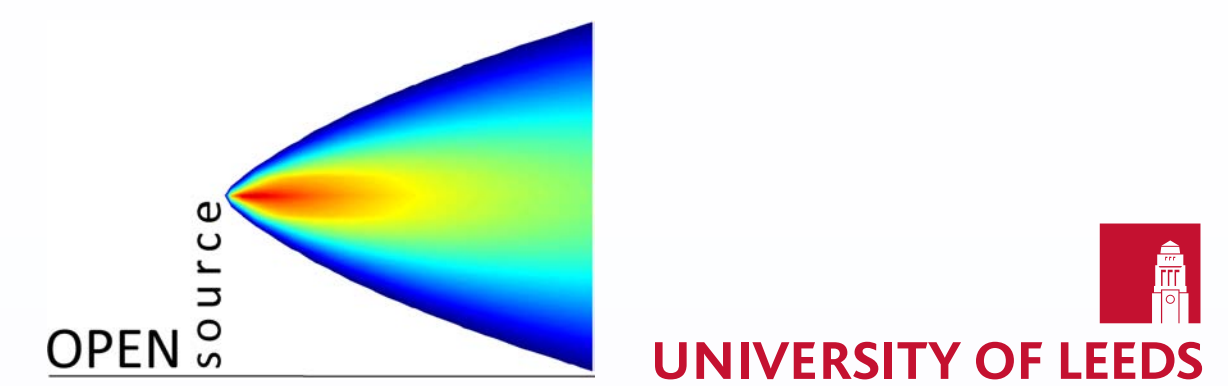


# Open Source Tools for Insightful Air Quality Data Analysis

David Carslaw and Karl Ropkins, University of Leeds



## Background and aims

In Europe, air pollution is the environmental factor with the greatest impact on health and is responsible for the largest burden of environment-related disease.<sup>a</sup> It is essential therefore that a comprehensive understanding of the sources, dispersion, chemical transformation and exposure to pollutants is developed, if air pollution is to be effectively controlled. One of principal tools used to achieve this aim is the measurement of pollutant concentrations. The **openair** project is designed to help better understand these issues.

The aims of the project are:

- Make available innovative tools for the analysis of air pollution data
- To help overcome many of the barriers that prevent the more insightful analysis of data
- To implement tools in **R** software
- Develop an approach that maximises uptake and participation by using entirely open-source software and methods that will maximise international adoption and participation

<sup>a</sup>European Environment Agency, 2005

## R Software

A key feature of our project is that it is based on the highly capable data analysis and statistical software called **R**. **R** is a programming language that has been designed for data analysis and is ideally suited to the **openair** project. In recent years its use has increased markedly across many sectors including biotechnology, environmental analysis and finance. Among the benefits to **openair** are:

- It is free and open-source — anyone can access the code, see how it works and improve it
- It is very well supported worldwide
- It implements leading analysis approaches and graphical methods
- The package system is ideal for disseminating software easily and widely — internationally
- It is updated twice-yearly and continues to develop rapidly

## Why open-source?

Openness and transparency should be at the heart of environmental regulation. Many believe that those affected by environmental decisions using models and data should have full and open access to the methods and data used — and our project reflects this thinking. There are also other advantages:

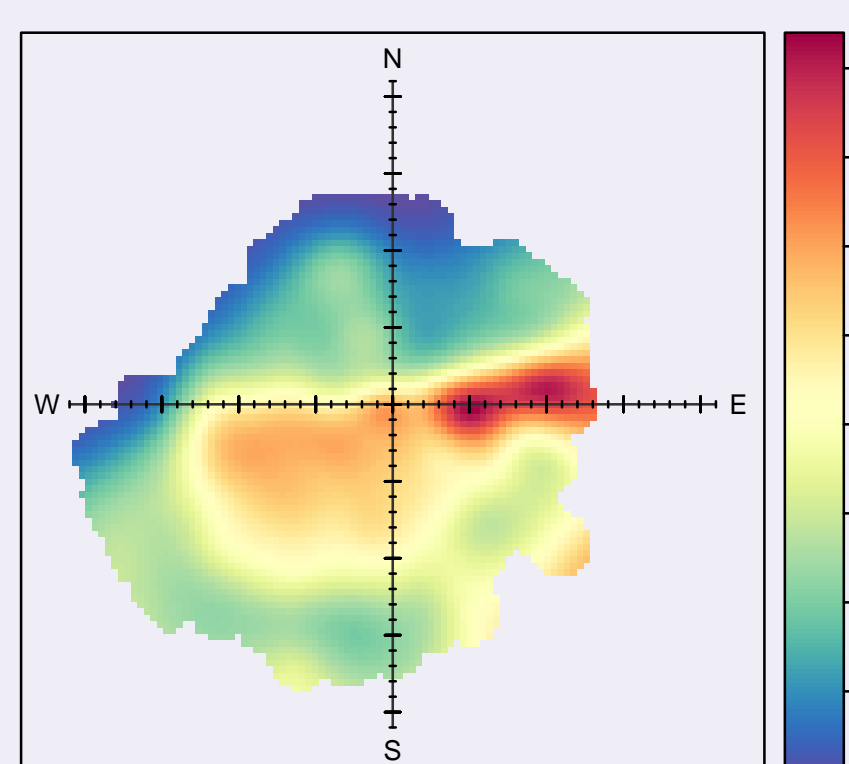
- Improved scrutiny of methods<sup>a</sup>
- The ability for others to contribute and improve the code
- Builds trust — no 'black boxes' and analysis can easily be made reproducible

<sup>a</sup>Successful **R** software packages have strong input from users

## What effort is required?

The aim is to make **openair** tools as simple as possible to use so that users can focus on questions and thinking about their data. Currently this is achieved by typing in some simple commands. These commands can easily be recorded to ensure analyses can be reproduced. In the example below, data are first read into **R** and a bivariate polar plot plotted [1].

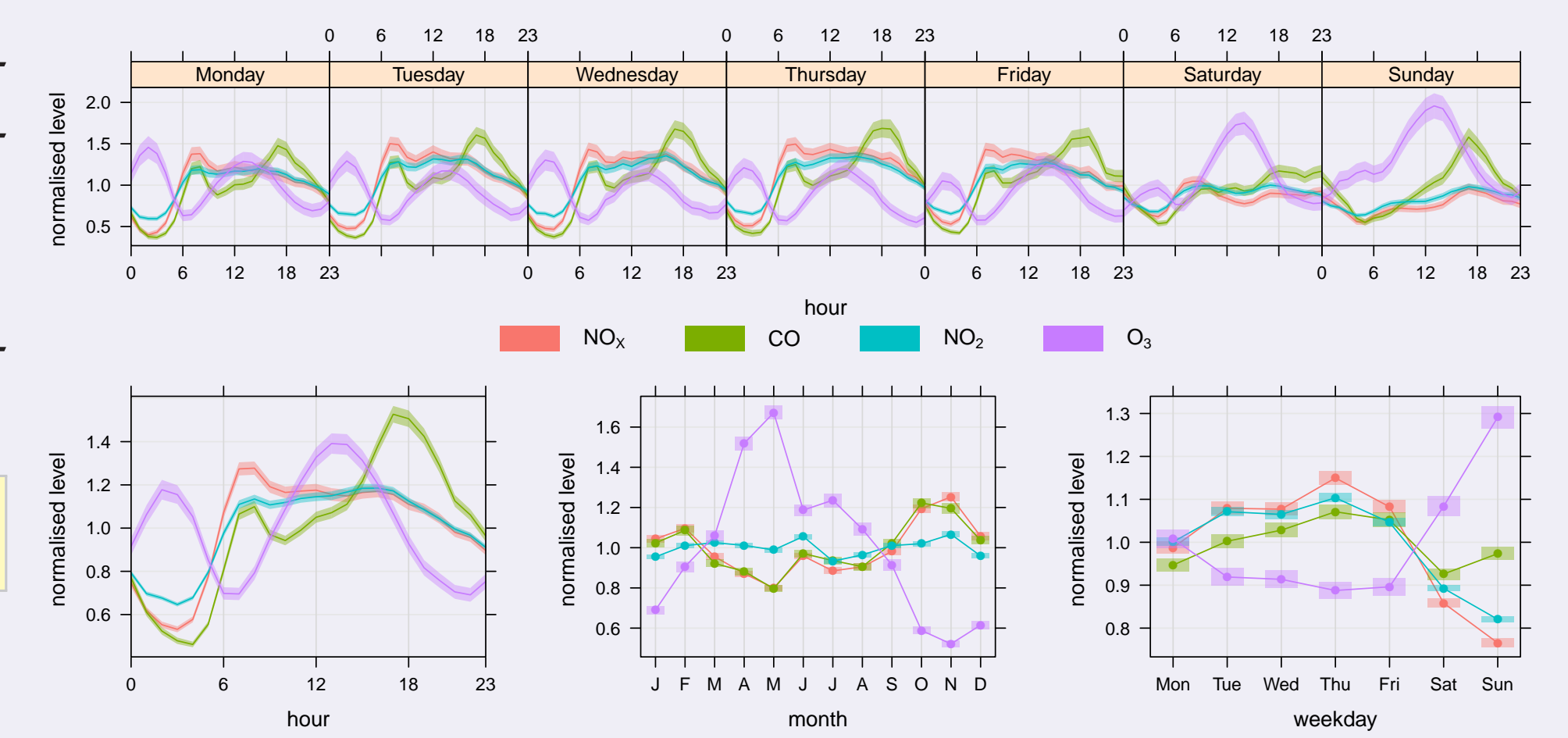
```
mydata = import("d:/data/air pollution data.csv")
polar.plot(mydata, pollutant = "so2")
```



## Examples

Considering how pollutant concentrations vary temporally can provide many insights into the factors controlling their variation. For example, road vehicle emissions vary by day of the week and hour of the day by vehicle type, some pollutants have strong seasonal cycles etc. The plot (right) is easily produced:

```
time.variation(mydata, pollutant = c("nox", "co", "no2", "o3"),
normalise = TRUE)
```



The shading shows the 95% confidence intervals in the mean and can be useful when trying to determine whether two conditions differ from one another. Setting **normalise = TRUE** makes it easier to compare different variables that are on very different scales; achieved by dividing by their mean value.

These types of approaches are very useful for *model evaluation* where model outputs can be compared with observations. It may be possible to easily see that a model tends to overestimate nighttime concentrations of ozone, which might provide clues as to where a model is deficient, for example.

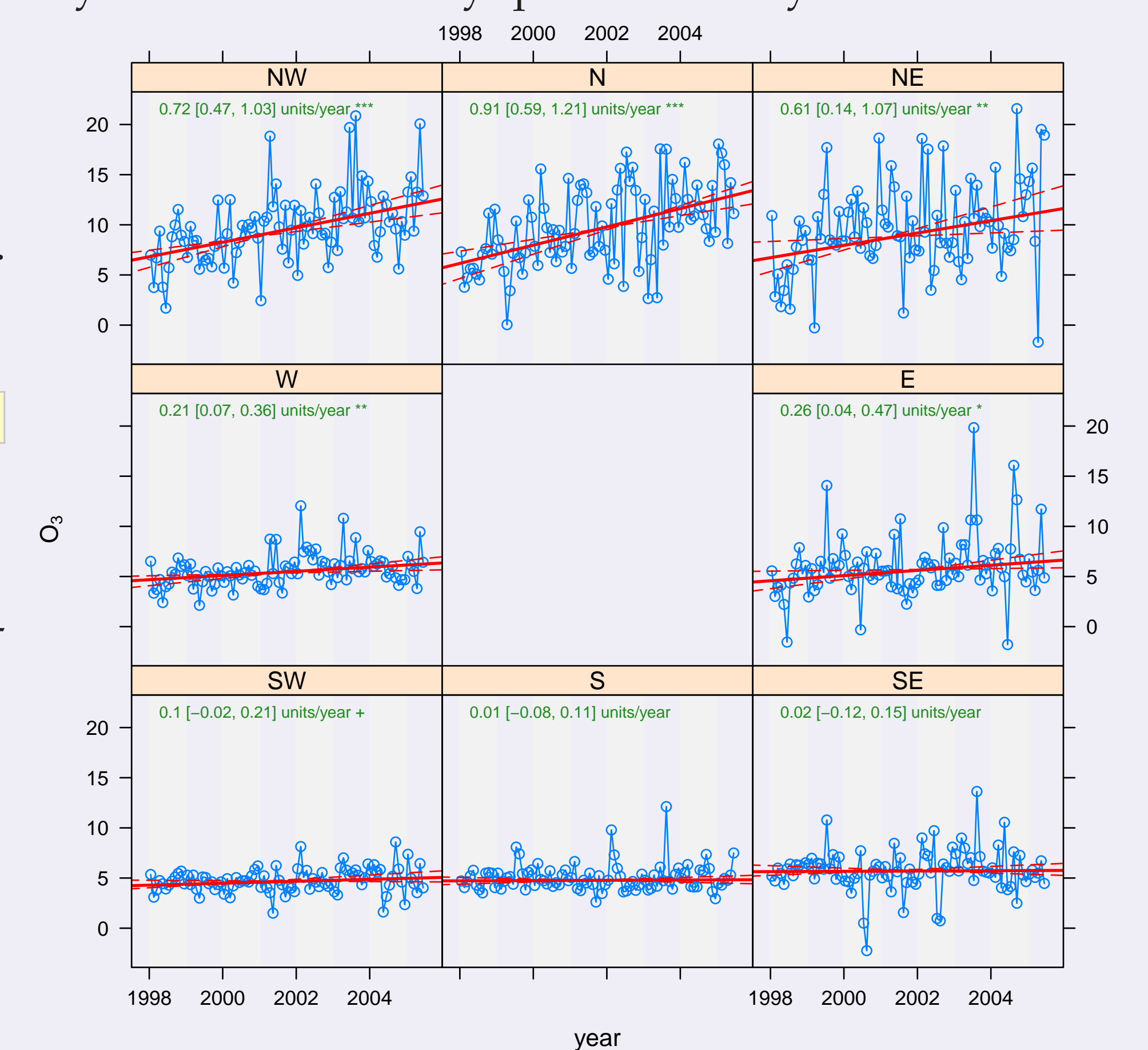
One of the important concepts in **openair** is the idea of *conditioning*. Rather than just considering how  $x$  varies with  $y$ , considerably more insight can often be gained by considering how  $x$  varies with  $y$  for different levels of a third variable,  $z$ . As an example, consider trends in ozone at Marylebone Road. In this plot, concentration is plotted against time, *dependent* on a third variable, wind direction. This way it is easy to see how strongly trends in ozone depend on wind direction.

```
MannKendall(mydata, pollutant = "o3", deseason = TRUE, type = "wd")
```

This plot:

- Uses Mann-Kendall estimates for trend with Sen-Theil slope estimates
- Deseasonalises the data to remove much of the monthly variation
- Plots the trends in a logical way — by points of the compass
- Uses re-sampling (bootstrap) techniques to estimate uncertainty intervals and account for autocorrelation

to condition data including by year, season, hour of the day, day of the week and by quantiles of any other variable...

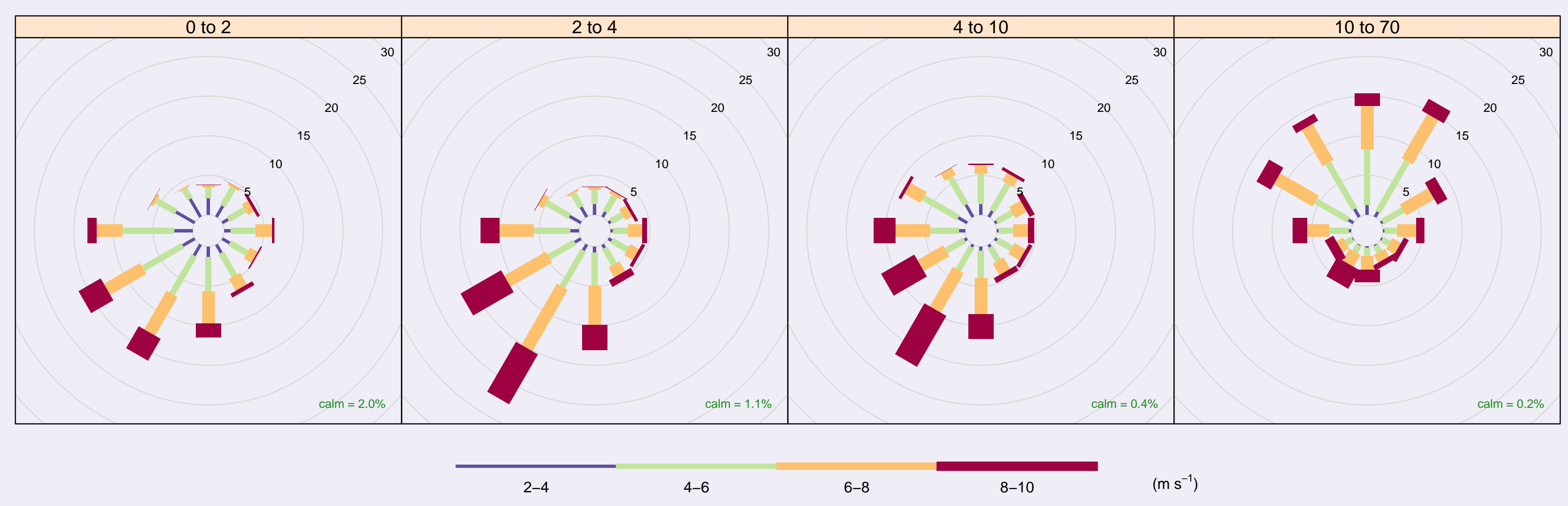


Most functions in **openair** provide extensive ways in which

In the example below, wind roses are plotted for different levels of ozone e.g. 0–2, 2–4, 4–10 and 10–70 ppb (quantiles of ozone). In this plot it is easy to see that the lowest concentrations of ozone (0–2 ppb) are dominated by winds

from the south-west, whereas the highest concentrations of ozone are dominated by winds from the north.

```
wind.rose(mydata, type = "o3")
```



## Developments

The **openair** project has been running for one year and we have many goals we wish to achieve, including:

1. Implement methods that have already been published by us or others e.g. change-point detection [2]
2. Develop new techniques — we are currently investigating how better information can be gained using higher temporal resolution measurements and the use of *quantile regression* for example
3. Ensure that excellent documentation is available
4. Implement methods for *reproducible research* — good progress has already been made
5. Develop **openair** on a remote repository for easy installation with a proper version control system
6. Develop case studies that show how the tools can be used to learn more about air pollution
7. Develop more functions for model evaluation
8. Run training courses to help folk out!

## References and acknowledgements

- [1] Carslaw, D.C., Beevers, S.D., Ropkins, K., Bell, M.E., 2006. Detecting and quantifying aircraft and other on-airport contributions to ambient nitrogen oxides in the vicinity of a large international airport. *Atmospheric Environment* 40 (28), 5424–5434.
- [2] Carslaw, D.C., Ropkins, K and M.C. Bell. (2006) Change-Point Detection of Gaseous and Particulate Traffic-Related Pollutants at a Roadside Location. *Environmental Science and Technology*. Vol. 40. Issue 22. 6912–6918.

We are very grateful to the following organisations that have so far provided funding towards this project:

- The Natural Environment Research Council (main sponsor), grant NE/G001081/1
- Defra
- AEA
- Sefton Council
- North Lincolnshire Council

We would very much like to hear from you, so please get in touch!