

# Source attribution of air pollution by spatial scale separation using high spatial density networks of low cost air quality sensors



I. Heimann, V.B. Bright, M.W. McLeod, M.I. Mead, O.A.M. Popoola, G.B. Stewart, R.L. Jones\*

University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

## HIGHLIGHTS

- Fast response, high spatial density measurements from a low cost sensor network.
- Technique developed to extract underlying pollution levels from high resolution data.
- Regional and local contributions to total pollution levels are quantified.
- Generally applicable technique for sensor networks (gases and particulates).

## ARTICLE INFO

### Article history:

Received 12 August 2014

Received in revised form

21 April 2015

Accepted 24 April 2015

Available online 25 April 2015

### Keywords:

Air quality

Baseline extraction

Emission scales

Source attribution

Electrochemical sensors

Sensor networks

## ABSTRACT

To carry out detailed source attribution for air quality assessment it is necessary to distinguish pollutant contributions that arise from local emissions from those attributable to non-local or regional emission sources. Frequently this requires the use of complex models and inversion methods, prior knowledge or assumptions regarding the pollution environment. In this paper we demonstrate how high spatial density and fast response measurements from low-cost sensor networks may facilitate this separation. A purely measurement-based approach to extract underlying pollution levels (baselines) from the measurements is presented exploiting the different relative frequencies of local and background pollution variations. This paper shows that if high spatial and temporal coverage of air quality measurements are available, the different contributions to the total pollution levels, namely the regional signal as well as near and far field local sources, can be quantified. The advantage of using high spatial resolution observations, as can be provided by low-cost sensor networks, lies in the fact that no prior assumptions about pollution levels at individual deployment sites are required. The methodology we present here, utilising measurements of carbon monoxide (CO), has wide applicability, including additional gas phase species and measurements obtained using reference networks. While similar studies have been performed, this is the first study using networks at this density, or using low cost sensor networks.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Numerous studies have demonstrated that certain gas-phase pollutants, such as nitrogen oxides (NO<sub>x</sub>), ozone (O<sub>3</sub>) and carbon monoxide (CO) can be physiologically toxic and may have adverse health effects even at low-level concentrations (e.g. Morris, 2000; Vreman et al., 2000; Wayne, 2000). As such, mitigating urban air pollution has gained considerable importance. Monitoring air quality within urban areas is vital in providing the necessary information to carry out detailed source attribution, and to inform

policy that allows the effective reduction of pollution levels as well as providing more detailed information for epidemiology.

A number of methods exist to carry out source apportionment to investigate the influence of emissions from varying sources that include methods based on emission inventories and dispersion models, and receptor models based on the statistical evaluation of chemical data acquired at measurement sites (Viana et al., 2008). Such methods however may be restricted by the accuracy, availability of emission inventories and pollution source information (Hopke et al., 2006).

The evaluation of monitoring data is a suitable alternative to those methods outlined above with the main advantage being the simplicity of the mathematical methods applied and the reduced effect of mathematical artefacts due to data processing or prior

\* Corresponding author.

E-mail address: [rj1001@cam.ac.uk](mailto:rj1001@cam.ac.uk) (R.L. Jones).

assumptions (e.g. Viana et al., 2008).

Of themselves, atmospheric measurements of pollutants provide only the total pollution levels and thus combine contributions from the underlying regional background sources as well as those from local emissions (e.g. Lenschow et al., 2001; Ketzel et al., 2003). In order to undertake more effective source attribution, measurements must be separated into several components, distinguishing between local plume events on short temporal and small spatial scales and underlying trends including the long range transport and variations in natural background pollution sources (Lenschow et al., 2001).

Within urban areas, a number of additional emission sources (such as vehicle exhaust and background heating) exist that contribute to total pollution levels when compared to the wider sub-urban and rural environments. These local source emissions can accumulate and thus increase pollution over longer time scales. Quantifying these contributions to the total pollution levels is useful for pollution mitigation.

A number of air quality monitoring networks exist that provide atmospheric measurements of key pollutants such as  $\text{NO}_x$ ,  $\text{O}_3$ , CO and particulate matter (PM). The Automatic Urban and Rural Network (AURN) is one such network and is deployed in the UK in order to comply with national and European air quality regulations (Defra, 2014). However, its temporal (1 h) and spatial (~100 sites in the entire UK) resolution is far too sparse to allow the contributions of different sources to total pollution levels across the UK to be quantified without significant additional constraints and assumptions such as the use of physical or statistical models. As outlined in Ketzel et al. (2003) measurements at both polluted and unpolluted sites in close proximity are necessary to effectively estimate the amounts that local plume events, emission build-up and background levels contribute to the overall pollution levels. This is particularly important in terms of advanced source attribution for relatively short-lived species that are chemically converted within close proximity (several hundred metres) of their sources (e.g.  $\text{NO}_x$  with a lifetime in the order of one day (Seinfeld and Pandis, 1998)). Low-cost electrochemical air quality sensors can be deployed in denser networks, potentially alleviating this problem (see for example Mead et al., 2013).

In this paper, we combine high temporal and spatial resolution data generated by a low-cost high density air quality network with a novel approach to data analysis to illustrate how, combined, source attribution can be achieved. The focus of this paper is CO, a moderately long lived chemical species with a lifetime in the order of weeks to months (Holloway et al., 2000; Zellweger et al., 2009). CO is thus subject to longer range transport and may be used as tracer molecule to investigate the influence of larger-scale meteorological events on tropospheric pollution. Because of its long lifetime, CO generally is a useful indicator of local pollutant emissions. CO is the main (70%) loss mechanism for the hydroxyl radical, OH, (Novelli et al., 1998) with increased CO levels enhancing the rate of OH removal, subsequently reducing the scavenging mechanism for other pollutants as well as augmenting tropospheric ozone production. Knowledge of CO emission sources may therefore be indirectly important in terms of reducing adverse health effects related to atmospheric pollutants. It also contributes to carbon emissions making it important in terms of climate change mitigation.

In the present work we will show that we are able to define a regional CO signal through a purely data-based approach (section 4.1), subsequently allowing detailed source attribution to be carried out based on the measurements alone. This new technique is used to separate the different contributing scales of air pollution namely regional, far field and near field (section 4.2). The methods presented may be applied to other pollutant species when differences

in abundance due to their chemistry and lifetimes are considered.

## 2. The Cambridge air quality monitoring network

In spring 2010, a network of 45 low-cost electrochemical sensor nodes was deployed in and around the city of Cambridge, UK during a period covering 2.5 months (11 March 2010 to 30 May 2010). The network provided measurements of CO, nitric oxide (NO) and nitrogen dioxide ( $\text{NO}_2$ ) as well as temperature and relative humidity at a high temporal (10 s) resolution. To account for stabilisation of the electrochemical cells within the surrounding ambient conditions, the initial 14 days of the full measurement campaign were excluded from further analysis. Thus, the study covers an eight week period from 28 March 2010 to 23 May 2010. Of the 45 sensor nodes deployed, 9 were discarded as a result of reduced data coverage (<1 month) due to battery issues or physical damage. An additional 4 nodes had technical problems or were clearly biased by external factors (Alphasense, 2005) thus 32 (71%) of the sensor nodes deployed were included in this study.

### 2.1. Accuracy of the electrochemical sensors to ambient concentrations

Mead et al. (2013) have reported on characterisation of electrochemical sensors, determining an instrumental detection limit (IDL) of 4 ppb (parts-per-billion) for CO and their sensitivity to ambient pollution levels. The electrochemical sensors' long-term stability allowed observed differences in the measured absolute mixing ratios for individual sensors to be corrected for during operation and data processing.

We reference the data of the sensor network to gas-chromatographic measurements, averaged to a 30 min resolution and calibrated daily against several NOAA standards. These data were obtained from the Greenhouse Gas Laboratory at Royal Holloway, University of London (RHUL), 75 miles south-west of Cambridge in Egham, Surrey. Every sensor node is referenced to this one station. To remove local influences on the measurements, a meteorological filter is applied. The individual sensor offset to ambient pollution levels is therefore defined as the difference between the node's minimum CO concentration during those nights (01:00–04:00, all times in BST) with a wind speed ( $U$ ) greater than  $2 \text{ ms}^{-1}$  and the corresponding value of the RHUL data set.

### 2.2. Deployment details of the electrochemical sensor network

The sensor nodes were mounted on lamp posts, 3 m above street level. A higher density of sensor nodes were deployed within the urban environment (Fig. 1) where higher variability of pollution levels were expected compared to rural areas. Sensor nodes were approximately evenly distributed in the sub-urban area of Cambridge in order to investigate the influence of varying wind direction on atmospheric pollution and to inter-compare rural environments.

The current of each electrochemical sensor was measured every 10 s, and then converted to counts (via a resistor to generate a voltage and an Analogue-To-Digital-Converter) and stored on-board the sensor. The collected (10 s time resolution) raw data were then transmitted in packets to a central computer server at two hour intervals, in order to reduce power consumption by the GPRS in each node.

The transmission process induces electrical interference on the sensor signals; thus, the initial 65 recordings, that is, ~11 min of data after each transmission were filtered out prior to further analysis. The data were then converted into mixing ratios using pre-defined, sensor-specific sensitivity factors.



**Fig. 1.** Spatial distribution of the Cambridge sensor network deployed from 14 March 2010 to 30 May 2010 (bottom panel), with a detailed view of the city centre area marked shown above.

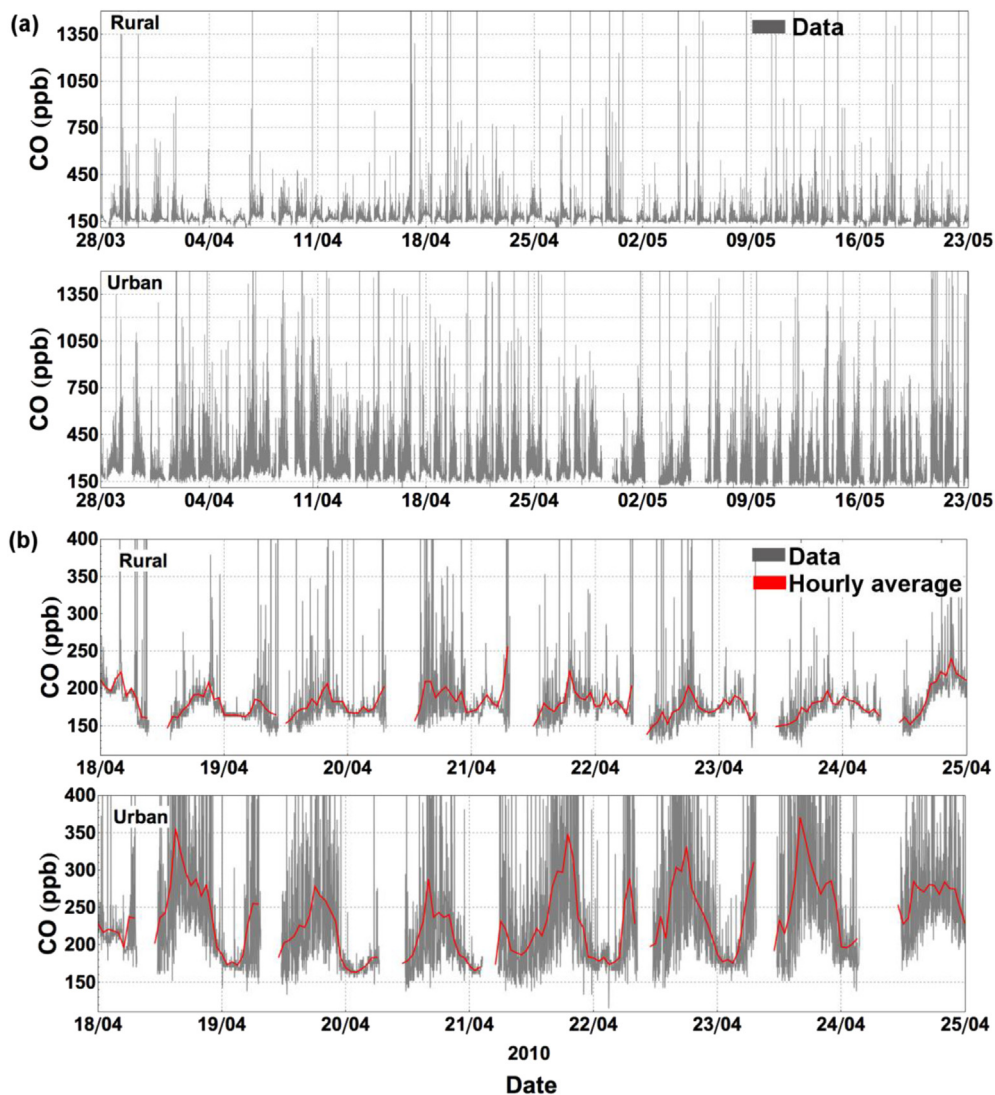
Fig. 2 shows a typical example of CO measurements during the analysis period, in this case for two sensor nodes deployed in contrasting rural and urban environments, illustrating the differences in the frequency and absolute levels of high pollution events (a). We use these two example sensor nodes throughout the paper to highlight the analysis methodology which we apply for all nodes. Note that the periods of missing data result from an imposed quality control criterion before the data analysis (details in section 2.3).

The measurements show CO pollution levels frequently above 1 ppm for point sources. This implies that hourly averages (as

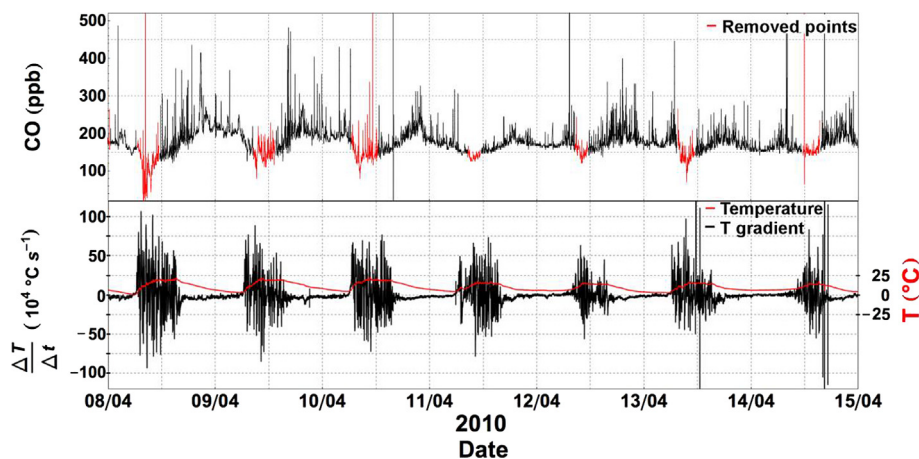
shown in b, red), with pollution levels below 500 ppb, are not likely to be sufficient to assess acute exposure of individuals to CO and potentially other pollutants.

### 2.3. Quality control of the measured data

Analysis of the CO data set reveals quasi-regular periods in the late morning and early afternoon (predominantly between 09:00 and 13:00) across the network where mixing ratios dropped suddenly (Fig. 3, top panel). These drops may be attributed to rapid changes in sensor temperature usually associated with solar



**Fig. 2.** Time series of CO (ppb) for one rural and one urban sensor node; covering the whole analysis period (a) and for one week (b). Shown in grey are data recorded at 10 s resolution (capped at 1500 ppb as less than 0.1% of the recorded data exceed this level), in red the hourly average of the measured data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



**Fig. 3.** Time series from a rural sensor node of (top panel) CO mixing ratios illustrating the data points that are removed due to the quality control criterion (red) and of (bottom panel) temperature  $T$  (red) and its temporal variation  $\Delta T/\Delta t$  (black). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

heating. However, no reliable quantitative correlation can be established (Fig. 3, bottom panel) as, while each node temperature was recorded, individual sensor temperatures were not. In addition, as a result of the rapid drops being observed at variable times on each day, the mechanism of removing these data points from further analysis was limited to the application of a statistical filter. On average  $6 \pm 4\%$  of the recorded data are removed ranging from 1% to 24% for the individual sensor nodes.

We developed an objective method to remove data in segments. For this we calculated the daily (midnight to midnight) 10th percentile of CO measurements and flagged those hours as biased where more than 20% of the 10 s data was below the daily 10th percentile. We proceeded in a cautious manner and removed all data between the first and last hour deemed biased plus a 30 min window on each side to account for the initial gradual changes in CO observed (see Fig. 3).

We are aware that the statistical filter applied is somewhat coarse and may result in the removal of reliable data as well as those affected by temperature changes. However, the baseline extraction method derived in this study is not affected by this removal process. Furthermore, the impact of the developed methodology outweighs the fact that some sensors of the prototype network are below the standard reference method data capture levels. We do not expect future generations of the sensor nodes to be affected by external interferences as changes have been made to their deployment boxes.

Fig. 4a illustrates that the hours between 09:00 and 13:00 are

removed most frequently. This emphasises the regularity in the low mixing ratio occurrence. The filtered data periods are mostly (~30%) four hours long and 70% are shorter than 7 h (Fig. 4b).

### 3. Estimation of underlying long time-scale contributions to a measured signal: baseline extraction methodology

In this section we outline a method for the separating the underlying large scale variations in pollutant concentrations associated, for example, with long range transport, from shorter time scale and often more pronounced events associated with local emission events. This method is similar to the removal of polluted signals in anthropogenic trend determination (e.g. Flandrin and Goncalves, 2004). We refer to the extracted, underlying signal as the ‘baseline’ concentration for each sensor node hereafter. By exploiting the fact that we were using a network of sensor nodes, rather than a single measurement site, we find that we could also distinguish temporally unresolved local emissions from both nearby individual local emission events and the regional pollution signal.

As stated in Ruckstuhl et al. (2012), the measured signal for an atmospheric pollutant,  $S(t)$ , at time  $t$  can be represented as the sum of a background concentration signal,  $B(t)$ , which is referred to as the baseline henceforth, and a local emission signal  $L(t)$  i.e.:

$$S(t) = B(t) + L(t). \quad (1)$$

With a high temporal resolution (i.e. 10 s) these baseline signals  $B(t)$ , i.e. events over longer time scales, can be separated from events with a significant influence of local sources  $L(t)$ , due to the fact that localised events occur over a short finite time even in an urban roadside environment. The difference in frequencies of local high pollution levels and lower background levels can therefore be observed.

To extract these various scales, the CO data over the full analysis period  $t_d$  were divided into a number ( $I$ ) of subsequent smaller data sets of equal time lengths  $\varphi$ . The signal  $S(t)$  over  $t_d$  was then defined as the sum of  $s_i$ , called fragments hereafter, where  $s_i$  represents the signal  $S(t)$  between two time points  $t$  and  $t + \varphi$ :

$$S(t) = \sum_i^I s_i. \quad (2)$$

It is important to choose a suitable value for  $\varphi$  carefully when considering the purpose of the analysis, as it defines the detail in which the baseline follows the fast response measurements. A shorter  $\varphi$  results in a baseline that is influenced by local emissions. In this study an optimal length,  $\varphi = 3$  h, was chosen to account for diurnal variability in the pollution levels.

For each fragment  $s_i$ , the distribution of the measured mixing ratios was calculated using discrete concentration intervals (bins). Adjusting the bin width  $\gamma$  to the range of the measured mixing ratios ensures that the distribution of the measurements is adequately represented. For the extraction of the CO baselines in this study,  $\gamma$  was set to 10 ppb i.e. greater than twice the IDL. The most probable mixing ratio was determined for each  $s_i$  representing the baseline signal  $b_i = B(t_i)$  of  $s_i$  where  $t_i$  is the mean fragment time (Fig. 5). Data points below the mode of each fragment  $s_i$  were attributed to noise and represent the associated baseline error. We attributed a minimum error associated to the baseline extraction method of 4 ppb, corresponding to the IDL (Fig. 6).

Taking advantage of the, by definition, smooth behaviour of a baseline (e.g. Ruckstuhl et al., 2012), the function  $B(t)$  over the analysis period  $t_d$  can be obtained through interpolation (here applying a third order polynomial function) between the individual

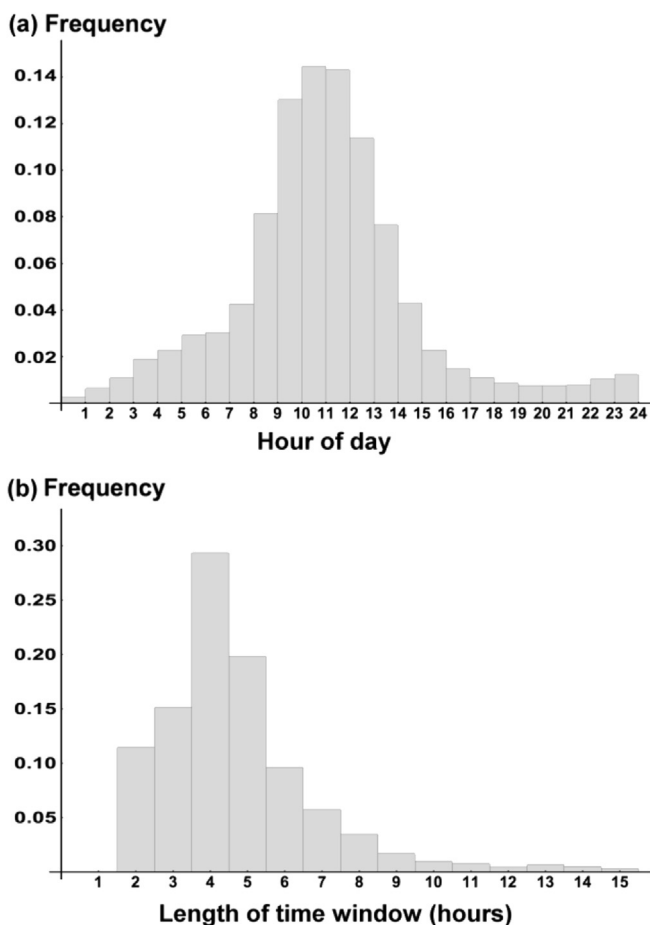
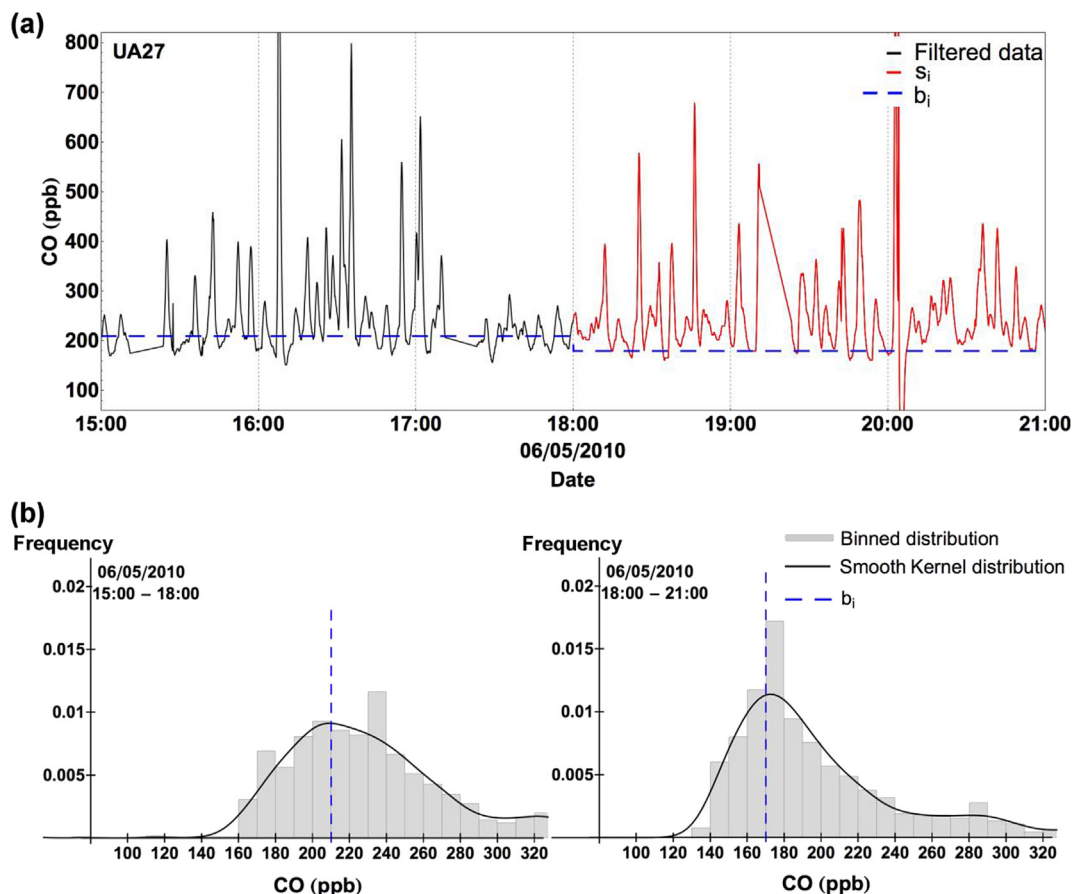


Fig. 4. Probability density functions (PDFs) of (a) the hours of the day that data are removed for all sensor nodes and (b) the length of filtered data windows (hours).



**Fig. 5.** (a) Time series of CO (ppb) measured by an example sensor node. The data highlighted in black and red are of two different fragments  $s_i$  ( $\varphi = 3$  h) for which the binned distribution ( $\gamma = 10$  ppb) of the mixing ratios is shown in (b). The derived baseline concentration  $b_i$  for the two  $s_i$  is shown by the dashed blue lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

baseline concentrations  $b_i$ . To account for missing data, baseline values were only deemed representative if more than 500 measurements (i.e. more half of the expected 1000 measurements at 10 s resolution within a three hour window  $\varphi$ ) were used to estimate  $b_i$ . No interpolation was applied between missing data points.

The extraction method devised allows fast computation on a desktop computer through its relatively straightforward mathematical approach. It is therefore suitable for application to large data sets generated by spatially dense high temporal resolution sensor networks as used here. It presents a flexible data analysis tool and can be used, as in this study, to analyse diurnal variation but can also be applied to investigate seasonal variations (increasing  $\varphi$ ) or to analyse shorter time scale temporal variability in atmospheric pollution levels (decreasing  $\varphi$ ).

High temporal resolution measurements have allowed the inference of a temporally varying local background that is impossible to obtain from hourly averages. The additional separation of scales that may be achieved using this method shows significant near field signals on top of the baselines, as is evident from the differences between the baseline and the hourly average that represents both the local and background components (Fig. 6).

#### 4. Results and discussion

We now show how the variability in baseline signals for sensor nodes deployed in different environments can be used to determine the regional pollution signal (in this case, obviously, for Cambridge) and highlight the differences in long term pollution levels present in urban and rural environments (4.1). Taking

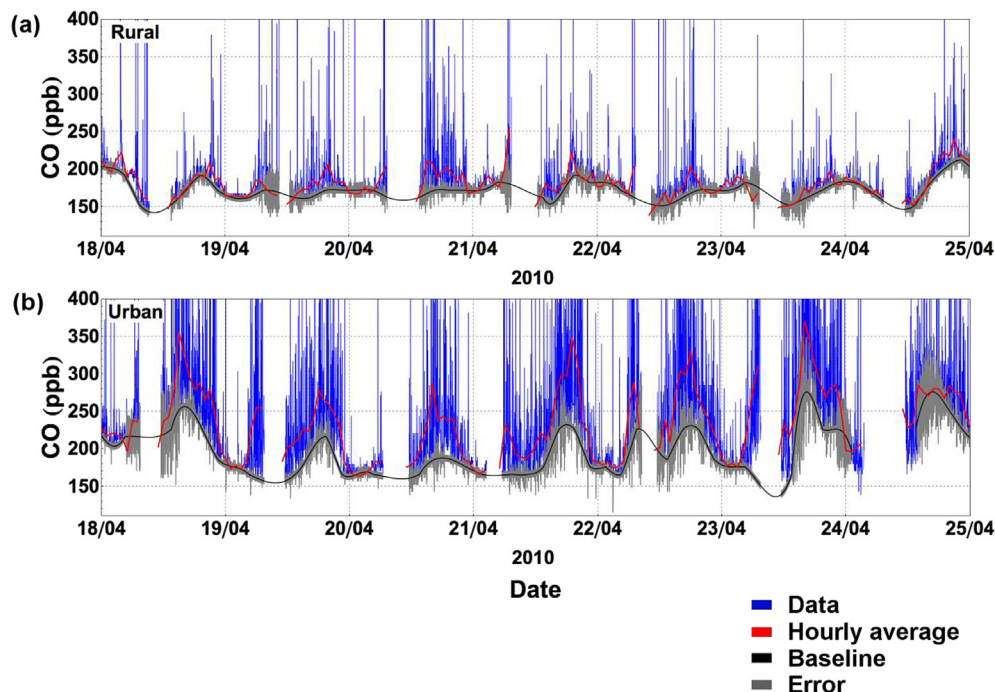
advantage of the high spatial density of measurements within the network we show how the regional pollution signal and baselines inferred from high temporal resolution measurements may be used to separate far field, i.e. suburban, near field, i.e. local, and regional contributions to pollution levels (4.2).

Lenschow et al., 2001 have previously investigated the different contributions to total PM levels. This study used a fixed, calibrated ensemble of 18 sites with prior knowledge/assumptions about the nature of local PM sources applied in order to estimate regional background levels, local background contributions (far field) and highly site dependent local sources. While we divide the pollution levels into three similar contributions, we demonstrate a different approach as to how the contributions are defined and apply the Lenschow technique to gas phase pollutants.

We show here how high-frequency measurements inherently carry information about the different time scales of pollution levels and show that low cost sensors provide the required temporal resolution. The collected data from our sensor network therefore allow the determination of a varying regional influence and local contributions without prior assumptions concerning the nature of the deployment site. As such, our approach to separate the different scales contributing to total pollution levels is fundamentally different to that presented in Lenschow et al. (2001).

##### 4.1. Extracting a regional pollution signal using a high spatial density sensor network

Fig. 7 illustrates the urban and rural average baselines calculated over the measurement period. A smaller range of CO baseline



**Fig. 6.** Time series of CO (ppb) for one rural (a) and one urban (b) sensor node for the period of a week. The data shown in blue are recorded at 10 s time resolution, the red curves are the hourly average and the black the extracted baselines with the associated error in grey (method described in section 3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pollution levels is observed for sensor nodes deployed in the outskirts of Cambridge when compared to those within the city centre. This may be explained by the additional pollution sources in urban environments which build up near their source, add to the regional background levels and thus increase the baseline pollution levels at such locations.

In order to estimate this contribution of local emission plumes and their build-up to the observed CO levels in urban environments, it is necessary to determine the regional pollution signal and its temporal variation. We define this regional pollution signal of Cambridge as the average of baselines attributed to rural environments which are, by definition, not influenced significantly by significant local emission sources (Fig. 7a, red).

To classify an environment as rural we apply two criteria, fulfilled by eight (i.e. 25%) of the sensor nodes included in this study. Firstly, the time series of high resolution data is required to show a standard deviation,  $\sigma$ , less than 40 ppb, i.e. the first quartile of the  $\sigma$  of the high resolution data provided by the 32 sensor nodes. This follows the method of Ruckstuhl et al. (2012) who show that measurements of pollutants in rural environments away from direct emission sources normally have small standard deviations. Secondly, the mean of the high resolution data is required to be within  $1\sigma$  of the corresponding baseline mean. This second criterion depends on the assumption that, with minimal influence of short term plume events and their accumulation, the distribution of the high resolution data is expected to be very similar to that of its baseline.

The regional signal (Fig. 7a, red) varies by approximately 80 ppb over the analysis period, with a minimum of ~135 ppb (11:30 on 22 May 2010) and a maximum of ~215 ppb (21:50 on 08 April 2010). Observing lower CO levels during daytime hours may arise as cleaner air mixes into the boundary layer from the free troposphere where generally lower CO mixing ratios reside, while a stable nocturnal boundary layer traps pollutants near the surface and thus increases their night time levels. The regional signal shows a

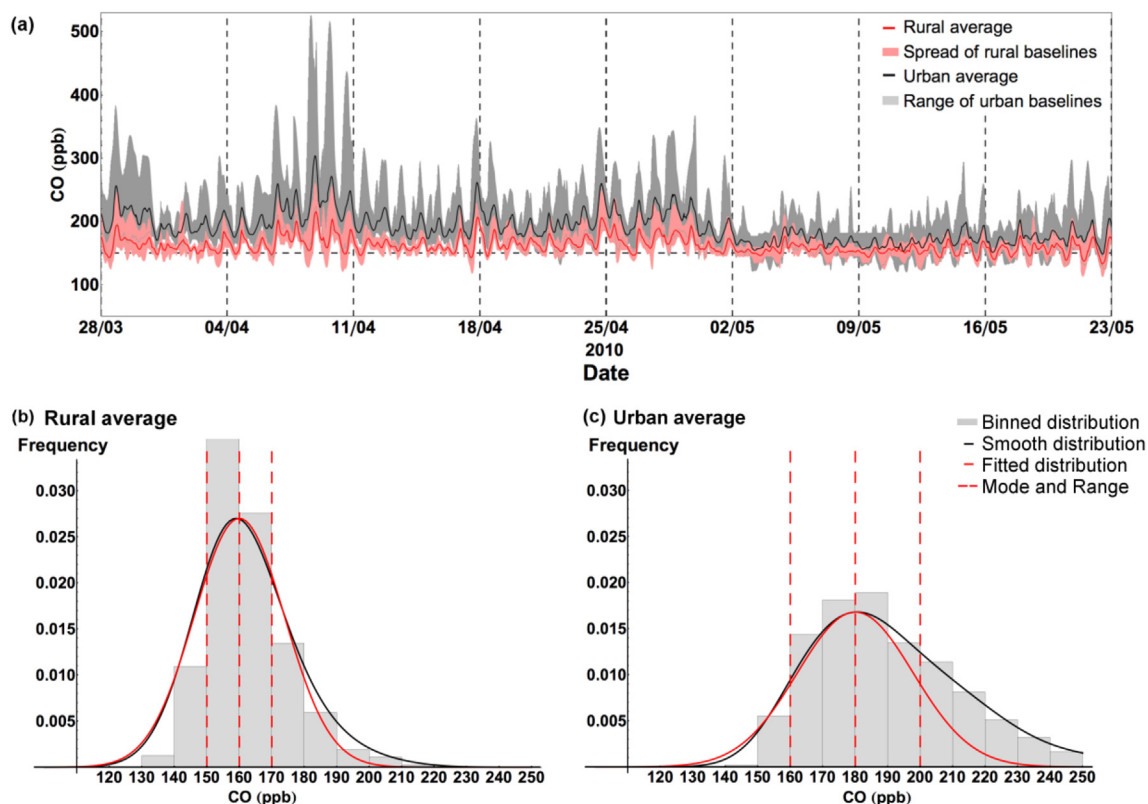
normal distribution with a mean CO level of  $160 \pm 10$  ppb over the analysis period (Fig. 7b). The small overall standard deviation ( $\sigma = 10$  ppb) demonstrates that the eight averaged baselines represent near identical environments with minimal local source contributions.

The diurnal variation changes over the course of the analysis period with, on average, smaller variations in May (40 ppb) compared to April (55 ppb). A 90% decrease can be observed between the maximum and minimum diurnal variation, 65 ppb (17 April 2010) and 5 ppb (8 May 2010) respectively. In addition to those factors discussed below, these differences are influenced by changes in meteorological conditions with a general shift in wind direction from late March and throughout April to May. While there is no predominant wind direction in April, a tendency towards northerly ( $337.5^\circ - 22.5^\circ$ ) winds is observed in May.

Fig. 7a illustrates the regional CO signal when compared to the average and range of the 24 sensor node baselines not deployed in rural environments, referred to as the urban average hereafter (black curve and grey filling). The urban average shows a range in CO baseline levels of 155 ppb, i.e. nearly double that of the regional signal. Both the minimum and maximum concentrations, 150 ppb and 305 ppb respectively, can be observed at the same times as the regional signal extremes.

Fig. 7c shows that the urban distribution, as might be expected, is skewed, more positively, with a larger mean CO level of 190 ppb over the analysis period. However, this increase is not significantly different from the regional signal when including the standard deviation of 25 ppb over the analysis period. Nonetheless, a significant decrease in mean CO from  $205 \pm 20$  ppb in April to  $170 \pm 10$  ppb in May can be observed for the urban average suggesting a 50% decrease in variability between urban locations in May compared to that of April. This decrease is less pronounced for the regional CO signal as these are less influenced by local emissions.

Although meteorological factors play a key role, pollutant build-



**Fig. 7.** (a) Time series of CO (ppb) of the rural average, i.e. regional signal, (red) and of the urban average (grey) with the respective ranges; PDFs of (b) the regional signal and (c) the urban average. In grey: the binned distribution of the baseline average ( $\gamma = 10$  ppb); in black: smooth distribution; in red: Gaussian function fitted to the maximum of the distribution. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

up in urban environments, as a result of increased vehicle emissions, heating and/or other combustion sources, may also decrease towards the summer months and contribute to this decrease.

Diurnal variations in the urban average are also less pronounced in May (average of 40 ppb) when compared to April (average of 90 ppb). This large reduction (>50%) may indicate the influence of changing synoptic conditions in May. A reduction of 85% can be observed between the maximum and minimum diurnal variation, 110 ppb (08 April 2010) and 15 ppb (03 May 2010) respectively. Even though these extremes do not coincide in time with those of the rural average, it can be concluded that the driver of the fluctuations in the diurnal variations may be attributed to factors such as a shift in wind direction and change in larger scale meteorological conditions.

For two distinct periods during the measurement campaign a notable difference in both the local and regional contributions is observed that may be attributed to the influence of synoptic conditions. Fig. 7a shows that over a three day period following 8 April 2010 there is an increase in both rural and urban baselines. At the same time a stable, high pressure system existed over the UK bringing stagnant air and low wind speeds ( $U \sim 1 \text{ ms}^{-1}$ ). Such conditions often lead to a reduction in vertical mixing after sunset and over-night following the formation of a stable nocturnal boundary that effectively traps pollutants near their sources. During the day however, with increasing solar insolation, high pressure conditions may lead to an increase in vertical mixing following the formation of a convective mixed layer, hence the large variation in CO levels observed in April. The accumulation of CO over this period is more pronounced in the urban sensor nodes, illustrating that despite pollution being transported into Cambridge, the local urban emissions have a higher influence on urban air quality.

In contrast, a period of relatively low baselines is observed for 3 days from 7 May 2010. During this time, low pressure conditions prevailed bringing higher wind speeds ( $U \sim 3 \text{ ms}^{-1}$ ) and thus increased horizontal mixing, thus reducing pollutant levels and their accumulation near source. Over this period conditions are less stable at night hence a reduction in the build of pollutants under the nocturnal boundary layer. During the day, there is likely to be less solar radiation and therefore a weaker convective boundary layer will reduce the extent of vertical mixing. In contrast to the period over which atmospheric conditions are relatively stagnant, where low pressure conditions prevail, there is no notable difference in pollution levels that can be observed between the regional CO signal and the urban baseline average (Fig. 7a). This arises due to less local emission build-up during the low pressure system, baselines in urban environments are less enhanced and experience similar pollution levels as those in rural environments.

#### 4.2. Scale separation: estimating contributions to the overall pollution levels

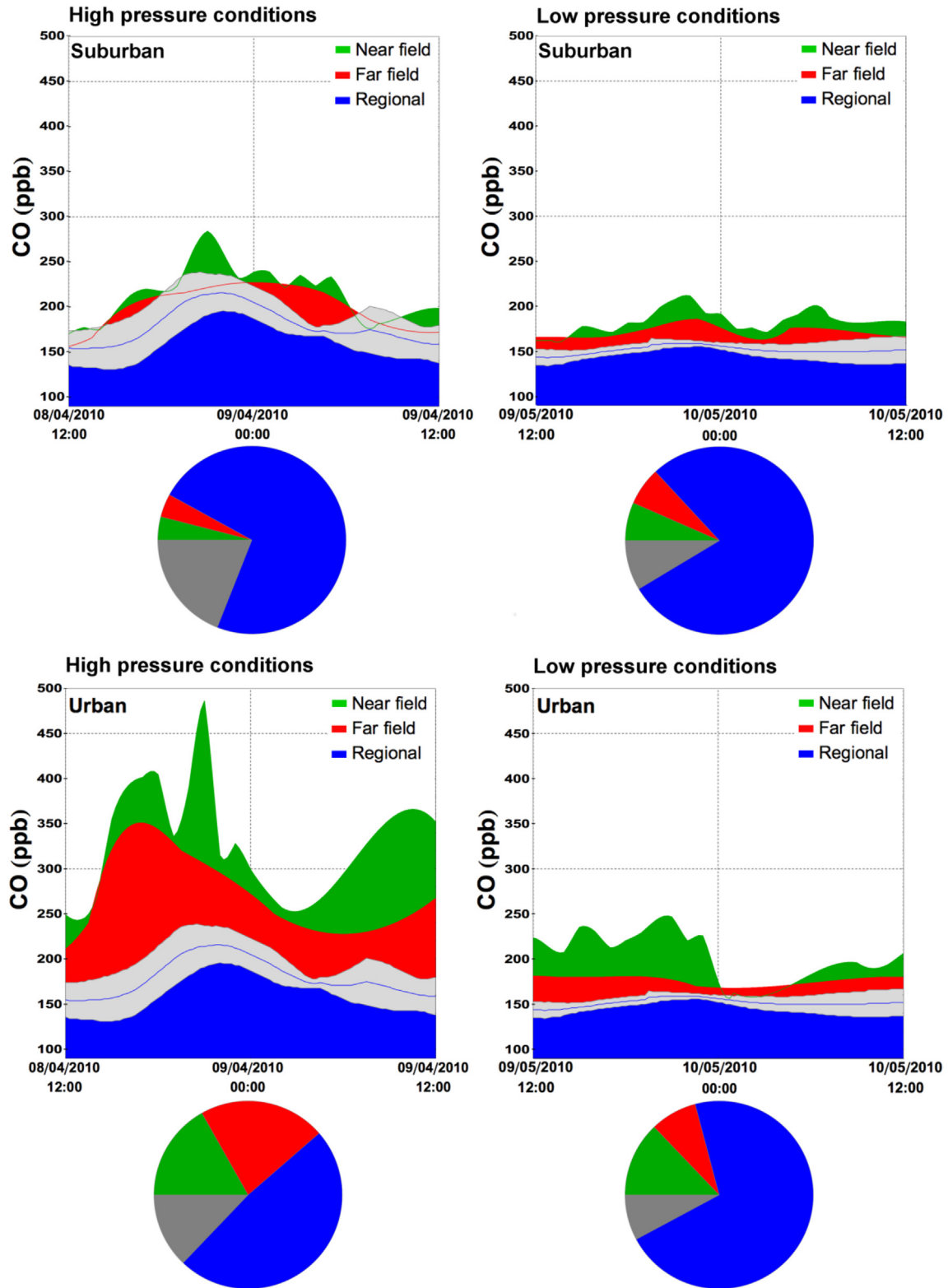
As discussed earlier in this paper, overall pollution levels are taken as the sum of a regional contribution and the accumulation of far field, e.g. build-up, and near field emissions. We demonstrate here the separation of these three components for the two periods of different meteorological conditions described in 4.1 and quantify their relative importance. The results are presented for two example sensor nodes, deployed in an urban and a suburban environment (Fig. 8).

While 10 s resolution data are used to derive the baselines, we generate hourly averages in order to estimate the contribution of short time scale events to the total pollution levels. We define the



regional influence for Cambridge as the area under the regional CO signal (blue fill). The contribution of far field emissions, which represent accumulated pollutants captured in the baselines, is defined as the area between the sensor node baseline and the CO

signal (red fill). The area between the baseline and the hourly averaged measurements represents the near field contribution of local pollution sources (green fill). We also estimate the influence of the uncertainty in the regional CO signal to the overall pollution



**Fig. 8.** Time series for CO (ppb) and pie-charts to illustrate the regional (blue fill), far field (red fill) and near field (green fill) contributions as well as the uncertainty in the regional CO signal (grey fill) for a suburban (top) and an urban (bottom) sensor node; (left) between 12:00 on 08 April 2010 to 12:00 on 09 April 2010 to highlight the effect of a high pressure system over the UK during this time and (right) between 12:00 on 09 May 2010 and 12:00 on 10 May 2010 illustrating the influence of a low pressure system over the UK. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

levels (grey fill). Both the regional and far field contributions are limited by the lower and upper range of the regional CO signal, respectively, and may therefore be underestimated.

Overall, the regional signal is the largest contributor to air pollution with 73% (78%) for the suburban and 49% (71%) for the urban sensor node under the high (low) pressure conditions studied here. However, the error associated with the estimation of the regional signal contributes as much as 19% to the total pollution levels for the suburban sensor node under high pressure conditions in April. Due to higher total pollution levels, this contribution represents only 13% for the urban node. During low pressure conditions in May this error only contributes to 9% and 8% to the levels for the suburban and urban sensor nodes respectively, again an indicator for the increased similarity in pollutant levels observed between the different measurement locations.

It is unlikely for local emission sources to change significantly between April and May for rural and suburban environments. Thus, the observed doubling of both near and far field contributions under low pressure conditions when compared to those under high pressure was used to estimate a 4% accuracy of the assessed individual contributions to total pollution levels.

The near field contribution for the urban sensor node decreases from 17% to 13% when changing from the high pressure conditions in April to low pressure in May, which generally lies within the associated error bars. However, the far field contribution decreases from 22% in April to 8% in May (i.e. 65% decrease). This highlights the effect of local pollutant sources on pollution levels over a longer time period when confined to urban environments and how different meteorological conditions may affect pollution levels.

## 5. Conclusions and future work

In this paper we have shown that low-cost sensors deployed in a dense network can provide the information required to carry out detailed source attribution. A high spatial and temporal resolution data set of CO concentration measurements, collected over a two month period (spring 2010) using 32 electrochemical sensor nodes included in a dense network deployed in Cambridge, UK, has been analysed.

A novel and flexible method has been developed to determine sensor baselines, i.e. underlying variation, of measurements (that represent non-local emissions) which is suitable for application to large data sets. Combining these baselines with high spatial resolution measurements made across the network we have demonstrated how to use these to separate and quantify levels of those pollutants that accumulate in urban environments and increase the long term pollution levels in these areas. The measured signals can thus be distinguished into three components: (1) local plume events, influenced mainly by traffic; (2) build-up of local events, influenced mainly by traffic queuing and general energy usage; (3) regional background, showing the large-scale effect of accumulations of the surrounding areas.

One limitation of the technique developed here is that the methodology outlined is based on the assumption that no chemical processing of pollutants occurs between each of the measurement sites and that pollutant levels are determined by emissions alone. Another source of uncertainty lies in the representativeness of each site as a degree of subjectivity may be applied when classifying these.

In addition, as a result of the filtering process applied to data for a proportion of the day for each sensor it is probable that some uncertainty surrounding source attribution for certain sites, i.e. those in particular affected by diurnal changes in traffic, will be introduced and thus treated with caution however in subsequent

generations of electrochemical sensors such drops in CO data are not observed.

Through this analysis, the variation in CO baseline pollution levels within an urban environment can be studied in greater detail, which may, for example, provide valuable information for both informing pollution mitigation measures and evaluating urban dispersion models in the future.

We also note that the technique has wider applicability, for example to low cost sensor networks of other species and particulates, and indeed could be applied effectively to reference instrument data (e.g. AURN) were it available at high time resolution rather than the hourly averages which are publicly disseminated.

## Acknowledgements

The authors would like to thank the following: Brian Jones of the Digital Technology Group, University of Cambridge Computer Laboratory for meteorological data; Christoph Hügliin of the Air Pollution and Environmental Technology Unit from EMPA for scientific advice; Euan Nisbet of the Department of Earth Science at the Royal Holloway University of London for calibrated CO reference data. The authors thank EPSRC (EP/E001912/1) for funding for the Message project. IH thanks the German National Academic Foundation for funding of MPhil degree.

## References

- Alphasense, 2005. Alphasense Application Note AAN 103: Shielding Toxic Sensors from Electromagnetic Interference.
- Defra, 2014. <http://uk-air.defra.gov.uk/networks/network-info?view=urn>, Searched Jul 2014.
- Flandrin, P., Goncalves, P., 2004. Empirical mode decompositions as data-driven wavelet-like expansions. *Int. J. Wavelets Multiresolution Information Process.* 2 (4), 477–496.
- Holloway, T., Levy, H., Kasibhatla, P., 2000. Global distribution of carbon monoxide. *J. Geophys. Res. Atmos.* 105, 12123–12147.
- Hopke, P.K., Ito, K., Mar, T., Christensen, W.F., Eatough, D.J., Henry, R.C., Im, E., Laden, F., Lall, R., Larson, T.V., Liu, H., Neas, L., Pinto, J., Stolzel, M., Suh, H., Paatero, P., Thurston, G.D., 2006. PM source apportionment and health effects: 1. Intercomparison of source apportionment results. *J. Expos. Sci. Environ. Epidemiol.* 16, 275–286.
- Ketzel, M., Wählin, P., Berkowicz, R., Palmgren, F., 2003. Particle and trace gas emission factors under urban driving conditions in Copenhagen based on street and roof-level observations. *Atmos. Environ.* 37, 2735–2749.
- Lenschow, P., Abraham, H.J., Kutzner, K., Lutz, M., Preuß, J.D., Reichenbacher, W., 2001. Some ideas about the sources of PM10. *Atmos. Environ.* 35 (Suppl. 1), S23–S33.
- Mead, M.I., Popoola, O.A.M., Stewart, G.B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J.J., McLeod, M.W., Hodgson, T.F., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J.R., Jones, R.L., 2013. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmos. Environ.* 70, 186–203.
- Morris, R.D., 2000. Low-level carbon monoxide and human health. In: Penney, D.G. (Ed.), *Carbon Monoxide Toxicity*. CRC Press LLC, Boca Raton, FL, pp. 381–392.
- Novelli, P.C., Masarie, K.A., Lang, P.M., 1998. Distributions and recent changes of carbon monoxide in the lower troposphere. *J. Geophys. Res. Atmos.* 103, 19015–19033.
- Ruckstuhl, A., Henne, S., Reimann, S., Steinbacher, M., Vollmer, M., O'Doherty, S., Buchmann, B., Hueglin, C., 2012. Robust extraction of baseline signal of atmospheric trace species using local regression. *Atmos. Meas. Tech.* 5, 2613–2624.
- Seinfeld, J.H., Pandis, S.N., 1998. In: Seinfeld, John H., Pandis, Spyros N. (Eds.), *Atmospheric chemistry and physics: from air pollution to climate change*. J. Wiley, New York; Chichester.
- Viana, M., Kuhlbusch, T.A.J., Querol, X., Alastuey, A., Harrison, R.M., Hopke, P.K., Winiwarer, W., Vallius, M., Szidat, S., Prévôt, A.S.H., Hueglin, C., Bloemen, H., Wählin, P., Vecchi, R., Miranda, A.I., Kasper-Giebl, A., Maenhaut, W., Hitznerberger, R., 2008. Source apportionment of particulate matter in Europe: a review of methods and results. *J. Aerosol Sci.* 39, 827–849.
- Vreman, H.J., Wong, R.J., Stevenson, D.K., 2000. *Carbon Monoxide in Breath, Blood, and Other Tissues, Carbon Monoxide Toxicity*. CRC Press, USA.
- Wayne, R.P., 2000. *Chemistry of Atmospheres*. Oxford University Press.
- Zellweger, C., Hügliin, C., Klausen, J., Steinbacher, M., Vollmer, M., Buchmann, B., 2009. Inter-comparison of four different carbon monoxide measurement techniques and evaluation of the long-term carbon monoxide time series of Jungfraujoch. *Atmos. Chem. Phys.* 9, 3491–3503.