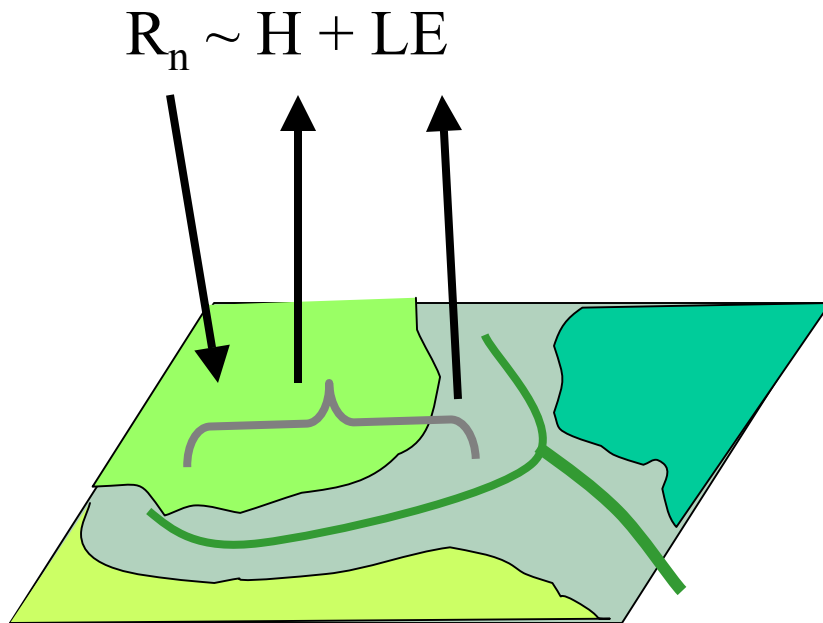


Surface water/energy budget coupling over heterogeneous terrain



$$LE = f_{veg} LE_{veg} + (1 - f_{veg}) LE_{soil}$$

$$LE = f(R_n, T, g_c, g_a, g_{soil}, VPD)$$

$$g_a = f(\text{canopy structure, wind, ...})$$

$$g_c = f(\text{soil water, VPD, PAR, T, LAI})$$

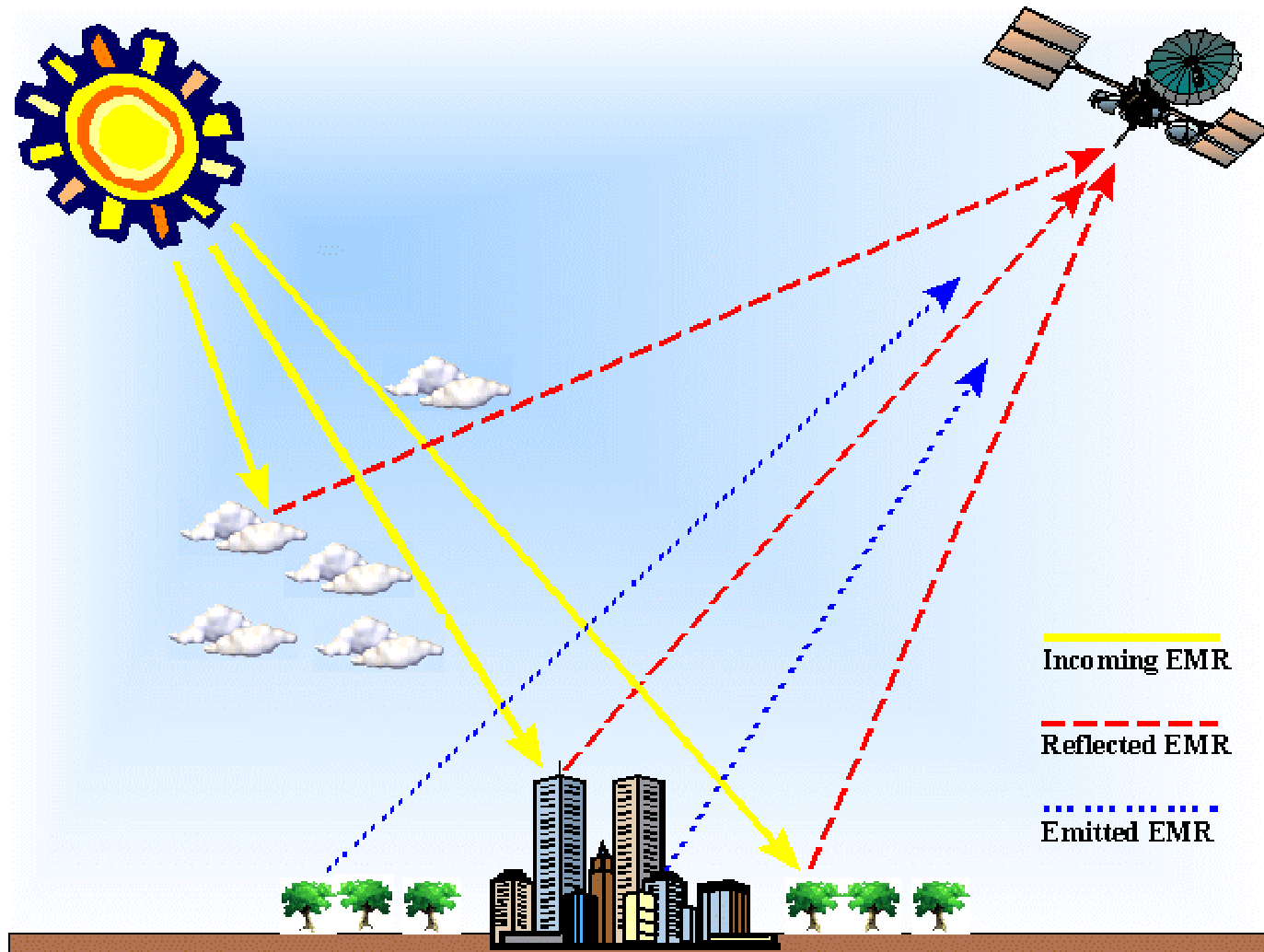
$$g_{soil} = f(\text{soil water, ...})$$

T_s lower with greater LE (evaporative cooling) as a function of soil water (other factors), greater canopy cover (higher NDVI)

T_s and NDVI estimated by a set of operational remote sensors

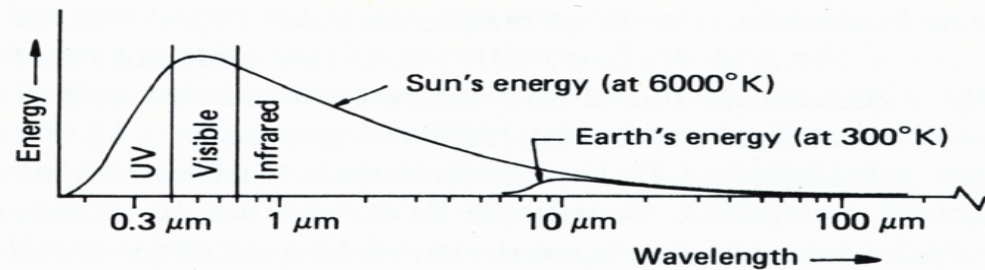
Satellite Imagery - Sensing EMR

- Digital data obtained by sensors on satellite platforms

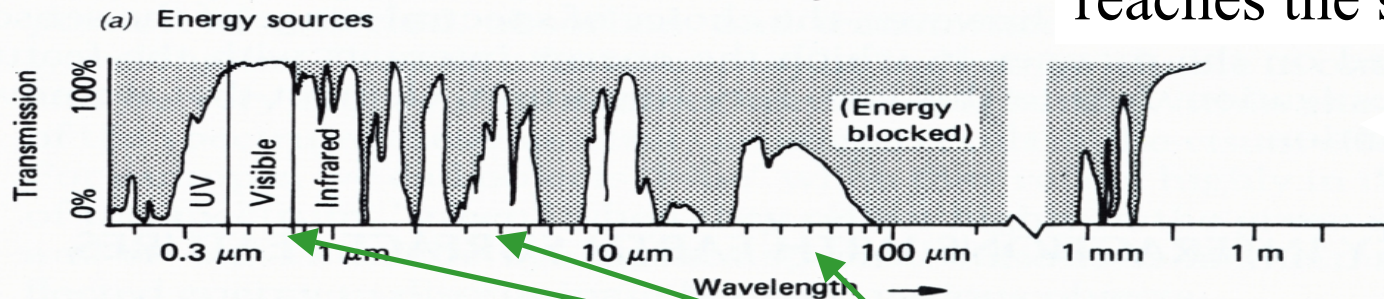


Solar Electromagnetic Radiation

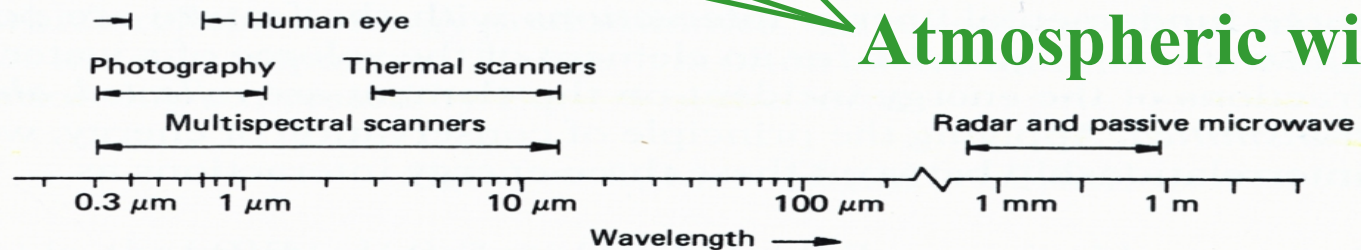
- The sun emits EMR across a **broad spectrum** of wavelengths:



But the atmosphere blocks much of the energy before it reaches the surface

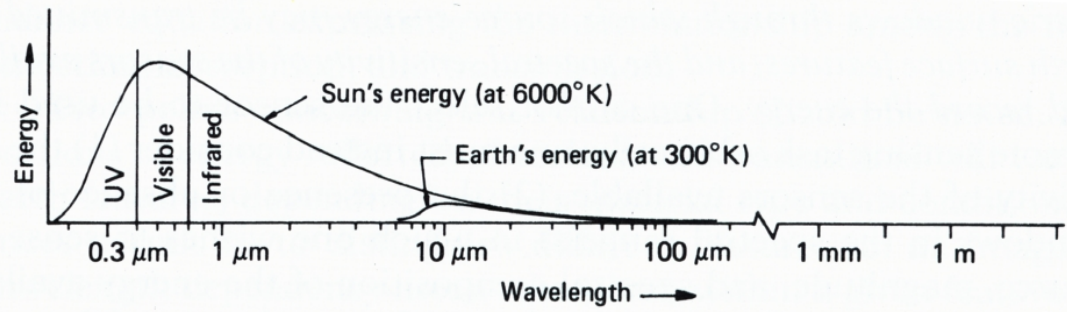


(b) Atmospheric transmittance

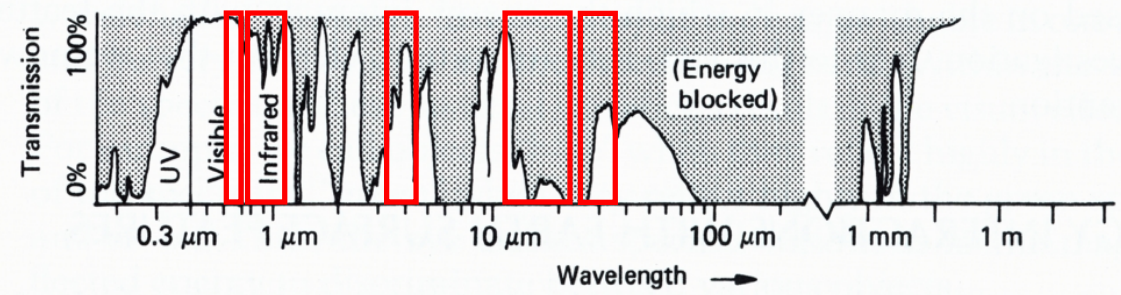


Atmospheric windows

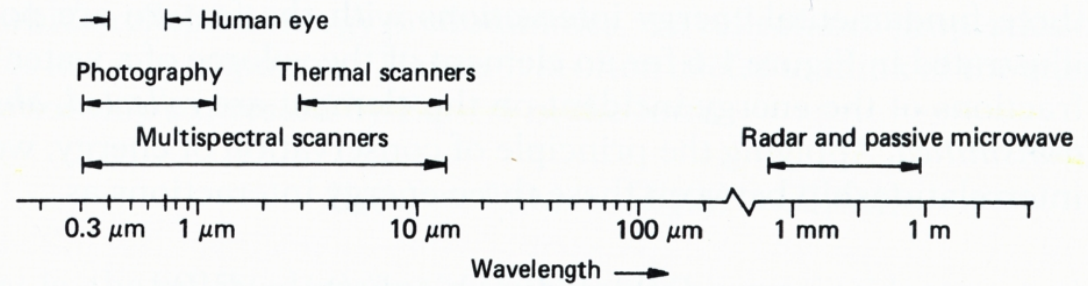
AVHRR Bands



(a) Energy sources



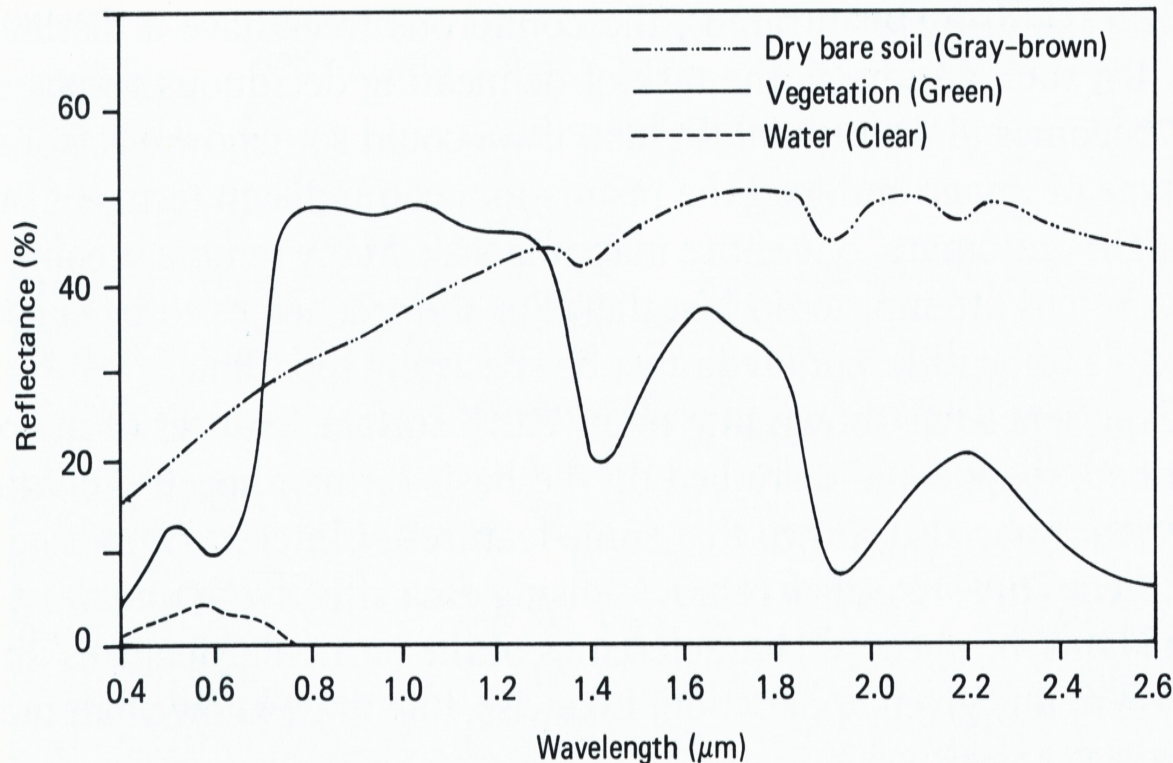
(b) Atmospheric transmittance



Sensing Vegetation and Temperature

- Can take ratios or other **combinations of multiple input bands** to produce indices, e.g.:
- **Normalized Difference Vegetation Index (NDVI)**
 - Designed to contrast heavily-vegetated areas with areas containing little vegetation, by taking advantage of vegetation's strong absorption of red and reflection of near infrared:
 - $NDVI = (NIR - R) / (NIR + R)$
- **Surface temperature (T_s)** from IR bands using Price (1984):
 - $T_s = TIR1 + 3.33 (TIR1 - TIR2)$
 - Wavelengths: $TIR1 = 10.8 \mu m$, $TIR2 = 11.9 \mu m$

Normalized Difference Vegetation Index

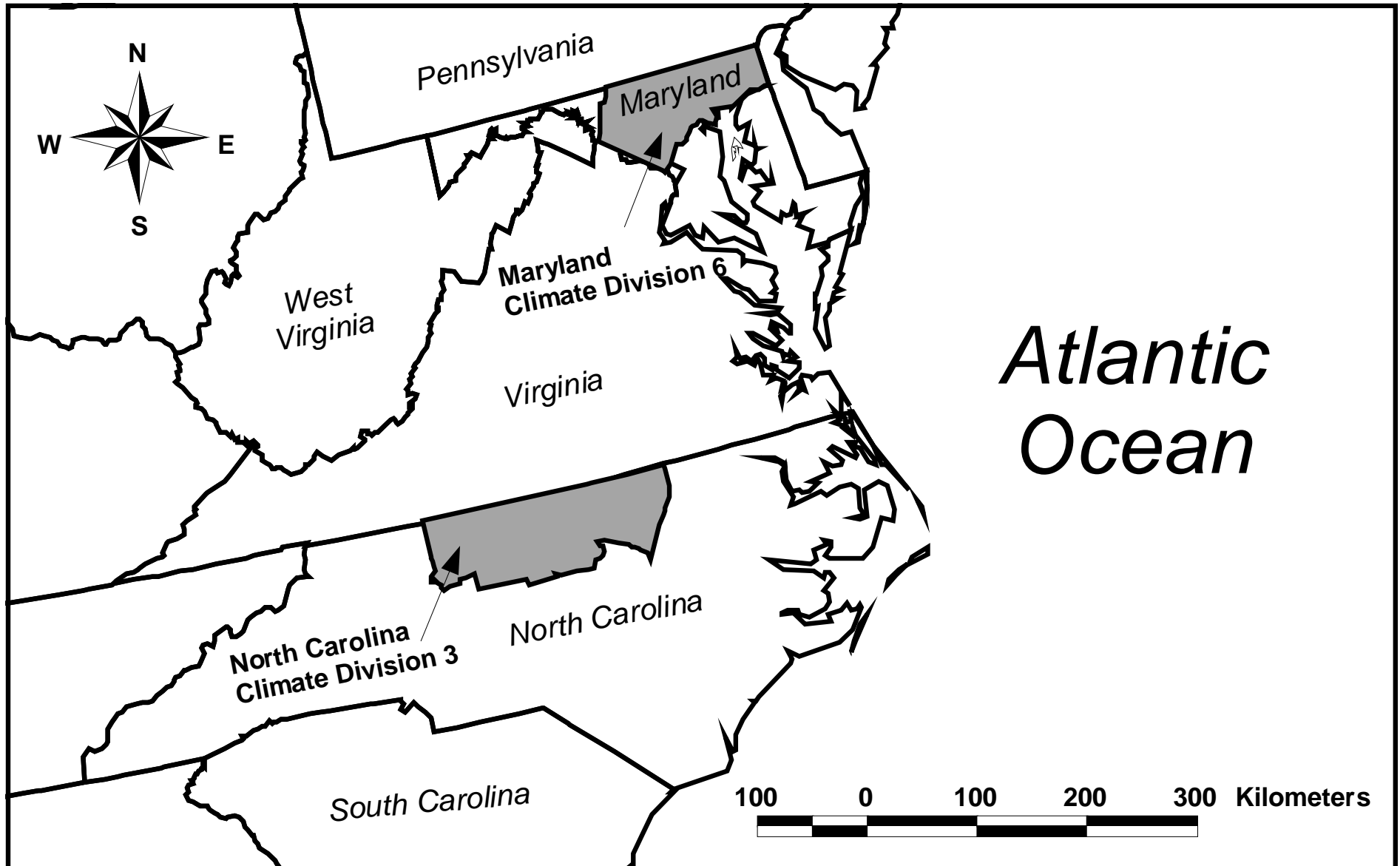


$$\text{NDVI} = \frac{(\text{NIR} - \text{R})}{(\text{NIR} + \text{R})}$$

$$\text{NDVI} [-1,1]$$

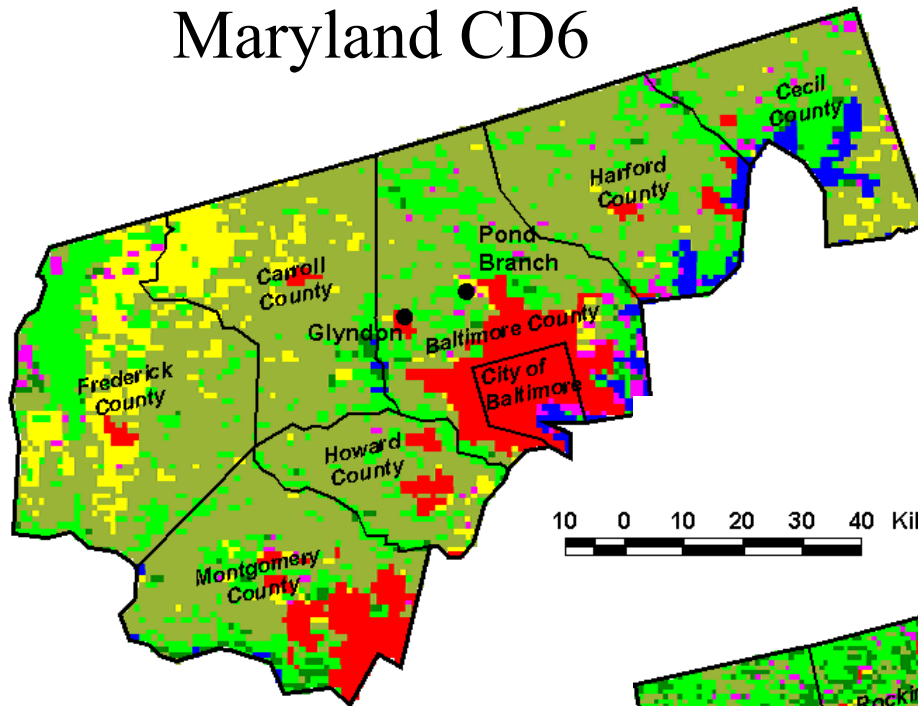
- Vegetation has a **strong contrast in reflectance** between red and near infrared EMR, and NDVI takes advantage of this to **sense the presence/density of vegetation**

Study Climate Divisions



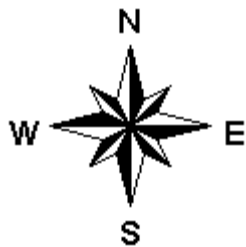
MODIS LULC In Climate Divisions

Maryland CD6

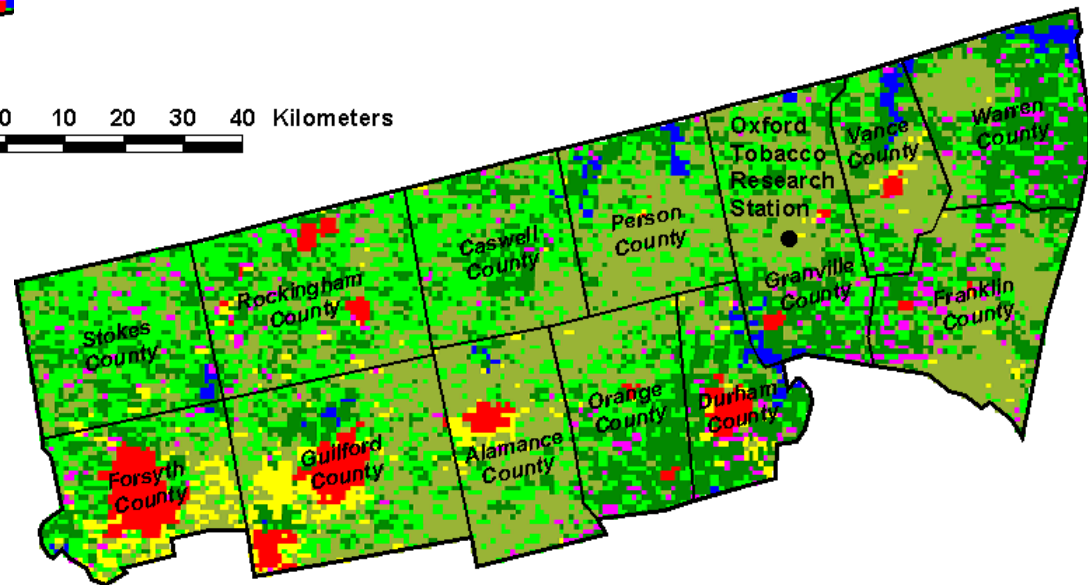


MODIS Land Cover

- Deciduous Broadleaf Forests
- Mixed Forests
- Cropland
- Urban and Built-Up
- Cropland/Natural Vegetation Mosaic
- Other
- Water
- Outside NC CD 3



10 0 10 20 30 40 Kilometers



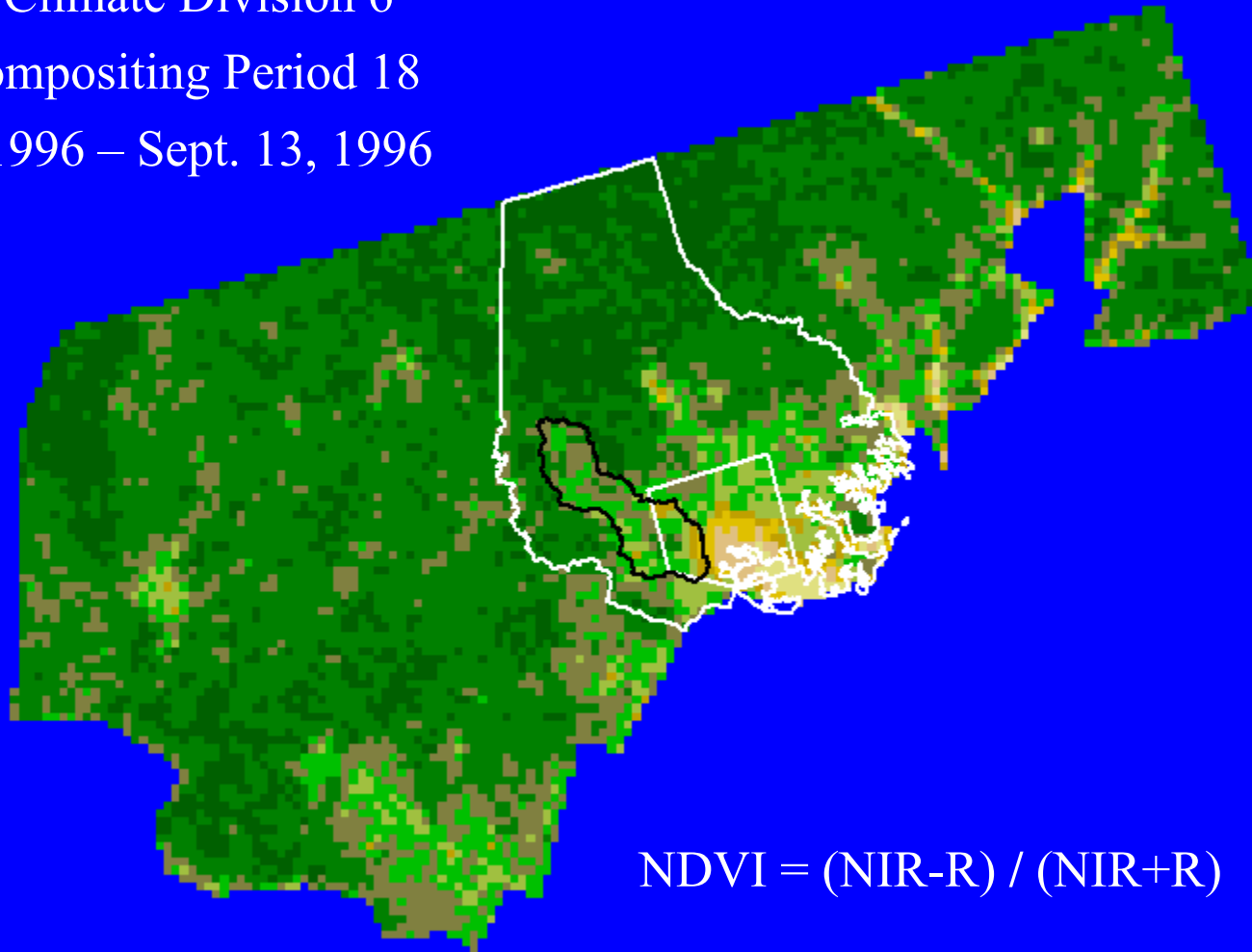
North Carolina CD3

AVHRR Satellite Imagery - NDVI

Maryland Climate Division 6

1996 – Compositing Period 18

Aug. 30, 1996 – Sept. 13, 1996



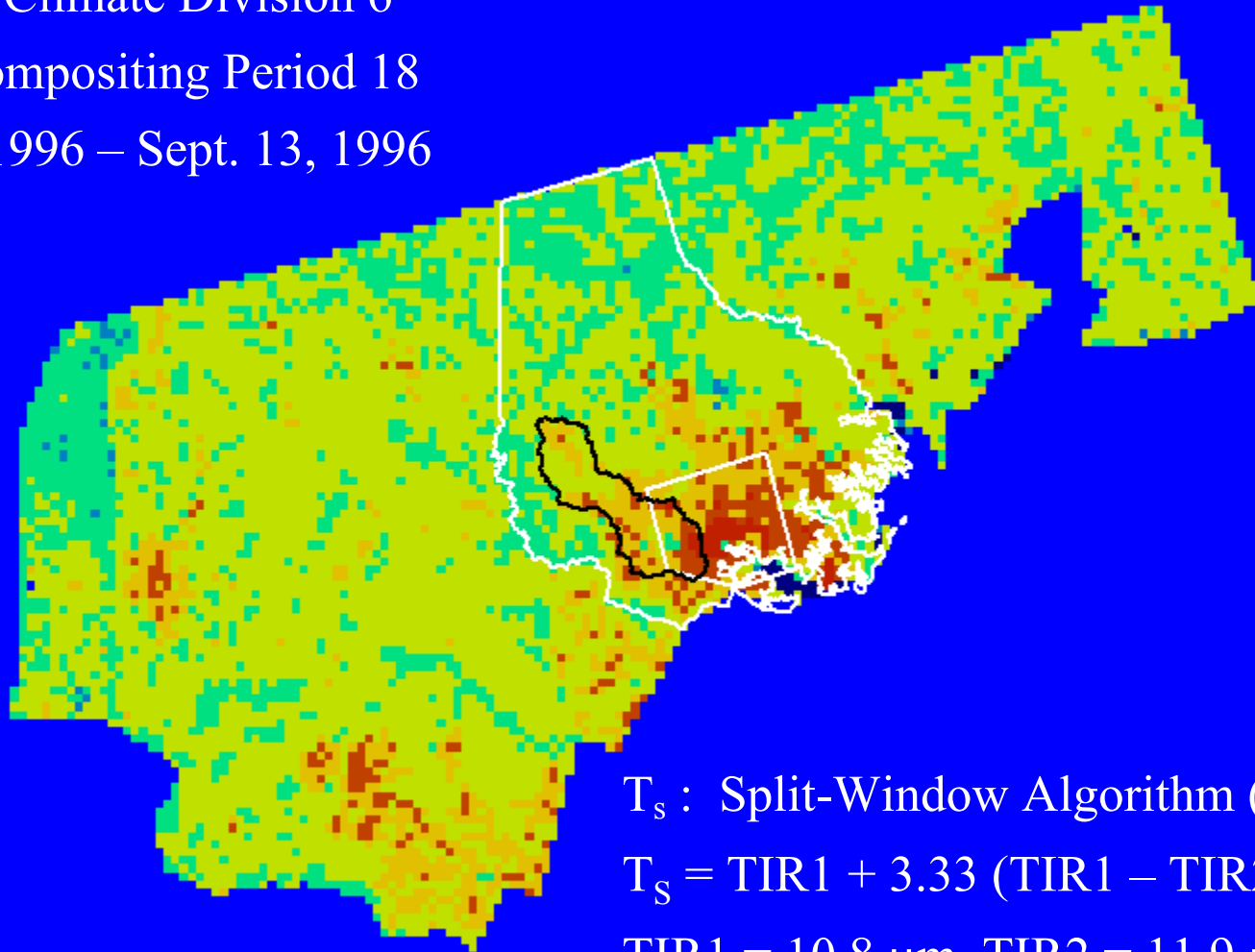
$$\text{NDVI} = (\text{NIR} - \text{R}) / (\text{NIR} + \text{R})$$

AVHRR Satellite Imagery - T_s

Maryland Climate Division 6

1996 – Compositing Period 18

Aug. 30, 1996 – Sept. 13, 1996

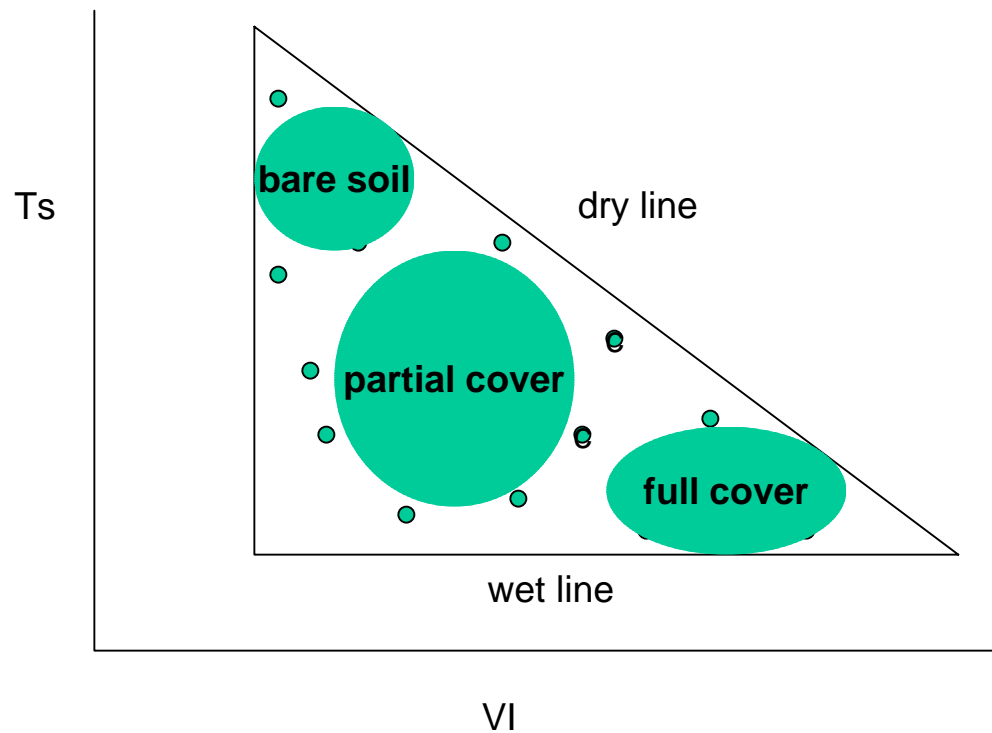


T_s : Split-Window Algorithm (Price 1984)

$$T_s = TIR1 + 3.33 (TIR1 - TIR2)$$

$$TIR1 = 10.8 \mu\text{m}, TIR2 = 11.9 \mu\text{m}$$

Interpretation of the VI-T_s Space

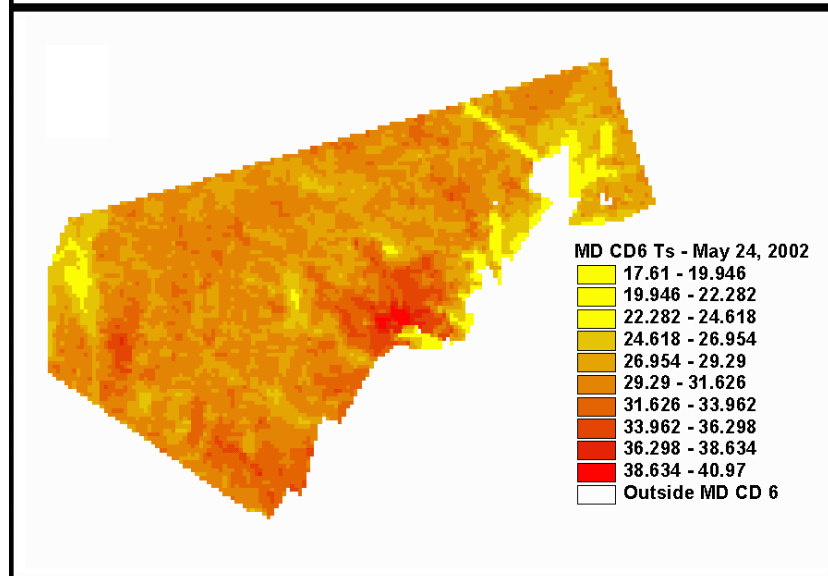
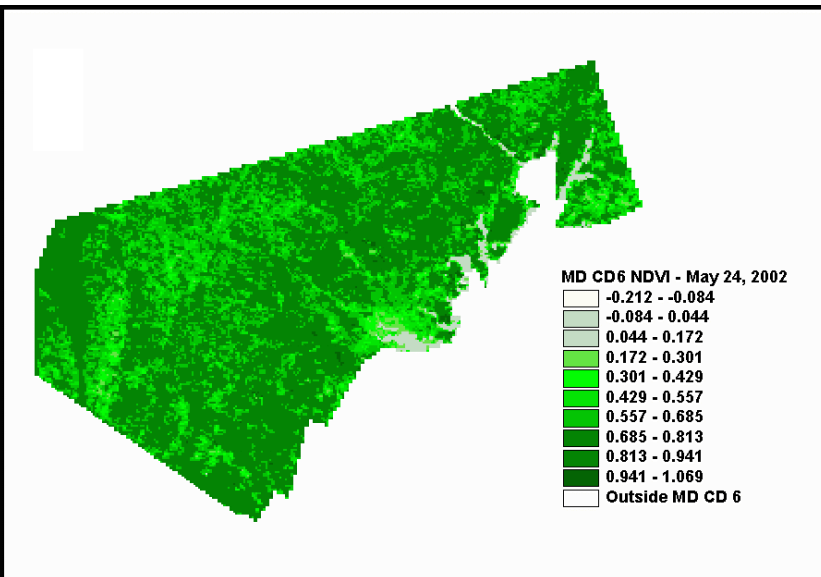
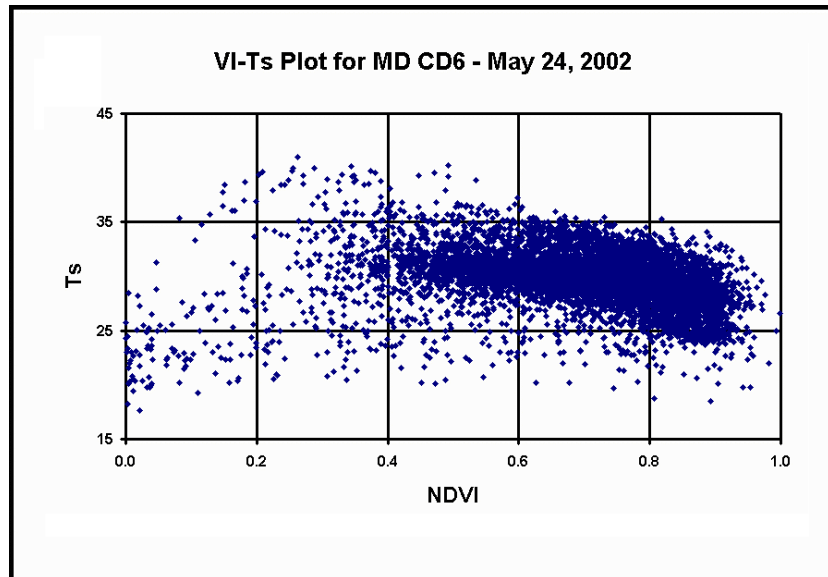


Adapted from Sandholt et al. 2002

Dry Line Slope – Sigma (σ)

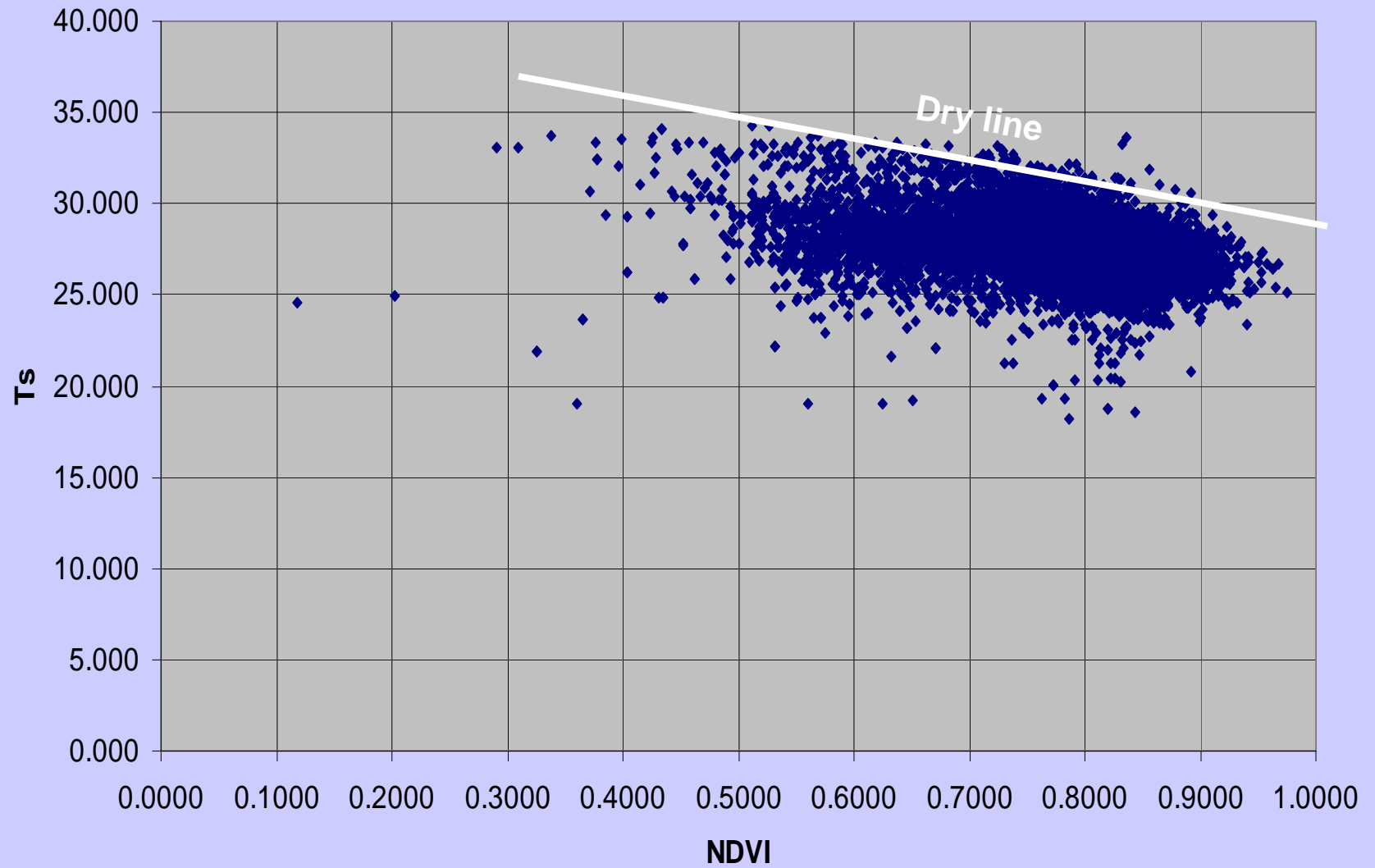
- Nemani and Running (1989) suggested, and later Nemani, Pierce, Running, and Goward (1993) demonstrated, that **the slope of the dry line** (symbolized using σ) is a good overall indicator of the **surface moisture condition** of a region (where the T_s and VI pixels that are drawn from to form the 2-D T_s -VI distribution) on the occasion when the imagery was collected
 - **Steeper, more negative** slopes represent **drier conditions** (where T_s disparities are greater)
- So **how** do we form the 2-D T_s -VI distribution and find the slope of the dry line?

Finding the Dry Line (σ) Slope



- We begin with T_s and VI data, ideally collected using the **same sensor at the same time** (e.g. from AVHRR bands 1, 2, 4, & 5)
- We then translate the values for each pixel into a **2-D parameter space**, the VI on the x-axis and the T_s on the y-axis

2001 MODIS Yearday 241 Climate Division 3 Ts-NDVI Plot



Finding the Dry Line (σ) Slope

- With a **real** T_s -VI distribution, **fitting a line** to the upper envelope of the distribution is **a little bit tricky!**
- We can break it down into a **two-part process**:
 - 1st, we must **identify a subset of all pixels** in the distribution that represent **the upper envelope**, that is those pixels with **the highest T_s for a given VI** → We can accomplish this through some sort of **classification/filtering method**
 - 2nd, once we have **identified the upper envelope pixels**, we must **fit a line through them** → We can accomplish this through fitting a **simple linear regression model**

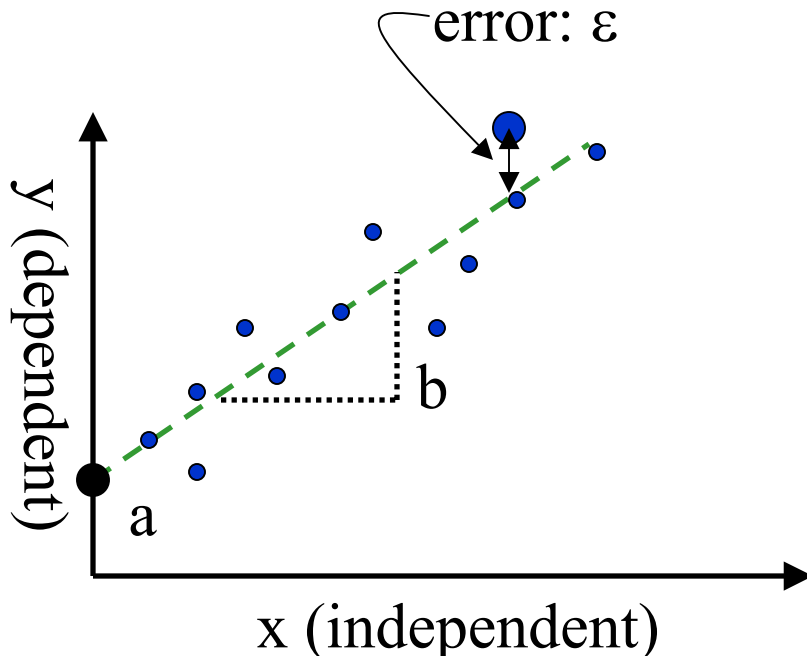
Simple vs. Multiple Regression

- Today, we are going to examine **simple linear regression**, where we estimate the values of a **dependent variable (y)** using the values of an **independent variable (x)**
- This concept can be extended to **multiple linear regression**, where **more** explanatory independent variables ($x_1, x_2, x_3 \dots x_n$) are used to develop estimates of the dependent variable's values
- For purposes of **clarity**, we will first look at the simple case, so we can more easily grasp the mathematics involved

Simple Linear Regression

- **Simple linear regression** models the relationship between an independent variable (x) and a dependent variable (y) using an equation that expresses y as a linear function of x , plus an error term:

$$y = a + bx + e$$



x is the **independent** variable

y is the **dependent** variable

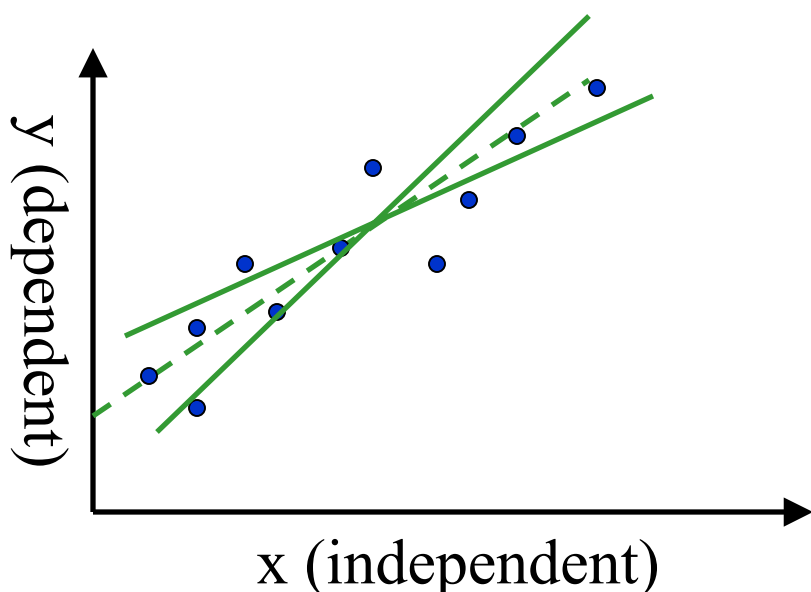
b is the **slope** of the fitted line

a is the **intercept** of the fitted line

e is the **error** term

Fitting a Line to a Set of Points

- When we have a data set consisting of an independent and a dependent variable, and we plot these using a scatterplot, to construct our model between the relationship between the variables, we need to **select a line** that represents the relationship:



- We can choose a line that fits best using a **least squares method**
- The least squares line is the line that **minimizes** the **vertical distances** between the points and the line, i.e. it minimizes the **error term ϵ** when it is considered for all points in the data set

Sampling and Regression II

- We usually operate using **sampled data**, and while we are building a model of the form:

$$y = a + bx + e$$

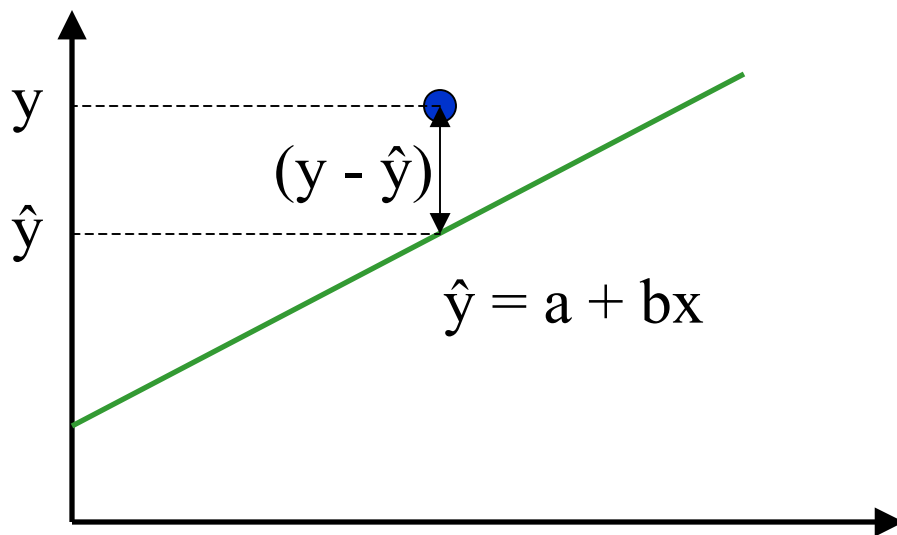
from our sample, in doing so we are attempting to estimate a “**true**” regression line, describing the relationship between independent variable (x) and dependent variable (y) for the entire **population**:

$$y = \alpha + \beta x + \varepsilon$$

- **Multiple** samples would yield several **similar** regression lines, which should approximate the population regression line

Least Squares Method

- The least squares method operates mathematically, **minimizing the error term e** for all points
- We can describe the line of best fit we will find using the equation $\hat{y} = a + bx$, and you'll recall that from a previous slide that the formula for our linear model was expressed using $y = a + bx + e$



- We use the value \hat{y} on the line to estimate the true value, y
- The **difference** between the two is $(y - \hat{y}) = e$
- This difference is **positive** for points **above** the line, and **negative** for points **below** it

Estimates and Residuals

- Our simple linear regression models take the **form**:

$$y = a + bx + e$$

which can **alternatively** be expressed as:

$$\hat{y} = a + bx$$

where \hat{y} is the **estimate** of y produced by the regression

- We can **rearrange** these equations to show:

$$e = y - \hat{y}$$

- The errors in the estimation of y using the regression equation are known as **residuals**, and express for any given value in the data set to what extent the regression line is either **underestimating or overestimating** the true value of y

Minimizing the Error Term

- In a linear model, the **error** in estimating the true value of the dependent variable y is expressed by the **difference** between the true value and the estimated value \hat{y} , $e = (y - \hat{y})$ (i.e. the **residuals**)
- Sometimes this difference will be **positive** (when the line **underestimates** the value of y) and sometimes it will be **negative** (when the line **overestimates** the value of y), because there will be points above and below the line
- If we were to simply **sum** these error terms, the positive and negative values would **cancel out**
- Instead, we can **square the differences** and then sum them up to create a **useful estimate** of the **overall error**

Error Sum of Squares

- By squaring the differences between y and \hat{y} , and summing these values for all points in the data set, we calculate the **error sum of squares** (usually denoted by SSE):

$$\text{SSE} = \sum_{i=1}^n (y - \hat{y})^2$$

- The **least squares method** of selecting a line of best fit functions by finding the parameters of a line (intercept a and slope b) that **minimizes** the error sum of squares, i.e. it is known as the least squares method because it finds the line that **makes the SSE as small as it can possibly be**, minimizing the vertical distances between the line and the points

Finding Regression Coefficients

- The **equations** used to find the values for the slope (b) and intercept (a) of the line of best fit using the least squares method are:

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Where:

x_i is the i^{th} **independent** variable value

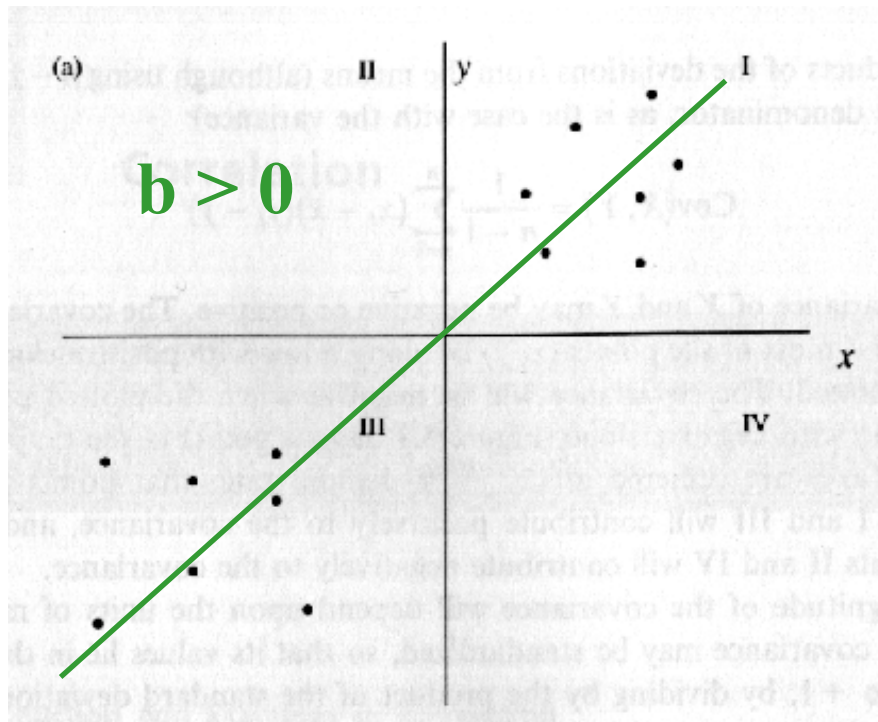
y_i is the i^{th} **dependent** variable value

\bar{x} is the **mean** value of all the x_i values

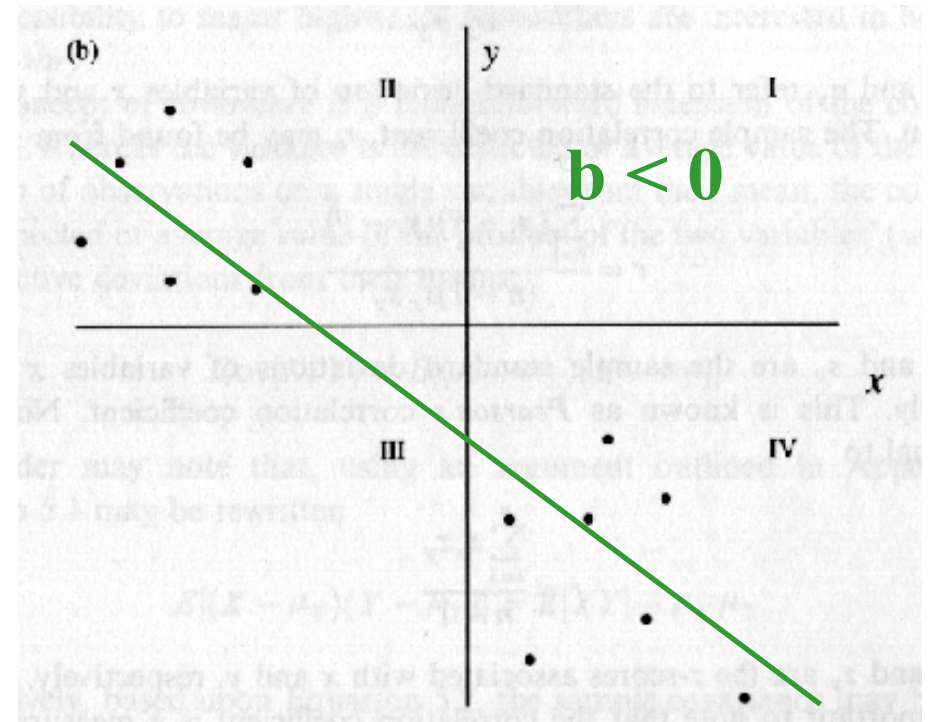
\bar{y} is the **mean** value of all the y_i values

Interpreting Slope (b)

- The **slope** of the line (b), gives the **change in y** (dependent variable) due to a **unit change in x** (independent variable):



Positive relationship – As the values of x **increase**, the values of y **increase too**



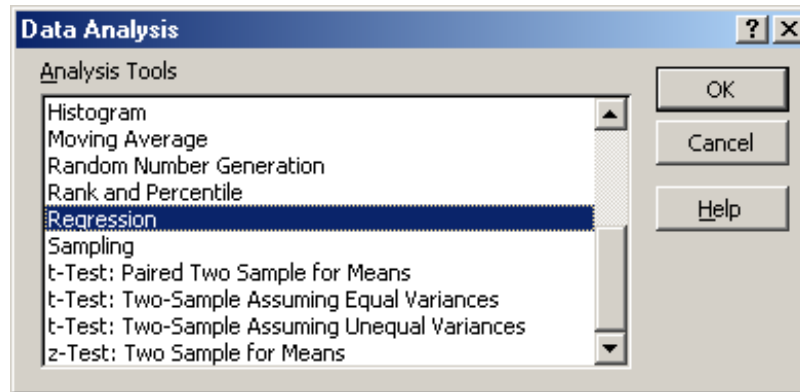
Negative (a.k.a. inverse) relationship – As values of x **increase**, the values of y **decrease**

Simple Linear Regression in Excel

- Excel can calculate regression parameters in **two** ways:
 - There are **built-in functions** that can be entered into a cell to specify the calculation of a regression slope or regression intercept:
 - `SLOPE(array1, array2)` can be used to calculate the **slope** of the least squares regression line, specifying the y values in array1 and the x values in array2
 - `INTERCEPT(array1, array2)` can be used to calculate the **intercept** of the least squares regression line, specifying the y values in array1 and the x values in array2
 - There is also a **Data Analysis Tool** that can be used to calculate the regression parameters

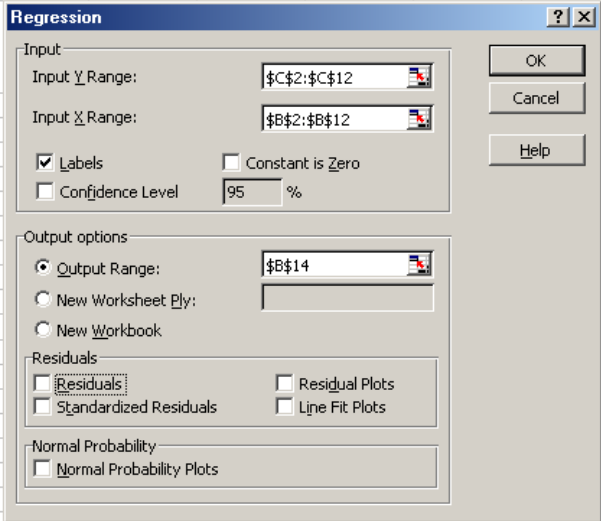
Regression Analysis Tool

- In the Data Analysis window, **select** the appropriate tool:



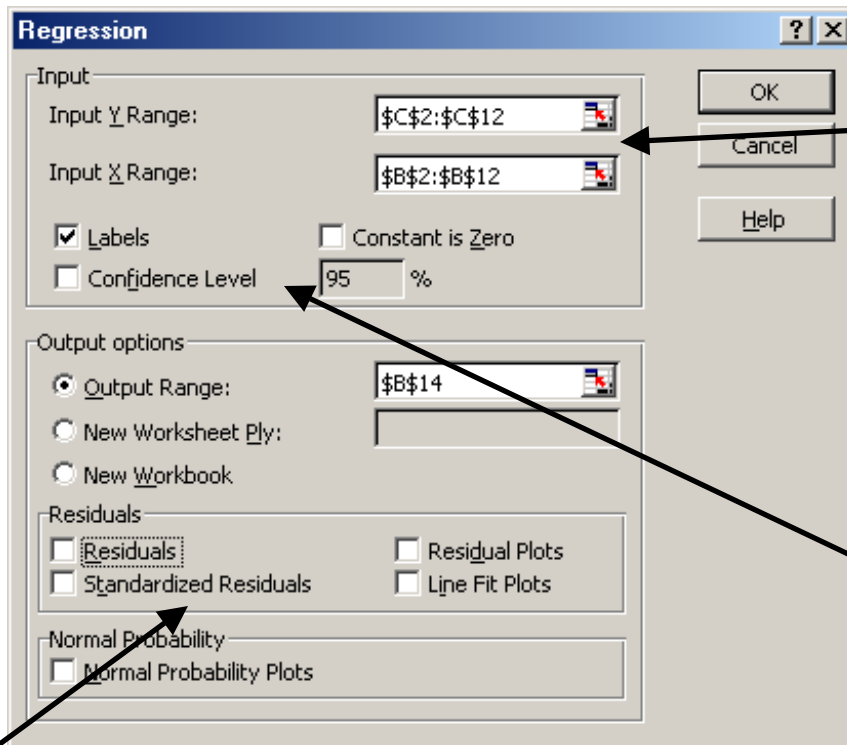
- After clicking OK, you'll be presented with the **tool** window:

	A	B	C	D	E	F	G	H	I
1									
2		TVDI (x)	Soil Moisture (y)						
3		0.274	0.414						
4		0.542	0.359						
5		0.419	0.396						
6		0.286	0.458						
7		0.374	0.350						
8		0.489	0.357						
9		0.623	0.255						
10		0.506	0.189						
11		0.768	0.171						
12		0.725	0.119						
13									
14									
15									
16									
17									
18									
19									
20									



The screenshot shows the 'Regression' dialog box overlaid on the spreadsheet. The 'Input' section has 'Input Y Range' set to '\$C\$2:\$C\$12' and 'Input X Range' set to '\$B\$2:\$B\$12'. The 'Labels' checkbox is checked, and 'Constant is Zero' is unchecked. The 'Confidence Level' is set to 95%. The 'Output options' section has 'Output Range' set to '\$B\$14'. The 'Residuals' section has 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' all unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. On the right side of the dialog, there are three buttons: 'OK', 'Cancel', and 'Help'.

Regression Analysis Tool



Of course, when specifying the input ranges, you must **distinguish** between the dependent variable (y) and the independent variable(s) (x); this tool can also be used for **multiple linear regression**, so more than one x variable can be used

The tool will **automatically** test the significance of the parameters at the 95% confidence level, but if you check the checkbox and **specify another confidence level**, it will test the significance of the regression parameters at that level of confidence **as well**

Checking boxes in the **Residuals** portion of the tool will produce other output including calculating the **residuals** for each value, calculating standardized residuals, and **plotting residuals** versus independent variables, and line fit plots as well

Regression Analysis Tool

- The **basic output** the tool produces includes:

The coefficient of determination (r^2)

The standard error of the estimate (e.g. the standard deviation of the residuals), s_e

An ANOVA table, including the minimum α where F would be **significant**

The regression coefficients produced by the least squares optimization (in the simple case, like this one, the intercept and the slope)

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R		0.87163053				
R Square		0.75973978				
Adjusted R Square		0.72970725				
Standard Error		0.05996834				
Observations		10				
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	0.090973945	0.09097394	25.2972303	0.001014626	
Residual	8	0.028769614	0.0035962			
Total	9	0.119743559				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.60320076	0.061926011	9.74066875	1.0324E-05	0.46039903	0.74600249
TVDI (x)	-0.5923931	0.117780521	-5.0296352	0.00101463	-0.863995597	-0.3207905

The standard error associated with each parameter (e.g. for the regression slope parameter, this is s_b , the standard deviation of the slope)

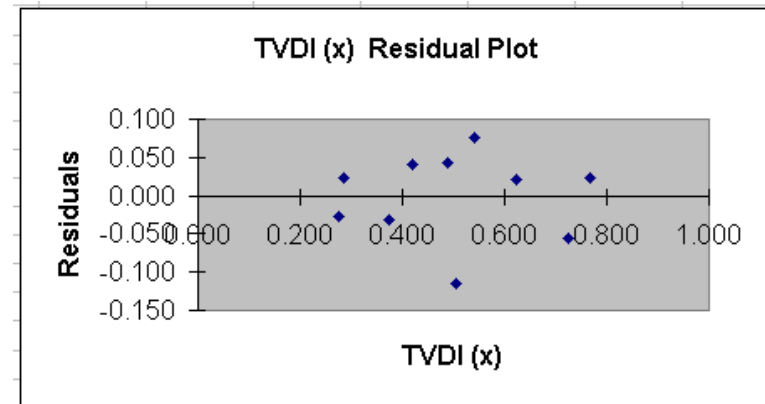
The t-statistic and the minimum α where each parameter would be significant

Regression Analysis Tool

- Checking the Residuals checkbox will produce a **table of the regression estimates** (the \hat{y}_i values) and **residuals**:

RESIDUAL OUTPUT		
Observation	Predicted Soil Moisture (y)	Residuals
1	0.441	-0.027
2	0.282	0.077
3	0.355	0.041
4	0.434	0.024
5	0.382	-0.031
6	0.313	0.043
7	0.234	0.020
8	0.304	-0.115
9	0.148	0.023
10	0.173	-0.055

- Residual Plots creates a **scatter plot of the residuals versus x** (this is useful for checking assumptions about the residuals):



- Line Fit Plots creates a scatter plot of the **actual and predicted values versus x** (this is useful for getting a visual sense of the accuracy of the estimates):

