What's the Goal??

- Exploratory or descriptive statistics:
  - Organize and interpret volumes of data
  - How big, how does one group differ from another

- Inferential statistics:
  - Assess the underlying physical processes that generate environmental data
  - Drug A cures cancer

Causes of Uncertainty

- 1. we can never measure the environment with complete accuracy and precision
- 2. the environment is a chaotic system, which is a maddening combination of randomness and order arising from the characteristics of a complex nonlinear system,
- 3. our understanding of the environmental system is imperfect, so physical (and certainly statistical) models do not capture the complete behavior of the system.

Population vs. Sample

- we never know the entire population of true values as the environmental conditions change in time or space.
- We hope that we choose a sample of observations for analysis such that each element in the population has an equal chance to be selected.
- Sampling issues
  - Trends
  - serial dependence of environmental data
  - model sample tend to be less variable than observed samples

- Selecting the sample for analysis is a critical aspect of organizing the data and depends on the question to be addressed by the study
- rule of thumb: sample should be large enough to capture the phenomenon of interest many times
- "Degrees of freedom": number of independent elements in the sample;
  - usually much smaller than the total number of members in the sample in environmental data sets
- Keeping your powder dry- saving data for an independent sample to evaluate and confirm your results.
- Tendency to assume sample is drawn randomly from the population, when sample grossly underestimates the variability inherent in the population

Why use statistics to describe the environment?

- Environment controlled by innumerable factors, which we hope to segregate into a few critical factors from the rest that, for the most part, simply contribute to background noise
- the characteristics of the system include linearly unstable processes such as baroclinic waves that cause growth of small features into larger ones
- the characteristics of the system (dynamics, thermodynamics) are nonlinear and include discrete step functions (i.e., rain/no rain) that can lead to the amplification of small errors into large ones
- the system is dissipative, which guarantees “stationarity”, i.e., the climate system will remain stable and not run away from the current state

matlab

- Need to know basic commands
- How to manipulate small vectors & arrays
- What .* means versus * and what ./ means

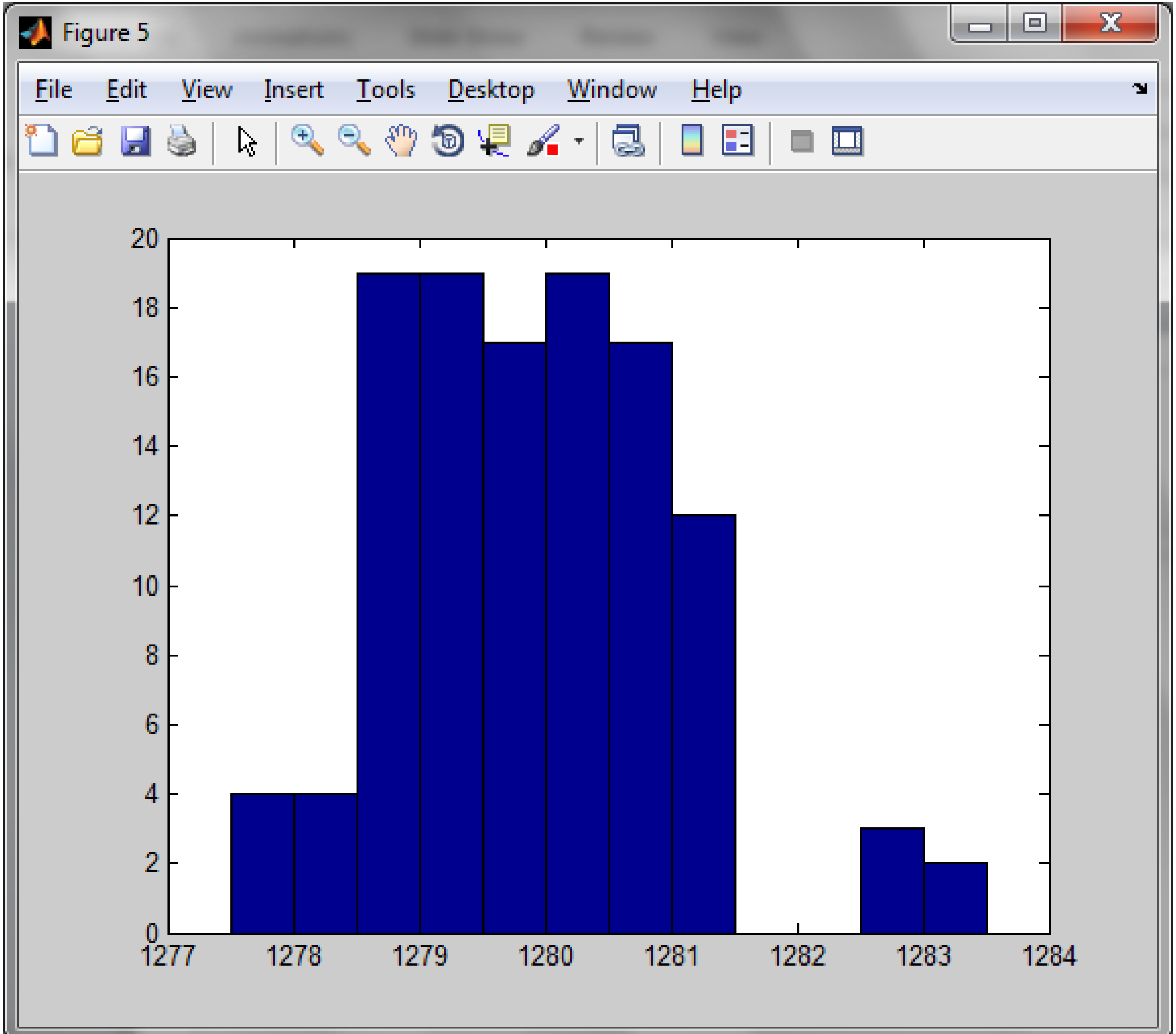
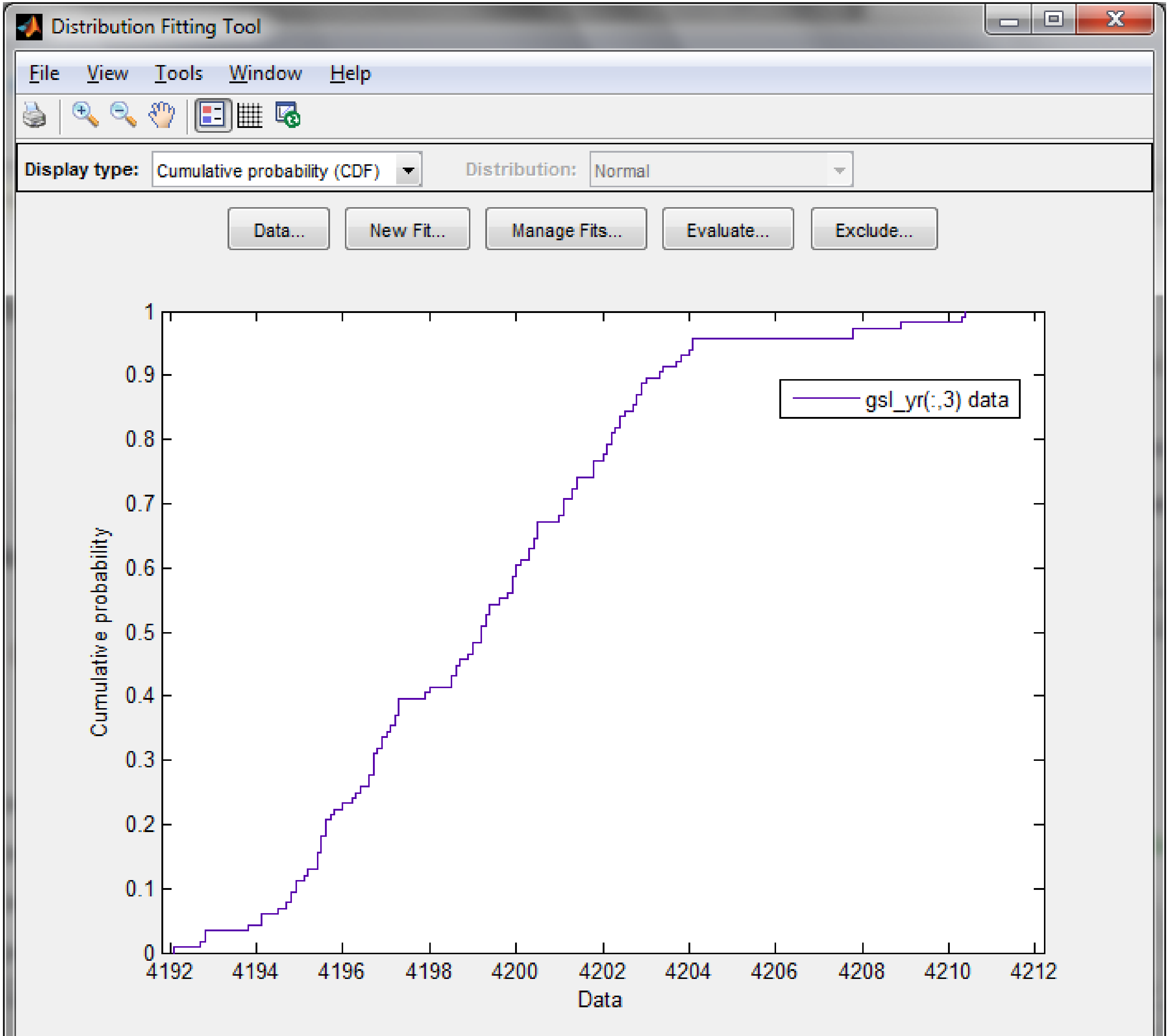


Figure 5
File Edit View Insert Tools Desktop Window Help
20
18
16
14
12
10
8
6
4
2
0
1277 1278 1279 1280 1281 1282 1283 1284


Distribution Fitting Tool
File
View
Tools
Window
Help
Display type:
Cumulative probability (CDF)
Distribution:
Normal
Data...
New Fit...
Manage Fits...
Evaluate...
Exclude...
gsl_yr(:,3) data
Cumulative probability
Data
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
4192
4194
4196
4198
4200
4202
4204
4206
4208
4210
4212

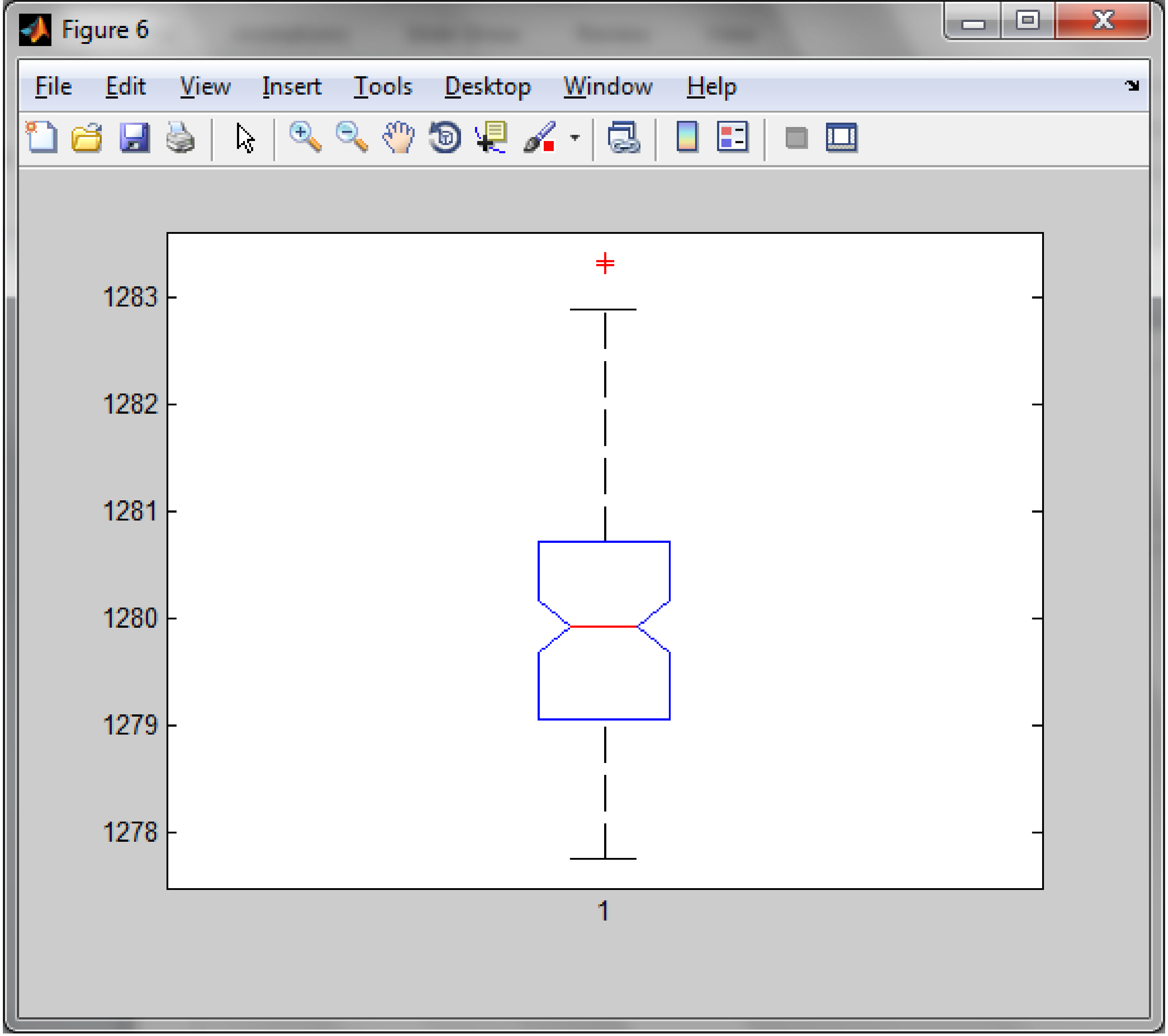
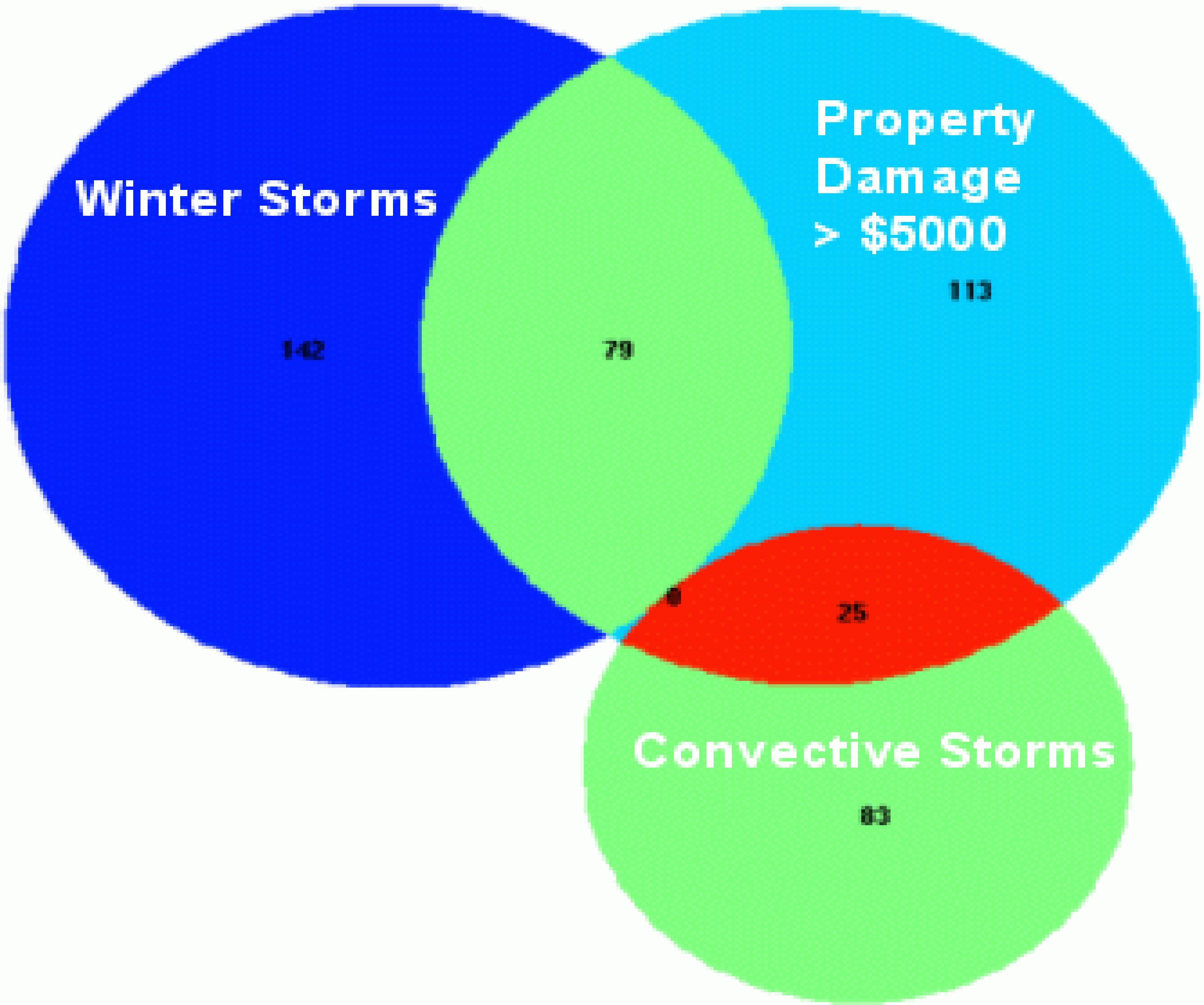


Figure 6
File
Edit
View
Insert
Tools
Desktop
Window
Help
1283
1282
1281
1280
1279
1278
1

- Measures of central value
- Spread
- Symmetry
- Robust and reliant
- How to compute mean and standard deviation of a sample

Storm Reports: Salt Lake County 1993-2005



Number of Opportunities: 2340 (180 days * 13 years)

Two Statistical Frameworks: Frequency vs. Bayesian

- Frequency- probability of an event is its relative frequency after many trials
- a- number of occurrences of E
- n- number of opportunities for E to take place
- a/n- relative frequency of E occurring
- $\Pr\{E\} \longrightarrow a/n$ as $n \rightarrow \infty$

More concepts

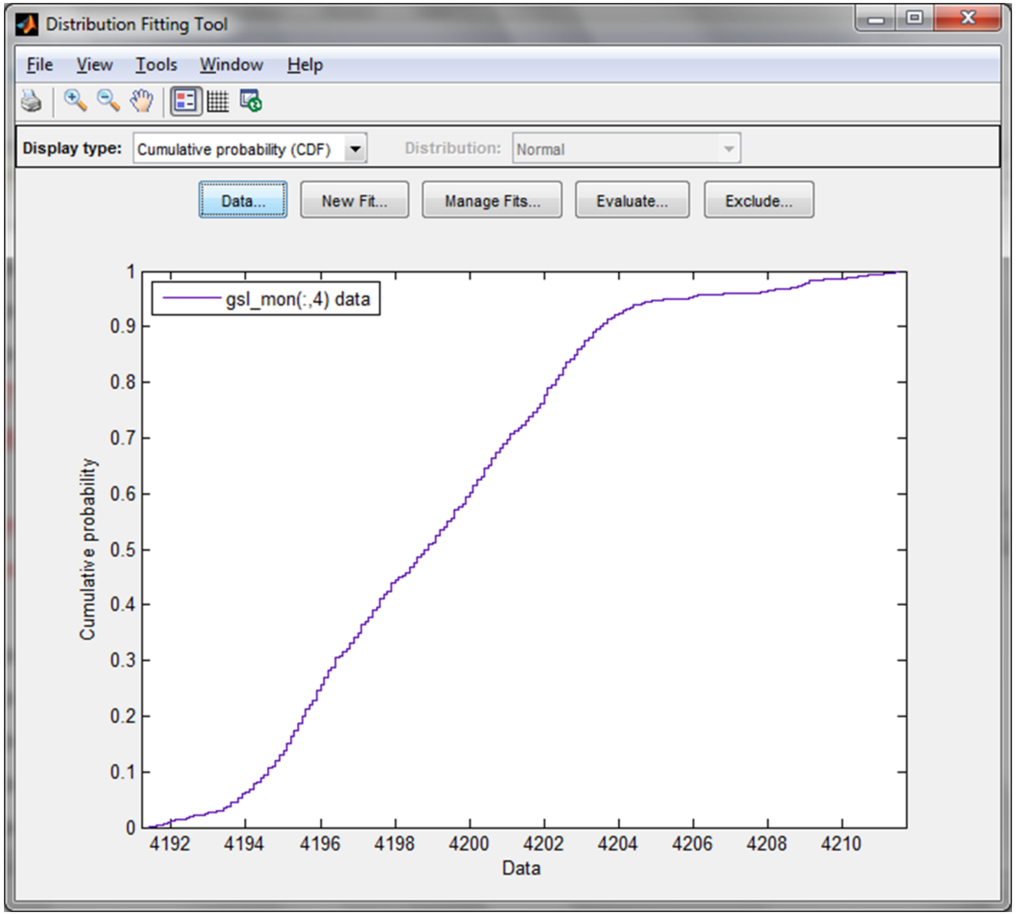
- $Pr\{E_1 \cap E_2\}$- joint probability that E_1 & E_2 occur
- $Pr\{E_1 \cap E_2\} = 0$ if E_1 & E_2 are mutually exclusive
- $Pr\{E_1 \cup E_2\}$- probability that E_1 OR E_2 occur
- $Pr\{E_1 \cup E_2\} = Pr\{E_1\} + Pr\{E_2\} - Pr\{E_1 \cap E_2\}$

Conditional Probability

- Conditional probability: probability of $\{E_2\}$ given that $\{E_1\}$ has occurred
- $Pr\{E_2 \mid E_1\} = Pr\{E_1 \cap E_2\} / Pr\{E_1\}$
- E_1 is the conditioning event
- If E_1 and E_2 are independent of each other, then
  $Pr\{E_2 \mid E_1\} = Pr\{E_2\}$ and $Pr\{E_1 \mid E_2\} = Pr\{E_1\}$
- Fair coin- Pr{heads} = 0.5
  - chance of getting heads on second toss is independent of the first
    Pr{heads | heads} = 0.5
    Pr{heads} twice = 0.5 * 0.5 = 025

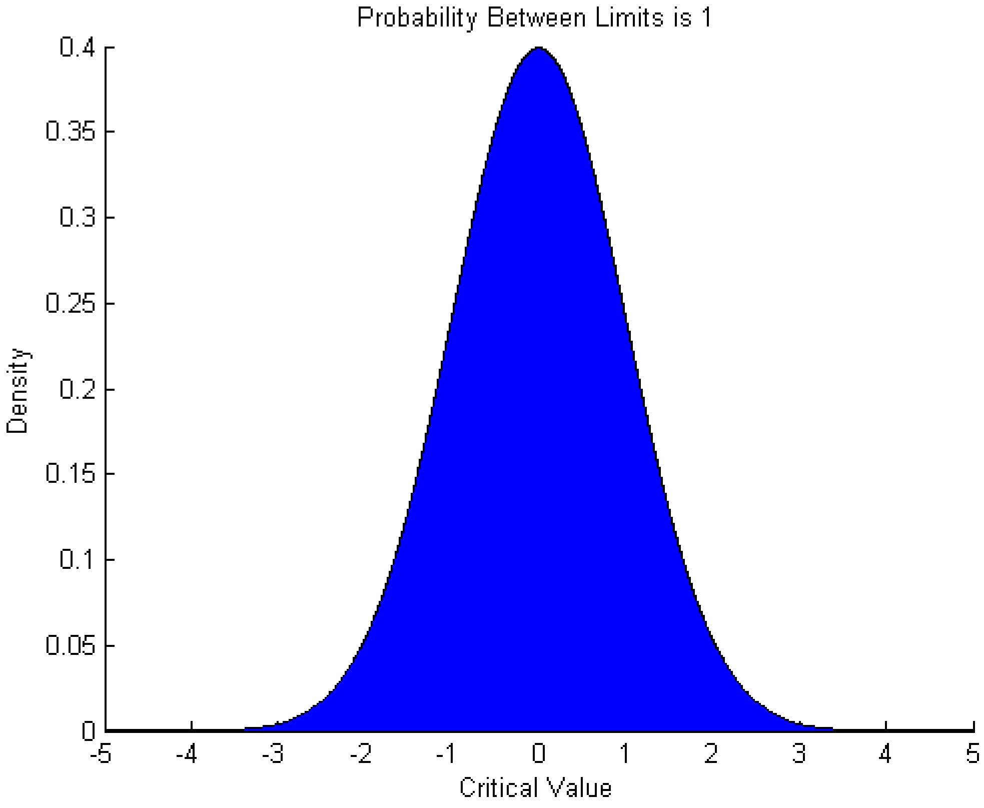
Empirical vs. Parametric Distributions

- Empirical distributions derived from a sample of a population



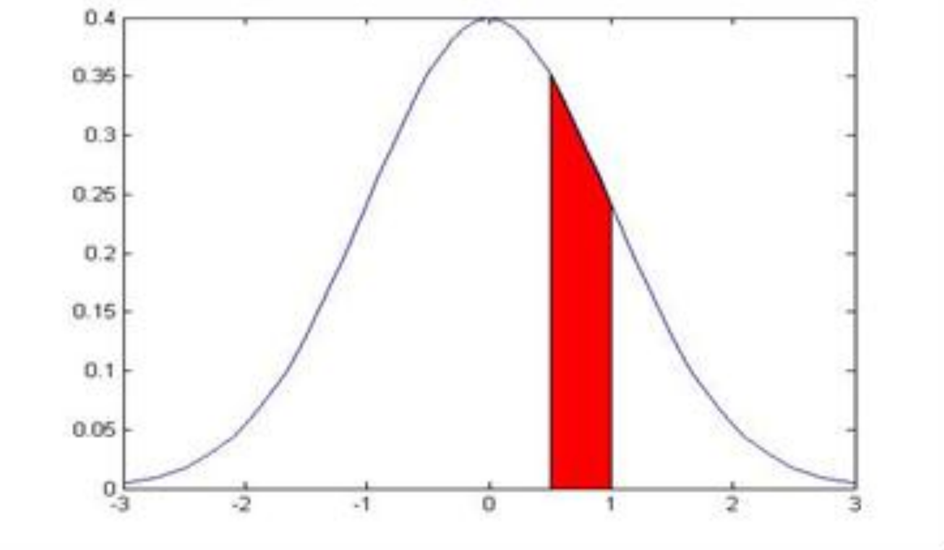
Empirical vs. Parametric Distributions

- Parameteric distributions:
  - Theoretical approach to define populations with known properties
  - Can be defined by a function with couple parameters and assumption that population composed of random events



Random Continuous Variable x

- f(x) probability density function (PDF) for a random continuous variable x
- f(x)dx incremental contribution to total probability



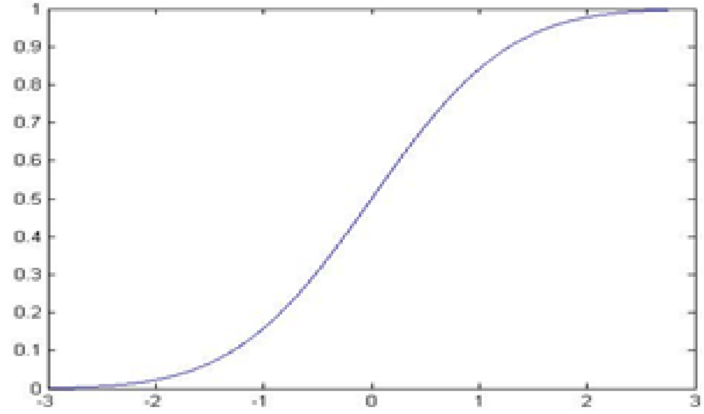
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Cumulative Density Function of Continuous Variable

- F(X)- total probability below a threshhold
- F(0) = 50%
- F(.66) = 75%

$$F(X) = \Pr\{x \le X\} = \int_{-\infty}^{X} f(x)dx$$

- X(F) – quantile function- value of random variable corresponding to particular cumulative probability
- X(75%) = 0.66



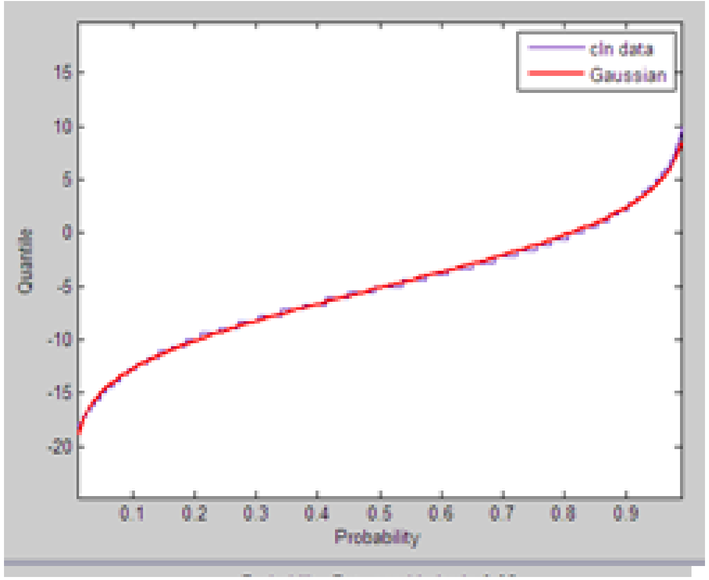
Gaussian Parametric Distribution

- PDF $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
- CDF $$F(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{X} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

- Two parameters define Gaussian distribution: μ and σ
- Nothing magic or “normal” about the Gaussian distribution- it is a mathematical construct

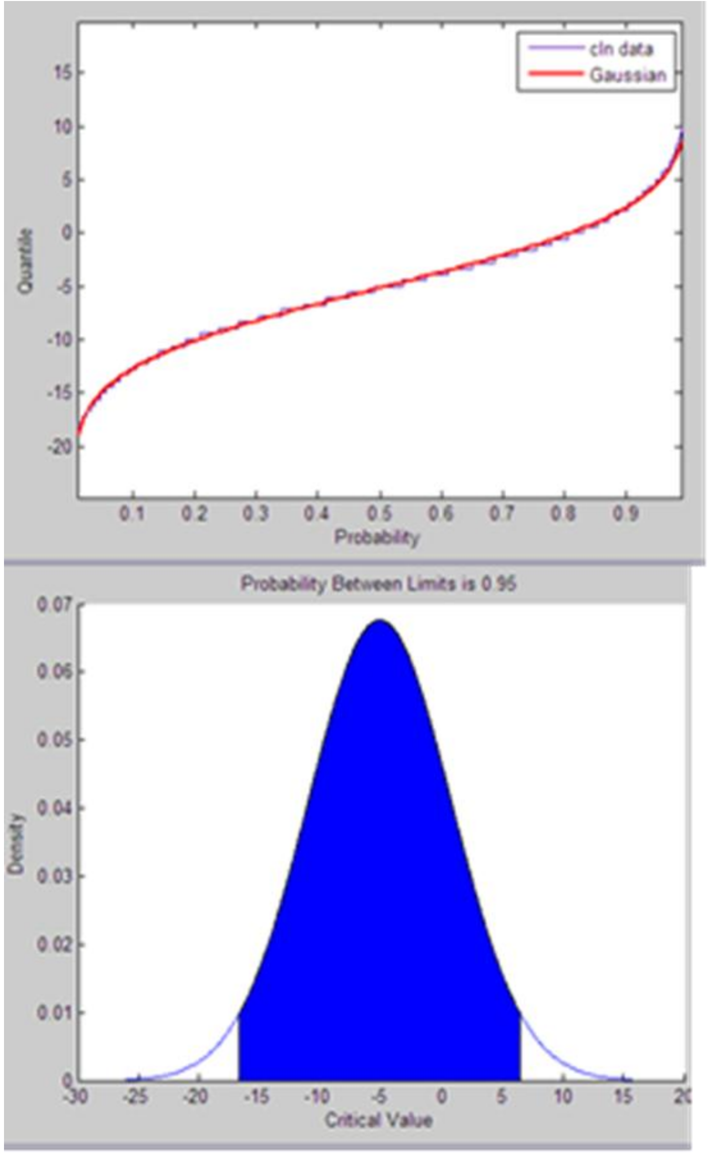
Hypothesis Testing

- Alta temperature:
- Empirically: probability of temperature less than -15 is low
- Empirical estimates:
  - Mean= -5.1C
  - Std dev = 5.9C
- What are chances of getting temp of -20 IF this was a population of random numbers with that mean and std dev?



Null hypothesis

- Null hypothesis: Temp of -20C does not differ significantly from mean of -5.1C
- 95% of time, random value would be within -16 and 6C
- So 5% of time, random value would be outside this range
- REJECT the null hypothesis accepting a 5% risk that we are rejecting the null hypothesis incorrectly
- If null hypothesis: Temp of -15C does not differ significantly from mean of -5.1C
- CANNOT reject the null hypothesis since 95% of the time the value could be within -16 and 6C



Students' t test

- $\sigma_{\overline{X}} = \frac{s_x}{\sqrt{n-1}}$
- Estimate of population variance from sample
- T value:
- Numerator: signal
- Denominator: noise
- At t gets larger, confidence in rejecting the null hypothesis (sample mean differs from population mean) gets higher
- T large IF:
  - Spread between sample and population means large
  - Degrees of freedom is large
  - Variability in sample is small

$$t = (\bar{x} - \mu)\sqrt{n-1} / s_x$$

Using t test

- **[h,p,ci,stat]= ttest(valy,0,.05,'left');**
- where on input valy is the vector of values in each 5-year sample
- 3.8, -3.9, -3.6, -2.9, 2.3
- 0 is the mean value for the null hypothesis
- .05 is the significance level chosen (5%)
- ‘left’ indicates that we are assuming that we have ruled out that large positive anomalies are relevant (the other options are ‘both’ a two-tailed test and ‘right’ where we rule out large negative anomalies, i.e., look for wet periods)
- Output:
  - h is a flag, 0 means the null hypothesis can not be rejected, 1 means it can be rejected
  - p is the significance level corresponding to the t value, the smaller the number the better
  - ci is the confidence interval
  - stat- is an array that returns the value of the t statistic, the number of degrees of freedom, and the estimated population standard deviation
- in this case, h= 0., p = .31 (really big), ci= -2.6, sample mean needs to be less than -2.6 to reject null hypothesis

Estimating Values of One Variable From Another

- X- Ben Lomond Trail
- Y- Ben Lomond Peak
- Want to estimate Peak from Trail
- Use pairs of observations from sample
- Need to determine coefficient b or r
- b- slope of linear estimate
- r- linear correlation

$$\hat{y}_i = \bar{y} + b(\hat{x}_i - \bar{x})$$

$$\hat{y}^*{}_i = r\hat{x}^*{}_i$$

Definitions

- Estimate $\hat{y}_i = \bar{y} + b(\hat{x}_i - \bar{x})$
- Error of estimate $e_i = y_i' - \hat{y}_i'$
- Want $\sum_{i=1}^{n} e_i^2$ to be a minimum
- Need to find the value of b that minimizes that sum

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} e_i^2 = 0$$

- The value of b that minimizes the total error in the sample $b = \overline{x_i' y_i'} / \overline{(x_i')^2} = \overline{x_i' y_i'} / s_x^2$

Covariance

- Relates how departures of x and y from respective means are related
- Units are the product of the units of the two variables x and y
- Large and positive if sample tendency for:
  - large + anomalies of x occurring when large + anomalies of y AND
  - large - anomalies of x occurring when large - anomalies of y
- Large and negative if sample tendency for:
  - large + anomalies of x occurring when large - anomalies of y AND
  - large - anomalies of x occurring when large + anomalies of y
- Near zero when tendency for cancellation
  - large + anomalies of x occurring when both large – and + anomalies of y AND
  - large - anomalies of x occurring when both large – and + anomalies of y

Linear Correlation

$$r^2 = b^2 s_x{}^2 / s_y{}^2 = (\overline{x_i' y_i'})^2 / (s_x{}^2 s_y{}^2) \qquad r = (\overline{x_i' y_i'}) / \sqrt{\overline{x_i'^2}\,\overline{y_i'^2}}$$

$$x_i{}^* = x_i' / s_x, y_i{}^* = y_i' / s_y, r = (\overline{x_i^* y_i^*})$$

$$1 = r^2 + \frac{\overline{e_i^2}}{s_y{}^2}$$

y's total sample variance = fraction of variance estimated by x + fraction of variance NOT explained by x

- Dimensionless number relates how departures of x and y from respective means are related taking into account variance of x and y
- r = 1. Linear fits estimates ALL of the variability of the y anomalies and x and y vary identically
- r = -1 perfect linear estimation but when x is positive, y is negative and vice versa
- r = 0. linear fit explains none of the variability of the y anomalies in the sample. Best estimate of y is the mean value

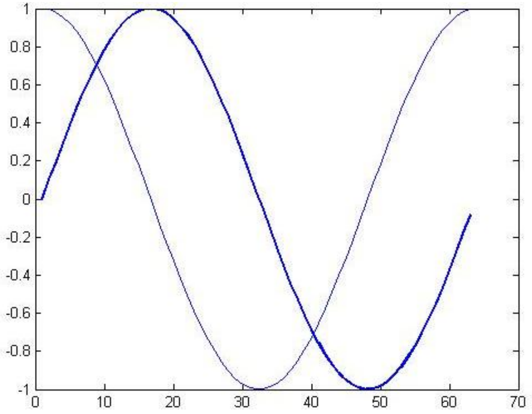
Linear Algebra is your friend

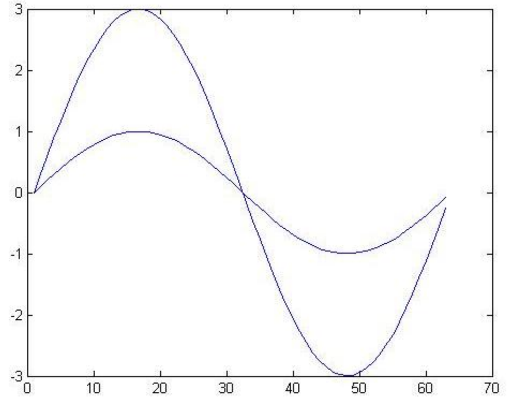
$$\vec{X}' = \begin{bmatrix} x'_1 \\ x'_2 \\ \ldots \\ x'_n \end{bmatrix} \quad \vec{Y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \ldots \\ y'_n \end{bmatrix}$$

$$\overline{x'_i y'_i} = \vec{X}'^T \vec{Y}' / n$$

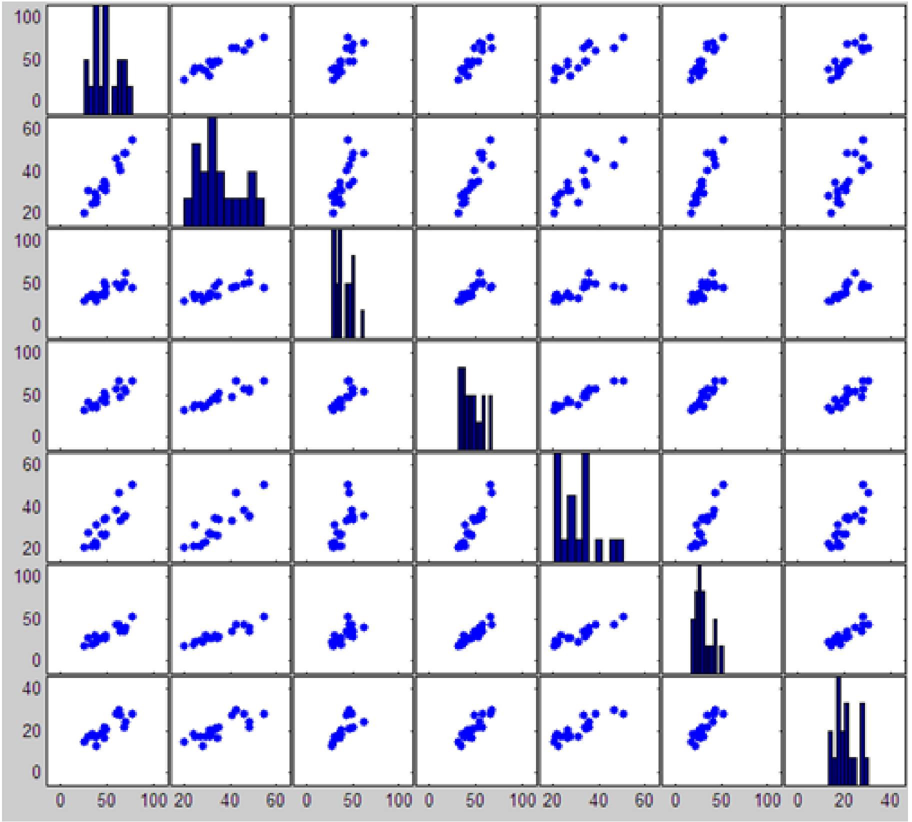
Stop and think before blindly computing correlations

- tendency to use correlation coefficients of 0.5-0.6 to indicate “useful” association.
  - 75%-64% of the total variance is NOT explained by a linear relationship if the correlation is in that range
- linear correlations can be made large by leaving in signals that may be irrelevant to the analysis. Annual and diurnal cyles may need to be removed
- large linear correlations may occur simply at random, especially if we try to correlate one variate with many, many others
- relationships in the data that are inherently nonlinear will not be handled well
- when two time series are in quadrature with one another then the linear correlation is 0
- Linear correlation provides no information on the relative amplitudes of two time series



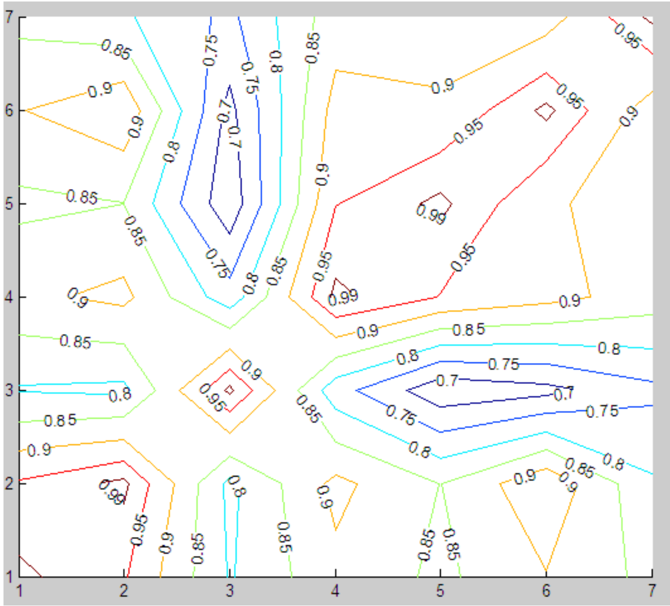


Multivariate Linear Correlations



$$\vec{X}^{*} = \begin{bmatrix} x^{*}_{11} & x^{*}_{12} & \cdots & x^{*}_{17} \\ x^{*}_{21} & x^{*}_{22} & \cdots & x^{*}_{27} \\ \cdots & \cdots & \cdots & \cdots \\ x^{*}_{n1} & x^{*}_{n2} & \cdots & x^{*}_{n7} \end{bmatrix}$$

$$\vec{R} = \vec{X}^{*T} \, \vec{X}^{*} / n$$



Describing the amount of variance explained by a linear relationship

$$s_y^{\ 2} = b^2 s_x^{\ 2} + \overline{e_i^2} \qquad ns_y^{\ 2} = nb^2 s_x^{\ 2} + n\overline{e_i^2}$$

- Sum of squares form:
- Total variance = explained variance + unexplained variance

ANOVA Table- Regression Form

Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST=n s_y^2	n-1	n s_y^2 /(n-1)	
Regression	SSR=n $b^2 s_x^{\ 2}$	1	MSR=n $b^2 s_x^{\ 2}$ /1	(n-2) $b^2 s_x^{\ 2}$ /($s_y^{\ 2} - b^2 s_x^{\ 2}$)
Error	SSE=n ($s_y^{\ 2} - b^2 s_x^{\ 2}$)	n-2	MSE= n ($s_y^{\ 2} - b^2 s_x^{\ 2}$)/(n-2)	

ANOVA

- summarize whether the variance of variable y explained by variable x is large in terms of three measures:
  - mean squared error of the regression (MSE),
  - variance explained by the regression (MSR),
  - and the F ratio that is assumed to have a known parametric form.
- We want:

1. the scatter around the line of best fit to be small, i.e., that SSE and MSE are small
2. the percent variance explained by the regression to be large (or MSR large)
3. F ratio is large, which is the ratio of the explained variance to that of the error

F Test of correlation coefficient

ANOVA Table- Correlation Form

Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST= n	n-1	n/(n-1)	
Regression	SSR=n r^2	1	MSR=n r^2 /1	(n-2)r ^{2} / $(1-r^2)$
Error	SSE=n (1- r^2)	n-2	MSE= n $(1-r^2)$/(n-2)	

- Signal: explained variance * (n-2)
- Noise: $(1-r^2)$

Compositing (Superposed Epoch)

- Identify common characteristics of a sample of events
- Simplest- average conditions before, during, and after some “rare” event
- Has an advantage over linear correlation since no linear assumption necessary
- Limitation- to what extent does sample mean used in composite differ from population?
- Day composites: http://www.cdc.noaa.gov/Composites/Day/
- Monthly/seasonal composites: http://www.cdc.noaa.gov/cgi-bin/data/composites/printpage.pl

Compositing Steps

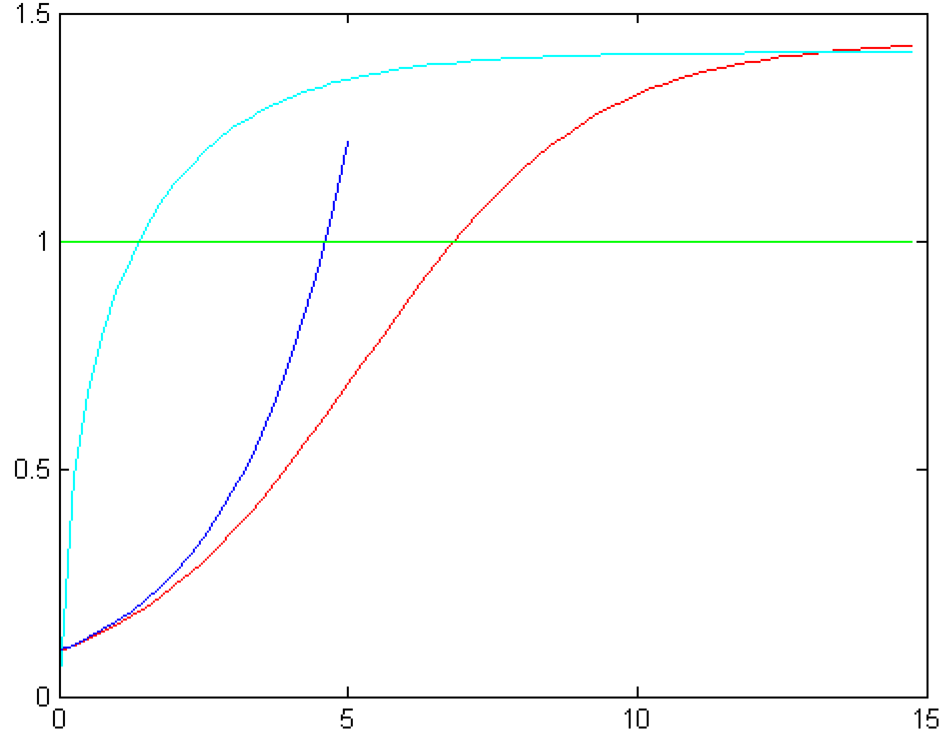
- Select the basis for compositing: why are you doing it? Physical reasoning hopefully?
- Define the categories on which you define the events: above, below normal? Or …?
- Compute the means and statistics for each category (minimum is standard deviation)
- Organize and display the results
- Validate the results:
  - Significance test? t test is the bare minimum to do
  - Reproduce in an independent sample?
  - Are the results sensible in space and time?
  - Is it consistent with theory?

Model: tool for simulating or predicting the behavior of a dynamical system such as the atmosphere

- heuristic: rule of thumb based on experience or common sense
  - Not strictly accurate or always reliable
  - Example: If the winds get strong, there'll be a lot of damage
- conceptual: framework for understanding physical processes based on physical reasoning
  - Very useful- that's what fills textbooks
  - Example: LIMBS
- empirical: prediction based on past behavior
  - Can tell us what has been likely in the past: record values, typical values, etc.
  - Example: average daily temperature in June vs. January
- analytic: exact solution to "simplified" equations that describe the atmosphere
  - Very useful to understand how things work
  - Example: many of the conceptual models described in the textbook rely on analytic models
- numerical: integration of governing equations by numerical methods subject to specified initial and boundary conditions
  - What is used for day-to-day weather forecasting
  - Example: Global forecast system (GFS) model

Comparing Error Growth of Perfect Model to Climo and Persistence

Green- climo forecast
cyan- persistence
Blue- exponentional error growth
Red- slower error growth



Look at forecast_error.m

Summary

persistence forecast is better empirical forecast than a climatological forecast at short lead times

numerical weather prediction models should out perform persistence and climatological forecasts at lead times out to some lead time

for all models: model accuracy is often evaluated in a least squared sense relative to the unknown truth.

Course Learning Objectives

- State and use basic statistical metrics to analyze environmental information
- Develop proficiency to program and use MATLAB software as a tool to analyze environmental data sets
- 5040: State and demonstrate the characteristics of effective research; organize, quality control, and find relationship(s) among data
- 6040:State and demonstrate the characteristics of effective research: distill a general interest in a subject into a specific question/hypothesis that can be evaluated; organize the data; find relationship(s) among the data; and examine the significance of the results

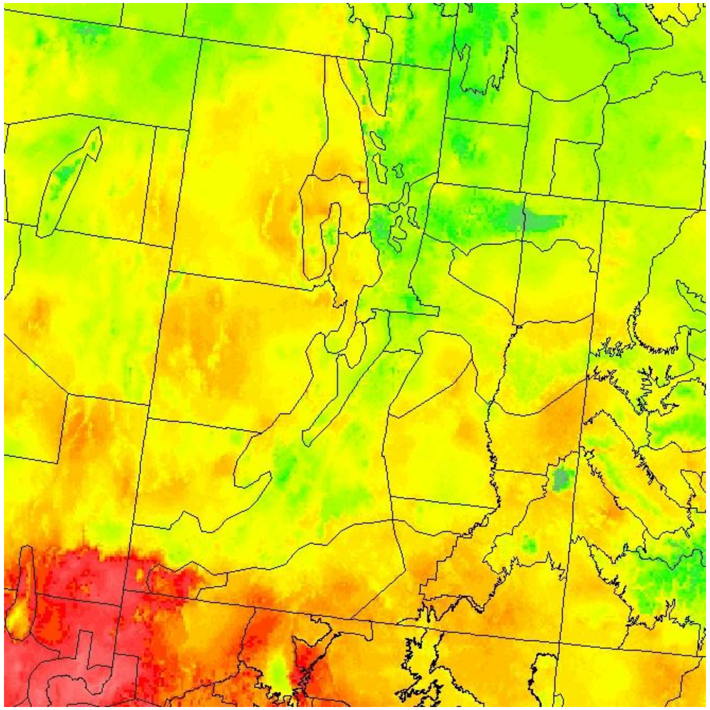
- Undergrads: perception of half semester classes
- All: perception of this class- pace, content, assignments, expectations, etc.

Rest not covered unless time

Statistical Interpolation

- A common goal in environmental fields is to take observations of environmental conditions scattered over a spatial domain and interpolate/extrapolate those values to a regular grid.
- Simple schemes (Cressman) were developed in the atmospheric sciences over 50 years ago to give greater weight to observations close to the location at which the analysis value is desired compared to more distant observations
- Rather than attempting to interpolate fields without any other information, early researchers recognized that defining a “first guess” or “background” from a source such as a model forecast and weighting corrections between the observations and first guess fields was a superior approach.

Objective Analysis



- A map of a meteorological field
- Relies on:
  - observations
  - background field
- Used for:
  - Initialization for a model forecast
  - Situational awareness
  - Verification grid

Discussion Points

- Why are analyses needed?
  - Application driven: data assimilation for NWP (forecasting) vs. objective analysis (specifying the present or past)
- What are the goals of the analysis?
  - Define microclimates?
    - Requires attention to details of geospatial information (e.g., limit terrain smoothing)
  - Resolve mesoscale/synoptic-scale weather features?
    - Requires good prediction from previous analysis
- How is analysis quality determined? What is truth?
  - Evaluating analysis by withholding observations

Discussion Points (cont.)

- What causes large variations in surface temperature, wind, moisture, precipitation over short distances?
  - Terrain, convection, etc.

- How well can we observe, analyze, and forecast conditions near the surface?
  - What errors should we tolerate?

- To what extent can you rely on surface observations to define conditions within 2.5 x 2.5 or 5 x 5 km^2 grid box?
  - Do we have enough observations to do so?

ABC's

Analysis value = Background value + observation Correction

- An analysis is more than spatial interpolation
- A good analysis requires:
    - a good background field supplied by a model forecast
    - observations with sufficient density to resolve critical weather and climate features
    - information on the error characteristics of the observations and background field
    - appropriate techniques to translate background values to observations (termed "forward operators")

Objective Analysis Approaches

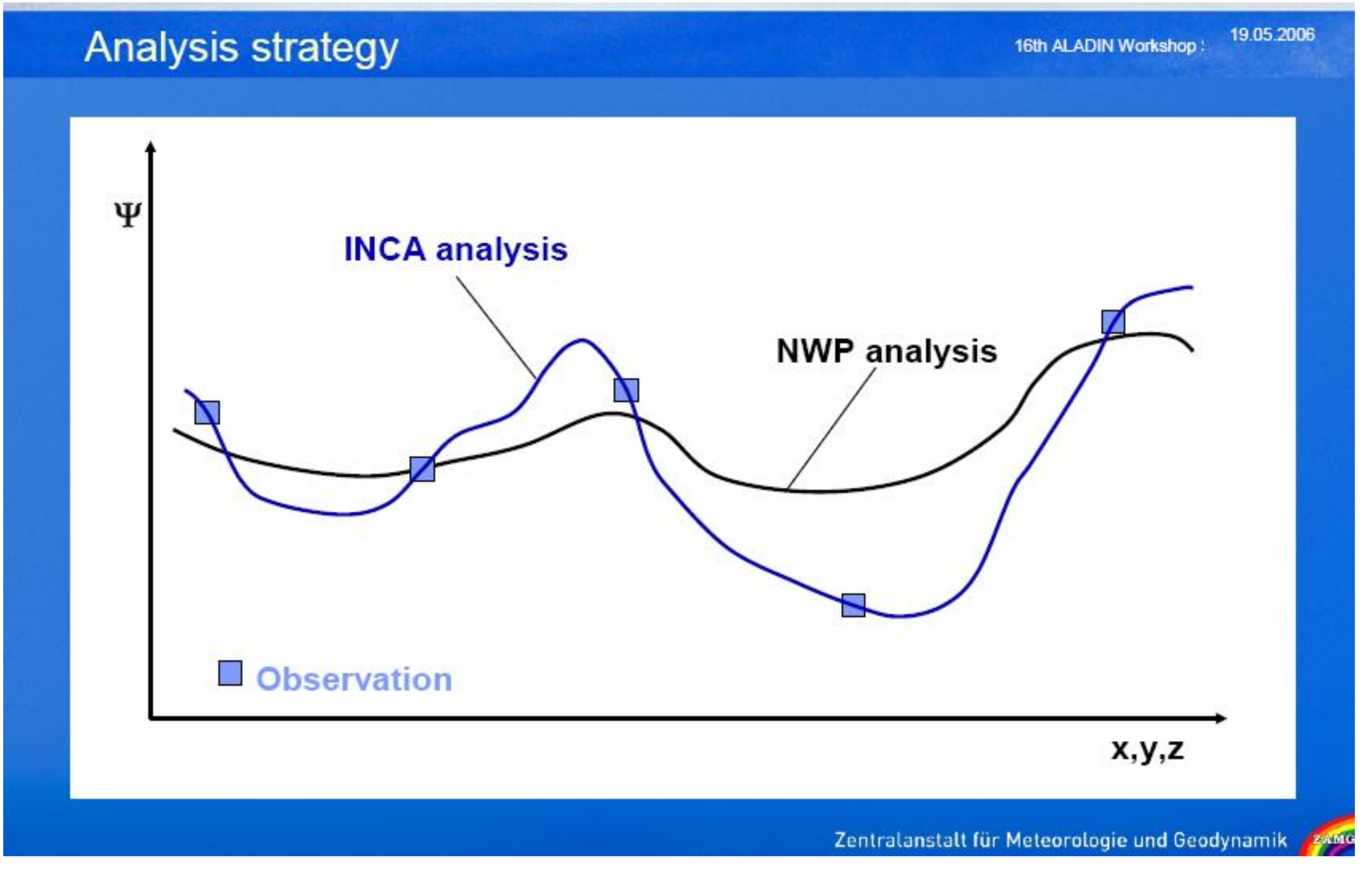
- Successive Corrections
- Optimal Interpolation
- Variational (2DVar,3DVar, 4DVar)
- Kalman or Ensemble Filters

simple

complex

- Kalnay (2003) Chapter 5 – good overview of different schemes

One Approach: Adjust Model Guidance to Match Observations (INCA and MatchObsAll)

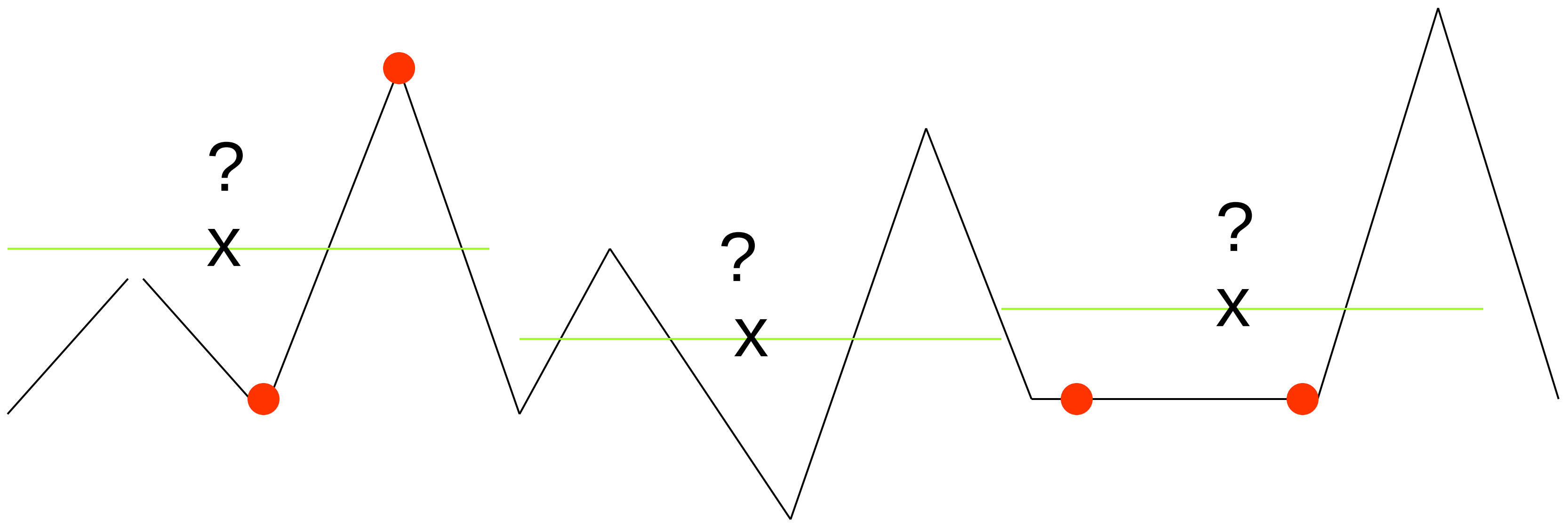


Potential for Confusion

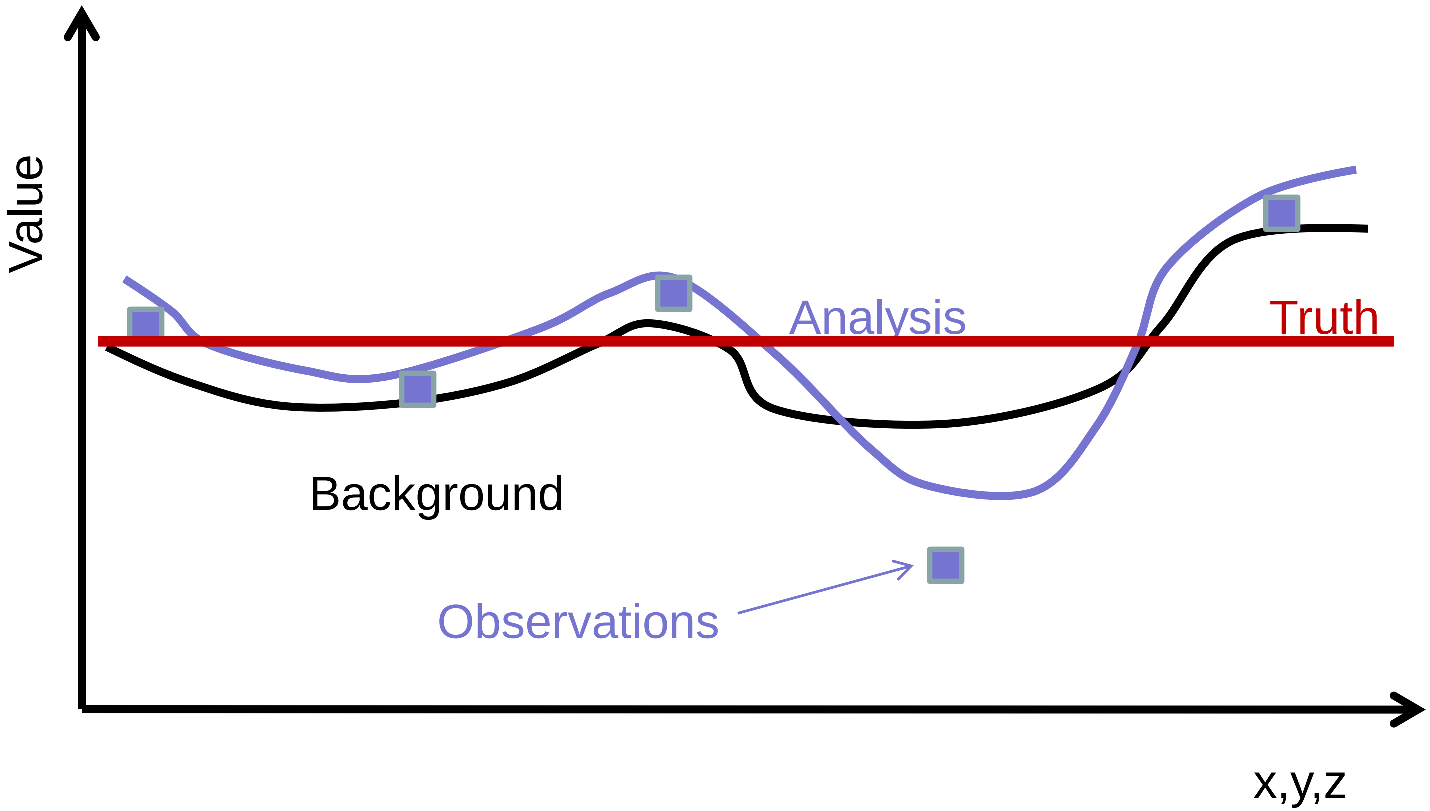
- Analysis systems like INCA suggest that the analysis should exactly match every observation
- Variational or other analysis values usually don't match surface observations
  - Analysis schemes are intended to develop the "best fit" to the *differences* between the observations and the background taking into account observational and background errors when evaluated over *a large sample* of cases

What are appropriate analysis gridpoint values?

- Inequitable distribution of observations
- Differences between the elevations of the analysis gridpoints and the observations

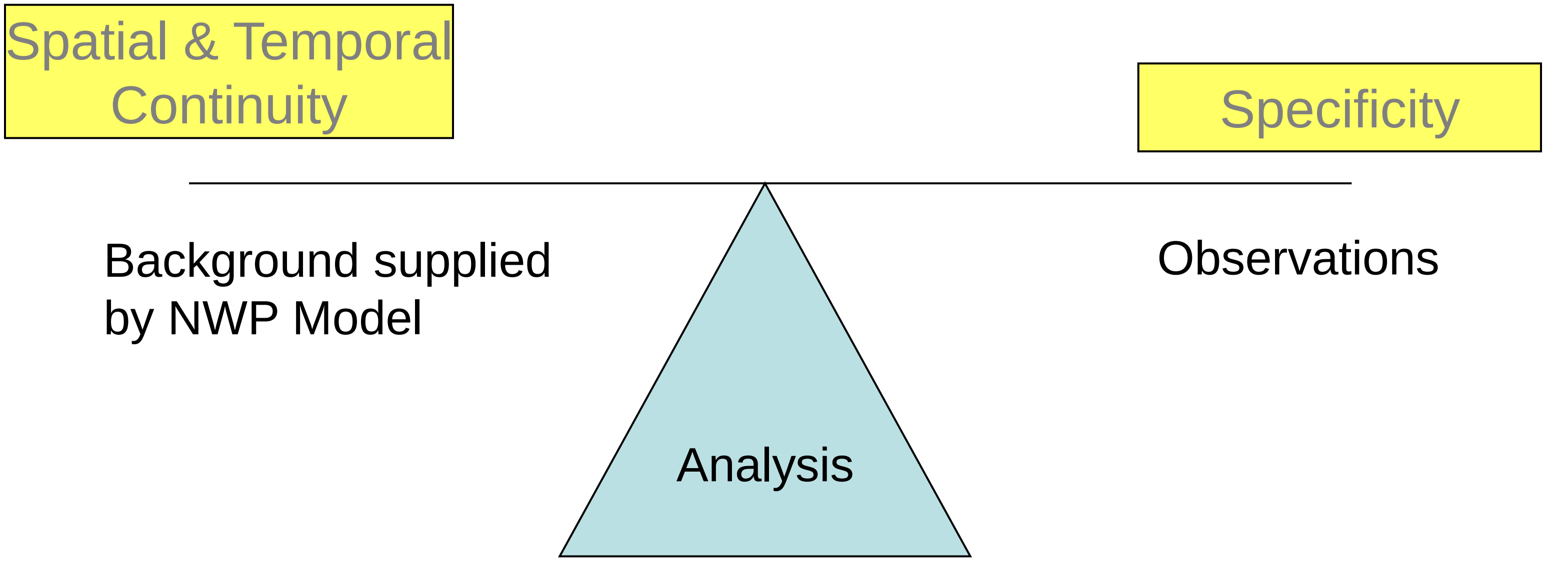


Predominant Approach: Constrain Imperfect Model Guidance by Imperfect Observations

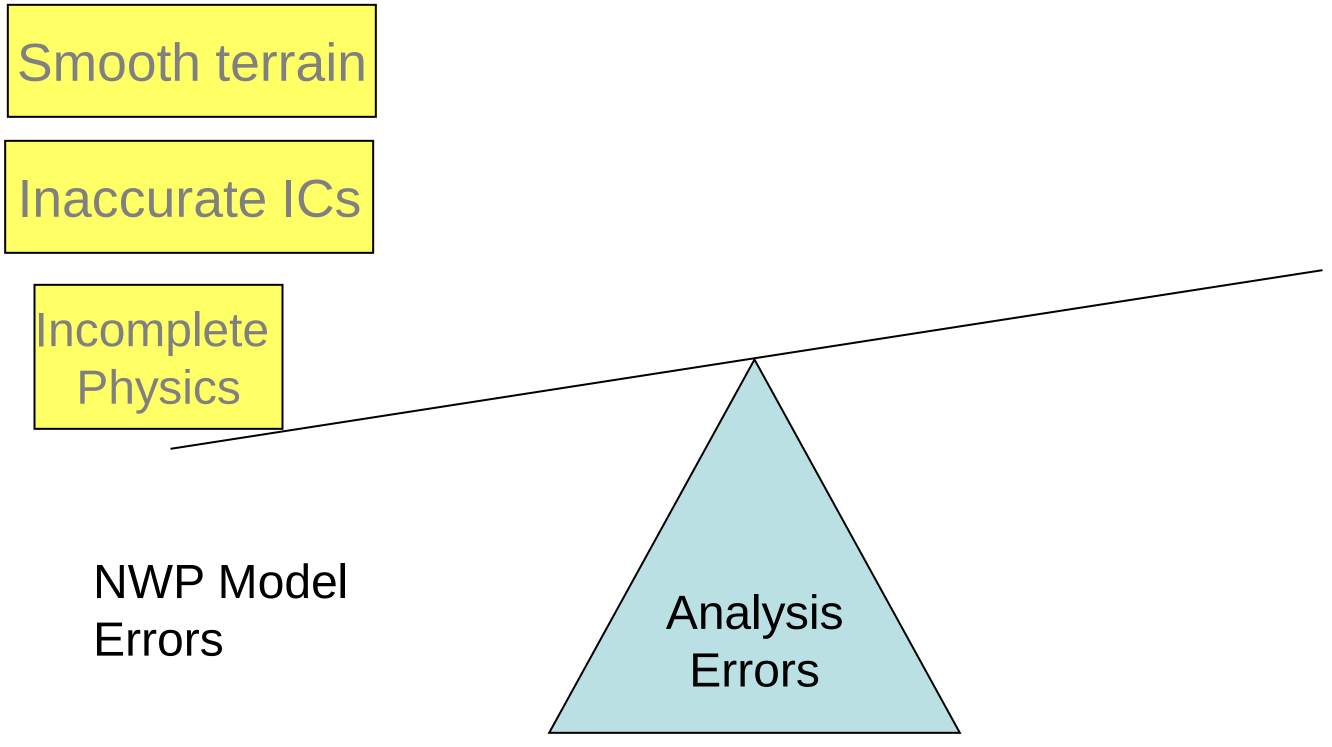


Need for balance…

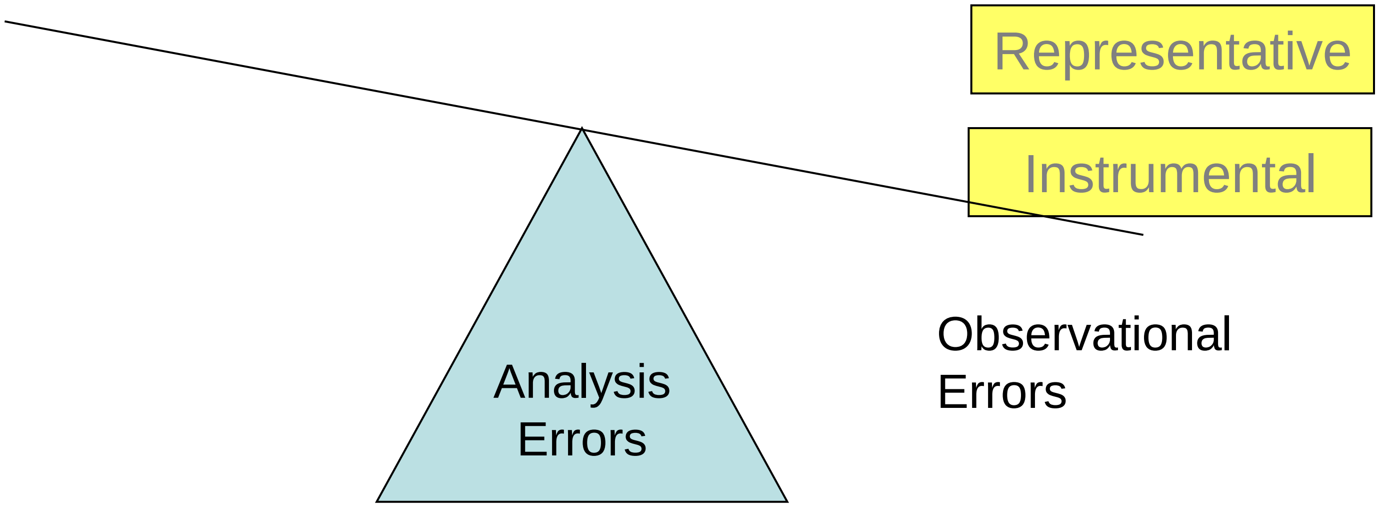
Models or observations cannot independently define weather and weather processes effectively



Recognition of Sources of Errors



Recognition of Sources of Errors

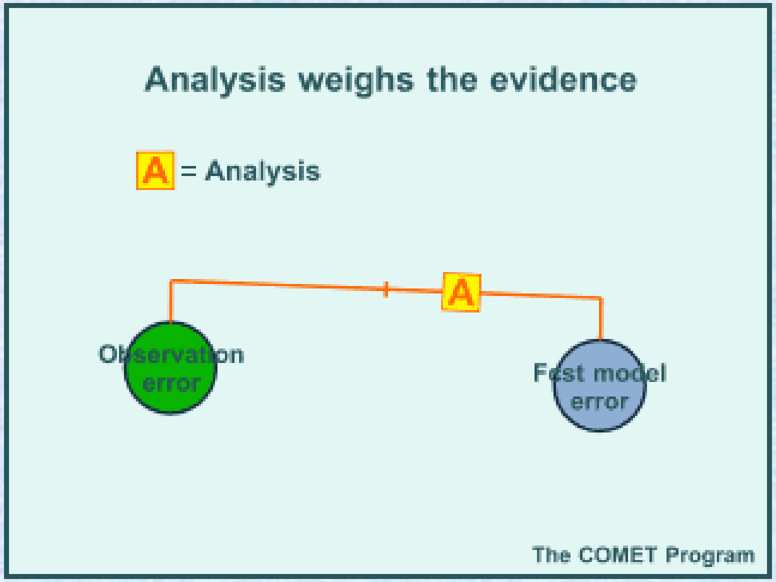


Which is weighted more – observation or model value?

- The analysis procedure (2D-VAR) "knows" the value and limitations of observations using expected observation errors for each data type
- It "knows" model's behavior by using model forecast error statistics at each grid point and spatial relationships of error patterns
- The analysis assesses penalties-
  - Penalty for deviations from observations
    - Larger penalty if observation type is known to have smaller error
  - Penalty for deviations from background
    - Larger penalty if model forecast is usually good
  - Scheme chooses analysis that pays the smallest total penalty for observations and model combined

- We want the analysis to:
  1. Draw closer to better quality data
  2. Retain more details in the background from a better quality model

But the weighting may be incorrect if error statistics are not appropriate for today's weather



Background Values

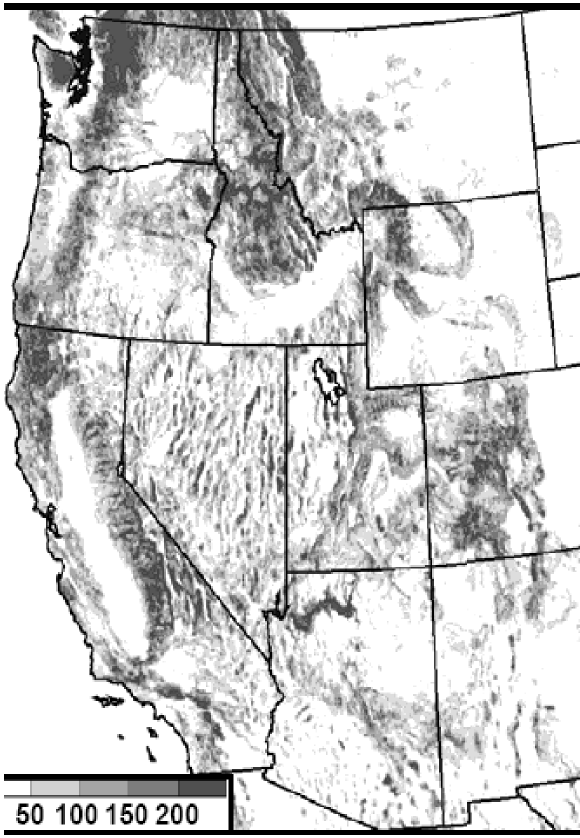
- Obtained from an analysis:
  - Climatology or analysis from prior hour
  - An objective analysis at a coarser resolution
  - Short term forecast
- Most objective analysis systems account for background errors but approaches vary

Observations

- Observations are not perfect…
  - Gross errors
  - Local siting errors
  - Instrument errors
  - Representativeness errors
- Most objective analysis schemes take into account that observations contain errors but approaches vary

Representativeness Errors

- Observations may be accurate…
- But the phenomena they are measuring may not be resolvable on the scale of the analysis
  - This is interpreted as an error of the *observation* not the analysis
- Common problem over complex terrain
- Also common when strong inversions
- Can happen anywhere



Sub-5km terrain variability (m)
(Myrick and Horel, WAF 2006)

Incorporating Errors

- Basic example:

$$T_a = T_b + W(T_o - T_b) \qquad W = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2}$$

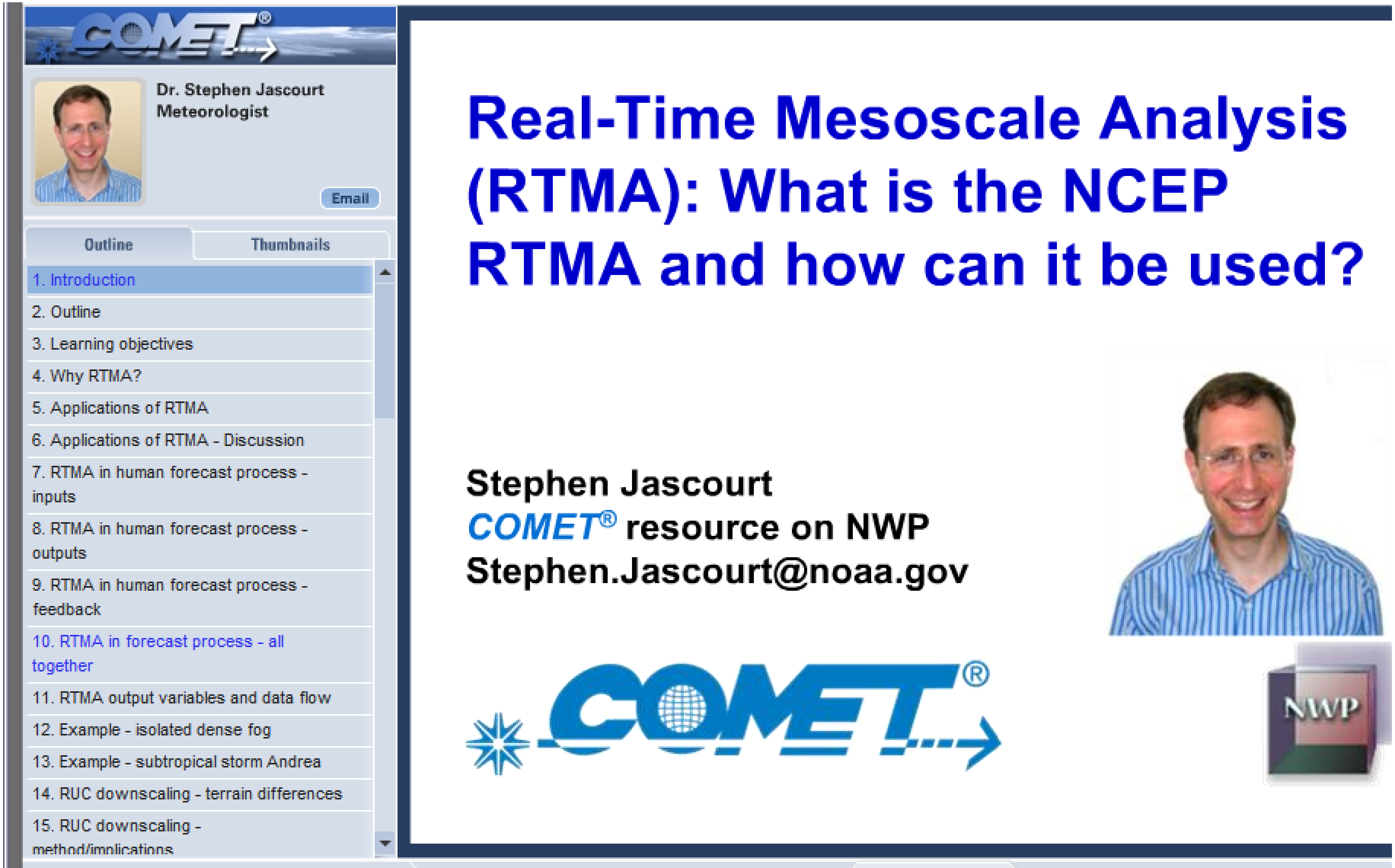
σ_b = background error variance

σ_o = observation error variance

W = 0, distrust observation

W = 1, trust observation

More Info… www.meted.ucar.edu



The actual ABCs…

- The RTMA analysis equation looks like:

$$\left(\vec{P}_b^{\,T} + \vec{P}_b^{\,T}\vec{H}^T\vec{P}_o^{-1}\vec{H}\vec{P}_b\right)\vec{v} = \vec{P}_b^{\,T}\vec{H}^T\vec{P}_o^{-1}\left[\vec{y}_o - \vec{H}\left(\vec{x}_b\right)\right]$$

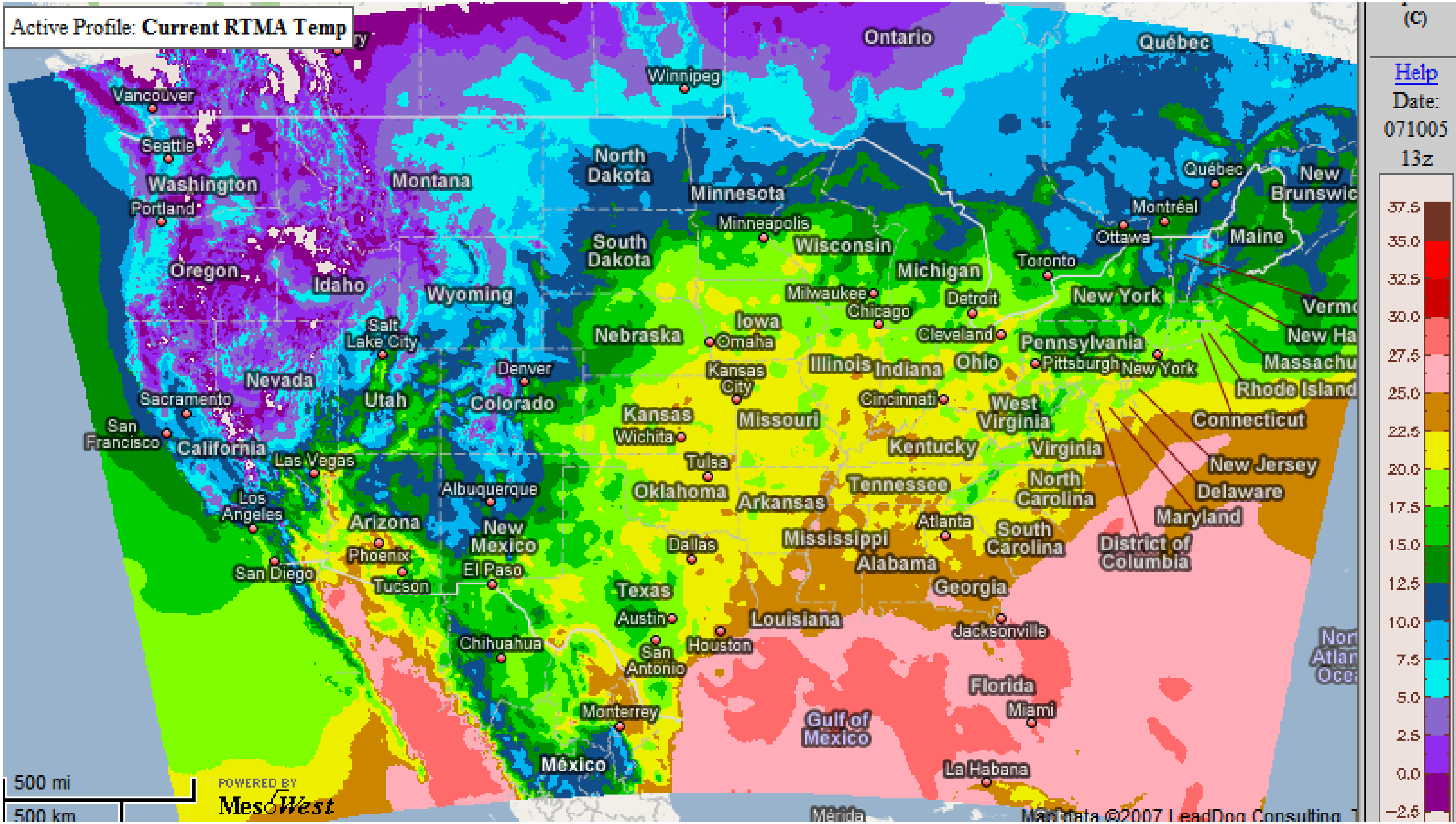
$$\vec{x}_a = \vec{x}_b + \vec{P}_b\vec{v}$$

- Covariances are error correlation measures between all pairs of gridpoints
- Background error covariance matrix can be extremely large
  - 2,900 GB memory requirement for continental scale
  - Recursive filters significantly reduce this demand

Estimation of Observation and Background Error Covariances

- Temperature errors at two gridpoints may be correlated with each other
- Error covariances specify the influence of observation innovations upon surrounding gridpoints
- RTMA used decorrelation lengths of:
  - Horizontal (*R*): 40 km
  - Vertical (*Z*): 100 m
  - Now increased to ~80 km and 200 m respectively
- *Significant limitation to specify error covariances rather than determine them through ensemble methods*

RTMA CONUS Temperature Analysis



Local Surface Analysis

- Solving linear system of form Ax=b using GMRES- generalized minimal residual method

$$\left(\overrightarrow{P_b}' + \overrightarrow{P_b}'\,\overrightarrow{H}'\,\overrightarrow{P_o}^{-1}\,\overrightarrow{H}\,\overrightarrow{P_b}\right)\vec{v} = \overrightarrow{P_b}'\,\overrightarrow{H}'\,\overrightarrow{P_o}^{-1}\left(\overrightarrow{y_o} - \overrightarrow{H}\left(\overrightarrow{x_b}\right)\right)$$

$$\overrightarrow{x_a} = \overrightarrow{x_b} + \overrightarrow{P_b}\vec{v}$$

- In matlab x= gmres(A,b)

Assumptions affecting the Analysis

1. **Statistical Assumptions**
   - Observation and model errors are assumed to follow normal distributions
     - Works well for common cases, but not extreme
     - Can’t distinguish different model performances in different regimes
     - Can’t distinguish different local conditions
2. **Assumed Observation Error**
   - Instrument (well known – an engineering matter)
   - Representativeness (not well known)
     - Accurate observation doesn’t represent average value over entire grid box
   - Observation error should vary by weather scenario – but no one knows how to do this
3. **Assumed Background Error**
   - Based on model performance statistics
   - If model performs differently than it usually does for this type of situation, model errors may be inappropriate
4. **Assumed Balance Constraint: Generally Not Known On Mesoscale**
   - Mass –Wind linkage is loosely enforced
   - An “initialization” or “spin-up” step is no longer necessary – balance is achieved within the analysis itself

Summary

- Improving current analyses such as RTMA requires improving observations, background fields, and analysis techniques
  - Increase number of high-quality observations available to the analysis
  - Improve background forecast/analysis from which the analyses begin
  - Adjust assumptions regarding how background errors are related from one location to another
- Future approaches
  - Treat analyses like forecasts: best solutions are ensemble ones rather than deterministic ones
  - Depend on assimilation system to define error characteristics of modeling system including errors of the background fields
  - Improve forward operators that translate how background values correspond to observations

Trauth Text: What Did/Didn't Get Covered

Chapter 1: all

Chapter 2: all

Chapter 3: all but discrete theoretical distributions, chi squared test

Chapter 4: 4.1-4.4

Grad students second half: Chapters 5, 6, 9 and…(?)