Analysis Methods in Atmospheric and Oceanic Science

AOSC 652

- Today: Multiple Linear Regression
- Wed: Description of Projects; Python & IDL breakouts

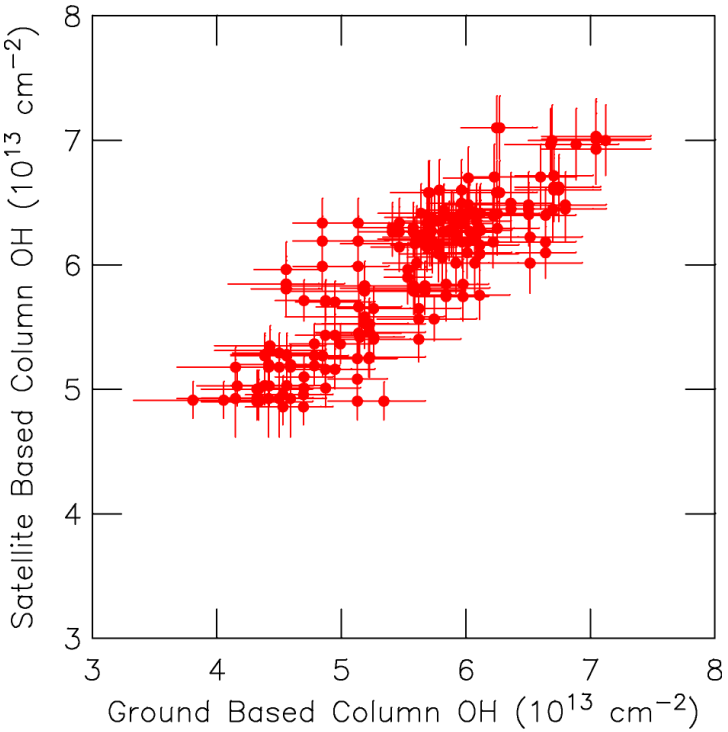
Week 10, Day 1

31 Oct 2016

AOSC 652: Analysis Methods in AOSC

Correlation and Regression

Suppose you have two sets of measurements (or data and model) that you'd like to relate.



What are some aspects of the data that are typically examined?

Correlation and Regression

Linear Correlation Coefficient:

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

r must lie between –1 and 1

If r = 1, the data are said to have a *complete positive correlation*

r = 0, the data are said to be *uncorrelated*

r^2 x 100 = percent of variance in common between x and y

See http://www.mega.nu/ampp/rummel/uc.htm#C2 for a nice explanation

AOSC 652: Analysis Methods in AOSC

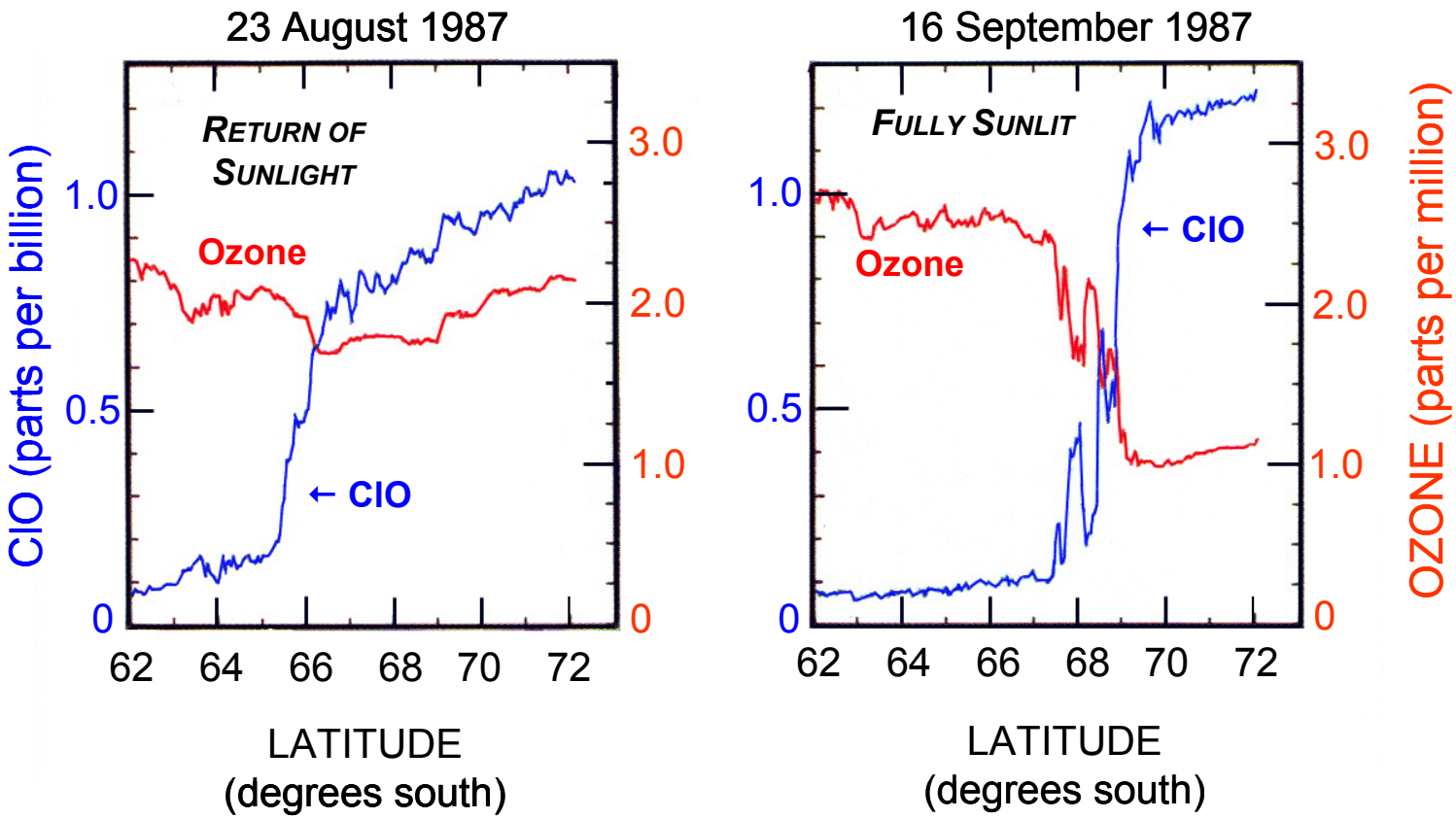
Correlation:

- Used to assess the “relationship” between two or more variables
- “Relationship” questions: **strong** or **weak** correlation ?
  linear and if not, **functional form** ?
- What else ?!?

AOSC 652: Analysis Methods in AOSC

What other knowledge, in addition to these observations, was needed to demonstrate that ClO (from CFCs) causes the ozone hole ?!?

Airborne Antarctic Ozone Expedition:
Punta Arenas, Chile,1987



Anderson et al., *Science*, 1991

AOSC 652: Analysis Methods in AOSC

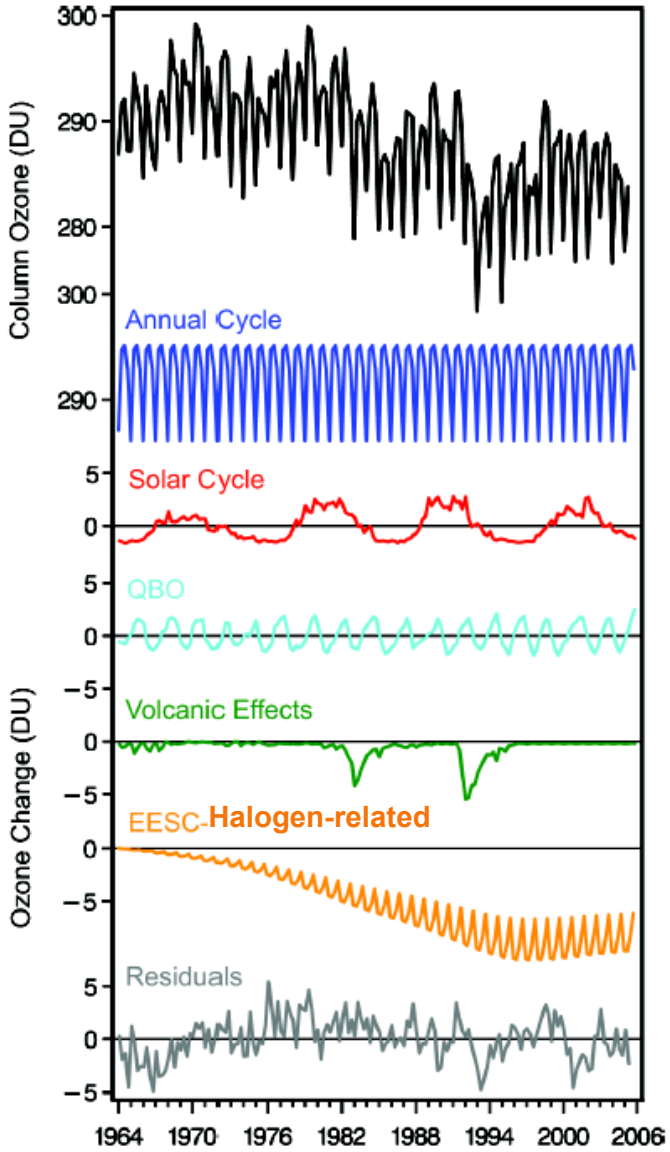
Regression (as used in AOSC):

- **Development of a mathematical model between a set of observations (i.e., a time series or a spatial distribution) and one or more predictor variables (usually measured or inferred from a proxy)**
- **Regression analysis almost always is preceded by a correlation analysis**
- ***Strong science* involves understanding (or development) of the underlying, causal relations between observations and predictor variables**

Multiple Linear Regression

$$Y = c_0 + c_1 X_1 + c_2 X_2 + \ldots + c_p X_p$$

AOSC 652: Analysis Methods in AOSC

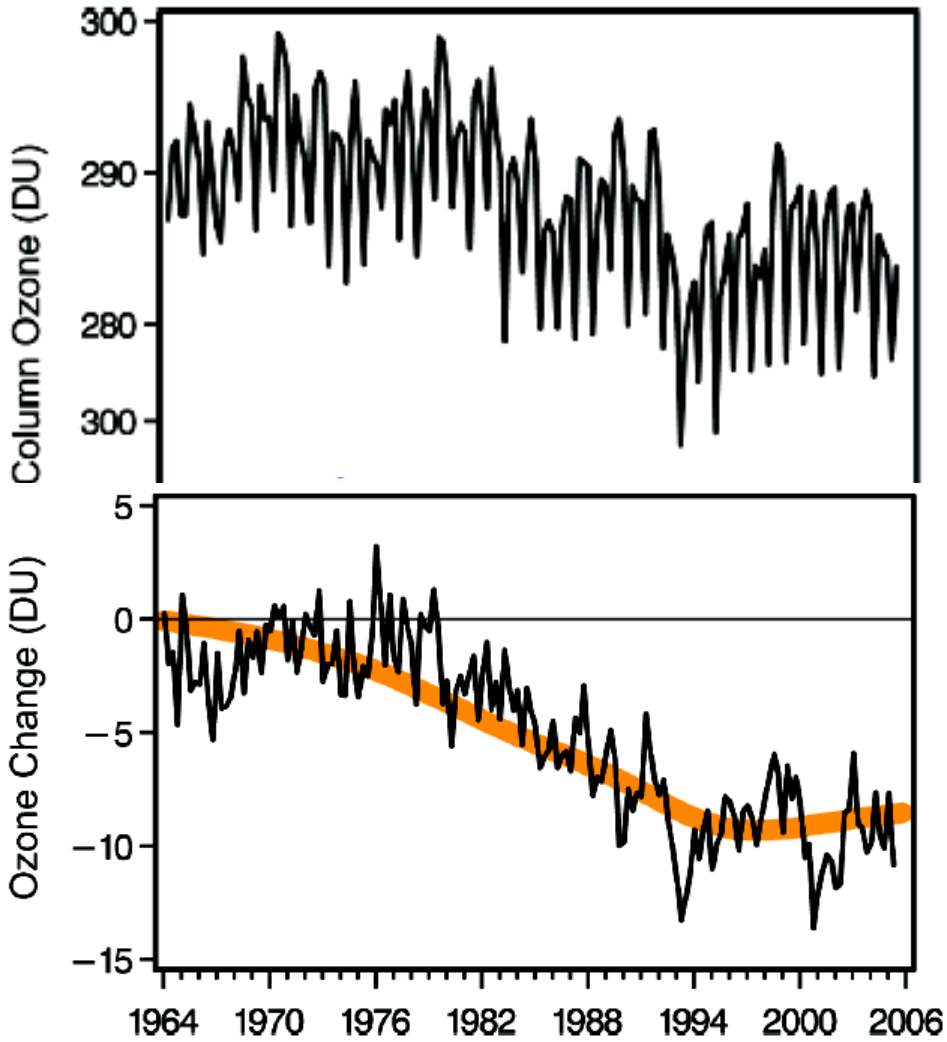


Variations in total column ozone (DU) for 60°S to 60°N (top panel) and individual components of a multiple linear regression of various explanatory variables times the regression coefficient. The residual between the data and the regression model is also shown (bottom panel).

Figure 3-1, 2006 WMO/UNEP Ozone Assessment Report

http://www.esrl.noaa.gov/csd/assessments/ozone/2006/images/Fig3-01.jpg

AOSC 652: Analysis Methods in AOSC



Raw Data (total ozone, 60°S to 60°N)

Adjusted data:
Deseasonalized total ozone deviations from 1964 value adjusted for effects of solar, volcanic, and QBO forcings (BLACK) compared to time series of stratospheric halogens (ORANGE) scaled to fit the data.

Figure 3-1, 2006 WMO/UNEP Ozone Assessment Report

http://www.esrl.noaa.gov/csd/assessments/ozone/2006/images/Fig3-01.jpg

Multiple Linear Regression

$$Y = c_0 + c_1 X_1 + c_2 X_2 + \ \ldots + c_p X_p$$

How do we solve for the regression coefficients ?

Multiple Linear Regression

$$\text{Cost Function} \Rightarrow \mathrm{F} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(c_0 + \sum_{j=1}^{p} c_j X_{i,j} - Y_i \right)^2$$

$$\frac{\partial F}{\partial c_0} = 2 \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(c_0 + \sum_{j=1}^{p} c_j X_{i,j} - Y_i \right) = 0$$

$$\frac{\partial F}{\partial c_1} = 2 \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(c_0 + \sum_{j=1}^{p} c_j X_{i,j} - Y_i \right) \left(X_{i,1} \right) = 0$$

$$\frac{\partial F}{\partial c_2} = 2 \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(c_0 + \sum_{j=1}^{p} c_j X_{i,j} - Y_i \right) \left(X_{i,2} \right) = 0$$

$$\vdots$$

$$\frac{\partial F}{\partial c_p} = 2 \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(c_0 + \sum_{j=1}^{p} c_j X_{i,j} - Y_i \right) \left(X_{i,p} \right) = 0$$

Multiple Linear Regression

$$c_0\ (n) + c_1 \sum_{i=1}^{n} X_{i,1} + c_2 \sum_{i=1}^{n} X_{i,2} + \ \ldots + c_p \sum_{i=1}^{n} X_{i,p} = \sum_{i=1}^{n} Y_i$$

$$c_0 \sum_{i=1}^{n} X_{i,1} + c_1 \sum_{i=1}^{n} (X_{i,1})^2 + c_2 \sum_{i=1}^{n} X_{i,1} X_{i,2} + \ \ldots + c_p \sum_{i=1}^{n} X_{i,1} X_{i,p} = \sum_{i=1}^{n} X_{i,1} Y_i$$

$$c_0 \sum_{i=1}^{n} X_{i,2} + c_1 \sum_{i=1}^{n} X_{i,1} X_{i,2} + c_2 \sum_{i=1}^{n} (X_{i,2})^2 + \ \ldots + c_p \sum_{i=1}^{n} X_{i,2} X_{i,p} = \sum_{i=1}^{n} X_{i,2} Y_i$$

$$\vdots$$

$$c_0 \sum_{i=1}^{n} X_{i,p} + c_1 \sum_{i=1}^{n} X_{i,1} X_{i,p} + c_2 \sum_{i=1}^{n} X_{i,2} X_{i,p} + \ \ldots + c_p \sum_{i=1}^{n} (X_{i,p})^2 = \sum_{i=1}^{n} X_{i,p} Y_i$$

Multiple Linear Regression

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_{i,1} & \sum_{i=1}^{n} X_{i,2} & \cdots & \sum_{i=1}^{n} X_{i,p} \\ \sum_{i=1}^{n} X_{i,1} & \sum_{i=1}^{n} (X_{i,1})^2 & \sum_{i=1}^{n} X_{i,1}X_{i,2} & \cdots & \sum_{i=1}^{n} X_{i,1}X_{i,p} \\ \sum_{i=1}^{n} X_{i,2} & \sum_{i=1}^{n} X_{i,1}X_{i,2} & \sum_{i=1}^{n} (X_{i,2})^2 & \cdots & \sum_{i=1}^{n} X_{i,2}X_{i,p} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} X_{i,p} & \sum_{i=1}^{n} X_{i,1}X_{i,p} & \sum_{i=1}^{n} X_{i,2}X_{i,p} & \cdots & \sum_{i=1}^{n} (X_{i,p})^2 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i,1}Y_i \\ \sum_{i=1}^{n} X_{i,2}Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{i,p}Y_i \end{pmatrix}$$

AOSC 652: Analysis Methods in AOSC

Multiple Linear Regression

$$\overline{\overline{A}} \times \vec{c} = \vec{b} \qquad \text{or} \qquad \vec{c} = \overline{\overline{A}}^{-1} \times \vec{b}$$

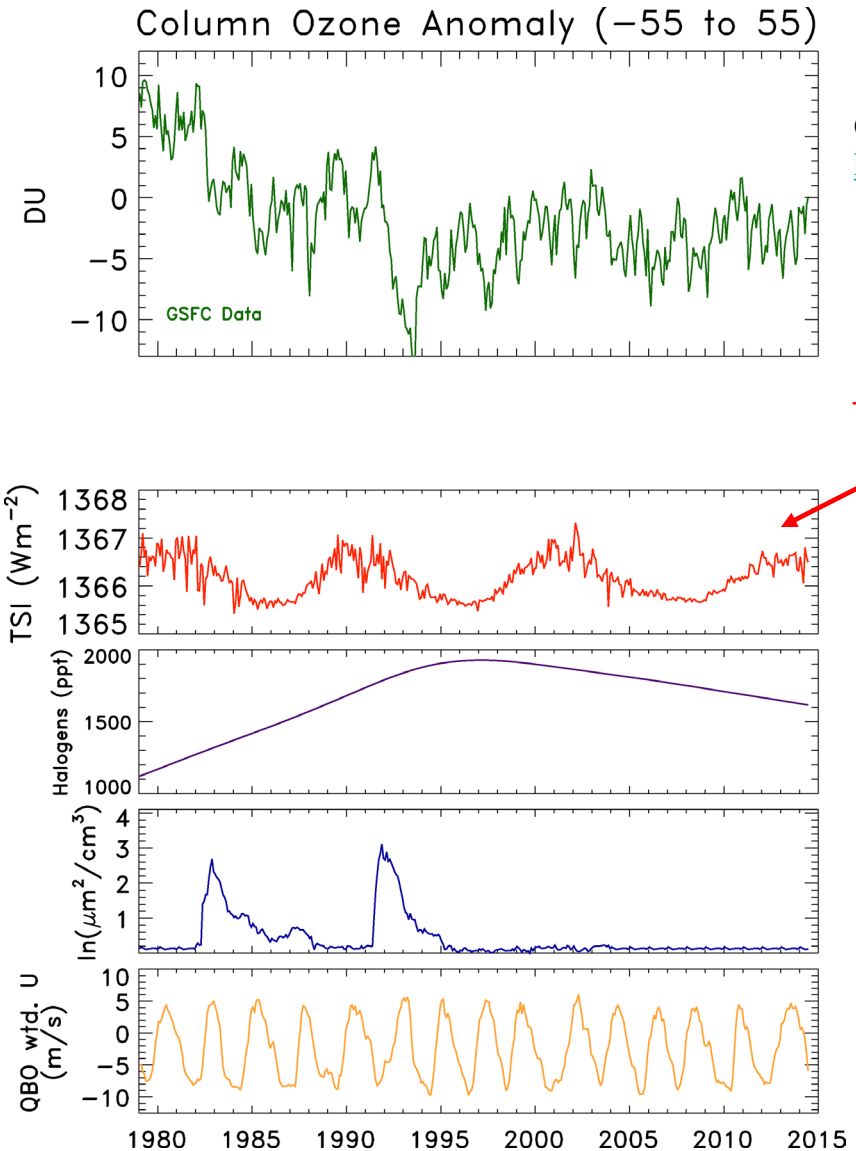
To regress an array of data points (e.g., global mean temperature) versus a set of predictor variables (e.g., ENSO, Volcanic Aerosols, Solar Irradiance, & Annual Avg CO_2) can read the data into a FORTRAN program and compute elements of matrix *A* and array *b*.

Then, can find the inverse of array *A.*

Then, can multiply the inverse of *A* by array *b* to find the array *c*.

Or, can use the "regress" function in IDL or MATLAB !

Global ozone anomaly versus time and 4 other, related quantities

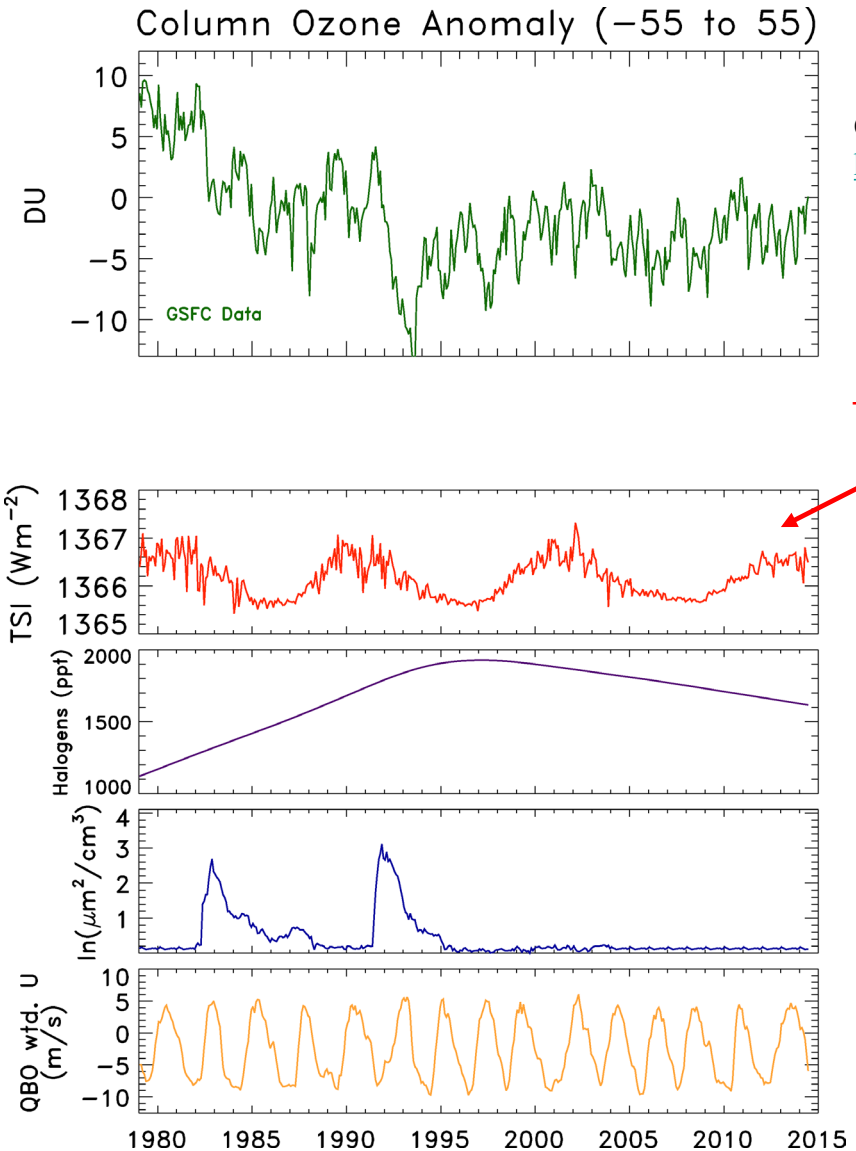


Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged

TSI : Total Solar Irradiance

The amount of energy Earth receives from the sun varies, with a periodicity of __ years

Global ozone anomaly versus time and 4 other, related quantities



Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged

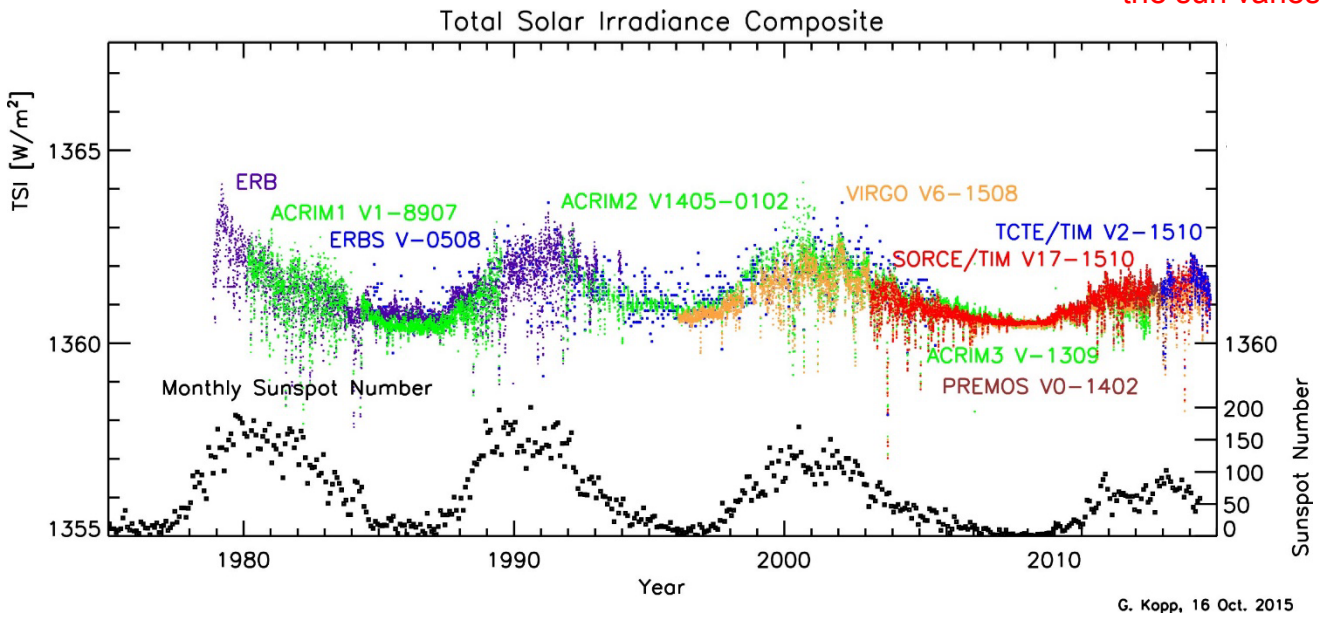
TSI : Total Solar Irradiance

The amount of energy Earth receives from the sun varies, with a periodicity of 11 years

Global ozone anomaly versus time and 4 other, related quantities

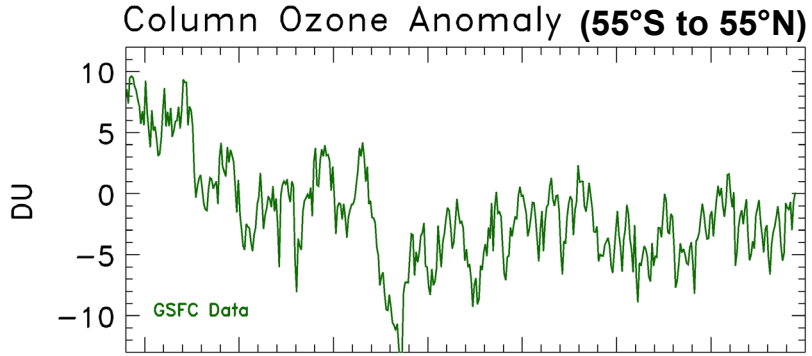
TSI : Total Solar Irradiance

The amount of energy Earth receives from the sun varies, with a periodicity of 11 years

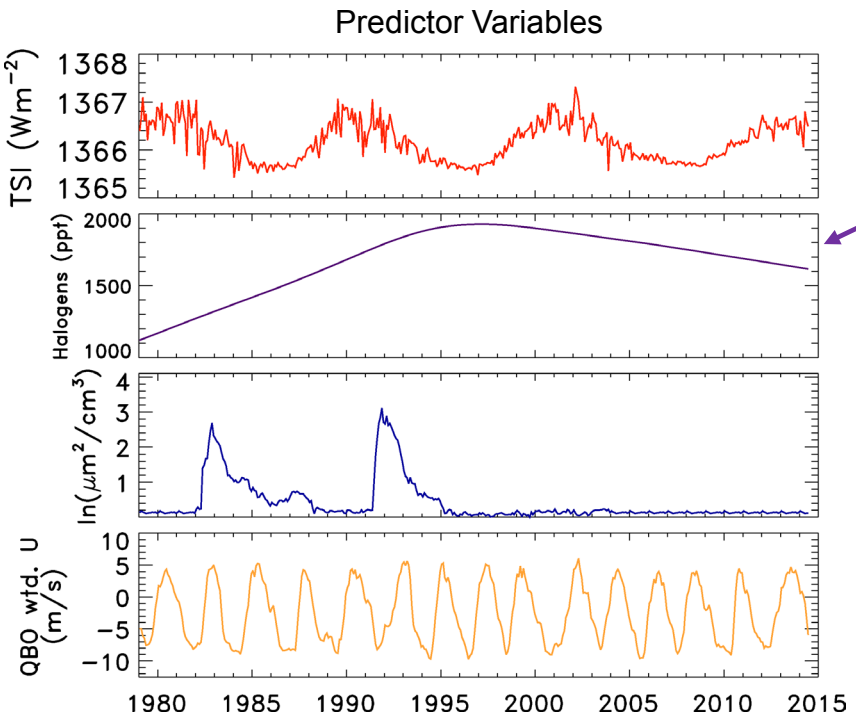


http://spot.colorado.edu/~koppg/TSI/TSI_Composite.jpg

Global ozone anomaly versus time and 4 other, related quantities

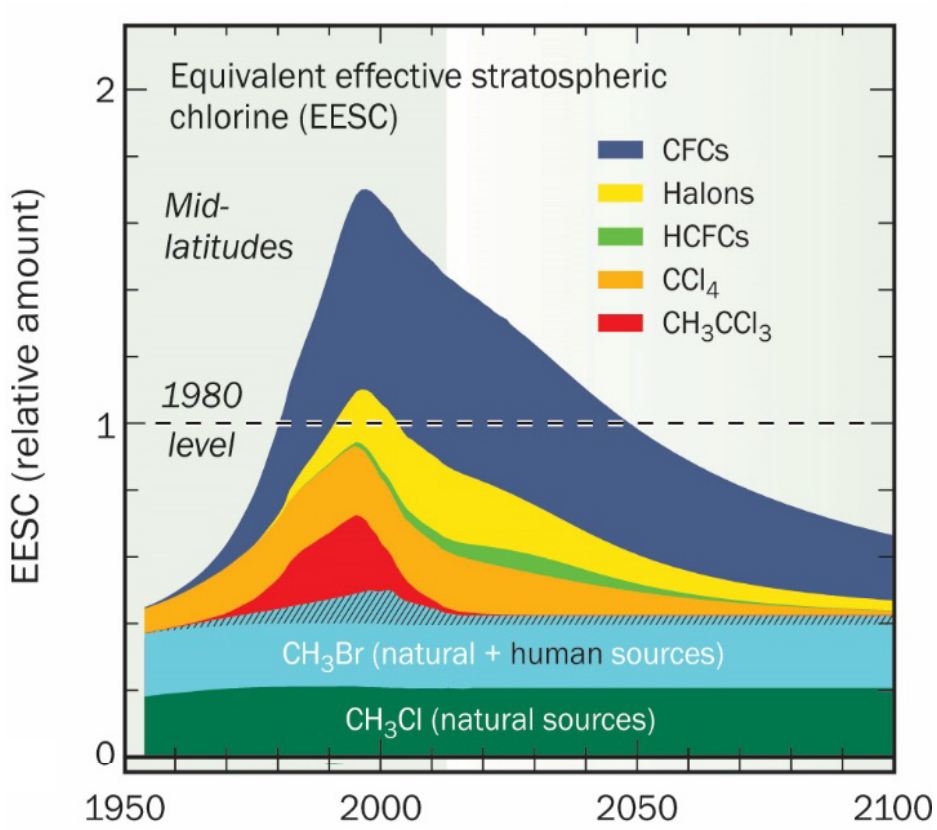


Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged



Stratospheric Halogen Loading:

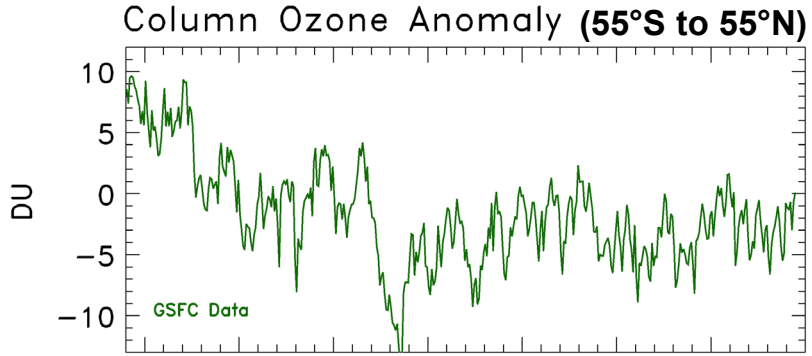
Global ozone anomaly versus time and 4 other, related quantities



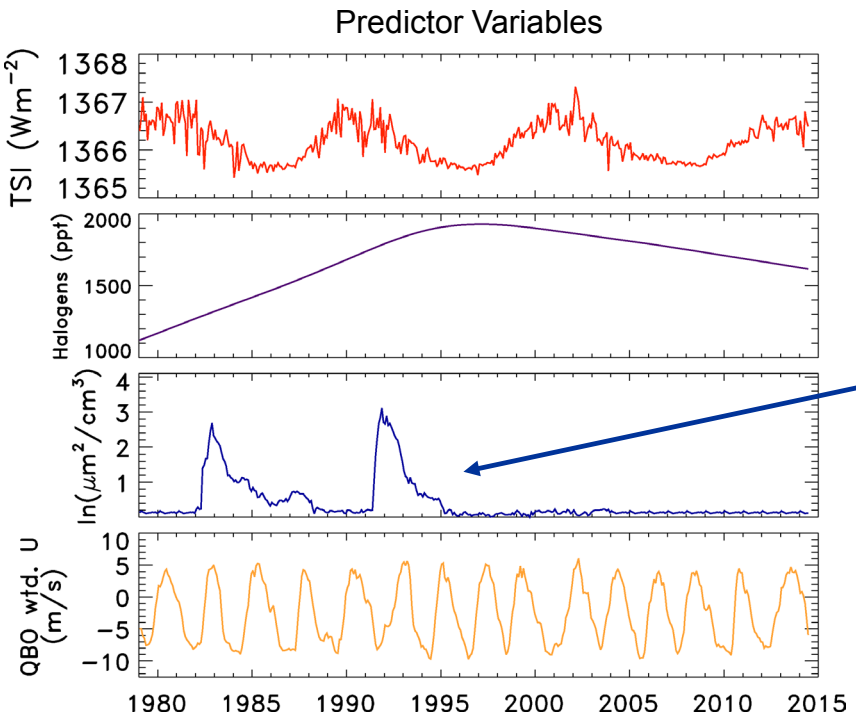
Stratospheric Halogen Loading:

http://www.esrl.noaa.gov/csd/assessments/ozone/2014/twentyquestions/images/Q16-1.jpg

Global ozone anomaly versus time and 4 other, related quantities

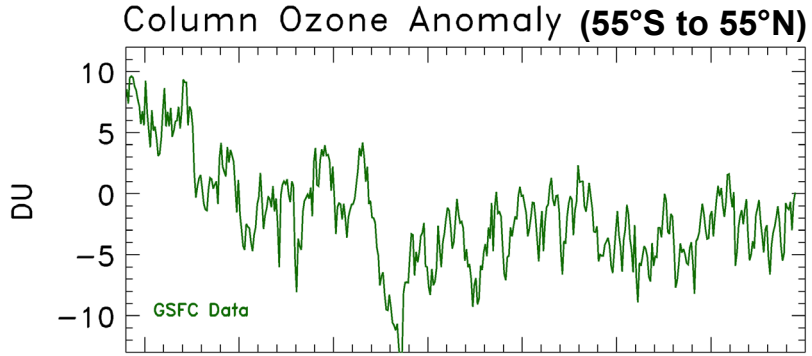


Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged

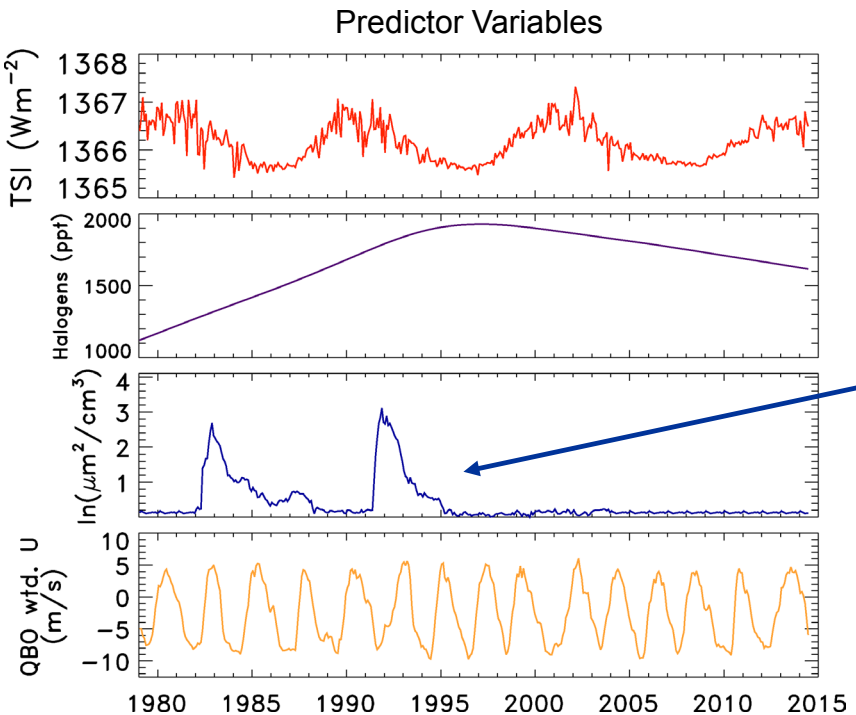


Volcanic Aerosol

Global ozone anomaly versus time and 4 other, related quantities



Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged



Volcanic Aerosol

Global ozone anomaly versus time and 4 other, related quantities

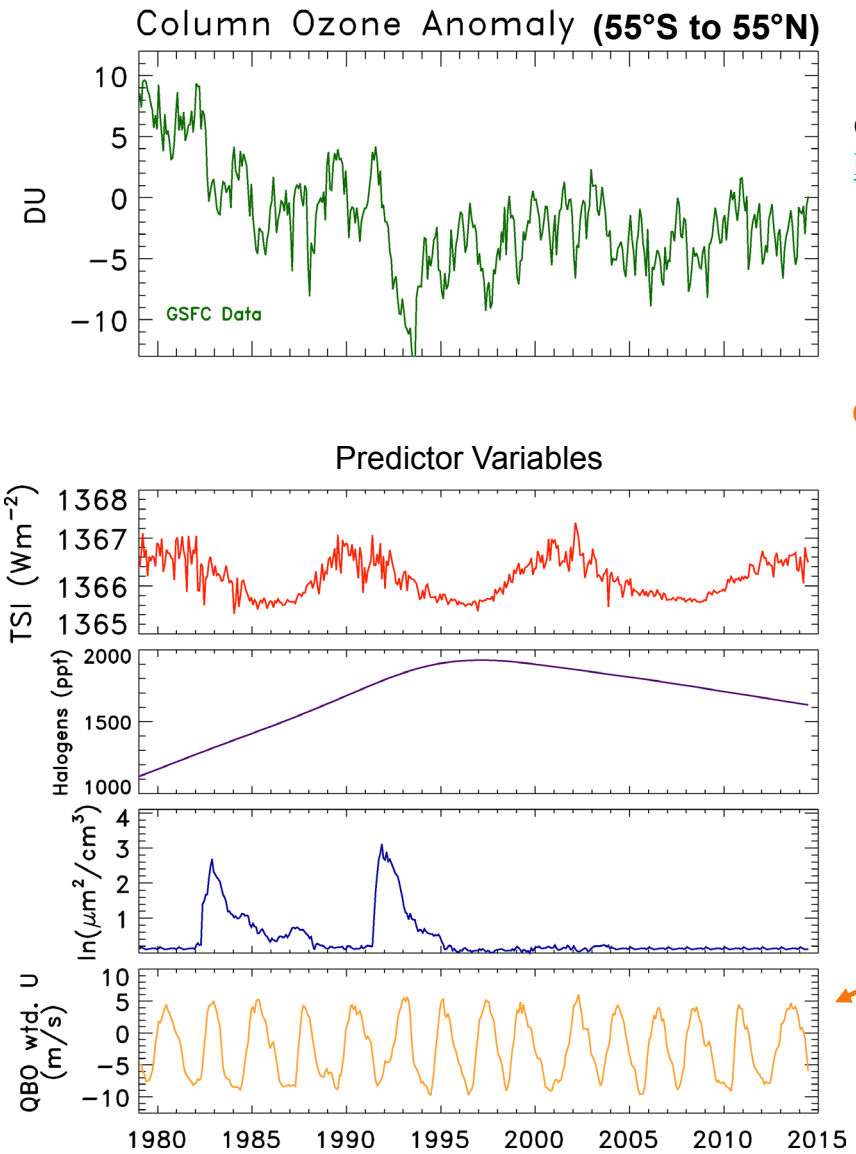
El Chichon, March 1982

https://sites.google.com/site/hesbearcat/fuego.jpg

Mt Pinatubo, June 1991

http://images.publicradio.org/content/2008/01/30/20080130_mount_pinatubo_23.jpg

Global ozone anomaly versus time and 4 other, related quantities



Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged

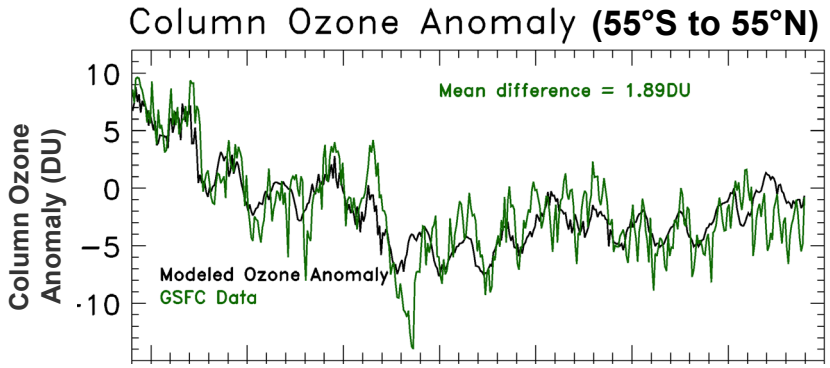
QBO : Quasi-Biennial Oscillation of direction of zonal wind in the tropical lower stratosphere

The zonal wind varies from prevailing easterlies to prevailing westerlies every ~2 yrs

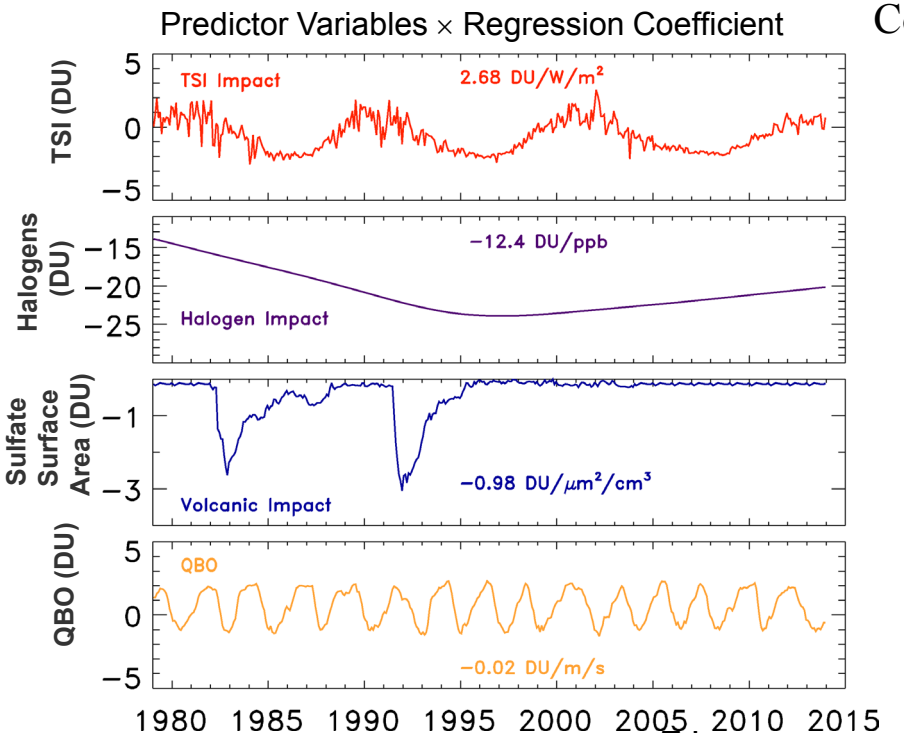
The direction of the wind affects propagation of waves and transport of ozone from source region (tropics) to accumulation region (mid-lats)

Many hundreds of papers have examined theory of QBO, effect of QBO on O_3, and generation of QBO in models

Multiple Linear Regression of Global Total Column Ozone Anomaly



Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged

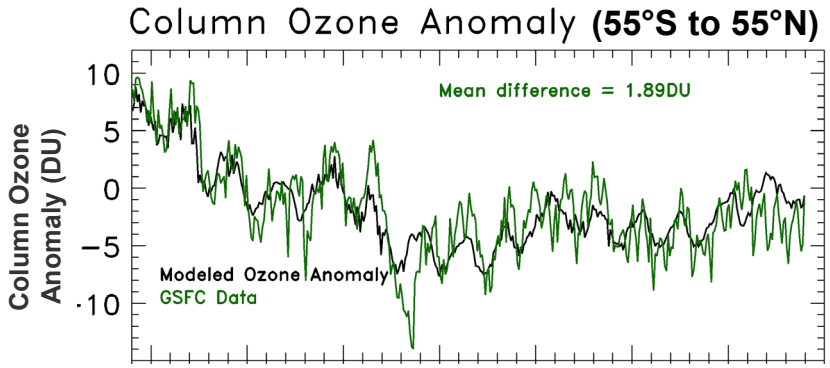


Column Ozone Anomaly (DU) =

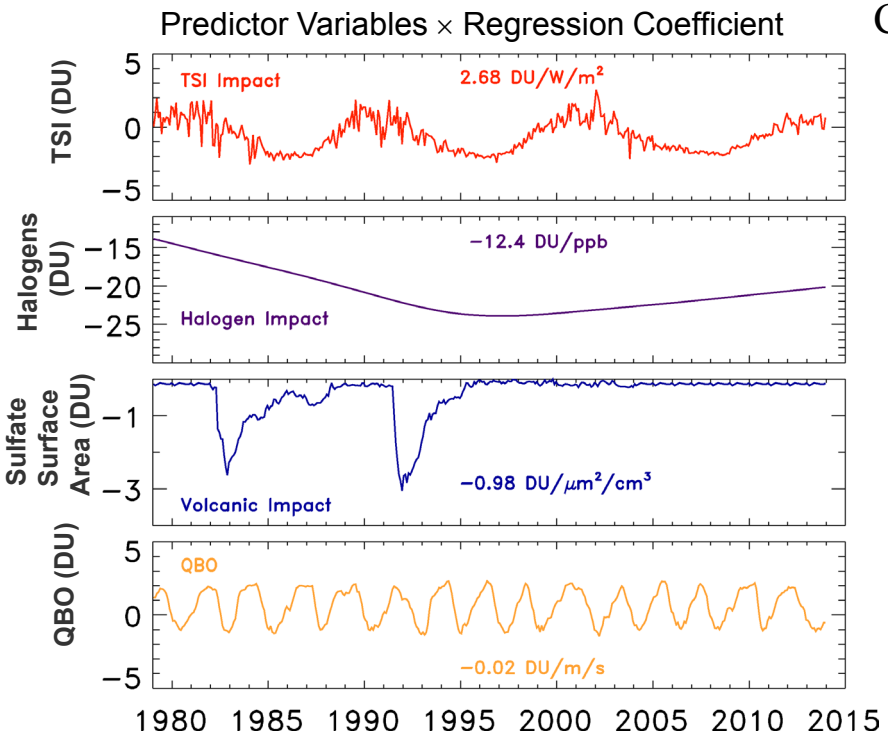
19.5	DU +	
2.68	DU / W m^{-2}	× TSI +
−12.4	DU / ppb	× Halogens +
−0.98	DU	× ln (SSA) +
−0.20	DU / m s^{-1}	× QBO

where
TSI = total solar irradiance
Halogens=stratospheric chlorine & bromine loading
SSA = Sulfate Surface Area
QBO = Quasi-biennial oscillation of the direction of winds in the tropical lower strat

Multiple Linear Regression of Global Total Column Ozone Anomaly



Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged



Regression Coefficients

Predictor Variables

Column Ozone Anomaly (DU) =

19.5	DU	+
2.68	DU / W m^{-2}	× TSI +
−12.4	DU / ppb	× Halogens +
−0.98	DU	× ln (SSA) +
−0.20	DU / m s^{-1}	× QBO

where

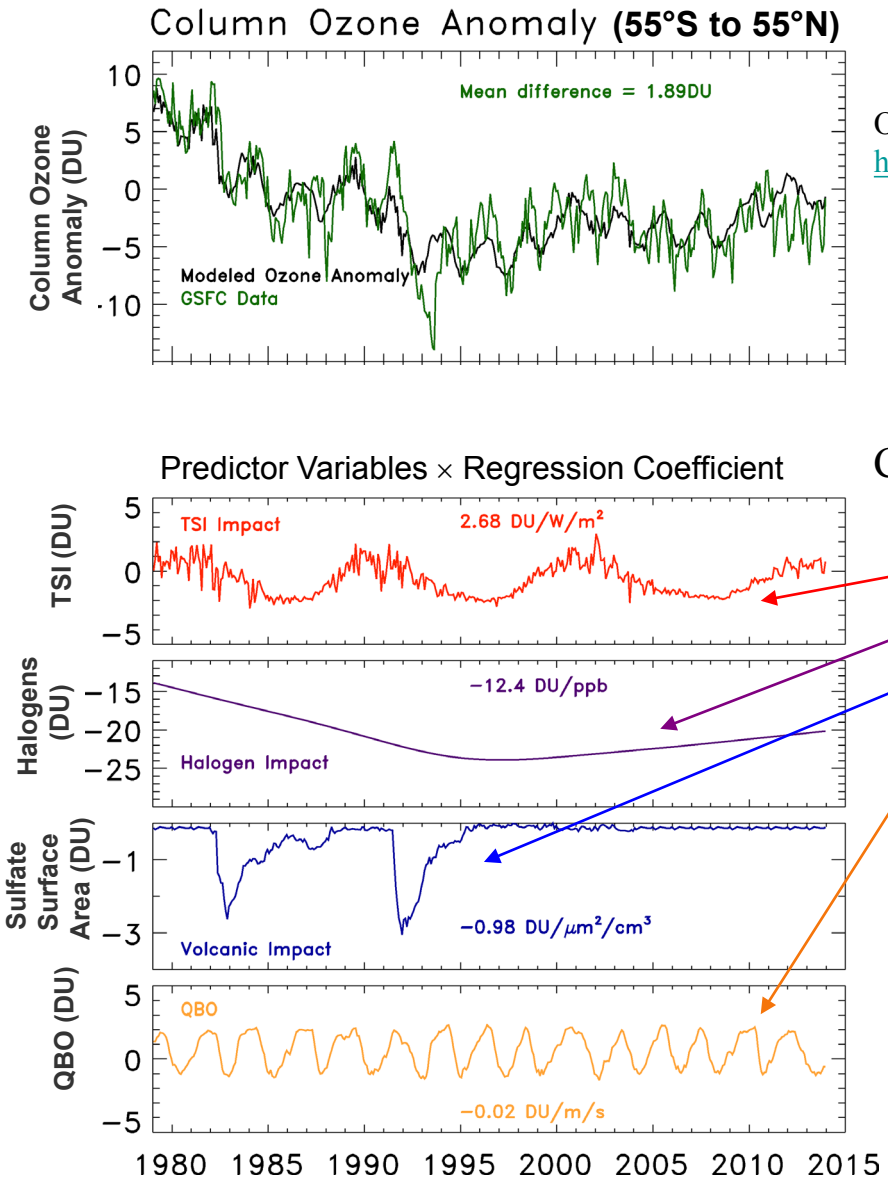
TSI = total solar irradiance

Halogens=stratospheric chlorine & bromine loading

SSA = Sulfate Surface Area

QBO = Quasi-biennial oscillation of the direction of winds in the tropical lower strat

Multiple Linear Regression of Global Total Column Ozone Anomaly



Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged

Regression Coefficients

Predictor Variables

Column Ozone Anomaly (DU) =

19.5	DU	+
2.68	DU / W m^{-2}	× TSI +
−12.4	DU / ppb	× Halogens +
−0.98	DU	× ln (SSA) +
−0.20	DU / m s^{-1}	× QBO

where

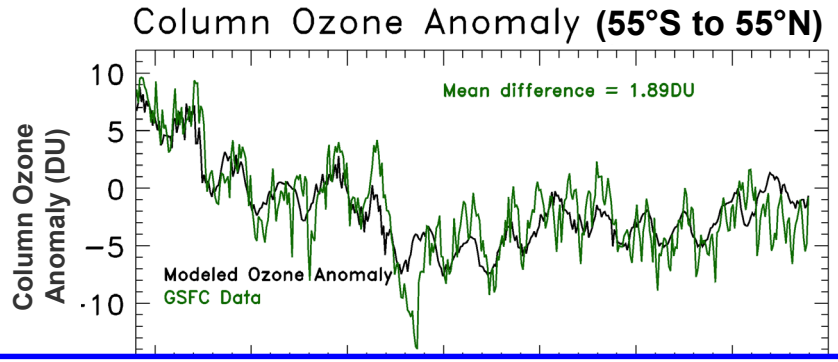
TSI = total solar irradiance

Halogens=stratospheric chlorine & bromine loading

SSA = Sulfate Surface Area

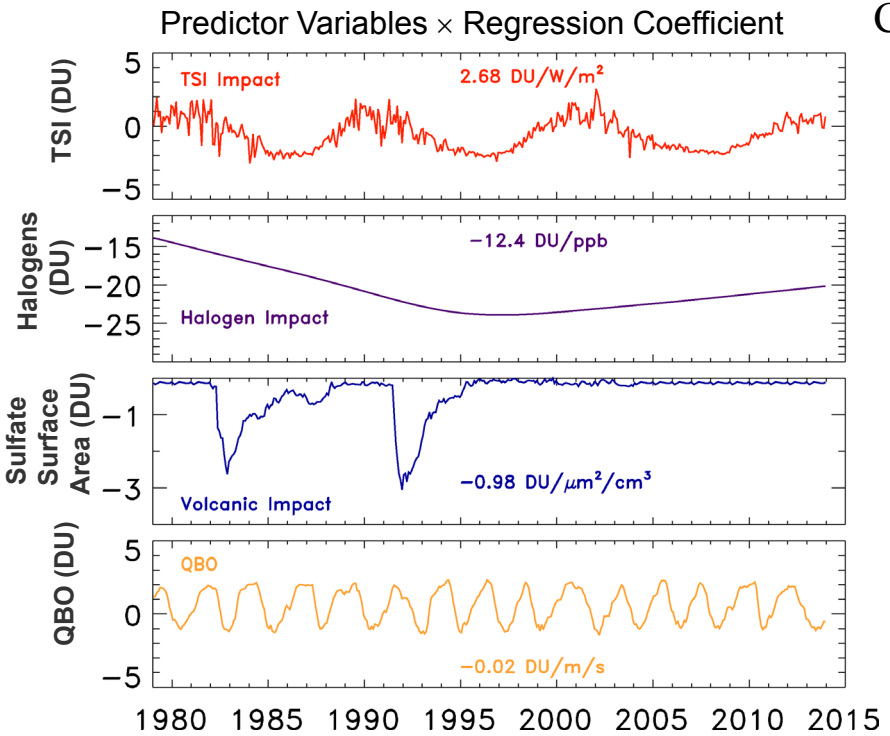
QBO = Quasi-biennial oscillation of the direction of winds in the tropical lower strat

Multiple Linear Regression of Global Total Column Ozone Anomaly



Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged

What other quantity should we examine ?

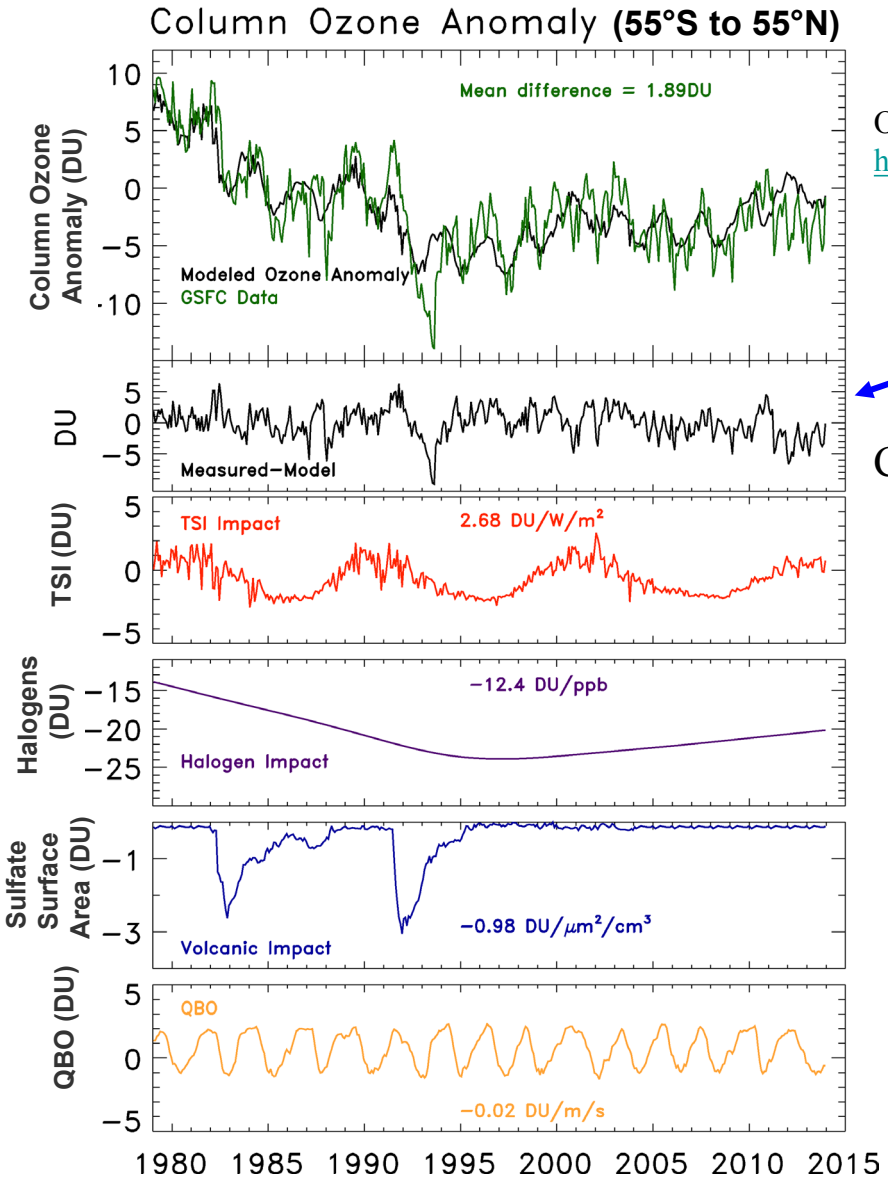


Column Ozone Anomaly (DU) =

19.5	DU	+
2.68	DU / W m^{-2}	× TSI +
−12.4	DU / ppb	× Halogens +
−0.98	DU	× ln (SSA) +
−0.20	DU / m s^{-1}	× QBO

where
TSI = total solar irradiance
Halogens=stratospheric chlorine & bromine loading
SSA = Sulfate Surface Area
QBO = Quasi-biennial oscillation of the direction of winds in the tropical lower strat

Global ozone anomaly versus time and 4 other, related quantites



Ozone data from
http://acdb-ext.gsfc.nasa.gov/Data_services/merged

The residual !

Column Ozone Anomaly (DU) =

19.5	DU +	
2.68	DU / W m^{-2}	× TSI +
−12.4	DU / ppb	× Halogens +
−0.98	DU	× ln (SSA) +
−0.20	DU / m s^{-1}	× QBO

where

TSI = total solar irradiance

Halogens=stratospheric chlorine & bromine loading

SSA = Sulfate Surface Area

QBO = Quasi-biennial oscillation of the direction of winds in the tropical lower strat

AOSC 652: Analysis Methods in AOSC

Multiple Linear Regression

$$\overline{\overline{A}} \times \vec{c} = \vec{b} \qquad \text{or} \qquad \vec{c} = \overline{\overline{A}}^{-1} \times \vec{b}$$

What are some of the mathematical concerns one must address when conducting regression analysis ?

What is the difference between multiple linear regression and multivariate regression?

AOSC 652: Analysis Methods in AOSC

Multiple Linear Regression

$$\overline{\overline{A}} \times \vec{c} = \vec{b} \qquad \text{or} \qquad \vec{c} = \overline{\overline{A}}^{-1} \times \vec{b}$$

What is the difference between multiple linear regression and multivariate regression?

"Multiple linear regression" is a model for one response variable ("y") and one or more predictor variables ("X").

"Multivariate linear regression" broadens that to more than one response variable ("Y"). The idea is that the response variables may be correlated for each "observation"

For example: a set of stocks, all reacting simultaneously to market factors.

MATLAB: Multiple Linear Regression ⇒ REGRESS
Multivariate Linear Regression ⇒ MVREGRESS
See http://www.mathworks.de/matlabcentral/newsreader/view_thread/154512

IDL: Multiple Linear Regression ⇒ REGRESS
Multivariate Linear Regression ⇒ No native version; many user defined versions available on line

Multiple Linear Regression

$$\overline{\overline{A}} \times \vec{c} = \vec{b} \qquad \text{or} \qquad \vec{c} = \overline{\overline{A}}^{-1} \times \vec{b}$$

Can we perform a regression using non-linear, multiple functions ?

Sure! The math is a bit more difficult, but not too bad. See Section 10.9 of Ayyub and McCuen for a description.