challenges and threats to human health and the sustainability of society. Mainly, suburban sprawl has caused loss of green open spaces, together with tremendous increase in vehicles number and energy consumption. As a result, the compressed city concept has been put forward as a form of sustainable urban development. Air pollution remains one of the critical environmental problems of close relevance to urban developments in all continents. Accurate air quality monitoring and precise modeling are two essential factors for city sustainability. Emissions inventory data administered by city authorities are essential for predicting air quality and hence, need to be accurate for all areas in a city. To this end, a two-stage inverse modeling framework for EI improvement has been proposed and implemented in this paper to remove uncertainties in EI with valid demonstration of air pollutant estimation in the Sydney basin. Indeed, the average value of R^2 between prediction and monitoring data for various pollutants using original EI is 0.5527. After implementing the proposed two-stage EI enhancement, the average coefficient of determination R^2 becomes 0.8432. The proposed technique can therefore be considered as promising for air pollution estimation in sustainable cities.



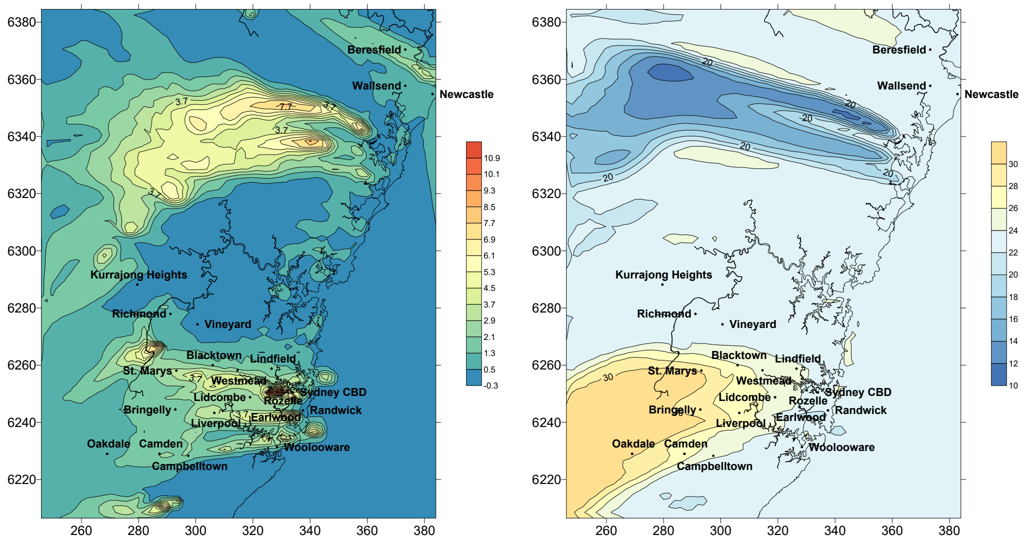
(a)

(b)

Figure 11: NO concentration (pphm) distribution on January 8, 2008 at 7 A.M., (a) based on UKF inventory, and (b) difference between UKF inventory and monitoring station data.

5. Conclusion

In this paper, we have presented a two-stage optimization approach to achieve emissions inventory improvement using statistical analysis and UKF emissions inventory. In the first stage, we have assumed the errors between air quality model output with original EI and monitoring station data is as Gaussian distributions. Emissions inventory is updated by using statistical analysis. In the second stage further refinements and improvements are achieved by using the UKF with Matérn function-based covariance taking into account the correlation length and smoothness of the spaciotemporal profiles of air pollutants. In the sequence, three runs are needed for optimizing EI. The first run is required to compare station data with TAPM-CTM output based on original EI. Another run is for comparing station data with TAPM-CTM output based on statistical EI. The final run is performed to further reduce the difference between station data and TAPM-CTM output



(a)

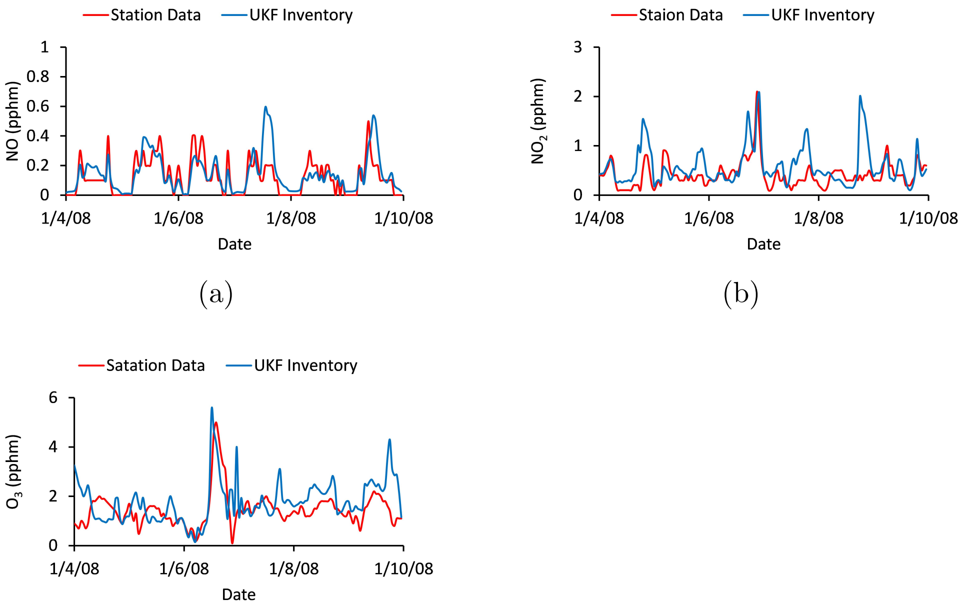
(b)

Figure 12: Pollutant concentration (pphm) difference between UKF inventory and monitoring station data on January 8, 2008, (a) NO_2 at 7 A.M., and (b) O_3 at 1 P.M.

based on UKF EI. Results obtained demonstrate that the proposed method has significantly improved the air quality prediction performance in suburbs of Sydney. The proposed technique can therefore be promising for inverse modeling of air pollution, which remains an important issue for air pollution control in sustainable cities. While being able to enhance the prediction performance, the proposed two-stage inverse modeling framework is still somewhat sensitive to the inputs from the air pollution and and chemical transport models used. This should be considered to further improve the estimation results. Future work will be conducted to address the effect of other inputs of TAPM-CTM, such as initial conditions, meteorological parameters and boundary conditions, where great uncertainties may exist.

Appendix

The generic UKF is summarized as follows:



(a)

(b)

(c)

Figure 13: Pollutant concentration (pphm) level at Earlwood station from 4^{th}-10^{th} January, 2008, (a) NO, (b) NO_2, and (c) O_3.

Step 1: Initialize filtering value

$$\begin{aligned} \hat{x}_0 &= E[x_0] \\ P_0 &= E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T], \end{aligned} \tag{10}$$

where x_0, $\hat{x}_0$ and P_0 represent the initial state vectors, predicted values and covariance, respectively.

Step 2: Calculate the Sigma points

$$\begin{aligned} \mathrm{x}_{k-1} &= \left[\hat{x}_{k-1} \quad \hat{x}_{k-1} \pm (\sqrt{(N+\lambda)p_{k-1}})_i\right] \\ \mathrm{w}_0^{(c)} &= \frac{\lambda}{(N+\lambda)} \qquad \mathrm{w}_0^{(m)} = \frac{\lambda}{(N+\lambda)} + (1 - \alpha^2 + \beta) \\ \mathrm{w}_i^{(m)} &= \mathrm{w}_i^{(c)} = \frac{\lambda}{2(N+\lambda)}, \qquad i = 1, 2, \ldots, 2N \\ \lambda &= \alpha^2(N + \mathrm{v}) - N \end{aligned} \tag{11}$$

where $\mathrm{w}_i^{(m)}$ and $\mathrm{w}_i^{(c)}$ represent the weights of predicted mean and covariance, respectively; λ is a scaling parameter and v is a secondary scaling parameter which is usually set to 0; the distribution of the sample points is represented by α and the approximate prior distribution of x is represented by β.

Step 3: Predict the system state matrix, the covariance matrix and the observation matrix

$$\begin{aligned}
\mathrm{x}_{i,k|k-1} &= \mathrm{x}_{i,k-1} \\
\hat{x}_{k|k-1} &= \sum_{i=1}^{2N} \mathrm{w}_i^{(m)} \mathrm{x}_{i,k|k-1} \\
\hat{P}_{k|k-1} &= \sum_{i=1}^{2N} \mathrm{w}_i^{(c)} g_{i,k|k-1} g_{i,k|k-1}^{T} + Q_{k-1} \\
y_{i,k|k-1} &= h_k(\mathrm{x}_{i,k|k-1}, x(k)) \\
\hat{y}_{k|k-1} &= \sum_{i=1}^{2N} \mathrm{w}_i^{(m)} y_{i,k|k-1},
\end{aligned} \tag{12}$$

where $g_{i,k|k-1} = \mathrm{x}_{i,k|k-1} - \hat{x}_{k|k-1}$.

Step 4: Calculate the covariance matrix of the system

$$\begin{aligned}
P_{\hat{x}_k \hat{y}_k} &= \sum_{i=1}^{2N} \mathrm{w}_i^{(c)} g_{i,k|k-1} m_{i,k|k-1}^{T} \\
P_{\hat{y}_k} &= \sum_{i=1}^{2N} \mathrm{w}_i^{(c)} m_{i,k|k-1} m_{i,k|k-1}^{T} + R_k,
\end{aligned} \tag{13}$$

where $m_{i,k|k-1} = y_{i,k|k-1} - \hat{y}_{k|k-1}$.

Step 5: Calculate the Kalman gain

$$\bar{K} = P_{\hat{x}_k \hat{y}_k} P_{\hat{y}_k}^{-1}. \tag{14}$$

Step 6: Update the system state matrix and covariance matrix

$$\begin{aligned}
\hat{x}_k &= \hat{x}_{k|k-1} + \bar{K}(y_k - \hat{y}_{k|k-1}) \\
P_k &= \hat{P}_{k|k-1} - \bar{K} P_{\hat{y}_k} \bar{K}^{T}.
\end{aligned} \tag{15}$$

Acknowledgment

This work is supported by a University of Technology Sydney and the Office of Environment and Heritage of NSW PhD post thesis industry scholarship 2016, for which the 1st author would like to acknowledge. Support from the Office of Environment and Heritage of NSW is gratefully acknowledged for their laboratory facilities at Lidcombe and the data used in this study.

References

[1] P. Sellitto, F. D. Frate, D. Solimini, and S. Casadio, "Tropospheric Ozone Column Retrieval From ESA-Envisat SCIAMACHY Nadir UV/VIS Radiance Measurements by Means of a Neural Network Algorithm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 998–1011, March 2012.

[2] J. Xu, O. Schüssler, D. G. L. Rodriguez, F. Romahn, and A. Doicu, "A Novel Ozone Profile Shape Retrieval Using Full-Physics Inverse Learning Machine (FP-ILM)," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 12, pp. 5442–5457, Dec 2017.

[3] L. T. Silva and J. F. Mendes, "City noise-air: An environmental quality index for cities," *Sustainable Cities and Society*, vol. 4, pp. 1 – 11, 2012.

[4] J. Yuan, Z. Chen, L. Zhong, and B. Wang, "Indoor air quality management based on fuzzy risk assessment and its case study," *Sustainable Cities and Society*, vol. 50, p. 101654, 2019.

[5] H. Liu, H. Wu, X. Lv, Z. Ren, M. Liu, Y. Li, and H. Shi, "An intelligent hybrid model for air pollutant concentrations forecasting: Case of Beijing in China," *Sustainable Cities and Society*, vol. 47, p. 101471, 2019.

[6] Q. Wu and H. Lin, "Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network," *Sustainable Cities and Society*, vol. 50, p. 101657, 2019.

[7] W. Sun and J. Sun, "Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm," *Journal of Environmental Management*, vol. 188, pp. 144 – 152, 2017.

[8] C. Ortolani and M. Vitale, "The importance of local scale for assessing, monitoring and predicting of air quality in urban areas," *Sustainable Cities and Society*, vol. 26, pp. 150 – 160, 2016.

[9] M. Yu, X. Cai, Y. Song, and X. Wang, "A fast forecasting method for PM2.5 concentrations based on footprint modeling and emission optimization," *Atmospheric Environment*, vol. 219, p. 117013, 2019.

[10] X. Tang, J. Zhu, Z. Wang, M. Wang, A. Gbaguidi, J. Li, M. Shao, G. Tang, and D. Ji, "Inversion of CO emissions over Beijing and its surrounding areas with ensemble Kalman filter," *Atmospheric Environment*, vol. 81, no. 0, pp. 676 – 686, 2013.

[11] R. Banks and J. Baldasano, "Impact of WRF model PBL schemes on air quality simulations over Catalonia, Spain," *Science of The Total Environment*, vol. 572, pp. 98 – 113, 2016.

[12] M. Li, D. Chen, S. Cheng, F. Wang, Y. Li, Y. Zhou, and J. Lang, "Optimizing emission inventory for chemical transport models by using genetic algorithm," *Atmospheric Environment*, vol. 44, no. 32, pp. 3926 – 3934, 2010.

[13] S. Metia, S. D. Oduro, H. N. Duc, and Q. Ha, "Inverse air-pollutant emission and prediction using extended fractional kalman filtering," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 5, pp. 2051–2063, May 2016.

[14] S. Metia, Q. P. Ha, H. N. Duc, and M. Azzi, "Estimation of power plant emissions with unscented kalman filter," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2763–2772, Aug 2018.

[15] C.-H. Hsu, F.-Y. Cheng, H.-Y. Chang, and N.-H. Lin, “Implementation of a dynamical NH3 emissions parameterization in CMAQ for improving PM2.5 simulation in Taiwan,” *Atmospheric Environment*, vol. 218, p. 116923, 2019.

[16] C. D. Bray, W. Battye, V. P. Aneja, D. Tong, P. Lee, Y. Tang, and J. B. Nowak, “Evaluating ammonia (NH_3) predictions in the NOAA National Air Quality Forecast Capability (NAQFC) using in-situ aircraft and satellite measurements from the CalNex2010 campaign,” *Atmospheric Environment*, vol. 163, pp. 65 – 76, 2017.

[17] D. V. Vladutescu, Y. Wu, B. M. Gross, F. Moshary, S. A. Ahmed, R. A. Blake, and M. Razani, “Remote sensing instruments used for measurement and model validation of optical parameters of atmospheric aerosols,” *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1733–1746, June 2012.

[18] S. Madala, K. H. Prasad, C. Srinivas, and A. Satyanarayana, “Air quality simulation of NOX over the tropical coastal city Chennai in southern India with FLEXPART-WRF,” *Atmospheric Environment*, vol. 128, pp. 65 – 81, 2016.

[19] K. B. Shaban, A. Kadri, and E. Rezk, “Urban air pollution monitoring system with forecasting models,” *IEEE Sensors Journal*, vol. 16, no. 8, pp. 2598–2606, April 2016.

[20] E. Nanaki, C. Koroneos, J. Roset, T. Susca, T. Christensen, S. D. G. Hurtado, A. Rybka, J. Kopitovic, O. Heidrich, and P. A. Lpez-Jimnez, “Environmental assessment of 9 European public bus transportation systems,” *Sustainable Cities and Society*, vol. 28, pp. 42 – 52, 2017.

[21] E. M. Constantinescu, A. Sandu, T. Chai, and G. R. Carmichael, “Assessment of ensemble-based chemical data assimilation in an idealized setting,” *Atmospheric Environment*, vol. 41, no. 1, pp. 18 – 36, 2007.

[22] E. Pisoni, D. Albrecht, T. Mara, R. Rosati, S. Tarantola, and P. Thunis, “Application of uncertainty and sensitivity analysis to the air quality

SHERPA modelling tool," *Atmospheric Environment*, vol. 183, pp. 84 – 93, 2018.

[23] S. Cheng, Y. Zhou, J. Li, J. Lang, and H. Wang, "A new statistical modeling and optimization framework for establishing high-resolution PM10 emission inventory - I. Stepwise regression model development and application," *Atmospheric Environment*, vol. 60, pp. 613 – 622, 2012.

[24] D. Romano, A. Bernetti, and R. D. Lauretis, "Different methodologies to quantify uncertainties of air emissions," *Environment International*, vol. 30, no. 8, pp. 1099 – 1107, 2004.

[25] S. K. Biswas, L. Qiao, and A. G. Dempster, "A novel a priori state computation strategy for the unscented kalman filter to improve computational efficiency," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1852–1864, April 2017.

[26] K. Xiong, H. Zhang, and C. Chan, "Performance evaluation of UKF-based nonlinear filtering," *Automatica*, vol. 42, no. 2, pp. 261 – 270, 2006.

[27] H. V. Poor and S. Verdu, "Probability of error in mmse multiuser detection," *IEEE Transactions on Information Theory*, vol. 43, no. 3, pp. 858–871, May 1997.

[28] W. Bich, "From errors to probability density functions. evolution of the concept of measurement uncertainty," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 8, pp. 2153–2159, Aug 2012.

[29] EPA, *The NSW Environment Protection Authority*, 2008 (accessed 2019.11.14), http://www.epa.nsw.gov.au/your-environment/air/air-emissions-inventory/air-emissions-inventory-2008.

[30] J. Hartikainen and S.Särkkä, "Kalman filtering and smoothing solutions to temporal gaussian process regression models," in *2010 IEEE International Workshop on Machine Learning for Signal Processing*, Aug 2010, pp. 379–384.

[31] S. K. Sahu and K. V. Mardia, "A bayesian kriged kalman model for short-term forecasting of air pollution levels," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 54, no. 1, pp. 223–244, 2005.

[32] S. Särkkä, "Unscented Rauch-Tung-Striebel Smoother," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, pp. 845–849, April 2008.

[33] Q. Ha, H. Wahid, H. Duc, and M. Azzi, "Enhanced radial basis function neural networks for ozone level estimation," *Neurocomputing*, vol. 155, pp. 62 – 70, 2015.