HYDROLOGY PROJECT

Technical Assistance

World Bank & Government of The Netherlands funded

Training module # SWDP -17

How to carry out secondary validation of climatic data

New Delhi, November 1999

CSMRS Building, 4th Floor, Olof Palme Marg, Hauz Khas,
New Delhi – 11 00 16 India
Tel: 68 61 681 / 84 Fax: (+ 91 11) 68 61 685
E-Mail: dhvdelft@del2.vsnl.net.in

DHV Consultants BV & DELFT HYDRAULICS

with
HALCROW, TAHAL, CES, ORG & JPS

Table of contents

1. Module context

While designing a training course, the relationship between this module and the others, would be maintained by keeping them close together in the syllabus and place them in a logical sequence. The actual selection of the topics and the depth of training would, of course, depend on the training needs of the participants, i.e. their knowledge level and skills performance upon the start of the course.

2. Module profile

Title	:	How to carry out secondary validation of climatic data
Target group	:	Assistant Hydrologist, Hydrologists, Data Processing Centre Managers
Duration	:	One session of 120 min
Objectives	:	After the training the participants will be able to*:* • Perform secondary validation of rainfall data
Key concepts	:	• spatial correlation structure of rainfall for various durations • spatial homogeneity • entries at wrong days • accumulated rainfall • transposed entries • number of rainy days • double mass analysis • auto correlation and spectral density functions
Training methods	:	Lecture, Software
Training tools required	:	Board, OHS, computers
Handouts	:	As provided in this module
Further reading and references	:	

3. Session plan

No	Activities	Time	Tools
1	***General*** • Overhead Highlighted text	5 min	
2	***Methods of Secondary validation*** • Overhead Header and bullet points	5 min	
3	***Multiple station validation*** ***3.1 Comparison plots*** • Overhead Highlighted text • Overhead Example figure using HYMOS *	5 min	
	3.2 Balance series • Overhead Example figure using HYMOS *	5 min	
	3.3 Regression analysis • Overhead Highlighted text • Overhead Example figure from HYMOS *	5 min	
	3.4 Double Mass curves • Overhead Highlighted text • Overhead Example figure using HYMOS *	5 min	
	3.5 Nearest neighbour analysis • Overhead Highlighted text • Overhead Example output from HYMOS (use example in Manual)	5 min	
4	***Single series test of homogeneity*** ***4.1 Trend analysis*** • Overhead Highlighted text and bullets	15 min	
	4.2 Mass curves		
	4.3 Residual mass curves		
	4.4 A note on hypothesis testing		
	4.5 Student's- t test for the stability of the mean		
	4.6 Wilcoxon-W test on the difference in the means		
	4.7 Wilcoxon-Mann-Whitney U-test		

4. Overhead/flipchart master

5. Handout

Add copy of Main text in chapter 8, for all participants.

6. Additional handout

These handouts are distributed during delivery and contain test questions, answers to questions, special worksheets, optional information, and other matters you would not like to be seen in the regular handouts.

It is a good practice to pre-punch these additional handouts, so the participants can easily insert them in the main handout folder.

7. Main text

Contents

How to carry out secondary validation of climatic data

1. General

- **Secondary validation will be carried out at Divisional offices.** Data which has been primary validated will be forwarded from the Sub-divisional office to the Divisional office in files generated by the primary module, and identifying and annotating all flagged values. Any hard copy field data relating to suspect data will also be forwarded.

- **Secondary validation is mainly concerned with spatial comparisons between neighbouring stations** to identify anomalies in recording at the station. Some secondary validation more appropriately carried out at Divisional level is concerned with investigation of a single series but with long data series rather than simply with current data. Such testing may not be possible until a significant amount of historical data has been added to the database.

- **Spatial validation of climatic data is not so much concerned with individual values as with the generality of values received from a station.** This is often best illustrated by the use of aggregated data

- **Procedures used in secondary validation apply over a range of variables.** However their applicability may vary with the spatial variability of the variable. Some of the methods have already been described for rainfall in Module 9. They will be mentioned but not described in full again.

- **In interpreting the results of secondary validation, it is essential to be aware of the physical properties, limits and controls of a variable and the method of measurement.** This has been outlined in more detail in Module 16 on primary validation.

- **Secondary validation is designed to detect anomalies in time series such as trends or changes in spatial relationships. The user should be warned that these may result from real environmental changes as well as data error.** Data processors should be careful not to adjust data unless they are confident that the data or a portion of the data are incorrect rather than due to a changed microclimate. The existence of trend should be noted in the station record and supplied to data users with the data. For some analytical purposes data users may wish to adjust for the trend, in others to retain it.

2. Methods of Secondary validation

- Multiple station validation

  - ❖ Comparison plots of stations
  - ❖ balance series
  - ❖ regression analysis
  - ❖ double mass curve
  - ❖ test of spatial homogeneity (nearest neighbour analysis)

- Single station validation tests for homogeneity

    - ❖ mass curves
    - ❖ residual mass curves
    - ❖ a note on hypothesis testing
    - ❖ Student's 't' test of difference of means
    - ❖ Wilcoxon W-test on the difference of means
    - ❖ Wilcoxon-Mann-Whitney U-test to determine whether series are from the same population

The section on multiple station validation is placed first in the text as it is generally chronologically first to be carried out at the Divisional office.

3. Screening of data series

After the data from various Sub-Divisional offices has been received at the respective Divisional office, it is organised and imported into the temporary databases of secondary module of dedicated data processing software. The first step towards data validation is making the listing of data thus for various stations in the form of a dedicated format. Such listing of data is taken for **two main objectives: (a) to review the primary validation** exercise by getting the data values screened against desired data limits and **(b) to get the hard copy of the data on which any remarks or observation about the data validation can be maintained and communicated** subsequently to the State/Regional data processing centre.

Moreover, for the case of validation of historical data for period ranging from 10 to 40 years this listing of the screening process is all the more important. This screening procedure involves, for example for daily pan evaporation, minimum or maximum temperature data, flagging of all those values which are beyond the maximum data limits or the upper warning level. It also prepares the data in a well-organised matrix form in which various months of the year are given as separate columns and various days of the month are given as rows. Below this matrix of data the monthly and yearly basic statistics like total and maximum pan evaporation etc. are listed. Also, the number of instances where the data is missing or has violated the data limits is also given.

This listing of screening process and basic statistics is very useful in seeing whether the data has come in the databases in desired manner or not and whether there is any mark inconsistency vis-à-vis expected hydrological pattern.

4. Multiple station validation

4.1 Comparison plots

The simplest and often the most helpful means of identifying anomalies between stations is in the plotting of comparative time series. This should generally be carried out first, before other tests. For climate variables the series will usually be displayed as line graphs of a variable at two or more stations where measurements have been taken at the same time interval, such as 0830 dry bulb temperature, atmospheric pressure or daily pan evaporation.

In examining current data, the plot should include the time series of at least the previous month to ensure that there are no discontinuities between one batch of data received from the station and the next - a possible indication that the wrong data have been allocated to that station.

For climatic variables which have strong spatial correlation, such as temperature, the series will generally run along closely parallel, with the mean separation representing some locational factor such as altitude. Abrupt or progressive straying from this pattern will be evident from the comparative plot which would not necessarily have been perceived at primary validation from the inspection of the single station. An example might be the use of a faulty thermometer, in which there might be an abrupt change in the plot in relation to other stations. An evaporation pan affected by leakage may show a progressive shift as the leak develops (Fig. 1). This would permit the data processor to delimit the period over which suspect values should be corrected.

Comparison of series may also permit the acceptance of values flagged as suspect in primary validation because they fell outside the warning range. Where two or more stations display the same behaviour there is strong evidence to suggest that the values are correct. An example might be the occurrence of an anomalous atmospheric pressure in the vicinity of a tropical cyclone.

Comparison plots provide a simple means of identifying anomalies but not of correcting them. This may be done through regression analysis, spatial homogeneity testing (nearest neighbour analysis) or double mass analysis.

4.2 Balance series

An alternative method of displaying comparative time series is to plot the differences. This procedure is often applied to river flows along a channel to detect anomalies in the water balance but it may equally be applied to climatic variables to detect anomalies and to flag suspect values or sequences. Considering Z_i as the balance series of the two series X_i and Y_i , the computations can be simply done as:

$$Z_i = X_i - Y_i$$

HYMOS provides this option under “Balances”. Both the original time series and their balances can be plotted on the same figure. Anomalous values are displayed as departures from the mean difference line.

4.3 Derivative series

For scrutinising the temporal consistency it is very useful to check on the rate of change of magnitude at the consecutive time instants. This can be done by working out a derivative series. The derivative of a series is defined as the difference in magnitude between two time steps. The series Z_i of a series X_i is simply the difference between the consecutive values calculated as:

$$Z_i = X_i - X_{i-1}$$

Together with the original series the derivative series can be plotted against the limits of maximum rate of rise and maximum rate of fall. This gives a quick idea of where the rate of rise or fall is going beyond the expected values.

4.4 Regression analysis

Regression analysis is a very commonly used statistical method. **In the case of climatic variables where individual or short sequences of anomalous values are present in a spatially conservative series, a simple linear relationship with a neighbouring station** of the form:

$$Y_i = a X_i + c$$

may well provide a sufficient basis for interpolation.

In a plot of the relationship, the suspect values will generally show up as outliers but, in contrast to the comparison plots, **the graphical relationship provides no indication of the time sequencing of the suspect values** and whether the outliers were scattered or contained in one block.

The relationship should be derived for a period within the same season as the suspect values. (The relationship may change between seasons). The suspect values previously identified should be removed before deriving the relationship, which may then be applied to compute corrected values to replace the suspect ones.

In HYMOS the validation section provides a section on “Relation Curves” which gives a number of variants of regression analysis including polynomial regression and introduction of time shifts, generally more applicable to river flow than to climate. A more comprehensive description of regression analysis is given in the Chapter on “Series completion and regression”

4.5 Double mass curves

Double mass curve analysis has already been described in the secondary validation of rainfall (Module 9) and a full description will not be repeated here. It may also be used to show trends or inhomogeneities between climate stations but it is usually used with longer, aggregated series. However in the case of a leaking evaporation pan, described above, the display of a mass curve of daily values for a period commencing some time before leakage commenced, the anomaly will show up as a curvature in the mass curve plot.

This procedure may only be used to correct or replace suspect values where there has been a systematic but constant shift in the variable at the station in question, i.e., where the plot shows two straight lines separated by a break of slope. In this case the correction factor is the ratio of the slope of the adjusted mass curve to the slope of the unadjusted mass curve. Where there has been progressive departure from previous behaviour, the slope is not constant as in the case of the leaking evaporation pan, and the method should not be used.

4.6 Spatial homogeneity (nearest neighbour analysis)

This procedure has already also been described in Module 9 for rainfall for which it is most commonly used and will not be covered fully again. Its advantage for rainfall in comparison to climate is that there are generally more rainfall stations in the vicinity of the target station than there are climate stations. The advantage for some climate variables is that there is less spatial variability and the area over which comparison is permitted (the maximum correlation distance R_{max}) may be increased.

Although there is strong spatial correlation, there may be a systematic difference due, for example to altitude for temperature. In these cases normalised rather than actual values should be used. This implies that the observations at the neighbour stations are multiplied by the ratio of the test station normal and the neighbour station normal:

$$T_{ci} = (N_{test} / N_i). T_i$$

where:
T_{ci} = Variable corrected for difference of mean at neighbour station
N_{test} = Mean of test station
N_i = Mean of neighbour station i

The analysis provides an estimate of a variable at a target station on the basis of a weighted mean of the neighbour stations, weighted as a function of the distance from the target station. It provides a list of those values (flagged values + or -) which are outside a specified range (mean + standard deviation times a multiplier), and provides the option of replacing the 'observed' values with the 'estimated' values.

5. Single series tests of homogeneity

Single series testing for homogeneity will normally only be used with long data sets and therefore will have to await the data entry of historical data. Once these are in place it will be necessary to inspect them for homogeneity and especially for trend. Even here it is expected that spatial comparison of two or more series will be more commonly, but not exclusively used.

It is not expected that these methods will be widely used for current data.

Series may be inspected graphically for evidence of trend and this may often be a starting point. However statistical hypothesis testing can be more discriminative in distinguishing between expected variation in a random series and real trend or more abrupt changes in the characteristics of the series with time.

5.1 Trend analysis (time series plot)

A series can be considered homogeneous if there is no significant linear or curvilinear trend in the time series of the climatic element. The presence of trend in the time series can be examined by graphical display and/or by using simple statistical tests. The data are plotted on a linear or semi-logarithmic scale with the climatic variable on the Y-axis and time on the X-axis. The presence or absence of trend may be seen by examination of the time series plot. Mathematically one may fit a linear regression and test the regression coefficients for statistical significance.

Trend generally does not become evident for a number of years and so the tests must be carried out on long data series, often aggregated into monthly or annual series. **Trend may result from a wide variety of factors including:**

- change of instrumentation
- change of observation practice or observer
- local shift in the site of the station
- growth of vegetation or nearby new buildings affecting exposure of the station
- effects of new irrigation in the vicinity of the station (affecting humidity, temperature and pan evaporation
- effects of the urban heat island with growing urbanisation
- global climatic change

The presence of trend does not necessarily mean that part of the data are faulty but that the environmental conditions have changed. Unless there is reason to believe that the trend is due to instrumentation or observation practices or observer, the data should not generally be altered but the existence of trend noted in the station record.

5.2 Residual mass curve

A residual mass curve represents accumulative departures from the mean. It is a very effective visual method of detecting climatic variabilities or other inhomogeneities. The residual mass curve is derived as:

$$Y_i = Y_{I-1} + (X_i - m_x) = \sum_{j=1}^{i} (x_j - 1/N \sum_{k=1}^{N} X_k)$$

where:
N = number of elements in the series
m_x = mean value of X_i , i=1,N

the curve can be interpreted as follows:

- an upward curve indicates an above average sequence
- a horizontal curve indicates an about average sequence
- a downward curve indicates a below average sequence

5.3 A note on hypothesis testing

Hypothesis testing forms a framework for many statistical tests; it is concerned with an assumption about the distribution of a statistical parameter. The assumption is stated in the null-hypothesis H_0 and is tested against and alternative formulated in the H_1 hypothesis. The parameter under investigation is presented as a *standardised variate* called a *test statistic.* Under the null-hypothesis the test statistic has some standardised sampling distribution, e.g. a standard normal or a Student's t-distribution. For the null hypothesis to be true the value of the test statistic should be within the acceptance region of the sampling distribution of the parameters under the null-hypothesis. If the test statistic does not lie in the acceptance region, expressed as a significance level, the null-hypothesis is rejected and the alternative is assumed to be true. Some risk is involved however in being wrong in accepting or rejecting the hypothesis. For a brief note on Type I and Type II errors refer to the HYMOS manual

5.4 Student's t tests of stability of the mean

The series may be tested for stability of variance and mean. Considering only the simpler case where the variance has not changed, the test for stability of the mean requires computing and then comparing the means of two or three sub-sets of the time series. The 't' value is computed as follows and is compared with the tabulated Student's 't' which is the standardised test statistic, for corresponding degrees of freedom, for testing its significance:

$$|t| = |x_1 - x_2| / S_{12}$$

and

$$S_{1,2} = \sqrt{\left\{ \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} * \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}}$$

where n_1 and n_2 are the number of data points in each sub-set, x_1 and x_2 are the corresponding means and S_1^2 and S_2^2 are the corresponding variance values. The number of values in each sub-set are usually taken as equal but if the graphical analysis indicates presence of trend in a particular part of the time series, the sub-sets can be decided accordingly.

Example:

Mean annual temperature is recorded at a station from 1978 to 1993 as follows:

Year	78	79	80	81	82	83	84	85	87	88	89	90	91	92
	86								93					
Annual mean temp. °C	18.5	16.5	17	19	18.5	18	18.5	17.5	17	15.7	16.5	17.3	16.5	15.5
	19.5								17.5					
				Mean 18.1							Mean 16.5			

By applying the above equation for stationarity it can be demonstrated that the data at station A is not-homogeneous.

A more robust test, can be carried out by comparison to homogeneity at adjacent stations which can in turn be used for adjustment of the non-homogeneous series if this seems appropriate.

5.5 Wilcoxon W-test on the difference of means

The Wilcoxon test again tests under the null-hypothesis that the means of two series A_i (i = 1,m) and B_j (j = 1,n) are the same. All values of A_i are compared with all B_j defining a value of $w_{i,j}$ as follows:

where $A_i < B_j$ then $w_{i,j} = 2$
$A_i = B_j$ $w_{i,j} = 1$
$A_i > B_j$ $w_{i,j} = 0$

The Wilcoxon statistic W is formed by:

$$W = \sum_{I=1}^{m} \sum_{j=1}^{n} w_{i,j}$$

Where the means are the same the W-statistic is asymtotically *normally* distributed with N (μ_w , σ_w)

where $\mu_w = mn$

$\sigma_w^2 = mn\,(N+1)/3$

$N = m + n$

The absolute value of the following standardised test statistic is computed

$$|u| = |W - \mu_w| / \sigma_w$$

and comparison is made against a tabulation of the Normal distribution to test the validity of the null- hypothesis at the significance level.

5.6 Wilcoxon Mann-Whitney U-test

The Wilcoxon-Mann-Whitney U-test investigates whether two series of length m and n are from the same population. The series may again may be split samples from a single series or series from paired instruments on the same site but as we would hardly expect series from different sites to be from the same population, this comparison is not recommended for such comparisons.

The data of both series are ranked together in ascending order. Tied observations are assigned the average of the tied ranks. The sum of ranks in each of the series, S_m and S_n, is then calculated. The U statistic is then computed as follows:

$$U_m = mn + m(m+1)/2 - S_m$$

$$U_n = mn - U_m$$

$$U = \min(U_m , U_n)$$

A correction for tied values is incorporated as follows:

(a) tied observations are given the average rank, and
(b) the variance of U is corrected by a factor F_1

$$F_1 = 1 - \Sigma (t^3 - t) / (N^3 - N)$$

where: t = number of observations tied for a given rank.

If the elements of the two series belong to the same population then U is approximately *normally* distributed with $N(\mu_U , \sigma_u)$:

$$\mu_u = mn / 2$$

$$\sigma_U^2 = F_1 . mn(N + 1)/12$$

where $N = m + n$

For the null hypothesis that the series are from the same population, the absolute value of the following standardised test statistic is computed:

$$| u | = | U + c - \mu_u | / \sigma_u$$

where c = a continuity correction; $c = 0.5$ if $U < \mu_u$, and $c = -0.5$ if $U > \mu_u$

and comparison is made against a tabulation of the Normal distribution to test the validity of the null- hypothesis at the significance level.