

< INTRODUCTION >

- Deterministic vs. Stochastic

- A variable X is deterministic if the output of X can be predicted with certainty
- A variable X is stochastic (= random) if the output of X cannot be predicted with certainty. X is governed by laws of probability.

- Time Series (or Process)

: sequential observations (= realizations) of a (deterministic or stochastic) variable X ,
i.e. X_1, X_2, \dots , where the subscript represents intervals of time.

- Hydrologic Time Series

: discrete time series of continuous hydrologic variables such as precipitation and runoff

- Types of Hydrologic Time Series

(1) single vs. multiple (= univariate vs multivariate)

(2) correlated vs. uncorrelated (cf dependent vs independent)

- autocorrelated = serially correlated = temporally correlated = correlated in time
- crosscorrelated = spatially correlated = correlated in space

(3) stationary vs. nonstationary

: a hydrologic time series is stationary if its statistical parameters such as the mean and variance remain constant through time.

- first order stationary
- second order stationary

(4) others: intermittent time series, counting time series, etc.

- Components of Time Series (see handouts)

$$X_t = T + I + P + S$$

where T = trend (gradual change)

I = intervention (sudden change)

P = periodicity (astronomic cycles)

S = stochasticity (ARMA model applied here)

Classification of Hydrologic Series.

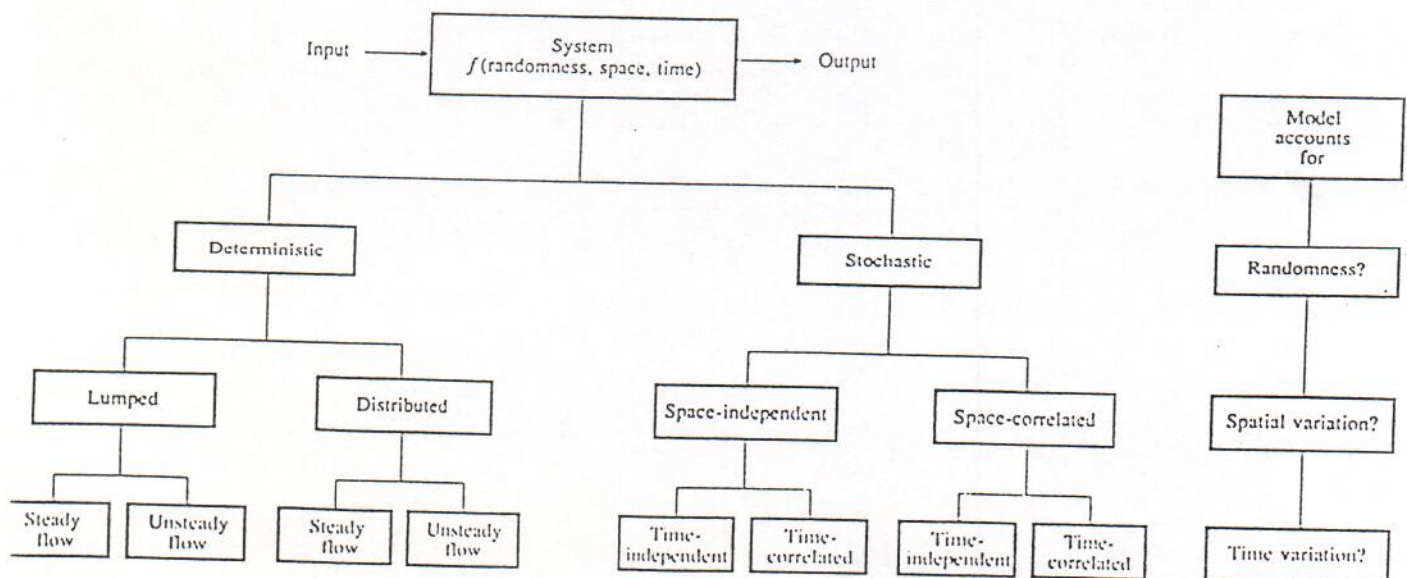
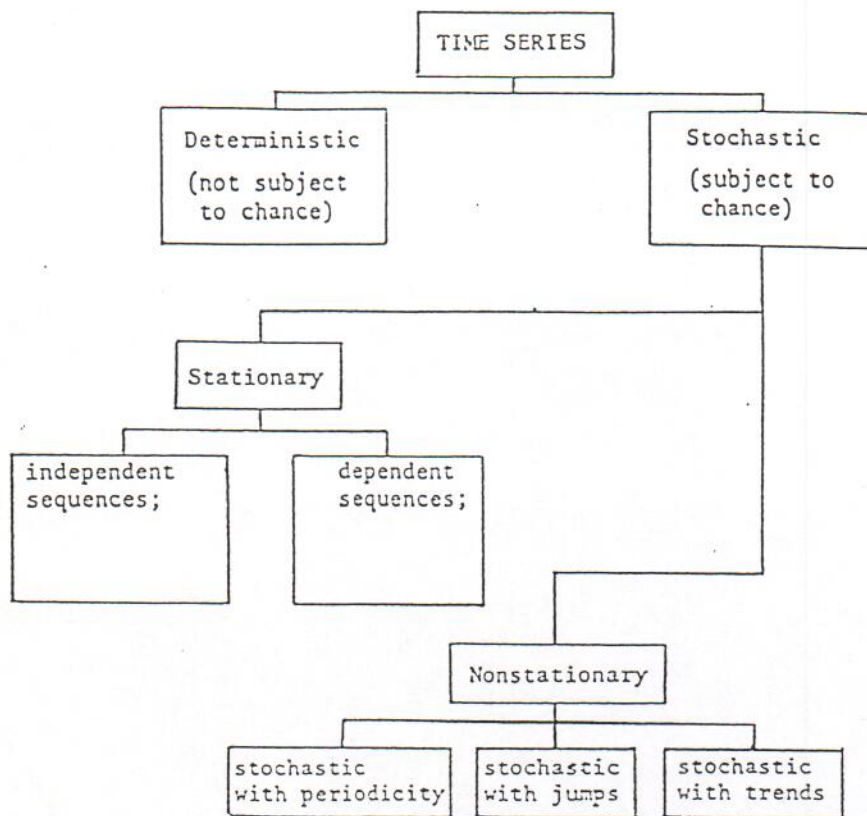
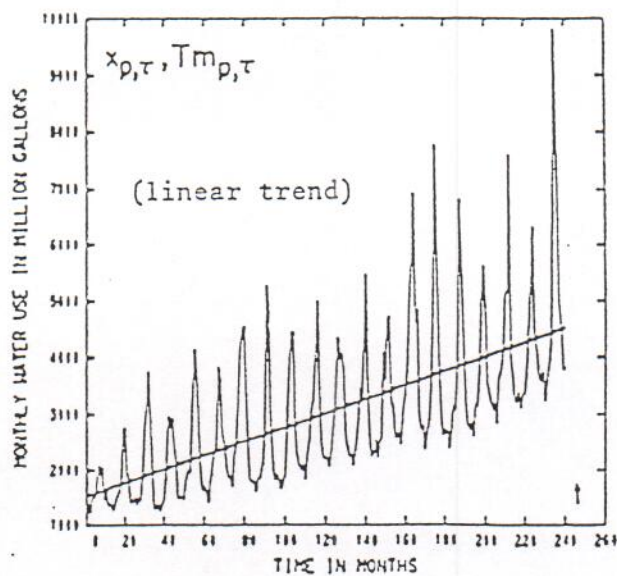


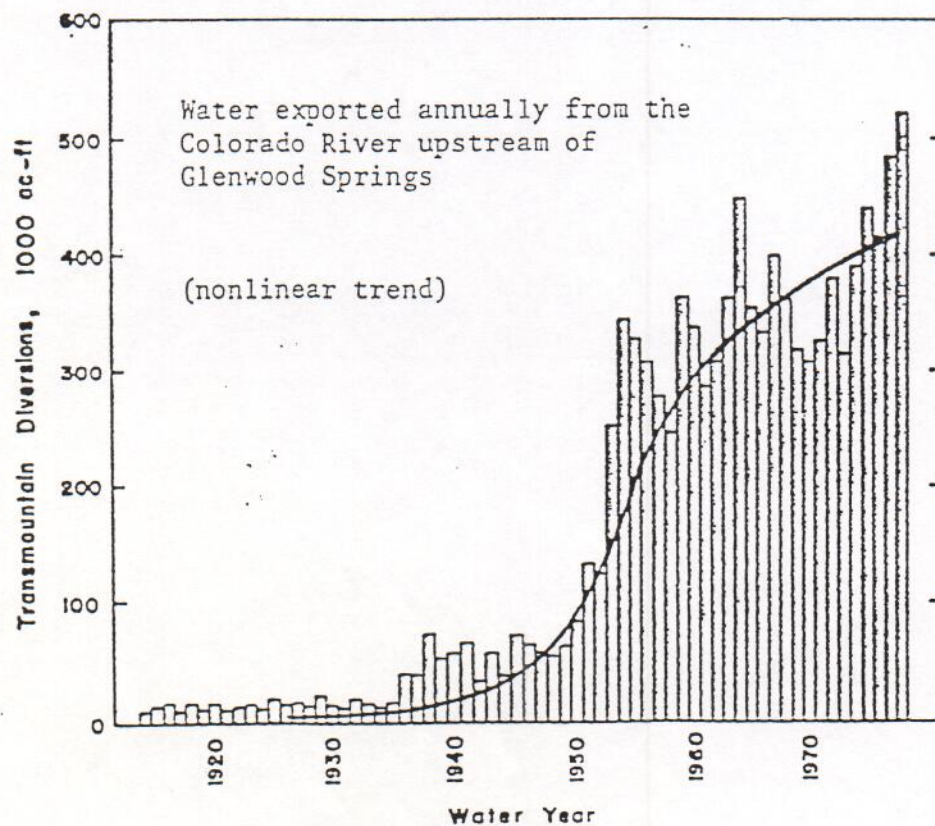
FIGURE 1.4.1
Classification of hydrologic models according to the way they treat the randomness and space and time variability of hydrologic phenomena.

Trend

Nonstationary Time Series with Trends

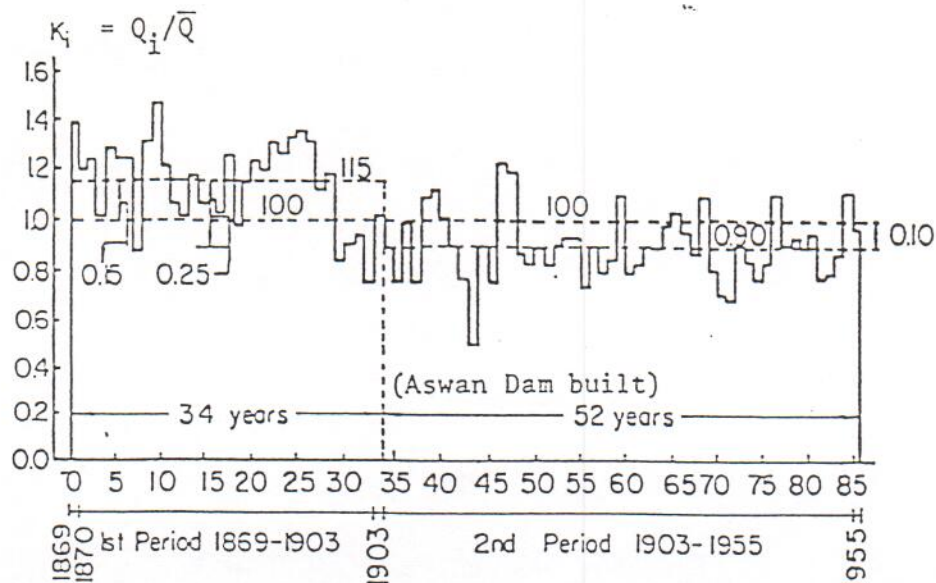


Observed monthly water use series $x_{p,\tau}$ and the linear trend in the mean $Tm_{p,\tau}$ for Dallas, Texas for 1950-1969.

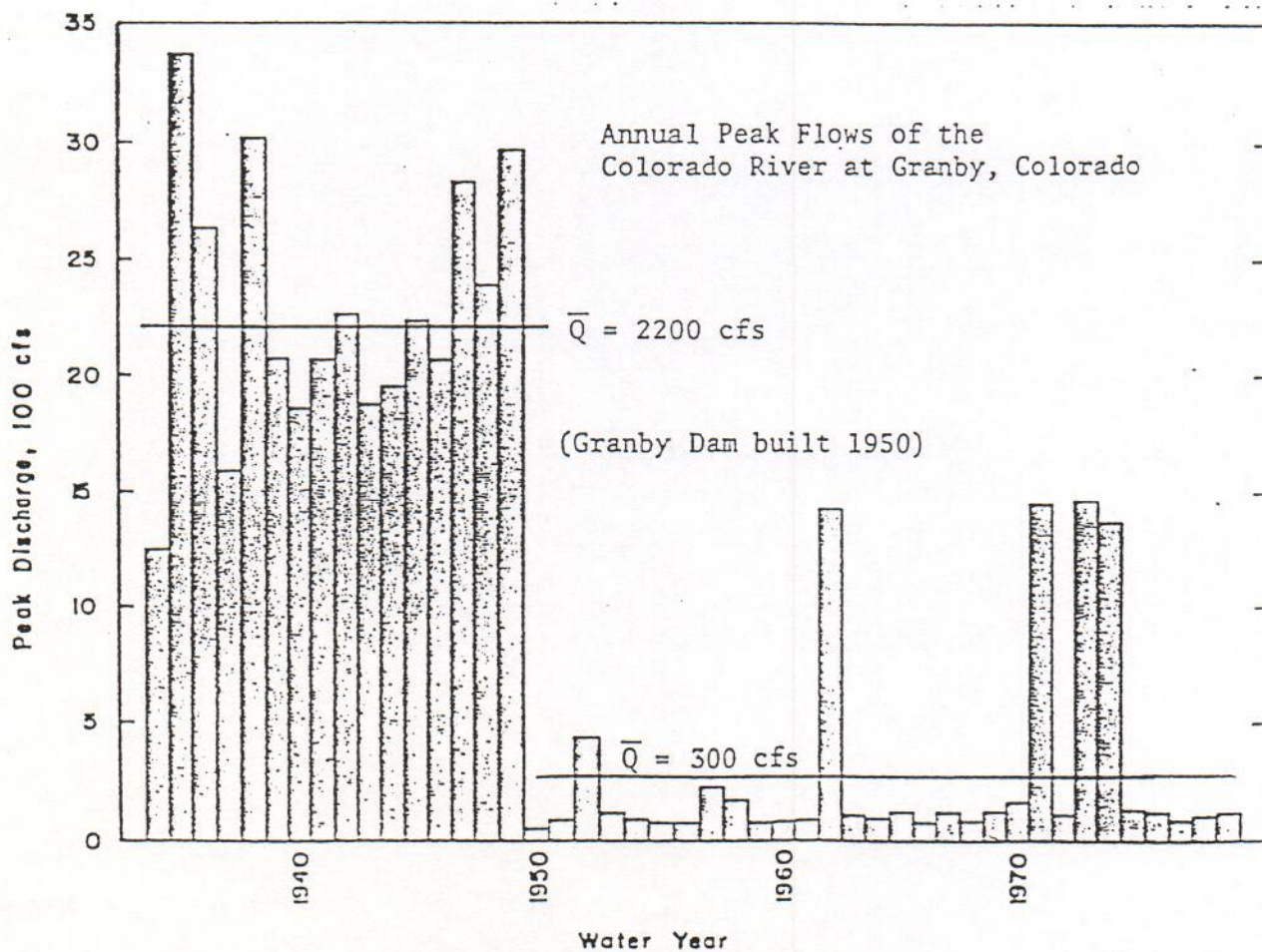


Intervention

Nonstationary Time Series with Jump



Fluctuations of annual flow of the River Nile as an example of inconsistency in data of the time series



Periodic Motion of the Earth.

4.2 The Basic Motions of the Earth

(1) Short Period Variations

Once each 24 hours the earth rotates on its axis, spinning from west to east on its path around the sun causing diurnal variations in many geophysical and hydrologic phenomena and producing the daily cycle. Once each month the moon makes a complete revolution around the earth causing almost-periodic oscillations in tides and related hydrologic phenomena and producing the lunar cycle. Once each year the earth moves in an elliptic orbit around the sun causing the annual periodic variation of incoming solar radiation received over the whole planet and producing the annual cycle.

(2) Long Period Variations

As determined from celestial mechanics the long-term variations in the elements of the earth's orbit are due to the interaction between the earth and eight other principal planets, namely, Mercury, Venus, Mars, Jupiter, Saturn, Uranus, Neptune and Pluto. Changes in the gravitational field accompanying motions of the planets result in perturbation of the earth's orbit and its axis of rotation. Three parameters are used to describe changes in the earth's motion as a function of time. These parameters are: eccentricity of the orbit, tilt or obliquity of the ecliptic, and precession. Vernekar (1972) has recomputed Milankovich's results following essentially the mathematical and theoretical treatment developed by Milankovich but with corrections, modifications and more accurate actions derived by subsequent researchers including Brouwer and van Woerkom, Sharaf and Budnikova, and Vernekar. Figure 12 shows diagrammatically the various motions of the earth. Variations of the three parameters, eccentricity, tilt, and precession, are shown as a function of time in Fig. 12.

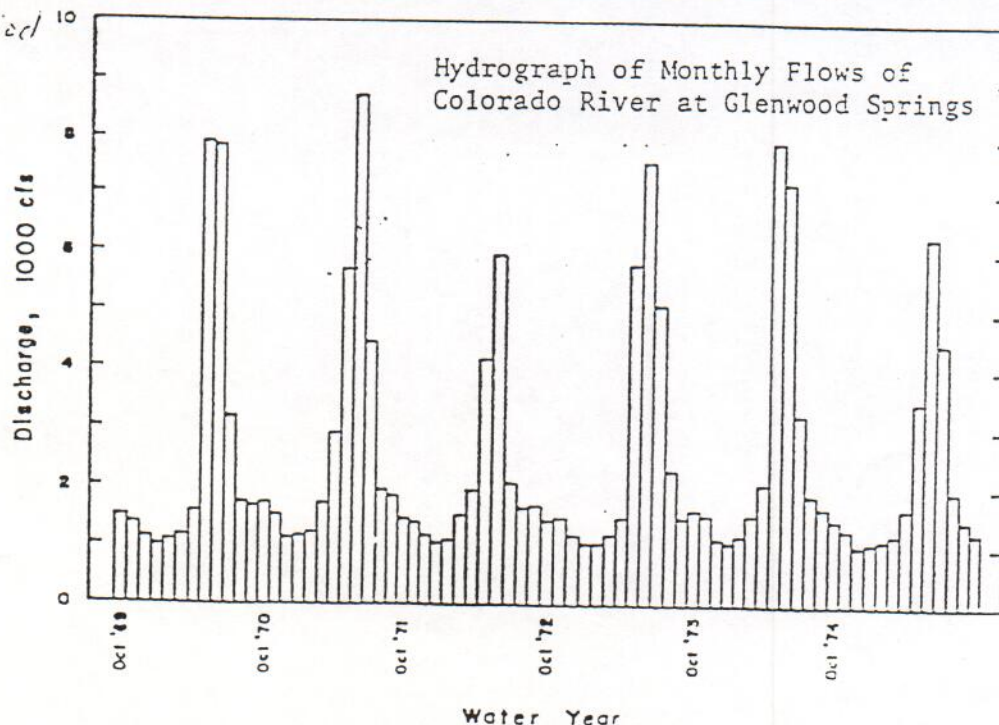
Eccentricity. The earth and moon trace an elliptical path in their annual orbit around the sun with the sun at one focus. The sun's gravity acts on the earth and moon as if they were coupled into a giant dumbbell. It is the center of the dumbbell mass, rather than the center of the earth which moves in a smooth elliptical orbit (Fig. 12, dashed line). The center of the earth follows a serpentine path (Fig. 12, dotted line). The earth-moon system does not quite retrace its orbit in successive years. With each revolution, it begins a path in a position slightly counterclockwise from the previous. Over the past 4 million years, Vernekar has calculated the eccentricity of the orbit to vary from a minimum value close to zero which was reached 534,000 year B.P.* to a maximum value of 0.061 which was obtained 2,775,000 years B.P. It can be seen on Fig. 13, upper graph, that variations in eccentricity exhibit an almost-periodic movement with maxima and minima changing in amplitude and spacing with time. The average time for one complete oscillation has been 93,000 years over the past 4 million years.

Tilt or Obliquity of the Ecliptic. The interactions between the earth and other bodies in the planetary system leads to an oscillatory change in the tilt of the earth's axis with respect to the orbital plane. The tilt has varied from 24.51 degrees to 22.10 degrees in the past 4 million years with an almost-periodic movement averaging 41,000 years per one complete oscillation. The present value of the tilt or obliquity is 23.44579 degrees.

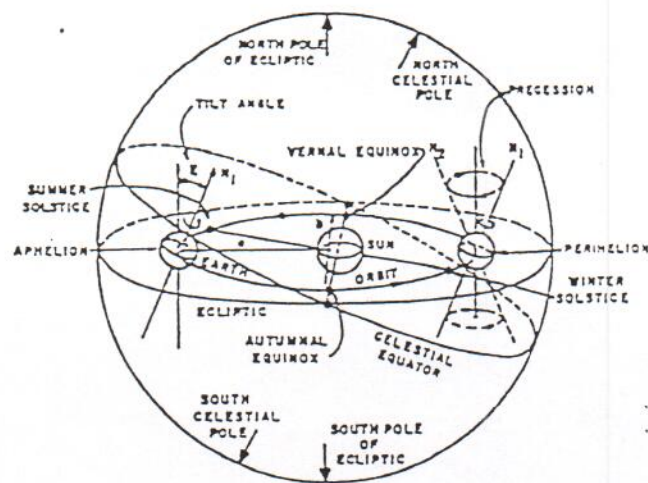
Precession of the Equinoxes. Because the earth has an oblate shape which is flattened at the poles and bulging at the equator, the sun's gravitational attraction of the equatorial belt tends to pull the earth's axis of rotation towards the vertical in relation to the plane of the orbit. The external force applied to a rotating body causes the tilted axis of the earth to swing ponderously around in a tight circle to trace a double conical figure much like the wobbling of a spinning top. The movement is called precession and has an almost-periodic movement averaging 21,000 years per one complete oscillation. This means that the line joining the equinoxes of the earth's orbit will make one complete revolution of the orbit in this time, on the average.

*B.P. means before present.

Short Period
Variation



Long Period Variation



CELESTIAL SPHERE



PRECESSION

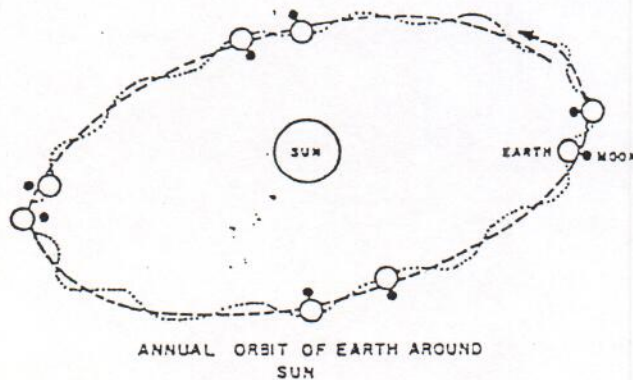


MAXIMUM TILT

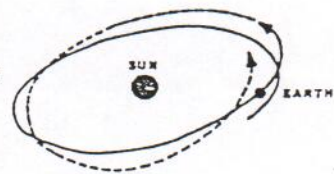


MINIMUM TILT

TILT OR OBLIQUITY



ANNUAL ORBIT OF EARTH AROUND SUN



PATH OF ECCENTRICITY

ECCENTRICITY

Fig. 12. The motions of the earth (from Imbrie and Kipp, 1971, after Vernekar, 1963).

< Statistical Properties of Time Series >

- Overall Sample Statistics

- sample autocorrelation
 - (i) correlogram
 - (ii) sample covariance
 - (iii) unbiased estimator
- observed annual statistics
 - (i) coefficient of variation
 - (ii) skewness coefficient
 - (iii) lag-1 autocorrelation

- Seasonal Sample Statistics

- seasonal autocorrelation
- observed seasonal statistics
 - (i) coefficient of variation
 - (ii) others

- Drought-related Statistics

- resiliency
- vulnerability
- reliability

- Storage-related Statistics

- rescaled range
- Hurst coefficient (1951)

- Hurst phenomenon

: the discrepancy between theoretical result ($h = 1/2$) and Hurst's empirical finding ($h = 0.73$)

- possible reasons of Hurst phenomenon

- (i) sampling variability
- (ii) nonstationarity
- (iii) long memory

The mean monthly flow for the Mita Mita is $101.1 \times 10^6 \text{ m}^3$; therefore, a 75% draft is $79.6 \times 10^6 \text{ m}^3/\text{month}$.

The first few months of the calculation are shown in Table 4-7. The draft has been subtracted from each monthly flow and the residuals accumulated. The first summed residual value is itself a peak (designated H_1), because the succeeding values decrease. The next peak is shown as H_2 . The storage required to cover this flow sequence is $H_1 - T_1$ where T_1 is the lowest cumulative residual value between the peaks, i.e., $0 - (-209.0) = 209 \times 10^6 \text{ m}^3$.

After traversing through the historical data twice, the successive H-T values are found to be:

209, 957, 960, 242, 1032, 227, 287, 104, 204, 220, 194, 197, 157, 16, 414, 323, 283, 682, 1107, 957, 960, 242, 1032, 227, 287, 104, 204, 220, 194, 157, 16, 414, 323, 283, 682 ($\times 10^6 \text{ m}^3$).

The required storage capacity is the largest value, say, $1110 \times 10^6 \text{ m}^3$.

This example shows the importance of running the historical flows through twice. If only one sequence had been used, the maximum storage capacity would not have been found.

4-21 HURST'S PROCEDURE

Hurst was concerned with storages on the River Nile. It was an unusual problem in that he was dealing with large equatorial lakes where a big increase in volume could be obtained by small structures at lake outlets. Consequently, storage sizes capable of providing very high regulation (approaching 100%) were being examined. Hurst's work (1951, 1956, 1965) is concerned more with mathematical experiment rather than theory, but from it, he was able to formulate a general solution to the reservoir capacity-yield problem.

From an analysis of some 700 natural time-series including streamflows, rainfalls, temperatures, atmospheric pressures, tree rings and lake varves, Hurst found that the range could be related to the length of record as follows:

$$R/s = (N/2)^K \quad (4-42)$$

where R = range defined as the sum of the cumulative departures from the mean,

s = standard deviation of the time-series data, N = length of the time-series, and

McMahon & Meade (1986) River and Reservoir Yield

K = exponent, and was found by Hurst to have a mean value of 0.72 and a standard deviation of 0.09.

As an aside it should be noted that for a purely random time-series process K tends towards 0.5 (Feller, 1951). The fact that natural time-series yield on the average values of K greater than 0.5 is known as the Hurst phenomenon; this aspect is considered again in Sec. 6-17.

In developing his general solution to the storage-yield problem, Hurst (1951) computed K and the storage size to "guarantee" a uniform draft less than mean \bar{x} using a mass curve analysis for records of natural phenomena. The generalized storage relations (for all observations) are:

$$(i) \text{ for draft} = \text{mean } (\bar{x}) \quad C = R \quad (4-43)$$

$$(ii) \text{ for draft} < \text{mean} \quad \log_{10}(C/R) = -0.08 - 1.05 (\bar{x} - B)/s \quad (4-44)$$

$$\text{or} \quad C/R = 0.94 - 0.96 \sqrt{[(\bar{x} - B)/s]} \quad (4-45)$$

where C = required reservoir capacity,

R = range defined by Fig. 4-5,

B = draft parameter which is defined as

$$\text{draft} = (\bar{x} - B)/s, \quad (4-46)$$

s = standard deviation of annual flows, and \bar{x} = mean annual flow.

Examining these relationships using only Hurst's river flow data, Joy (1970) found that the generalized curves were not lines of best fit and that the results exhibited large scatter as shown in Fig. 4-28. Hurst admitted the generalized nature of his results in replying to discussions by Chow (Hurst, 1951, p. 800). A further limitation of the procedure is that only the rank 1 values of storage are considered and no other estimates of risk of storage failure can be estimated. Nevertheless, Hurst's work is of monumental significance because of the importance of the value of K in stochastic data generation models (Sec. 6-17).

Fathy and Shukry

Fathy and Shukry (1956) agreed with Hurst that for 100% regulation the equivalent mass curve reservoir capacity was given by Hurst's range R . But for lower drafts they disagreed, and developed their own method. However, as shown by Joy (1970) their technique is no more than a mathematical representation of the minimum flow approach of Wallis (1945) given in Sec. 4-8.

LONG MEMORY MODELS

6-16 BACKGROUND

Stochastic or synthetic streamflows have become widely used in the design and operation of water resources systems. Models for stochastically generating streamflows can be broadly classified into short- and long-memory types. Short memory (or high frequency) models are those of the Markovian type in which the Hurst exponent (h) in Eq. 6-77 approaches $1/2$ as n (the number of items in the time series) becomes large.

$$\frac{R}{s} = (0.5 n)^h \quad (6-77)$$

where R = range of cumulative departures from the mean,
 s = standard deviation of flows, and
 n = number of items of data.

Short memory models have been criticized for their inability to simulate the relatively high values of the Hurst coefficient (say $h = 0.7$) observed in long time sequences of streamflow.

As the magnitude of h is related to the size of storage required for regulation of streamflow, the use of short memory models may lead to underestimation of storage capacities when used in reservoir design. In comparing storage estimates based on Markovian models and fast fractional Gaussian noise models that preserve h , Wallis and Matalas (1972) observed that differences occurred only for drafts greater than 80%. More recent studies by Klemes et al. (1981) of the reliability of reservoir systems indicate that, except for very large storages and high drafts, the effect of the Hurst coefficient is not of practical importance particularly when considered in terms of other potential errors in the hydrology (see Fig. 6-4). We have noted in Sec. 2-3 that errors in storage estimates would be rarely less than 25 percent. Hence, it is reasonable that the conclusion reached by Klemes et al. (1981) is realistic, namely that "... the use of long-memory models will, in principle, remain equivalent to the use of a small safety factor in the intrinsically inaccurate estimate of reservoir reliability performance".

6-17 PRESERVING h IN TIME SERIES MODELS

For Markov type models, as n in Eq. 6-77 becomes large, $h \rightarrow 1/2$. Thus more complex time series models have been developed to preserve h at values greater than $1/2$ along with the other historical parameters. These include the autoregressive integrated moving average (ARIMA) model

(O'Connell, 1971), fractional Gaussian noise process (ffGn), particularly fast fractional Gaussian noise (ffGn) (Mandelbrot, 1972) and the Broken Line (BL) model (Mejia et al., 1972). For practical use, we would recommend the latter model.

An initial problem with respect to preserving h in Eq. 6-73 relates to estimating its value from historical data. Four methods have been proposed:

- (i) Hurst's estimate K (Hurst, 1951) (Eq. 2-25);
 - (ii) Mandelbrot and Wallis (1969) H . (Two versions F Hurst and G Hurst are given in Wallis and Matalas, 1970). The usual method is G and is denoted by H_1);
 - (iii) Gomide's estimate YH (Gomide, 1975);
 - (iv) Maximum likelihood estimate \hat{H} (McLeod and Hipel, 1978).
- Srikanthan (1979) computed the above four estimates for 16 Australian rivers and observed wide variations. Results for four rivers are given in Table 6-6. From the analysis he concluded that YH and H were unsatisfactory procedures and adopted K because it is less variable than H .

Furthermore Srikanthan (1979) carried out a very extensive study of the three models - ARIMA, ffGn and BL - and concluded that the BL model produced the most satisfactory synthetic streamflows. The ARIMA model was the least satisfactory of the three. Some details of this work are published in Srikanthan and McMahon (1978a, b, c). Sen (1977) developed a theoretical procedure for evaluating h in small samples and concluded that both H and K are biased estimators of h . He also found that H has the smaller bias of the two estimators.

Table 6-6 Estimates of Hurst coefficient

River (Australian national stream gauge number)	C_v	r_1	Hurst coefficient			
			K	H	YH	\hat{H}
South Johnston (112101)	0.38	-0.10	0.49	0.37	0.34	0.26
South Esk (318001)	0.47	0.01	0.60	0.52	0.43	0.45
Peel (419004)	0.82	0.24	0.71	0.78	0.53	0.65
Wide Bay Creek (138002)	1.26	0.28	0.77	0.81	0.58	0.66

(Extracted from Srikanthan, 1979)

< Testing & Removing Nonstationalities >

- Testing Trends

- Parametric Test: Polynomial Fitting Technique

(i) testing a linear trend if $i = 1$

(ii) hypothesis testing for linear trends

- Nonparametric Test: Mann-Kendall Test

- Testing Shifts

- Parametric Test

(i) split the original sample X

(ii) hypothesis testing

- Nonparametric Test: Mann-Whitney Test

- Hypothesis Test for Differences in Means When the Variances are Not Different

- assumptions

- (i) $X_1, X_2 \sim \text{normal}$

- (ii) $\sigma_1 = \sigma_2$

- (iii) σ_1 and σ_2 are unknown

- hypothesis

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

- test statistic

- Hypothesis Test for Differences in Means When the Variances are Not Same

- assumptions

- (i) $X_1, X_2 \sim \text{normal}$

- (ii) $\sigma_1 \neq \sigma_2$

- (iii) σ_1 and σ_2 are unknown

- hypothesis

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

- test statistic

- Hypothesis Test for Differences in Variances

- assumptions: $X_1 \sim N_1(\mu_1, \sigma_1)$ and $X_2 \sim N_2(\mu_2, \sigma_2)$

- hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_a: \sigma_1^2 \neq \sigma_2^2$$

- test statistic

- Hypothesis Test for Goodness of Fit

- Chi-Square Test

- the Kolmogorov-Smirnov Test

- the Filliben test

< Normality Assumption>

- Testing Normality

- Box & Cox Transformation

- 3 LN Transformation

- Removing Trends & Shifts

- Removing Seasonality

- Seasonal Standardization

- Fourier Series

- (i) mean square deviation

- (ii) cumulative periodogram (P_i vs i)

The plot of P_i vs. i is called the cumulative periodogram. A graphical criteria using the cumulative periodogram for obtaining the significant harmonics is given below.

The criteria is based on the concept that the variation of P_i versus i is composed of the two distinct parts: (1) a periodic part of a fast increase of P_i with i and (2) a sampling part of a slow increase of P_i with i . Two approaches are feasible for determining those two parts. (First) the two parts are approximated by smooth curves that intersect at a point, which corresponds to the critical harmonic h^* that gives the number of significant harmonics. The (second) approach is to assume approximate mathematical models of these two parts, to estimate their parameters and to find the intersection of two equations. The ordered harmonic nearest to the intersection point is then the critical harmonic. In the second approach, when $z_{v,t}$ of Eq. (3.25) is an independent series, the sampling part of the cumulative periodogram, as referred above, is a straight line, whereas when $z_{v,t}$ is a linearly dependent series, the sampling part is a curve.

Figures 3.1 and 3.2 show the intersection point A for a periodic series with either an independent or a dependent stochastic component, respectively. The value of P_i at point A is determined by the sample size, while the value of i is much less affected by the sample size and sampling variation. Difficulties arise when the point A for a dependent stochastic component is in such a position that both curves (3) and (4) of Fig. 3.2 come out to be nearly one continuous curve, implying that the separation of the two parts of the cumulative periodogram becomes uncertain. Examples show that this case is less common in practice.

Figures 3.3 through 3.7 show the cumulative periodograms for five statistical characteristics: mean y_t , standard deviation s_t , and the first, second and third serial correlation coefficients, $r_{1,t}$, $r_{2,t}$, and $r_{3,t}$, for five discrete series: (1) 69 years of daily precipitation at Fort Collins, Colorado, from 1898 to 1966, Fig. 3.3; (2) 70 years of 3-day precipitation at Austin, Texas, from 1898 to 1967, Fig. 3.4; (3) 18 years of 7-day precipitation at Ames, Iowa, from 1949 to 1966, Fig. 3.5; (4) 40 years of daily discharge of the Tioga River near Erwins, New York, from 1921 to 1960, Fig. 3.6; and (5) 37 years of 3-day discharge of the McKenzie River at McKenzie Bridge, Oregon, from 1924 to 1960, Fig. 3.7. The harmonic i ranges from 1 to 182 for daily series, from 1 to 60 for three-day series, and from 1 to 26 for 7-day series. Because other precipitation and river gaging stations for

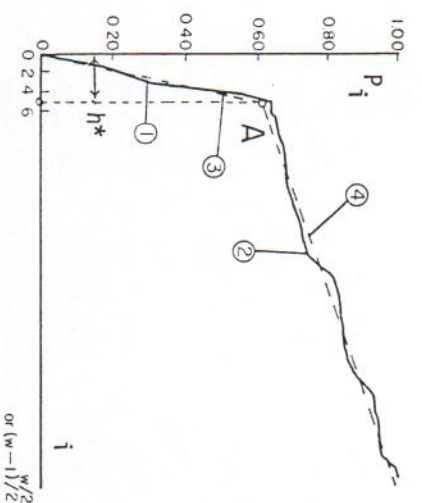


Figure 3.1.

Separation of the cumulative periodogram into the periodic part, for both the observed (1) and the fitted (3), and the sampling variation part, also for both the observed (2) and the fitted (4), in case of a periodic series with an independent stochastic component.

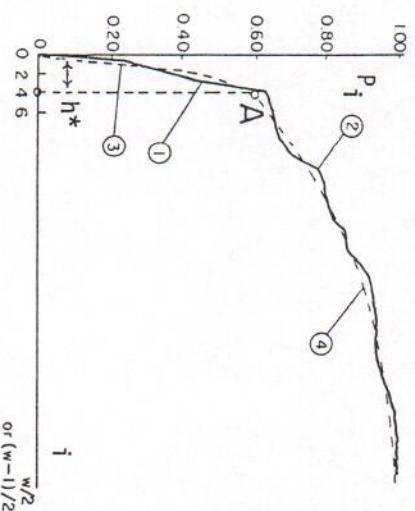


Figure 3.2.

Separation of the cumulative periodogram into the periodic part, observed (1) and fitted (3), and the sampling variation part, observed (2) and fitted (4), in case of a periodic series with an autoregressive stochastic component.

< Forecasting and Generation >

- Applicability of Time Series Models in Hydrology
 - Forecasting
 - Generation
- Forecasting vs. Prediction
 - forecasting: the estimate of conditions at a specific future time or during a specific time interval
 - prediction: the estimate of future conditions
- Measure of Forecast Error
 - (1) bias
 - (2) variability
 - (3) mean square error
 - (4) root mean square error
 - (5) mean absolute error
 - (6) relative bias
 - (7) relative mean absolute error
 - (8) forecast efficiency
 - (9) R squared
- Precision vs. Accuracy
 - precision: the ability of an forecaster to provide repeated estimates that are close together.
 - ((note)) 1. due to random error
 - 2. measured with the variability
 - accuracy: precision+unbiasedness
 - ((note)) measured with the $MSE = \text{variability} + \text{bias}^2$
- Forecast Worth
 - no forecast
 - naive forecast
 - imperfect forecast
 - perfect forecast

- Deterministic vs. Probabilistic Forecasts

- Short-term vs. Long-term Forecasting

- Monte Carlo Simulation

- generation of uniform r.v.

- generation of normal r.v.

- generation of correlated inputs

- length of generated samples

- numbers of samples to generate

PROGRAM RNG

```

C
C .....
C
C Simple FORTRAN random number generator adapted from:
C   Bratley et al., (1983) A Guide to Simulation
C   Springer-Verlag, New York, 383 pp.
C .....
C
C   dimension U(5000)
C   open(Unit=6,File='C:output.rng')
C .....
C   Interactive prompt for generator seed.
C .....
C
C 600 write(0,700)
C 700 format(//,10x,'Enter seed (integer between 0 and 2**31-1)',/)
C   read(0,800) iseed
C 800 format(I12)
C   if(iseed.le.0 .or. iseed.gt.2147483647) go to 705
C   go to 708
C 705 write(0,707)
C 707 format(//,10x,'Seed is out of acceptable range; try again',/)
C   go to 600
C 708 continue
C .....
C   Initialize and then call the random number generator unif(*).
C .....
C
C   dummy = unif(iseed)
C   do 100 i=1,2000
C     u(i) = unif(0)
C     write(6,900) u(i)
C 100 continue
C 900 format(5x,f12.6)
C   stop
C   end
C   FUNCTION UNIF(kx)
C .....
C
C   This function generates a random number distributed uniformly
C   over the interval [0,1] using the recursion
C
C     u(i+1) = 16807 * u(i) mod[(2**31)-1].
C
C   Input:  kx is a random integer, 0 < kx < (2**31)-1, the first
C           time this function is called. Thereafter kx=0.
C
C   Output: ix is a new psuedo-random integer.
C           unif is a random fraction, 0 < unif < 1.
C
C   Notes:  (2**31)-1 is 2147483647.
C           This function is valid only with a 32-bit word.
C           IBM uses 8 bits/byte, 2 bytes/word, or 16-bit words.
C           Need double precision to achieve 32-bit word on PC.
C .....
C
C   if(kx.gt.0) ix=kx
C   ix = dmod(16807d0*ix, 2147483647d0)
C   unif = ix/2147483647d0
C   return
C   end

```

Probabilistic Verification Measures

1. Scalar Measures

Brier Score (BS)

- probabilistic forecasts of dichotomous events
- the mean square error of the probabilistic forecasts
- $0 \leq BS \leq 1$
- perfect forecasts: $BS = 0$

$$BS = \frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2$$

where i : a numbering of the n forecast/event pairs

p_i : the forecast probability

o_i : the verifying observation (1 if the event occurs, 0 if it does not)

Half-Brier Score

- an extension of the Brier Score to the multi-category probabilistic forecasts
- the mean square error of the probabilistic forecast
- a perfect forecast: $HBS = 0$
- a worst forecast: $HBS = 2$

$$HBS = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^n (p_{ij} - o_{ij})^2$$

where J : the number of the forecast categories

p_{ij} : the forecast probability that the event will occur in category J

o_{ij} : the verifying observation

Ranked Probability Score

- an extension of the Brier Score to the multi-category probabilistic forecasts
- the squared error with respect to the cumulative probabilities in the forecast and observation vectors
- a perfect forecast: $HBS = 0$
- a worst forecast: $HBS = J - 1$

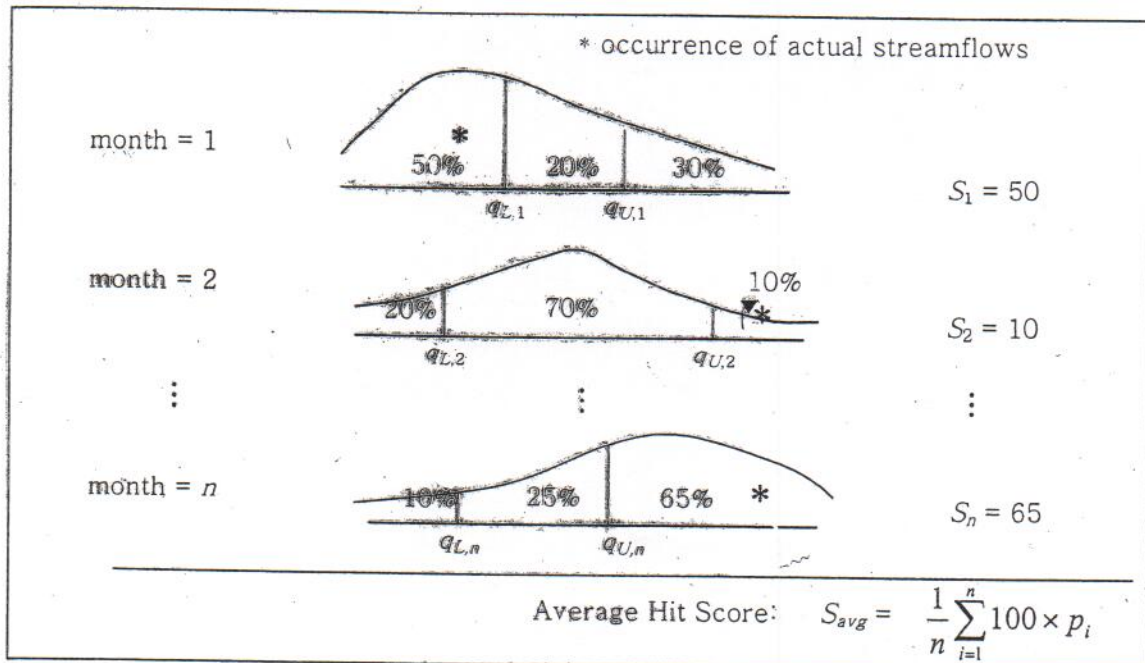
$$RPS = \sum_{m=1}^J (P_m - O_m)^2$$

where

$$P_m = \sum_{j=1}^m p_j, \quad m = 1, \dots, J,$$

$$O_m = \sum_{j=1}^m o_j, \quad m = 1, \dots, J.$$

Average Hit Score



Skill Score

- a percentage improvement over the reference forecasts

- $A = A_{perf}$: $SS_{ref} = 100\%$

- $A = A_{ref}$: $SS_{ref} = 0\%$

$$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\%$$

where A_{ref} : the value of the accuracy measure of reference forecast

A_{perf} : the value of the accuracy measure of perfect forecast

2. Contingency Table

- categorical forecasts of most-probable event

	O_L	O_M	O_H
f_L	a	b	c
f_M	d	e	f
f_H	g	h	i

hit rate

- a perfect forecast: $H = 1$

$$H = \frac{a+e+i}{n}$$

where $n = a+b+c+d+e+f+g+h+i$

Bias Ratio

- a unbiased forecast: $B_L, B_M, B_H = 1$

$$B_L = \frac{a+b+c}{a+d+g}$$

$$B_M = \frac{d+e+f}{b+e+h}$$

$$B_H = \frac{g+h+i}{c+f+i}$$

3. Reliability Diagram

- the categories of a forecast variable: low (0~30%), middle (30~70%), high (70%~100%)
- the forecasts are considered perfectly reliable if the relative frequency of the observation equals the forecast probability:

$$p(o_i = 1 | p_i) = p_i$$

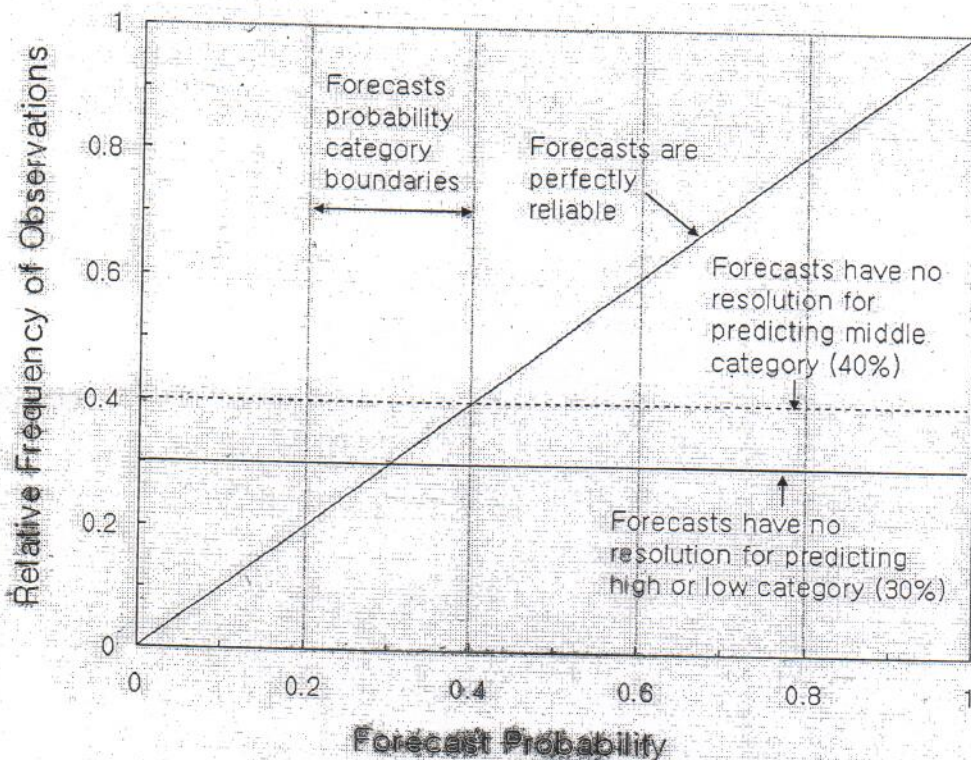
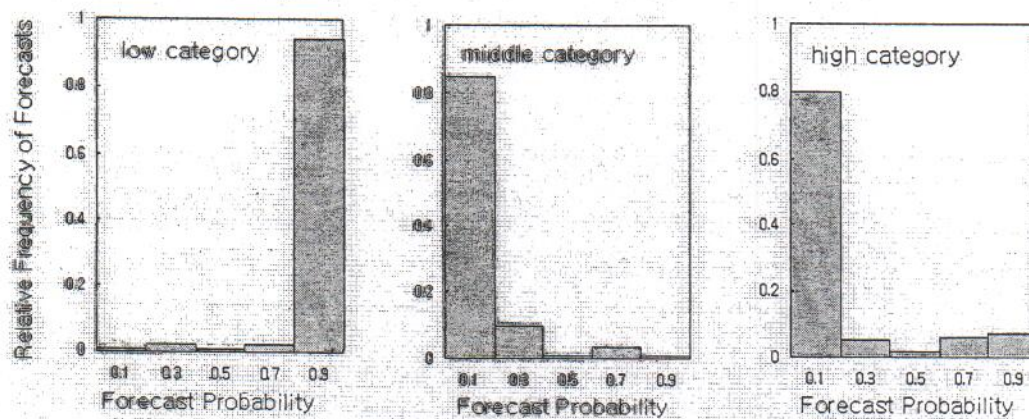


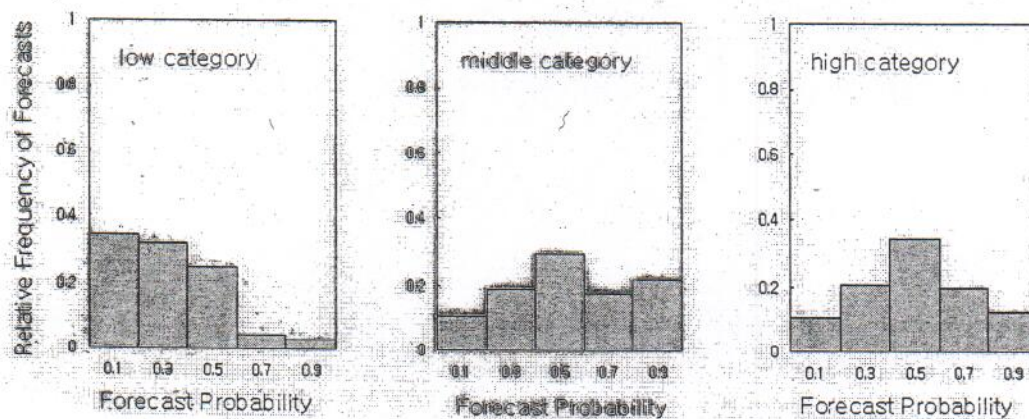
Figure 1 Example reliability diagram describing the behavior of forecasts that fall in particular regions of the diagram. (Franz et al. 2003).

4. Discrimination Diagram

- the conditional distribution of the forecasts given the observed category, $p(p_i|o_i)$
- if the $p(p_i|o_i)$ equals zero for all possible observations except one, the forecast procedure is perfectly discriminatory for forecasts of that observation



(a) Good discrimination



(b) No discrimination

Figure 2 Example of discrimination diagrams

< Introduction to Time Series Modeling >

- Assumptions of ARMA(AutoRegressive Moving Average) Models
 - normality
 - the second order stationarity
- General Procedure of ARMA Modeling
 - (1) preliminary data analysis
 - graphical display of data
 - summary statistics
 - removal of nonstationarities
 - transformation to normal
 - (2) identification for determining p and q of ARMA(p,q)
 - ACF
 - PACF
 - (3) parameter estimation
 - unconditional/ conditional least square method
 - method of moment
 - maximum likelihood method
 - (4) diagnostic check
 - parsimony of parameters: AIC
(model with minimum number of parameters which represents or preserves certain statistics)
 - independence of residuals: ACF of residuals, Portemanteau Test
 - normality of residuals
 - (5) application
 - forecasting
 - generation
- Physical Basis of ARMA(1,1) Models
 - Notations
 - x_t = the precipitation in the year t
 - S_t = the groundwater storage at the end of year t
 - z_t = the streamflow in the year t
 - S_{t-1} = the groundwater storage at the beginning of year t

- Assumptions

- (i) the amount ax_t infiltrates, percolates, and reaches the groundwater storage
- (ii) the amount bx_t evaporates from the soil, plants, and surface storage
- (iii) the amount $(1-a-b)x_t = dx_t$ represents the surface runoff reaching the stream

- the streamflow is made up of

- the mass balance equation for the groundwater storage

< AutoRegressive Modeling: Identification >

- Formulation

- Notations

y_t = autocorrelated stationary time series $\sim N(\mu, \sigma^2)$

ε_t = residual series (white noise) $\sim N(0, \sigma_\varepsilon^2)$

NOT autocorrelated

NOT crosscorrelated with y_{t-1}, \dots, y_{t-p}

ϕ_1, \dots, ϕ_p = autoregressive coefficients

- AR(p) process

((note)) Fiering & Jackson model

- Parameters

{

- relationship σ & σ_ε

- ACF for $AP(p)$
- Yule-Waker Equation

- Partial ACF for AP(p)

- Notations

$\phi_j(k)$ = jth AR coefficient of AR(k) model, $j = 1, 2, \dots, k$

$\phi_k(k)$ = the last term in the AR(k)

- Stationary Condition

: the stationary condition of the AR(p) model is satisfied if the roots of the following characteristic equation lie inside the unit circle, i.e. $|u_i| < 1$

- AR(1)

- AR(2)

< AutoRegressive Modeling: Parameter Estimation >

Say where N observations x_1, x_2, \dots, x_N .

$$x_t = f(x_{t-1}, x_{t-2}, \dots, a_1, \dots, a_1) + \varepsilon_t$$

- Moment Estimator: Use the Yule-Walker equation
- Least Squares Estimator

- Maximum Likelihood Estimator

< AutoRegressive Modeling: Diagnostic Checks >

- Test for Normality of Residuals

 - Independent Test for Residuals
 - residual correlogram

 - cumulative periodogram test

 - Porte Manteau lack of fit test
- For ARMA(p,q) models,
H0: ε_t is a white noise series, i.e. an independent process
Ha: ε_t is not a white noise series.

- Parameter Parsimony

- Akaike Information Criteria (AIC)

- : Among competing ARMA models, select the one which minimize AIC,

- $$AIC(p,q) = N \ln(\sigma_\epsilon^2) + 2(p+q)$$

- modified AIC

- $$AICC(p,q) = N \ln(\sigma_\epsilon^2) + 2(p+q+1)N/(N-p-q-2)$$

((note)) Need $N/k > 15 \sim 20$ where k is th number of parameters

< AutoRegressive Modeling: Forecasting >

< Moving Average Modeling >

- Formulation of a MA(q) model

- Parameters of MA(q)

- ACF

- the autocovariance

- the variance

- the autocorrelation coefficient

((note)) The ACF is truncated or cut off at lag q with a "memory" limited to q lags.

- Partial ACF

The PACF are of infinite extent and tail off

- Invertibility Condition

The roots of the following characteristic equation must lie inside the unit circle:

3.1.3 Stationarity and invertibility conditions for a linear process

Stationarity. The convergence of the series (3.1.9) ensures that the process has a finite variance. Also, we have seen in Section 2.1.3, that the autocovariances and autocorrelations must satisfy a set of conditions to ensure stationarity. For a linear process these conditions can be embodied in the single condition that the series $\psi(B)$, which is the generating function of

the ψ weights, must converge for $|B| \leq 1$. That is, on or within the unit circle. This result is discussed in Appendix A3.1.

Spectrum of a linear stationary process. It is shown in Appendix A3.1 that if we substitute $B = e^{-i2\pi f}$, where $i = \sqrt{-1}$, in the autocovariance generating function (3.1.11), we obtain one half of the power spectrum. Thus the spectrum of a linear process is

$$\begin{aligned} p(f) &= 2\sigma_a^2 \psi(e^{-i2\pi f}) \psi(e^{i2\pi f}) \\ &= 2\sigma_a^2 |\psi(e^{-i2\pi f})|^2 \quad 0 \leq f \leq \frac{1}{2} \end{aligned} \quad (3.1.12)$$

In fact, (3.1.12) is the well known expression [27], which relates the spectrum $p(f)$ of the output from a linear system, to the uniform spectrum $2\sigma_a^2$ of a white noise input by multiplying by the squared gain $G^2(f) = |\psi(e^{-i2\pi f})|^2$ of the system.

← **Invertibility.** We have seen above, that the ψ weights of a linear process must satisfy the condition that $\psi(B)$ converges on or within the unit circle, if the process is to be stationary. We now consider a restriction applied to the π weights to ensure what is called "invertibility." The invertibility condition is independent of the stationarity condition and is applicable also to the non-stationary linear models, which we introduce in Chapter 4.

To illustrate the basic idea of invertibility, consider again the model

$$\tilde{z}_t = (1 - \theta B)a_t \quad (3.1.13)$$

Expressing the a 's in terms of the \tilde{z} 's, (3.1.13) becomes

$$a_t = (1 - \theta B)^{-1} \tilde{z}_t = (1 + \theta B + \theta^2 B^2 + \dots + \theta^k B^k)(1 - \theta^{k+1} B^{k+1})^{-1} \tilde{z}_t$$

that is

$$\tilde{z}_t = -\theta \tilde{z}_{t-1} - \theta^2 \tilde{z}_{t-2} - \dots - \theta^k \tilde{z}_{t-k} + a_t - \theta^{k+1} a_{t-k-1} \quad (3.1.14)$$

and, if $|\theta| < 1$, on letting k tend to infinity, we obtain the infinite series

$$\tilde{z}_t = -\theta \tilde{z}_{t-1} - \theta^2 \tilde{z}_{t-2} - \dots + a_t \quad (3.1.15)$$

and the π weights of the model in the form of (3.1.4), are $\pi_j = -\theta^j$. Whatever the value of θ , (3.1.13) defines a perfectly proper stationary process. However, if $|\theta| \geq 1$, the current deviation \tilde{z}_t in (3.1.14) depends on $\tilde{z}_{t-1}, \tilde{z}_{t-2}, \dots, \tilde{z}_{t-k}$, with weights which increase as k increases. We avoid this situation by requiring that $|\theta| < 1$. We shall then say that the series is *invertible*. We see that this condition is satisfied if the series

$$\pi(B) = (1 - \theta B)^{-1} = \sum_{j=0}^{\infty} \theta^j B^j$$

converges for all $|B| \leq 1$, that is, on or within the unit circle.

In Chapter 6, where we consider questions of uniqueness of these models, we shall see that a convergent expansion for a_t is possible when $|\theta| \geq 1$, but only in terms of $z_t, z_{t+1}, z_{t+2}, \dots$ (that is in terms of present and *future* values of the process). The requirement of invertibility is needed if we are interested in associating present events with *past* happenings in a sensible manner.

In general, the linear process

$$\pi(B)\tilde{z}_t = a_t$$

is invertible if the weights π_j are such that the series $\pi(B)$ converges on, or within the unit circle.

To sum up, a linear process is *stationary* if $\psi(B)$ converges on, or within the unit circle and is *invertible* if $\pi(B)$ converges on, or within the unit circle.

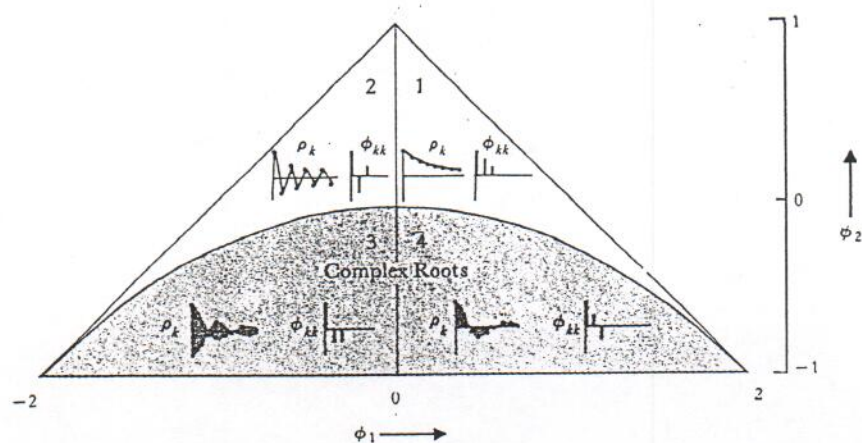


FIG. 3.2 Typical autocorrelation and partial autocorrelation functions ρ_k and ϕ_{kk} for various stationary AR(2) models

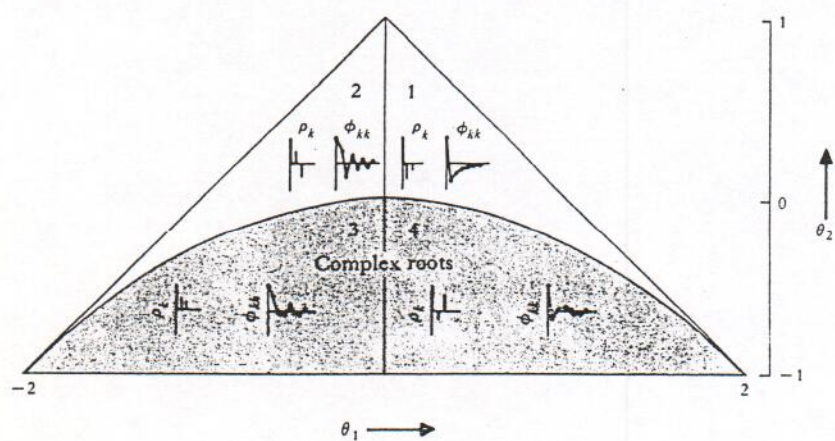


FIG. 3.9 Autocorrelation and partial autocorrelation functions ρ_k and ϕ_{kk} for various MA (2) models

cf) ACF

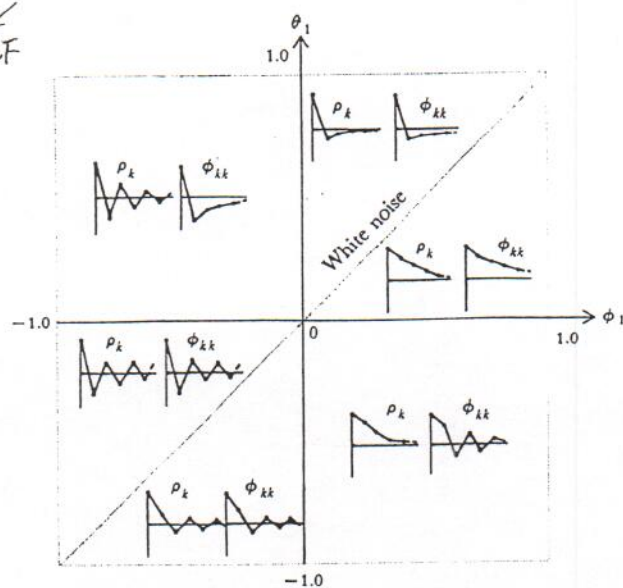


FIG. 3.11 Autocorrelation and partial autocorrelation functions ρ_k and ϕ_{kk} for various ARMA (1, 1) models

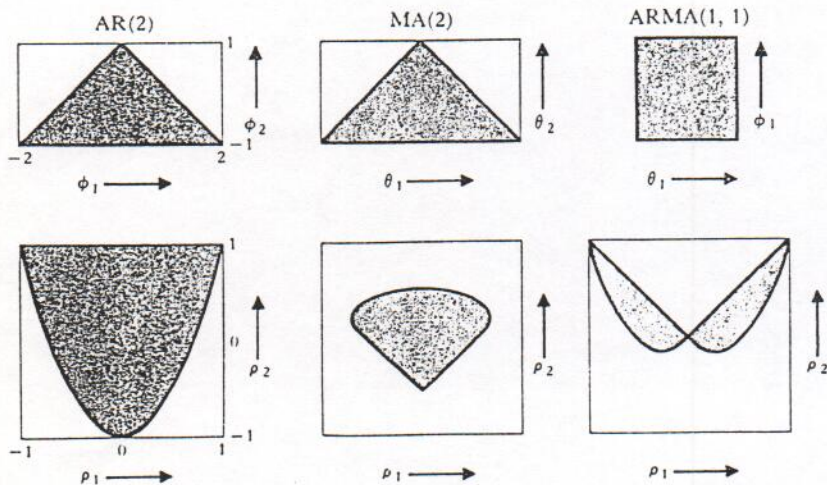


FIG. 3.12 Admissible regions for the parameters ρ_1, ρ_2 for AR (2), MA (2), and ARMA (1, 1) processes which are restricted to being both stationary and invertible

< AutoRegressive Moving Average Modeling >

- Formulation of a ARMA(p,q) model

- Parameters of ARMA(p,q)

((note1)) ARMA(1,1) vs. AR(1)

((note2)) ARMA models may come out with a smaller number of parameters to estimate than the AR models of higher order

- ACF

- Partial ACF

The PACF attenuates as damped waves or exponential decay

- Stationarity & Invertibility Conditions

- Properties of ARMA(1,1) models

- Forecasting
 - L-step ahead forecasting function for ARMA(p,q)

((example)) ARMA(1,1)

< Modeling of Single Periodic Series >

- Handling Periodic Series

- (1) standardizing --> modeling with ARMA
- (2) seasonal differencing --> modeling with ARMA (SARIMA)
- (3) modeling with periodic ARMA

- Periodic ARMA

- PAR(p)

((example)) PAR(1)

- PARMA(p,q)

((example)) PARMA(1,1)

- (Seasonal) AutoRegressive Integrated Moving Average
 - SARIMA(p,d,q)(P,D,Q)_s

((example)) ARIMA(2,00)(0,1,1)₁₂

< Transfer Function Noise (TFN) Models >

- Transfer Function (TF)

- Consider AR(p)

- deterministic Box-Jenkins TF model

$$y_t = v_0 x_t + v_1 x_{t-1} + v_2 x_{t-2} + \dots$$

$$y_t = v(B)x_t$$

where

$v(B)$: transfer function

$v_0, v_1, v_2 \dots$: impulse response function

x_t : forcing function (exogeneous variable)

- Theory of TFN Models

$$y_t = v(B)x_t + n_t$$

- assumption

- (i) x_t & y_t are stationary

- (ii) x_t & n_t are independent, so

- (iii) x_t, y_t, n_t are normal

- estimation of $v(B)$

To estimate $v_0, v_1, v_2 \dots$

we will first "prewhiten" the x_t series

This will make parameter estimation easier.

Multiply (1) by a_{t-k} and take expectation

• Identification

(1) prwhiten the input series by a univariate ARIMA model

(2) calculate the sample cross-correlation function (CCF) and the impulse response weight v_k

(3) identify the orders (r,s,b) of the parsimonious equation,

on equating coefficient of B, we find

$$\begin{array}{ll}
 v_j = 0 & \text{if } j < b \\
 v_j = \delta_1 v_{j-1} + \delta_2 v_{j-2} + \dots + \delta_r v_{j-r} + \omega_0 & \text{if } j = b \\
 v_j = \delta_1 v_{j-1} + \delta_2 v_{j-2} + \dots + \delta_r v_{j-r} - \omega_{j-b} & \text{if } j = b+1, b+2, \dots, b+s \\
 v_j = \delta_1 v_{j-1} + \delta_2 v_{j-2} + \dots + \delta_r v_{j-r} & \text{if } j > b+s
 \end{array}$$

Thus, in general, v_j consists of

- (i) b zero values v_0, v_1, \dots, v_{b-1}
- (ii) a further s-r+1 values $v_b, v_{b+1}, \dots, v_{b+s-r}$ following no fixed pattern (but no such values occur if $s < r$), and
- (iii) for $j \geq b+s-r+1$, values v_k following the pattern dictated by an rth order difference equation with r starting values $v_{b+s}, \dots, v_{b+s-r+1}$

(4) generate the noise series

(5) identify an univariate ARIMA model for n_t by using the general identification procedure for ARIMA

- Parameter Estimation

- Diagnostic Checks

- Is a_t not serially correlated?

: check if $r_a(k) = 0$ by using the univariate PM test.

If not, respecify the noise model.

- Are x_t & a_t not cross-correlated?

: check if $r_{xa}(k) = 0$ by using the bivariate PM test.

If not, respecify the TF model.

- Forecasting

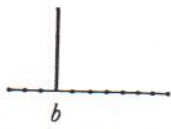
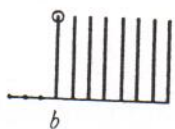
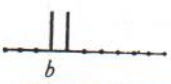
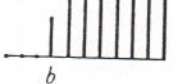
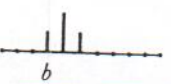
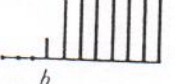
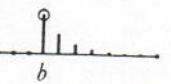
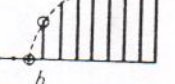
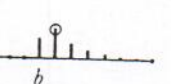

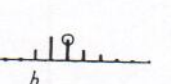


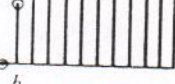
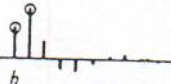
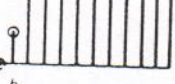
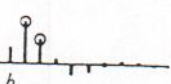
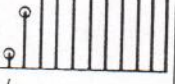
r, s, b	∇ Form	B Form	Impulse Response v_j	Step Response $V_j = \sum_{i=0}^j v_i$
003	$Y_t = X_{t-3}$	$Y_t = B^3 X_t$		
013	$Y_t = (1 - .5\nabla) X_{t-3}$	$Y_t = (.5 + .5B) B^3 X_t$		
023	$Y_t = (1 - \nabla + .25\nabla^2) X_{t-3}$	$Y_t = (.25 + .50B + .25B^2) B^3 X_t$		
103	$(1 + \nabla) Y_t = X_{t-3}$	$(1 - .5B) Y_t = .5B^3 X_t$		
113	$(1 + \nabla) Y_t = (1 - .5\nabla) X_{t-3}$	$(1 - .5B) Y_t = (.25 + .25B) B^3 X_t$		
123	$(1 + \nabla) Y_t = (1 - \nabla + .25\nabla^2) X_{t-3}$	$(1 - .5B) Y_t = (.125 + .25B + .125B^2) B^3 X_t$		
203	$(1 - .25\nabla + .5\nabla^2) Y_t = X_{t-3}$	$(1 - .6B + .4B^2) Y_t = .8B^3 X_t$		
213	$(1 - .25\nabla + .5\nabla^2) Y_t = (1 - .5\nabla) X_{t-3}$	$(1 - .6B + .4B^2) Y_t = (.4 + .4B) B^3 X_t$		
223	$(1 - .25\nabla + .5\nabla^2) Y_t = (1 - \nabla + .25\nabla^2) X_{t-3}$	$(1 - .6B + .4B^2) Y_t = (.2 + .4B + .2B^2) B^3 X_t$		

FIG. 10.6 Examples of impulse and step response functions with gain $g = 1$

< Modeling of Multivariate ARMA Models >

- Formulation
 - MAR(1)

$$Z_t = A_1 Z_{t-1} + B \varepsilon_t$$

where

$$Z_t = Y_t - \mu$$

$$E[\varepsilon_t \varepsilon_t^T] = I$$

$$E[\varepsilon_t \varepsilon_{t-k}^T] = 0 \text{ for all } k \neq 0$$

$$\varepsilon_t \sim \text{Normal}$$

$$A_1 = [n \times n]$$

$$B = [n \times n]$$

- MARMA(1,1)

$$Z_t = A_1 Z_{t-1} + B \varepsilon_t - C_1 \varepsilon_{t-1}$$

where

$$C_1 = [n \times n]$$

- Parameter Estimation

$$M_t = A_1 M_{t-1} \quad \text{or} \quad M_t = A_1 M_{t-1}$$

where

$$M_t = E[Z_t Z_{t-k}^T] =$$

$$A_1 = M_1 M_0^{-1}$$
$$B B^T = M_0 - A_1 M_1^T$$

((note)) A and B can be solved if M & M satisfy the following condition

(i) M must be positive definite (generally satisfied when the sample sizes for all sites are the same)

(ii) $B B^T$ must be positive definite

(See p.19.30 ((T1)) for the details)

- Diagnostic Checks

- normality: $\varepsilon_t \sim \text{Normal?}$

- independence in space: $E[\varepsilon_t \varepsilon_t^T] = I$

- independence in time: $E[\varepsilon_t \varepsilon_{t-k}^T] = 0$ for all $k \neq 0$

((note)) The elements of $M_0(\varepsilon)$ & $M_1(\varepsilon)$ except the diagonal of $M_0(\varepsilon)$ must be within the limits of

< Contemporaneous ARMA Models >

• Formulation

$$Z_t = \sum A_j Z_{t-j} + \varepsilon_t - \sum C_j \varepsilon_{t-j}$$

where A_j & C_j are diagonals

$$Z_t^{(i)} = \sum a_j^{(i)} Z_{t-j}^{(i)} + \varepsilon_t^{(i)} - \sum c_j^{(i)} \varepsilon_{t-j}^{(i)}$$

((note))

(i) p and q do not have to be the same for each site.

(ii) The model components at each site are simply univariate ARMA(p, q) models

where each $\varepsilon_t^{(i)}$ is uncorrelated in time but correlated in space. Thus

$$\varepsilon_t = B \xi_t$$

where ξ_t is normal such that

$$E[\xi_t \xi_t^T] = I$$

$$E[\xi_t \xi_{t-k}^T] = 0 \text{ for all } k \neq 0$$

• Parameter Estimation

- estimate the elements of A & C through the univariate parameter estimation
- estimate B by solving $BB^T = G$ where the elements of G can be estimated as

$$\text{eq(19.3.50) ((T1))}$$

• Diagnostic Checks

- normality: $\xi_t \sim \text{Normal?}$
- independence in space: $E[\xi_t \xi_t^T] = I$
- independence in time: $E[\xi_t \xi_{t-k}^T] = 0 \text{ for all } k \neq 0$

< Disaggregation Models (DM) >

- Disaggregation Modeling: A process by which time series are generated dependent on a time series already available
 - temporal disaggregation
- ((note)) Approach for generation of seasonal & annual flows
 - (i) generation of seasonal flows with a periodic model
 - > aggregation of the seasonal flows to annual flows
 - (ii) generation of annual flows with any an annual model
 - > disaggregation of the annual flows into seasonal flows
- spatial disaggregation

- Purpose: to reproduce statistics at more than one level of aggregation

- Advantage of the Disaggregation Approach

- (i) DM allows for a reduction in the number of parameters with little or no corresponding loss of desirable properties in generated data
- (ii) DM allows for increased flexibility in the methods used for generation

- Basic DM (Valencia & Schaake, 1973)

- formulation

$$Y = AX + B\varepsilon$$

where Y : the current observation of the series being generated (subseries) $\sim N(0, \sigma_y^2)$

X : key series $\sim N(0, \sigma_x^2)$

ε : stochastic term $\sim N(0, 1)$

- parameters

$$\hat{A} = S_{XY} S_{XX}^{-1}$$

$$\hat{B} \hat{B}^T = S_{YY} - A S_{XY}$$

- moments: Eq. (A8.1) \sim (A8.12) ((T2))

- features

- (i) preserves covariance between annual values and its seasonal values
- (ii) preserves variance and covariances among the seasonal values
- (iii) preserves cross-covariance between the values at the various sites in case of a multi-site case

- advantage : basic and clean

- disadvantages

- (i) the number of parameters is large

(ii) the moments being preserved are not consistent

(inconsistency in calculation of seasonal covariance - See ((T2))Eq. A.8.2)

- Extended DM (Mejia & Rousselle, 1976)

- formulation

$$Y = AX + B\epsilon + CZ$$

where Z: a column matrix containing as many seasonal values from the previous year as are desired (ex: A8.13 ((T2)))

- parameters: Eq. (8.16) ~ (8.18) ((T2))

- moments: Eq. (A8.13) ~ (A8.19) ((T2))

- additional advantage: preserves the covariances of the first season of a year and preceding season

- disadvantages

- (i) the number of parameters becomes larger

- (ii) inconsistency in moment calculation still exists

- Condensed DM (Lane, 1979)

- formulation

$$Y_{\tau} = A_{\tau}X + B_{\tau}\epsilon + C_{\tau}Y_{\tau-1}$$

where τ : the current season being generated

- parameters: Eq. (8.19) ~ (8.21) ((T2))

- moments: Eq. (A8.20) ~ (A8.24) ((T2))

- features

- (i) some parameters of the extended DM are set to zero

- (ii) one-season-at-a-time

- (iii) ignores all but the lag-0 and lag-1 correlations among the monthly flows and therefore requires fewer parameters

- advantage: the number of parameters is reduced considerably

- disadvantages

- (i) not clean and straightforward

- (ii) since all seasons are not generated jointly, the seasonal data will not add exactly to give the annual time series (nonpreservation of additivity)

- ((note)) the second disadvantage is outweighed by its advantages because the problem are common for all DM models if the data have undergone any transformations

- software: LAST (Lane & Frevert, 1990)