MYAIR TOOLKIT FOR MODEL EVALUATION

Amy Stidworthy[1]*, David Carruthers*[1]*, Jenny Stocker*[1]*, Dimitris Balis*[2]*, Eleni Katragkou*[2]*, Jaakko Kukkonen*[3]

[1]Cambridge Environmental Research Consultants (CERC), Cambridge, UK
[2]Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece
[3]Finnish Meteorological Institute (FMI), Helsinki, Finland

Abstract: PASODOBLE is the GMES downstream service project, producing local-scale air quality services for Europe under the name 'Myair' (http://www.myair.eu/). The local forecast model evaluation support work package of PASODOBLE has developed, demonstrated and evaluated a toolkit for evaluating local air quality forecasts: the Myair Toolkit for Model Evaluation. A key aim in the design of the toolkit was to build on existing tools and methodologies wherever possible. A state-of-the-art review carried out in 2010 and updated in 2011 identified two key initiatives upon which the toolkit was later built: firstly the work of the FAIRMODE community in developing the DELTA tool and secondly the openair suite of tools (Carslaw and Ropkins, 2012). In addition event-based methods used in weather forecasting evaluation were identified which could be applied to pollution forecasting.

The resulting toolkit consists of four tools: a questionnaire tool offering structured advice on the advisability of the proposed evaluation; a data input tool able to import a wide range of modelled and in-situ monitored data formats; a model evaluation tool that analyses the performance of the model at predicting concentrations and pollution episodes; and a model diagnostics tool that compares modelled and monitored data at individual stations in more detail. The Myair Toolkit is easy to use and produces statistical data and attractive graphs. It is coded in the widely used statistical language R with an interface for user inputs; it also has a command-line mode giving scope for automating its use, for example in batch files. The Toolkit has been evaluated during the PASODOBLE project by a panel of air quality forecasting users. It has also been used to evaluate the performance of the *air*TEXT pollution forecasts for Greater London (http://www.airtext.info/). In addition to its local forecasting assessment capability, the Toolkit has the potential to be exploited more generally in the field of air pollution model evaluation and has therefore been used in the validation of the ADMS suite of air dispersion models. This paper gives an introduction to the Myair Toolkit, including examples of its use, within PASODOBLE and for *air*TEXT and ADMS validation.

Key words: *PASODOBLE, airTEXT, Myair, forecasting, air dispersion, validation*

INTRODUCTION

A number of tools currently exist for air dispersion model evaluation. These can be used to assess the accuracy of air quality forecasting systems to a certain extent, but they have limitations when applied to forecasting. Air quality forecasting systems often predict air quality levels in terms of indices, which are related to fixed concentration ranges. A range of indices may correspond to a set of index bands, for instance in the UK's Daily Air Quality Index (DAQI) system, the indices are numbered one to ten, and these are associated with index bands: 'low' (1-3), 'moderate' (4-6), 'high' (7-9) and 'very high' (10). The total pollution index is usually taken to be the maximum over all pollutants. It is common for pollution alerts to be triggered when the forecast pollution index exceeds the moderate, high or very high thresholds. In order to assess the accuracy of a forecasting system therefore, the accuracy of both predicted indices and alert threshold exceedences needs to be assessed.

The local forecast model evaluation support work package of PASODOBLE has developed, demonstrated and evaluated a toolkit for evaluating local air quality forecasts. This 'Myair Toolkit for Model Evaluation' is a free, open-source tool for evaluating air quality models, which can be downloaded from http://www.myair.eu/products-services/local-model-evaluation. The specification for the Myair Toolkit was developed based on existing methodologies and tools, including those currently used to evaluate air dispersion modelling outputs, and those used to assess the accuracy of meteorological forecasts.

The existing methodologies and tools are reviewed in the first section below. Some of the key Myair Toolkit capabilities are discussed in the following section alongside example outputs. The main features of the Toolkit are then summarised.

EXISTING METHODOLOGIES AND TOOLS

A number of existing methodologies were reviewed prior to the development of the Myair Toolkit. These include the Model Validation Kit (also known as 'BOOT', Olsen, 2001), which was developed as part of the Harmo initiative. The American Society for Testing and Materials model evaluation methodology (Irwin *et al.*,2002) differs from the Model Validation Kit in two main respects: consideration of more than one observed value per arc and classification of data into regimes with similar physical properties prior to statistical analyses. The methodology used for the analysis of the performance of meteorological forecasting models is also highly relevant to a tool for evaluating local air quality forecasts, in part because forecast meteorological data are an essential input to air quality forecasting models. The indicators used at the European Centre for Medium-range Weather Forecasting for assessing meteorological forecasts were reviewed as part of this work. Ongoing initiatives and currently available tools were also reviewed in detail prior to the development of the Myair Toolkit. The most relevant of these were the FAIRMODE initiative and the openair tools; details of these projects are given below.

FAIRMODE

FAIRMODE is a joint European Environment Agency and European Commission Joint Research Centre (JRC) initiative. The FAIRMODE working group WG2 have produced a benchmarking report (Thunis *et al.*, 2011) promoting the use of common tools and metrics that are able to assess model output performance relating to the Air Quality Directive (2008/50/EC). FAIRMODE demonstrates a procedure for the benchmarking of air quality models to evaluate their performance and to indicate possible improvements. The procedure has been developed to support both model users and model developers. The type of pollutants, period of interest and spatial scales are determined by the requirements of the Air Quality Directive. One element of the procedure is the 'DELTA' model evaluation tool, which is able to rapidly calculate diagnostics of model performance based on the set of variables of direct relevance for the AQ directive; this tool has been developed at the EU JRC.

openair

This UK project for the air pollution community (Carslaw and Ropkins, 2012) provides free, open-source, innovative data analysis tools. The tools are written in R (http://www.r-project.org/), a programming language developed specifically for the purposes of analysing data, and allow plotting of wind roses, pollution roses and time series of data. Whilst the package is not designed specifically for model evaluation it can help users visualise, analyse and interpret data resulting from a measurement campaign or model simulation.

MYAIR TOOLKIT CAPABILITIES

The Myair Toolkit for Model Evaluation has a range of capabilities that make it suitable for assessing local air quality forecasts. Key features are discussed below, with output examples given for both forecasting applications and general air dispersion model validation.

Model forecast skill assessment

The forecast index evaluation tool produces three types of output: *forecast index accuracy*, *forecast alert accuracy* and *forecast index data* output files. The *forecast index accuracy* is related to the performance of the model's forecast index predictions against forecast indices calculated from observed concentrations. An example forecast index accuracy graph is shown in Figure 1. The bar chart shows, for each station, the percentage of calculated forecast indices where the modelled index was equal to the observed index (green) and where the modelled index was equal to the observed index plus or minus one band (grey). Only forecast periods for which both modelled and observed indices can be calculated are included in the assessment. The stations are sorted by the number of indices valid for comparison, which is also shown on the chart by the blue circles and the right-hand y-axis. A key to the stations is given below the graph.

It is usual for operational pollution forecasting services to use pollution bandings, rather than pollution indices, to communicate pollution levels to the public. Pollution alerts are often triggered when the forecast pollution index changes band. The *forecast alert accuracy* can be assessed by the tool using a number of metrics: the odds ratio skill score, the probability of a correct forecast, the probability of detection, the false alarm ratio, and the probability of false detection. The definitions of these are given in the Myair Toolkit user guide. An example graph showing one of these, the probability of a correct forecast, is shown in Figure 2. The stations are sorted by the number of observed alert threshold exceedences. Note that the probability of a correct forecast is the total number of correct forecasts (both alerts and non-alerts) divided by the total number of forecasts. As well as the graphs, the Toolkit output includes text files of all raw, processed and statistical data presented on the graphs, to provide an audit trail and to allow further analyses to be performed.

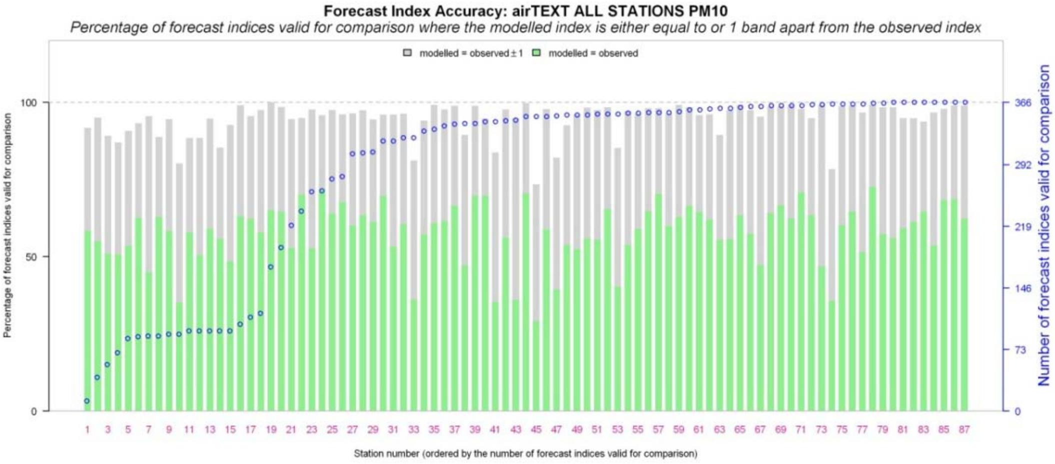

Figure 1. Example of the forecast index accuracy graph (PM_{10} indices), taken from the 2012 evaluation of the *air*TEXT forecasting system (station key omitted)

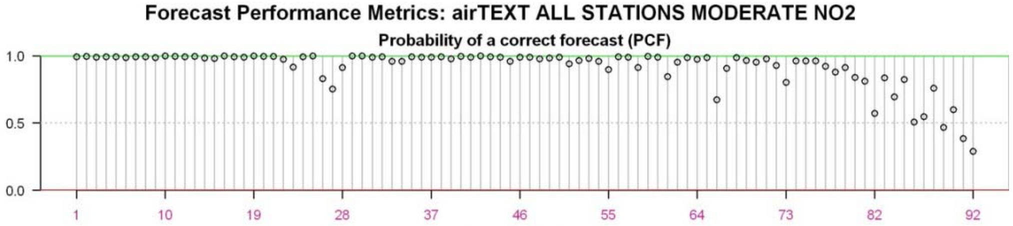

Figure 2. Example of the probability of a correct forecast graph (moderate NO_2 alerts), taken from the 2012 evaluation of the *air*TEXT forecasting system (station key omitted)

Modelled concentration assessment

The model evaluation part of the Myair Toolkit is able to calculate the usual summary statistics of the concentrations calculated by the model: mean, standard deviation, mean bias, normalised mean square error, correlation coefficient, fraction of modelled values within a factor of two of the observed, fractional bias and fractional standard deviation. Note that the sign of the bias and fractional bias calculated by the Toolkit is consistent with openair and the DELTA tool, but not with the BOOT package. Statistics generated by the Myair Toolkit have been used to assess and compare road source model results (Stocker *et al.*, 2013).

Models may under-predict average concentrations from annual campaigns where there are few receptors and the plume impacts with the receptors for a relatively small proportion of the experiment. The modelled pollutant for these experiments is usually SO_2, and the lack of inclusion of background concentrations leads to the under-prediction. For these experiments, therefore, it is the maximum concentrations, and the robust highest concentration (RHC) that are of interest. RHC is defined by $\chi(n)+(\chi-\chi(n))\ln((3n-1)/2)$ where n is the number of values used to characterise the upper end of the concentration distribution, χ is the average of the n-1 largest values, and $\chi(n)$ is the n^{th} largest value; n is taken to be 26. These statistics are calculated by the Toolkit, over any averaging time specified by the user.

The model evaluation tool can create scatter, quantile-quantile and box and whisker plots of results. An example quantile-quantile plot of ADMS 5 modelled results for the Lovett power plant field study (US EPA 2003) is shown in Figure 3 a); this was an annual campaign, with SO_2 measurements at nine receptors located in the complex terrain surrounding the power station. The tool can also create a ‘target’ plot based on the DELTA tool target plot (version 1.2). Figure 3 b) shows an example target plot for one of the Idaho Falls experiments (Stocker *et al.*, 2013); ADMS-Roads model results are displayed.

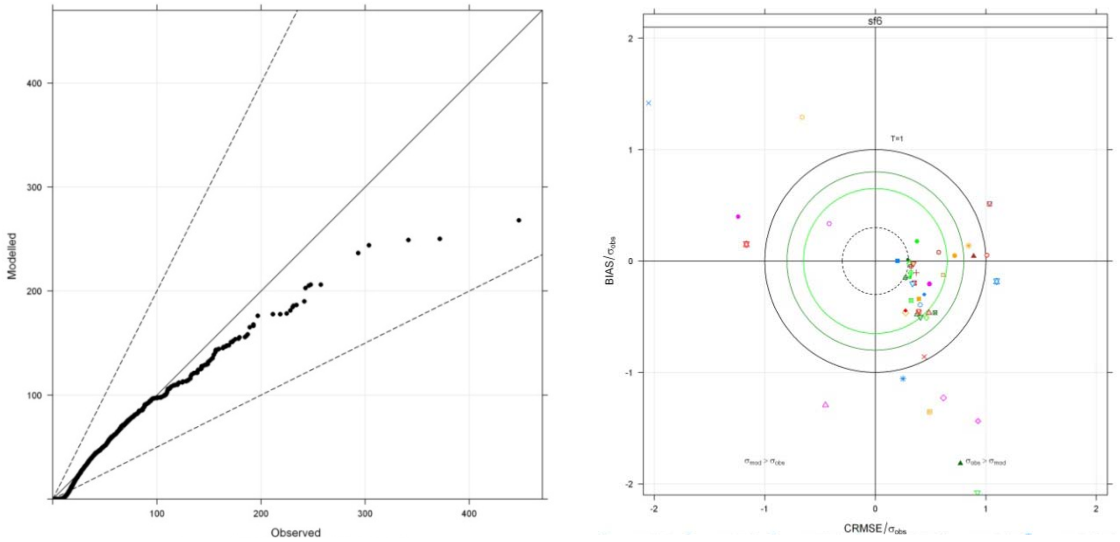

Figure 3. Example ADMS model evaluation output from the tool: a) Quantile-quantile plot for Lovett validation study; and b) Target plot (Delta version 1.2) for Test 1 of the Idaho Falls (titles and station key omitted).

Model diagnostic tool

The model diagnostics tool part of the Myair Toolkit allows detailed analysis of model performance for a single monitoring site and pollutant based on tools developed by the openair project. It has three output options, a *time variation plot*, *frequency scatter plot* and *time series plot*. In addition to raw concentration data, concentration statistics or forecast index values can be plotted on the frequency scatter plot and time series plot. The *time series plot* shows the variation of concentration over time, which can identify periods of missing data. The *frequency scatter plot* uses colour to represent the number of points in a particular 'bin' of modelled and observed concentration values, and can be useful for interpreting densely clustered data. The *time variation plot* shows averaged daily and monthly profiles for modelled and monitored concentrations. An example is shown in Figure 4. This urban site has good agreement between the annual average modelled and monitored concentrations, but the time variation plot shows that there are discrepancies in the daily profiles, particularly at weekends.

Import of different modelled and observed data formats

The Myair Toolkit supports a wide range of input modelled data file formats for modelled data, the majority of which are netCDF files, for instance: AIRSHEDS, MACC Ensemble, CMAQ and DELTA; full details are given in the Myair Toolkit user guide. In addition, point receptor output from the ADMS suite of atmospheric dispersion models (CERC, 2013) can be imported, as well as data formatted in a text file. Gridded data are interpolated to monitoring site locations. In the UK, monitoring data from the London and UK-wide networks of automatic monitoring stations can be downloaded and imported automatically to match modelled data. Alternatively, observed data may be formatted in a text file.

Other Toolkit capabilities

The Myair Toolkit user can specify whether to analyse individual monitoring stations, groups of stations or all stations. The station 'types' are defined by the user and are useful for detailed inspection of model performance, for instance, for field studies such as Kincaid and Indianapolis (included in the Model Validation Kit), model performance can be categorised in terms of quality of measurement, or receptor arc. Graphical output can be saved as PNG, JPG or PDF files. The Myair tools can be run from an intuitive user interface or from a batch file.

SUMMARY

The Myair Toolkit for Model Evaluation has been developed during the PASODOBLE project to build on existing tools and methodologies for model evaluation. Specifically, the indicators used in the *forecast* *model skill assessment* tools are consistent with those used in the assessment of meteorological forecasts. The *modelled concentration assessment* tools calculate similar statistics to other packages, such as BOOT and the DELTA tool and the *model diagnostic tool* uses features from the openair package. The combination of these state-of-the-art features makes the Myair Toolkit very useful for the evaluation of both forecast and hindcast air dispersion model output.

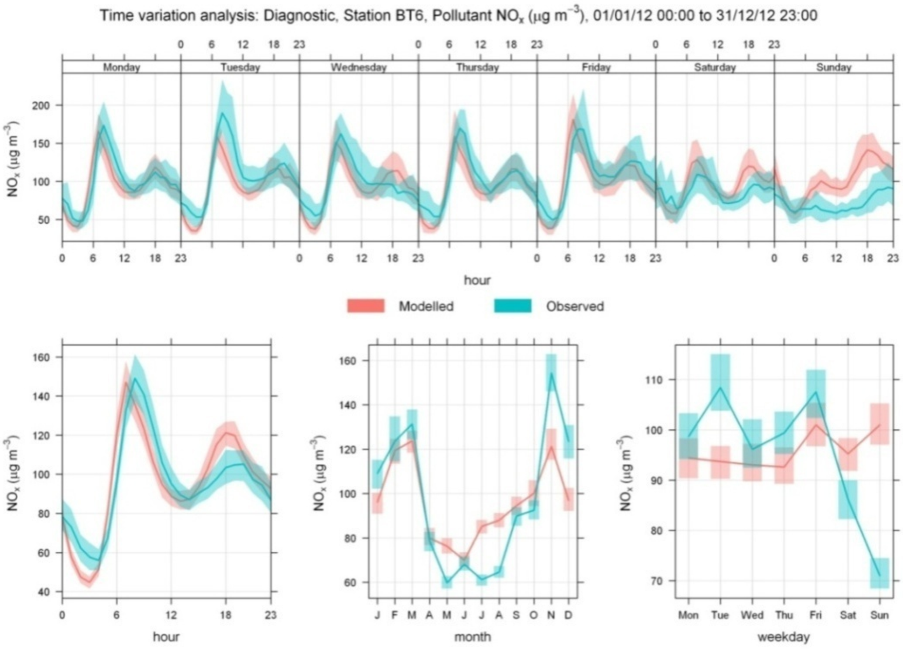

Figure 4. Time variation plot taken from the 2012 *air*TEXT forecasting system evaluation.

ACKNOWLEDGEMENTS

The authors wish to thank the PASODOBLE project partners, in particular the users in the 'Local forecast model evaluation support' work package and those involved in evaluating the AIRSHEDS products, for their valuable feedback during development and demonstration of the Myair Toolkit. The authors also wish to thank the EU FP7 programme for funding the 3-year PASODOBLE project, which finished at the end of April 2013.

REFERENCES

Carslaw, D.C. and Ropkins, K., 2012: openair — an R package for air quality data analysis. *Environ. Model. & Softw*., **27-28**, 52-61.

CERC, 2013: ADMS Technical Specifications. [online] Available at http://www.cerc.co.uk/environmental-software/model-documentation.html#technical (accessed March 2013)

Olesen, H.R., 2001: A platform for model evaluation. 7th international conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Belgirate, Italy.

Irwin J.S., Carruthers D.J., Paumier J. and Stocker J., 2002: Application of ASTM D6589 to evaluate dispersion model performance to simulate average centerline concentration values. *Int. J. of Environ. & Pollution*, **20**, 4-10.

Stocker, J., Heist, D., Hood, C., Isakov, V., Carruthers, D., Perry, S., Snyder, M., Venkatram, A., Arunachala, S., 2013: 'Road source model intercomparison study using new and existing datasets' 15th International Conference on Harmonisation, Madrid, Spain.

Thunis, P., Georgieva, E. and Galmarini, S., 2011: A procedure for air quality models benchmarking. FAIRMODE WG2 SG4 report, available at http://fairmode.ew.eea.europa.eu/models-benchmarking-sg4/wg2_sg4_benchmarking_v2.pdf (accessed March 2013)

US EPA, 2003 'AERMOD: Latest Features and Evaluation Results' report reference EPA-454/R-03-003