Supplement of Atmos. Chem. Phys., 15, 79–98, 2015
http://www.atmos-chem-phys.net/15/79/2015/
doi:10.5194/acp-15-79-2015-supplement

Atmospheric Chemistry and Physics
Open Access

Supplement of

Tracing the second stage of ozone recovery in the Antarctic ozone-hole with a "big data" approach to multivariate regressions

A. T. J. de Laat et al.

Correspondence to: A. T. J. de Laat (laatdej@knmi.nl)

QBO.

The Quasi-Biennial Oscillation (QBO) of the winds in the equatorial stratosphere has been discovered in the 1950s through the establishment of a global, regularly measuring radiosonde network [Graystone, 1959; Ebdon, 1960]). The Free University of Berlin has compiled a long-term record from 1953 onwards of daily wind observations of selected stations near the equator. From these daily measurements monthly mean zonal components were calculated for pressure levels of 70, 50, 40, 30, 20, 15, and 10 hPa. For the period after 1979 only measurements from Singapore are used. The QBO data set is supposed to be representative of the equatorial belt since various studies have shown that longitudinal differences in the phase of the QBO are small [Hood, 1997]. It should be noted, however, that some uncertainties arose at higher levels during the early years from the scarcity of observations. More information on the original data and their evaluation can be found in Naujokat [1986].

As proxy for the regressions we will use the 40-hPa QBO index, also used in Kuttippurath et al. [2013]. Salby et al. [2011, 2012] chose to use 30-hPa winds instead. The relevancy of the choice of QBO index will be evaluated later. Information on the uncertainties in the monthly QBO data is not available. One indirect method to estimate the uncertainties is by examining QBO index variability close to the maximum and minimum of the QBO cycles, where the QBO index values remains more or less constant for some months. Assuming that during the maximum or minimum in the QBO phase variations from month to month are indicative of uncertainties in the QBO, we come up with estimated uncertainties of around 1.5-2.0 m/s in the zonal mean wind speeds.

Solar flux

Variations in incoming solar radiation – in particular the shorter ultraviolet wavelengths – have an effect on stratospheric ozone [Haigh, 1996; McKormack and Hood, 1996; Soukharev and Hood, 2006; Anet et al., 2013]. A standard proxy for variations in incoming solar radiation in ozone regression studies is to use the monthly mean 10.7 cm radio flux, as also used in Kuttippurath et al. [2013]. This data set was obtained via NOAA/NESDIS/NGDC/STP.

However, there are other solar activity proxies available. Ideally, in absence of true UV spectral measurements, one would like to use a proxy that is representative for solar activity at those wavelengths where stratospheric ozone formation occurs, which is of roughly between 200 and 300 nm. Dudok de Wit et al. [2009] tried to identify the best proxy for solar UV irradiance, and concluded that proxies derived from a certain wavelength range best represent the irradiance variations in that wavelength band. Thus, the 10.7-cm radio flux might not fully represent solar UV variability. Using the results from Dudok and de Wit et al. [2009] to analyze a set of seven solar activity proxies dating back to at least 1979 based on the solar2000 model and obtained from NOAA/NESDIS/NGDC/STP (F10.7, Lyman-alpha, E10.7, and the solar constant S), we will assume in our regressions that the uncertainty range associated with the solar proxy is approximately 15% of the root-mean-square of the anomaly values.

Why do standard errors of an ordinary linear regression relative to the regression slope not depend on the actual regression itself?

This analysis is based on the "Data Analysis Toolkit" document (chapter 10), written by Prof. James Kircher, Professor of Earth and Planetary Science at the University of California, Berkley and emeritus Goldman Distinguished Professor for the Physical Sciences.

http://seismo.berkeley.edu/~kirchner/

The standard error of the regression slope **b** of an ordinary linear regression of two variables **x** and **y,** and the regression slope **b** itself can be written as:

$$s_b = \frac{b}{\sqrt{n-2}}\sqrt{\frac{1}{r^2}-1} \qquad and \qquad b = r\frac{S_y}{S_x} \tag{S1}$$

In which $\mathbf{s_b}$ is the standard error of the regression slope, **n** the number of data points of the variables **x** and **y, r** is the Pearson correlation coefficient between the variables **x** and **y**, and $S_{x,y}$ is the standard deviation of the variables **x** and **y**.

For a statistically significant trend one generally defines that the trends (slopes) should exceed two times the standard error. Or, in other words, the standard error of the regression slope divided by the regression slope itself should be less than 0.5

The standard error of the regression slope relative to the regression slope itself – which directly relates to statistical significance of the trend - becomes, based on the equation above:

$$s_b / b = \frac{1}{\sqrt{n-2}} \sqrt{\frac{1}{r^2} - 1} \tag{S2}$$

which only depends on the correlation between the variables **x** and **y** and the number of data points of variable **x** and **y** (record length).