Forecasting Accuracy for Emergency Measures

Gary Fuller, Timothy Baker, Heather Walton

Environmental Research Group

King's College London

(Imperial College London from July 2020)

November 2020

Title	Forecasting accuracy for emergency measures

Customer	Greater London Authority / Transport for London

Environmental Research Group
King's College London
4th Floor
Franklin-Wilkins Building
150 Stamford St
London SE1 9NH
Tel 020 7848 4044
Fax 020 7848 4045

	Name	Date
Authors	Gary Fuller	29th May 2019 24th June 2019* 29th November 2019* 3rd November 2020*
	Timothy Baker	29th May 2019 24th June 2019
Reviewed by	Heather Walton	5th June 2019 24th June 2019* 29th November 2019* 4th November 2020*
Approved by	Heather Walton	15th June 2019 29th November 2019* 4th November 2020*

** Additional material added following GLA comments and workshop.*

Contents

1. Summary

1.1. Introduction

This report is part of a wider set of work packages that gather evidence on the health impacts of emergency / short-term action plans to control air pollution episodes in London. Combined, they explore the information that might be needed if the Mayor wanted to consider introducing a scheme of emergency measures. The work packages a- e :

a. Summarised the health effects of short term exposure to high levels of air pollution.
b. Estimated the magnitude of the health impact of high air pollution episodes in London
c. Reviewed the evidence on the effectiveness of emergency measures elsewhere (e.g. Madrid, Paris, Beijing)
d. Assessed the accuracy of existing air quality forecasting for use in triggering emergency air quality measures (this report).
e. Convened an expert workshop that considered the work packages a to d and the conclusions that could be drawn from them.

Air pollution forecasts can provide valuable warning of impending periods when air pollution might place an additional health burden on London's population. If these forecasts are sufficiently accurate, they can be used as the basis of short-term action plans. These plans can advise people to take extra precautions and mandate changes to decrease emissions.

The London Mayor's air quality forecast service has been operating since 2017. In February 2018, operation of the system was transferred to King's College London and its scope was extended.

The system sends out advisory messages by email to schools and other stakeholders if air pollution is forecast to be moderate, high or very high according to Defra's daily air quality index. When high or very high air pollution is forecast, information is also displayed on Transport for London infrastructure including electronic displays at bus stops, in the underground and beside trunk roads.

This report analyses the performance of the Mayor's air quality forecast system. The current system focuses on providing public information and takes a precautionary approach. The report identifies changes that would allow the system to evolve towards the provision of information for emergency or short-term actions to reduce air pollution during episodes.

The Mayor's forecast uses an ensemble approach, combining the publicly available forecasts from three providers: Defra (the Met Office), LondonAir (King's College London / Imperial College London) and AirText (Cambridge Environmental Research Consultants - CERC). Each provider uses a different input data and a different forecast methodology.

Commented [SM1]: Overall reads well.
Same formatting comments as per report 2.
Review the use of uppercase letters when referring to high/moderate days etc. as this is inconsistent in the document.

Same comments from report 2 apply to the summary section here.

Commented [FG2R1]: Thanks. There were some differences between the capitalization in the main text and in the appendices. I've searched through and sorted these.

Commented [SM3]: Define

Commented [FG4R3]: done

1.2. Methods

Air pollution forecasts are a categorical prediction; they provide a "yes" or "no" prediction that can be assessed against a set of "yes" or "no" observations. It is important to assess not only the times when the model correctly warns of an episode (true positives) but also the times that it misses episodes (false negative), incorrectly predicts an episode (false positives) and correctly predicts no episode (true negatives). These are normally evaluated using a 2 x 2 contingency table, as shown in Table 1.

		Pollution Observed	
		Yes	No
Pollution Forecast	Yes	a (true positive)	b (false positive)
	No	c (false negative)	d (true negative)

Table 1 2x2 contingency table showing the possible forecasts and outcomes

Air pollution episodes (moderate, high or very high) are relatively rare events, occurring on around 10% of days. A prediction of low air pollution everyday would achieve substantial accuracy (the ratio of true events to the total), but it would have no skill nor utility. Instead air pollution forecasts should be assessed using a skill score. Here the Gilbert skill score was used. This ranges from zero (no true positives) to one (all true positives), relative to true and false positives and false negatives.

Forecasts from 1st February 2018 to the end of April 2019 were compared to times when there were air pollution episodes at background locations (away from local sources) and additionally at times when episodes also happened close to roads.

1.3. Results

For predictions of moderate air pollution on the following day, the system had accuracies of 90% and 91% for background and on the roadside proximity metric respectively. Using the preferred skill-based assessment, the forecast had skill scores of 0.5 and 0.53; mid-range values for this metric. The system takes a precautionary approach by design, commensurate with its objectives to supply public information. It assumes a worst-case outcome and therefore over-predicts but misses very few episodes.

Over the analysis period, five forecasts of high air pollution were made but high was only measured on one of these days according to the episode criteria. Widespread moderate air pollution was measured on each day when high was forecast and high was measured in parts of London or on the fringes of the capital on three of these days.

High air pollution was measured on four days but not predicted on three of these. One of these days was predicted to be moderate and two were predicted to be low by all three forecast providers. This highlights the intrinsic difficulties in predicting extremes of air pollution where small changes in weather cause marked changes in concentrations. This may reflect the limitations of the current state of the art in air pollution forecasting.

1.4. Discussion and conclusions

The Mayor's forecasting service is an ensemble of forecasts from three different providers who use different forecasting techniques. This is a great strength of the service. There were clear improvements in the skill score when there was agreement between the forecasters. Only issuing forecast alerts when there is good or medium agreement between forecasters would make the service less precautionary. This would increase the skill scores from 0.5 and 0.53 to 0.67 and 0.8.

However, this would lead to times when a publicly available forecast was predicting moderate air pollution, but this was not relayed by the Mayor's service.

The current service is meeting a requirement to provide precautionary advice to the public. However, if the forecasts are to be used to enact emergency or short-term measures, such as traffic restrictions or reduced cost public transport, a clearer definition of a pollution episode will be needed, and a less precautionary service may be required. An alternative would be to trigger such actions on the combination of severity of the episode and a probability, based on the degree of agreement between providers.

2. Purpose

This report is part of a wider set of work packages that gather evidence on the health impacts of emergency / short-term action plans to control air pollution episodes in London. Combined, they explore the information that might be needed if the Mayor wanted to consider introducing a scheme of emergency measures.

The work packages a to e :

a. Summarised the health effects of short term exposure to high levels of air pollution.
b. Estimated the magnitude of the health impact of high air pollution episodes in London

c. Reviewed the evidence on the effectiveness of emergency measures (e.g. Madrid, Paris, Beijing).
d. Assessed the accuracy of existing air quality forecasting for use in triggering emergency air quality measures (this report).

e. Propose evidence-based suggestions of where to set a threshold for triggering emergency measures to tackle air pollution
f. Convened an expert workshop that considered the work packages a to d and the conclusions that could be drawn from them.

This report for work package e) draws on data from the Mayor's air quality forecasting system that is operated at King's. It compares forecast air pollution to that measured at Defra monitoring sites and those in the London Air Quality Network. It uses this comparison to assess the current forecasting system and to make recommendations for changes that would be needed to allow the service to move towards a tool for triggering emergency or short-term actions.

3. Introduction

The health impacts of air pollution are well recognised and have been the motivation for air quality legalisation in states around the globe. Within the European Union, Directive 2008/50/EC sets out air quality limits (which are legally binding) and target values (that are not binding) along with dates for compliance forming a pan-European framework for the assessment and management of air pollution.

The Directive also includes alert thresholds when short-term action plans or emergency measures need to be triggered. These may,

> "… provide for effective measures to control and, where necessary, suspend activities which contribute to the risk of the respective limit values or target values or alert threshold being exceeded. Those action plans may include measures in relation to motor-vehicle traffic,

construction works, ships at berth, and the use of industrial plants or products and domestic heating. Specific actions aiming at the protection of sensitive population groups, including children, may also be considered in the framework of those plans."

Commented [FG5R3]: There seems to be a limit on the size of foot notes. I've therefore moved this into the main text.

Here, 'limit value' shall mean a level fixed on the basis of scientific knowledge, with the aim of avoiding, preventing or reducing harmful effects on human health and/or the environment as a whole, to be attained within a given period and not to be exceeded once attained;

'target value' shall mean a level fixed with the aim of avoiding, preventing or reducing harmful effects on human health and/or the environment as a whole, to be attained where possible over a given period;

'alert threshold' shall mean a level beyond which there is a risk to human health from brief exposure for the population as a whole and at which immediate steps are to be taken by the Member States.

As with other member states the UK transposed the directive into domestic law. This was achieved through the Air Quality Regulations 2010. These *require* the Secretary of State to set up short-term action plans if there is a risk of alert thresholds being exceeded and they *may be drawn up* if there is a risk of exceeding the target values.

Within the UK, short-term actions in the event of an air pollution episode provide public information only and do not contain measures to reduce emissions or concentrations. As well as the legal notifications required by the Directive the UK operates an air quality index to help the public translate air pollution concentrations into an index number and a health risk. The index number 1 to 10 and the health risk bands are described as low, moderate, high or very high. The index applies to those pollutants where there is good evidence of a health risk from short-term exposure: nitrogen dioxide (NO_2), ozone (O_3), sulphur dioxide (SO_2), PM10 and PM2.5. Carbon monoxide was removed from the index when it was last revised in 2012. The index is summarised in Table 2.

Band	Index	Ozone	Nitrogen Dioxide	Sulphur Dioxide	PM2.5 Particles	PM10 Particles
		Running 8 hourly mean	Hourly mean	15 minute mean	24 hour mean	24 hour mean
		µg m-3	µg m-3	µg m-3	µg m-3	µg m-3
Low						
	1	0-33	0-67	0-88	0-11	0-16
	2	34-66	68-134	89-177	12-23	17-33
	3	67-100	135-200	178-266	24-35	34-50
Moderate						
	4	101-120	201-267	267-354	36-41	51-58
	5	121-140	268-334	355-443	42-47	59-66
	6	141-160	335-400	444-532	48-53	67-75
High						
	7	161-187	401-467	533-710	54-58	76-83
	8	188-213	468-534	711-887	59-64	84-91
	9	214-240	535-600	888-1064	65-70	92-100
Very High						
	10	241 or more	601 or more	1065 or more	71 or more	101 or more

Table 2 Summary of the bands and index values in the UK Daily Air Quality Index (COMEAP, 2011) for more information see: http://londonair.org.uk/london/asp/airpollutionindex.asp?IndexDate=2012

By contrast the short-term action plans used in some European cities include controls on emissions. Outside Europe short-term actions have been enacted in a reactive mode to control pollution episodes or in a pre-planned way to minimise air pollution when a city is hosting a prominent

international event, the Olympic Games for instance. Short-term action plans are reviewed in the accompanying report from work package c.

An important part of any protocol is the decision process for enacting the measures in a short-term action plan and, equally important, when to withdraw them. When emergency measures have been used for events such as Olympic games or parades the timing of short-term actions are dictated by the dates of the event. However, this is not the case when the short-term actions are part of a responsive programme during times of adverse air pollution.

A review of short-term action plans across the EU was undertaken in 2012 (Conlon et al 2012). The decision to enact emergency measures varied between areas and between countries. A reliance on measurements alone may mean that the short-term action plan is not enacted until the pollution episode has passed. Most cities therefore rely on a combination of measurements and forecasts, with forecast playing a large role in the decision to continue or terminate the actions. However, there is no broad agreement on how the combination of measurement and modelling are used for these purposes. Conlon et al (2012) also found a large range of forecasting models in use. Most were deterministic dispersion and chemical-transport models that predict air pollution using data on forecast weather and emissions, but others were using statistical approaches that made predictions based on previous weather and past air pollution.

The purpose of this report is to look at the forecasting system that is operating in London and to consider its role in any future short-term action plans.

Commented [SM6]: Some examples (or a table of examples) would be beneficial in this section.

Commented [SM7R6]:

Commented [FG8R6]: These are reviewed in detail in the report for work package b. It would be a divergence to include them in this report that is focusing on forecasts rather than action plans.

Commented [WHA9R6]: The review of schemes report is not report c.

4. The London Mayor's air quality forecast system

Air pollution in London is measured by several stakeholders; Defra, local authorities, universities, Transport for London and private organisations such as business improvement districts. Air pollution is measured at a range of locations including kerbs and roads as well as background locations away from pollution sources. The majority of these measurements are collated by King's College London as part of the London Air Quality Network. These are provided to the public via the *LondonAir* website and apps.

Forecasts for air pollution in London are available from several providers. The longest established services are provided by Defra (run by the Met Office), LondonAir (King's College London) and AirText (CERC).

The three forecast providers use different input data and different systems:

- The Met Office forecast is a national deterministic air pollution model that uses measured data to "nudge" its results. It forecasts for specific points based on a 12 x 12 km grid (Neal et al 2014). The Met Office makes clear that its forecast does not represent the very localised increases in pollution that you might find close to roads or in urban centres. The forecast represents the background and regional air quality away from these strong sources of pollution.
- The forecast from King's is based firmly on air pollution measurements. These include source tracers and particle composition to infer sources (wood burning, traffic, secondary particles etc). Forecast weather and air paths (back trajectories) are also used.
- AirText use a London-focused deterministic air pollution model with a spatial resolution of 7 x 7 m. Air pollution from sources outside London are provided by the EU Copernicus system.

All forecasting systems rely on weather predictions. Air pollution episodes occur during a narrow range of weather conditions which are hard to predict. A further challenge is the lack of time

resolved data on air pollution emissions. Instead models have to rely on annual emissions information that does not reflect seasonal, day of week and even hour by hour changes.

Since 2017 these forecasts have been used as part of the London Mayor's air pollution alerting system that provides forecasts according to Defra's Daily Air Quality Index. This index describes air pollution on a one to ten scale and also divides air pollution into four overall bands: low (one to three), moderate (four to six), high (seven to nine) and very high (ten). These bands are linked to air pollution advice for the public (COMEAP, 2011).

Commented [SM10]: Check the bands. Six is included in both moderate and high.

Commented [FG11R10]: My error – thank you.

Although many roads in London breach legal limits for nitrogen dioxide, the localised nature of this pollutant means that it is rarely the cause of widespread moderate, high or very high air pollution. Widespread moderate, high and very high pollution are most frequently caused by O_3, $PM_{2.5}$ and PM_{10}. O_3 is not emitted directly into our air but instead it forms from chemical reactions between other air pollutants. The same is true of much of the PM2.5 and PM10 that we experience in London. Predicting the conditions when these chemical reactions will occur presents a further challenge for air pollution forecasting.

The Mayor's system sends out advisory messages by email to schools and other stakeholders if air pollution is forecast to be moderate, high or very high on the following day. When high or very high air pollution is forecast, information is also displayed on Transport for London infrastructure including electronic displays at bus stops, in the underground and beside trunk roads.

An early version of the system was operated by staff from the Greater London Authority who drew upon information from forecast providers to create the Mayor's forecast. Since early 2018 the system has been operated by King's and expanded to include the automated emails, an application programming interface (API) and Twitter links. The final decision to activate notices on TfL infrastructure is still taken by the Mayor's office following a recommendation by King's.

Each day staff within the operations centre at King's retrieve forecasts from Defra, AirText and the London Air Quality Network (King's). An ensemble forecast is created using a combination of a rules-based system and expert judgement that provides a forecast band and a confidence:

- Following the rules of the Daily Air Quality Index the band is determined by the most pessimistic forecast from the three providers.
- For moderate and high banding, the confidence is determined by the level of agreement between the forecast providers. If all agree then the confidence is good. If two agree then the confidence is medium and if there is no agreement to support the most pessimistic forecast, then the confidence is poor.
- For low banding, medium and poor confidence are used by the King's forecaster to indicate an expert view of uncertainty in the forecast.
- Staff in the King's operations centre draw upon the most recent measurements to check the sensibility of the forecast. They can over-ride the ensemble if the most pessimistic forecast is not supported by current measurements and forecast weather conditions or if the spread between the forecasts is greater than one band. This decision is usually undertaken in consultation with GLA staff.

Forecasts are made on the day, one, two and three days ahead. Ahead of public holidays, forecasts are also made four days ahead. The forecast one day ahead is used for sign boards and other alerts.

Time spent in travel environments and alongside roads makes a disproportionate contribution to daily exposure. For instance, personal sampling undertaken as part of the EU PASTA project found

that participants in five cities (including London) spent on average 7.5% of their time in transport modes but this contributed to over 18% of their daily black carbon exposure (de Nazelle, personal communication). In an earlier study in Barcelona participants spent 6% of their time in transit but this contributed 24% of their daily NO_2 exposure (de Nazelle et al 2013). It may therefore be expedient to advise people to avoid polluted roads where they can. Forecasts are therefore issued for roadside and background environments.

This report will look at the ensemble forecasts that have been issued since the system began operating at King's on 1st February 2018. This report assessed the outputs from Mayor's system and not at the performance of the input data from each forecast provider.

5. Assessing forecast accuracy

The performance of air pollution models is typically evaluated by comparing predicted and measured concentrations. The model performance can then be expressed in terms of metrics of correlation, bias or the fraction of predictions that have fallen within a predefined range of the measured value; the proportion that are within ± 50% for example.

However, air pollution forecasts are a categorical prediction; they provide a "yes" or "no" prediction that can be assessed against a set of "yes" or "no" observations. The performance of such models needs to be considered carefully. It is important to assess not only the times when the model correctly warns of an episode (true positives) but also the times that it misses episodes (false negative), incorrectly predicts an episode (false positives) and correctly predicts no episode (true negatives). These are normally evaluated using a 2 x 2 contingency table, as shown in Table 3~~Table 2~~.

		Pollution Observed	
		Yes	No
Pollution Forecast	Yes	a (true positive)	b (false positive)
	No	c (false negative)	d (true negative)

Table 3 2x2 contingency table showing the possible forecasts and outcomes.

These can also be expressed in terms of a receiver - operator characteristic or the correct detection of a signal. Rather than the true / false and negative / positive terminology in Table 3~~Table 2~~, this frames the outcomes in terms of hit, miss, or false alarm or no event, as shown in Table 4~~Table 3~~.

		Pollution Observed	
		Yes	No
Pollution Forecast	Yes	Hit	False Alarm
	No	Miss	No event

Table 4 Alternative terminology for 2x2 contingency table showing the possible forecasts and outcomes.

Various metrics exist to determine the accuracy or skill of such forecasts. The debate around this is almost as old as weather forecasting itself and this is still relevant today. Weather forecasting using synoptic maps began in Europe and the US in the period 1850 to 1870 along with the founding of national and regional metrological services. Issues around the assessment of the accuracy of UK storm forecasts were debated in the 1860s but this issue received surprisingly little attention until the 1880s with the so-called Finley affair. In 1884, Sergeant John Finley of the US Army Signal Corps published a paper on the accuracy of his tornado forecasts for 18 areas to the east of the Rocky Mountains (Murphy, 1996). Tornados are, thankfully, rare events but they can cause considerable

damage. Finley claimed an accuracy of 96% based on a period with 56 tornados. Here accuracy was defined as the ratio of true events to the total, in terms of Table 1 as:

Accuracy = 100 (a+d) / (a+b+c+d)

This prompted an important debate on how to assess these types of forecasts, which still resonates over 100 years later. With the rarity of tornados, the prediction of no tornado every day would achieve a very high accuracy of 98%, actually better than Finley's forecasts, but would not be a skillful predictor nor a useful one.

Like tornados, air pollution episodes are relatively rare events and the prediction of low air pollution everyday would achieve substantial accuracy, but it would have no skill nor utility. Appropriate skill scores for categorical weather forecasting were considered by Agnew et al (2009) for air quality, and their use for the air pollution bands was reviewed against air pollution data in the UK by King's (in COMEAP (2011)) when the Daily Air Quality Index was created. COMEAP (2011) selected the Gilbert skill score (GSS) as the most appropriate metric. This has the advantage of placing less weight on those occasions when low air pollution was correctly forecast. The Gilbert skill score is defined in terms of the Table 3~~Table 2~~ terminology as:

Gilbert skill score (GSS) = a / (a+b+c)

6. Methods

Air pollution forecasts made under the London Mayor's system were retrieved from the database at King's. These were then compared with measured air pollution from the air pollution database at King's. A total of 449 days were evaluated from February 2018 (the start of operation at King's) to the end of April 2019 (when analysis began).

The Mayor's forecast is one pollution band for the whole of London, although the accompanying text often describes areas of specific concern. With over 80 measurement sites operating in London a decision had to be made to determine what constitutes an air pollution episode in order to match forecast and measured outcome. For this study the following criteria were used:

- Episode at background: the episode (moderate, high or very high) had to be measured at one background or suburban site and one other monitoring site. It is this criterion that was used in the accompanying report on the health impact of air pollution episodes (Walton et al in prep). If an episode criterion was not reached, then air pollution was recorded as low. Please note the Mayor's forecasting system also includes air pollution close to roads and whilst this is an appropriate episode definition for protecting health, it does not match the criteria against which the forecasts are issued (Section 4).

- Widespread roadside or background episode: an episode (moderate, high or very high) that was measured according to the background criteria OR was measured at 10% of road or kerbside monitoring sites. Again, if this criterion was not reached then air pollution was recorded as low. This is the best match for the criterion for which the forecasts are made.

Whilst each forecast can be considered as a binary categorical prediction with a binary outcome, the forecast system as a whole is more complex. A forecast of moderate would not be a false alarm if the measured outcome was high or very high for instance. The forecast and observed 2 x 2 matrix was therefore redefined as shown in Table 5~~Table 4~~.

		Pollution Observed	
		Yes	No
Pollution Forecast	Yes	Forecast and observed same banding	Observed lower than forecast
	No	Observed higher than forecast	No forecast for this band and didn't occur

Table 5 2x2 contingency table showing the possible forecasts and outcomes.

Separate assessments were made for each confidence level and for each time period; forecasts on the day, one day, two days and three days in advance. Forecast made four days in advance to cover public holiday weekends were also included.

7. Results

7.1. Forecasts issued

The forecasts for air pollution one day in advance are summarised in Figure 1.

These forecasts predict the worst-case conditions at either background or close to roads. The clear majority of the 449 forecasts were for low air pollution. There were 93 forecasts for moderate or high air pollution. No very high forecasts have been issued.

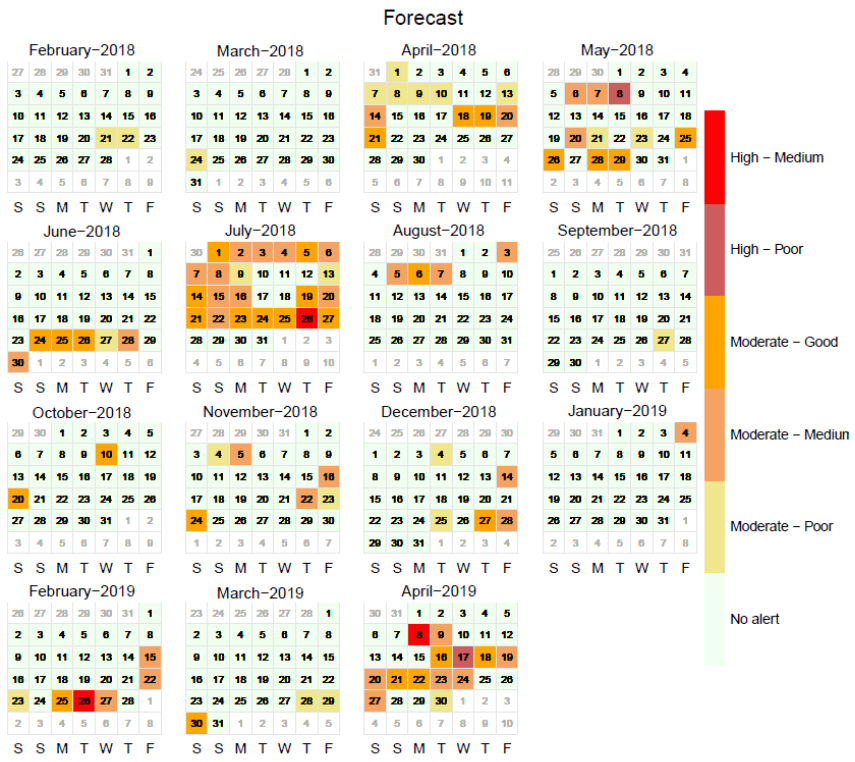

Figure 1 Calendar showing forecasts issued Feb 2018 to April 2019for one day ahead, showing both forecast band (moderate and high) and confidence (good, medium and poor).

Forecasts were also issued on the day of the episode and several days ahead. These are detailed below.

7.2. Episode at background

Figure 2 shows the times when moderate or high pollution was measured according to the background episode criteria. These show a good correspondence to the time periods when the moderate or high air pollution forecast were issued, though clearly the match with Figure 1 is not complete.

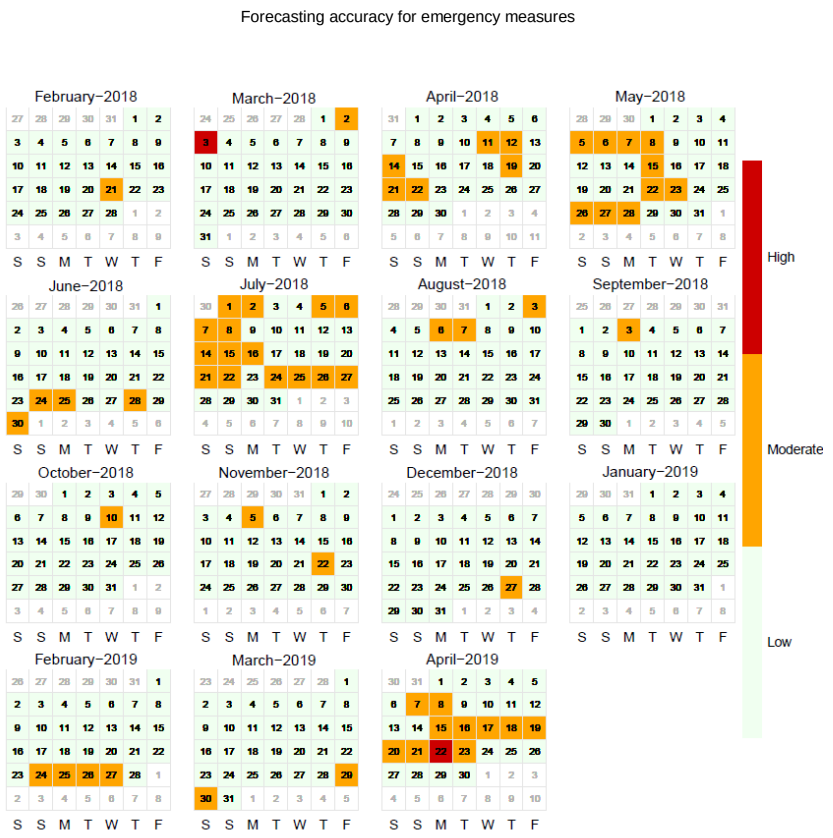

Figure 2 Calendar showing air pollution bands measured during Feb 2018 to April 2019for the background episode criteria.

Moderate

Table 6~~Table 5~~ shows the forecast and outcomes for moderate air pollution forecasts issued one day in advance against the background episode criteria. It must be remembered that the forecasts are also taking into account conditions close to roads. The forecast is therefore more pessimistic than this criterion. This was reflected in the results. When predicted, moderate air pollution was observed on 44 days, i.e. the forecast for moderate was correct. Forecasts were incorrect on 44 days, with 43 being false alarms and an episode was missed on one day. High air pollution was measured once when that day was forecast as moderate. The accuracy was 90% and the skill score was 0.5[1].

[1] Accuracy is included for comparability to other forecast assessments, but skills-based metrics are more applicable for categorical forecasts.

Moderate forecast one day in advance (All Confidence Levels) GSS=0.5		Pollution Observed	
		Yes	No
Pollution Forecast	Yes	44	43
	No	1	361

Table 6 2x2 contingency table for moderate forecast one day ahead assessed against the criteria for background episodes.

High and very high

Table 7~~Table 6~~ shows the results for high forecasts assessed against the background episode criteria. The forecast is more pessimistic than this criterion. This was reflected in the results. Five high forecasts were made, and these were not measured at background. The accuracy was 98% but the skill score was zero. There were no forecasts of very high.

High Forecast one day in advance (All Confidence Levels) GSS=0		Pollution Observed	
		Yes	No
Pollution Forecast	Yes	0	5
	No	0	444

Table 7 2x2 contingency table for high forecast one day ahead assessed against the criteria for background episodes.

Confidence intervals and time ahead

Table 8~~Table 7~~ shows the GSS scores by confidence interval and different days ahead. Skill scores show the full range between zero and one but scores at these extremes of the GSS range are for small numbers of forecasts. Each score of one is from a single forecast. With one exception, accuracy scores were all greater than 95%[2].

The number of forecasts and accuracy scores are shown in the appendix.

[2] Accuracy is included for comparability to other forecast assessments, but skills-based metrics are more applicable for categorical forecasts.

GSS		Days Forecast in advance				
Forecast	Confidence	On day	1	2	3	4
Low	Good	0.99	0.98	0.98	0.93	-
	Medium	0.80	0.81	0.86	1.00	-
	Poor	1.00	0.67	0.67	1.00	1.00
Moderate	Good	0.76	0.70	0.72	0.67	-
	Medium	0.54	0.57	0.56	0.86	0.00
	Poor	0.17	0.13	0.36	0.50	1.00
High	Good	1.00	-	-	-	-
	Medium	0.00	0.00	0.00	-	-
	Poor	0.00	0.00	0.00	-	-
Very High	Good	-	-	-	-	-
	Medium	-	-	-	-	-
	Poor	-	-	-	-	-

Table 8 Gilbert's Skill scores for background episodes. The table shows the air pollution band forecast and the confidence. Dashes denote no forecast for this criterion.

7.3. Episode at background or roadside

Figure 3 shows the times when moderate or high pollution was measured according to the background or roadside episode criteria. These show a good correspondence to the time periods when the moderate or high air pollution forecast were issued, though again the match with Figure 1 is not complete. Compared with Figure 2, additional episodes and more severe episodes can be seen. These were mainly during the winter months when episodes are generally caused by the poor dispersion of air pollution in the cold, winter air. At these times concentrations build up close to sources, such as roads.

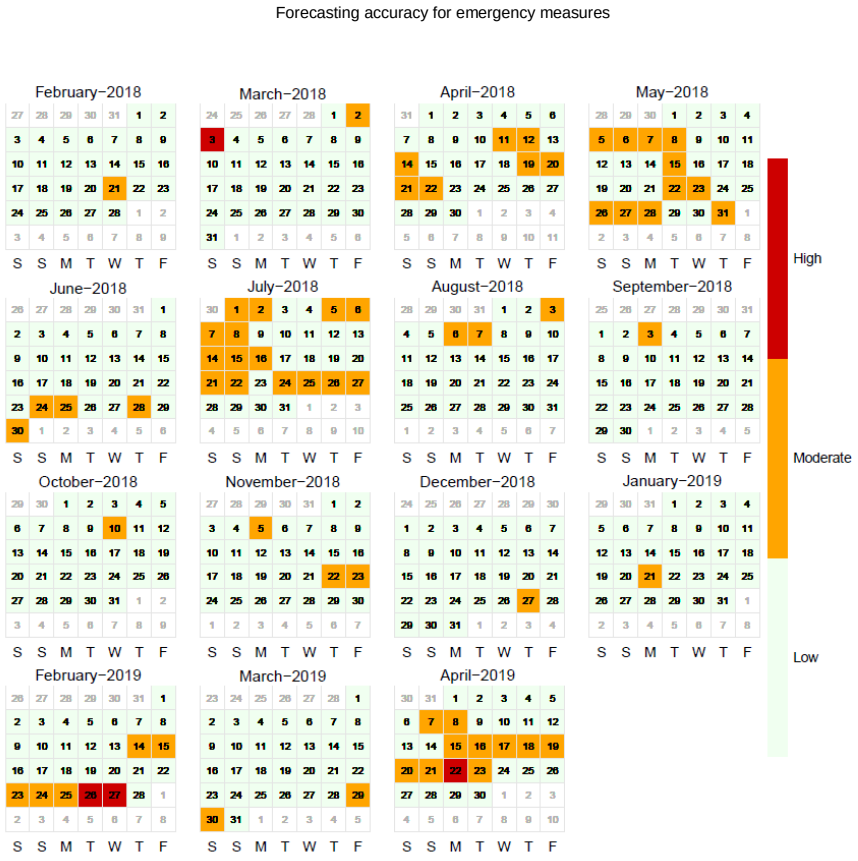

Figure 3 Calendar showing air pollution bands measured during Feb 2018 to April 2019for the background or roadside episode criteria.

Moderate

Table 9~~Table 8~~ shows forecast and outcomes for moderate air pollution forecasts issued one day in advance assessed against background or roadside criteria. When predicted, moderate air pollution was observed on 47 days. Forecasts were incorrect on 41 days, 39 being false alarms and episodes were missed on two days. High air pollution was measured on two days that were forecast to be moderate. The accuracy was 91% and the skill score was 0.534[3].

Moderate forecast one day in advance (All Confidence Levels) GSS=0.534		Pollution Observed	
		Yes	No
Pollution Forecast	Yes	47	39
	No	2	361

Table 9 2x2 contingency table for moderate forecast one day ahead assessed against the criteria for background or roadside episodes.

High

Table 10~~Table 9~~ shows the results for high forecasts assessed against the roadside or background episode criteria. Five high forecasts were made, and one of these was measured. The accuracy was 91% and the skill score was 0.2. There were no forecasts of very high.

High forecast one day in advance (All Confidence Levels) GSS=0.2		Pollution Observed	
		Yes	No
Pollution Forecast	Yes	1	4
	No	0	445

Table 10 2x2 contingency table for high forecast one day ahead assessed against the criteria for background or roadside episodes.

Confidence intervals and time ahead

Table 11~~Table 10~~ shows the GSS scores by confidence interval and different days ahead. Skill scores show the full range between zero and one, but these events are for small numbers of days. Each score of one is from a single forecast.

Accuracy scores are shown in the appendix. With one exception these were all greater than 94%.

[3] Accuracy is included for comparability to other forecast assessments, but skills-based metrics are more applicable for categorical forecasts.

GSS		Days Forecast in advance				
Forecast	Confidence	On day	1	2	3	4
Low	Good	0.98	0.97	0.97	0.91	1.00
	Medium	0.76	0.81	0.82	1.00	-
	Poor	1.00	0.67	0.67	1.00	
Moderate	Good	0.76	0.70	0.80	0.67	0.00
	Medium	0.62	0.60	0.59	0.86	1.00
	Poor	0.21	0.22	0.36	0.50	-
High	Good	1.00	-	-	-	-
	Medium	0.67	0.33	0.00	-	-
	Poor	0.00	0.00	1.00	-	-
Very high	Good	-	-	-	-	-
	Medium	-	-	-	-	-
	Poor	-	-	-	-	-

Table 11 Gilbert Skill scores for background or roadside episodes. The table shows the air pollution band forecast and the confidence. Dashes denote no forecast for this criterion.

8. Discussion

Air pollution forecasts can provide valuable warning of impending periods when air pollution might place an additional health burden on London's population. If these forecasts are sufficiently accurate they can be used as the basis of short-term action plans. These plans can advise people to take extra precautions and mandate changes to decrease emissions.

Evaluating the accuracy or skill of an air pollution forecasting system is more complex than it might initially appear.

The first complexity is the so-called change of support problem. Forecast models generally predict air pollution over a wide area but measurements are only made at specific locations. For this project we processed the measurement data using two criteria to define air pollution episodes, as described in Section 6. One was a criterion for episodes at background that were likely to cover a wide area. A second criterion defined episodes that were affecting either a wide area or many roads. However, none of these matched the outputs from the individual forecast providers. The Defra (Met Office) forecast is for background areas of 12 x 12 km (Neal et al, 2014), AirText is forecast on a finer scale grid of around 7 x 7 m and the King's forecast estimates concentrations measured at monitoring sites. Whilst this diversity of approaches strengthens the current Mayoral forecast system it makes evaluation difficult when none of the forecasters are working to the same criteria as that being used in the evaluation.

The second complexity arises from the relative rarity of air pollution episodes as they are defined by the UK index. This study looked at 449 days in London, since the start of the Mayor's forecasting system in its current form until the end of April 2019. Using our background or roadside episode criteria, high air pollution was measured on four days. It was measured on two days at background only. Moderate air pollution was more frequent. This was measured 66 days if we use the background plus roadside episode criteria and on 61 days at background only (14% of days) and. It is therefore possible to create an accurate forecasting system by simply predicting low air pollution every day, but this would have no skill or utility. Looking at forecasts up to four days ahead the

Commented [SM12]: Is this 6 in total or is this 2 part of the 4 days?

Commented [FG13R12]: Yes, the first criteria is roadside or background. The second criteria is background only. I changed the wording. I changed the order of the next sentence to match the order of this one.

Commented [WHA14]: 14% of all days? I presume it is. It just read a bit as though it was going to be the % background of background plus roadside. But 61/66 is not 14%!

accuracy of the Mayor's forecasts was almost always greater than 94% but this is not the same as the skill.

The Mayor's forecasting system is based on an ensemble of forecasts from three providers. It adopts a precautionary approach by issuing a forecast based on the worst outcome from the ensemble. The skill of the forecast was greatly enhanced by this ensemble approach compared with the alternative of using just one of the three providers. As detailed Table 8~~Table 7~~ and Table 11~~Table 10~~, there were clear improvements in the skill of the forecast when all providers agreed (good confidence) compared with times when the forecast was made on the basis of a single provider (poor confidence). Skills scores for forecasts of moderate air pollution were greater than 0.67 when all providers agreed but never greater than 0.5 when based on a single provider. It was expected that forecast issued several days ahead would have less skill than those issued the day before the episode or on the day, but no clear pattern was seen for moderate forecasts.

As shown in Figure 1, forecasts clearly clustered during specific periods of adverse weather for air pollution. These include springtime, especially April and May in both 2018 and 2019. This seasonal pattern is due to the combination of emissions from traffic and industry along with the increased ammonia from agriculture at this time of year. These pollutants react to form secondary airborne particles. The ammonia emissions are mainly due to seasonal fertilizer use and muck spreading as manure accumulated over winter is applied to fields. In the UK, these events were first highlighted in 1996 (Stedman, 1996). There is however good evidence that they also prevailed in the earlier parts of the 20th century but were not detected by the pollution measurement equipment that was routinely operated at that time (Fuller, 2018). Summer 2018 was a further period of adverse air pollution. This is the normal season for ozone episodes.

Totalling the rows in Table 7~~Table 6~~ and Table 11~~Table 10~~, five forecasts of high air pollution were made but high was only measured on one of the days when it was forecast, according to our episode criteria. The outcome on each day is detailed in Table 12~~Table 11~~. High air pollution met the episode criteria on the 16th February 2019. High air pollution was measured at single monitoring sites on two of the days (8th May 2018 and 26 July 2018) and high air pollution was measured in towns just beyond the M25 (Dartford and Sevenoaks) on 17th April 2019. This shows the complex nature of near-miss circumstances for forecast evaluation that are not well reflected in the categorical assessment.

Date	Confidence	Outcome
8th May 2018	Poor	High was measured at one road.
26th July 2018	Medium	High was measured at one background site. Widespread moderate.
26th February 2019	Medium	Widespread high at eight sites and very high at one road.
8th April 2019	Medium	No high was measured but widespread moderate.
17th April 2019	Poor	Widespread moderate. No high within London but high at three sites in west Kent, just outside the M25.

Table 12 Outcome when high air pollution was forecast.

High air pollution was measured on four days, according to the background and roadside episode criteria[4], and on two days using the background episode criteria as shown in Figure 2, Figure 3 and detailed in Table 13~~Table 12~~. Concentrations can be found in Appendix C – London mean pollution concentrations on high days~~Appendix C – London Mean Pollution Concentrations on High Days~~. Moderate air pollution often occurs for consecutive days, but high air pollution days were isolated, see Appendix D – Persistence of pollution episode conditions~~Appendix D – Persistence of pollution episode conditions~~ . For two of the high four days, all three providers forecast that air pollution would be low. This highlights the difficulties of predicting extremes of air pollution.

Measured background criteria	Measured background or roadside criteria	Forecast band	Forecast confidence
3rd March 2018	3rd March 2018	Low	-
	26th February 2019	High	Medium
	27th February 2019	Low	-
22nd April 2019	22nd April 2019	Moderate	Good

Table 13 Days on which high air pollution was measured.

Moderate air pollution was measured on around 10% of days allowing a more representative evaluation for this band. Overall, moderate air pollution was predicted with a skill of 0.5 for background and 0.53 for the roadside and background episode criteria. However, if we consider only those forecasts with good confidence, the skill increased to 0.67 for the background criterion and 0.8 for the roadside and background criterion.

Moderate air pollution was forecast one day ahead on 87 days and was measured on around half of these. This gave skill scores of around 0.5 overall but almost no episodes (only two on the roadside or background episode criteria) were missed by the service. An alternative approach to only issue forecast when two or more forecasters agree (good and medium confidence) would have a skill score of 0.63 for the background episode criterion and 0.65 for the background or roadside episode criterion. The number of missed episodes would be unchanged.

Conlon et al (2012) noted that some cities relied on measurements to trigger their emergency measures / short term action plans. This is the case in Madrid, for instance, where emergency measures are triggered for the following day if a threshold is exceeded and if weather conditions are

[4] Number of measured days for each banding using both criteria are tabulated in Appendix B - Number of days in each pollution banding (measurements)~~Appendix B~~

expected to persist (Borge et al 2018). Looking at Figure 2 and Figure 3, it is clear that London often experiences consecutive days of moderate air pollution, but this is not the case for high. Triggering actions on this basis would have resulted in one day of emergency measures for high air pollution and three days of high air pollution where emergency measures would not have been triggered.

Many modellers choose to rely on scalar assessments of skill rather than categorical ones that match the forecast purpose. Little data is therefore available on the skill of categorical air pollution predictions from other air pollution forecasting systems to compare with this assessment of the Mayor's system.

- A measurement-based predictive system was devised by COMEAP (2011) to make very short-term predictions of PM10, PM2.5 and O_3 for public information systems. These so-called trigger events were created to inform the public of a building air pollution episode, before it fully developed. Skills scores (GSS) for the COMEAP trigger predictions for moderate air pollution were 0.55 for PM10, 0.57 for PM2.5. This is similar to the overall skill for the Mayor's forecasts. At 0.79, the GSS for COMEAP triggers for O_3 were within the good confidence band of the Mayoral forecasts.

- Honoré et al (2008) evaluated the French Prev Air forecast system against measurements in range of locations from urban to rural across France and the neighbouring countries. The GSS was around 0.21 for O_3 forecasts, one day ahead. This is a lower skill score than that of the Mayor's system.

9. Conclusions and recommendations

This report has considered the performance of the Mayor's air quality forecast from 1st February 2018 to 30th April 2019, a period of 449 days.

Commented [SM15]: Date – 01 Feb 2018?

Commented [FG16R15]: Yes -changed here and elsewhere in the report.

Care needs to be taken when assessing categorical forecasts of relatively rare events. Due to the large number of low pollution days, simple metrics based on accuracy can be very high even when the model has no skill or utility in the prediction of air pollution episodes. We recommend that skill metrics are used instead and specifically those that are less prone to skew by the large number of non-episode days. The Gilbert Skill Score GSS was used in this report in line with COMEAP (2011).

There is no agreed definition of an air pollution episode. The definition of an episode needs to be tightly tied to the purpose of the forecasting system. Two criteria were used here, one for background and one for a roadside or background episode. These required the predicted pollution to be measured at multiple locations for the episode to be confirmed. Most epidemiological studies use measurements at home address or, most commonly for short-term exposure studies, at a nearby background monitoring site as a proxy for air pollution exposure. Based on the evidence from this type of study, an episode would need to be widespread at background locations before action was justified. Studies that focus on personal exposure highlight that time spent in traffic environments makes a disproportionate contribution to our daily air pollution dose and it may therefore be prudent to also include roadside air pollution within the episode definition.

The definition of a pollution episode should be transparent for the public. It may be difficult to justify ignoring measurements or forecasts of high air pollution because these are not in the right places or not sufficiently widespread. The Mayor's current system is designed to warn the public about air pollution episodes. It therefore takes a precautionary approach. However, a very clear definition of

an episode is needed if the system is used to trigger emergency / short-term actions such as traffic restrictions or reduced cost public transport. We therefore recommend that the definition of a pollution episode needs to be reviewed by the GLA, if such actions are planned.

The Mayor's forecasting service is an ensemble of forecasts from three different providers who use different forecasting techniques. This is a great strength of the service. There were clear improvements in the skill score when there was agreement between the forecasters.

The forecasts ensemble is created in a precautionary manner, taking the worst-case prediction from each provider. This may explain the tendency of the forecasts to over-predict the measurements. The Mayor's system could adopt a matrix similar to the UK's National Severe Weather Warning Service when deciding to activate warnings on TfL infrastructure or take further actions. The combined use of impact and confidence (or in this case likelihood) for severe weather is shown in Figure 4. This is linked to different advice for each warning level / colour.

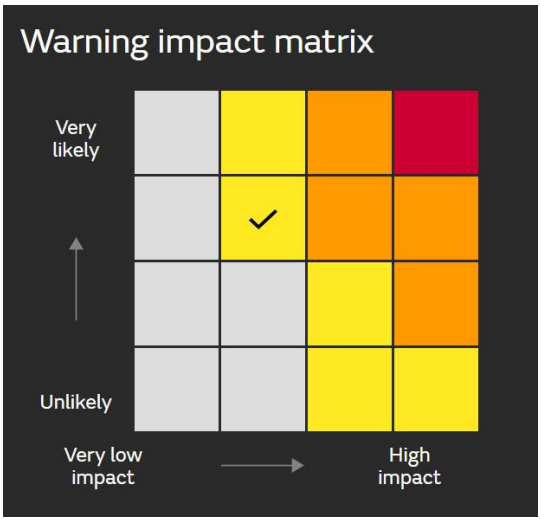

Figure 4 Likelihood and impact is combined in the warning impact matrix from the UK National Severe Weather Warning Service. Warnings are graduated from yellow, amber and red, with red being the highest level (image from UK Met Office).

High air pollution is rare, and it is difficult to robustly evaluate the service with the small number of events seen. Although high was predicted on five days it was only measured at sufficient monitoring sites to reach the episode criteria on one of these days. Isolated measurements of high air pollution or very widespread moderate occurred on each of these days.

Using measurements to trigger actions on the following day would be effective for many episodes of moderate air pollution and those times where high air pollution occurs as part of a longer polluted period. However, they would not be effective for triggering actions in a protocol that focused on high episodes only as these tend to occur on isolated days. This lack of persistence adds to the challenges of forecasting high and very high air pollution in London.

Commented [SM17]: Would it not work if you did moderate +. Because the high would be then included and not isolated.

Commented [FG18R17]: Yes, good point. I've reworded to include this. I hope it make sense.

The alerts issued as part of the Mayor's forecasting system therefore fit the purpose of providing suitable warnings to the public of adverse air pollution. High air pollution was measured on four days but not predicted on three of these. One of these days was predicted to be moderate and two were predicted to be low by all three forecast providers. This highlights the intrinsic difficulties in

predicting extremes of air pollution where small changes in weather cause marked changes in concentrations. This reflects the limitations of the current state of the art in air pollution forecasting and current limitations in inputs, including a lack of data on time varying emissions and uncertainties in weather forecasting.

The Mayor's current service is therefore meeting a requirement to provide precautionary advice to the public. However, if the forecasts are to be used to enact emergency or short-term measures, such as traffic restrictions or reduced cost public transport, a less precautionary service and one with a greater skill score would be required. One clear, simple step would be to trigger such actions based on the combination of severity of the episode and a probability, based on the degree of agreement between providers.

10. Acknowledgements

This work was funded by Transport for London and the Greater London Authority.

Heather Walton's post was part funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Health Impact of Environmental Hazards at King's College London in partnership with Public Health England (PHE) and Imperial College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health & Social Care or Public Health England.

11. References

Conlan, B., 2012. Best practices for short term action plans. AEA, Didcot.

Committee for the Medical Effects of Air Pollution (COMEAP), 2012. Review of the UK Air Quality Index. Department for Health, London.

De Nazelle, A., Seto, E., Donaire-Gonzalez, D., Mendez, M., Matamala, J., Nieuwenhuijsen, M.J. and Jerrett, M., 2013. Improving estimates of air pollution exposure through ubiquitous sensing technologies. Environmental Pollution, 176, 92-99.

Fuller, G. W, 2018. The Invisible Killer – the rising global threat of air pollution and how we can fight back. Melville House Books, London.

Honoré, C., Rouil, L., Vautard, R., Beekmann, M., Bessagnet, B., Dufour, A., Elichegaray, C., Flaud, J.M., Malherbe, L., Meleux, F. and Menut, L., 2008. Predictability of European air quality: Assessment of 3 years of operational forecasts and analyses by the PREV'AIR system. Journal of Geophysical Research: Atmospheres, 113(D4).

Murphy, A.H., 1996. The Finley affair: A signal event in the history of forecast verification. Weather and Forecasting, 11(1), 3-20.

Neal, L.S., Agnew, P., Moseley, S., Ordóñez, C., Savage, N.H. and Tilbee, M., 2014. Application of a statistical post-processing technique to a gridded, operational, air quality forecast. Atmospheric Environment, 98, 385-393.

Stedman, J.R., 1997. A UK-wide episode of elevated particle (PM10) concentration in March 1996. Atmospheric Environment, 31(15), 2381-2383.

12. Appendix A – Forecast accuracy skill scores

Accuracy		Days Forecast in advance				
Forecast	Confidence	On day	1	2	3	4
Low	Good	0.99	0.98	0.98	0.96	1.00
	Medium	0.99	0.99	0.99	1.00	-
	Poor	1.00	1.00	1.00	1.00	-
Moderate	Good	0.98	0.98	0.98	0.99	0.67
	Medium	0.96	0.97	0.96	0.99	1.00
	Poor	0.96	0.96	0.96	0.97	-
High	Good	1.00	-	-	-	-
	Medium	0.99	0.99	1.00	-	-
	Poor	0.99	1.00	1.00	-	-
Very high	Good	-	-	-	-	-
	Medium	-	-	-	-	-
	Poor	-	-	-	-	-

Table 14 Accuracy for background episodes. The table shows the air pollution band forecast and the confidence. Dashes denote no forecast for this criterion.

Accuracy		Days Forecast in advance				
Forecast	Confidence	On day	1	2	3	4
Low	Good	0.99	0.98	0.98	0.94	1.00
	Medium	0.99	0.99	0.99	1.00	-
	Poor	1.00	1.00	1.00	1.00	-
Moderate	Good	0.98	0.98	0.99	0.99	0.67
	Medium	0.97	0.97	0.97	0.99	1.00
	Poor	0.96	0.96	0.96	0.97	-
High	Good	1.00	-	-	-	-
	Medium	1.00	1.00	1.00	-	-
	Poor	0.99	1.00	1.00	-	-
Very high	Good	-	-	-	-	-
	Medium	-	-	-	-	-
	Poor	-	-	-	-	-

Table 15 Accuracy for background or roadside episodes. The table shows the air pollution band forecast and the confidence. Dashes denote no forecast for this criterion.

Number of forecasts		Days Forecast in advance				
Forecast	Confidence	On day	1	2	3	4
Low	Good	326	327	278	46	1
	Medium	25	26	22	6	0
	Poor	3	3	3	1	0
Moderate	Good	29	30	25	3	1
	Medium	37	35	32	7	1
	Poor	24	23	25	4	0
High	Good	1	0	0	0	0
	Medium	3	3	1	0	0
	Poor	4	2	1	0	0
Very high	Good	0	0	0	0	0
	Medium	0	0	0	0	0
	Poor	0	0	0	0	0

Table 16 Number of forecasts issued. The table shows the air pollution band forecast and the confidence. Blank cells denote no forecast for this criterion.

13. Appendix B - Number of days in each pollution banding (measurements)

The tables below show the number of days in each banding each year based on measurements during the forecast period analysed.

Table 17~~Table 16~~ shows the number of days in each banding when using the episode at background episode criteria, the maximum DAQI banding measured at one background site and one other site across the network.

Table 18~~Table 17~~ shows the number of days in each banding when using the widespread roadside criteria, as per the background criteria used in Table 17~~Table 1~~ OR was measured at 10% of road or kerbside monitoring sites

It should be noted that the overall column is not the sum of the individual pollutants in that row since on some days more than one pollutant may have been in that banding, e.g on an overall moderate day both NO_2 and PM_{10} may have been moderate, equally on a high day PM_{10} and $PM_{2.5}$ may have been high but NO_2 only moderate

	Overall	NO_2	O_3	PM_{10}	$PM_{2.5}$
Low Days	**389**	**452**	**413**	**431**	**426**
2018	286	332	297	325	320
2019	103	120	116	106	106
Moderate Days	**61**	**0**	**39**	**21**	**24**
2018	45	0	35	7	11
2019	16	0	4	14	13
High Days	**2**	**0**	**0**	**0**	**2**
2018	1	0	0	0	1
2019	1	0	0	0	1

Table 17- number of days in each banding based on one background site plus one other

	Overall	NO_2	O_3	PM_{10}	$PM_{2.5}$
Low Days	**382**	**450**	**413**	**422**	**419**
2018	283	331	297	319	316
2019	99	119	116	103	103
Moderate Days	**66**	**2**	**39**	**27**	**31**
2018	48	1	35	12	15
2019	18	1	4	15	16
High Days	**4**	**0**	**0**	**3**	**2**
2018	1	0	0	1	1
2019	3	0	0	2	1

Table 18 - number of days in each banding based on the background criteria or seen at ten percent of roadside or kerbside monitoring locations

14. Appendix C – London mean pollution concentrations on high days

The mean across all sites on the LAQN on days that were classified as high (no very high days occurred) during the forecast analysis period. This includes days high was measured according to the background or roadside episode criteria as detailed in Table 13~~Table 12~~.

	Daily PM_{10}	Daily $PM_{2.5}$	Max hour NO_2	Max Rolling 8 hour O_3
03-Mar-18	67	63	69	26
26-Feb-19	60	45	110	13
27-Feb-19	63	48	120	27
22-Apr-19	57	54	91	102

Table 19 - pollution concentrations on high days in ug/m^3

15. Appendix D – Persistence of pollution episode conditions

The conditions that result in moderate pollution occur far more frequently than those for high. As a result, several consecutive moderate days can occur whereas high days tend to be single isolated incidents. No very high days occurred during the analysis period.

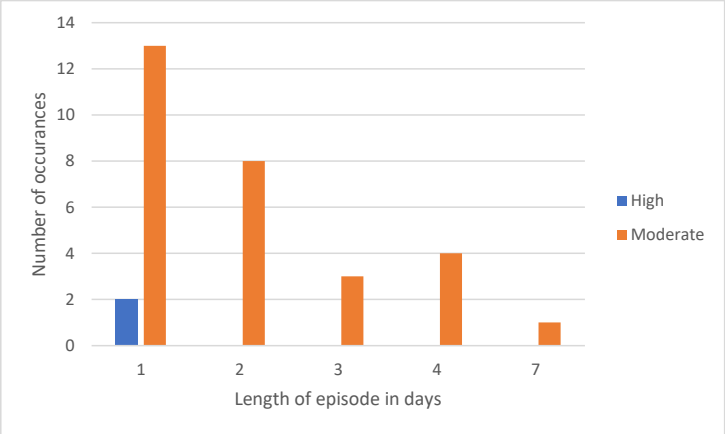

Figure 5 - occurrences of episodes using background criteria

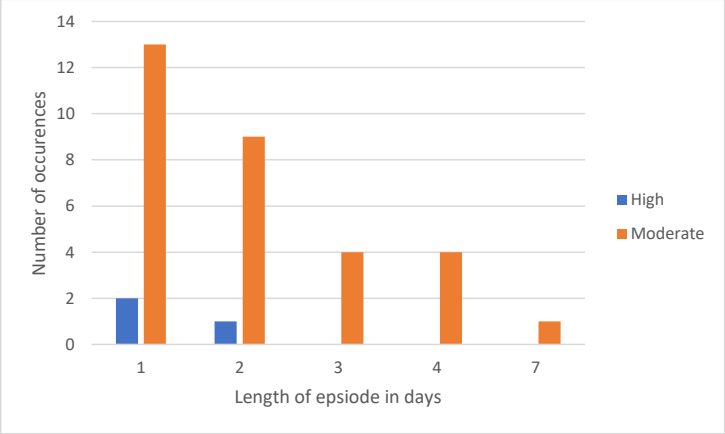

Figure 6 - occurrences of episodes using background or roadside criteria