

## Chapter 6

### Step 4 - Historical Data Analysis and Processing

#### Introduction

This step involves analyzing and processing those data that are not yet in the form needed for model calibration. These primarily are precipitation, temperature, and evaporation data. Lumped hydrologic models require mean areal estimates of these variables, thus this step involves the computation of mean areal values from the station data retrieved in step 1. Even if point model computations are to be done, the station data need to be checked for consistency and any missing data gaps filled in. In addition to the areal estimates of model input variables, mean daily discharge data may need to be checked and modified for the effects of diversions or other control structures. The other data needed for model calibration mentioned in chapter 3 should already be in the form needed for model calibration and do not require further processing. These additional data are typically used to verify or modify simulation results.

This is a very important step in the calibration process. There is a tendency in many cases for people to rush through this step in a mechanical, cookbook fashion without carefully evaluating data consistency, understanding the variability in the data fields, and determining how to properly use the available data to generate physically realistic estimates of these input variables. This can easily result in time series that are inconsistent and biased. This is especially true in areas with significant spatial variability in data values on a seasonal or annual basis such as mountainous regions. The snow and rainfall-runoff models are designed to represent those processes and are not designed to deal with the interpolation and extrapolation of data fields. The models do contain some very simple factors for making minor data adjustments (e.g. the SAC-SMA operation contains a precipitation and evaporation multiplier and the SNOW-17 operation includes a precipitation multiplier and a lapse rate adjustment), but these factors cannot make seasonal modifications or deal with inconsistencies. In order for the models to work properly, the input data should represent as closely as possible the scale and seasonal variations that actually occur in nature.

There is also a tendency for people to immediately start to process precipitation and temperature data after downloading the station information. It is very important to take the time for steps 2 and 3, i.e. assessing spatial variability and selecting flow points and period of records. An understanding of the variability of hydrologic and climatic conditions is essential to choosing the proper data analysis procedure. A careful determination of the flow points to be modeled and the period of record to use will avoid having to regenerate time series for different subdivisions of the river basin and different data periods during the calibration process. By carefully doing steps 2 and 3, a considerable amount of time and effort can be saved in the long run and the entire process made more successful and efficient.

The techniques used for data consistency checks and estimating missing data and the procedures recommended for determining mean areal values and assigning station weights referred to in this chapter are not highly sophisticated. This chapter uses well established techniques and proced

ures based on engineering judgement. Most of the methods referred to in this chapter were devised by people primarily interested in hydrologic modeling. In order to use the models it is obviously necessary to get the data into the form needed for calibration and operational applications.

The procedures described in this chapter for data analysis and processing are an engineering solution to the problem, not a scientific hydro-meteorological solution. The underlying objectives for coming up with these methods for preparing the data were to have estimates that reflect the proper scale of what actually occurs in nature and to be able to reasonably guarantee that there would be minimal bias between historical and operational estimates of data quantities. This last objective is critical to successfully using calibration results for operational applications. There is more discussion of potential sources of bias between historical and real time analyses in chapter 8.

This chapter provides a discussion of general topics and tasks that are important in the analyzing and processing of all types of historical data. Details for performing these tasks and recommendations on the steps to follow and procedures to use when working with individual data types are contained in the subsequent sections.

### Precipitation, Temperature, and Evaporation

#### Effect of Data Bias on Model Results

The Sacramento and snow models are very sensitive to biased data. Thus it is very important to remove as much bias as possible when generating areal estimates of input variables. This not only includes overall bias as reflected in mean annual values, but also seasonal biases that can occur due to such things as variations in the precipitation versus elevation patterns and lapse rates.

Again remember that the models are not designed to represent how variables like precipitation and temperature vary over the area. These variations need to be handled by the data analysis procedures.

Table 6-1. Change in runoff based on a change in precipitation

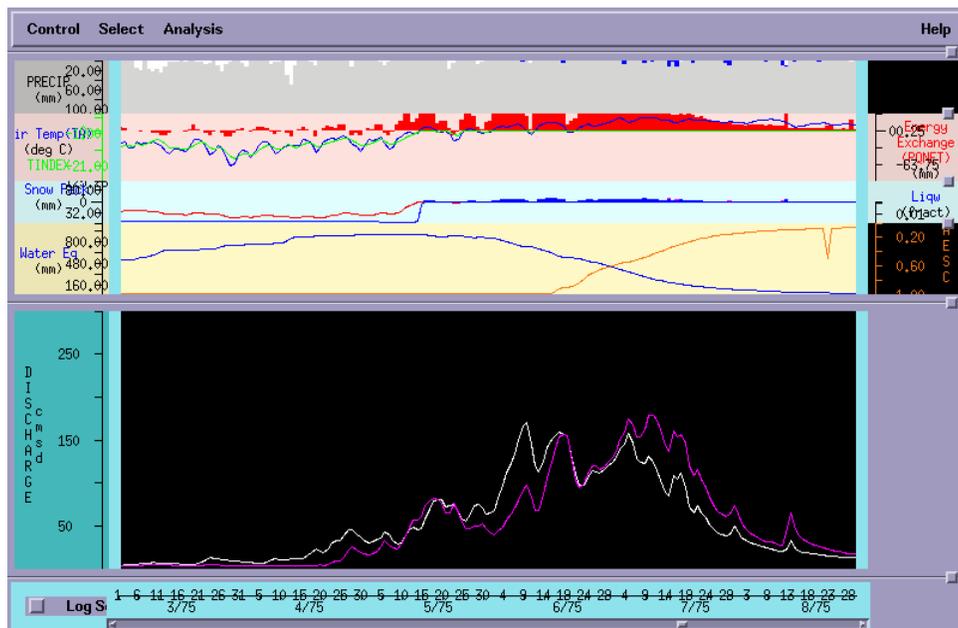
Watershed	Change in Precipitation	Change in Runoff
Leaf R. nr. Collins, MS	+10 %	+24.9 %
Bird Ck. nr. Sperry, OK	+10 %	+28.8 %
Smith R. nr. Bristol, NH	+10 %	+17.8 %
Animas R. at Durango, CO	+5 %	+9.7 %

Table 6-1 shows how a change in the overall amount of precipitation will effect the amount of runoff generated. As can be seen the percentage change in runoff is greater than the relative change in the amount of precipitation. Typically the drier the region the greater the difference. It can be seen that a bias in the amount of precipitation will have a significant effect on the amount of runoff produced by the models. In order to compensate for data bias, either evaporation amounts will have to be warped or model parameters, especially ones affecting the water balance such as soil tension water capacities, will have to be changed. Neither of these adjustments will produce the same results as if the precipitation data were unbiased. Also, evaporation demand curves and soil tension water capacities that should vary in an understandable pattern over the river basin will vary unrealistically if different amounts of precipitation bias exists from one watershed to another within the basin.

Figure 6-1 shows the effect of changing the temperature by only 3°F on the timing of snowmelt for the Animas River at Durango, Colorado. This response is typical of a similar magnitude temperature change on other watersheds where snowmelt is significant. In this case the temperature for both the upper and lower elevation zones were changed by the same amount. As can be seen this relatively small change in temperature causes a fairly large shift in the timing of the snowmelt. If biased temperature data are used in a calibration, the non-rain melt factor parameters and parameters that affect the heat deficit and liquid water storage prior to melt could be changed to partly compensate for the biased temperature input, however, the results would not be nearly as good as if the data were properly corrected. In extreme cases one would be tempted to use a melt base (MBASE parameter) of other than 0°C in order to correct the timing of the snowmelt.

The use of a non 0°C MBASE value almost always is an indicator of a bias problem with the temperature time series. Also, as with precipitation, one would find that the pattern of melt factors across the river basin would be unrealistic if temperature inputs were biased by different amounts from watershed to watershed.

Figure 6-1. Effect of a 3°F change in temperature - Animas River at Durango, CO  
 Biased evaporation estimates will also change the amount of runoff computed though the effect is not as great as with precipitation. Model runs on several basins indicated that the change in runoff due to modifying evaporation was about 40% of the change in runoff due to varying precipitation by a given percentage. Again, if such bias exists, it will cause model parameters to be distorted and the variations from watershed to watershed within the river basin will likely not follow



expected patterns.

It is very important to carefully analyze the data during this step in the calibration process to minimize any data bias. This is most important in areas where the data values vary from location to location, typically mountainous areas. In such areas it is relatively easy to create biased input time series especially when extrapolating data to ungaged portions, such as high elevation zones.

#### Mountainous versus Non-Mountainous Area Procedures

Different techniques and procedures are recommended for use in analyzing and processing data based on whether the area is considered to be mountainous or non-mountainous. A non-mountainous area is defined as:

*An area where the long term average values of the variable being analyzed are essentially the same at all locations within the area. This refers to both annual and seasonal averages.*

Typically for a non-mountainous area the mean values of the variable being considered should be within a range of about  $\pm 5\%$  over the entire basin. There can be a general trend of increasing or decreasing average values as one moves across the area, but there should not be any rapid transitions. For non-mountainous areas the recommended techniques and procedures assume that any station can be used to estimate missing data at another station without making any adjustments for differences in magnitude and that areal averages can be computed by weighing stations merely as a function of their x,y plane location (sum of weights will always equal one).

A mountainous area is defined as:

*An area where there are significant differences in the long term average values of the variable being analyzed over the area. These can be differences in annual or seasonal averages.*

Generally an area that needs to use mountainous area procedures will have terrain differences as these are the most common reason that long term means of a variable will differ from one location to another, however, areas with flat terrain can in some cases require the use of mountainous area methods. For example, long term averages of precipitation, temperature, or evaporation could vary significantly based on the distance from the coast or a large water body even though the terrain is generally level. In the mountainous area procedures, techniques for estimating missing data account for the long term difference in the means between stations. Also, factors other than merely location are used to determine the station weights used to calculate mean areal values (sum of weights do not have to equal one).

It is very important to determine whether mountainous or non-mountainous area procedures should be used prior to analyzing and processing the data for model computations. As far as estimating missing data, if any portion of the river basin conforms to the mountainous area definition, then the entire river basin should use mountainous area techniques. When determining station weights, the use of mountainous versus non-mountainous area methods can vary from one area to another within the basin. To use non-mountainous area methods to determine station weights, al

If the stations that will be assigned weight for the watershed or subarea should have the same long term average value and all points within the boundaries should have the same mean.

### Subdivision of Watersheds

When the flow points to be modeled are chosen, this defines the headwater and local areas for which data estimates need to be provided. However, in some cases these drainages need to be divided into subareas in order to properly model the variations that occur in such quantities as runoff, snow cover, and soil moisture. While areas could be subdivided in many different ways based on such factors as elevation, distance from gage, vegetation cover, soil type, and land use, there are two ways that are commonly used. The first is to divide the drainage into elevation zones and the second is to divide the area based on hydrograph response. Since the conceptual models are being applied in a lumped rather than distributed manner in this manual, the headwaters and local areas are divided into the minimum number of subareas needed to handle the effects of the spatial variability. Typically when subdivision is necessary only 2 zones are required, though in a few cases with extreme variations in conditions, 3 or even 4 zones may be required. From an operational standpoint, it has been found that the fewer the zones, the easier it is to manage real time adjustments to model states and computations.

Elevation zones are used in mountainous areas when the distribution of climatic variables and possibly physiographic factors create significant differences in the amount and timing of runoff. In the United States these differences are almost always related to varying amounts of snow cover and differences in the timing of snowmelt over the drainage. While these differences are not completely based on elevation, elevation is the dominant reason for the variations and elevation zones can easily be delineated. In most of the country when elevation zones are needed, two zones are sufficient with the models used in NWSRFS due to the inclusion of the areal depletion curve concept within the snow model. There are other snow models that divide a mountainous area into many elevation zones. These models don't use an areal depletion curve, but instead model areal cover based on which zones have snow and which are bare. However, even with the use of the depletion curve there are some areas that will require more than two elevation zones in order to get satisfactory results. These are typically drainages with large elevation variations or watersheds that are partly covered by glaciers. Also, if significant rainfall events are common late in the melt season, additional zones may be needed. Typically there is more snowmelt than ET during much of this period so that the tension water storages in the soil model remain full, even though only a small portion of the area is covered by snow. In reality, when rain occurs late in the melt season, areas that have been bare of snow for some time have dried out and thus, the amount of runoff tends to be over computed. Adding additional zones can reduce this problem.

Section 6-1 describes some criteria for determining when elevation zones are needed and guidelines for selecting the elevations between zones.

The second main reason for subdividing a headwater or local area into subareas is based on variations in hydrograph response due primarily to differences in the spatial distribution of rainfall. If the time to peak and/or the shape of the hydrograph can vary significantly based on where most of the rainfall occurs, then the drainage should possibly be subdivided so that these differences

can be modeled. Besides differences in response, a precipitation gage network that can adequately define the spatial variation in the rainfall is necessary for the subdivided drainage to produce improved model results. One of the common cases for subdividing based on hydrograph response is when the shape of the drainage is long and narrow. In such a watershed, the response is normally quite different when most of the runoff comes from the upper end of the area as compared to a storm centered just above the gage. In this case the drainage would be subdivided by travel time. In other cases where the shape of the response varies, the subdivision might be based on the configuration of the channel network, e.g. one main branch of the stream network might be separated from another branch.

After making decisions as to whether any of the headwater and local areas need to be subdivided and, if so, delineating subareas based on elevation or drainage boundaries, the areas for which precipitation, temperature, and evaporation estimates must be computed are now defined. Also, at this point it should have been decided whether mountainous or non-mountainous area methods will be used to produce the areal estimates. The next step is to check the consistency of the data prior to computing the mean areal values.

### Data Consistency Checks

The station data used to compute areal estimates of precipitation, temperature, and evaporation should first be checked for consistency and adjusted if necessary. By consistency we mean whether the data record is consistent over time, i.e. whether the record at a station maintains the same relationship over time to the other stations in the basin. Station consistency is checked by using a double mass analysis. In this technique the data for each station is plotted against the average of the data for a group of other stations. If the group is reasonably large, the average for the group is not significantly affected by inconsistencies in individual stations, but changes in the record for the station plotted against the group will be revealed. The main problem associated with consistency checks is separating real changes in the stations data record from natural variations that occur. When making consistency checks and adjustments it is very important not to remove the natural variability of the data. The underlying rule when making adjustments should be that if there is any doubt that the correction should be made, then don't make the adjustment.

Inconsistencies in data records occur primarily for two reasons. First, the station is moved to a new location with a somewhat different climatic regime and a different exposure. This is the most common cause of an inconsistency. Second, there is a change that occurs at the site, such as an equipment change or a change in the exposure of the gage. This can include a new type of gage being installed, a wind shield installed or removed from a precipitation gage, a new building built near the site, or vegetation growing or removed from near the instrumentation. For many of the newer automated networks, moves and site changes are infrequent, thus inconsistencies in the data record are rare. Within the NCDC climatic network there are also stations that have remained in one place with the same equipment for many years, however, there are also stations that are moved periodically due to changes in observers. These are the stations that most frequently require adjustments for inconsistencies. Unfortunately, the records of station moves and equipment changes for the climatological network are not totally reliable. Comparisons made bet

ween different sources of station changes such as the Daily Climatological Bulletins, NCDC metadata files, B-44 forms filed by network managers, and records from state climatologists are not always consistent. Thus, changes may have occurred at a station even if the records you have available at the time do not indicate a move or equipment change. However, going along with the underlying rule mentioned earlier, if there is no documentation of a station change, don't make an adjustment unless it is very clear that a correction is needed.

The reason for making adjustments and removing inconsistencies prior to using the data to compute areal estimates of the input variables is to avoid having differences or bias between one portion of the historical record and another. If a reasonably sizeable inconsistency exists and is not removed and that station has a large weight in the computation of an areal estimate, then a significant bias will result from one portion of the areal estimate to another. As we have seen, the hydrologic models are quite sensitive to data bias. Model parameters determined using the record prior to the inconsistency will not be the same as parameters based on the period after. Results obtained using parameters determined from one period will be biased when those parameters are run on the other portion of the period of record. If the model parameters are based on a period that doesn't reflect the current status of the station, then operational results will be biased compared to those obtained during calibration.

Details on the procedures used to make consistency checks and adjustments and guidelines for using these procedures are contained in Section 6-2.

### Computation of Mean Areal Values

Once the data have been checked for consistency and any corrections made, it is time to generate mean areal estimates of precipitation, temperature, and evaporation for use in model calibration. The analysis methods and techniques used vary as to whether an area is classified as mountainous or non-mountainous. The objective is the same in all cases, i.e. to obtain an estimate that is as unbiased as possible compared to what occurred in nature and to minimize the random errors in those estimates. For precipitation and temperature the methods used in NWSRFS involve estimating all missing station data and then determining appropriate weights to assign to the stations to compute a mean areal estimate of the variable. This is also true for evaporation in some cases, but more frequently average climatic estimates of evaporation are used when calibrating the models. The recommended procedure for generating mean areal estimates of precipitation are in Section 6-3, temperature is in Section 6-4, and evaporation is discussed in Section 6-5.

## Mean Daily Discharge Data

### Possible Adjustments

Mean daily discharge data are used primarily to evaluate simulation results by comparing the model computed values to the observed data. For streams that do not have any man-made control structures, the model simulations can be compared directly against the observed natural flow in the river. When control structures exist, there are two options. The first is to model the operati

ons of all of the controls so that a direct comparison can be made between the computed discharge and the observed controlled flow. The second is to adjust the observed daily discharge to remove the effects of the control structures so that these time series now reflect “observed” natural flow conditions. Prior to using the streamflow data, it is probably a good idea to check its consistency. Such a check could uncover variations in measurement techniques, but more likely will provide insight to alterations in runoff produced by such things as forest fires, changes in land use, different agricultural practices, variations in irrigated acreage, and changes to diversions which might otherwise not be noticed. Instantaneous streamflow data, when available, could also be adjusted for control structures, but this is more difficult and is usually not done. Section 6-6 describes possible checks and adjustments to mean daily discharge data.

### Extension of Historical Data Record

As time goes by there likely will be a need to extend the historical data record. Reasons for extending the record were listed in chapter 2. When doing a data extension, it is very important to insure that the additional record is unbiased and consistent with the previous record. Since model parameters are based on the data used in the original historical analysis, the time series extensions must be unbiased as compared with the initial period used for calibration. When generating extensions, it is critical to do the following:

- New data for stations that were included previously should be checked for consistency with the prior data for these stations and adjusted if necessary (this typically requires that station data for some period prior to the extension, typically about 10 years, be available to use when generating consistency check plots - any consistency adjustments that were used for these stations for this period should be applied before producing the plots). Also, data for any new stations that are being added to the analysis should obviously be checked for consistency.
- Station and areal means for precipitation, temperature, and evaporation that were determined for the original complete historical data analysis should continue to be used until another complete data analysis is performed (e.g. if 30 years of data were used when the river basin was first calibrated, and later the record was extended by 10 years, and now it is being extended by another 10 years, the means for the original 30 years of data should be used when processing this latest extension), and
- Mean values for new stations need to be determined based on their relationship to stations with a long historical record that were used in the original analysis using the techniques included in PXPP for precipitation (uses the average ratio of precipitation at each new station to the amount recorded at a well established station) or as outlined in Section 6-4 for temperature data (uses the average difference between each new station and a station with a long record). Typically, based on experience, about 5 years of data are needed at a new station in order to compute a stable estimate of the mean values. Less data can be used, but the estimate of the mean values will not be as reliable.

Historical data extensions are possible when the types of measurements, processing methods, an

d networks are the same as used in the original analysis. When new types of measurements (e.g. radar based estimates of precipitation), new processing methods (e.g. switching from non-mountainous area to mountainous area procedures), or significant changes in the network (e.g. the addition of a high elevation gage network in an area where previously there were low elevation stations) occur, the historical data analysis should generally be completely redone. In some of these situations it may be possible to produce values that are consistent and unbiased compared to the previous analysis and thus can be used to extend the existing record, but this is difficult to accomplish. In addition, for many of these cases the period of the historical data record that can be used will change since the new measurements or network are only available for a limited time. Unless the new data can be proven to be unbiased and consistent with the previous record, the models must be recalibrated so that the parameters are compatible with the new data. This new historical data period then becomes the base period to use for any future extensions of the record.