APPENDIX I

ANN TRAINING DATASET DESCRIPTION

EXPLANATORY VARIABLES

The ANN explanatory variables described in this appendix were designed to represent the system state and provide the ANN with pertinent information about the current state and magnitude of stresses change with respect to previous states. Explanatory variables are totalized by grouping areas and depending on its nature further summarized by area-buffer. In the following reference, the internal *River GeoDSS* keyword for the variables is supplied in parenthesis for each explanatory variable. Keywords were used in the training dataset, MATLAB training and testing tools, as well as in the *River GeoDSS* ANN module to consistently keep track of the variables. The examples provided in this section relate to the ANN-based stream-aquifer interaction modeling in the LARV. The results of the explanatory variables processing is stored in the *buffers* database, which is a MS-Access database that combines geo-referenced layers and data tables to store the processed explanatory variables and processing support information.

Grouping Area-Based

These explanatory variables are extracted per grouping area. They provide information related to the main stream segment and the overall modeling conditions.

Main Stream Length (StreamLengthArk)

This variable contains the grouping area main stream length in kilometers. For most of the modeling areas in the Arkansas Valley, it is 15 km except the areas intercepting John Martin Reservoir where the main stream is submerged and the first one that extends longer to completely cover the irrigated valley (Figure I-1).

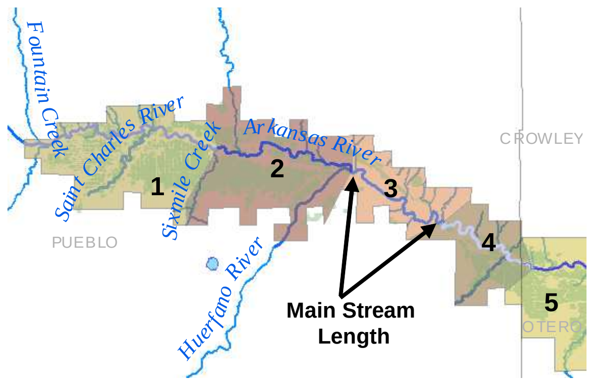

Figure I-1 – Main stream length example in upstream grouping areas

Tributary Stream Length (StreamLengthTrib)

This variable indicates the total length of tributary lines in the grouping area in kilometers.

Average River Elevation (AveRiverElev)

The groundwater flow analysis (Chapter 4 – *Spatial Variable Grouping*) exhibits the importance of the relationship between the elevation of water bodies, canals and the river elevation in the magnitude and direction of the flows. This explanatory variable measures the average elevation in the main river segment in the grouping area. This variable is calculated using the GIS Zonal Statistics function for the main stream line on the USGS Digital Elevation Model (DEM).

River Flow Indicator (RiverFlow)

The average flow in the river and tributaries links that are flagged for ANN return is computed as indicator of flow state in the grouping area. This value is computed with the assistance of the table *RiverLinksBufferID* (located in the data-model) that contains the links and corresponding grouping area's IDs with a place holder for a single flow value that is updated at run time for each time step. The average calculation includes flow of both the links completely contained inside the grouping area and links intersecting the boundaries of the grouping area.

Pumping Increase Indicator (PercPumped)

This explanatory variable reflects the percent of pumping increased, which is used only in vertical drainage alternatives.

Subsurface Drainage Indicator (DrainSpc)

An indicator is included to provide the ANN with information about the spacing density in sub-surface improvement management alternatives. The drain spacing value, used to identify the management scenario, is entered in the ANN training dataset to capture the effect of drainage improvement in the stream-aquifer interaction modeling. The variable doesn't take into account the percent of the irrigated fields that are improved.

Area-buffer-Based

The following explanatory variables are extracted per area-buffer; an explanatory variable is created in the training dataset for each area-buffer

Area-Buffer Area (BufArea_)

This variable consists of the total area of the area-buffer in square kilometers (km^2). The variable includes all polygons in the area-buffer even if they are not physically connected (see Figure I-2).

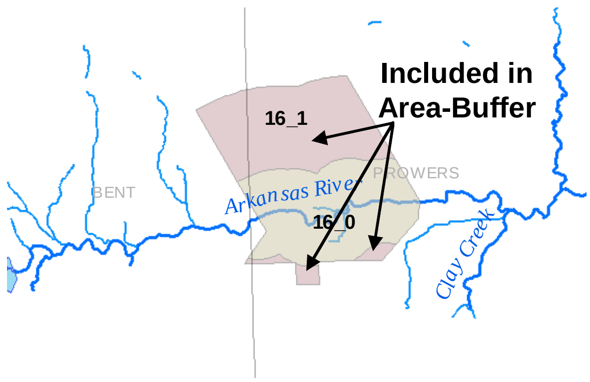

Figure I-2 – Area-buffers configuration in grouping area 16

Average Terrain Elevation (AveElev_)

The average elevation of the area-buffers terrain is used to calculate the ANN explanatory variable. The average elevation is computed using GIS-Zonal Statistics applied on the DEM for the area-buffer polygons. The processed averaged elevations are stored in the database table *AverageElev*. The explanatory variable *AveElev_b* is calculated, for area-buffer *b*, as the difference between the average elevation in the area-buffer (E_b) and the average elevation in the main river[b] (RE_i) for grouping area *i* normalized by the river elevation.

$$AveElev\_b = \frac{E_b - RE_i}{RE_i}$$

[b] Explanatory Variable *AveRiverElev* for area-buffer *b*.

Water Bodies Areas (WBArea_)

This explanatory variable is created to represent the extension of the surface water accumulation in the area-buffer. GIS geo-processing tools are used to intersect the polygons representing water bodies and area-buffers. The total area of polygons (or part of polygons) inside the area-buffers are summarized as the explanatory variable. The value itself provides an idea of the extent of the water bodies and the grouping in area-buffers provides insight on the location of them respect the river system. The results are compiled in the *BWBodies* prefixed tables in the *buffers database*.

Water Bodies Elevation (WBElev_)

The elevation of the water bodies with respect to the mean sea level is calculated based on the DEM using the GIS-Zonal Statistics function. The average values are summarized in the database table *BWBodiesElev.* This explanatory variable captures the average elevation of the water bodies in the area-buffer with respect to the river elevation. The explanatory variable is computed as the difference between the average water bodies elevation (WBE_b) in the area-buffer *b* minus the average elevation of the river (RE_i) in grouping area *i* divided by the average elevation in the river (RE_i).

$$WBElev\_b = \frac{WBE_b - RE_i}{RE_i} \quad (\text{I.1})$$

Diversion (AveDiversion_)

An explanatory variable is created to provide information of the magnitude of the diversion water applied to the irrigated fields in the area-buffers. The diversion per area-buffer (D_b) is calculated as the combination of the fractions of the diversions proportional to their area-buffer irrigated area with respect their total irrigated area.

$$D_b = \sum_{C_b} D_c \, {A_{b,c}}/{A_c} \quad (I.2)$$

where, D_c = the diversion of canal c. C_b = the set of canals that irrigate fields in buffer b. $A_{b,c}$ = the area irrigated by canal c in buffer b, and A_c = the total area irrigated by canal c.

Precipitation (Precip_)

Spatially-varied precipitation is processed at weekly time steps for the *River GeoDSS* use. The spatially-varied (raster) maps can be generated either from NEXRAD processed rainfall or using spatial interpolation from the point-measurement data (CoAgMet and NWS stations). The *River GeoDSS* implements a tool using GIS functions that allows processing point-measured variables to generate climate variable raster maps using the inverse distance method. The weekly precipitation is stored in a database that is accessed for both training and simulation. GIS Spatial Analyst functions are used to create weekly summary tables with the mean precipitation per area-buffer. The table name uses a user-defined prefix and the beginning of time step date (*ddmmYY*). Summary tables are reliant on area-buffers and time step, alteration in any of these elements requires re-generation of the precipitation summary tables. For example, if ANN training time steps are shifted from the simulation time steps, different tables will need to be generated for training and simulation.

Canals (Canals_)

The length of canals in the area-buffer will bring in an indicator of the amount of potential seepage interfaces and their location with respect to the main stream. This explanatory variable is computed clipping the canal lines by area-buffers and totalizing all the lengths.

Canal Elevation (BCanalsElev_)

In the groundwater regional-scale modeling (Burkhalter and Gates 2005; Burkhalter and Gates 2006), the relationship between the canals stage and water table depth shows a significant influence in the modeling results. At basin scale, we can approximate the canal stage using the DEM-derived canal elevation respect to the mean sea level. This explanatory variable captures the relationship between the average elevation of all canal lines in the area-buffer with respect to the average elevation in the river. The variable is computed using GIS-Spatial Analyst tools and stored in the table *BCanalsElev*. The explanatory variable is computed as the difference between the average canals elevation (CE_b) in the area-buffer b minus the average elevation in the river (RE_i) in grouping area i divided by the average river elevation (RE_i).

$$BCanalElev\_b = \frac{CE_b - RE_i}{RE_i} \tag{I.3}$$

Irrigated Area (IrrgArea_)

This explanatory variable captures the amount of irrigated area in the area-buffer in squared kilometers. These values are calculated using the Colorado Division of Water Resources potential irrigated fields map and GIS Spatial Analyst tools.

Canal Seepage Indicator (AveVolSeep_)

An explanatory variable is computed to provide an estimate of the amount seeped in the area-buffers to the groundwater. The seepage volume calculation is reliant on a user defined coefficient ($\bar{s}_c$) that specifies, for canal c, the fraction of the total diversion (D_c) that will flow to the groundwater along the run of the canal (from the head gate to the end of the canal). The total seepage (S_c), in canal *c*, is computed using the diversion flow (D_c) and the canal seepage coefficient.

$$S_c = D_c \cdot \bar{s}_c \tag{I.4}$$

The estimated seeped amount in the area-buffer is calculated as a fraction of the total seepage of the canal. The area-buffer seepage ($S_{c,b}$), for canal *c*, is computed as the total seepage per unit length times the length of the canal intersecting the area-buffer.

$$S_{c,b} = D_c \cdot \bar{s}_c \frac{L_{c,b}}{L_c} \tag{I.5}$$

where, L_c = the total length of canal *c* and $L_{c,b}$ = the length of the canal *c* that lays in area-buffer *b*. The length of each canal per area-buffer is calculated during the data preparation summing the canal line lengths in the area-buffer and storing the results grouped by the corresponding canal *HydroID* and the *bufferID*.

The explanatory variable (*AveVolSeep_*) for each area-buffer, it is calculated as the average seeped volume of the canals in the buffer. The average seeped volume (S_b) in the buffer *b* is calculated as:

$$S_b = \frac{\sum_{i}^{C_b} S_{i,b}}{C_b} \tag{I.6}$$

where, C_b = number of canals present in the area-buffer *b*, $S_{i,b}$ = the seepage of the canal command area *i* in the buffer *b*.

This indicator is factored by the percent reduction when simulating management alternatives in *River GeoDSS*. The seepage reduction fraction for canal c ($R_{S\,c}$) is incorporated in the explanatory variable as follow:

$$S_b = \frac{\sum_{C_b}\left[\frac{D_c \cdot \bar{s}_c \cdot L_{c,b}}{L_c} \cdot \left(1 - R_{S\,c}\right)\right]}{C_b} \qquad (\text{I.7})$$

Canal Seepage Reduction (PercSeep_)

An indicator for the canal seepage reduction fraction in the area-buffer is incorporated into the explanatory variables. The seepage reduction fraction corresponds to the fraction of the baseline seepage that was reduced by improvements in the canal conveyance efficiency. The area-buffer indicator is computed as the average of the fractions for the set of canals intersecting the area-buffer (C_b). The indicator for area-buffer b can be calculated as:

$$PercSeep\_b = \frac{\sum_{i=1}^{C_b} R_{S_i}}{C_b} \qquad (\text{I.8})$$

where, $R_{S\,i}$ = Seepage reduction fraction for canal i.

Aquifer Recharge (Rech_)

An explanatory variable is introduced to provide the ANN with information about the amount of water that is deep-percolated to the groundwater from irrigation practices. The aquifer recharge is approximated as a fraction of the water available to the fields. Water available for irrigation (F_c) for canal command area c, is approximated as the diverted water minus the seepage losses. Assuming an average leaching fraction (lf), the aquifer recharge volume (R_c) from canal c is calculated as:

$$R_c = F_c(lf) = (D_c - D_c \cdot \bar{s}_c)lf \qquad \text{(I.9)}$$

where, D_c = the diversion of canal *c*, $\bar{s}_c$ = the average seepage coefficient for canal *c*.

The recharge indicator for area-buffer *b* is calculated using the proportional aquifer recharge amount for all canals that irrigate the buffer. The average recharge for area-buffer *b* ($\bar{R}_b$) is calculated as:

$$\bar{R}_b = \frac{\sum_{i=1}^{C_b} \frac{lf(D_i - D_i \cdot \bar{s}_i)A_{i,b}}{A_i}}{C_b} \qquad \text{(I.10)}$$

where, D_i = the canal *i* diversion in acre-ft/week, C_b = the number of canals that irrigate fields in area-buffer *b* and $A_{i,b}$ = the area irrigated by canal *i* in buffer *b*. A_c = the total area irrigated by canal *i*.

During the management scenarios simulation, the aquifer recharge reduction is implemented using the recharge reduction factor (R_R). The recharge indicator is calculated as:

$$\bar{R}_b = \frac{\sum_{i=1}^{C_b} \frac{lf(D_i - D_i \cdot \bar{s}_i)A_{i,b}}{A_i}(1 - R_R)}{C_b} \qquad \text{(I.11)}$$

Notice that the recharge indicator becomes Equation I.10 when the recharge reduction factor R_R equals to zero (e.g., baseline and management alternatives that don't include recharge reduction).

Recharge reduction is achieved by improvement of application efficiency at the field operations, resulting in a new leaching fraction (lf'). The recharge indicator can be expressed as a function of the scenario current diversion and improved leaching fraction as:

$$\overline{R}_b = \frac{\sum_{i=1}^{C_b} \frac{lf'\left(D_i' - D_i' \cdot \overline{s}_i'\right)A_{i,b}}{A_i}}{C_b} \quad \text{(I.12)}$$

Where D_i'= Diversion in the simulated recharge reduction scenario for canal i, and $\overline{s}_i'$ = average adjusted canal seepage coefficient to produce the management scenario seepage. The adjusted leaching fraction can be calculated assuming that the scenario aquifer recharge (R') equals the reduced baseline recharge ($R(1 - R_R)$).

$$F' \cdot lf' = F \cdot lf\left(1 - R_R\right) \quad \text{(I.13)}$$

$$lf' = \frac{F}{F'} lf \cdot \left(1 - R_R\right) = \frac{F}{F - F \cdot lf \cdot R_R} lf \cdot \left(1 - R_R\right) = \frac{lf \cdot \left(1 - R_R\right)}{\left(1 - lf \cdot R_R\right)} \quad \text{(I.14)}$$

Aquifer Recharge Indicator (PercRech_)

An indicator for the recharge reduction factor in the area-buffer is used as explanatory variable. The indicator is the average of the individual recharge reduction fraction (R_R) overall canals that irrigate fields in the area-buffer.

$$Perc\,\mathrm{Re}\,ch\_b = \frac{\sum_{i=1}^{C_b} R_{R_i}}{C_b} \quad \text{(I.15)}$$

Where $R_{R\,i}$ = Recharge reduction fraction for canal i. C_b = the number of canals that irrigate area-buffer b.

Groundwater Pumping (AvePumped_ and NoPumps_)

Historical pumping data in the area-buffers is used as explanatory variable for the ANN stream-aquifer interaction modeling. Two explanatory variables are created to indicate the magnitude and density of the pumping in the area-buffers. The first one (*AvePumped_b*) is the total volume pumped from the aquifer and the second one (*NoPumps*_b) is the number of active pumps in the area-buffer *b*.

The pumped volume is increased in by the factor indicated by the simulated management scenario. The vertical drainage alternative pumped water (*P'*) is calculated as:

$$P' = P \cdot \left(1 + I_p\right) \qquad \text{(I.16)}$$

where, P = historical pumped water, and I_p = the factor in which the pumping is increased.

The additional pumping (ΔP) can be calculated from the historical records of the pumps located in the grouping areas using the I_p factor or it can be extracted from the total scenario pumping (in MODFLOW output) using the following relationship:

$$P' = P + \Delta P = P\left(1 + I_p\right) \qquad \text{(I.17)}$$

where P' = management scenario pumping and P = baseline pumping. Rearranging the equation terms ΔP is computed as:

$$\left(P' - \Delta P\right)\left(1 + I_p\right) = P' \qquad \text{(I.18)}$$

$$\Delta P = P' \frac{I_P}{1 + I_P} = P'\left(1 - \frac{1}{1 + I_P}\right) \qquad \text{(I.19)}$$

MODFLOW-MT3DMS VARIABLES FOR ANN TRAINING

The following variables are processed and included in the ANN training dataset, for the baseline and management alternatives, using Geo-MODFLOW with the user specified binary MODFLOW-MT3DMS output files. The Geo-MODFLOW queries the individual cell outputs. The geo-processed MODFLOW-MT3DMS grid cells are clipped to the grouping areas. All cells outputs are adjusted proportional to the area inside the grouping area. A proportionality factor, defined as the ratio between the cell area inside the grouping area and total area of the cell, is multiplied times the cell output to compute the corresponding grouping area portion.

MODFLOW-MT3DMS River Cells Grouped Variables

The following output variables of the training datasets are populated for both the Arkansas River (Main River) and the tributaries using the cells the MODFLOW-MT3DMS flagged river cells. These variables summarized per grouping area.

Main River Aquifer Return (OUTPUTInRiverArk)

This variable includes the total water volume that flows from the aquifer to the river during each time step in the grouping area.

Tributary Aquifer Return (OUTPUTInRiverTrib)

This variable equals the total water volume that flows from the aquifer to the tributary streams during a time step in the grouping area.

Main Stream depletion (OUTPUTOutRiverArk)

This variable contains the sum of the water depleted from the main river into the aquifer during a time step in the grouping area.

Tributary depletion (OUTPUTOutRiverTrib)

This variable contains the sum of the water depleted from the tributaries into the aquifer during a time step in the grouping area

Main Stream/Aquifer Salt Load (OUTPUTQualityArk/OUTPUTQualityTrib)

These variables summarized the total salt load to the main river and tributaries from the aquifer in the grouping area.

Main Stream/Aquifer Salt Load (OUTPUTConcArk/OUTPUTConcTrib)

These variables contained the computed concentration of water returned to the main river and tributaries respectively. It is calculated as the ratio between mass returned and volume returned.

All MODFLOW-MT3DMS Cells Grouped Variables

Using all the cells in the grouping area, the following variables are computed for the ANN training dataset:

Aquifer Recharge (AquifRech)

This variable contains the sum of the total aquifer recharge amount for all the cells and all the MODFLOW-MT3DMS layers in the area-buffers. An output variable is created per area-buffer in the grouping area.

Simulated Drainage Volume (DrainReturn)

This variable summarizes the total drained water volume in the grouping area.

Simulated Drained Salt Load (DrainSaltLoad)

This variable contains the salt load contributed by the drained water that flows to the surface system.

MODFLOW-MT3DMS Pumped Water (OUTPUTPumped)

This variable summarized the MODFLOW-MT3DMS groundwater pumped in a time step from the grouping area.

Pumped Salt Load (OUTPUTPumpedSalt)

This variable contains the total salt load from the pumping activity in MODFLOW-MT3DMS.

Canal Seepage (CanalSeepage_)

For each area-buffer in the grouping areas the total seepage is summarized to the training dataset.

ANN TRAINING DATABASE

The tool produces a database containing the ANN training dataset. The database includes three tables:

General Information Table (ANN_InputOutput_Original_GeneralInf)

This table contains general information about the source tables, preferences, variables generated description and date of execution.

Data Sources Information Table (ANN_InputOutput_Original_DataSources)

This table holds information about the scenarios used to generate the dataset. It contains the scenarios number and names keys (unique identifiers) as well as the MODFLOW file name used to generate the cases for each management alternative.

Data Sources Information Table (ANN_InputOutput_Original)

This table contains the sets of inputs/output for ANN training and simulation. The training cases will have sets of inputs and corresponding outputs. The simulation cases correspond to pre-calculated explanatory variables that will be available basin-wide at simulation time. The input/output cases (table rows) have information about the grouping area, the scenario where it came from as well as the time step for which it was generated.

ANN TRAINING DATASET FILES DESCRIPTION

The *ANN Database Management* utility exports the dataset to a set of comma separated files (*.CSV) files for training, testing and simulation. The files are created in the folder specified in the *Output Folder* box using the prefix name given by the user in the *File Base Name* (Figure 4.12). The files created are:

ANN_to_MODSIM_DataSources.csv contains the list of scenario number id and name.

ANN_To_MODSIM_CONST.csv holds a table with values for the explanatory variables that are constant for all the modeled time steps. It is structured with the explanatory variables in the columns and the grouping areas in the rows.

ANN_To_MODSIM_AvgOutputs.csv provides the average values for the output variables to be predicted.

ANN_To_MODSIM_AvgInputs.csv contains a summary table with average values for recurrent explanatory variables.

ANN_OutputVarLabels.csv provides the labels and units information for the output variables. These values are user-defined in the *ANN Database Management* utility support tables.

ANN_To_MODSIM.csv contains the generated dataset including all the explanatory variables (and available recurrent variables) for all the scenarios. This table is used for *River GeoDSS* testing purposes.

ANNTrain_X.csv contains the training dataset for each group *X* defined in Figure 4.11.

ANNTest_X.csv contains the testing dataset for each group *X* defined in Figure 4.11. It includes all available time steps (some of the initial time steps are excluded from training for not having recurrent variables) and it might not include the training dataset filters, if selected by the user in the interface. This dataset might also include training cases since it will provide an overall performance measurement.

ANNIOInfo.csv contains information for MATLAB ANN training. This file includes the number of previous time steps included in the dataset, the number of area-buffers, the training and testing grouping areas, the name of explanatory variables, name of recurrent variables and output variables.