2016-09-08

Short Term Forecasting of Nitrogen Dioxide (NO2) Levels Using a Hybrid Statistical and Air Mass History Modelling Approach

Aoife Donnelly
Technological University Dublin, aoife.donnelly@tudublin.ie

Owen Naughton
National University of Ireland, Galway, Naughto@tcd.ie

Brian Broderick
Trinity College Dublin, Ireland

See next page for additional authors

Recommended Citation

Donnelly, A., Naughton, O., Broderick, B. & Misslear, B. (2016). Short-Term Forecasting of Nitrogen Dioxide (NO2) Levels Using a Hybrid Statistical and Air Mass History Modelling Approach. *Environmental Modeling & Assessment*, vol.22(3), pp.231-241.

Funder: EPA Ireland

Authors
Aoife Donnelly, Owen Naughton, Brian Broderick, and Bruce Misstear

Short term forecasting of nitrogen dioxide (NO_2) levels using a hybrid statistical and air mass history modelling approach

Aoife Donnelly[1], Owen Naughton[2], Brian Broderick[3], Bruce Misstear[3]

[1] School of Food Science and Environmental Health, Dublin Institute of Technology, Dublin, Ireland

[2] Civil Engineering, School of Engineering and Informatics, University of Ireland, Galway,

[3] Department of Civil, Structural and Environmental Engineering, Museum Building, Trinity College Dublin, College Green, Dublin 2, Ireland

Abstract

A novel hybrid model has been developed to support the provision of real time air quality forecasts. Statistical techniques have been applied in parallel with air mass history modelling to provide an efficient and accurate forecasting system with the ability to identify high NO_2 events, which tend to be the episodes of most significance in Ireland. Air mass history modelling and k-means clustering are used to identify air mass types that lead to high NO_2 levels in Ireland. Trajectory matching techniques allow data associated with these air masses to be partitioned during model development. Nonparametric regression (NPR) has been applied to describe nonlinear variations in concentration levels with wind speed, direction and season and produce a set of linearized factors which, together with other meteorological variables, are employed as inputs to a multiple linear regression. The model uses an innovative integrated approach to combine the NPR with the air mass history modelling results. On validation, a correlation coefficient of 0.75 was obtained and 91% of daily maximum (hourly averaged) NO_2 predictions were within a factor of two of the measured value. High pollution events were well captured, as indicated by strong agreement between measured and modelled high percentile values. The model requires only simple input data, does not require an emission inventory and utilises very low computational resources. It represents an accurate and efficient means of producing real time air quality forecasts and, when used in combination with forecaster experience, is a useful tool for identifying periods of poor air quality 24 hours in advance. The hybrid approach outlined in this paper can easily be applied to produce high quality forecasts of both NO_2 and additional pollutants at new locations/countries where historical monitoring data are available.

Highlights

- A novel hybrid model is presented for quick and efficient air quality forecasts
- Combines parametric and non-parametric regression with air mass history modelling
- 24 hour forecasts of daily maximum NO_2 are produced
- Model validation produced an r value of 0.75 and a FAC2 value of 0.91.

Keywords: Air quality forecasting, air mass history modelling, nonparametric regression, air pollution, NO2

1. Introduction

Across Europe NO_2 has been one of the most problematic air pollutants whilst in Ireland, a trend of increasing NO_2 has been identified at traffic-impacted sites in the major cities [1]. Air quality modelling and short term forecasting of NO_2 could reduce the risk of exceeding EU limit values through adoption of air quality action plans and also provide important information to the public in advance of a poor air quality episodes. The diverse range of scientific and regulatory applications such as forecasting, public information provision, compliance assessment and air quality management in which air quality modelling is employed, is reflected in the multiplicity of modelling approaches used. An air quality model can be conceptual, empirical or process oriented, with each approach demanding varying levels of technical, scientific and computational resources. The more physical processes that are included in the model, the more comprehensively it will generally be able to describe reality. However, increasing the inputs and model processes leads to high demands on the quantity and quality of information needed to drive the models [2]. The complexity of atmospheric chemical and transport processes, as well inaccuracies in emission estimations, means that significant approximations and uncertainties exist within deterministic models. As a result it is often argued that chemical transport models (CTMs) are currently less accurate than well-developed, site-specific empirical air quality forecast models trained with local air quality and meteorological data [3, 4]. Furthermore, the use of large-scale deterministic models as, for example, operational air quality forecast systems is frequently beyond the reach of many national authorities. Instead, statistical modelling has been found by many countries to offer a viable and attractive alternative [5-7].

Statistical models are built on existing links between pollutant concentrations, meteorological parameters (e.g. wind speed and direction, temperature, air pressure, rainfall) (eg. [8]) and physical parameters (e.g. emissions, land use, road density). They tend to be more suitable for the description of complex site specific variations and generally have a higher accuracy than deterministic models [9]. Numerous studies have used statistical techniques to develop air quality forecasts. Techniques adopted include multiple linear regression ([10, 11]), ARIMA modelling [4, 12], neural networks [13-15], nonlinear regression [16, 17], Kalman filtering [18] and various combinations of these [4, 14]. Other integrated or hybrid approaches have also previously been adopted (e.g. [9, 19, 20]). However, most of these methods suffer from the disadvantage that they cannot capture the contribution of distant weather-dependent sources and regional air mass movement since the local forcing variables often do not account for regional scale weather variations and transboundary air pollution. Air mass history modelling has successfully been used in previous studies to address this limitation. High pollution events have been assessed by calculating a single back trajectory ending on the day on which concentrations were high ([21]). Vukmirovic *et al.* [22] used the ETA model to generate forward trajectories to identify the paths and evaluate removal mechanisms following bombings at industrial sites in northern Serbia and the Belgrade region during the Kosovo war. However, such methods give only a rough approximation of the actual air mass history, having uncertainties associated with the resolution and accuracy of the meteorological data [23, 24].

Anastassopoulos *et al.* [25] characterised HYSPLIT (HYbrid SingleParticle Lagrangian Integrated Trajectory) back trajectories by air mass path direction and regions traversed. In conjunction with measured NO_x and SO_x data, the regions which most frequently influenced air quality in Windsor, Ontario, Canada were identified. Riccio *et al*. [26] used a combination of *k*-means clustering and PCA (Principal Component Analysis) approaches of HYSPLIT-4 back trajectory classification to assess the influence of large-scale meteorological conditions on ozone and PM_{10} concentrations in Naples, Italy. They found that ozone and PM_{10} exhibited increased concentrations during anti-cyclonic, subsiding conditions due to stagnation and recirculation effects.

In Ireland, a trajectory sector frequency analysis and subsequent vector addition was used to define the path of the air mass and identify the effects of regional air mass movement on air quality [27]. Recent work by the authors used k-means clustering techniques on air mass back trajectories to identify air mass movement corridors which lead to poor air quality [28]. Average NO_2 levels were found to vary by 124% and 239% of the seasonal mean between clusters.

The specific aim of this work was to produce a fast, non-resource intensive and reliable means of forecasting daily maximum hourly averaged NO_2 concentrations. This paper describes a novel hybrid model which combines air mass history modelling results with a statistical forecasting model previously developed by the authors [17] to predict concentrations of the priority pollutant NO_2 in Ireland 24 hours in advance. The statistical forecasting model previously developed by the authors, while producing good results with regards to mean concentration variations, was limited in its ability to capture high pollution events because it could not describe the effects on concentrations due to distance weather depend factors and regional air mass characteristics. The innovative air quality forecasting methodology described in this paper brings together and enhances two major areas of work in a novel manner to provide an elegant solution for capturing high NO_2 events and routinely forecasting air quality levels. The low data and computational resource requirements needed to implement this model, together with the open-access nature of its components, make it well suited to providing fast and reliable real-time air quality forecasts in regulatory environments where resources are limited.

2. Methods

2.1 Overview

The air quality forecasting methodology described in this paper is a two-step process comprising a model development phase followed by an operational (or model validation) phase. Air mass history modelling is used during both phases to identify periods of high and low contributions from background regional and transboundary pollutant sources. It is employed first as a data partitioning tool during model development; multiple statistical models representing high and low regional background concentrations are developed using the partitioned data and regression techniques described in Donnelly *et al.* [17]. Trajectory forecasting during the hybrid model operation is then applied to select the appropriate high or low background model.

Two monitoring sites are defined in the process: the background reference site and the local site. The background site is not influenced by any major local sources in the area and is considered to be representative of regional air pollution. The local site is the location for which the forecast is required, e.g. an urban area, and is influenced by both local and regional air pollution.

During model development air mass history modelling is used to identify and quantify trajectories which lead to higher concentrations at a background site. For example in Ireland, air masses originating over the United Kingdom and mainland Europe contribute to higher background NO_2 concentrations than do the relatively clean air masses from the Atlantic Ocean. Back-trajectories are calculated for each time step within the monitoring data time series from the background site. Trajectory cluster analysis is employed to group trajectories based on their three-dimensional similarities and identify the primary meteorological pathways influencing the site. The objective here is to identify groups or clusters of air masses which contribute to high background NO_2 concentrations in the region of interest; thus the output is a set of trajectory clusters which are assigned into the “High” or “Low” background concentration group.

It is assumed that air masses contributing to higher concentrations at the background reference site will produce comparable increases in concentrations at the local site. Monitoring time series at the background and local sites are then partitioned into high and low regional pollutant concentrations based on the cluster arriving at the background site at each time step. Statistical regression models are then developed representing different regional background conditions at both sites.

The operational phase (the air quality forecast) is a two-step process involving a trajectory forecast coupled with a series of statistical regression models. A schematic diagram of the hybrid model is displayed in Figure 1. Firstly, a set of trajectory forecasts are computed using HYSPLIT to predict if the air mass arriving at the background site over the coming forecasting period belongs to a low- or high-pollutant concentration cluster. Separate statistical forecast models are then invoked specifically for low and high background conditions. The forecast model itself comprises three regression sub-models, all of which are based on the same multiple linear regression methods described in Donnelly *et al.* [17] but trained on different datasets derived during the preliminary data partitioning phase. The model development, calibration and validation are described in the following sections, and it is helpful to read this in the context of the model decision tree, which is illustrated in Figure 1. In this paper the model has been developed using a rural background site (Kilkitt) with results obtained on high cluster days added to basic predictions at a representative urban site (Rathmines). Combining the two models allows local and regional effects to be accounted for individually.

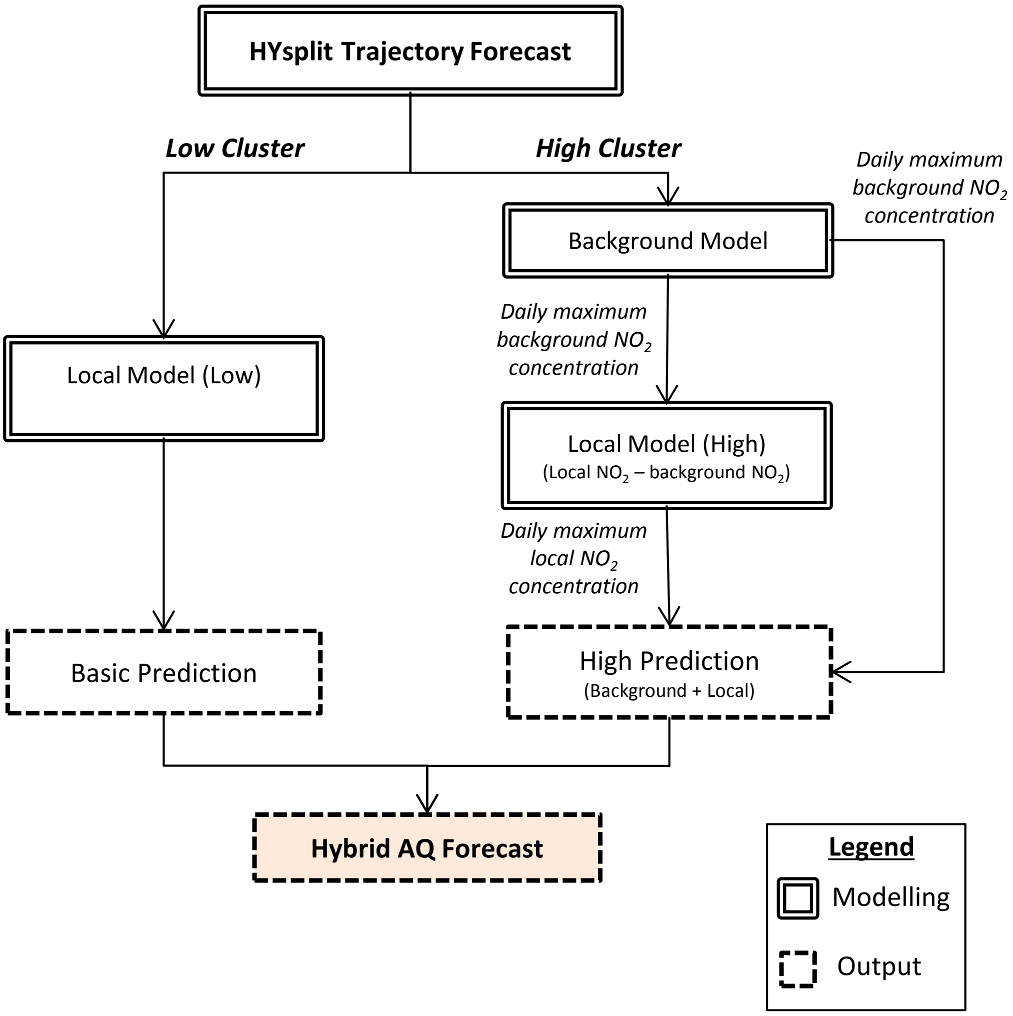
HYsplit Trajectory Forecast
Low Cluster
High Cluster
Daily maximum background NO_2 concentration
Background Model
Daily maximum background NO_2 concentration
Local Model (Low)
Local Model (High)
(Local NO_2 – background NO_2)
Daily maximum local NO_2 concentration
Basic Prediction
High Prediction
(Background + Local)
Hybrid AQ Forecast
Legend
Modelling
Output

Figure 1 Hybrid model operational decision tree

2.2 Data

The data used in the research have been collected by the Irish Environmental Protection Agency (EPA), the agency responsible for maintaining the air quality monitoring network in Ireland. Kilkitt is a remotely located monitoring site in an agricultural region. This site was chosen as the background site for air mass history modelling as there are no major local emission sources in the area and hence it is considered to be representative of regional air pollution. The Rathmines monitoring site is used in the validation of the modelling technique. It is classified as an urban background site and located in a residential and commercial area to the south of Dublin City Centre. This site is influenced by both local sources such as roads and domestic burning and distant transboundary sources. Data from 2011 and 2012 were used in model development.

NO_2 is monitored and recorded at hourly resolution at each site using chemiluminesence samplers (as recommended for demonstration of compliance with EU limit values). NO_2 is measured indirectly by reduction to NO by means of a molybdenum converter. The instrument has a range of 0 to 20,000ppb and a lower detection limit of 0.4ppb with a precision of 0.5% of the reading. Average concentrations at all sites are well within an appropriate range for the instrument used.

2.3 Back-Trajectory Analysis

The HYSPLIT model was developed by the U.S. National Oceanic and Atmospheric Administration (NOAA)'s Air Resources Laboratory (ARL) and combines the Eulerian and Lagrangian approaches to track air mass movement [29]. Whereas the model is capable of calculating concentrations of pollutants, it is applied in this study to calculate back trajectories. With the HYSPLIT model, air mass paths from one region to another can be calculated and it can therefore be demonstrated whether or not the vector necessary for air pollutant transport is present [25]. When the model is run in back trajectory mode, the movement of a parcel of air can be calculated backwards in time from the receptor where concentrations were measured, allowing the origin of the pollution to be identified.

Forty-eight hour trajectories were calculated for two full calendar years (2011 and 2012) with hourly end-points located at the background monitoring station. These trajectories were divided into seasonal groups. The choice of trajectory duration is important because too short a duration may miss the actual source of the emissions and important path crossings while too long a run induces a large amount of uncertainty into the analysis and may produce misleading results. Previous work by the authors showed that clustering of 48-hour trajectories could usefully partition data into high and low pollution conditions [28] and this trajectory duration was adopted here. As an island with no nearby land mass to the west and south west, and significant nearby land mass to the east and south, the appropriate trajectory duration may differ in Ireland than in land-locked countries. A simple analysis of air masses and clusters should reveal the appropriate trajectory duration for a given country.

2.4 Trajectory Analysis: k-means clustering

Trajectory cluster analysis was employed to group trajectories based on their three-dimensional similarities and to identify the primary meteorological pathways influencing the background site. This technique groups similar trajectories together, with the aim of minimising differences within clusters and of maximizing the differences between clusters. It allows for the inclusion of re-circulated trajectories and trajectories with rapidly varying directionality. The hierarchical cluster method adopted in this paper initially assumes that the number of clusters is equal to the total number of trajectories (N) and thus the spatial variance (SV) (the sum of the squared distances between end points of the clusters component trajectories and the mean of that cluster) is zero. In the first iteration, each combination of trajectory pairs is tested to compute the cluster SV. The total spatial variance (TSV) is then calculated by summing all of the clusters SV's. The two trajectories with the lowest SV are combined into a single cluster, thus reducing the total number of clusters after the first iteration to N-1. Once paired, clusters remain together in subsequent iterations. In the second iteration, the clusters are either individual trajectories or the cluster of the initial pairing of trajectories. Again, every combination is assessed and the two clusters combined are those that result in the lowest increase in TSV. The iterations continue in this manner until the last two clusters are combined resulting in all N trajectories in one cluster. In the first number of iterations the TSV increases greatly as the number of clusters combined increases. Thereafter, it tends to increase gradually up to a point beyond which it increases sharply, indicating that the clusters being combined are not very similar. A plot of TSV against the number of clusters will clearly indicate this change and suggest where clustering should be stopped as shown in Figure 2.

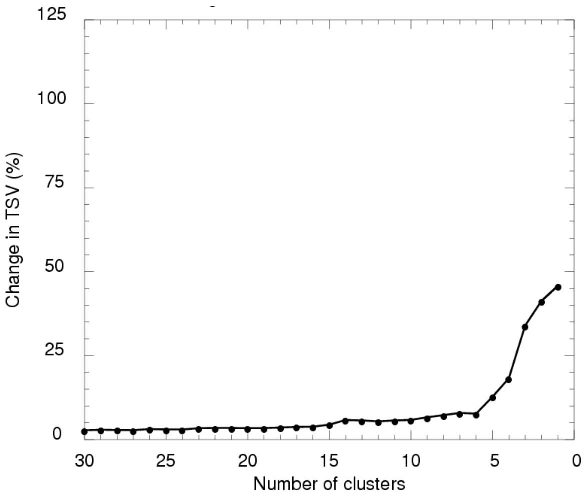

Figure 2 TSV plot showing the appropriate number of clusters for the Jan-Mar period as 6

The trajectories were first separated into four distinct seasons to account for known variability in both synoptic scale variations and air pollution levels between the winter (January – March), spring (April-June), summer (July – September) and autumn (October – December) periods. Clustering was carried out on each data set individually and the optimum number of clusters was chosen in each case by visual inspection of the TSV plots.

2.5 Data partitioning and model development

After clustering, the variability in hourly NO_2 between clusters was determined and an analysis of variance (ANOVA) technique was applied to assess which cluster types led to increased concentrations. Using these results, date/hours from the 2011-2012 time period were partitioned into “high” and “low” groups at both the background and local sites. Two unique local forecast models were then developed for each background condition using the techniques outlined in section 2.6:

Low Cluster Model: This model for “low” conditions was trained using non-partitioned raw measured data from urban site (Rathmines). It provides forecasts at the urban site under “low” background conditions.

High Cluster Model: This comprises two interrelated regression equations to forecast the background and local contributions to air quality. The first, the “high” background model, is trained using partitioned “high” data from the background site. The second, the “high” local model, is trained using partitioned “high” data from the local site **minus** the corresponding background concentration. The forecast concentration at the urban site on “high” days is thus the concentration computed by the “high” background model plus the concentration computed by the “high” local model.

2.6 Parametric and non-parametric regression

A previous publication describes in detail the statistical modelling approach [17] so only a summary of the key points will be presented here.

The general form of the model is:

$$C = b_0 + \sum_{i=1}^{n} b_i\, x_i + \sum_{i=1}^{m} d_i\, y_i + \varepsilon$$

where C is the response variable (Daily maximum hourly NO_2 concentration), b_0 is the regression constant, x_i are meteorological predictor variables with coefficients b_i, and y_i are predictor variables output from the non-parametric and time series models with coefficients d_i. The parameter ε is the stochastic error associated with the regression.

A least squares technique was employed to determine the coefficients for each of the predictor variables, as shown in Table 1.

Table 1 Description of model coefficients

Variable Category	Predictor Variable	Description
Model (Y$_i$)	$WSWD_f$	Wind speed, wind direction factor
	S_f	Non-parametric seasonal factor
	D_f	Non-parametric diurnal factor
Meteorological (X$_i$)	Sat	Dummy variable: 1 if Saturday, 0 otherwise
	Sun	Dummy variable: 1 if Sunday, 0 otherwise
	$Temp$	Hourly Temperature
	$SunHr$	Sunshine Hours
	$RelHum$	Relative Humidity
	$AtmPres$	Atmospheric Pressure
	$StabilityCl$	Stability Class
	NO_{2h-24} NO_{2h-48}	Daily average NO$_2$ concentration at 24/48 hour lag
	$NO_{2max-24}$ $NO_{2max-48}$	Daily maximum NO$_2$ concentration at 24/48 hour lag
	O_{3d-24} O_{3d-48}	Daily average O$_3$ concentration at 24/48 hour lag
	$O_{3min-24}$ $O_{3min-48}$	Daily minimum O$_3$ concentration at 24/48 hour lag

The predictor variables $WSWD_f, S_f$ and D_f were developed using non-parametric kernel regression (as described in [17, 30, 31, 32 and 34].

A stepwise process was used to isolate the significant explanatory variables during model development. Predictor variables were examined for co-linearity as multicolinearity increases the standard errors of the coefficients, meaning that coefficients for some independent variables may be erroneously found to be significant. The variance inflation factor (VIF) was used to assess how much the variance of an estimated regression coefficient increases if predictors are correlated; it is equal to 1 if no factors are correlated [33]. Variables with high VIF were removed from the model in a stepwise manner ensuring that each variable removed is redundant in the explanation of NO$_2$ concentration and assessing the correlation coefficient (r) value at each pass Definition of the final model also relied on knowledge of the directions of influence for particular parameters and graphical techniques.

Various statistical performance methods have differing strengths and weaknesses and therefore in this study, as discussed in [34] and following a similar model development and validation procedure as presented in [17] a number of methods have been applied to assess model performance: r (Correlation coefficient), IA (Index of Agreement), the fraction of predictions within a factor of two of observations (FAC2), the fractional bias (FB) and the standard error (s). In addition to these statistical metrics, the accuracy of the proposed prognostic model to predict the "exceedances" days has been assessed by means of the hit rate (H) or the probability of detection (this indicates the percentage of actual exceedances that are correctly forecast) [35]. Since the concentrations in the area assessed all fall below the EU hourly limit value, a more stringent value of the annual mean limit value of 40µg/m^3 has been used for this test. In order to assess the likelihood of false alarms the False Positive Rate (FPR) has also been applied (using the same theoretical exceedance rate.

3. Results

3.1 Model development

3.1.1 Air mass clustering

The HYSPLIT model was run for two full calendar years (2011 and 2012) with the end point at Kilkitt. Results of the cluster analysis are shown in Figure 3. Six clusters were defined from January to March (Figure 3(A)). These include a slow moving east/south easterly cluster and a moderate moving easterly cluster. These two clusters are associated with the highest NO_2 concentrations, (averaging 196% and 168% of the mean for this time period). From April to June only one easterly cluster is defined and NO_2 concentrations for these air masses average 161% of the mean for the time period (Figure 3(B)). A similar result is observed between July and September where concentrations for the easterly cluster average 191% of the mean for the period (Figure 3(C)). The easterly cluster results in average concentrations of 196% between October and December (Figure 3(D)). During this time period Ireland is also frequently affected by slow moving northerly air masses representative of cold winter weather conditions. The defined cluster is of much shorter length in this season than in other seasons and its slow moving nature and its land track over parts of the UK result in average NO_2 concentrations of 153% of the mean for the time period.

Average concentrations for each cluster are displayed in Table 2. Shaded clusters illustrated in the table represent the clusters which were identified by the ANOVA test as significantly different and used to partition the data into “high” and “low” sets.

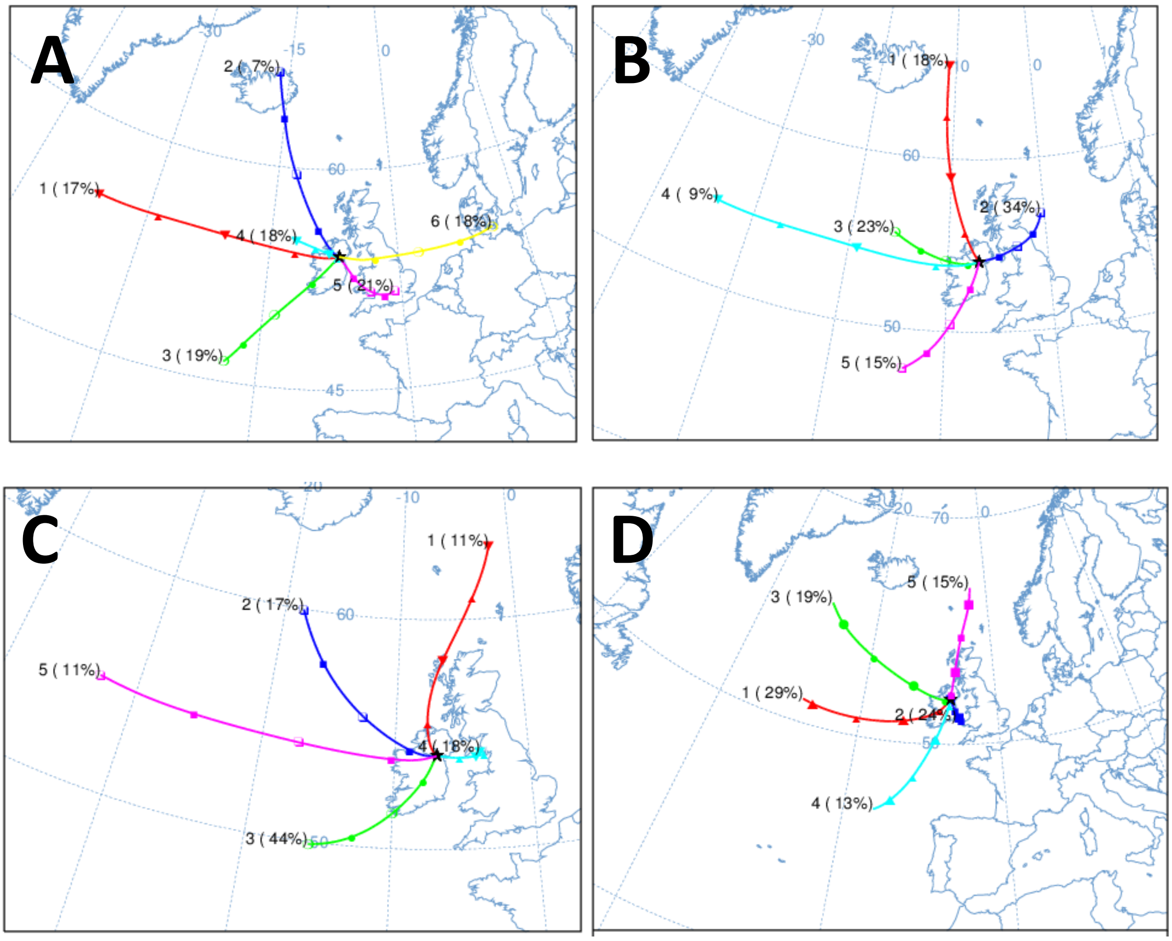

Figure 3 Cluster results from Kilkitt for 2012 and 2013 data Jan-Mar (A), Apr-Jun (B), Jul-Sep (C) and Oct-Dec (D)

Table 2 NO_2 concentrations for air mass clusters arriving at Kilkitt

Jan-Mar			Apr-Jun			Jul-Sep			Oct-Dec		
No.	Average NO_2 ($\mu g/m^3$)	% of mean	No.	Average NO_2 ($\mu g/m^3$)	% of mean	No.	Average NO_2 ($\mu g/m^3$)	% of mean	No.	Average NO_2 ($\mu g/m^3$)	% of mean
5	4.29	1.96	2	3.38	1.61	4	2.65	1.96	2	2.85	1.91
6	3.68	1.68	3	2.07	0.99	1	1.23	0.91	5	2.29	1.53
4	2.46	1.12	5	1.37	0.65	3	1.09	0.81	4	0.88	0.59
2	1.58	0.72	1	1.11	0.53	5	0.93	0.69	1	0.78	0.52
1	0.34	0.15	4	0.79	0.38	2	0.92	0.68	3	0.68	0.45
3	0.09	0.04	-	-	-	-	-	-	-	-	-

3.1.2 Model fitting using partitioned data

Using the partitioned "high" data, the background model was fitted using a least squares regression. Exploratory analysis indicated that the predictor variables which should be considered for this regression analysis were as follows: two dummy variables indicating if it was a Saturday or a Sunday, 24-hour average temperature, $WSWD_f$ and the 24-hour lagged NO_2 concentration. (Note that this will allow 24-hour forecasts to be made. In the event that 48 hour forecasts are to be made the 48-hour lagged NO_2 concentration would be used. The authors have also tested the model performance over this time period but these results are not included here for the purpose of brevity.) An analysis of residuals (actual value minus predicted value) was carried out at each pass of the regression to determine compliance of the model with the assumptions of an ordinary least squares regression. The objective in model fitting was to maximise the correlation coefficient (r value) while minimising assumption-based errors. Residuals were found to depart from a Gaussian probability distribution and therefore a square root transformation was applied to the dependent variable and to the lagged variable. During summer months there was less of a distinction between weekday and weekend concentrations at the background site which was indicated by high VIF and r-values for the dummy variables. To minimise autocorrelation of the errors and improve parameter estimation, these parameters were not used in the Jul-Sep model and only the Sunday dummy variable was used in the Apr-Jun model. The complete dataset of modelled NO_2 was plotted against observed NO_2 as shown in Figure 4 (A). A straight line fitted to the data indicates the intercept is very close to 0 and the slope of the line is close to 1, suggesting that the bias in the model is very low. An r value of 0.74 was obtained.

Again using the partitioned "high" data, the model was fitted to the Rathmines data (background subtracted). Exploratory analysis indicated that the variables which should be considered for the *local model* were relative humidity, air pressure, sunshine duration, $WSWD_f$, S_f, D_f, and NO_{2-24}. Residual analysis indicated departure from normality and a square root transformation was applied to the dependent variable and NO_{2-24}. The complete dataset of modelled NO_2 was plotted against observed NO_2 as shown in Figure 4 (B). The non-zero intercept and slope was less than one indicating some small systematic bias in the model for NO_2 concentrations above 50ppb. Following Hubbard and Cobourn [36], to eliminate this bias a second stage correction was applied to the model using a polynomial curve fit such that:

$$\hat{C}_{adj} = a_0 + a_1\hat{C} + a_1\hat{C}^2$$

where $\hat{C}_{adj}$ is the new adjusted forecast, $\hat{C}$ is the unadjusted forecast and the a terms are regression coefficients. Figure 4 (B) shows the adjusted scatter plots. The latter model contains no bias and the plot illustrates an intercept of zero and a line with slope equal to 1. An r value of 0.764 was obtained (p-value <0.05).

Finally, a model was developed using the non-partitioned raw data from Rathmines (total concentrations for "low" conditions) and measured versus modelled values are plotted in Figure 5.

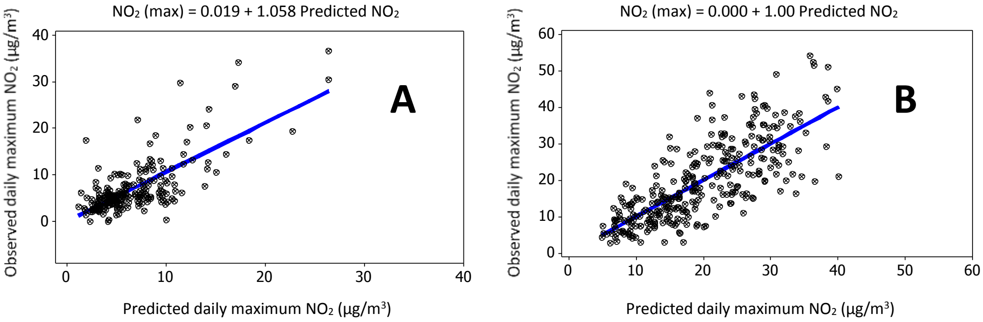

Figure 4 Model development at background site (A) and local source adjusted site (B) for partitioned "high" data

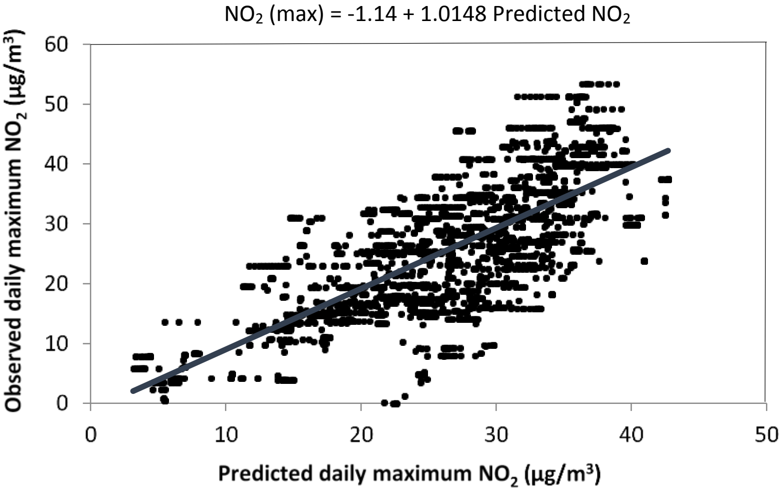

Figure 5 Model development for non-partitioned raw data at Rathmines for "low" conditions

3.2 Model validation

For validation purposes, the hybrid modelling system was run operationally for 2014 at Rathmines using the decision tree outlined in Figure 1 (note that these data were not used during model development) to produce a year of 24 hour forecasts of daily maximum hourly NO_2 concentrations. In addition, forecasts were produced for the same time period using the statistical model alone, i.e. air mass history modelling and data partitioning were not invoked. A plot of the two sets of predictions against observed values is shown in Figure 6. There is a clear improvement in predictions using the hybrid system. In particular, the model performs better in the prediction of high pollution events.

Model performance statistics for the validation of the hybrid model are shown in Table 3. Based on the validation data the model has a coefficient of determination (R^2) of 0.52 and an r value of 0.752 is obtained. This value is similar to that observed during model fitting indicating a similar explanation of variation by the model in extrapolated data. The FAC2 value shows that 91% of the predictions fall within a factor of two of the observed values and an IA value of 0.81 is achieved which was similar to that obtained by the basic model [17]. A standard error (s) for the regression model of 7.3 µg/m^3 was obtained.

A hit rate of 0.63 is achieved using the stringent EU annual mean limit value of 40µg/m^3 as the theoretical exceedence rate. While the standard model performed well in predicted overall concentration variations, it did not capture very high pollution events and was unable to predict any of these "exceedence" days. Therefore, this new hybrid model presents a significant improvement in this regard. This improvement has been achieved by incorporating air mass history as a predictor term. This improvement in the prediction of high pollution events also leads to the false positive rate increasing slightly. The hybrid model showed a false positive rate (FPR) of 0.16 in this validation data which is an increase on the standard model which had a false positive rate of just 0.02. The magnitude of these false positives is small however, with an average over exceedance of 12 µg/m^3 on false positive days. As a result there is some positive bias in the hybrid model overall but since it is only invoked on high cluster days this is not considered problematic. This slight bias is reflected in the overall mean predictions as shown in Table 4. While the hybrid model over predicts the overall mean (119%) its performance far exceeds the basic model in the prediction of the high percentile values in each season and the annual maximum concentration. The hybrid model predicts the 95th and 98th percentile of measured daily maximum concentrations with an impressive degree of accuracy (102.5% and 101.1% of actual respectively). This prediction of the peak values by the model is important and while there is sometimes a slight positive bias in results the model statistics have shown that over-predictions are small and no significant false alarms were produced.

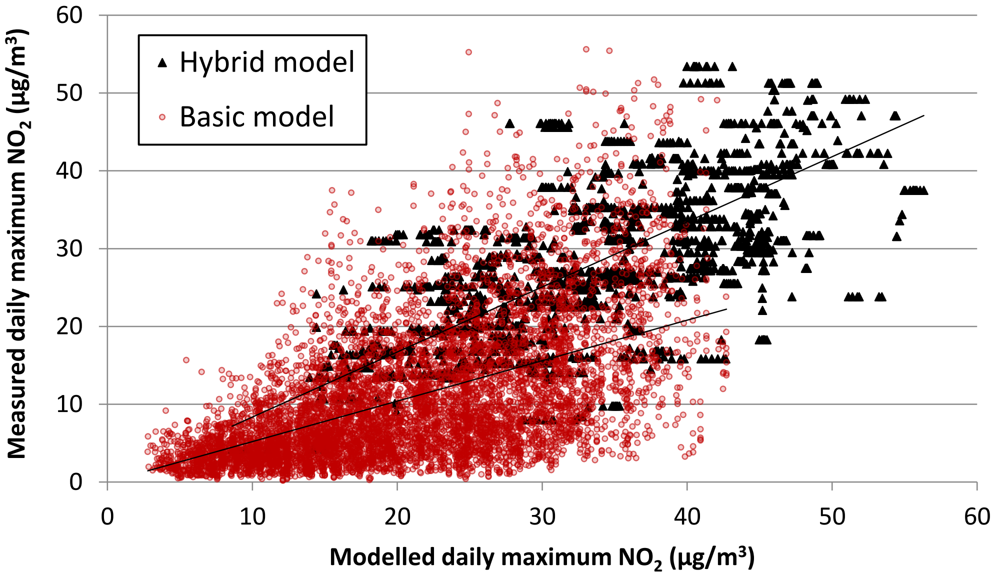

Figure 6 Observed versus modelled daily maximum hourly NO$_2$ concentrations (24 hour maximums) at Rathmines

Table 3 Model validation statistics

IA	r	FAC2	S	FB	NMSE	H	FPR
0.81	0.752 (p-value<0.05)	0.91	7.30	0.17	0.11	0.63	0.16

Table 4 Model performance for the basic model and hybrid model

	Measured	Basic model	Hybrid model
Mean (µg/m^3)	25.70	26.44	30.62
Percentage of mean	100	103	119
Maximum (µg/m^3)	53.43	42.72	56.33
Percentage of maximum	100	80	105
95th Percentile (µg/m^3)	45.60	37.21	46.75
% of measured	100	81.6	102.5
98th Percentile (µg/m^3)	49.19	39.35	49.75

% of measured	100	90.0	101.1

4 Discussion

Real time air quality forecasting has become an area of much interest in recent years and various deterministic and statistical techniques have been used to produce functional forecasts. While deterministic models can account for air parcel history, they are computationally intensive, require detailed emissions inventories over the modelled domain and have a high operational cost [4] which can make them unsuitable for real time air quality forecasting in many situations. Furthermore, many applications of real time air quality forecasting only require predictions at certain locations and in such instances, the processing required by deterministic models to provide detailed spatial variations may be a beyond the practicalities of available resources. As noted by Zhang *et al*. [4] statistical models often have a better capability for describing complex site specific variations in concentrations than deterministic models, often with a higher accuracy than deterministic models.

The hybrid model presented in this paper represents a novel way to forecast air quality routinely and accurately with minimal resource requirements. Previous work by the authors involved the development of a statistical forecast model and produced good predictions of average daily maximum NO_2 at urban, suburban and rural sites. By its nature, that model had a smoothing effect since it used median values to compute the seasonal indices and non-parametric regression to describe variation with wind speed and direction. Since local meteorological parameters were used as forcing variables it could not fully account for unusually high or low events caused by external factors such as long range transport of pollution or regional air mass movements. The hybrid model described in this paper incorporates the advantages of the statistical model outlined in Donnelly *et al*. [17] which include its low bias, its ability to forecast cyclical and anthropogenic effects without the need for an emissions inventory, its speed of computation and ease of operation and combines it with the (open-source) deterministic HYSPLIT model. This allows regional effects to be included in the forecasts without the need for a complex deterministic (and computationally-demanding) air quality model to be used. The model performance metrics presented in section 3 show that this new hybrid model offers a significant improvement over the standard non parametric statistical model developed by the authors [17]. This was particularly relevant in the high percentile levels and also in the model hit rate in predicting exceedences.

A key underlying assumption in this hybrid approach is that the transboundary contribution to air quality at the background site is representative of that at the forecasting site. The geographic location, prevailing climatic conditions and relatively low urbanisation characteristic of Ireland make this a reasonable assumption to make in the case presented here. However, when applying the method in areas influenced by heavy urbanisation, industrialisation or more complex regional air mass transport, care should be taken in the selection of the appropriate background site to ensure representativeness. Multiple background sites may thus be required when applying the model across a national monitoring network, with parallel trajectory forecasts necessary to enable model selection and forecasting. Due to the low computational resources of the statistical model and the ease with which trajectory forecasts can be produced this does not represent a substantial increase in resource requirements and so this approach remains a viable option for producing fast and reliable real-time air quality forecasts.

A typical approach to modelling is often to try a number of different models and select that which gives the most accurate result. However, the risk in doing so is that the final selected model may not be the most applicable under future conditions. The hybrid model combines different methods to reduce this model selection error and account for the fact that relationships between variables are rarely purely linear or nonlinear. As a result the impact of the error due to any single assumption is decreased. Furthermore, as with all forecasting models, periodic retraining is required throughout its operational cycle to ensure national and international trends in emissions are accurately represented within the forecasting process. The hybrid model retains the ability of the statistical techniques to describe complex site specific variations while including the effects of regional weather patterns through the use of air mass history analysis. Simplicity of the required input data is retained. No emissions data are required, since the model infers the influence of emissions from historical variations in concentration values. The authors' recommend that based on the results presented in this paper this model should be employed in cases where fast, efficient and accurate forecasts of daily maximum concentrations are required, particularly in cases where computational resources are limited. In other cases the model could also be used as a screening method to flag a requirement to run a more intensive deterministic modelling system in the event that more detail on forecast concentrations was required. The method has been shown to provide a useful means of forecasting high pollution events without bringing a permanent positive bias into the model results.

5 Conclusion

A model that combines air mass history modelling and parametric and non-parametric statistical modelling techniques to provide real-time daily maximum forecasts of NO_2 has been presented. The model uses a stepwise decision making process to produce 24 hour forecasts of daily maximum NO_2 concentrations at background and urban locations. A regional component is included in the model to better capture high pollution events through the use of the HYSPLIT model to assess air mass history. The model produces excellent prediction of the 95th and 98th percentile of daily maximum NO_2 concentrations out to 24 hours. On validation a correlation coefficient of 0.75 was obtained and 91% of values were within a factor of two of the measured value.Applied individually each technique is not sufficient to provide useful daily forecasts of peak concentration events but the novel combination of air mass history modelling and parametric and non-parametric techniques allows a large degree of temporal variation in concentration levels to be explained and forecast. The approach is therefore concerned with both natural and anthropogenic emissions, their transport through the environment and resulting pollutant concentrations. High pollution days are typically the most challenging to forecast and yet remain the most important to identify and this paper present a simple solution to predicting this high events where the risk of over prediction is also minimised. The simplicity of the input data together with very low computational resource requirements and minimisation fo assumption based errors make this model ideally suited to providing fast and reliable real time air quality forecasts at locations where monitoring data are available. This paper demonstrates model development for the purpose of forecasting NO_2 but the method is applicable to other priority pollutants including $PM_{10/2.5}$ and ozone. The model has been tested at a suburban monitoring site in Dublin, Ireland but the techniques are internationally applicable.

References

1. Environmental Protection Agency, *Irelands Environment - An assessment*. 2012: Johnstown Castle.
2. European Environment Agency (EEA), *The application of models under the European Union's Air Quality Directive: A technical reference guide*. 2011.
3. Cobourn, W.G., *An enhanced PM< sub> 2.5</sub> air quality forecast model based on nonlinear regression and back-trajectory concentrations.* Atmospheric Environment, 2010. **44**(25): p. 3015-3023.
4. Zhang, Y., et al., *Real-time air quality forecasting, part I: History, techniques, and current status.* Atmospheric Environment, 2012. **60**: p. 632-655.
5. Lissens, G., C. Mensink, and G. Dumont, *SMOGSTOP: A new way of forecasting ozone concentrations at ground level.* International Journal of Environment and Pollution, 2000. **14**(1): p. 418-424.
6. Chaloulakou, A., et al., *Measurements of PM10 and PM2. 5 particle concentrations in Athens, Greece.* Atmospheric Environment, 2003. **37**(5): p. 649-660.
7. Cobourn, W.G., *Accuracy and reliability of an automated air quality forecast system for ozone in seven Kentucky metropolitan areas.* Atmospheric Environment, 2007. **41**(28): p. 5863-5875.
8. Ragosta, M., M. D'Emilio, and G. Giorgio, *Input strategy analysis for an air quality data modelling procedure at a local scale based on neural network.* Environmental monitoring and assessment, 2015. **187**(5): p. 1-8.

9. Zhang, Y., et al., *Real-time air quality forecasting, part I: History, techniques, and current status.* Atmospheric Environment, 2012.

10. Genc, D.D., C. Yesilyurt, and G. Tuncel, *Air pollution forecasting in Ankara, Turkey using air pollution index and its relation to assimilative capacity of the atmosphere.* Environmental monitoring and assessment, 2010. **166**(1-4): p. 11-27.

11. Vlachogianni, A., et al., *Evaluation of a multiple regression model for the forecasting of the concentrations of NO x and PM 10 in Athens and Helsinki.* Science of the total environment, 2011. **409**(8): p. 1559-1571.

12. Kumar, U. and V. Jain, *ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO).* Stochastic Environmental Research and Risk Assessment, 2010. **24**(5): p. 751-760.

13. Moustris, K.P., I.C. Ziomas, and A.G. Paliatsos, *3-Day-ahead forecasting of regional pollution index for the pollutants NO2, CO, SO2, and O3 using artificial neural networks in Athens, Greece.* Water, Air, & Soil Pollution, 2010. **209**(1-4): p. 29-43.

14. Voukantsis, D., et al., *Intercomparison of air quality data using principal component analysis, and forecasting of PM 10 and PM 2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki.* Science of the Total Environment, 2011. **409**(7): p. 1266-1276.

15. Feng, Y., et al., *Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification.* Atmospheric Environment, 2011. **45**(11): p. 1979-1985.

16. Singh, K.P., et al., *Linear and nonlinear modeling approaches for urban air quality prediction.* Science of the Total Environment, 2012. **426**: p. 244-255.

17. Donnelly, A., B. Misstear, and B. Broderick, *Real time air quality forecasting using integrated parametric and non-parametric regression techniques.* Atmospheric Environment, 2015. **103**: p. 53-65.

18. Hoi, K., K. Yuen, and K. Mok, *Optimizing the performance of Kalman filter based statistical time-varying air quality models.* Global NEST J, 2010. **12**: p. 27-39.

19. Beelen, R., et al., *Development of NO< sub> 2</sub> and NO< sub> x</sub> land use regression models for estimating air pollution exposure in 36 study areas in Europe–The ESCAPE project.* Atmospheric Environment, 2013. **72**: p. 10-23.

20. Tamas, W., et al., *Hybridization of Air Quality Forecasting Models Using Machine Learning and Clustering: An Original Approach to Detect Pollutant Peaks.* Aerosol and Air Quality Research, 2015.

21. Viana, M., et al., *Influence of African dust on the levels of atmospheric particulates in the Canary Islands air quality network.* Atmospheric Environment, 2002. **36**(38): p. 5861-5875.

22. Vukmirović, Z.B., et al., *Regional air pollution caused by a simultaneous destruction of major industrial sources in a war zone. The case of April Serbia in 1999.* Atmospheric Environment, 2001. **35**(15): p. 2773-2782.

23. Borge, R., et al., *Analysis of long-range transport influences on urban PM10 using two-stage atmospheric trajectory clusters.* Atmospheric Environment, 2007. **41**(21): p. 4434-4450.

24. Jorba, O., et al., *Cluster analysis of 4-day back trajectories arriving in the Barcelona area, Spain, from 1997 to 2002.* Journal of Applied Meteorology, 2004. **43**(6): p. 887-901.

25. Anastasspoulos, A., S. Nguyen, and X. Xu *On the use of the HYSPLIT model to study air quality in Windsor, Ontario, Canada*. Environmental Informatics Archives 2004. **2**, 517-525.

26. Riccio, A., G. Giunta, and E. Chianese, *The application of a trajectory classification procedure to interpret air pollution measurements in the urban area of Naples (Southern Italy).* Science of the Total Environment, The, 2007. **376**(1-3): p. 198-214.

27. Donnelly, A., B. Broderick, and B. Misstear, *Relating background NO2 concentrations in air to air mass history using non-parametric regression methods: application at two background sites in Ireland.* Environmental Modeling & Assessment, 2012. **17**(4): p. 363-373.

28. Donnelly, A.A., B.M. Broderick, and B.D. Misstear, *The effect of long-range air mass transport pathways on PM10 and NO2 concentrations at urban and rural background sites in Ireland: Quantification using clustering techniques.* Journal of Environmental Science and Health, Part A, 2015. **50**(7): p. 647-658.

29. Draxler, R. and G. Hess, *An overview of the HYSPLIT_4 modelling system for trajectories, dispersion, and deposition.* Australian Meteorological Magazine, 1998. **47**(4): p. 295-308.

30. Yu, K.N., et al., *Identifying the impact of large urban airports on local air quality by nonparametric regression.* Atmospheric Environment, 2004. **38**(27): p. 4501-4507.

31. Donnelly, A., B. Misstear, and B. Broderick, *Application of nonparametric regression methods to study the relationship between NO< sub> 2</sub> concentrations and local wind direction and speed at background sites.* Science of the Total Environment, 2011. **409**(6): p. 1134-1144.

32. Silverman, B.W., *Density estimation for statistics and data analysis*. 1986: Chapman & Hall/CRC.

33. Mansfield, E.R. and B.P. Helms, *Detecting multicollinearity.* The American Statistician, 1982. **36**(3a): p. 158-160.

34. Willmott, C.J., *Some comments on the evaluation of model performance.* Bulletin of the American Meteorological Society, 1982. **63**(11): p. 1309-1313.

35. Kang, D., et al., *New categorical metrics for air quality model evaluation.* Journal of applied meteorology and climatology, 2007. **46**(4): p. 549.

36. Hubbard, M.C. and W.G. Cobourn, *Development of a regression model to forecast ground-level ozone concentration in Louisville, KY.* Atmospheric Environment, 1998. **32**(14): p. 2637-2647.

ACKNOWLEDGEMENTS

This paper has been written as part of the Science, Technology, Research and Innovation for the Environment (STRIVE) Programme 2007–2013. The programme is financed by the Irish Government under the National Development Plan 2007–2013 and administered on behalf of the Department of the Environment, Heritage and Local Government by the Environmental Protection Agency which has the statutory function of co-ordinating and promoting environmental research.