HYDROLOGY PROJECT
World Bank & Government of The Netherlands funded

Training module # SWDP - 46

# *How to organise data into temporary databases*

New Delhi, February 2002

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: hydrologyproject@vsnl.com

DHV Consultants BV & DELFT HYDRAULICS

with
HALCROW, TAHAL, CES, ORG & JPS

# *Table of contents*

**Page**

# 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 Organise Data into Temporary Databases |
| **Target group** | : | HIS function(s): …… |
| **Duration** | : | x session of y min |
| **Objectives** | : | After the training the participants will be able to*:* |
| **Key concepts** | : | • |
| **Training methods** | : | Lecture, exercises |
| **Training tools required** | : | Board, flipchart |
| **Handouts** | : | As provided in this module |
| **Further reading and references** | : | |

# 3. Session plan

| No | Activities | Time | Tools |
|---|---|---|---|
| 1 | ***Preparations*** | | |
| 2 | ***Introduction***: | min | OHS x |
| | | | |
| | ***Exercise*** | min | |
| | ***Wrap up*** | min | |

# 4. Overhead/flipchart master

# 5. Handout

**Add copy of the main text in chapter 7, 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**

# 1 Organisation of data into temporary databases

## 1.1 Separation between data processing and data storage functions

Under HIS the data processing and the data storage functions are separated; data processing is carried out in various Data Processing Centres in a distributed manner at three distinct levels, whereas the data archival is accomplished in the Data Storage Centres. This separation between the data processing and the data storage functions is illustrated in Figure 6.1.

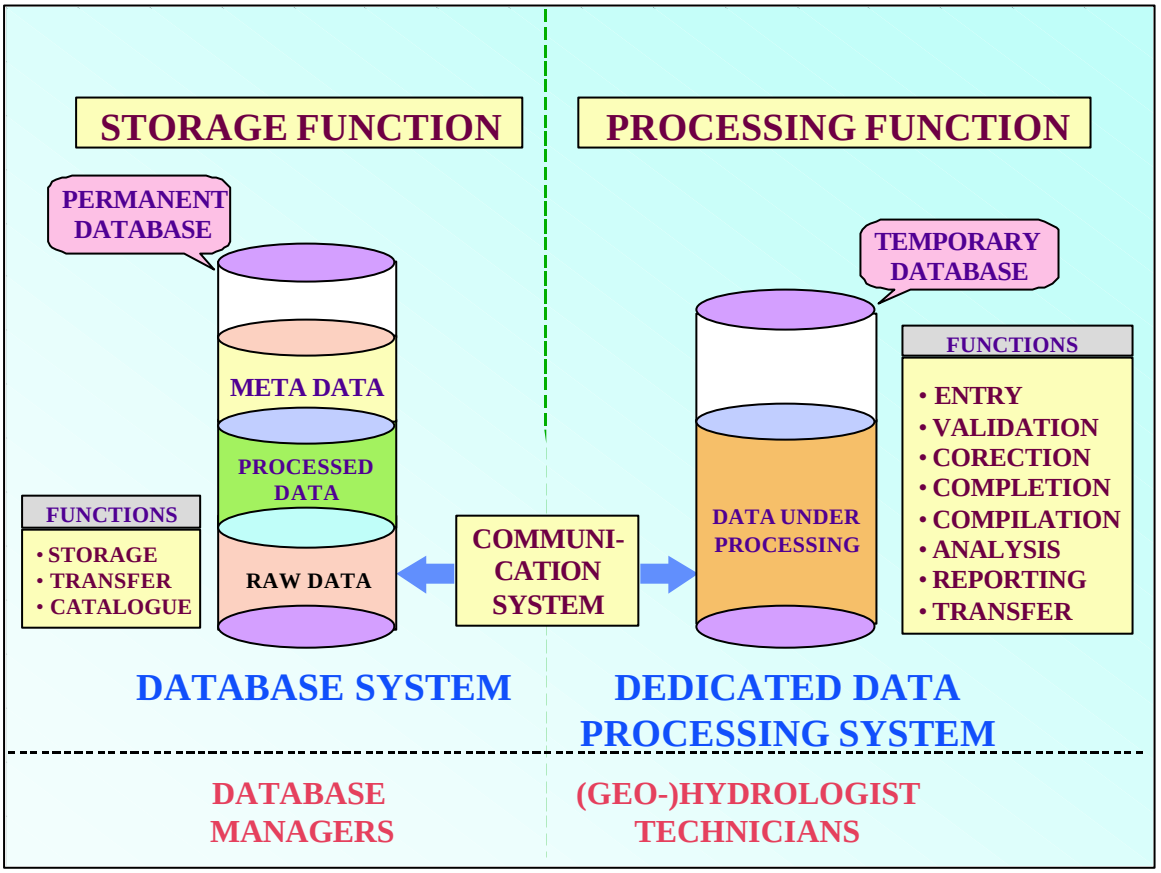


**Figure 1.1: Illustration of clear separation between data processing and data storage functions**

As can be seen in the illustration above, functions like data entry, validation, correction, completion, compilation, analysis and reporting are required to be undertaken under data processing activities whereas functions like storage, retrieval and cataloguing are required for data storage activities. Data in the databases under data processing systems are processed hydrologically for the purpose of validation, completion, correction and compilation and hence may get revised. On the other hand, data in the data storage systems is not required to be revised at all. The functions needed in the data processing system are the ones needed for a comprehensive hydrological data processing whereas those needed in the data storage system are purely for an efficient storage and management of data. It is also to be realised that the personnel required for the data processing activities are those knowing the hydrological data processing concepts whereas those required for the data storage function are those who are specialists on data management and information technology aspects.

Such a distinct separation will be very helpful in making an unambiguous hydrological data archival which is essential for its sustainability and would also incidentally provide an adequate scope for any upgradation which might be required for either of the two systems independently.

## 1.2 What are temporary, permanent and transfer databases?

As described above, the data processing centres are expected to provide excellent hydrological data processing opportunities whereas data storage centres are to function as effective electronic hydrological data libraries. Since the data under data processing centres are under processing and are liable for a change the character of data sets held are called as temporary and hence the name ***temporary databases***. On the other hand, the data sets in the data storage centres are not to be modified at all and hence the name ***permanent databases***.

Thus all the databases which are initiated by and for the dedicated hydrological data processing systems (SWDES and HYMOS) at the three levels; Sub-Divisional, Divisional and State Data Processing Centres, would be called as temporary databases. Similarly, the databases in which all the data is archived at the data storage centres are called as permanent databases.

All other databases which are used for transferring/exchanging data between two data centres would be termed as ***transfer databases***. These transfer databases may be required for transfer/exchange of data from a Sub-Divisional Data Processing Centre to a Divisional Data Processing Centre or from a Divisional Data Processing Centre to State/Regional Data Processing Centre or from one Data Storage Centre to another Data Storage Centre. This is as illustrated in Figure 6.2 below.

## 1.3 What are Field and processed databases?

The data as observed in the field by the observers is entered in the field note books and the data entry forms. The same data is keyed-in in the computer using SWDES. During the data entry a few mistakes may occur and the data entered and available on computer may not be exactly same as in the field forms. Upon detection of such mistakes, by employing simple data entry checks and graphical viewing, these errors are corrected such that the data in the computer database is the same as is available in the field forms. However, sometimes the data available on the field form itself bear some error that is very clear and self-evident. These are due to calculation mistakes or slippage in the decimal places while scribbling on the forms by the observer. Highly self-evident errors are corrected in the computer database and an equivalent correction and remark made on the manuscript. Thus the data as observed in the field by the observers and finally available in the computerised databases is termed as the ***field database***.

This field data as available in the databases may have gaps or inconsistencies. Such gaps and inconsistencies are looked into at the time of processing the data. Wherever sufficient related data is available for filling-up these data gaps and correcting the inconsistencies, appropriate data in-filling and data correction procedures are employed. In this process of in-filling and correction, some of the data values get filled up or modified. This modification is carried out on the copy of the raw data set and not on the original itself. Such a of modified set of data is termed a ***processed database***.

A set of ***filed data*** is always to be stored for the purpose of future reference and archival. This availability of field data set is maintained at all DPCs, as ***temporary databases***, for the requirement of day-to-day reference to the field data at the time of data processing. The availability of field data at the Data Storage Centre (DSC), as ***permanent databases***, is

purely for the purpose of long term archival and for limited dissemination for specific purposes, if required.

***Processed data*** sets in the DPCs are the ***temporary databases*** that are worked upon during validation and processing whereas those in the DSCs are the final data sets as ***permanent databases***.

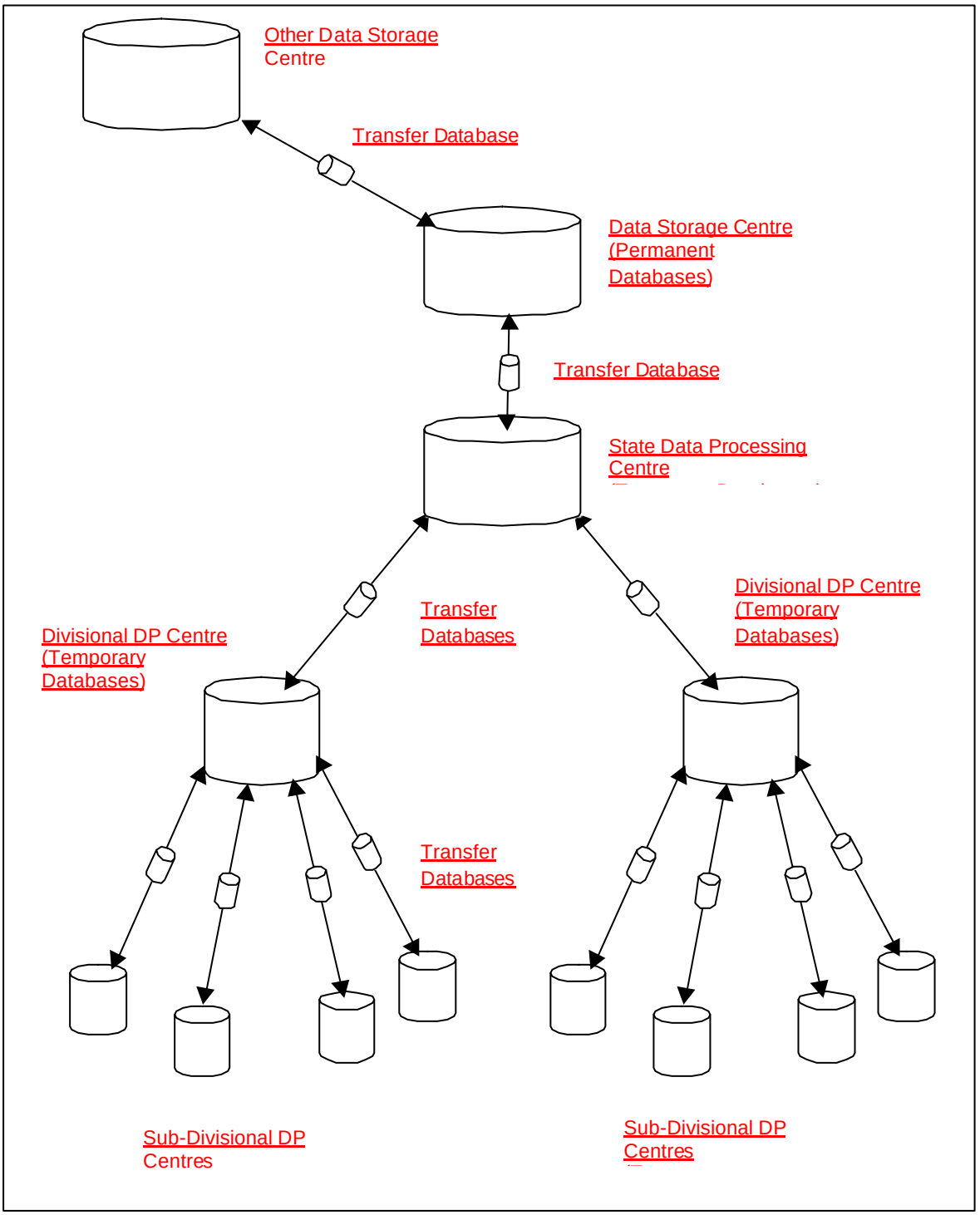


**Figure 1.2: Illustration of permanent, temporary and transfer databases**

Organising temporary databases at Sub-Divisional Data Processing Centres

Surface Water Data Entry System (SWDES) is to be employed at the Sub-Divisional Data Processing Centres of all surface water agencies for entering the required observed meteorological and hydrological data. SWDES has been designed with a view to have a software using a standard, inexpensive and commonly available database software. Microsoft ACCESS software, being in the family of Microsoft Office suite, is expected to be available on all the computers being used by the agencies. This Microsoft ACCESS database, therefore, has been used as the back-end database in SWDES.

All the data entered using SWDES is organised in a **well-defined database** which is also called as ***workarea***. The software can initiate and maintain one or more workareas as per the requirement of the user. However, only one workarea is active at any point of time while working with the software. There are no restrictions put on the size of these workareas other than the one of the ACCESS databases themselves which is not restrictive in anyway for such small front-end applications. Though all the data from all the stations under a Sub-Divisional Data Processing Centre can be stored in one workarea itself but it is preferable to use smaller workareas for the purpose of better management, maintaining adequate speed while working and to categorise data in well-defined units. The suggested approach to be practised at the Sub-Divisional Data Processing Centres is as given here.

A **Sub-Divisional Data Processing Centre** (SDDPC) of the state SW agencies and the Central Water Commission (CWC) normally has a few river gauging stations, a few rainfall stations and a few full climatic stations under its jurisdiction. For the CWC, there are a maximum of 10 river gauging stations, on an average, in any SDDPC. The SDDPCs of CWC does not have any other type of station besides a few flood forecasting gauge stations. The number of observation stations under the SDDPCs of the state agencies is relatively larger. On an average, the state SDDPCs may have a maximum of 10 river gauging stations, 20 rainfall stations and 3 HP full climate stations.

From the hydrological data processing point of view it is useful to keep the data from all the stations together in one workarea so that the data of any of the station can be viewed immediately. But from the management and clearness sake it is preferable to have multiple databases categorised by the type of station; river gauging station, rainfall station and HP full climate station. Though the prevailing number of stations under a SDDPC can be organised in a single workarea, it is preferable to use multiple workareas whenever the numbers are very high. As a guideline, when there are upto about 10 river gauging stations, 10 raingauge stations and 4 HP Full Climatic stations, data from all the stations can be organised in one single workarea.

If the numbers are exceedingly high as compared to the ones given above then data can be organised in more than one workarea. While categorising data in multiple workareas one underlining principle has to be followed and that no station must appear in two different workareas. As a guideline, such multiple workareas can be based on the type of the station; river gauging station or rainfall station or HP full climate station. It is obvious that some of the river gauging stations would have observations on rainfall or even climatic variables and if multiple workareas are maintained at that SDDPC then all such rainfall and climatic data have to be entered with the concerned station which has been created and maintained in the workarea pertaining to river gauging stations.

Workareas for entering and managing historical data must not be combined with the current data (to be considered from 1/1/1999 onwards). For historical data, the organisation of stations in the workarea(s) may be different than as mentioned above, if it is so desired. This might be needed due to historical data being organised district wise and also in view of large number of years of historical data available with most of the stations. In such cases of historical data organisation, stations can be group on the basis of districts and also for the

three types of stations separately as the volume of data would be very high. For the current data, workareas can be envisaged for a period of 5 years so that the same workarea remains operational for a considerable time and the burden of maintaining a huge number of workareas is avoided.

Every workarea or database is a single file containing all types of data, of all the stations within it. These files bear filenames by which the workareas are identified. All these workarea files are to compulsorily have the extension as ".MDB" in the filenames. Though there is no restrictions in naming the workareas but it is good practice to follow certain guidelines. This would bring uniformity in naming the workareas and also be helpful for an easy identification and recognition of the workarea. As a guideline, the name of the workarea may have reference, in the form of acronyms, to (a) the name of DDPC, (b) the name of SDDPC, (c) category of stations and (d) the period for which the data is maintained in that workarea. It would be a good practice to intersperse all these four acronyms with an "underscore" ("_") for making the name a contiguous and unambiguous string of characters. Though this guideline is not a rigid instruction to be always followed but if done in this manner it would make it standardised, uniform and self-explanatory all over.

The acronyms for the DDPCs and the SDDPCs are to be either the names of the offices or the names of the places whichever is more commonly used for other purposes also. In some agencies, the DDPCs and SDDPCs are based on distinct sub-basins/basins and they are named on the basis of the names of these sub-basins/basins and that is why are better recognised by such names as compared to the name of the place. In other cases, the name of the cities/towns can be more common in use for identifying a certain DDPC or SDDPC. The basic idea is to use some acronym which can be easily recognised and related with the DDPC or SDDPC rather than using some abstract numbering as DDPC 1, DDPC 2 or SDDPC 1, SDDPC 2 etc.

For the portion of the name on the category of the station a common guideline of using "GD", "CLIM" and "RAIN" can be followed for the three types viz. river gauging (gauge-discharge), HP full climatic and rainfall stations (both SRGs and ARGs included) respectively.

The period for which the data is maintained in a particular workarea can be simply given by stating the start and end years in the form of two digit numbers in succession. For example, if a certain workarea is to maintain data for a period of hydrological year 1999-2000 to year 2003-2004, it can be simply represented in the name as "9903".

Three examples below illustrate how the data for a Sub-Division in different situations can be maintained.

**Case 1:** Consider that the Lower Godawari Sub-Division No. 2 (Rajamundry) of Lower Godawari Division (Hyderabad) of CWC has 9 river gauging stations in all. Since the number of stations is not very large (even less than 10 as mentioned above) it is appropriate to have only one workarea at the Sub-Division for storing data of all these 9 stations together. If it is thought to have the workarea for a period of five years from 2000-2001 to 2004-2005 then the name of the file can be given as "LGD_RJY_0004.MDB". It is important to note here that for the DDPC and SDDPC "LGD" and "RJY" have been used which implies that DDPC is Lower Godawari Division and SDDPC is at Rajamundry. Both of them are though different in the way that one indicates the name of the DDPC and other the place of the SDDPC but since these are the more commonly understandable names for these DPCs they are used.

**Case 2:** Consider that there are 27 rainfall stations (including both, SRGs & ARGs), 8 HP full climatic stations and 17 river gauging stations under Kolhapur Sub-Division of Pune Division of Maharastra state. Since the number of stations are comparatively more it is good to organise these stations in three different workareas as per the category of stations. Thus

three workareas can be organised for the three types of stations. If is required to organise 5 years of data from 1999-2000 to 2003-2004 then the names of these workareas can be given as: "PUNE_KOLH_RAIN_9903.MDB", "PUNE_KOLH_CLIM_9903.MDB" and "PUNE_KOLH_GD_9903.MDB" respectively.

**Case 3:** Consider that there are 17 rainfall stations (including both, SRGs & ARGs) and 8 river gauging stations under Nellore Sub-Division of Guntur Division of Andhra Pradesh. Since the total number of stations in each category is not very large it is better to organise all the stations in one workarea itself. If 5 years of data from 1999-2000 to 2003-04 is required to be organised then the name of workarea filename can be "GUNT_NELL_9903.MDB".
Normally, all the stations under the SDDPC belong to the same sub-basin of the independent river basin. It is unlikely that a SDDPC's jurisdiction is spread over parts of more than one sub-basin. If in case the jurisdiction is indeed spread over more than one sub-basin and at the DDPC also these sub-basins are to be organised in separate workareas then the stations of each sub-basin are to be kept in separate workareas at the SDDPC, on the lines similar to as mentioned above.

Thus at the SDDPCs there will be ***only one workarea*** if the number of stations are not excessive as discussed above or ***at the most three*** if there are very many stations under a Sub-Division. Whatever is the number of workareas at the SDDPCs, these are to be taken as ***temporary databases*** which contains ***field data***. At the level of SDDPCs there is no requirement to keep the processed databases as no data is expected to be modified at this level. Only the suggested values in case of any inconsistency or gaps in the data, whenever and wherever required, are to be recorded in the slots for remarks or in writing on the reports taken from SWDES and communicated to the DDPCs. This communication of hardcopy is made as soon as the transfer database is sent from SDDPC to the DDPC at monthly interval.

## 1.4 Organising temporary databases at Divisional Data Processing Centres

At the DDPCs, both, the primary and the secondary modules of hydrological data processing system would be operational. The primary module (SWDES) would be used to consolidate the field data arriving from various SDDPCs at regular intervals and transfer data from primary to secondary module. At the DDPCs, it is very essential to have an exact replica of the contents of the databases available at SDDPCs. For this purpose, it is required to have a copy of all workareas which are used at the SDDPCs available at DDPC. The names of these workareas must also be exactly same as used at the SDDPCs except for a suffix "_DDPC" to indicate that it is maintained at DDPC level. When the incremental data from SDDPCs is regularly transferred to the DDPCs, it is consolidated in the respective workarea. This would also ensure a full backup at DDPCs, of the works as carried out at various SDDPCs. Such workareas can keep working for a considerable period (say 5 years).

In case there is only one workarea at each SDDPC then the number of workareas at DDPC would be small and manageable. However, if there are separate workareas for different categories then the number would be larger and less compact. To make the transfer of data from primary to secondary module at the DDPCs with adequate brevity and comfort, it is required to consolidate individual SDDPC workareas into one common workarea for each of the SDDPC. That is, the contents of the multiple workareas, if available at any SDDPC, are to be consolidated into a single workarea for that SDDPC. However, in case an SDDPC is having separate workareas on account of having jurisdiction spread to two or more distinct sub-basins (which is a highly unlikely case) similar distinction can be retained while consolidating into common workareas.

The acronyms for the workareas in primary module at the DDPCs can be on similar lines as those at the SDDPCs except that sometimes there can be separate workareas at the SDDPCs for different types of stations whereas at DDPCs they would all be combined into

one workarea. Thus the acronym for the type of station would not be required at the DDPCs. The name of the workareas for the primary module at the DDPCs would be based on (a) name of DDPC, (b) name of SDDPC, (c) period for which the data is to be maintained at the DDPC and (d) the suffix "_DDPC". Normally, a period of 5 years can be a suitable length of time for which this organisation can be done.

Thus, at every DDPC, there will be exact replicas of all the workareas being used at different SDDPCs (with only the suffix "_DDPC" added to the file names) and if there are multiple workareas in use at one or more SDDPC then one combined workarea for each SDDPC would be required to be established additionally. A full view of each SDDPC's workarea can be obtained from these combined workareas for each SDDPC.

In most cases the jurisdiction of the DDPCs would be within an independent river basin. In only rare case it may cover more than one distinct sub-basins or part thereof of the independent river basin. Organisation of the work and the databases in the secondary module at the DDPCs is to be done such that all the data (rainfall, climate and hydrological) pertaining to each distinct sub-basin or group of sub-basins resides in a separate workarea. That is to say that the tributary to the independent river (i.e. the sub-basin) is to be taken as the smallest unit for organising the database at the DDPCs and thereby all stations within must be together. On the other side, if the DDPC's jurisdiction is extending upto a location on the independent river itself or covers the whole independent river then the workarea is to be for that extent of basin coverage. It is needless to say that, if a DDPC covers more than one independent river basin then the sub-basins with these basins must be organised in different workareas separately.

Thus at DDPCs there would be usually one workarea in the secondary module on the basis of drainage area under its jurisdiction. Obviously, all the data from one or more SDDPCs pertaining to this drainage area (of the independent river) is to be pooled together in this every time when incremental data from SDDPC is available. Only in rare cases, it may have two or more workareas for distinct sub-basins/zones of independent river basin/zone.

The name for the workareas in the secondary module at the DDPCs can be based on the name of the DDPC and on the name of the drainage area covered. Since these workareas in the secondary module are to serve for considerably long period of time there is no specific purpose solved by including the period also in the name of the workarea. However, any extra qualifier required for greater distinction from any other workareas available at the DDPC can always be included. Thus the name of the workareas in the secondary module at the DDPCs would be based on the (a) name of DDPC and (b) name of drainage basin/area.

The examples below illustrate how the data in the secondary module, at the DDPC, could be maintained.

**Case 1:** Consider the case of Pune Division of Maharastra state which has six SDDPCs and there are two distinct drainage areas, Bhima and Krishna upto ???, under its jurisdiction. Since these forms two distinct portions for the independent river Krishna, it is appropriate to organise all the data of the DDPC in two parts: one for Bhima sub-basin and another for sub-basin of Krishna at ???. The names of the workarea and its directory can be given as: "PUNE_BHIMA" and "PUNE_KRISHNA".

**Case 2:** Consider the case of Dowalaiswaram Division of AP state which has lower east Godawari region and a sizable region in northern Andhra Pradesh called "North Circars" draining directly into the sea. At DDPC, workareas can be maintained for these distinct regions. The names of the directory and workareas for these regions at the DDPC can be DOWL_LOWEAST_GODAWARI and DOWRM_NORTHCIRCARS.

## 1.5 Organising temporary databases at State/Regional Data Processing Centres

At the SDPCs/RDPCs, the whole state or a very large drainage region is under the jurisdiction. The state would include parts or full of one or more independent river basins whereas the regions of CWC being based on the river basins would include one or more complete independent river basins. Both primary and full package would be operational at the SDPCs/RDPCs.

The primary module would be employed to consolidate the raw data emanating from all the DDPCs. The DDPCs will have regular consolidation all data of every SDDPC under it in one single workarea. Similar workarea is to be available at the SDPC/RDPC for the purpose of reference and most importantly for transfer of raw data to the Data Storage Centre. This is done by regular consolidation of incremental data sets being sent by the DDPCs to SDPC/RDPC. The names of these workareas have to be exactly similar to what is used at the DDPCs with the exception that instead of "_DDPC" as the suffix "_SDPC/_RDPC" is to be used.

The full package at the SDPCs/RDPCs would be employed to aim at hydrological data validation and reporting for the complete river basin(s) or part thereof within the state. The SDPCs will also require to pool data from all the river basin(s) within the state boundary together to get an overall view for the whole state or smaller administrative units, especially for rainfall information. The contents of such unified workarea may not include all the data to the maximum available details but will largely have finalised summary data on daily/monthly/yearly levels for various variables and as per the requirement of the SDPCs.

Thus, for the organisation of data in the full package at the SDPCs/RDPCs, two type of workareas have to be maintained. The first type of workareas will be for the individual independent river basins or part thereof within the state. Thus as many workareas would have to be established as the number of independent river basins within the SDPC/RDPC's jurisdiction. In special circumstances, a group of very small independent rivers (as will be the case with the coastal rivers which directly drain into the sea) may be clubbed together in one workarea. The second type of workarea is required at the SDPCs to pool up all the data for the whole state together.

The names of the workareas of the first type have to be based on the name of the independent river basin or the group of smaller basins taken together with the indication of name of the concerned state, if it is including only a part of it within its boundary. Thus the name for the workarea would comprise of acronyms for (a) name of the independent river basin or group of smaller basins and (b) name of the state. The name of the workarea of second type is obviously to be based on name of the state itself since this would contain data for the whole of the state.

## 1.6 Organising temporary databases at the National Data Processing Centres

At the national level the Central Water Commission will have a National Data Processing Centre (NDPC) together with a National Data Storage Centre (NDSC). At this NDPC, the required data of all the CWC observation stations for different river basins/zones of the HP area will be organised. Apart from the CWC observation stations, data of the selected observation stations of various states could also be available at the NDPC depending on the requirement and protocol between various states and the CWC.

The NDPC will require to organise the authenticated data from the respective RDPCs . The contents of the workareas at the NDPC may not include all the data to the maximum

available details but will largely have finalised summary data on daily/monthly/yearly levels for various variables. Such an organisation is essential at the national level for providing an integrated view on a macro level of the peninsular region as a whole. Also, NDPC will be required cater to any requirement which may come to it for providing hydrological information on any river basin in the peninsular region of the country.

The name of the workareas for various river basins at the NDPC is to be exactly similar to those used at the RDPCs except for the fact that a suffix "_NDPC" can be put at the end so as to make a distinction from the similar workareas operative at the RDPCs.