Analysis Methods in Atmospheric and Oceanic Science

AOSC 652

Least Squares Analysis, Statistical Regression, and Spline Fitting: Day 1

26 Sep 2016

AOSC 652: Analysis Methods in AOSC

Assignment 4a: Integer Sorting

Only 9 or 15 students turned in code that was working: output to file fort.99 designed to *let you check* if your code worked

If you know your code is not working, please ask me, Jeff, or Walt for help

We had hoped that the answers to the functional dependence of the timing would be drawn from the readings: please try to do the readings

Review Mon:

Call to piksrt in our code:

```
call piksrt(iarray_in,iarray_out,npts)
```

New subroutine piksrt to comply with our call statement

```
      subroutine piksrt(arr_in,arr_out,n)
      integer n,i,j
      real a,arr_in(n),arr_out(n)
      do i=1,n
         arr_out(i)=arr_in(i)
      enddo
      do j=2,n                         ! Pick out each element in turn
         a=arr_out(j)
         do i=j-1,1,-1                 ! Look for the place to insert it
            if(arr_out(i).le.a) goto 10
            arr_out(i+1)=arr_out(i)
         enddo
         i=0
10       arr_out(i+1)=a                ! Insert it
      enddo
      return
      end
```

AOSC 652: Analysis Methods in AOSC

Review Mon:

Call to piksrt in our code:

```
call piksrt(iarray_in,iarray_out,npts)
```

New subroutine piksrt to comply with our call statement

```
      subroutine piksrt(arr_in,arr_out,n)
      integer n,i,j
      integer a,arr_in(n),arr_out(n)
      do i=1,n
         arr_out(i)=arr_in(i)
      enddo
      do j=2,n                    ! Pick out each element in turn
         a=arr_out(j)
         do i=j-1,1,-1            ! Look for the place to insert it
            if(arr_out(i).le.a) goto 10
            arr_out(i+1)=arr_out(i)
         enddo
         i=0
10       arr_out(i+1)=a           ! Insert it
      enddo
      return
      end
```

AOSC 652: Analysis Methods in AOSC

Assignment 4a: Integer Sorting

Only 9 or 15 students turned in code that was working: output to file fort.99 designed to *let you check* if your code worked

If you know your code is not working, please ask me, Jeff, or Walt for help

We had hoped that the answers to the functional dependence of the timing would be drawn from the readings: please try to do the readings

Reading: piksrt speed varies as N^2 ⇨ sort of 1 million points would take $10 \times 10 \times$speed of 100,000 point sort or ~100 min

AOSC 652: Analysis Methods in AOSC

Assignment 4a: Integer Sorting

Only 9 or 15 students turned in code that was working: output to file fort.99 designed to *let you check* if your code worked

If you know your code is not working, please ask me, Jeff, or Walt for help

We had hoped that the answers to the functional dependence of the timing would be drawn from the readings: please try to do the readings

Reading: piksrt speed varies as N^2 ⇨ sort of 1 million points would take $10 \times 10 \times$speed of 100,000 point sort or ~100 min

heapsort speed varies as $N \log_2 N$

Assignment 4a: Integer Sorting

Only 9 or 15 students turned in code that was working: output to file fort.99 designed to *let you check* if your code worked

If you know your code is not working, please ask me, Jeff, or Walt for help

We had hoped that the answers to the functional dependence of the timing would be drawn from the readings: please try to do the readings

Reading: piksrt speed varies as N^2 ⇨ sort of 1 million points would take $10 \times 10 \times$speed of 100,000 point sort or ~100 min

heapsort speed varies as $N \log_2 N$

Ratio of run times either:

$N^2 / N \log_2 N = 10^6 / \log_2 (10^6) =$ ~50,000 (theory)

or

100 min × 60 sec / min / 2.6 sec = ~2300 (observation)

Assignment 4a: Integer Sorting

It is remarkable that, for a little bit more effort (~30 min for most students, hopefully), we can obtain an algorithm thousands of times more efficient for sorting a million numbers than a "brute force" method

There are many times in your research life you may need to make a choice between "clock time" in getting a computation started and "run time" on the computer

Reading: piksrt speed varies as N^2 ⇨ sort of 1 million points would take $10 \times 10 \times$speed of 100,000 point sort or ~100 min

heapsort speed varies as $N \log_2 N$

Ratio of run times either:

$$N^2 / N \log_2 N = 10^6 / \log_2 (10^6) = \sim 50{,}000 \text{ (theory)}$$

or

$$100 \text{ min} \times 60 \text{ sec} / \text{min} / 2.6 \text{ sec} = \sim 2300 \text{ (observation)}$$

AOSC 652: Analysis Methods in AOSC

Please take Assignment 4B, check your numbers, and return on Wed at start of class

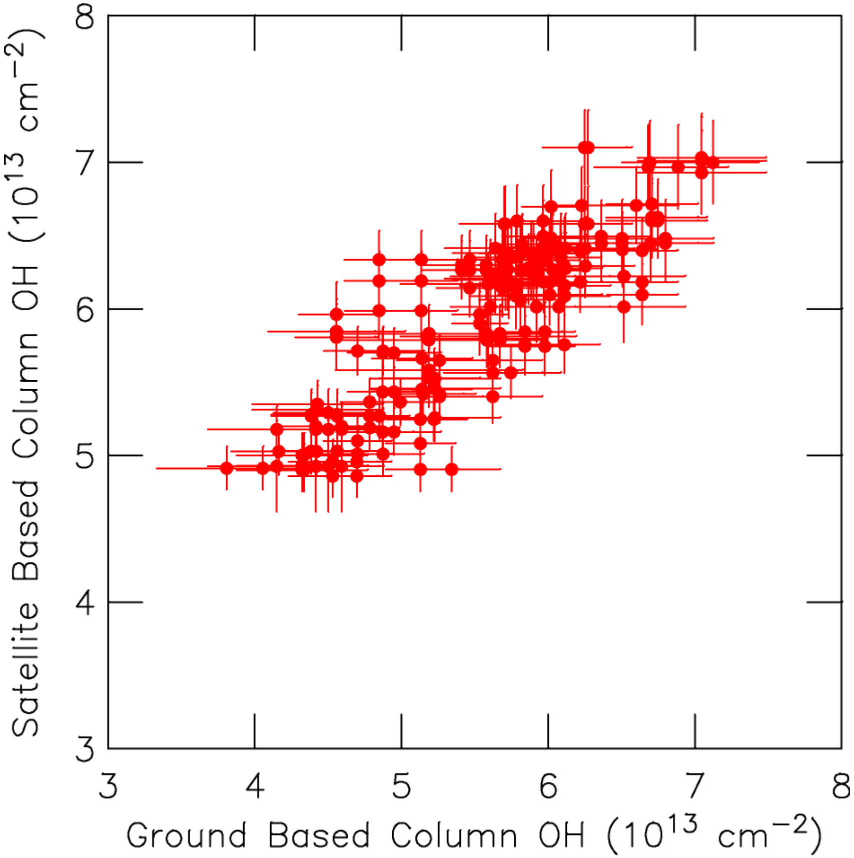
Name: ______________________________

Year	Temperature
1951	
1952	
1953	
1954	
1955	
1956	
1957	
1958	
1959	
1960	
1961	
1962	
1963	
1964	
1965	
1966	

Year	Temperature
1967	
1968	
1969	
1970	
1971	
1972	
1973	
1974	
1975	
1976	
1977	
1978	
1979	
1980	
Baseline Hand	
Baseline Code	

AOSC 652: Analysis Methods in AOSC

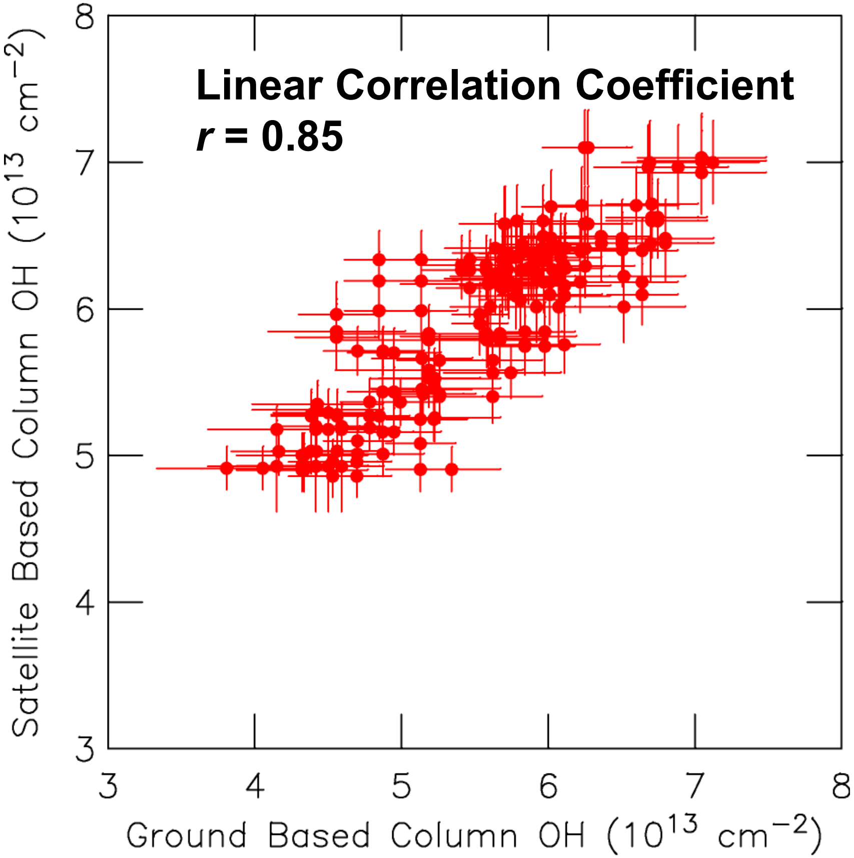
Suppose you have two sets of measurements (or data and model) that you'd like to relate.



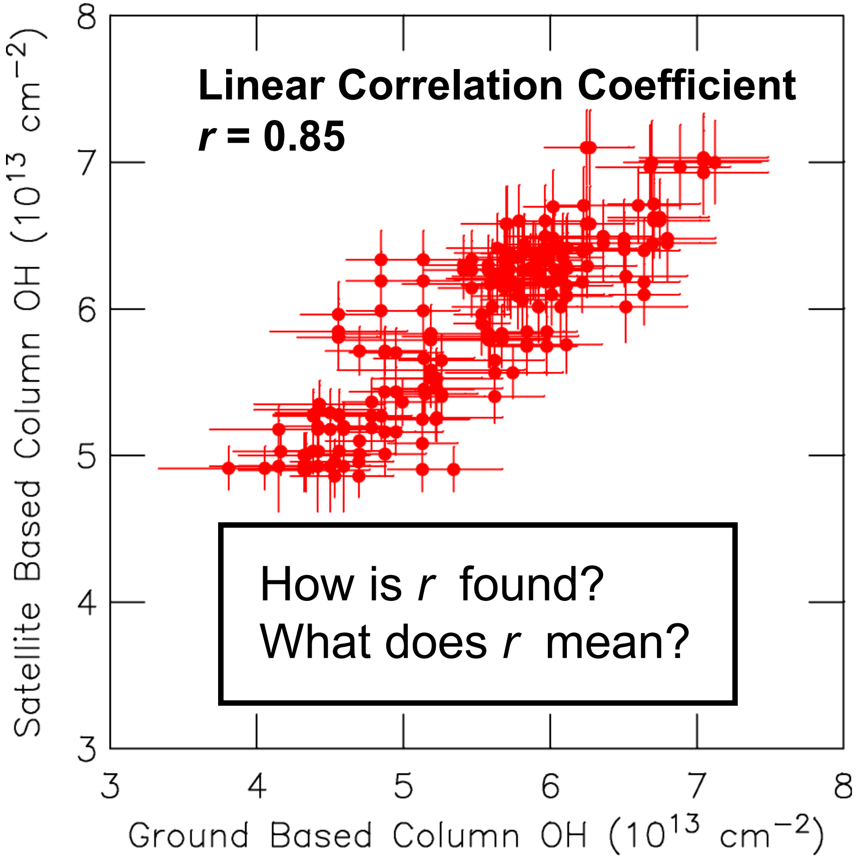
What are some aspects of the data that are typically examined?

AOSC 652: Analysis Methods in AOSC

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



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



Linear Correlation Coefficient:

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

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*

Linear Correlation Coefficient:

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

r must lie between –1 and 1

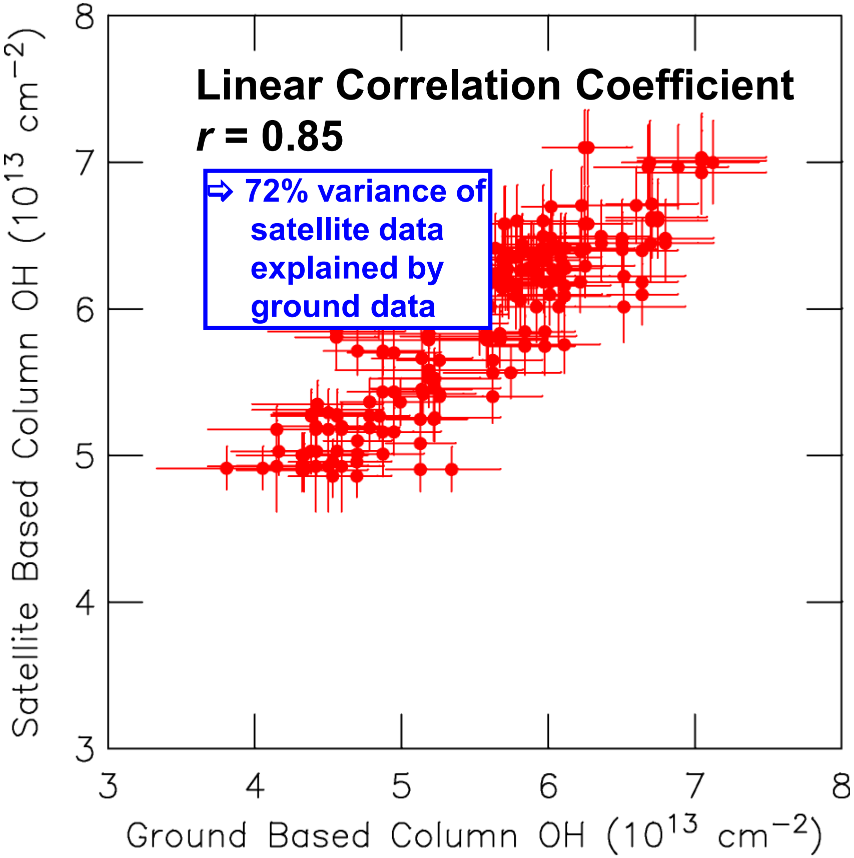
If r = 1, the data are said to have a *complete positive correlation*

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

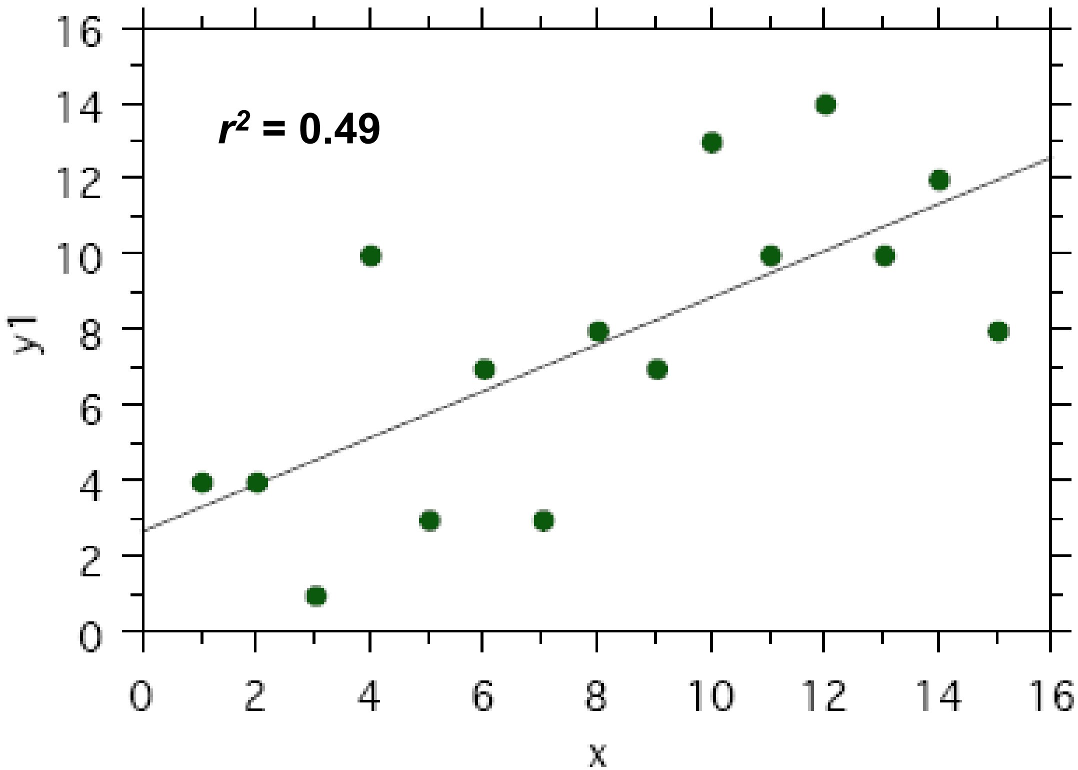
r^2 x 100 = percent of variance in common between x and y

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

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

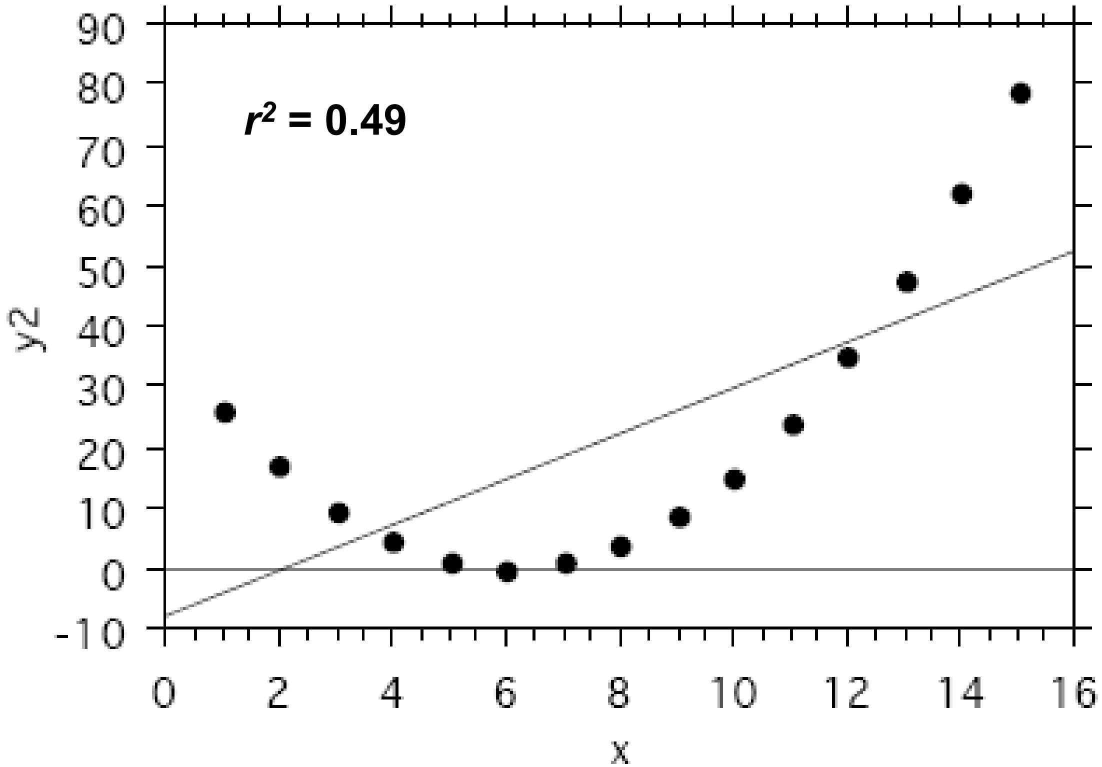


Visualization of data is vital



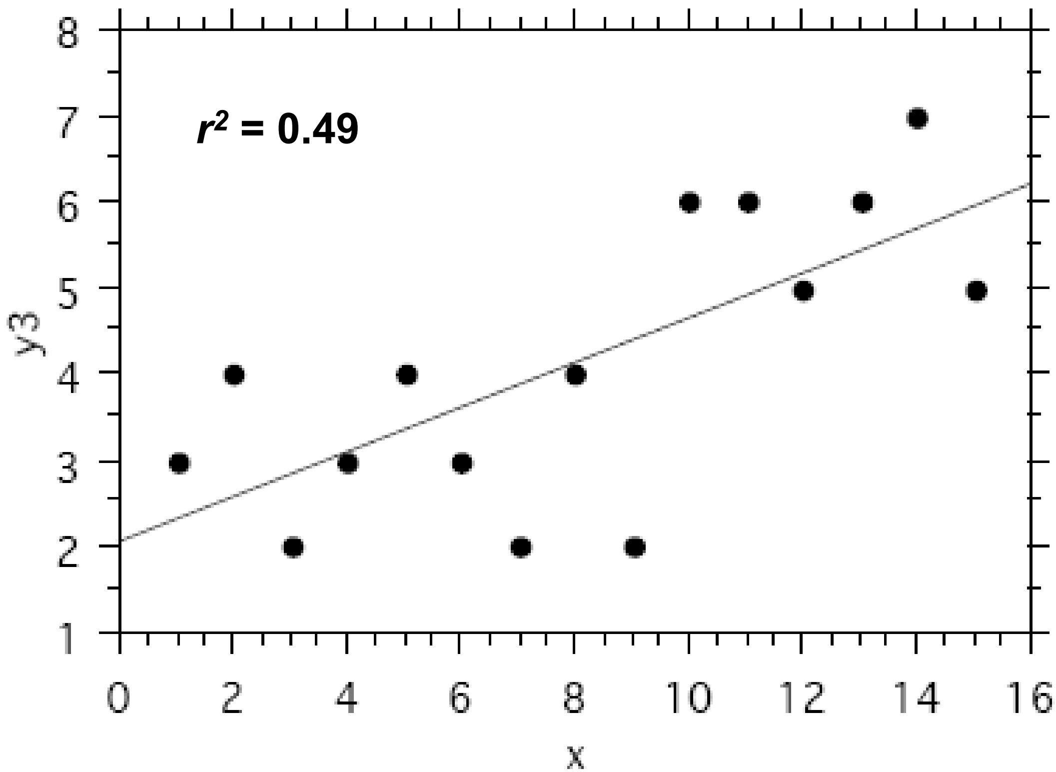
From Dennis Hartmann's class notes:
http://www.atmos.washington.edu/~dennis/552_Notes_3.pdf

Visualization of data is vital



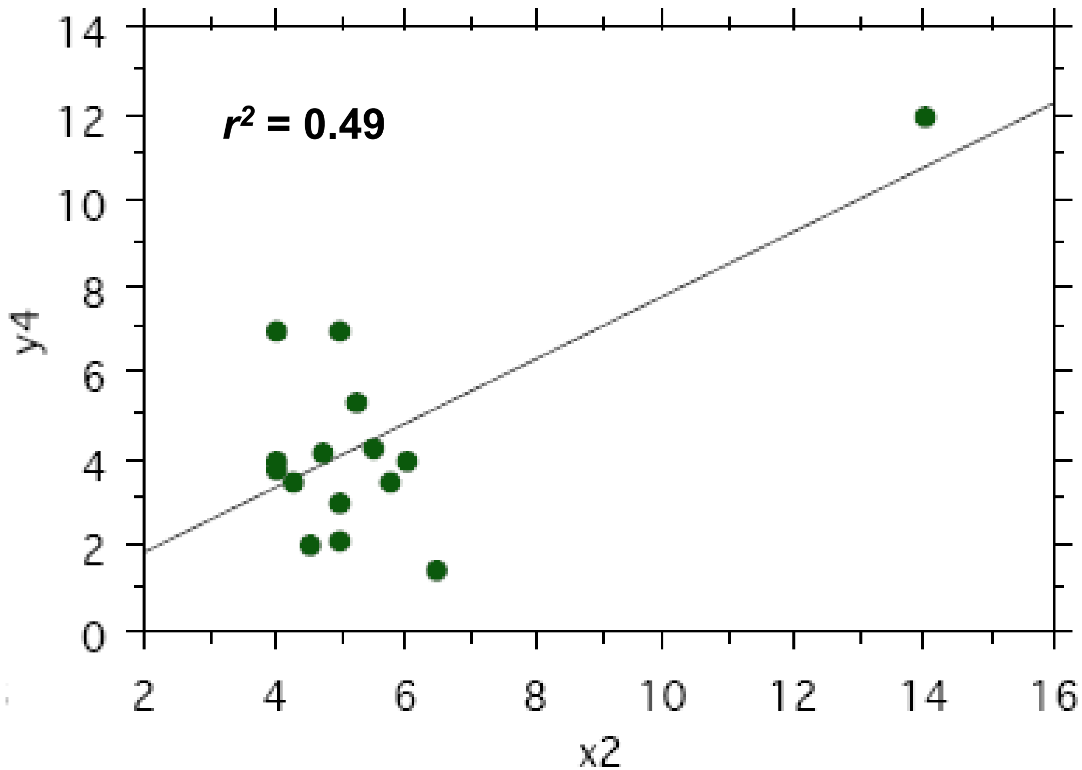
From Dennis Hartmann's class notes:
http://www.atmos.washington.edu/~dennis/552_Notes_3.pdf

Visualization of data is vital



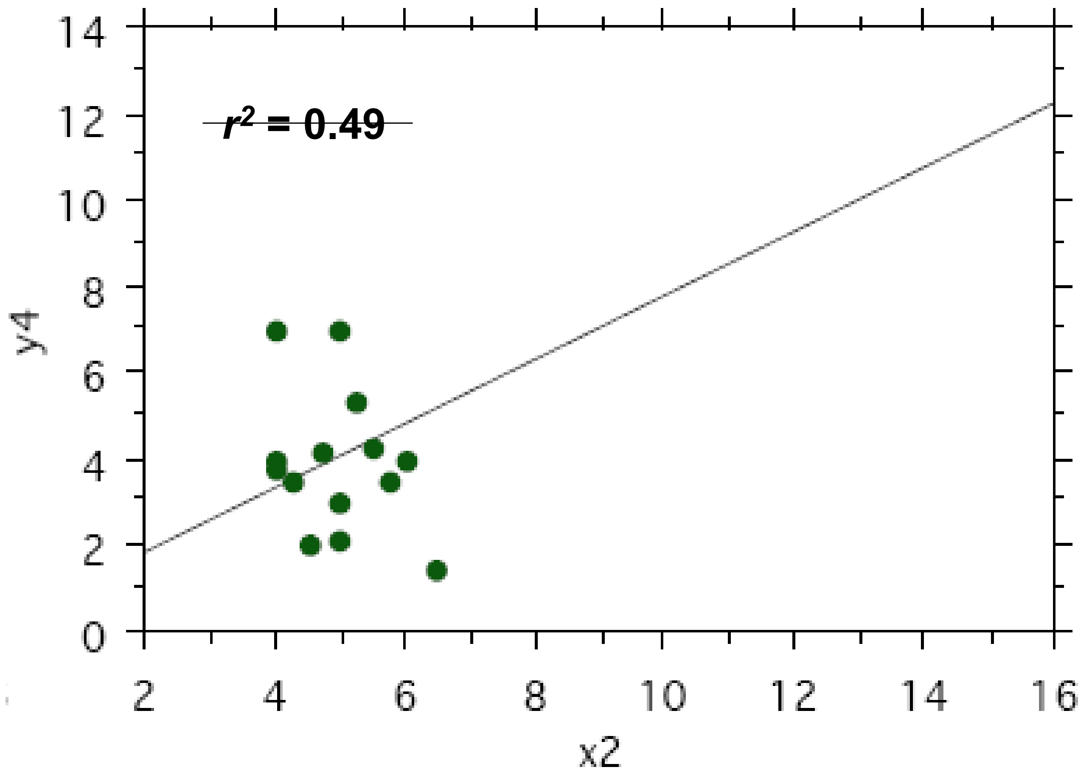
From Dennis Hartmann's class notes:
http://www.atmos.washington.edu/~dennis/552_Notes_3.pdf

Visualization of data is vital



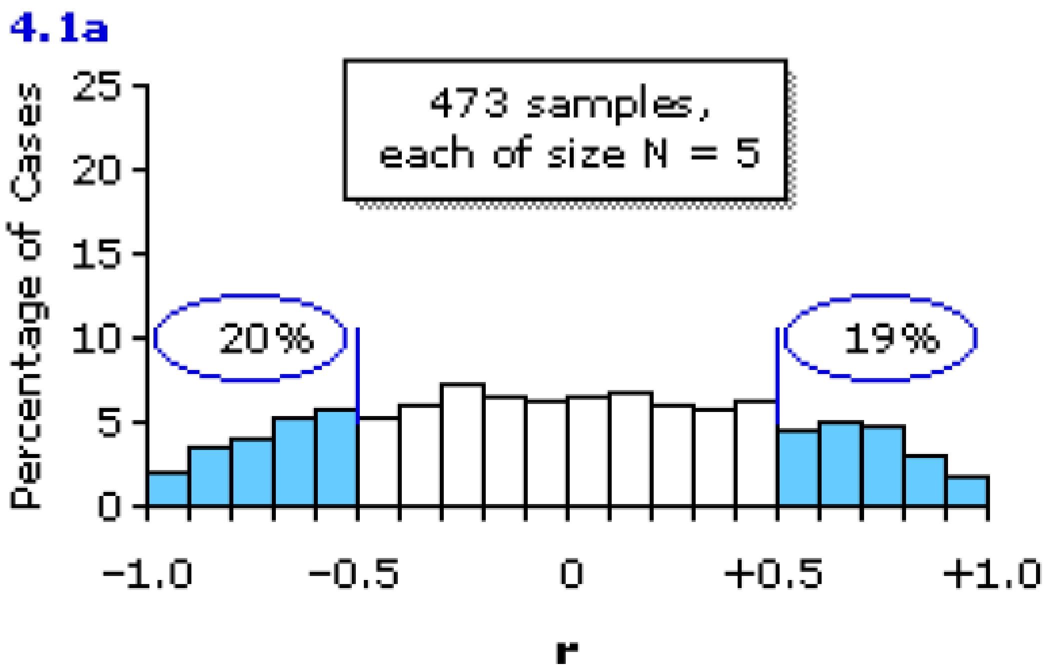
From Dennis Hartmann's class notes:
http://www.atmos.washington.edu/~dennis/552_Notes_3.pdf

Visualization of data is vital



From Dennis Hartmann's class notes:
http://www.atmos.washington.edu/~dennis/552_Notes_3.pdf

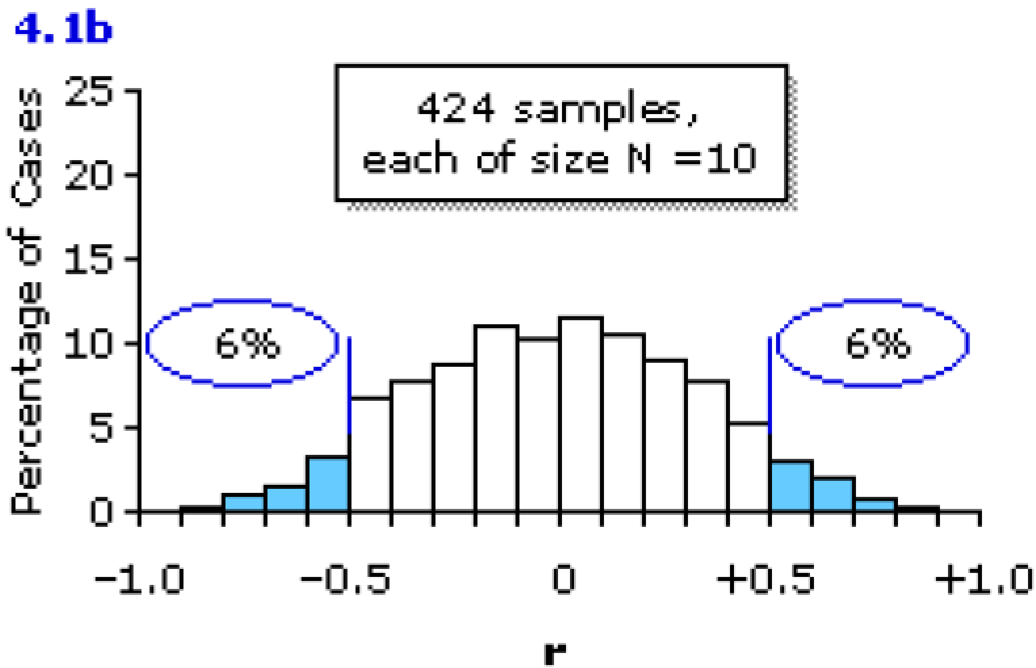
Test for statistical significance of correlation vital



Practice: Linear Correlation Coefficient for 5 throws of a pair of dice, done 473 times

From Richard Lowry's class notes:
http://www.vassarstats.net/textbook/ch4pt1.html

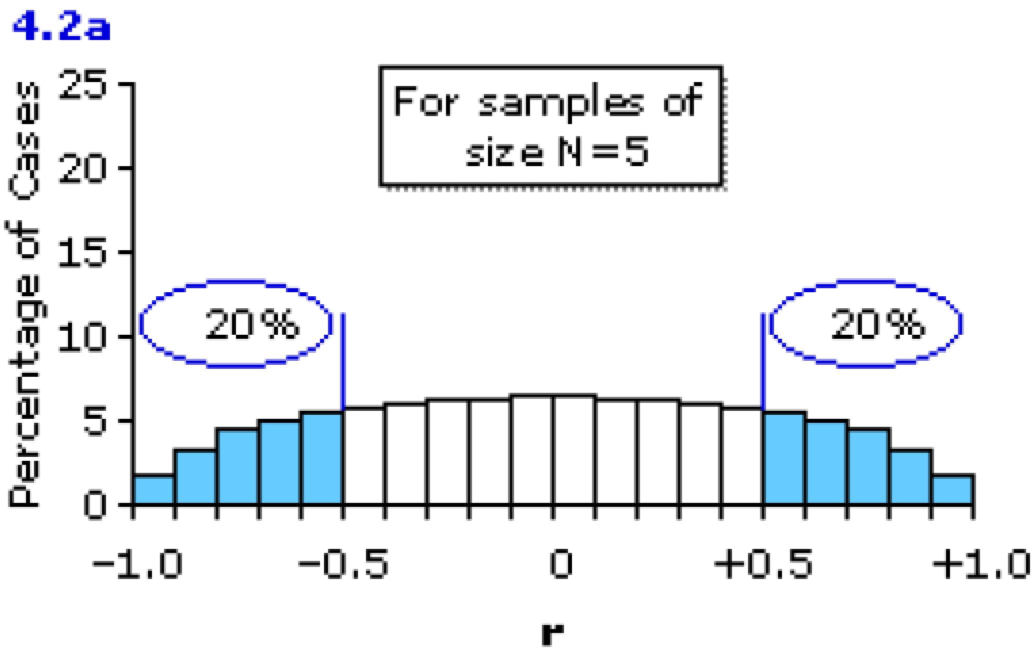
Test for statistical significance of correlation vital



Practice: Linear Correlation Coefficient for 10 throws of a pair of dice, done 424 times

From Richard Lowry's class notes:
http://www.vassarstats.net/textbook/ch4pt1.html

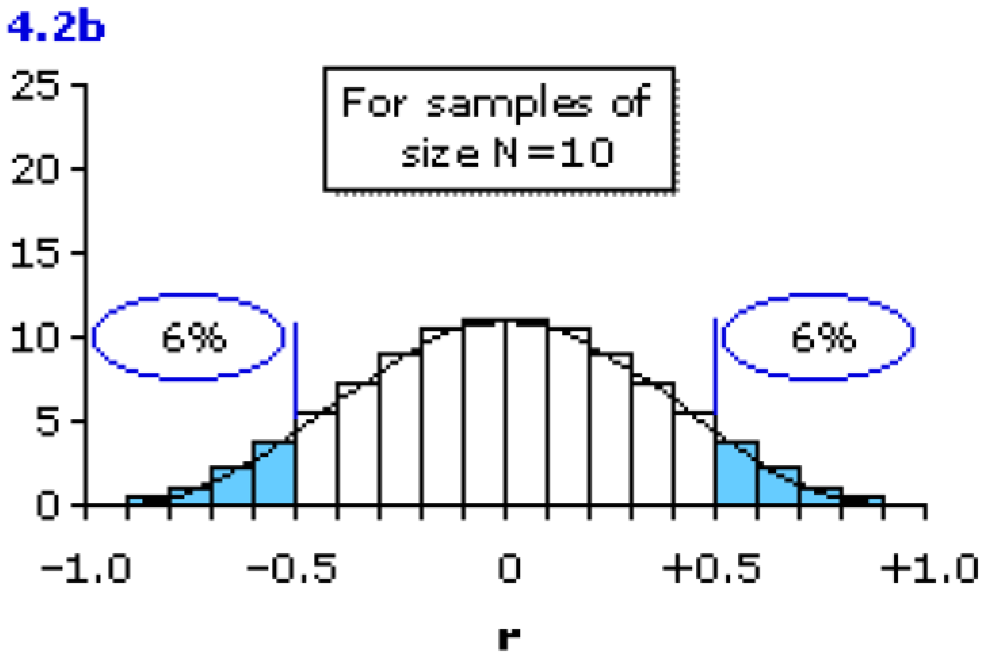
Test for statistical significance of correlation vital



Theory: Correlation Coefficient for 5 throws of a pair of dice, done infinite # of times

From Richard Lowry's class notes:
http://www.vassarstats.net/textbook/ch4pt1.html

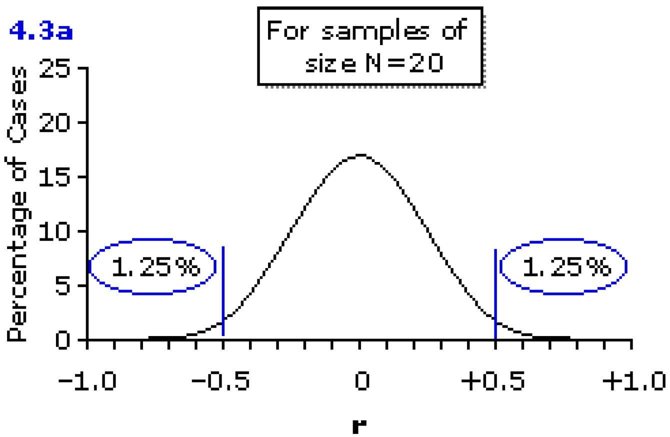
Test for statistical significance of correlation vital

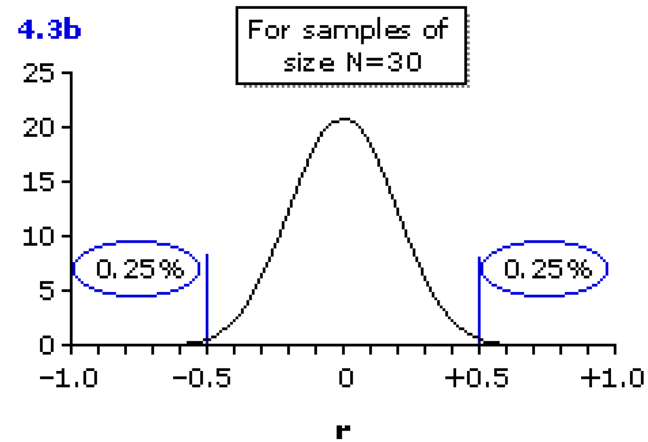


Theory: Correlation Coefficient for 10 throws of a pair of dice, done infinite # of times

From Richard Lowry's class notes:
http://www.vassarstats.net/textbook/ch4pt1.html

Test for statistical significance of correlation vital





Theory: Correlation Coefficient for 20 throws (left) and 30 throws (right) of a pair of dice (left)

From Richard Lowry's class notes:
http://www.vassarstats.net/textbook/ch4pt1.html

Test for statistical significance of correlation vital

Absolute value of r needed to show two quantities bear a statistically significant relation at the 95% confidence interval, as a function of sample size

N	Directional	Non-Directional
5	0.81	0.88

From Richard Lowry's class notes:
http://www.vassarstats.net/textbook/ch4pt1.html

Test for statistical significance of correlation vital

Absolute value of r needed to show two quantities bear a statistically significant relation at the 95% confidence interval, as a function of sample size

N	Directional	Non-Directional
5	0.81	0.88
10	0.55	0.63
15	0.44	0.51
20	0.38	0.44
30	0.30	0.35

From Richard Lowry's class notes:
http://www.vassarstats.net/textbook/ch4pt1.html

Test for statistical significance of correlation vital

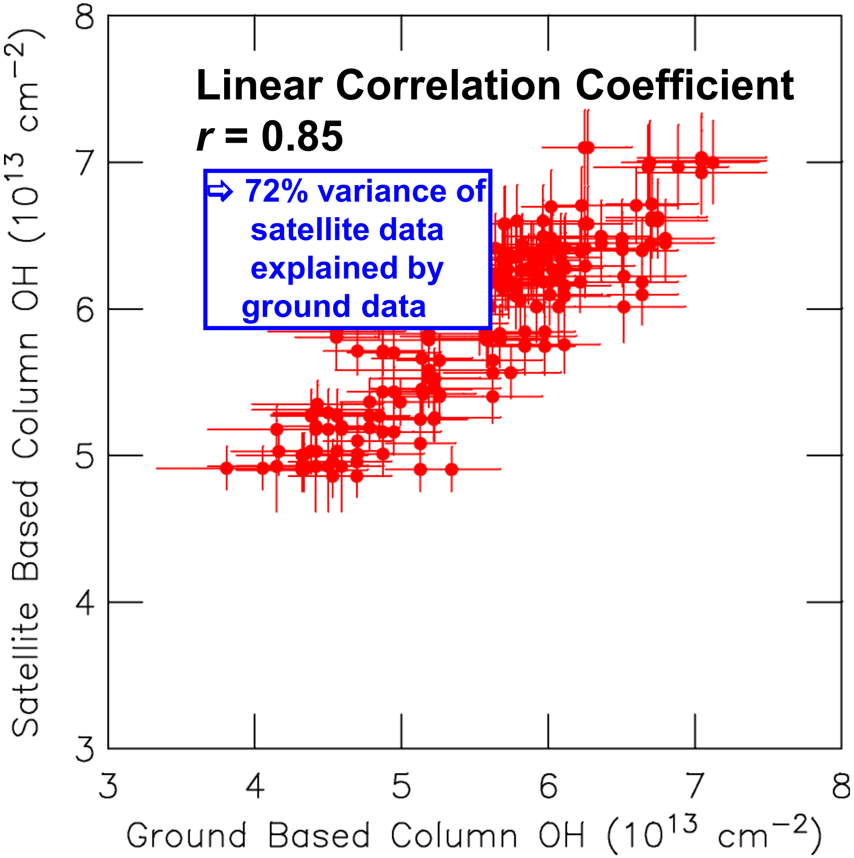
Absolute value of r needed to show two quantities bear a statistically significant relation at the 95% confidence interval, as a function of sample size

N	Directional	Non-Directional
5	0.81	0.88
10	0.55	0.63
15	0.44	0.51
20	0.38	0.44
30	0.30	0.35

See Section 8.4 of von Storch and Zwiers, *Statistical Analysis of Climate Research* for methodology for assessing statistical significance of a linear correlation

http://www.leif.org/EOS/vonSt0521012309.pdf

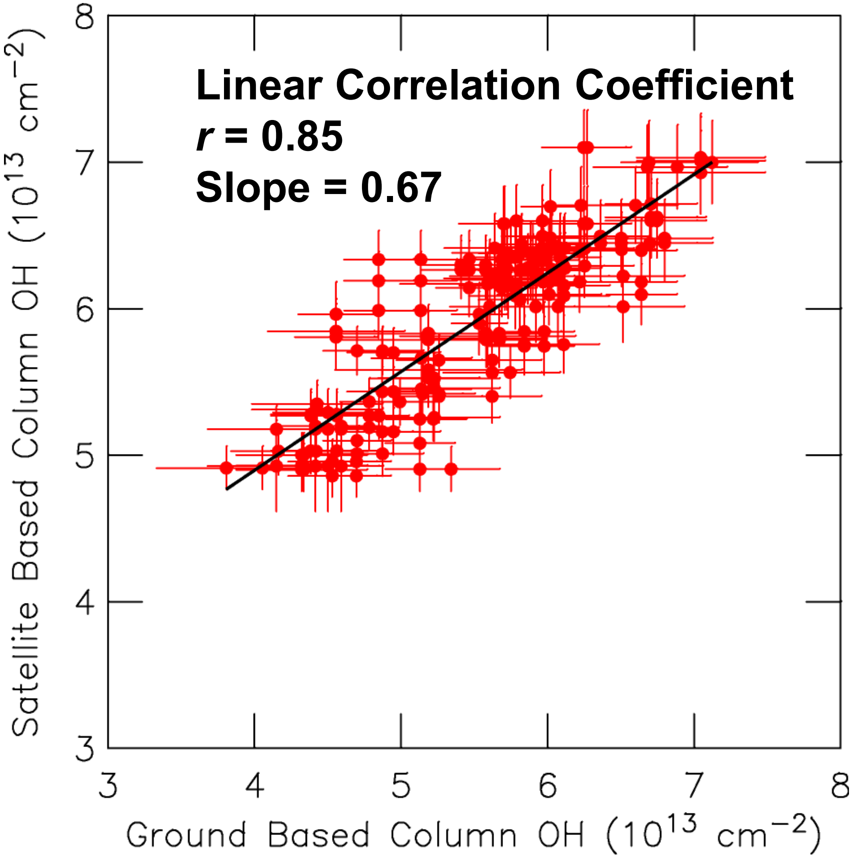
Suppose you have two sets of measurements (or data and model) that you'd like to relate.



What else might we want to know?

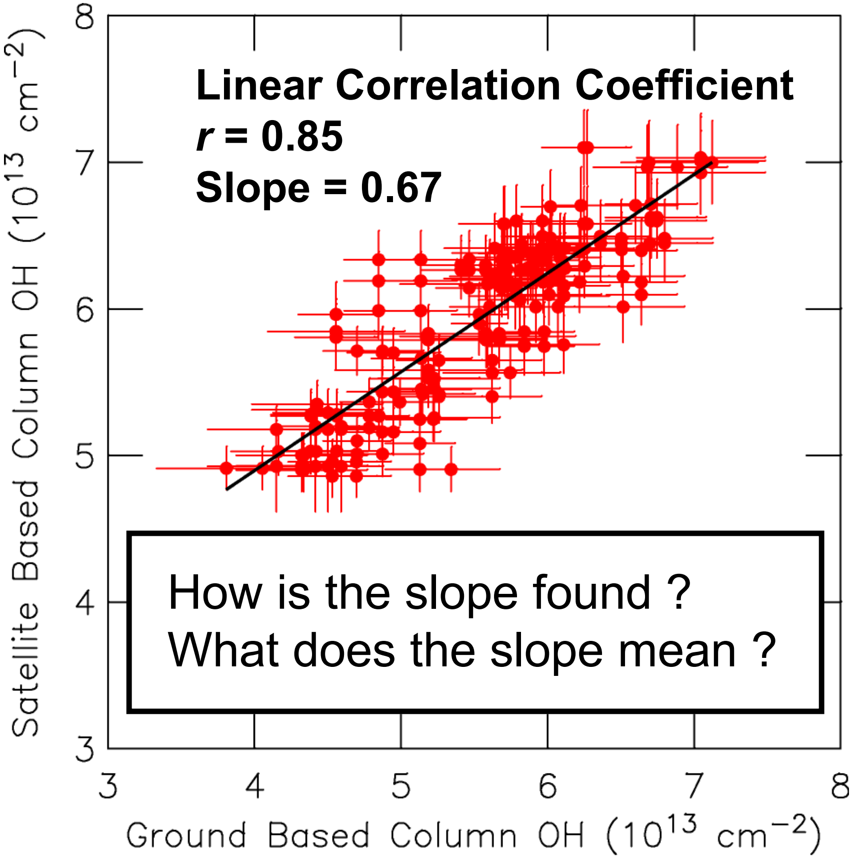
AOSC 652: Analysis Methods in AOSC

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



AOSC 652: Analysis Methods in AOSC

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



Linear Least Squares Fitting:

$$y = a + b\,x$$

Minimize:

$$\sum_{i=1}^{N} (a + b\,x_i - y_i)^2 \equiv Cost\ Function$$

$$\frac{\partial\ Cost\ Function}{\partial a} = 2\sum_{i=1}^{N}(a + b\,x_i - y_i) = 0$$

$$\frac{\partial\ Cost\ Function}{\partial b} = 2\sum_{i=1}^{N}(a + b\,x_i - y_i)\,x_i = 0$$

$$\frac{\partial\ Cost\ Function}{\partial a} = 2\sum_{i=1}^{N}(a + b\ x_i - \ y_i) = \ 0$$

$$\frac{\partial\ Cost\ Function}{\partial b} = 2\sum_{i=1}^{N}(a + b\ x_i - \ y_i)\ x_i = \ 0$$

$$a\ N + b\ \Sigma x_i - \ \Sigma y_i = \ 0$$

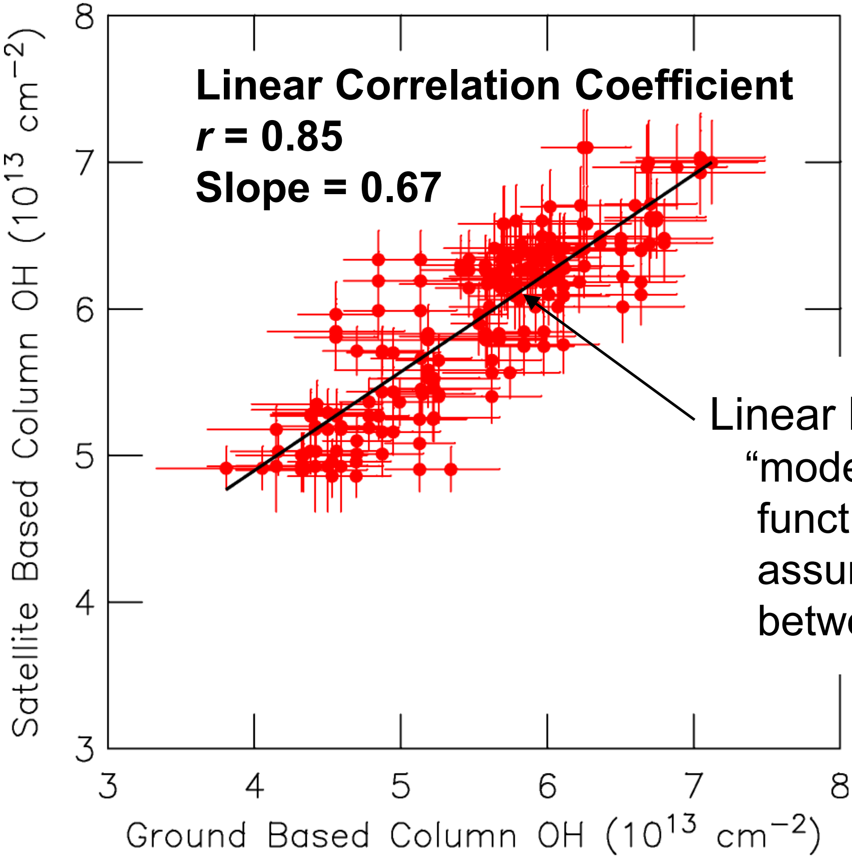
$$a\ \Sigma\ x_i + b\ \Sigma\ x^2_i - \ \Sigma\ x_i \cdot y_i = \ 0$$

Can show:

$$b = \frac{\Sigma x_i \cdot \Sigma y_i - \ N \cdot \Sigma x_i y_i}{(\Sigma x_i)^2 - \ N \cdot \Sigma x_i^2}$$ ⇦ Slope of linear least squares fit

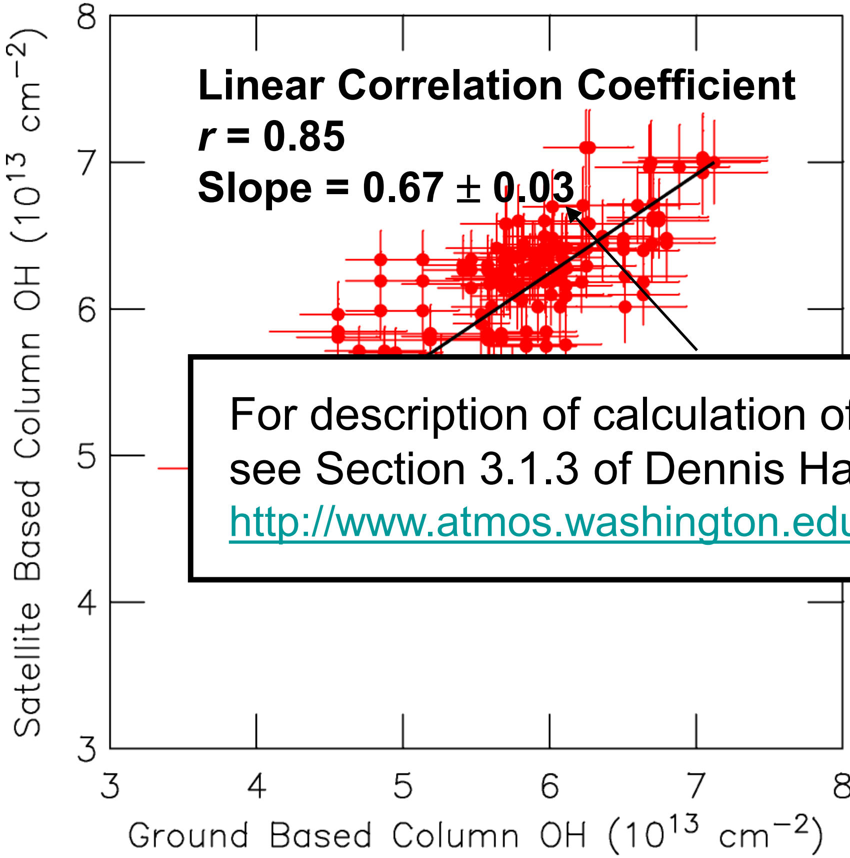
$$a = \frac{\Sigma y_i - \ b \cdot \Sigma x_i}{N}$$ ⇦ Intercept of linear least squares fit

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



Linear least squares fit:
"model" of satellite data as a function of ground-based data, assuming a linear relation between the two quantities

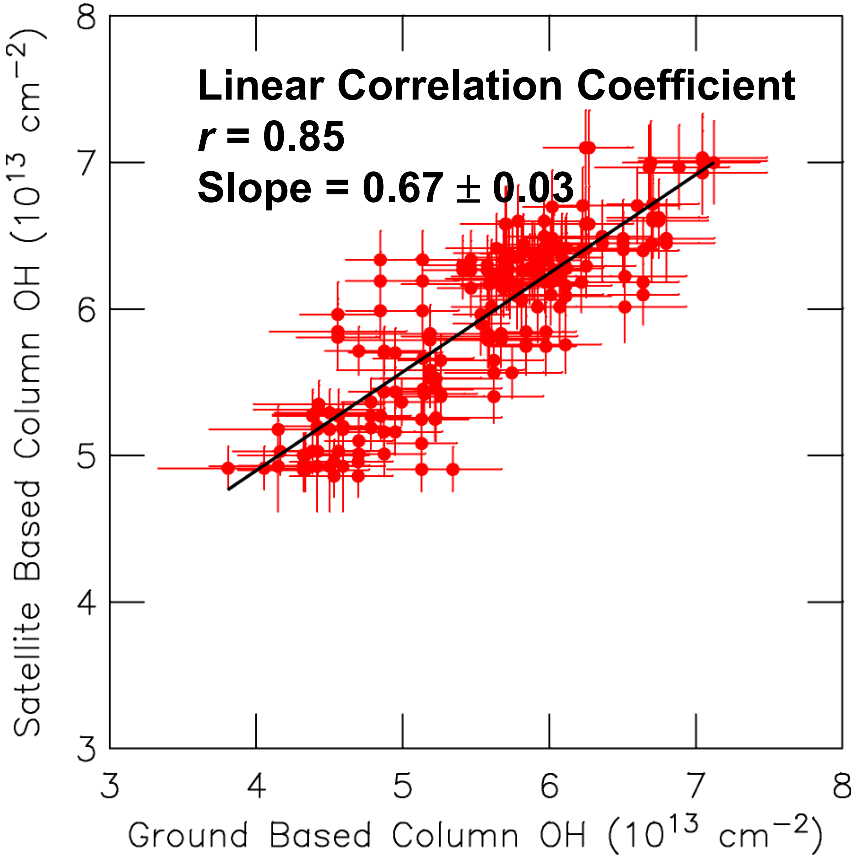
Suppose you have two sets of measurements (or data and model) that you'd like to relate.



For description of calculation of the uncertainty of slope see Section 3.1.3 of Dennis Hartmann's class notes:
http://www.atmos.washington.edu/~dennis/552_Notes_3.pdf

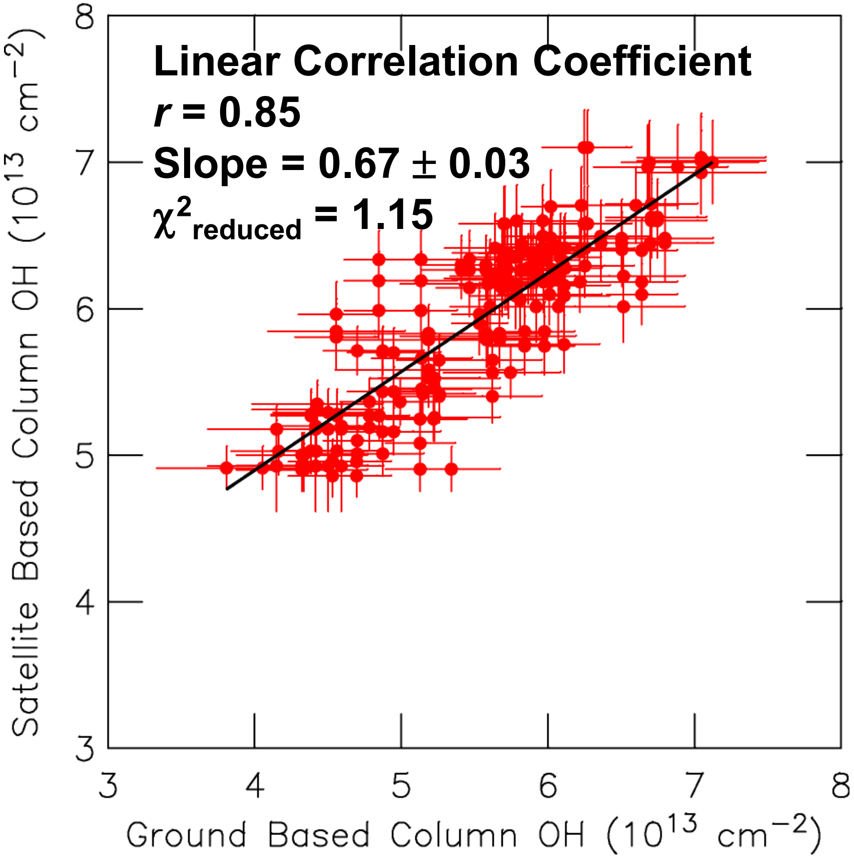
AOSC 652: Analysis Methods in AOSC

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



How can we quantify the "goodness of the fit" ?

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



Reduced χ^2 :

$$\chi^2_{\text{Fit}} = \frac{1}{\nu} \sum_{i=1}^{N} \left(\frac{y_i - Fit(x_i)}{\sigma_i} \right)^2$$

where σ_i is the uncertainty associated with each measurement and

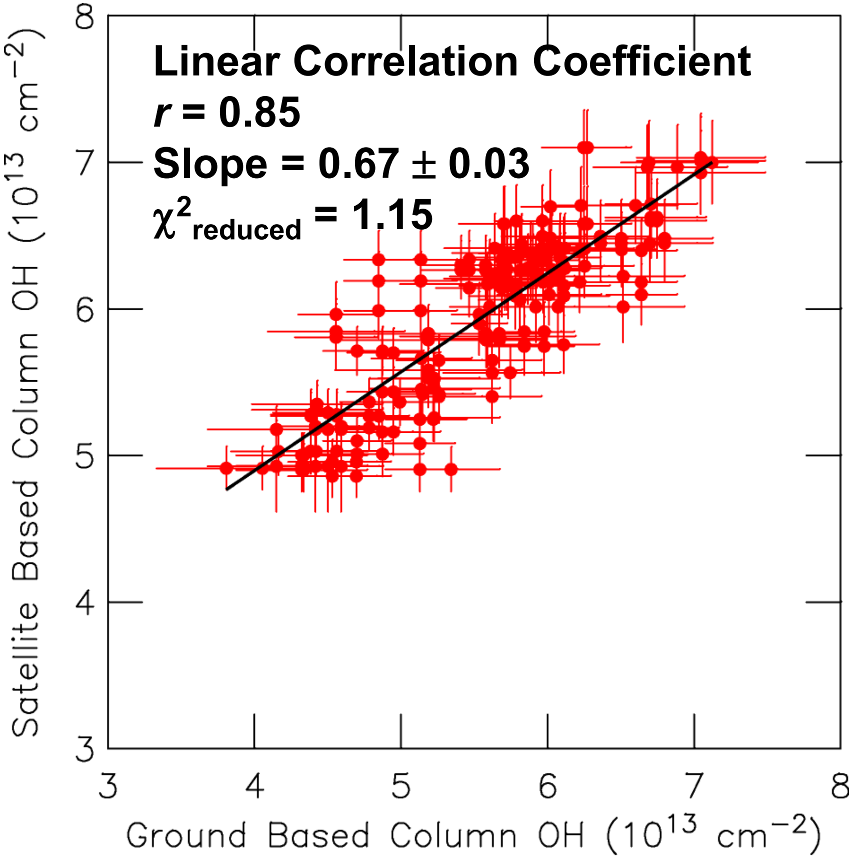
$\nu = \text{N} - p - 1$

for N = number of observations

p = number of fitting parameters

Often, one sets ν equal to N

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



Reduced χ^2 :

$$\chi^2_{\text{Fit}} = \frac{1}{\nu} \sum_{i=1}^{N} \left(\frac{y_i - Fit(x_i)}{\sigma_i} \right)^2$$

where σ_i is the uncertainty associated with each measurement and

$\nu = \mathrm{N} - p - 1$

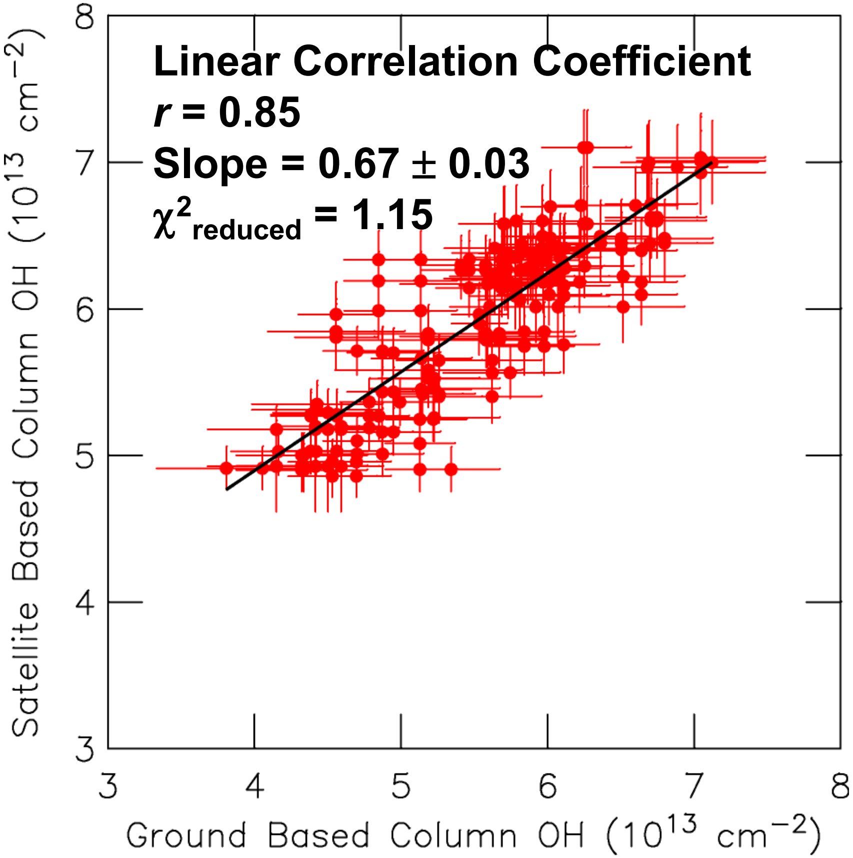
for N = number of observations

p = number of fitting parameters

Often, one sets ν equal to N

Degrees of freedom

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



Reduced χ^2 :

$$\chi^2_{\text{Fit}} = \frac{1}{\nu} \sum_{i=1}^{N} \left(\frac{y_i - Fit(x_i)}{\sigma_i} \right)^2$$

where σ_i is the uncertainty associated with each measurement and

$\nu = \text{N} - p - 1$

for N = number of observations

p = number of fitting parameters

Often, one sets ν equal to N

Reduced χ^2 is an under-used, often unappreciated, quantitative measure of the degree of relation between a model and data, or two types of measurements (or, two types of models, if we can define uncertainties for each)

Reduced χ^2 is an under-used, perhaps unappreciated quantitative measure of the degree of relation between a model and data, or two types of measurements (or, two types of models, if we can define uncertainties for each)

Reduced χ^2 is an under-used, perhaps unappreciated quantitative measure of the degree of relation between a model and data, or two types of measurements (or, two types of models, if we can define uncertainties for each)

Reduced χ^2 is very commonly used in the physics community

Indeed, in physics lab courses, students are sometimes cautioned to be critical of an experiment where reduced χ^2 lies close to zero. Why is this?

Reduced χ^2 is an under-used, perhaps unappreciated quantitative measure of the degree of relation between a model and data, or two types of measurements (or, two types of models, if we can define uncertainties for each)

Reduced χ^2 is very commonly used in the physics community

Indeed, in physics lab courses, students are sometimes cautioned to be critical of an experiment where reduced χ^2 lies close to zero. Why is this?

See http://www.physics.csbsju.edu/stats/chi_fit.html for more for reduced χ^2

A method for evaluating bias in global measurements of CO_2 total columns from space

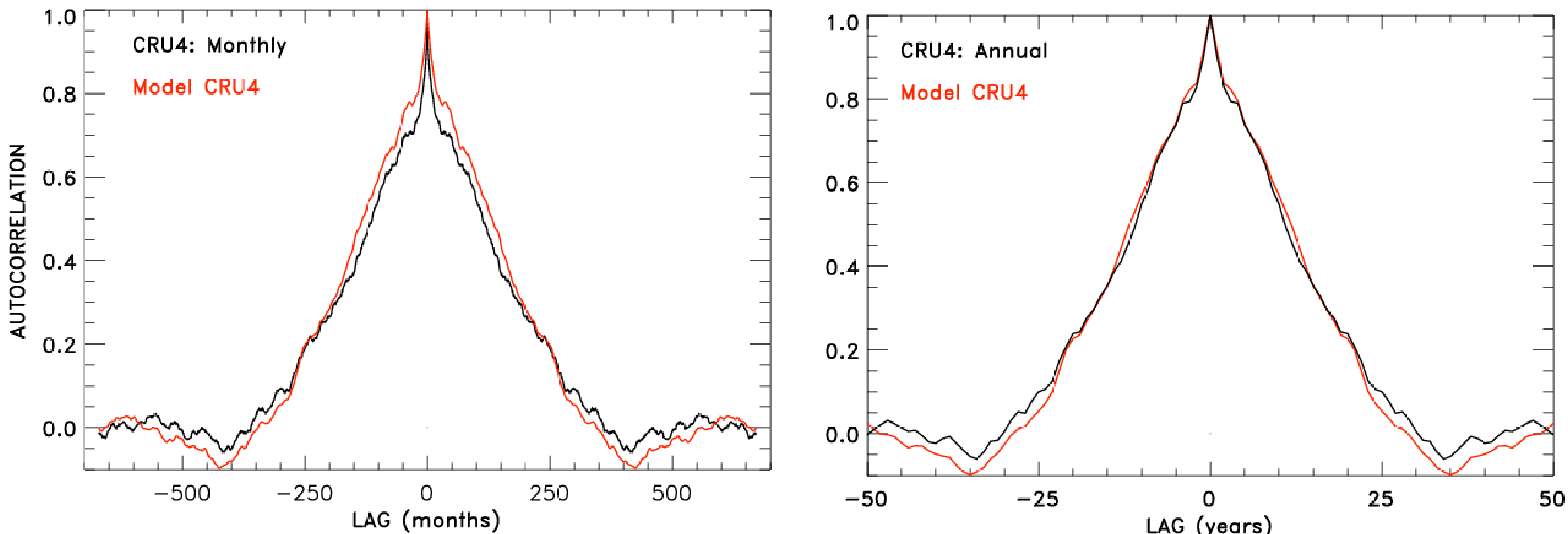
D. Wunch[1], P. O. Wennberg[1], G. C. Toon[1,2], B. J. Connor[3], B. Fisher[2], G. B. Osterman[2], C. Frankenberg[2], L. Mandrake[2], C. O'Dell[4], P. Ahonen[5], S. C. Biraud[14], R. Castano[2], N. Cressie[6], D. Crisp[2], N. M. Deutscher[7,8], A. Eldering[2], M. L. Fisher[14], D. W. T. Griffith[8], M. Gunson[2], P. Heikkinen[5], G. Keppel-Aleks[1], E. Kyrö[5], R. Lindenmaier[15], R. Macatangay[8], J. Mendonca[15], J. Messerschmidt[7], C. E. Miller[2], I. Morino[9], J. Notholt[7], F. A. Oyafuso[2], M. Rettinger[10], J. Robinson[12], C. M. Roehl[1], R. J. Salawitch[11], V. Sherlock[12], K. Strong[15], R. Sussmann[10], T. Tanaka[9,*], D. R. Thompson[2], O. Uchino[9], T. Warneke[7], and S. C. Wofsy[13]

Retrievals are defined as successful by the master quality flag when they satisfy $\chi^2 < 1.2$. However, the χ^2 values have increased linearly over time, because the time-dependent radiometric calibration caused by a sensitivity degradation of the O_2 A-band channel was not applied to the noise model. To compensate for this, we adjust the cutoff value so that it starts at 1.2 and evolves with a linear increase in time, matching the increase in minimum χ^2. As a result, a similar number of scenes are retained over time.

Auto-correlation

$$\phi(L) = \frac{1}{N-2L}\sum_{k=L}^{N-L} x'_k \, x'_{k+L} = \overline{x'_k \, x'_{k+L}}; \;\; L = 0, \pm1, \pm2, \pm3, \ldots$$

where x' is deviation from mean



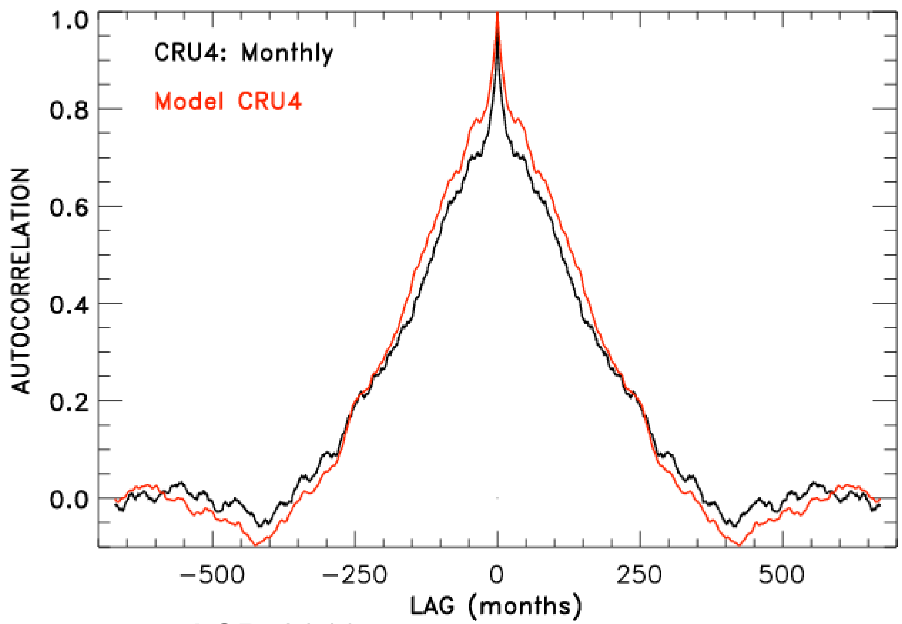
Canty et al., ACP, 2013

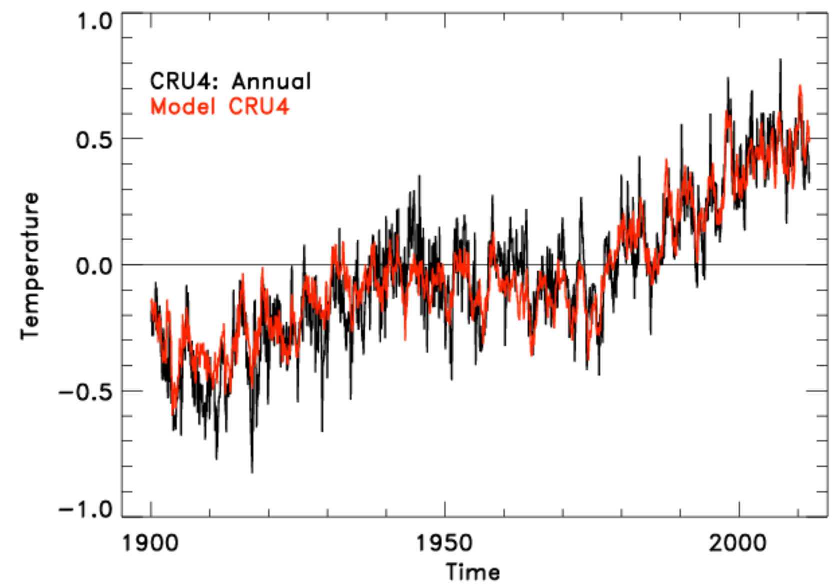
AOSC 652: Analysis Methods in AOSC

Auto-correlation

$$\phi(L) = \frac{1}{N-2L}\sum_{k=L}^{N-L} x'_k\, x'_{k+L} = \overline{x'_k\, x'_{k+L}};\;\; L = 0, \pm1, \pm2, \pm3, \ldots$$

where x' is deviation from mean





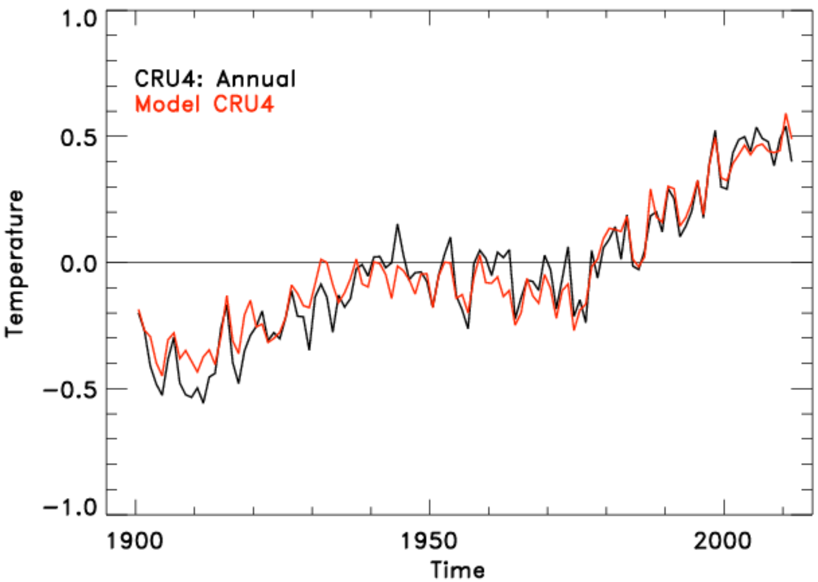
Canty et al., ACP, 2013

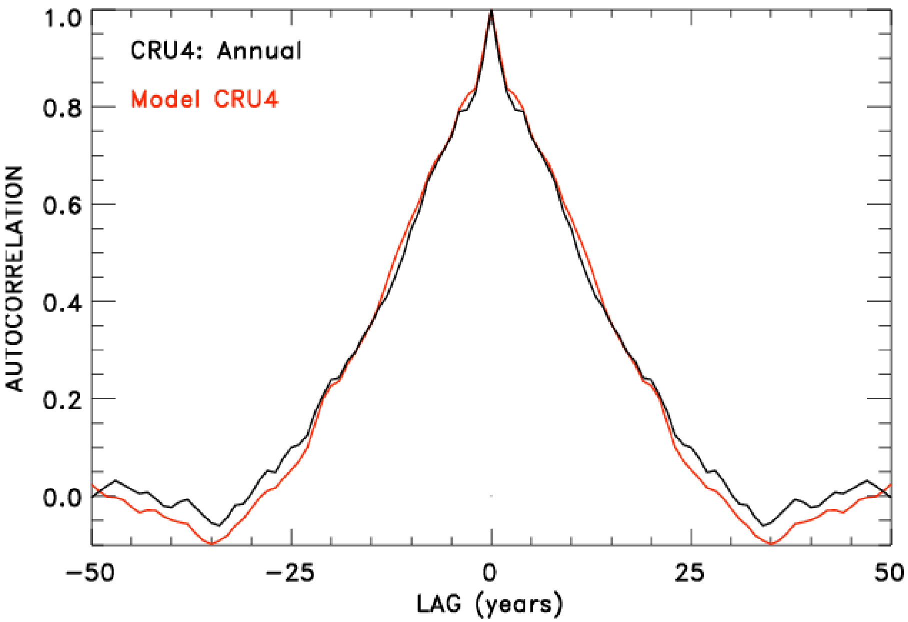
AOSC 652: Analysis Methods in AOSC

Auto-correlation

$$\phi(L) = \frac{1}{N-2L}\sum_{k=L}^{N-L} x'_k\, x'_{k+L} = \overline{x'_k\, x'_{k+L}};\;\; L = 0, \pm 1, \pm 2, \pm 3, \ldots$$

where x' is deviation from mean





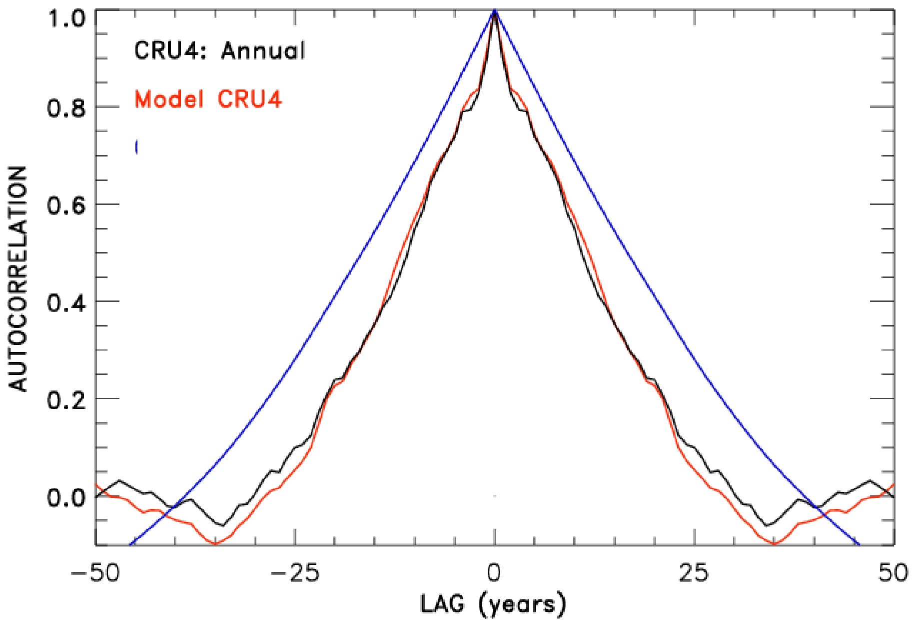
Canty et al., ACP, 2013

AOSC 652: Analysis Methods in AOSC

Auto-correlation

$$\phi(L) = \frac{1}{N-2L}\sum_{k=L}^{N-L} x'_k\, x'_{k+L} = \overline{x'_k\, x'_{k+L}};\;\; L = 0, \pm 1, \pm 2, \pm 3, \ldots$$

where x' is deviation from mean



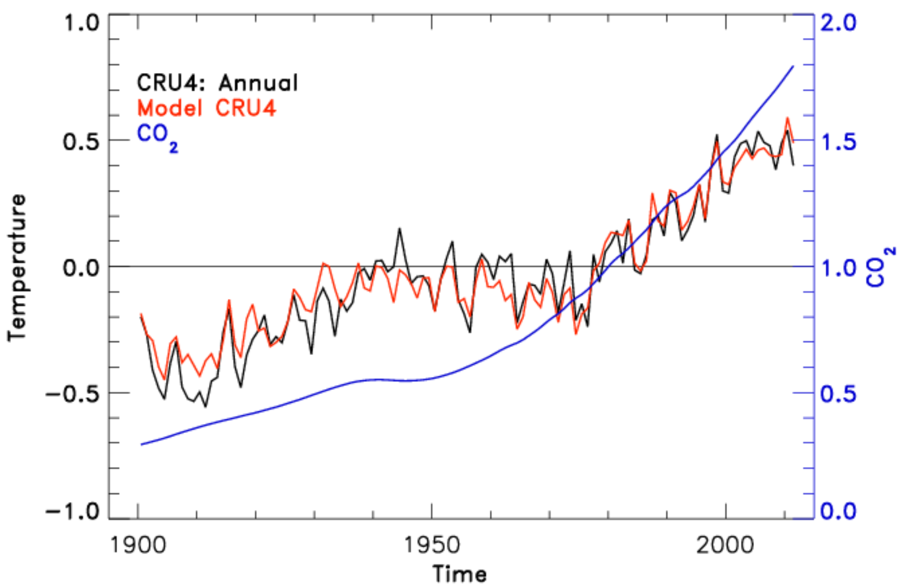
Canty et al., ACP, 2013

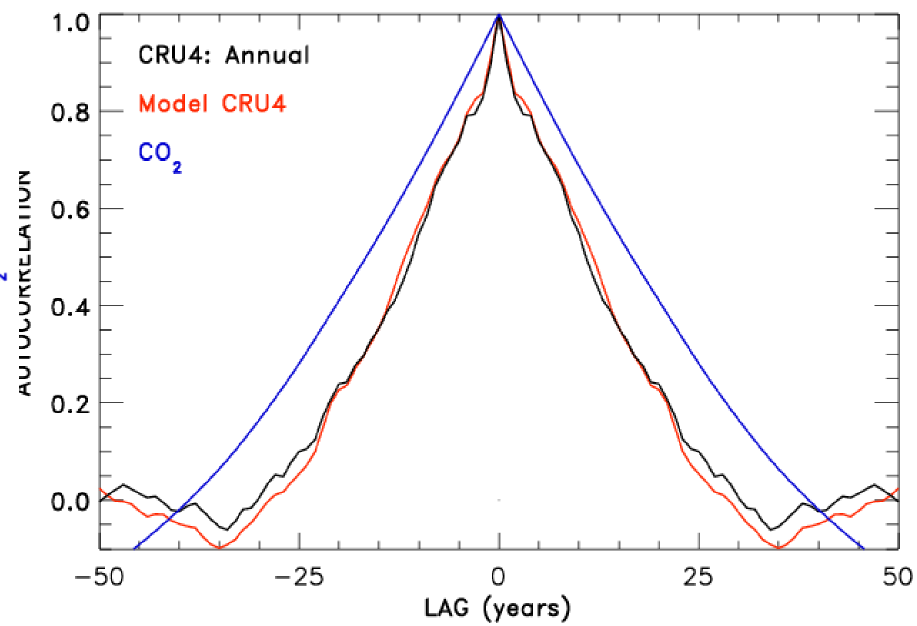
AOSC 652: Analysis Methods in AOSC

Auto-correlation

$$\phi(L) = \frac{1}{N-2L}\sum_{k=L}^{N-L} x'_k\, x'_{k+L} = \overline{x'_k\, x'_{k+L}};\; L = 0, \pm 1, \pm 2, \pm 3, \ldots$$

where x' is deviation from mean





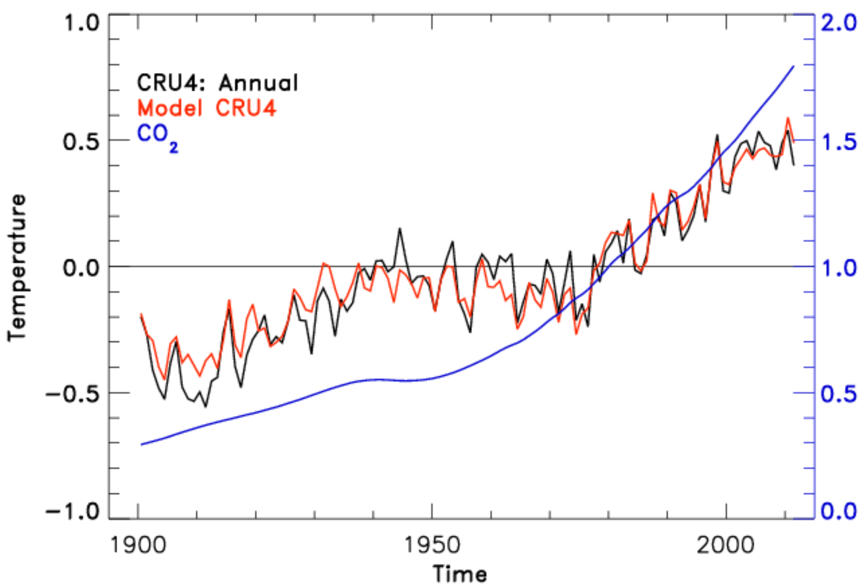
Canty et al., ACP, 2013

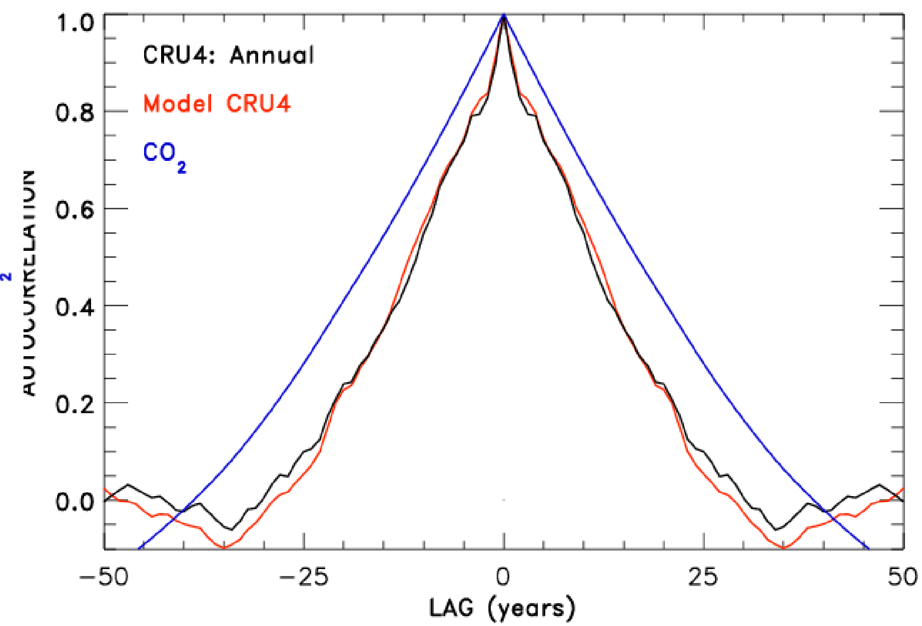
AOSC 652: Analysis Methods in AOSC

Auto-correlation

$$\phi(L) = \frac{1}{N-2L}\sum_{k=L}^{N-L} x'_k\, x'_{k+L} = \overline{x'_k\, x'_{k+L}};\;\; L = 0, \pm 1, \pm 2, \pm 3, \ldots$$

where x' is deviation from mean





Canty et al., ACP, 2013

The behavior of the auto-correlation of a signal is often used to infer degrees of freedom: see for example Dennis Hartmann's class notes: http://www.atmos.washington.edu/~dennis/552_Notes_6a.pdf