Climate Modeling

Chapter 3. Fundamental analysis of climate data

Chapter 4. Regional climate change projections

Tan Phan-Van

VNU Hanoi University of Science

Outline

Lecture 6

- 3.1 Statistical analysis of four-dimensional atmospheric data

Lecture 7

- 3.2 Statistical analysis of global climate projection data

Lecture 8

- 4.1 Statistical downscaling

Lecture 6:

3.1 Statistical analysis of four-dimensional atmospheric data

Some important statistical characteristics

- Two approaches:
  - **Descriptive statistics**: Relates to the organization and summarization of data
    - Pre- vs Post-Processing
    - Observed vs Modeled data
    - Station based vs Gridded data
    - Predicted vs Reanalysis data
    - Surface vs Satellite Observation
  - **Inferential statistics**: Methods and procedures used to draw conclusions regarding underlying processes that generate the data
    - Physical understanding of atmospheric phenomena
    - Uncertainty analysis

Some important statistical characteristics

- Elements of Probability:
  - Events: An event is a set, or class, or group of possible uncertain outcomes
    - Rain (heavy rain/hot weather/cold weather,…) might or might not on a given day in Hanoi
  - Probability of events: The probability of an event is a number describing the chance that the event will happen
    - The probability of an event is the likelihood of that event occuring
    - The probability of event A is approximately the ratio between the number of times that A is observed and the number of repetitions of the experiment
    - $0 \leq P(A) \leq 1$
    - If A is certain to happen then $P(A) = 1$, if A cannot possibly happen then $P(A) = 0$

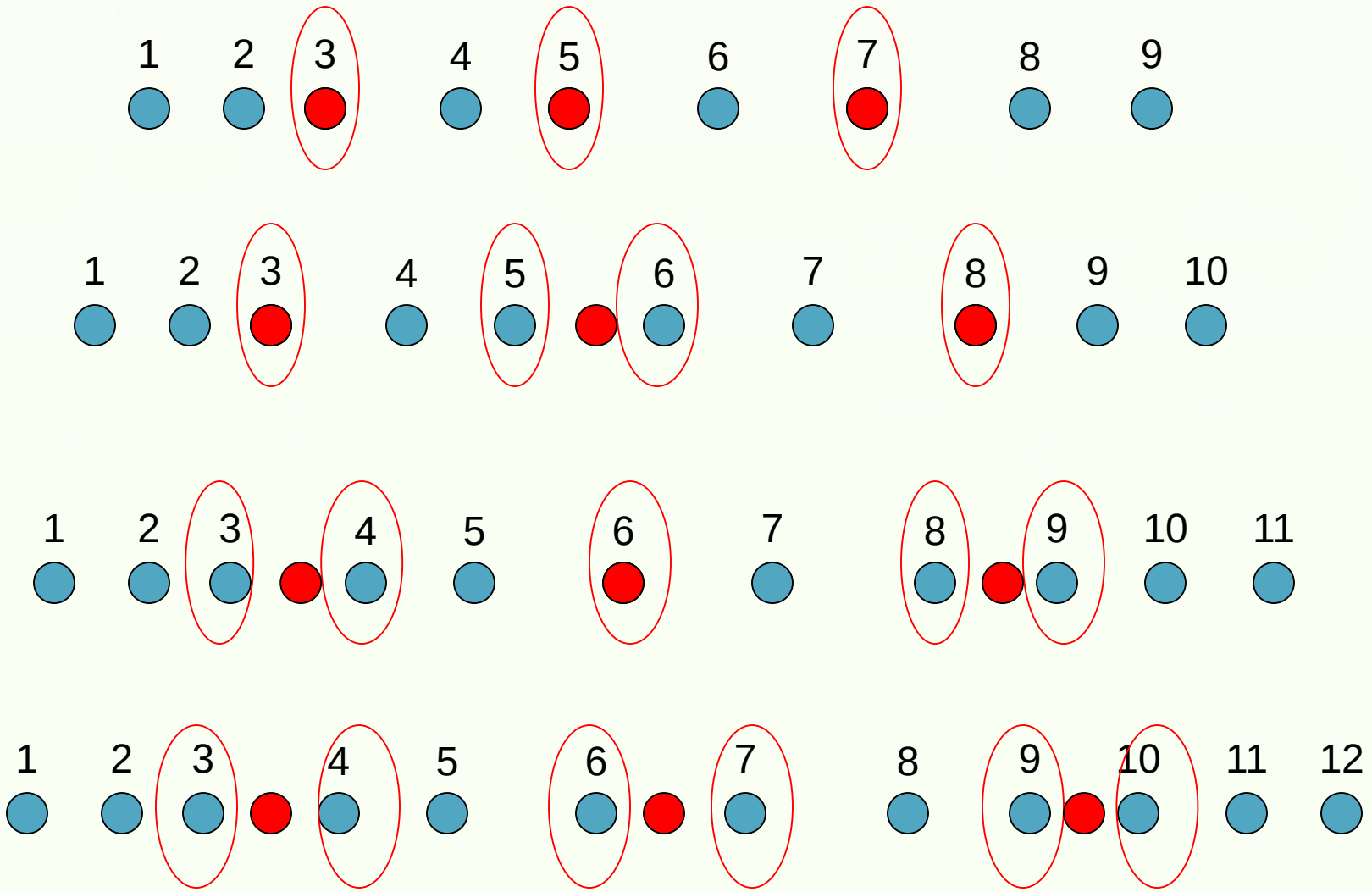
Some important statistical characteristics

- Elements of Probability:
  - Conditional Probability: A conditional probability is the probability of an event, given some other event has already occurred
    - Probability of coastal wind speeds above some threshold, given that a typhoon makes landfall nearby
  - Frequency: is the number of times that event occurs
    - Number of rainy days in the year
  - Relative Frequency: is the ratio between the number of times that event occurs and the number of repetitions of the experiment
    - Number of rainy days divides to the number of observed days
    - Relative Frequency is usually used to estimate the probability of events
  - Cumulative relative frequency: is the accumulation of the previous relative frequencies

Some important statistical characteristics

- Statistical characteristics:
  - Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - Standard Deviation: $S_x = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
  - Percentiles: A percentile is a number (of data series) where a certain percentage of scores fall below that number
    - The 25th percentile is called the first quartile (q_{25})
    - The 50th percentile is the median (*Me*), that's the second quartile
    - The 75th percentile is called the third quartile (q_{75})
    - The difference between the third and first quartiles is the interquartile range (IQR)
    - Other percentiles usually used: q_{01}, q_{05}, q_{10}, q_{90}, q_{95}, q_{99} which are the 1th, 5th, 10th, 90th, 95th, 99th percentiles

How to find the first, second (Me) and third quantiles?



Calculating the percentile

Origin data									
$\mathbf{x_i}$	x_1	x_2	…					x_{n-1}	x_n

Sorted data									
$\boldsymbol{x_{(i)}}$	$x_{(1)}$	$x_{(2)}$	…					$x_{(n-1)}$	$x_{(n)}$
$\boldsymbol{i}$	1	2	…					n-1	n
$\boldsymbol{p_i}$	p_1	p_2	…					p_{n-1}	p_n

$$p_i = 100\frac{i - 0.5}{n}$$

For a given p_i: $\quad i = \frac{np_i}{100} + 0.5$

If i is an integer, $x_{(i)}$ is simply the p^{th} percentile.

If i is not an integer, we can interpolate as follows:

- let k = the integer part of i , (i.e., if i = **10.375**, then k = 10)
- let f = the fractional part of i, (i.e., if i = **10.375**, then f = 0.375)
- let $x_{(int)}$ = the value we want to interpolate between $x_{(k)}$ and $x_{(k+1)}$:

$$x_{(int)} = (1-f)\, x_{(k)} + f x_{(k+1)}$$ ***is the p^{th} percentile***

Some important statistical characteristics

- Statistical characteristics:
  - Correlation Coefficient between two data series:

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

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Origin data							
$\mathbf{x_i}$	x_1	x_2	…		x_{n-1}	x_n	$\bar{x}$
$\mathbf{y_i}$	y_1	y_2	…		y_{n-1}	y_n	$\bar{y}$
$\mathbf{x'_i = x_i - \bar{x}}$	x'_1	x'_2	…		x'_{n-1}	x'_n	$\sum_{i=1}^{n}(x_i - \bar{x})^2$
$\mathbf{y'_i = y_i - \bar{y}}$	y'_1	y'_2	…		y'_{n-1}	y'_n	$\sum_{i=1}^{n}(y_i - \bar{y})^2$
$\mathbf{x'_i y'_i}$	$x'_1y'_1$	$x'_2y'_2$	…		$x'_{n-1}y'_{n-1}$	$x'_ny'_n$	$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Some important statistical characteristics

- Statistical characteristics:
  - Correlation Matrix
    - Let $X_1, X_2, \ldots, X_m$ be m variables
    - Let r_{jk}, $j,k=1,2,\ldots, m,$ be correlation coefficients between X_j, X_k
    - The correlation matrix is denoted by

$$(R) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{pmatrix}$$

    - $r_{11} = r_{22} = \ldots = r_{mm}$; $r_{jk} = r_{kj}$

Important atmospheric variables

- Temperature: T(xlon,ylat,plev,t), in C or K
- Precipitation: P(xlon,ylat,t) or R(xlon,ylat,t), in mm/day or mm/month
- Mean Sea Level Pressure: PMSL(xlon,ylat,t), in mb, hPa
- Geopotential high: HGT(xlon,ylat,plev,t), in m, km
- Surface Wind: U10m(xlon,ylat,t), V10m(xlon,ylat,t), in m/s
- Wind: U(xlon,ylat,plev,t), V(xlon,ylat,plev,t), W(xlon,ylat,plev,t), in m/s
- Humidity(xlon,ylat,plev,t):
  - Specific humidity (q), Mixing Ratio: g/kg
  - Relative Humidity: % or -

Analysis of horizontal means

- Let *F(xlon,ylat,plev,t)* be a certain variable
- Given a certain *plev*, say 1000mb, or 850mb,…
- Given *xlon,ylat* for a specific area or whole globe
- Calculating mean values of $F(xlon_i,ylat_j,plev,t)$ at each point (station or grid box) $(xlon_i,ylat_j)$ over entire time series $t_1, t_2, \ldots, t_n$:

$$F(xlon_i, ylat_j, plev) = \frac{1}{n}\sum_{k=1}^{n} F(xlon_i, ylat_j, plev, t_k)$$

where $i=1,N$; $j=1,M$ are number of stations or of grid box in the interested area; t_k can be chosen depending on objectives of analysis

  - Annual mean, seasonal mean, monthly mean,…
  - Long term climatology
  - …

Analysis of horizontal means

- Annual mean fields
- Monthly mean fields
- Zonal Means
- Meridional Means
- Time-Zonal cross section
- Time-Meridional cross section
- Correlation Maps

Analysis of vertical means

- Mean Profiles
- Zonal-Vertical cross section
- Meridional-Vertical cross section

Time series Analysis

- Changes vs Variabilites
- Anomalous time series
- Trend Analysis
- Periodic Analysis

Examples and Homeworks

- Reanalysis data: Era Interim
- Station data: Rainfall and Temperature
- Data structures: Text files, NetCDF
- Calculations: Fortran Programs
- Display, plot, graphics: Grads, EXCEL
- Login to the Cluster Linux system at HUS: