CHAPTER 4

ANN Assisted Water Resources Modeling

Stream-Aquifer Interaction Modeling

Stream-aquifer interaction is the fundamental component to address river basin conjunctive use modeling, for rivers that are hydraulically connected to a phreatic aquifer. Simplified stream-aquifer response models are generally incorporated into river basin models [e.g., Dai and Labadie 2001 and Kansas Hydrologic-Institutional Model (Burkhalter 1997)], but they fail to adequately capture the complex dynamic and spatial characteristics of the system response (Fredericks et al. 1998). An alternative methodology for modeling stream-aquifer interactions at the river basin scale is presented that integrates artificial neural networks (ANN), geographical information system (GIS) and regional-scale MODFLOW-MT3DMS groundwater modeling. The methodology is based on the development of dynamic, spatially dependent relationships between basin scale measurable system characteristics, which directly or indirectly trigger aquifer stresses, and the stream-aquifer interaction. GIS provides the framework for managing and preprocessing the extensive spatio-temporal database required for building the dataset used to explore and derive the relationships. An ANN is used to extract the relationships between the explanatory variables and the stream-aquifer interaction, with MATLAB™ (MathWorks, Inc) providing the foundation for training, validating and analyzing the ANN. A calibrated regional-scale finite difference groundwater model (MODFLOW) and the associated water quality model (MT3DMS) are used to represent stream-aquifer interaction in both historical system operations and simulated management alternatives.

The methodology is designed and implemented to be coupled with the river basin network flow model Geo-MODSIM. The coupling is rooted in the linkage of geo-referenced system features and characteristics to geo-referenced river basin network flow model objects. Complex stream-aquifer interactions are embodied in the trained ANN, which is embedded in Geo-MODSIM for providing accurate water quantity and quality conjunctive surface and groundwater modeling. The methodology presented herein is aimed to the construction of a robust and realistic river basin scale decision making support tool. The Lower Arkansas Valley in Colorado, from Pueblo Reservoir to the Colorado-Kansas State Line, is used to illustrate the method and demonstrate the potential of its application.

Regional Scale Groundwater Modeling

The Department of Civil and environmental Engineering at Colorado State University has been conducting research in the Lower Arkansas River, defined from Pueblo Reservoir to the Colorado-Kansas State Line, for many years (Gates et al. 2006). Figure 4.1 shows the location of the current (upstream) and in-progress groundwater modeled areas, where the upstream modeled area is approximately one-fourth of the Lower Arkansas River Valley.

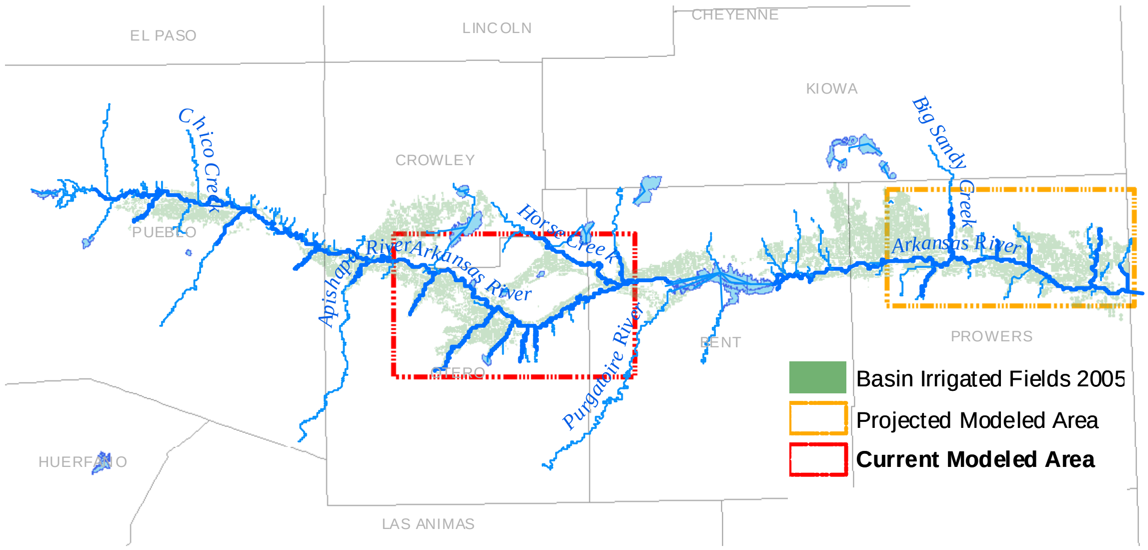


Figure 4.1– Regional-scale groundwater modeled areas

Management alternatives were engineered and modeled at the regional-scale to mitigate salinity and waterlogging problems and sustain agriculture in the area (Burkhalter 2005). Management alternatives included combinations of: aquifer recharge reduction, canal seepage reduction, groundwater pumping increase, and sub-surface drainage improvements. Burkhalter (2005) modeled the baseline and 38 alternatives including: (1) nine scenarios of aquifer recharge reduction from 10% to 90%, (2) seepage reduction levels of 50%, 70% and 90%; (3) 90% seepage reduction over 20% of the irrigation canals length, (4) canal lining scenarios of 90% seepage reduction in Holbrook, Ft. Lyon, Rocky Ford, Catlin, Otero and Highline canals, (5) drainage improvement scenarios with tile drains installed in selected fields at spacing of 50m, 75m, 100m, and 150m, (6) groundwater pumping increases of 25%, 50%, 100%, and 200%, (7) combinations of recharge reduction and seepage reduction of 30%-50%, 50%-90%, and 80-90% respectively, (8) recharge reduction and drainage improvement of 30%-100m, 50%-50m, and 80%-50m respectively, (9) combination of seepage reduction and drainage improvements of 50%-100m and 90%-

50m respectively, and (10) combination of recharge reduction, seepage reduction and drainage improvement of 30%-50%-100m, 50%-90%-50m, and 80%-90-50m. This transient model is an important resource for understanding stream-aquifer interaction and evaluating the aquifer response to salinity remediation strategies in the basin (Burkhalter and Gates 2005; Burkhalter and Gates 2006).

Approach

The methodology developed herein learns from the detailed regional groundwater modeling effort to model the stream-aquifer interaction at a larger-scale (i.e., the basin scale). The ANNs have been demonstrated as a powerful tool to describe complex relationships between sets of explanatory variables and observed data (Rogers 1992; Maskey et al. 2000; Govindaraju and Ramachandra 2000). In particular, Bowers and Shedrow (2000), Sandhu et al. (1999), and Parkin et al. (2007) have successfully applied ANNs to predict combined effects of the river-aquifer system. Detailed modeling of groundwater basins requires an extensive set of system characteristics that are pertinent to the specific modeled area. Many of these characteristics are not available at basin scale, or in adequate resolution to be reliable. Therefore, traditional groundwater modeling parameters cannot be used for the ultimate goal of basin scale modeling. However, there are events that indirectly influence stream-aquifer interaction such as precipitation, canal diversion, river flow, etc. Combining the occurrence of these events together provides are indicators of not only the temporal, but the spatial system states. A methodology is needed to define explanatory variables that adequately represent system state changes to provide guidance in predicting stream-aquifer interaction. The selected explanatory variables must be available (i.e., measurable) at basin scale. The innovative methodology proposed herein predicts stream-aquifer interactions using an ANN which is trained with current and previous basin-wide-measurable system states. The ANN is trained using (1) a detailed well-calibrated quantity and quality groundwater model representing the regional response of the aquifer (i.e., river return flows, river depletions and salt loads to the river) and (2) basin-wide quantifiable system state variables. The ANN training ("learning") process develops dynamic relationships between the inputs and outputs as embodied in the training data set that captures the complex nonlinear spatially distributed stream-aquifer response to system stresses. The relationships learned by the ANN can be used to prescribe the stream-aquifer interaction in areas where detailed groundwater modeling is not available. GIS is used in building the ANN training and testing datasets as spatially grouped by area-buffers. The ANN output variables are queried using the *River GeoDSS* Geo-MODFLOW extension.

The developed ANN can be spatially linked with the surface basin scale model (Geo-MODSIM) for efficient and practical conjunctive use modeling. In addition, embedding the ANN within river basin decision support tools eliminates the computational burden of directly incorporating realistic numerical finite difference models.

ANN Development

Spatial Variable Grouping

Aquifer responses can vary significantly, even within proximate locations or contiguous cells. When stream-aquifer interaction is analyzed over larger areas, some of the local variability is smoothed, resulting in increased predictability. Return flow volumes and concentrations are modeled in aggregated areas that extend to the adjacent alluvial irrigated valley around the main stream (i.e., the Arkansas River). These areas groupings are created from 15-km river segments in an attempt to maintain comparable predictions per unit length. The boundary of each grouping area follows, for the most part, the sub-watershed boundary created from the most downstream point of the river segment. Figure 4.2 shows the stream-aquifer interface modeling grouping areas in the Lower Arkansas Valley. The grouping areas are sequentially numbered from upstream (east of Pueblo Reservoir) to downstream (Colorado-Kansas State Line). ArcHydro tools were used to delineate the sub-watersheds to guide the grouping areas boundary definition (see Appendix III – *Stream-Aquifer Interaction Grouping Areas* for details).

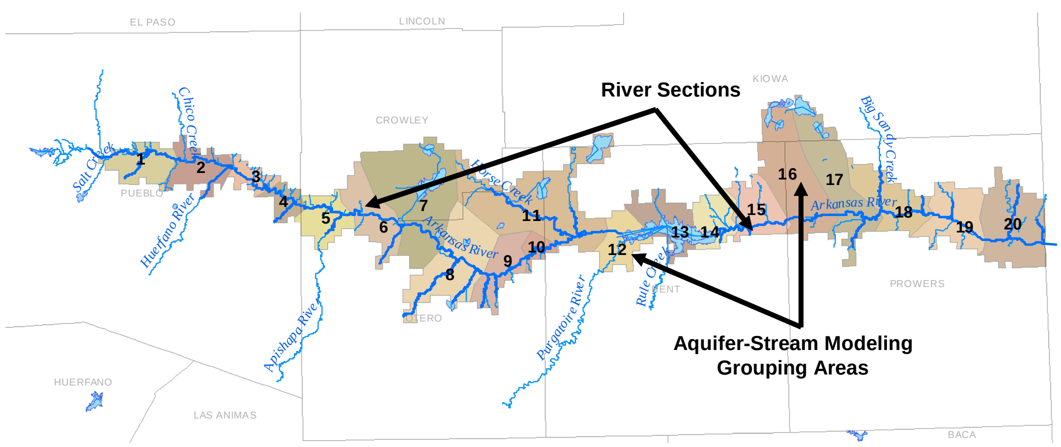


Figure 4.2 – Stream-Aquifer Modeling Grouping Areas in the Lower Arkansas River Valley

Since the aquifer responds to stresses as a function of distance from the stress, it is expected that more proximate stresses have more impact on the aquifer response. The Arkansas River is expected to be the most influential system element for the aquifer in the grouping area. Figure 4.3 shows an example of flow directions in MODFLOW groundwater modeled cells in the Arkansas Valley regional-scale model. In this example, the influence of the main stream (Arkansas River cells) on groundwater flow direction in the vicinity of the river (including areas close to the tributaries where arrows are parallel to the tributaries) is clearly evident. In contrast, for cells father from the river, the direction of the flow is toward the tributaries. The canal lines (solid light blue) seem to have less influence on the direction of the groundwater flow, although canal stage will likely have a significant impact on the magnitude of flows in/out of the aquifer.

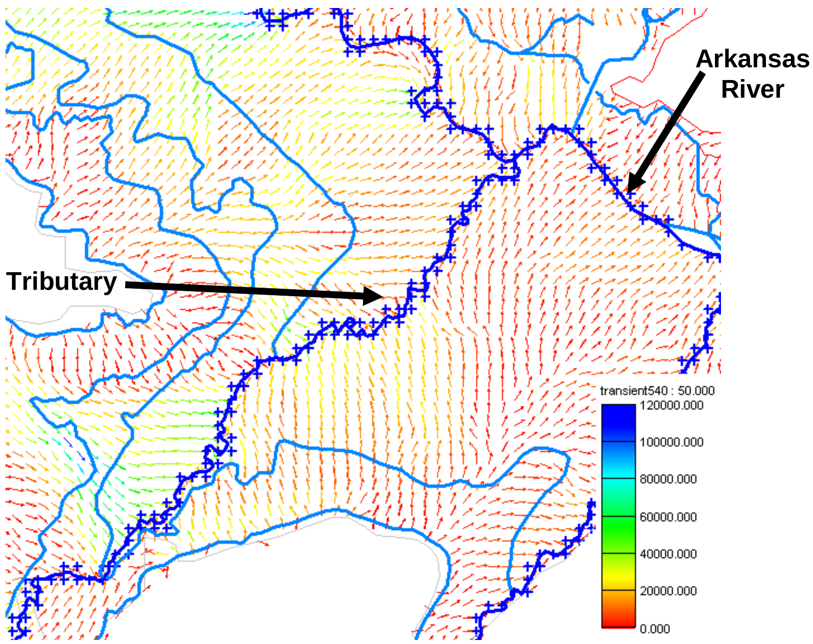


Figure 4.3 – MODFLOW Groundwater flow direction example

Surface water bodies in the system also play an important role in the groundwater flow direction. Figure 4.4 shows another example of the direction of the groundwater model cells, where it is evident that surface water interaction with the aquifer influences the groundwater flow direction.

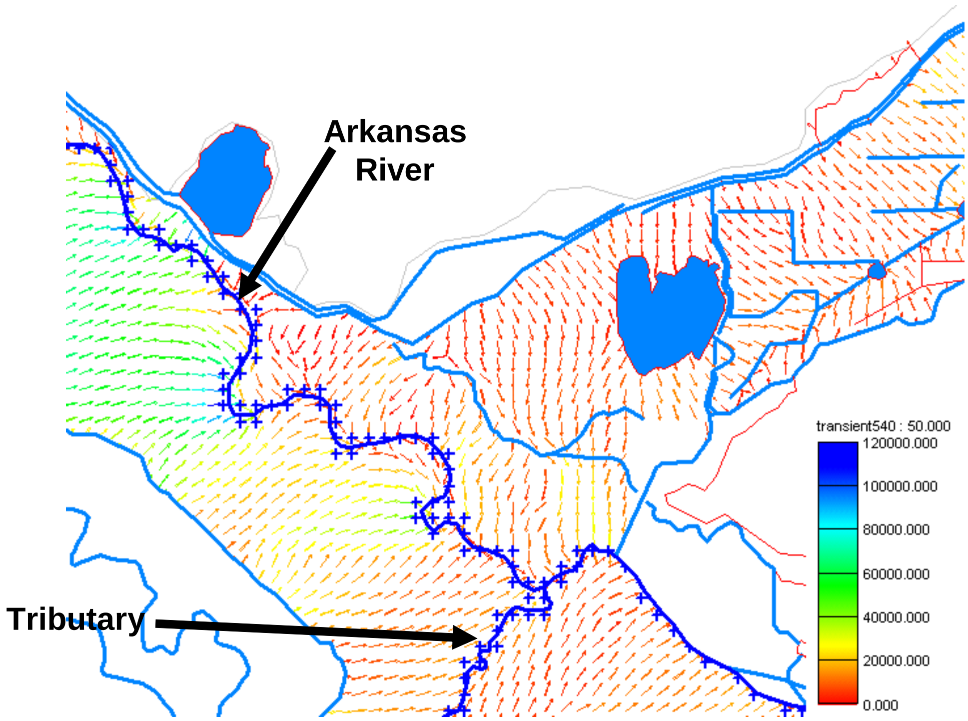


Figure 4.4 – Groundwater flow direction example

From the previous analysis, it can be inferred that stresses located in close vicinity to the main stream will have the most significant effects on stream-aquifer interaction. In the case of tributary stream-aquifer interface modeling, stresses farther away from the main stream will play a more important role than in the main stream-aquifer interaction modeling. The spatial relevance of system state changes is incorporated by grouping variables according to the locations where they occur. For this purpose, area-buffers surrounding the stream are created using incremental buffer zones along the main stream segment. The first area-buffer extends 3 km east and west of the stream, while the second area-buffer extends 6 km east and west of the first area-buffer external boundary. Figure 4.5 illustrate the area-buffers for the grouping areas in the groundwater modeled region. The stream-aquifer interface for the main river is located in the first buffer, whereas tributary stream-aquifer interactions can occur in any of the area-buffers.

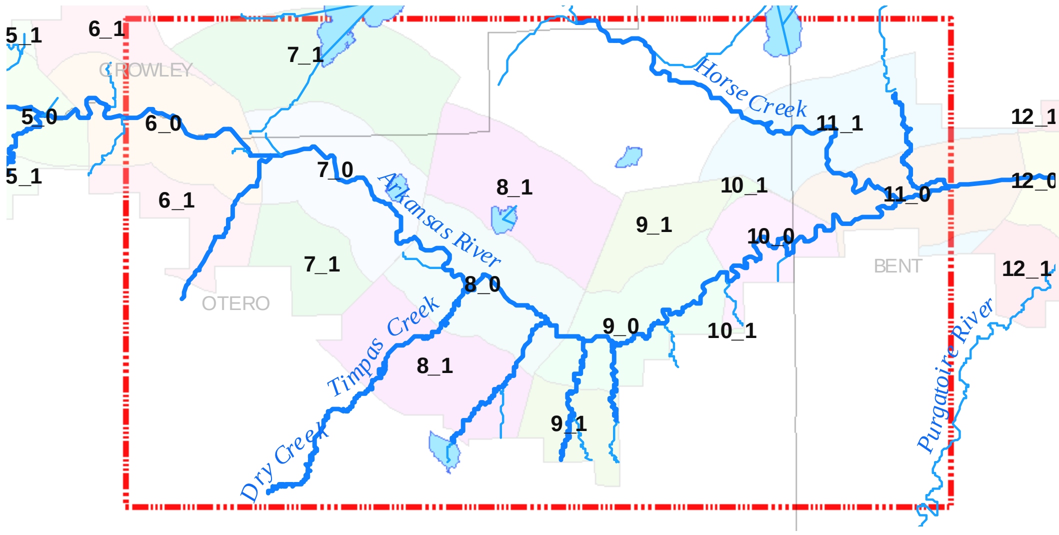


Figure 4.5 – Buffer-Areas for variable aggregation inside the stream-aquifer modeling grouping areas

Explanatory Variables

The explanatory variables are designed to capture the system state and provide the ANN with pertinent information about the current state and magnitude of change with respect to previous states in order to predict aquifer return flows and salt loadings to the surface water system.

Explanatory variables are aggregated by grouping areas and depending on the nature of the explanatory variable further sub-divided by area-buffer. Explanatory variables aggregated per grouping area include (1) average main stream elevation from sea level, (2) stream length in the grouping area, (3) tributary stream lengths in the grouping area, and (4) the average system stream flow in the grouping area. The explanatory variables summarized per area-buffer in the grouping areas include (1) average terrain elevation with respect to the main stream elevation, (2) lengths of the canals, (3) canal average elevations with respect to the average main stream elevation, (4) areas of water bodies in the area buffer, (5) average elevation of the water bodies with respect to the main stream elevation, (6) extent of irrigated fields, (7) average diversion per irrigated area for fields in the area-buffer, (8) the number of active pumping wells, (9) total groundwater pumped volume, (10) total precipitation over the area-buffer, (11) average canal seepage in the area-buffer, and (12) average aquifer recharge in the area-buffer, as computed for a canal irrigating land in the area-buffer as a fraction of water available to the area-buffer fields. Intensity indicator variables for potential changes in the modeling of various management scenarios are included as explanatory variables. Variables indicating percentage of increase pumping from the baseline and drainage intensity are computed for each grouping-area. Variables reflecting seepage reduction from the baseline and percentage of recharge reduced from the baseline are calculated for each area-buffer as a function of the corresponding overlying canals and irrigated fields in the area-buffers. Appendix I describes the explanatory variables in detail including *River GeoDSS* database keywords and processing methods.

Training in Passes

As discussed previously, explanatory variables that are dependent on Geo-MODSIM simulations can create instability in the predictions if the training values significantly differ from the simulation values. The aforementioned training technique attempts to alleviate this effect. All the scenarios to be included in the training are executed in Geo-MODSIM prior the first training as an initial approximation in order to mimic all important model conditions, such as scenario demands, seepage, and reservoir storage. In the baseline network calibration, the *River GeoDSS* provides gains and losses to closely match the measured flows that bring flows in the system closer to a conjunctive use simulation flow. The ANN training dataset is then generated using the initial set of Geo-MODSIM results. Using the trained ANN, a new execution of the Geo-MODSIM networks is performed that allocates water according to the new system conditions, including the ANN predicted stream-aquifer interaction. The explanatory variables for the ANN are computed based on the first MODSIM run in order to lessen the impact on the predictions. A second ANN training session can then take place using the most recent MODSIM results to build the new training dataset. Figure 4.6 shows the sequence used to train the ANN in several passes or sequences.

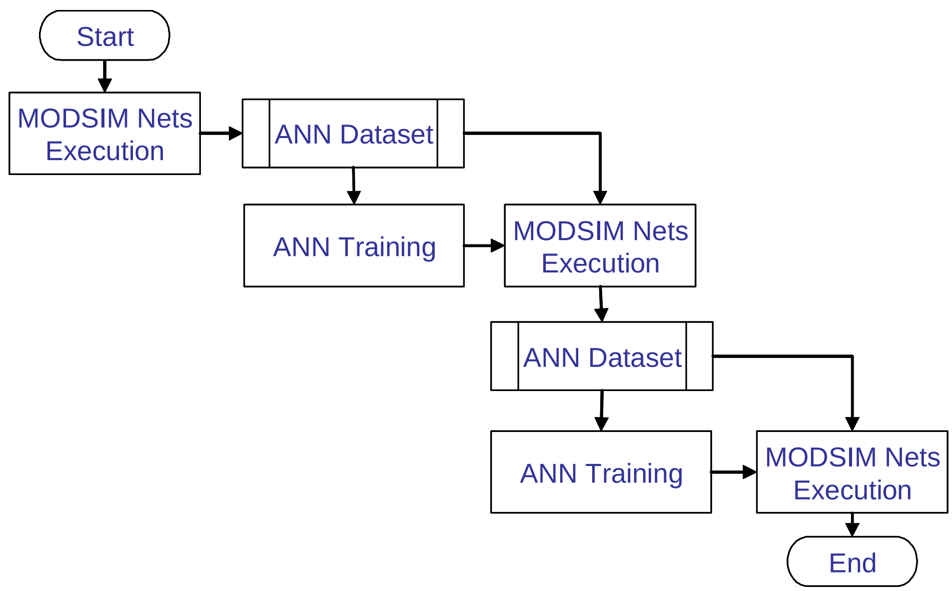


Figure 4.6 – ANN training in passes diagram

Using this iterative process between the ANN training and *River GeoDSS* simulation, it is possible to refine the explanatory variable dataset to more accurately capture the complexity of the system model. In the future, this method can provide a valuable contribution to refine the groundwater model by providing a better representation of spatial and temporal water availability, especially for scenarios where no historical measurements are available.

Training Dataset

The ANN training dataset is created using a set of tools that (1) extract the spatially dependent variables from the GIS spatial database, (2) query baseline and management alternative Geo-MODSIM results for the modeling dependent variables, and (3) summarize temporal-varied variables from the time series database per modeling time step.

Geo-Processing Tools

A set of tools have been implemented in VB.NET using ESRI-ArcObjects to geo-process and extract explanatory variables aggregated by grouping area or area-buffers. These tools are packed under a user interface (Figure 4.7), which is accessed from the *River GeoDSS* toolbar using the button .

The data management interface allows system elements to be associated with the name of representing feature class (from the available in the ArcMap™ project). The system elements defined in the data management tool include irrigated fields, pumping wells, canals lines, the Digital Elevation Model (DEM), water bodies, the main stream and the grouping areas. In addition, the data management interface allows selection of file paths and file names for the time series database, the precipitation database, water quality database, and the ANN buffers and training datasets.

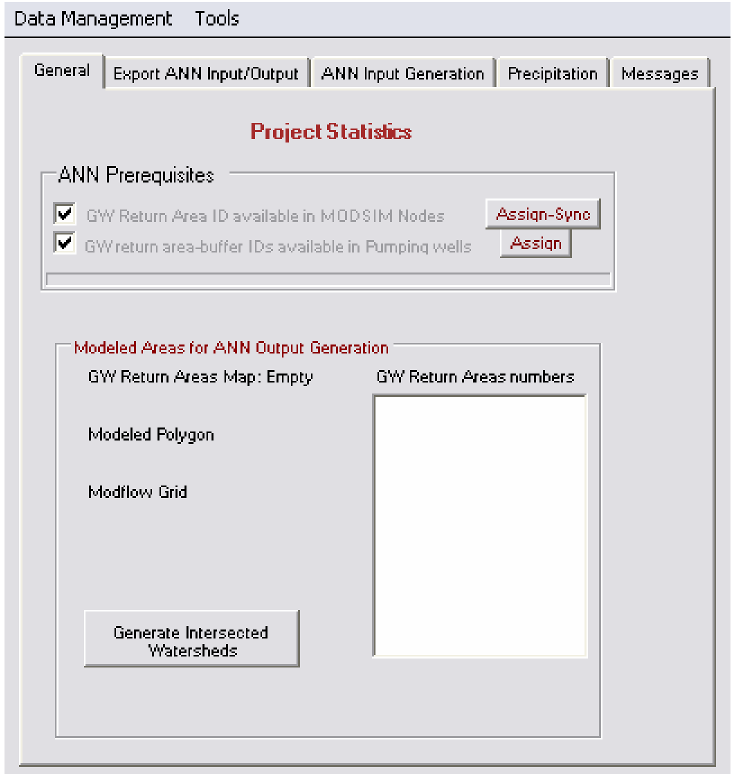
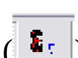


Figure 4.7 – Geo-tool user interface for ANN training dataset processing in ArcMap^{TM}

Spatial Precipitation Tool

The *precipitation* tab (Figure 4.7) gives access to a tool for generating a database with tabular precipitation summarized per area-buffer and per time step. This tool summarizes precipitation based on a set of precipitation raster maps, which are the result of processed NEXRAD data or raster maps generated from point-measured precipitation, using a *River GeoDSS* tool () to generate raster maps from National Weather Service (NWS) and Colorado Agricultural Meteorological network (CoAgMet) stations (Figure 4.8). The raster maps folder location is specified in the data management interface. The generated raster files are named with user defined prefix (data management interface) and the simulation time step start date.

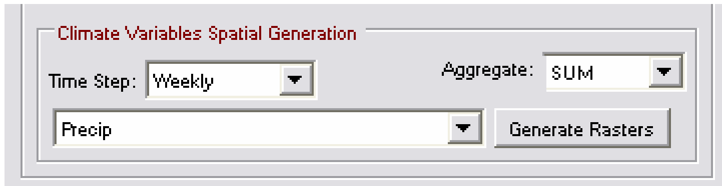


Figure 4.8 – *River GeoDSS* climate raster maps from point-based data generation interface

ANN Input Generation Tool

The *ANN Input Generation* tool processes the available spatio-temporal data and stores the summaries in a database referred as the *buffers database* (assigned in the Data Management interface). The user specifies the number of area-buffers and the base size of the buffer, i.e., the first buffer (Figure 4.9). These parameters are used in the grouping area polygons to create the area-buffers and generate a set of support summary tables and processed GIS feature classes for the creation of the ANN training dataset.

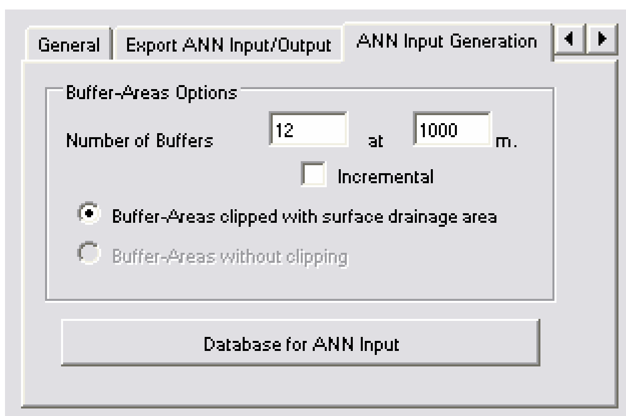


Figure 4.9 – ANN training dataset preprocessing interface

MODSIM Inputs Tool

This tool uses MODSIM output files to extract the model related explanatory variables. This tool allows using in the ANN training dataset results of complex *River GeoDSS* modeling, including water rights allocation, calibration flows, temporal and spatial water availability and accurate modeling of scenarios dependent variables such as local inflows (vertical drainage scenario), diversions, seepage and even predicted return flows.

Geo-MODFLOW in ANN Training

Geo-MODFLOW tools are used to build the ANN training dataset. The geo-referenced grid cells are geo-processed (clipped) to provide the groups of cells included in the analysis of each of the grouping areas. The geo-processing results include cells per area-buffer (*BGridCells* feature class) and MODFLOW river cells classified by grouping area for both the main stream and the tributaries (*BRiverCells_XX* feature class). These processed feature classes are stored in the *buffers* database. The variables included in the training dataset are: (1) total volume returned to the main stream and tributaries, (2) total salt returned to the main stream and the tributaries, (3) calculated concentration of the return flows, (4) grouping area aquifer recharge, (5) grouping area pumping and corresponding salt load, (6) simulated drained volume and corresponding salt load, and (7) the area-buffer based canal seepage. Details on these MODFLOW modeled variables are provided in Appendix I – *MODFLOW-MT3DMS Variables for ANN Training.*

Export ANN Input\Output Tool

The *Export ANN Input/Output* tool uses the *buffers* database support tables and GIS processed feature classes to generate datasets with inputs and outputs for all simulation time steps and all modeled management alternatives. The user specifies the type of data to be included in the dataset (i.e., input or output data) and the type of output (i.e., quantity or quality), as well as the type of grouping areas to include. Although explanatory variables for simulation grouping areas can be generated, some may lack modeled output variables. Figure 4.10 shows the user interface for the ANN training dataset generation. The *River*

GeoDSS Simulation Scenario Manager is used to display/enter preferences for the scenarios to be used for training, including the MODFLOW file locations and characteristics of alternatives for pumping increase, recharge reduction, seepage reduction and drainage density. The tool generates an MS-Access database containing the ANN training dataset; details on the generated database structure are found Appendix I – *ANN Training Database*.

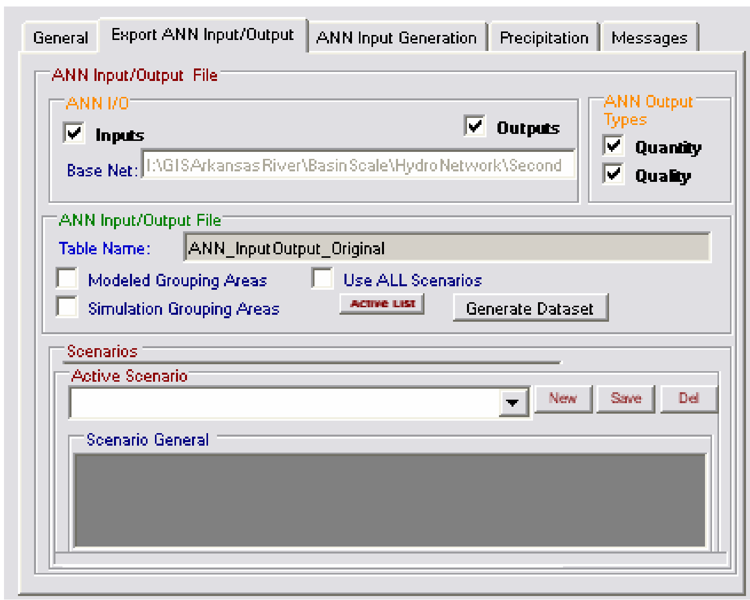


Figure 4.10 – ANN training dataset generation user interface

ANN Database Management Utility

This tool filters and processes the dataset generated by the geo-processing tools to create files for ANN MATLAB training and *River GeoDSS* simulation.

The *ANN Database Management* utility adds four variables to the dataset, making them available for the ANN training. The variables introduced are: (1) *NETRetFlow_CalcArk* and *NETRetFlow_CalcTrib* that use the MODFLOW output variables to compute the net volume from the aquifer to the river and to the tributaries, respectively, where a negative sign indicates water leaving the aquifer, (2) *RechReduction*, which uses the user defined baseline to compute reduction in recharge to the grouping areas, and (3) *SeepReduction_* that computes, per area-buffer, the seepage reduction from the user-defined baseline scenario.

This tool implements two MS-Access user interfaces. The first one (Figure 4.11) allows data grouping, filtering the initial time steps (i.e., excluding them from training) and applying factors to the dataset variables to facilitate the training. In addition, this user interface allows splitting the dataset into groups based on the month number for separated ANN training under different seasonal conditions (e.g., irrigation and non-irrigation). Experimentation on the ANN stream-aquifer interaction modeling showed insignificant improvement when training the ANN for irrigating and non-irrigating seasons; therefore, a single period was selected for this study and this ANN split option is not available in the *River GeoDSS*.

The second interface (Figure 4.12) allows selection of the training dataset generation parameters, including (1) the grouping areas for training and testing, (2) explanatory variables, (3) output variables to be predicted by the ANN, (4) previous time step variables and number of previous time steps to be included in each explanatory variable set; (5) dataset generation options, and (6) naming conventions for the output files. Explanatory variables with "_" indicate that they will be included for all the area-buffers specified in the preferences. This interface processing engine implements an algorithm that includes input/output variables from previous time steps to simulate a recurrent effect that allows the neural net to capture time-varying patterns. Although, the feedback is static, and it will not change during training based on the actual outputs, as it would in a Jordan network (Jordan 1990). The advantage of having the outputs included in the inputs is that the training dataset can be randomly extracted from the entire dataset. This utility exports a group of text files for MATLAB ANN training and *River GeoDSS* simulation support; the files names and description are available in Appendix I – *ANN Training Dataset Files Description.*

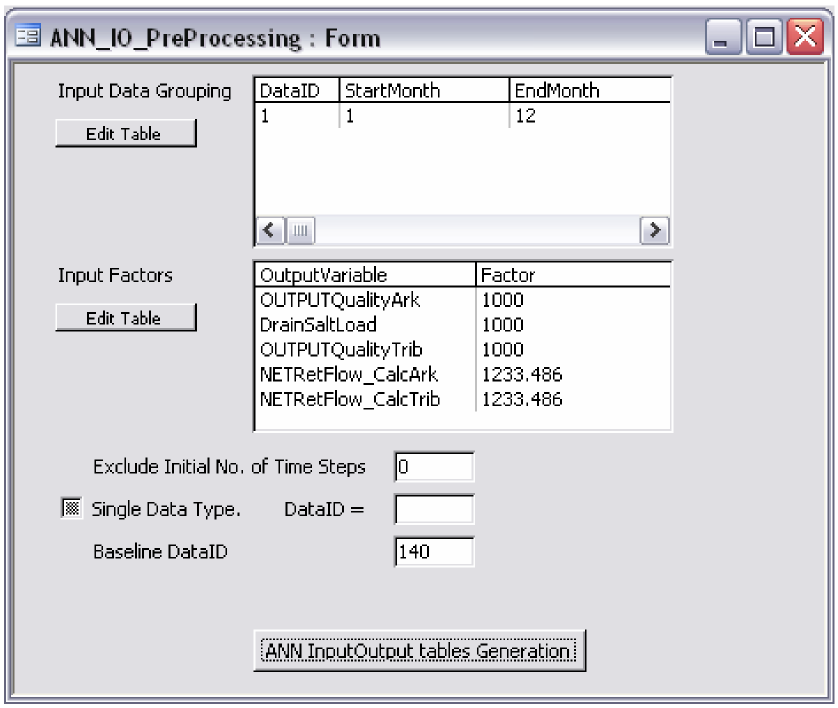


Figure 4.11 – ANN training dataset pre-processing interface

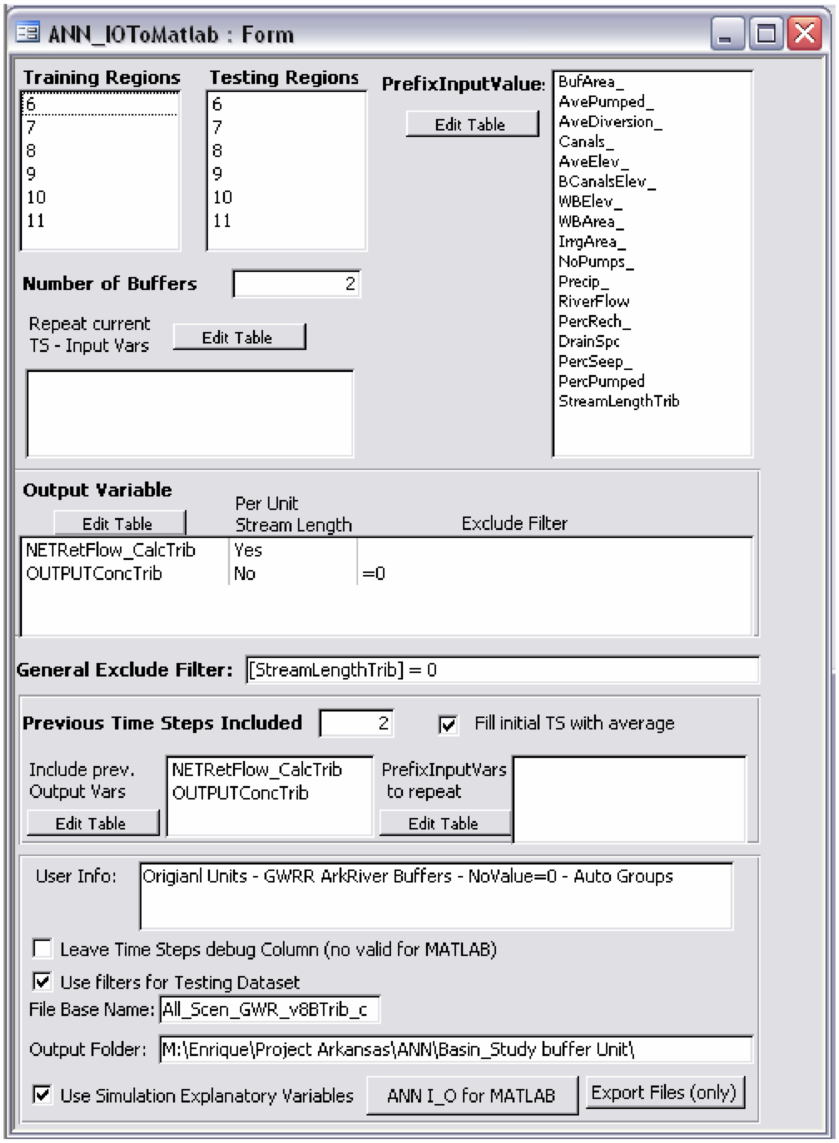


Figure 4.12 – ANN training dataset preferences interface

MODSIM Reservoir Operation in the Arkansas River System

Explanatory variables such as the flow and diversion (from the Geo-MODSIM modeling) play an important role in the set of explanatory variables. In addition, the reservoir operation rules are important for water allocation and, consequently, in the Geo-MODSIM flow calculation. Two reservoir operating rules were implemented to generate the ANN training datasets. The first one sets the targets to historical levels for Pueblo and John

Martin Reservoirs. The second operating rule uses reservoir layers with incremental costs to balance storage water between the two reservoirs. Using historical reservoir levels as targets prevents storage of water above the historical levels, but will satisfy historical water rights and river compact requirements by releasing storage water if needed and replenishing it as soon as possible. The reservoir layer balancing approach uses incremental costs in the user-defined reservoir layers to dictate the allocation of water in the reservoir system. Calibration of layer balancing operating rules includes setting the reservoir layer costs to match as close as possible the historical levels in the reservoir, so that system flows in the baseline network with layer balancing operating rules are close to the baseline flows with the first operating rule. Difference in MODSIM simulations are noticeable in the scenario simulations where "free" water can be stored in the reservoirs rather than released down the system.

The sets of baseline and scenario MODSIM runs for each operating rule are used to generate two separate ANN training datasets. The dataset generated with the set of MODSIM runs using the historical level as reservoir targets is referred to herein as *Dataset_A* and the dataset generated with reservoir layer balancing is referred to as *Dataset_B*.

ANN Training

Basin-scale stream-aquifer interaction modeling is achieved with two independent ANNs. The ANN training characteristics and relevant explanatory variables for stream-aquifer interaction modeling of the Arkansas River versus its tributaries are significantly different. Therefore, an ANN for each type of training is implemented to accurately model the system water conjunctive use. This separated training process gives the ability to better filter the dataset, thereby improving performance by training and predicting over more homogeneous input\output cases. For both the Arkansas River ANN and the tributaries ANN the explanatory variables are spatially group using area-buffers in Figure 4.5.

Grouping areas (6 and 11) located near the edges of the groundwater modeled area require special handling since if only the modeled area is considered, it results in explanatory variables values of smaller magnitude due to the smaller number of features considered, rather than if the entire area were taken into account. Predictions per unit length in grouping areas 6 and 11 are assumed to remain similar to the modeled values per unit length, which includes only a portion of these grouping areas. Therefore, the ANN training dataset is built assuming that explanatory variables for the entire grouping areas 6 and 11 (including no modeled areas) will produce the same results per unit length of stream as the values calculated in each modeled area. Using this approach all the grouping areas for training will have a common ground for comparison in terms of their river lengths and area-buffer extents.

Both *Dataset_A* and *Dataset_B* are used to train ANNs for the basin-scale stream-aquifer interaction modeling. Each of the datasets produces ANNs that attempt to predict the same phenomena, thereby giving an indication of the sensitivity of the predictions to changes in the MODSIM-reliant explanatory variables.

The training dataset for the ANN that models the stream-aquifer interaction is prepared using the preferences displayed in Figure 4.11 for the ANN Database Management Utility, including (1) all months in a single group, (2) a factor of 1000 applied to the ANN predicted salt load related variables [i.e., salt loads to the Arkansas River

(*OUTPUTQualityArk*), salt loads to the tributaries (*OUTPUTQualityTrib*), and salt loads in drained water(*DrainSaltLoad*)] to reduce their magnitude on the ANN training, and (3) a factor of 1233.486 to convert the computed net return flows (*NETRetFlow_Calc*) from m^3 to acre-ft.

Custom MATLAB Training Tools

A set of customized tools were developed using the MATLAB® neural network Toolbox libraries. These tools allow importing the training datasets, selecting training parameters, preparing the data, training and testing the ANNs. The training process begins by browsing for the ANN Database Management utility exported files. A set of dialogs allows the user to select options for the ANN training. The user-selected options include scaling type, training/testing/validation groups to be generated, the size of the training dataset, the type of network, the network structure, and the training parameters.

The Training Tool uses the training dataset created by the Database Management Tool (the testing dataset is used later for simulation performance analysis). This training dataset is divided in training, testing and validation groups according to the user preferences. The testing/validation groups are used for the backpropagation ANNs to avoid overtraining. These groups consist of training cases that are not presented to the ANN for training but rather used for stopping the process when overtraining is detected. If later in the training process the user selects a type of network that does not require the testing and validation groups, these cases are merged back with the training cases.

ANN Explanatory Variables Automatic Scaling

The pre-processing algorithm automatically selects between two scaling methods: (1) Min/Max scaling (*MnMx*) and (2) Standard Deviation (*Std*) scaling.

In *MnMx* scaling, the explanatory variables are scaled between -1 and 1 using the following transformation equation:

$$pn_i = 2 * \frac{(p_i - min_p)}{(max_p - min_p)} - 1$$

where min_p = the minimum of explanatory variable p and max_p = the maximum value for the explanatory variable p.

The *Std* scaling transforms the data so that its mean is 0 and a standard deviation of 1 using the following equation to transform the variables:

$$pn_i = \frac{(p_i - mean_p)}{std_p}$$

where pn_i = the transformed value; p_i = the explanatory variable value; $mean_p$ = the mean of the explanatory variable p and std_p = the standard deviation of the explanatory variable p.

The *MnMx* scaling option is used for explanatory variables when the mean value of the *MnMx* transformed data falls between 0.25 and -0.25; otherwise, *Std* scaling is used. This rule attempts to maintain a symmetrical spread of the scaled data around the 0 value. In an attempt to reduce the effect of errors in the previous time step explanatory variables during simulation, a tighter interval of -0.1 and 0.1 is used for assigning *MnMx* scaling to these variables, resulting in a *Std* scaling in almost all cases. The output variables are always scaled using the *MnMx* method.

ANN Training Dataset Size

The combination of about 130 weekly time steps for the baseline and 36 scenarios produces a very large dataset for training. Training an ANN with large dataset becomes impractical or impossible due to computing limitations. Therefore, the custom ANN training tool allows the user to randomly select a specified number of cases for training (Figure 4.13). The remaining training cases are added to the performance testing dataset.

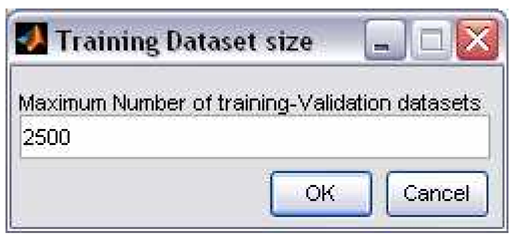


Figure 4.13 – MATLAB ANN training tool dataset size user-dialog

Neural Nets Types and Training Preferences

Four types of neural nets are implemented in the MATLAB training tool: (1) feed-forward backpropagation, (2) Elman backpropagation neural network, (3) Generalized Regression Neural Net (GRNN) and (4) Radial Basis Neural Net (RBNN).

Backpropagation type networks use a training method that relies on sequential improvement of weights and biases to minimize errors between the predicted and observed values; therefore, the minimization result is a function of the initial parameters condition. A backpropagation training event is difficult to reproduce, since the initial training parameters are randomly selected, and the training might require many trials to acquire the

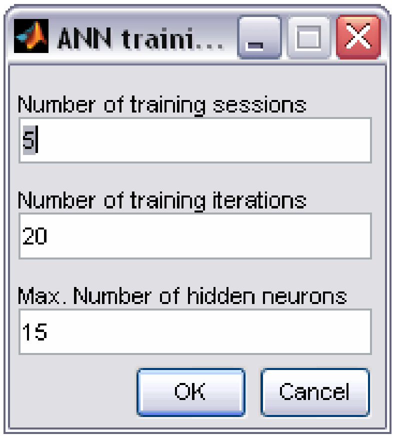


Figure 4.14 – Backpropagation Neural Net Training Dialog

desired performance. The developed training tool implements a sequential training process in which many training events in several sessions are carried out, with the best performance training selected per session. Figure 4.14 shows the user preferences interface for the backpropagation sequential tool. For each training event, the algorithm selects a random number of neurons (up to the user specified maximum); in addition, it changes the MATLAB training method (i.e., traincgf, traincgb, trainscg, traingdx, trainbfg and trainlm), and randomly selects the layers transfer function type (i.e., purelin, tansig and logsig). A weighted performance function is used to select the best trained network per session. The performance function includes training and validation mean squared errors (MSE), which are combined using a factor of 0.3 and 0.7, respectively. The performance function (f_p) is computed as:

$$f_p = 0.3 \cdot MSE_T + 0.7 \cdot MSE_V \tag{4.1}$$

where MSE_T = training dataset mean squared error and MSE_V = validation dataset mean squared error. The training process stores the best networks for all the training sessions, which are available for comparison and analysis in the custom MATLAB post-processing tools. In the feed-forward network training, additional options are available to the user such as: training stopping methods and types of network structure. The training tool implements three MATLAB training stopping methods: Early stopping, Regularization and Bayesian regularization. The user can select from two types of feed-forward neural network architectures: Feed forward or Cascade forward.

The Elman Network training follows the backpropagation training procedures described above, but is significantly more difficult and time consuming than the other backpropagation networks. Therefore, training is restricted to smaller number of sessions.

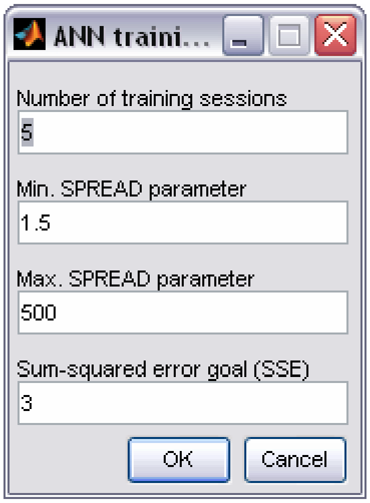


Figure 4.15 – Radial Basis NN training preferences

The RBNN type networks are trained based on the user-specified sum-squared error (SSE) goal and network *spread*. Figure 4.15 shows the user interface for the radial basis ANN training. The custom training tool trains the user-defined number of networks between the maximum and minimum *spread* until the SSE goal is reached or more than 400 neurons are required in the training event, size found to be impractical for training and prediction generalization. The training sessions are then stored for further visualization and analysis. Use of various network *spreads* results in different separations of the neurons and consequently requires different numbers of neurons to achieve the error goal. Parsimony in the number of neurons required results in better generalization capabilities of the ANN.

The Generalized Regression neural network (GRNN) is a type of radial basis network, but is rather insensitive to the *spread* parameter, as with the RBNN. Since the range of spreads with good performance is small (1 to 4), the training process becomes almost unique for each set of training datasets. In this training process, the user is only asked for the *spread* value. Commonly, the GRNN architecture results in a network with larger numbers of neurons than the regular radial basis network.

Neural Network Selection

Neural Network selection involves training different types of implemented ANNs using several configurations and sizes of datasets in order to observe and compare their performance in training/verification, overall performance (using the testing dataset) and network structure complexity. The best ANNs are compared in detail by comparing predictions for the baseline scenario at individual grouping areas. Finally, selective screening is performed to observe return flow and concentration predictions for scenarios compared with the baseline predictions. These prediction differences are essential for analysis of the management alternatives using the ANN conjunctive use modeling.

The feed-forward ANN training is disadvantageous because of the large number of training events needed to achieve good performance. Prediction performances are highly variable depending on the training event, which results in a wide range of prediction errors for the best networks. Even though one-third of the available cases (8000+) are used for training, prediction performance on the testing/validation dataset was extremely poor, in most of the cases. Based on the limited number of training events performed, the feed-forward NN seem to lack the ability to predict this complex phenomenon.

The Elman recurrent network was not used for this modeling because the nature of the training dataset is unsuitable for this type of ANN. The training cases were derived from small sets (i.e., grouping areas and scenarios) of chronological events, but these sets are not all sequential. Since, the network training relies on the order in which the explanatory variables are presented, combining cases from different sets results in misinterpretation of the training dataset.

The GRNN prediction is limited by the size of the training dataset, with the maximum number of training cases found to be less than 500 due to computer memory limitations. The GRNN training produced ANNs with the number of hidden neurons equal to the number of training datasets, which indicates that only the randomly selected cases will influence the prediction. The GRNN training is repeatable and rapidly processed, with acceptable overall performance but poor generalization. The ANN testing results show smooth predictions but including large prediction errors in some cases. The differences between the baseline and scenario predictions were inconsistent with the modeled differences.

This experimentation with different ANN architectures and configurations pointed to selection of the radial basis neural network as best suited for the modeling based on prediction performance and generalization capabilities. Upon adequate combination of parameters and error goals, the RBNN was able to model the process with a small number of neurons, and produce predictions with the smallest errors. Although for some of the test cases the correlation between the predicted and the modeled values was low, the average predictions seemed to have a relatively small error based on visual inspection of the graphical results. In addition, differences between the baseline and the management scenario predictions produced the correct prediction change direction in most cases. In many instances, the predicted differences have a similar magnitude to the differences between the modeled baseline and the corresponding management scenarios.

Arkansas River Stream-Aquifer Interaction ANN

For these ANNs, explanatory variables are gathered from only the first area-buffer (i.e., closer to the Arkansas River). Figure 4.16 shows the preferences used to generate the ANN training dataset that predicts Arkansas River stream-aquifer interactions using *Dataset_A*. Values from four previous time steps are used in the dataset to provide the ANN with time-varying "memory" effects. The variables made recurrent are: (1) grouping area average flows, (2) net return flows to the Arkansas River and (3) return flow concentrations. Notice that the return flow output is normalized per length of the main stream in the grouping area. Figure 4.17 shows the preferences used to generate the ANN training dataset for modeling the interaction between the aquifer and the Arkansas River using *Dataset_B*. In this case, since only data from two previous time steps are used in the dataset, the increase in the previous time steps included in the network "memory" can potentially magnify the error propagation during simulation.

The networks used to model stream-aquifer interactions in the Arkansas River basin are trained using a random set of 2500 cases. The radial basis ANN is trained for *spreads* between 2 and 150 and a SSE error goal of 15. The best performance (Equation 4.1) was achieved in both trainings using a *spread*[1]=39. The ANN internal configuration is defined in the training process based on the training parameters (i.e., *spread* and error goal). The training results in an internal configuration of 11 hidden neurons for the ANN from *Dataset_A* (*AllScen_GWR_v8BArk_b*) and 34 hidden neurons for the ANN from the *Dataset_B* (*AllScen_GWR_v8Ark*).

[1] Best *spread* from a discrete search using the user-specified maximum, minimum *spread* and the number of training events.

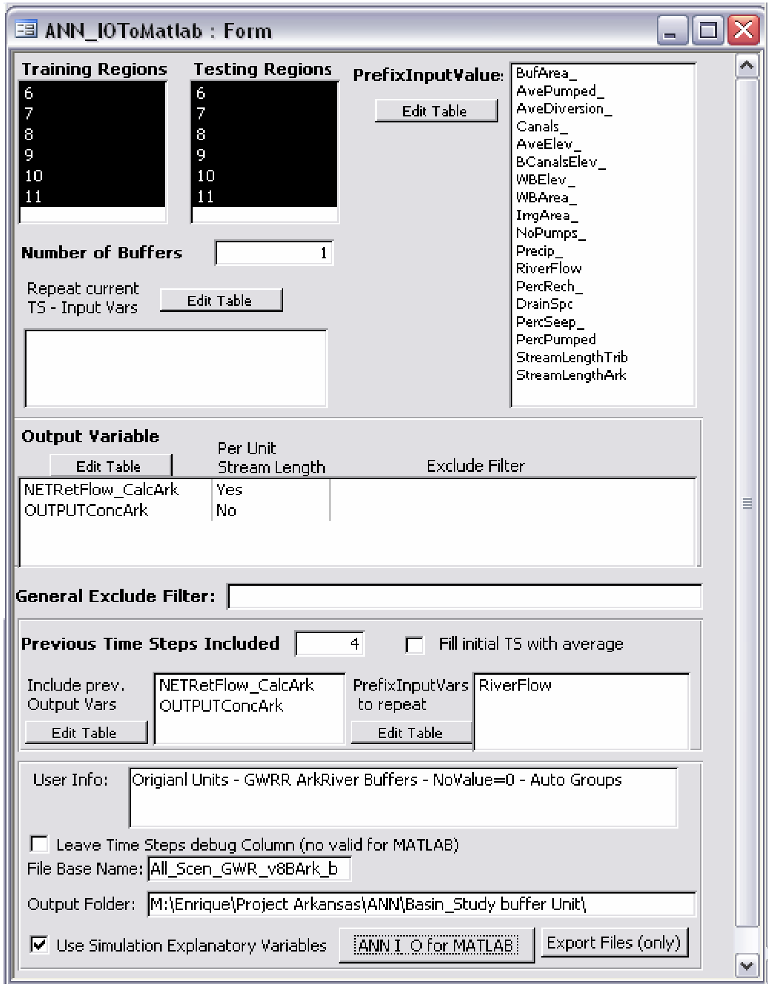


Figure 4.16 – ANN training dataset preferences for the Arkansas River stream-aquifer interaction modeling (Dataset_A)

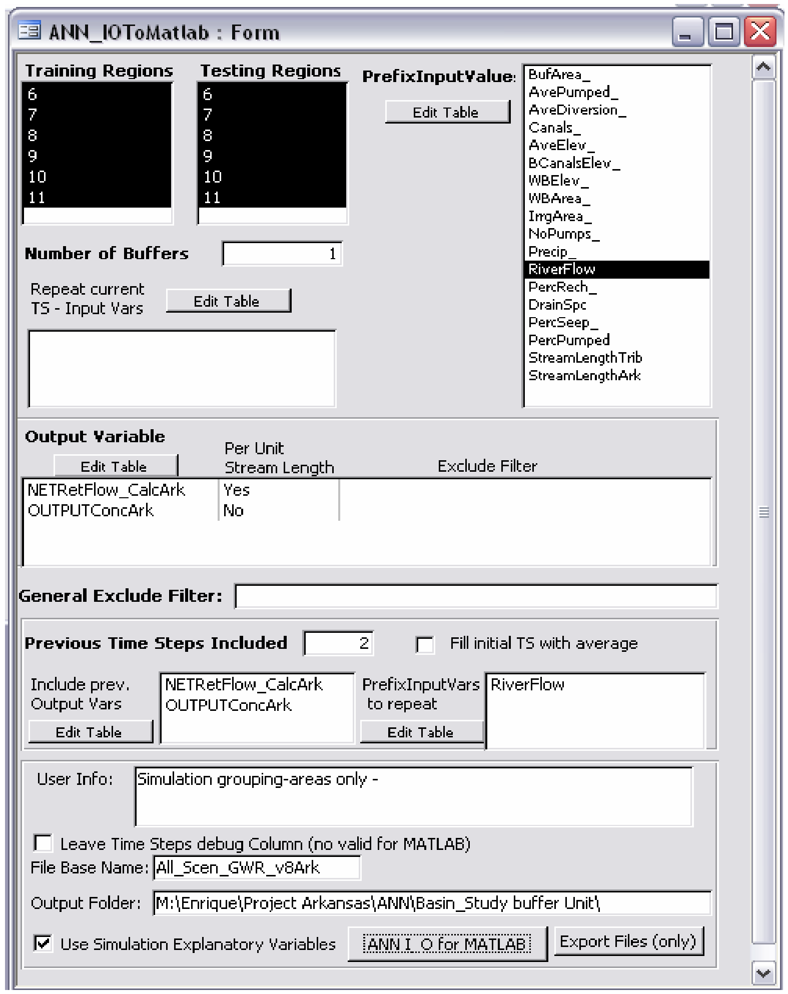


Figure 4.17 – ANN training dataset preferences for Aransas River stream- aquifer interaction modeling (Dataset_B)

AllScen_GWR_v8BArk_b Validation for Arkansas River Stream-Aquifer Interaction

The training performance analysis report (Figure 4.18) is used to visualize the main elements and statistics of an ANN training event. This report belongs to a set of reports generated by the custom ANN training tool. The performance analysis report includes plot of predictions vs. modeled values in both the training and original scales and predictions statistics in both training and testing\validation. Figure 4.18 shows the ANN training and testing prediction performance of the *AllScen_GWR_v8BArk_b* neural net to prescribe the return flow per unit length to the Arkansas River. The analysis shows the prediction performance in both the randomly selected 2500 training cases and the remaining 25000+ cases (testing\validation cases). The testing forecast bias *m*(*e*)= -0.002 acre-ft/km, indicating that the prediction mean is close to the mean modeled values. The testing root mean squared error per week *s*(*e*) = 7.2 acre-ft/km, which provides an indication of the variability of the forecast errors. The noise-to-signal ratio *s*(*e*)/*s*(*y*)=0.002 indicates good accuracy since a small amount of information hidden by the noise (i.e., *s*(*e*) much smaller than the modeled data variance *s*(*y*)). In addition, the coefficients of correlation *R*=0.96 and determination R^2=0.92 during training and testing, indicate a good forecast and do not show signs of overtraining.

Figure 4.19 shows the training and validation analysis for the Arkansas River return flow concentrations as predicted by the *All_Scen_GWR_v8BArk_b* ANN. The prediction statistics indicate an unbiased and accurate prediction with a high coefficient of determination of 0.99 and low noise-to-signal of 0.0001. The root mean squared error *s*(*e*)=48.5 mg/L.

AllScen_GWR_v8Ark Validation for Arkansas River Stream-Aquifer Interaction

Figure 4.20 shows the *AllScen_GWR_v8Ark* return flow prediction performance analysis, and Figure 4.21 displays the *AllScen_GWR_v8Ark* return flow concentration validation analysis. The prediction performances of the *AllScen_GWR_v8Ark* ANN is similar in all aspects to the *AllScen_GWR_v8BArk_b* ANN. The return flow root mean squared error *s*(*e*)=7.80 acre-ft/km and the return flow concentration *s*(*e*)=50.9 mg/L The major difference is a more complex internal structure in the *AllScen_GWR_v8Ark* network. It is believed that the main contributor to increases in complexity is usage of fewer previous time steps explanatory variables.

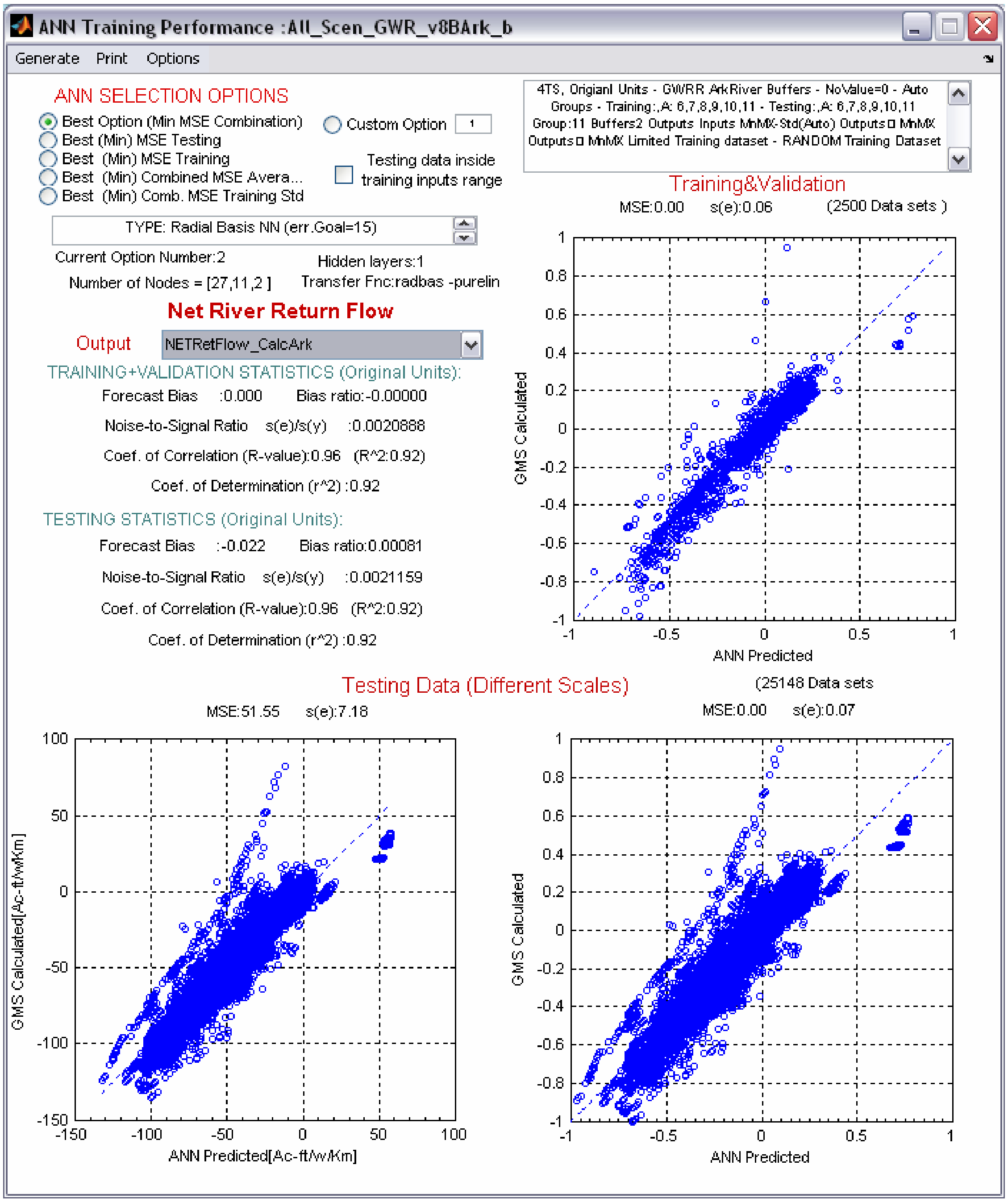


Figure 4.18 – All_Scen_GWR_v8Ark_b ANN training\testing performance

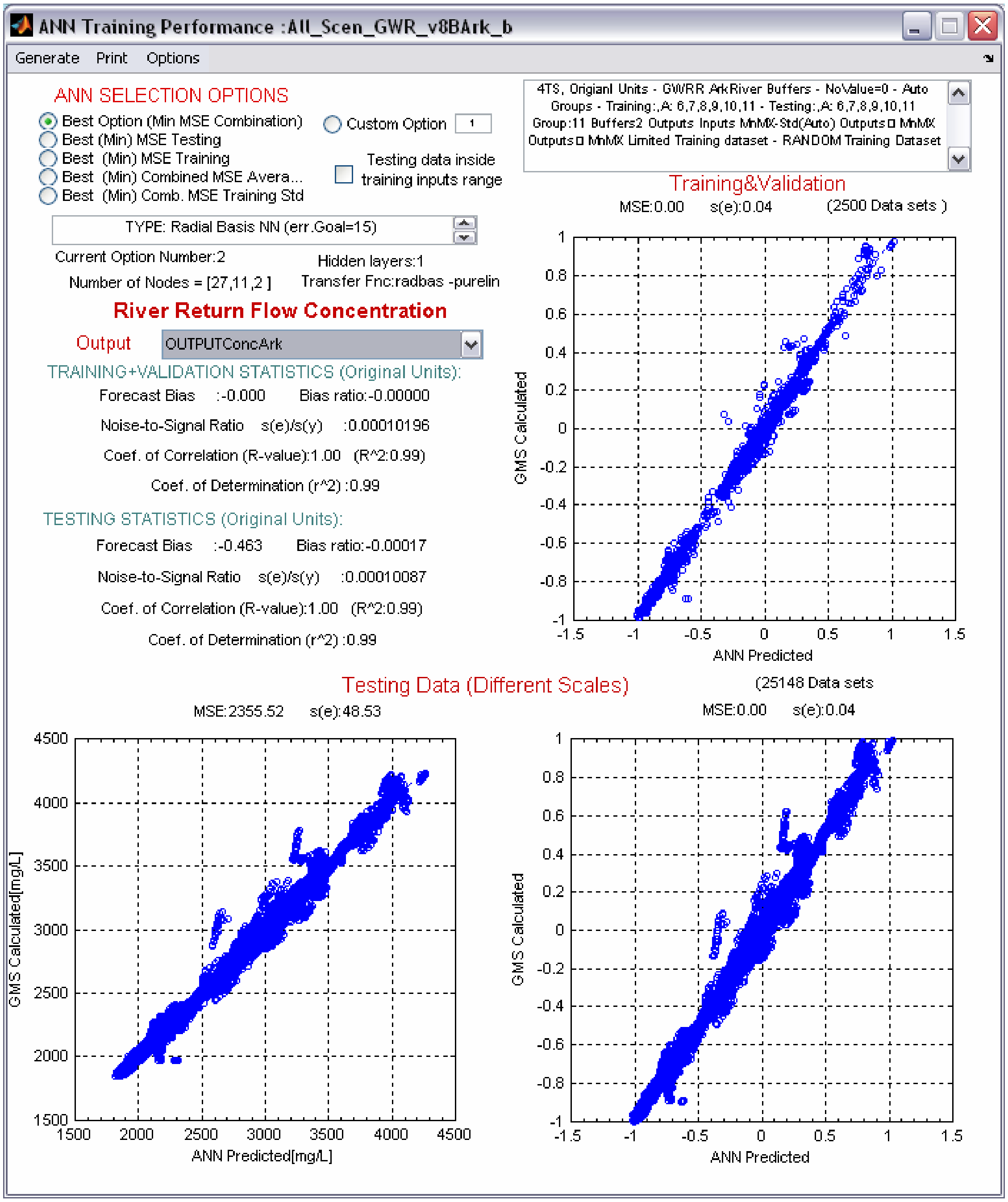


Figure 4.19 – All_Scen_GWR_v8BArk_b return flow concentration training and validation analysis

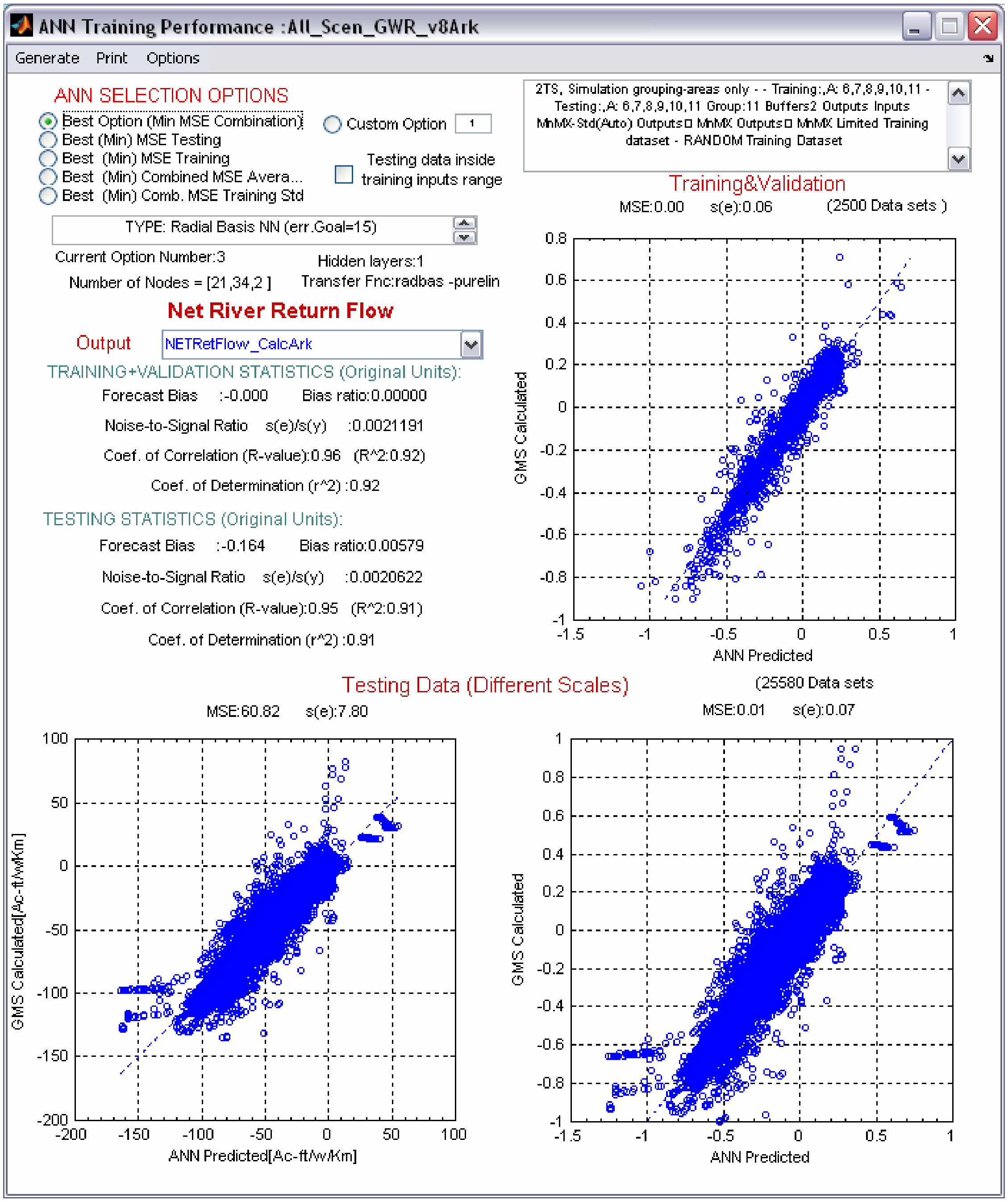


Figure 4.20 – All_Scen_GWR_v8Ark ANN training\testing performance analysis

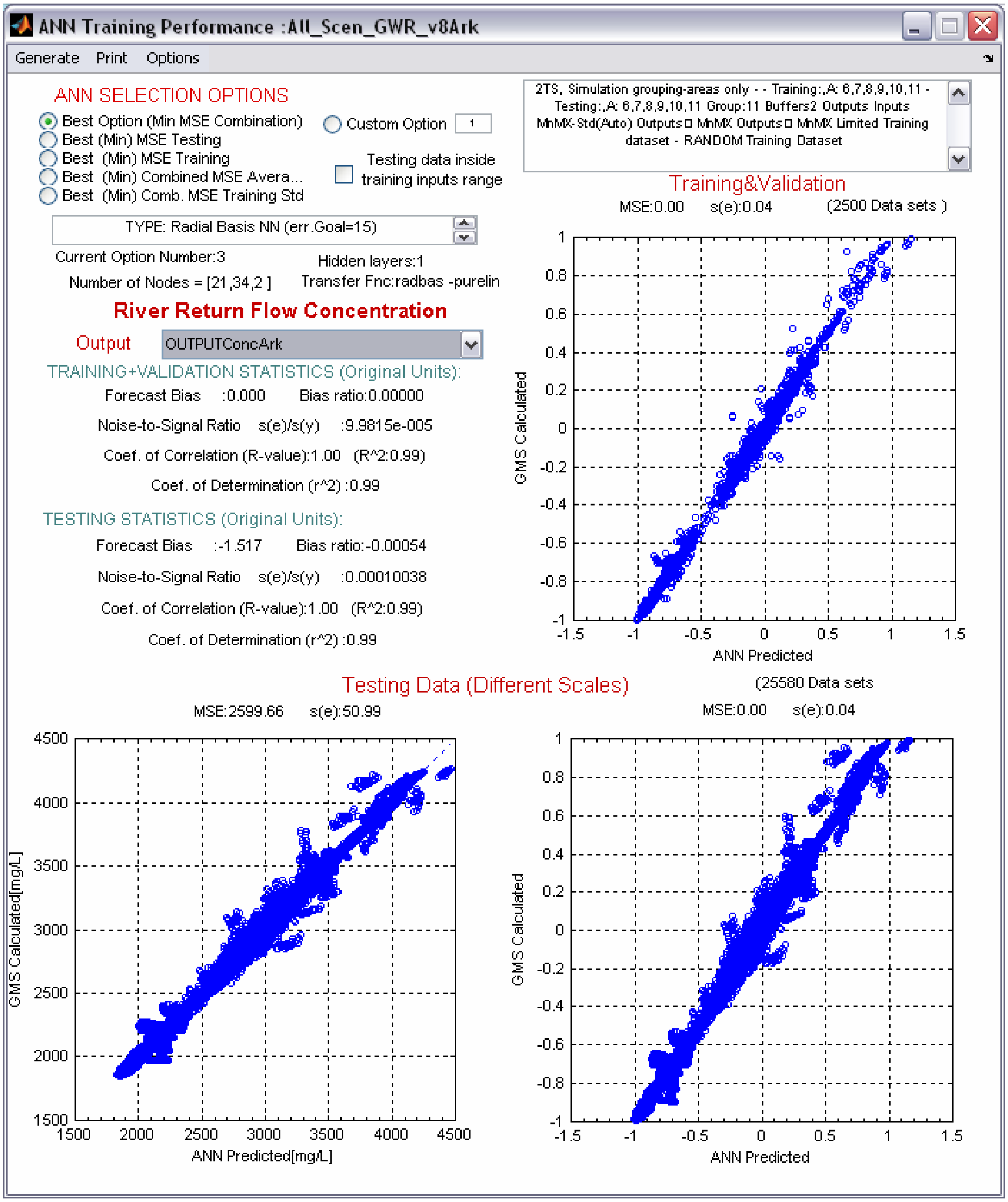


Figure 4.21 – Return Flow Concentration predicted by All_Scen_GWR_v8Ark

Tributaries Stream-Aquifer Interaction ANN

For the tributary stream-aquifer interaction ANNs, the explanatory variables are gathered from both available area-buffers to capture system stresses in areas far from the main stem but important to describe return flows to the near by tributaries. The variables made recurrent are only the output variables: (1) net return flow to the Arkansas River and (2) return flow concentrations. The return flow output is normalized per length of tributary streams marked with an active return flow flag in the corresponding grouping area. Two previous time steps of the return flows and their concentrations are used to provide data on time-varied influences. The training dataset excludes grouping area number 7 since it has a tributary length less than one kilometer. In addition, cases having tributary stream lengths or return flow concentrations equal to zero are excluded from the training. Data filters are also used in the testing dataset for realistic simulation performance measurements. These filters are handled when simulating with these ANNs in the *River GeoDSS* simulation tool.

ALL_Scen_GWR_v8BTrib_c Validation for Tributaries Stream-Aquifer Interaction

Figure 4.12 shows the preferences for the generation of the ANN training datasets using the Geo-MODSIM historical reservoir simulation (*Dataset_A*). The return flow prediction analysis shows good prediction ability of the tributary aquifer return flow (Figure 4.22). The computed coefficients of determination are R^2=0.97 and R^2=0.98 in training and validation respectively. The testing set mean squared error $s^2(e)$=5.67 (acre-ft/km)^2 and $s(e)$=2.38 acre-ft/km; the noise-to-signal ratio=0.003 and the forecast bias shows a negligible over-prediction $m(e)$=0.054 acre-ft/km. Figure 4.23 shows the *ALL_Scen_GWR_v8BTrib_c* ANN return flow concentration prediction analysis. The analysis show accurate predictions with high coefficients of determination of R^2=0.99 and $m(e)$= -1.87 mg/L; with a prediction root mean square error of $s(e)$=46.4 mg/L.

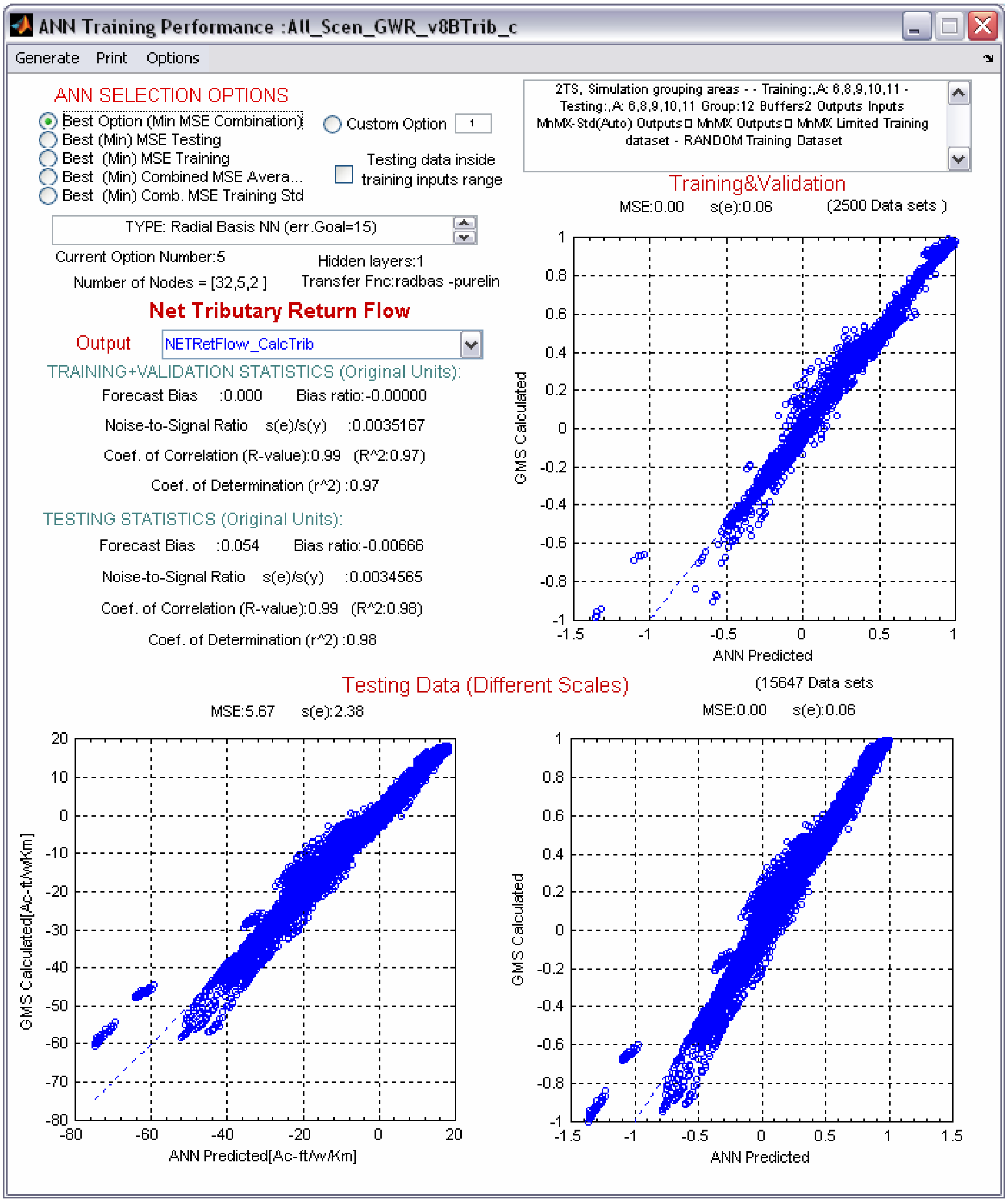


Figure 4.22 – All_Scen_GWR_v8BTrib_c return flow prediction performance analysis

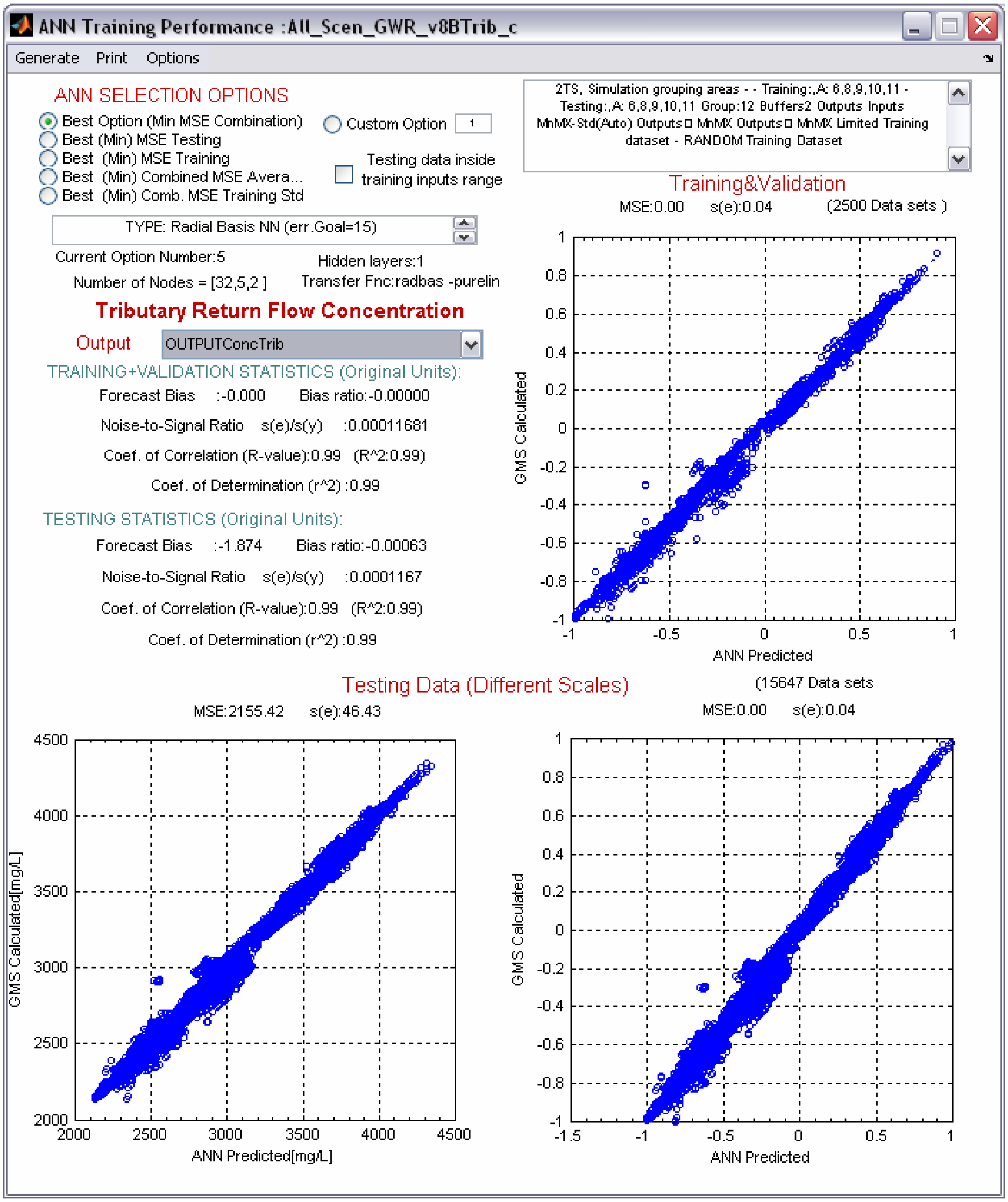


Figure 4.23 – All_Scen_GWR_v8Trib_c return flow concentration prediction performance analysis

ALL_Scen_GWR_v8Trib_a Validation for Tributaries Stream-Aquifer Interaction

This Neural Net is trained to predict the tributaries stream-aquifer interaction using *Dataset_B*. Figure 4.24 shows the preferences used to build the training dataset for this ANN.

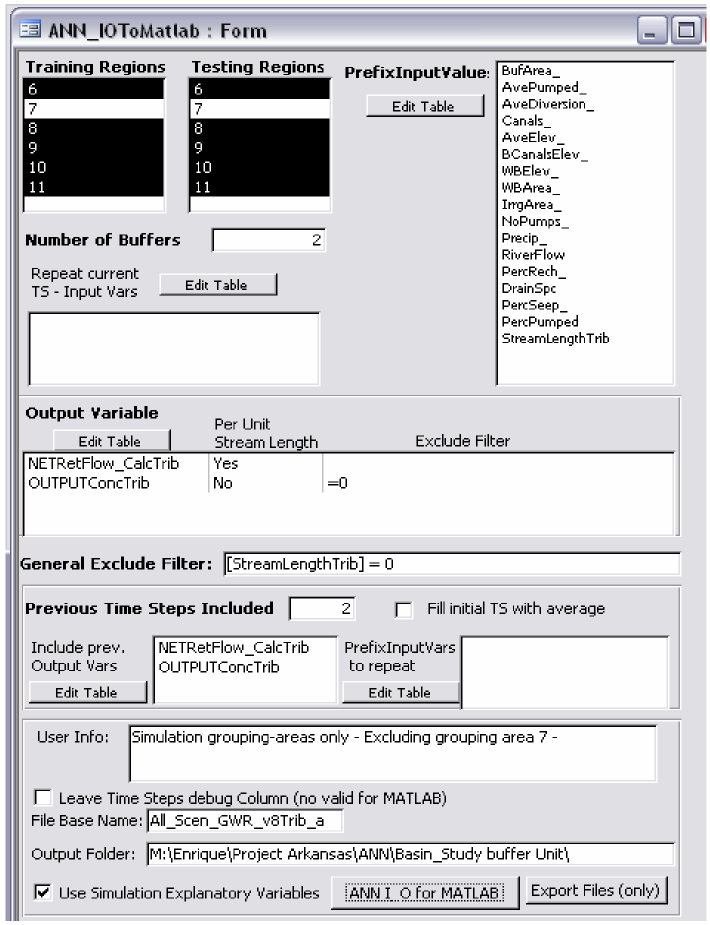


Figure 4.24 – ANN training dataset preferences for aquifer-tributary interaction modeling with Dataset_B

The ANN prediction performances are analyzed on Figures 4.25 and 4.26. The prediction statistics show an accurate return flow prediction with R^2=0.98 in both training and validation; $s(e)$=2.18 acre-ft/km; the noise-to-signal ratio equals 0.003 and $m(e)$= -0.09 acre-ft/km. The return flow concentration results give R^2=0.98; $m(e)$ =1.88 mg/L; and $s(e)$=69.2 mg/L.

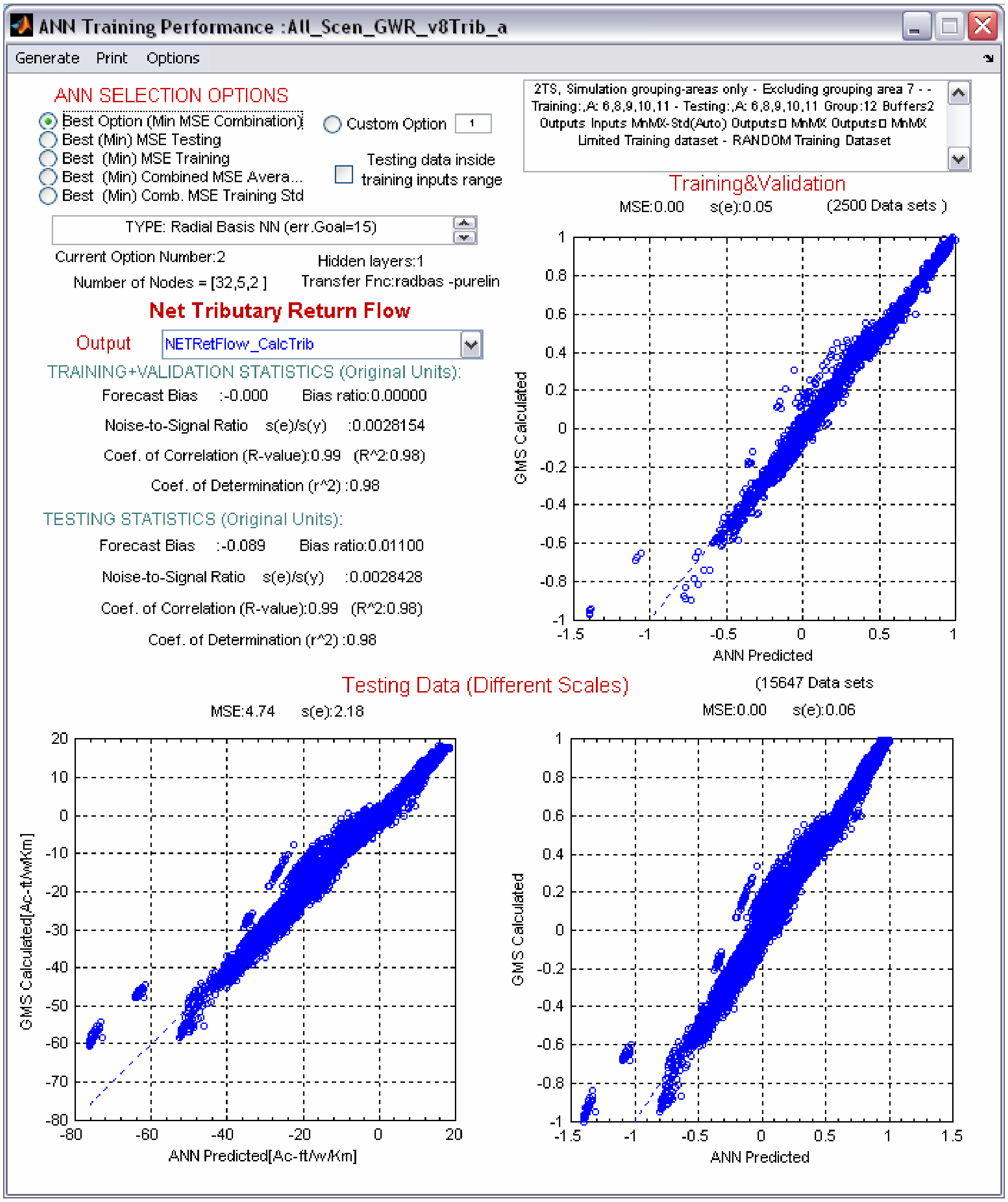


Figure 4.25 – All_Scen_GWR_v8Trib_a return flow prediction performance analysis

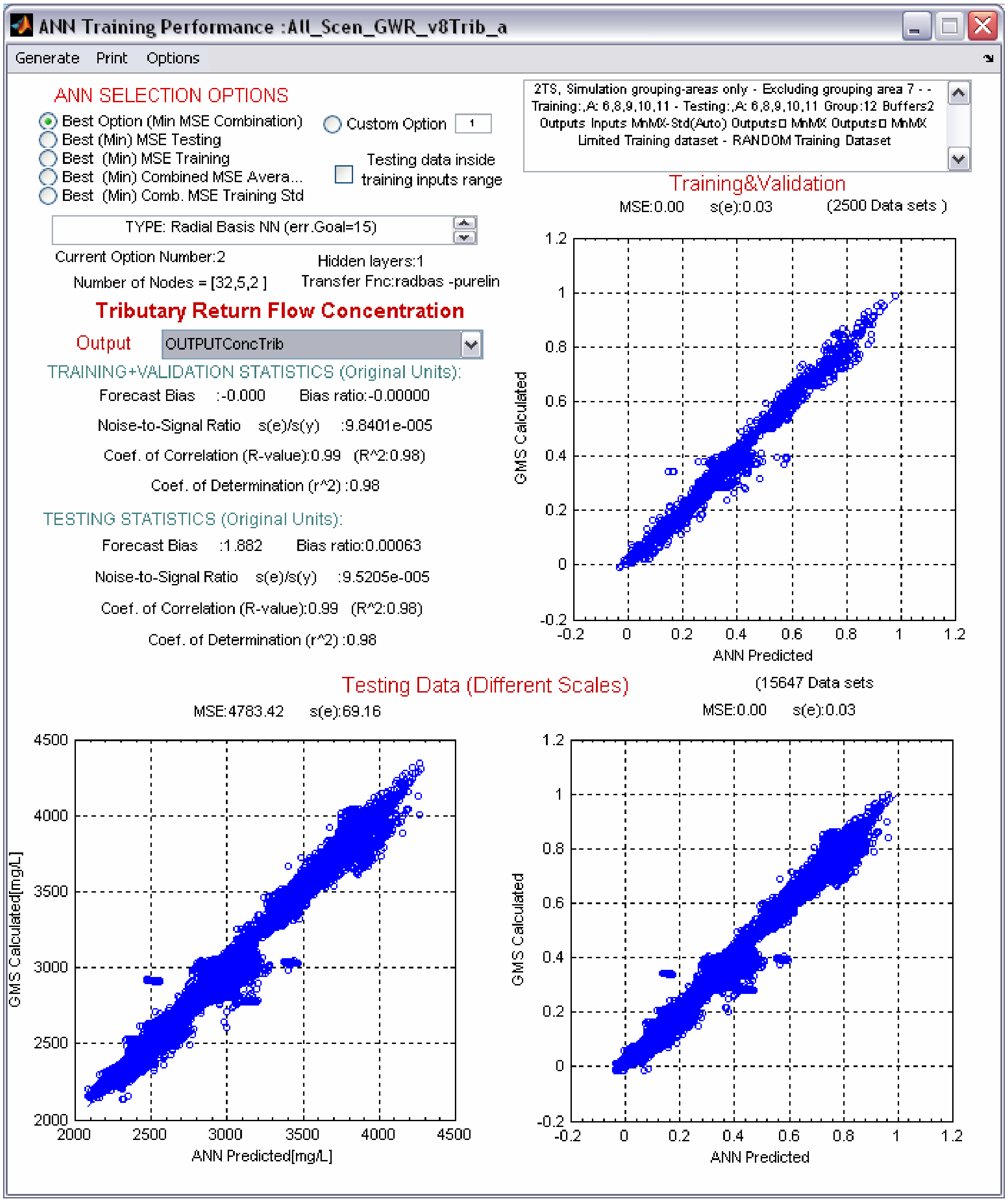


Figure 4.26 – All_Scen_GWR_v8Trib_a return flow concentration prediction performance analysis

ANN Performance Evaluation

The trained ANN performances are evaluated in detail on the baseline and some selected management scenarios. A set of interactive MATLAB-based reports are implemented in the custom ANN training tool to compute and display the predictions for analysis and comparison with observed values (i.e., modeled values in this case). This evaluation consists of comparison of predicted and modeled values for both return flows and concentrations at several detailed levels. The first level is overall performance including all simulated scenarios, along with the baseline, and all modeled grouping areas. The next level is performance evaluation for all grouping areas in the baseline scenario. The final level includes prediction comparisons for each individual grouping area. Appendix II presents the baseline detailed performance evaluation for the four trained ANNs for stream-aquifer interaction modeling and selected comparisons of the prediction for the management alternatives as compared with the baseline prediction.

ANN-based Stream-Aquifer Modeling Analysis

The methodology for stream-aquifer interaction modeling based on artificial neural networks as introduced herein is demonstrated as a robust alternative to traditional methods such as the SDF method (Jenkins 1968) or MODRSP approaches (Maddock and Lacher 1991). Congruent with Suen and Eheart (2003), the ANN type selection gave preference to the radial basis networks to represent complex problems. In this study, the radial basis ANN succeeded in modeling an extremely complex phenomenon with the simplest network architecture, while performing exceptionally well in the overall prediction of the large number of testing cases, including the management scenarios. The largest errors are located at the initial time steps of the groundwater simulation where erratic behavior is visible in the modeled values. The trained ANNs are able to closely duplicate the MODFLOW modeled return flows per grouping area and the corresponding MT3DMS modeled concentrations. Therefore, the trained ANN can be used to substitute the groundwater model in the *River GeoDSS* for conjunctive use modeling. The main advantages of implementing the trained ANN in the *River GeoDSS* modeling system (often referred to as “machine learning” approach) are: (1) the ANNs are able to predict return flows and concentrations faster than a finite difference model, (2) the predictions based on this methodology avoid compromising the quality of the predictions with approximations; e.g., linear superposition, as would be required in simplified groundwater modeling methods (Glover 1974; Jenkins et al. 1972; Maddock and Lacher 1991), (3) its design allows applying the learned relationships between system state changes and stream-aquifer system response in contiguous areas that lack detailed groundwater models.

Traditional coupling between groundwater and surface water models for conjunctive use modeling can be performed directly if the groundwater model covers the basin modeled area and all the simulated alternatives are modeled in the groundwater model. This unique approach can be used to substitute the MODFLOW-MT3DMS groundwater models in their coupling with MODSIM by not only replacing the groundwater model in the modeled area, but also predicting stream-aquifer interactions (1) for areas where groundwater models are unavailable, and (2) for management alternatives not included in the groundwater model. This approach allows conjunctive use basin scale modeling in the Lower Arkansas River Valley, where detailed groundwater models are not available for the entire basin.

The trained ANNs for the two reservoir operations do not show significant differences in predictions performance. It can be concluded that using more than two weeks of memory only marginally improve, if any, the prediction performance and can cause larger error-accumulated effects on the simulation.

Uncertainty Discussion

The developed ANN attempts to reproduce the MODFLOW-MT3DMS modeled stream-aquifer interaction, and therefore inherits all the uncertainty associated with the groundwater model in predicting the stream-aquifer interaction. Additional uncertainty is introduced by the implemented methodology, with the main sources of uncertainty in the ANN representation of the groundwater model being: (1) discrepancies between the values calculated for the regional-scale groundwater model and the homologous variables computed for ANN training (e.g., pumping, precipitation, canal seepage, recharge), (2) the spatial aggregation of variables and predictions, (3) differences in stresses due to shifts in weekly time steps between MODSIM and MODFLOW-MT3DMS (after the second year of simulation there is a shift between stresses and responses), and (4) the ANN model structure and error goals. Statistics on the reports included in Appendix II give a sense of the uncertainty associated with the ANN predictions.

Non-Modeled Areas Prediction

In addition to uncertainty in the modeled area, additional uncertainty is introduced while using the learned relationships outside the groundwater modeled area. Since the ANN is trained based on conditions that might be unique for the modeled area, such as aquifer characteristics, the assumption for basin scale application is that those unique characteristics are not going to significantly change in the non-modeled area to as to invalidate the predictions. Based on this assumption, the uncertainty is potentially going to increase with distance between the modeled area and the prediction location. John Martin Reservoir, located downstream of the groundwater modeled region, could influence system characteristics in the downstream area by increasing the prediction errors in this region. Assessment of these errors is only possible if there are data available to compare against the prediction. Statistics could be derived removing by individual grouping areas from training and assessing the prediction errors. Since the modeled grouping areas are all clustered together, however, the ultimate error evaluation is only possible when the calibrated transient model for the downstream area becomes available. Analysis and validation of the baseline predictions outside the modeled area is performed using surface water measurements in the Arkansas River basin scale modeling presented in Chapter 7.

Limitations and Additional Thoughts

Since the ANN combines the effects of the explanatory variables to predict the stream-aquifer system response, these predictions are biased by the hydrologic conditions for which the network was trained. In this case, the groundwater modeling encompasses the transition from an extremely wet period to a drier period. Therefore, the predictions are expected to behave in a similar fashion to the aquifer response under these conditions. The magnitude and sequence of the explanatory variables will dictate, to some degree, the prediction but it must be noted that the underlying behavior was a wet-to-dry transition.

RESERVOIR WATER QUALITY TRANSPORT MODELING

Valerie (2001) analyzed the regional transport of water and dissolved constituents through heavily regulated river systems, showing that the system is influenced in varying extents by the presence of reservoirs. Transport of chemicals through heavily regulated river systems is influenced by the interaction between the chemistry of the inflowing water and processes occurring within reservoirs. These processes are determined by (1) reduced water velocities and concomitant loss of sediment, (2) the timing and extent of thermal stratification, (3) advective transport characteristics and the occurrence of density currents, (4) the location of outflow portals, (5) hydraulic retention times, and (6) biological activity within the reservoir (Thornton et al. 1990). Additionally, concentrations of dissolved materials from evaporation is a major factor in controlling salinity in reservoirs due to increased residence time, temperature, and surface area, especially in arid basins. Accurate basin scale water quality modeling with in-line reservoirs requires the modeling of these processes in the reservoir. Modeling the complex processes in the reservoir, using the most up-to-date tools such as CE-QUAL-2E (Wells 2000b), requires a large number of input variables and detailed information on the reservoir system that is not readily available. A need for a simplified but robust approach is identified to assist the basin scale salinity modeling efforts of this research project.

John Martin Reservoir Salt Transport Analysis

Salt transport in John Martin Reservoir is analyzed based on available historical data from November 1985 to October 2004. The two stations providing water quality data on inflows include ARKLASCO and PURLASCO, with the ARKJMRCO station providing continuous records for water quality of the reservoir outflows. The ARKLASCO station is located on the Arkansas River, and it records data regularly at constant time intervals, whereas the PURLASCO station is located on the Purgatoire River and has intermittent total dissolved solid samples. The TDS samples at PURLASCO are used to fit an equation to represent concentrations as a function of the flow during periods where no data are available. The data indicate a wide range of concentrations for a given flow (especially low flows), with the best fit equation from the set of tested equations determined to be logarithmic (Figure 4.27). Since this equation shows an extremely low portion of the variance as explained by flow, indication that there exist additional factors that influence concentration measurements at the station. For this demonstration, the concentrations are assumed to result from the fitted logarithmic equation.

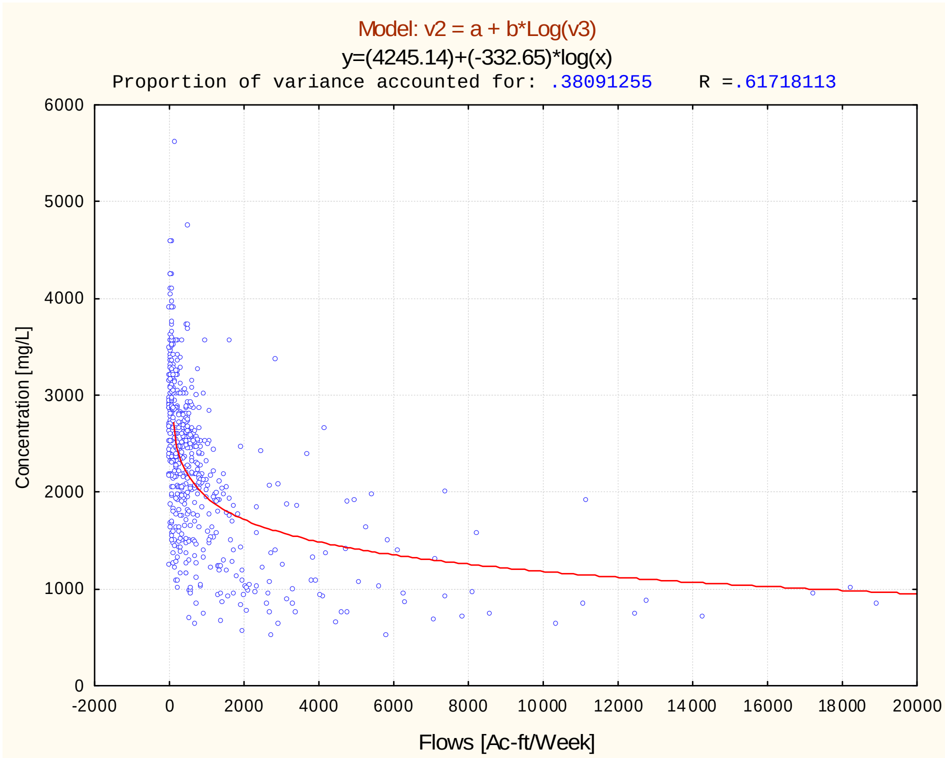


Figure 4.27 – Flow vs. Concentration at Purgatoire River near Las Animas

Results from Valerie (2001) indicate that sites located above the major reservoirs on the Colorado River and the Rio Grande have different characteristic and distinct chemical signatures, with passage through the reservoirs acting to merge and collapse these individual signatures. Figure 4.28 shows a comparison of the combined salt load coming into the reservoir from the Arkansas and Purgatoire Rivers and load coming out the reservoir to the Arkansas River. This figure shows a seasonal behavior, in which during the irrigating season (peak flows), the salt mass released from the reservoir is greater than the salt mass coming into the reservoir; conversely, the salt mass conveyed downstream during the non irrigation season is smaller than the mass load entering the reservoir. In addition to the increased evaporation occurring during the summer months, it appears that the high flow season increases the reservoir layer mixing and therefore contributes higher salt mass loadings downstream. The aggregated annual salt transport is summarized in Figure 4.29, showing that the net salt mass entering/leaving the reservoir changes from year to year. Figure 4.30 shows that the addition of salt to the reservoir is correlated with the net increase in storage (positive *TotalFlowIn*).

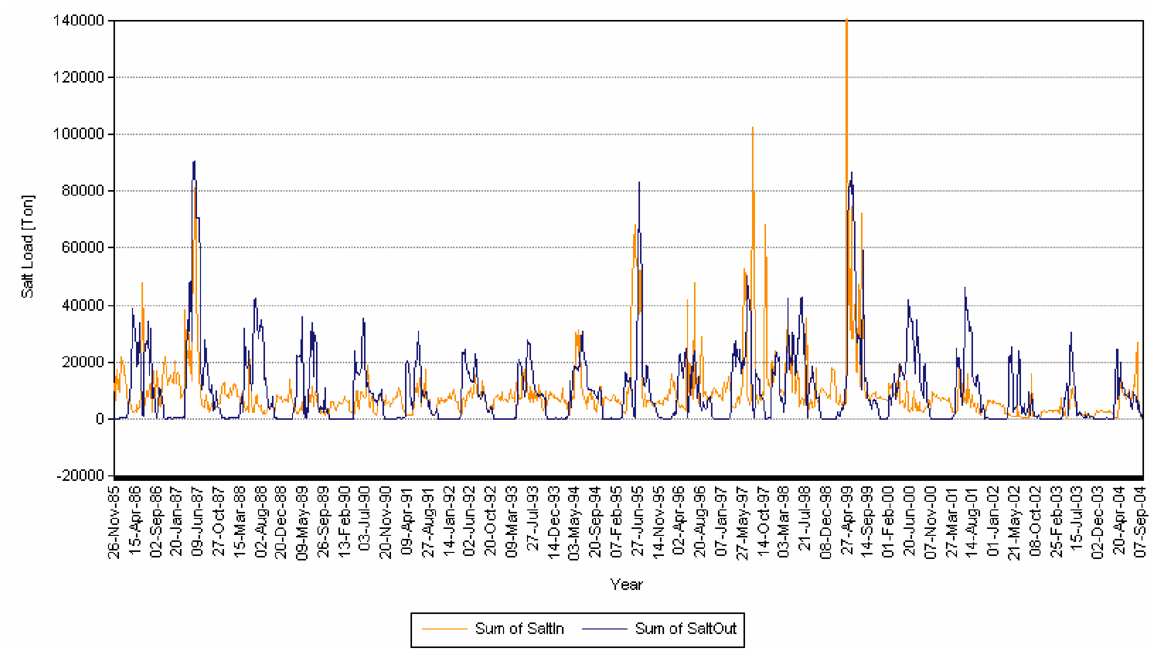


Figure 4.28 – Salt transport in John Martin Reservoir

Analysis of salinity concentrations reveals a different behavior than the salt loadings. Salinity concentrations of the reservoir inflows fluctuate more than the outflow concentration, indicating that the reservoir performs a smoothing effect on the concentrations. Figure 4.31 shows the Arkansas and Purgatoire inflows concentrations, along with the John Martin Reservoir outflow concentrations. In general, the Purgatoire concentrations are larger than the Arkansas concentrations, with corresponding averages of

2608.08 mg/L and 2168.56 mg/L respectively. Concentrations of the outflows are less variable and average 1861.83 mg/L, showing an overall reduction in concentration for the analyzed period.

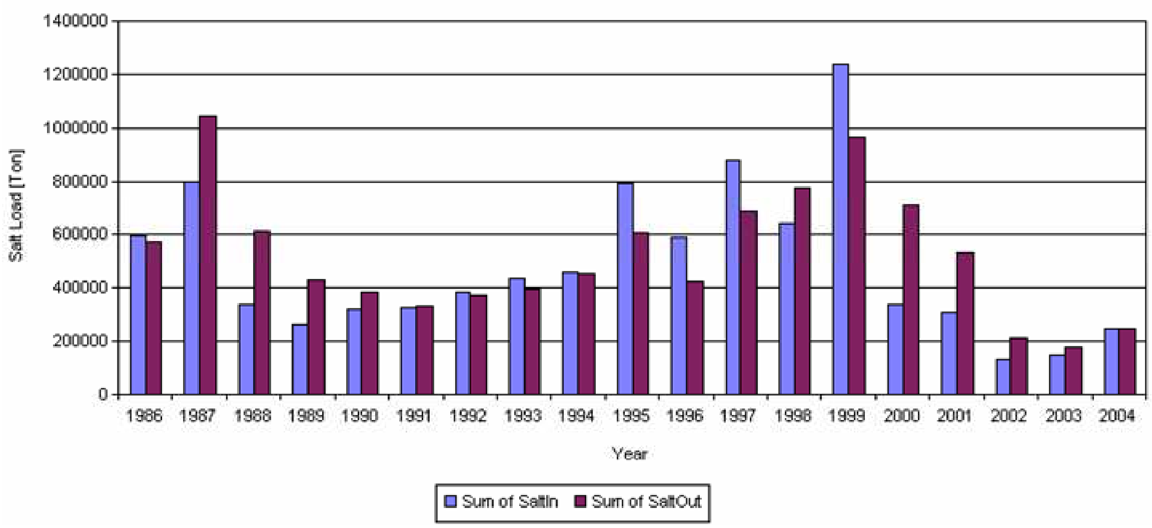


Figure 4.29 – Annual Salt Load In/Out John Martin Reservoir

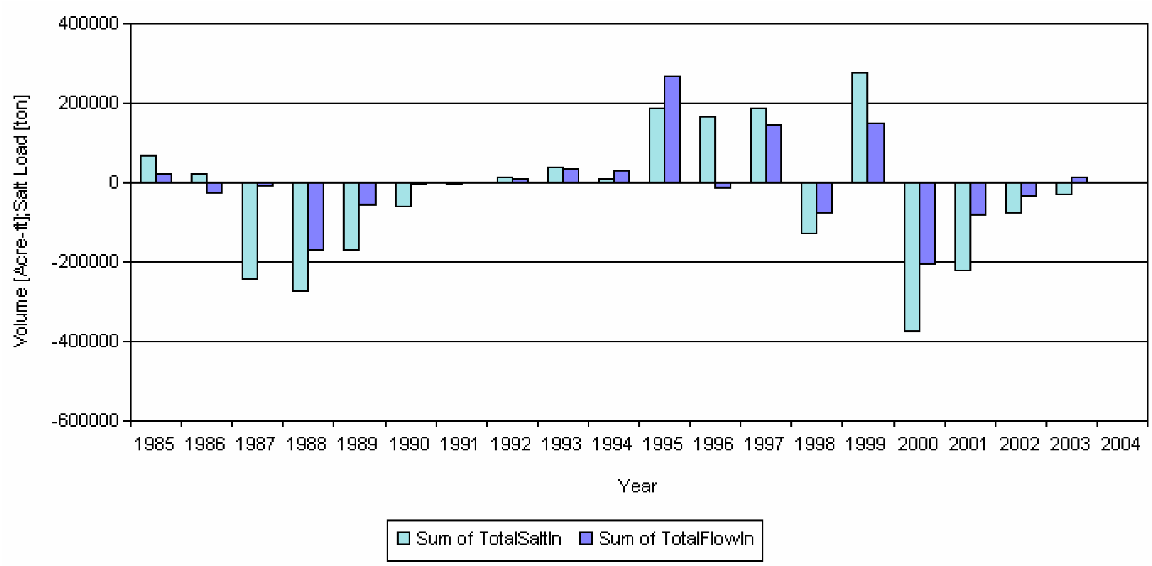


Figure 4.30 – John Martin net annual change in storage and salt mass

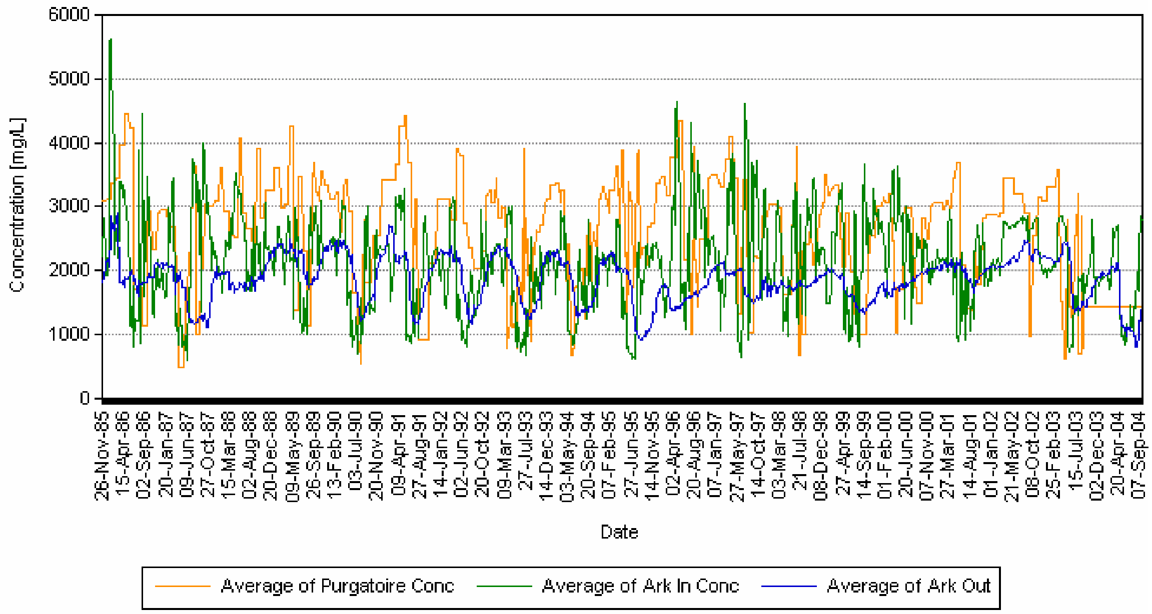


Figure 4.31– John Martin Reservoir concentration In\Out

Concentrations at the outlet of the reservoir are consistently lower during the high flow seasons, and inversely correlated to the salt loadings. Observing the historical data, there is a clear tendency for low concentrations to occur right after a high flow season has began. Low concentration values for the reservoir outflows show a strong lagged-relationship with peak discharges (Figure 4.32), pointing to a significant lagged correlation between them.

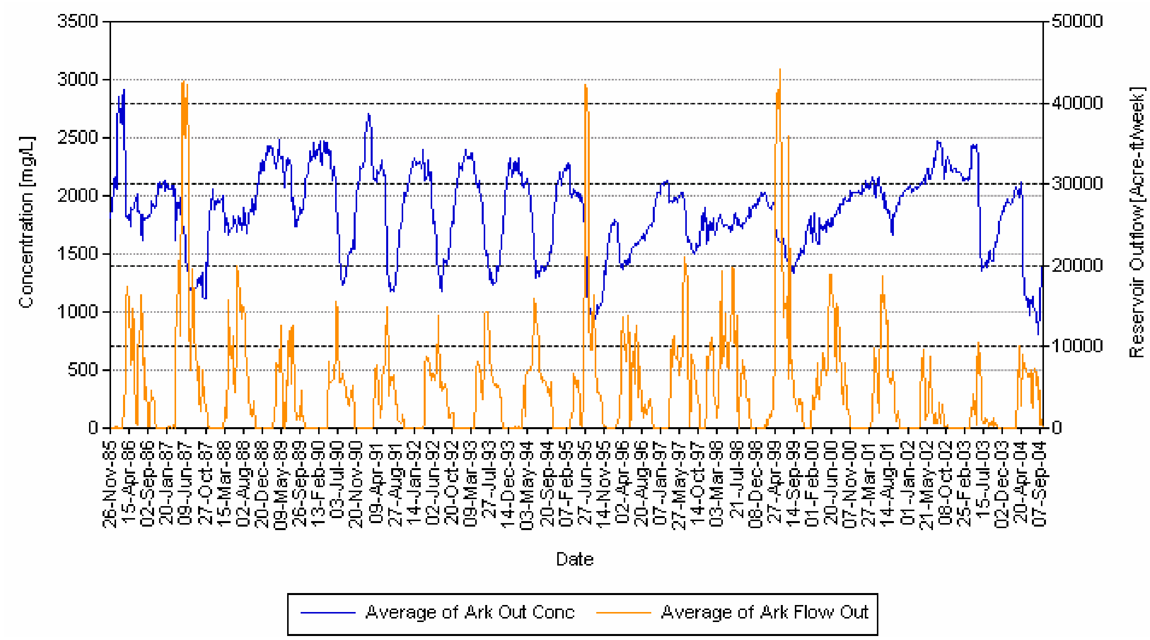


Figure 4.32 – John Martin Reservoir outflow and concentration

A cross-correlation analysis was performed between the measured variables at the inlets and outlet of the reservoir. The reservoir release (*ARKJMRCOSurfIn*) and the outflow concentration (*OUTPUTInRiver*) cross-correlation corroborated the observed strong correlation between these two variables (Figure 4.33-A). The statistical analysis also illustrated a strong autocorrelation of the reservoir outlet concentrations. The results show significant correlations up to 15 lags (weeks) based on values higher than 0.7 up to the first 6 lags (Figure 4.33-B).

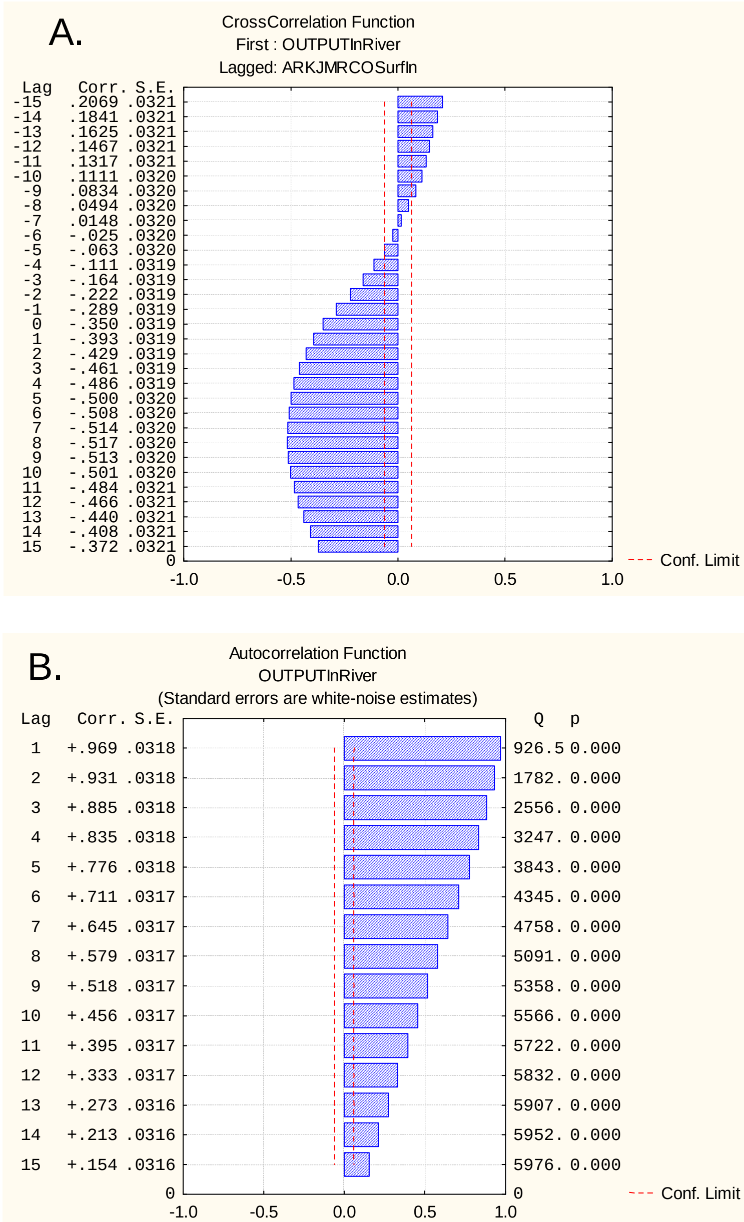


Figure 4.33 – Outlet variables correlations

This analysis reiterates the necessity for modeling salt transport through John Martin Reservoir to better quantify water quality effects of management alternatives in the LARV downstream of John Martin.

Approach

An ANN-based approach is proposed to model salt transport through John Martin Reservoir, which avoids simplifications that would relax the data requirements and modeling complexity. The ANN-based modeling approach is rooted in measured reservoir states, inputs and outputs by attempting to match the historical overall results based on a full range of processes occurring in the reservoir. The ANN model is designed to describe average weekly output concentrations as a function of the reservoir inputs (i.e., flow and salinity loadings), reservoir releases and changes in storage. The weekly explanatory variables for the outlet concentrations are the flows and concentrations at the Arkansas and Purgatoire Rivers near Las Animas (Colorado), the beginning and ending reservoir storage for each simulated week, and the corresponding reservoir release. The explanatory variables include selected previous week inputs and outputs as time-varying “memory”, taking advantage of the observed lagged correlation/autocorrelation.

ANN Development

A training dataset is created from the weekly processing of the available measured data using the fitted curve to represent concentrations in the Purgatoire River. The dataset is created processing measured values for the weeks in three periods: (1) from 01/1980 to 03/05/1999, (2) from 03/05/1999 to 08/2003, and (3) the weeks from 04/01/99 to 10/14/01 (weeks shifted from data group 2); the cases are flagged with a group identifier using the field *RetHydroID*. The first and second groups are to be used for ANN training and validation; the third group will test the NN performance on a shifted weekly data set to observe sensibility to the weekly groping of variables.

ANN Database Management Tool

A customized version of the *ANN Database Management* tool is implemented for John Martin salt transport modeling. The MS-ACCESS user interfaces are used to generate the MATLAB ANN training/testing files. Figure 4.34 shows the preferences for building this ANN training dataset. Three previous weeks of outlet concentrations, reservoir ending storage and Arkansas flow entering the reservoir are used to simulate a recurrent effect allowing the neural net to use time-varying patterns in the predictions. Using these preferences, the Database Management Tool creates the set of files for MATLAB ANN training.

ANN Training

Relationships between the explanatory variables and the outlet concentration were searched training different types of ANN and performing multi-variable regression analysis, with the Elman NN producing the most robust and consistent set of predictions. The network is trained using a time-sorted dataset of 920 cases (2/3 of the training dataset) and tested/validated with the remaining 307 cases. The testing/validation dataset mainly corresponds to cases in the second group of the dataset (flagged with *RetHydroID* = 2). The training was performed using the *River GeoDSS* custom MATLAB Training Tools with 10 training sessions of 15 trainings each. For each training event, the network is built with a single hidden layer with random number of neurons (maximum 45 neurons) and random transfer functions. For each session, the representative network is the one with best combined performance (Equation 4.1). In agreement with the autocorrelation analysis, the inclusion of previous outlet concentrations is demonstrated to positively influence the predictions. The Elman Network training time for these sessions is much longer than similar training of the standard feed forward ANN. From all the sessions, the best combined performance ANN has 32 neurons and two tan-sigmod transfer functions. Figure 4.35 shows the ANN performance during training and testing/validation for the best training event.

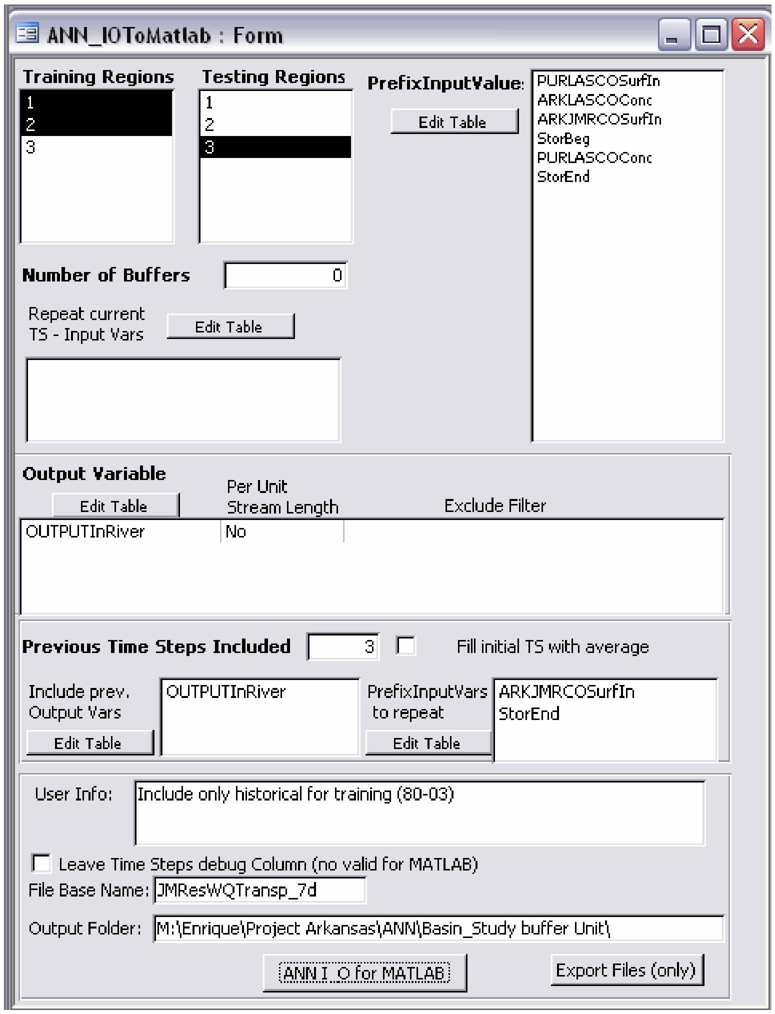


Figure 4.34 – John Martin Reservoir salt transport ANN training dataset preferences interface

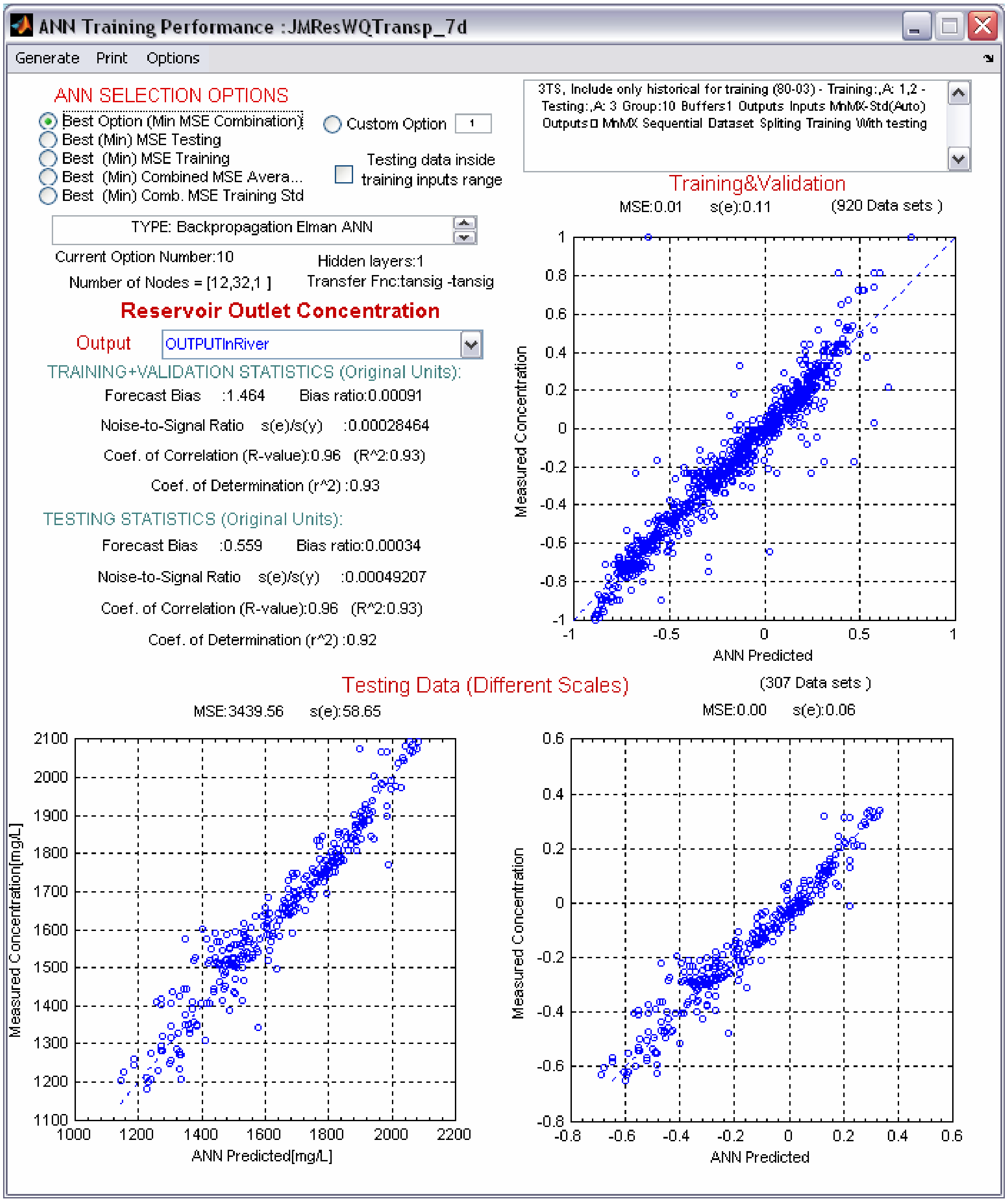


Figure 4.35 – ANN training and validation performance summary for the John Martin Reservoir salt transport

The predictions show a coefficient of determination of 0.93 and 0.92 in training and testing respectively, with a testing prediction root mean squared error of 58.5 mg/L, indicating a good ability to predict the John Martin outlet concentration. The predictions were tested in the same original test period with the modeled weeks shifted to observe the explanatory variables sensibility. Figure 4.36 shows the MATLAB-based performance report in the modeled weeks from 4/1/1999 and 10/14/2001. The calculated coefficient of determination (0.93) and the root mean squared error (58.16 mg/L) indicate little sensitivity to variations in the weekly-grouped explanatory variables. Application of the Elman ANN with the generated training dataset combines the dynamic feedback from the previous hidden layer output (Elman ANN) with a static feedback from the previous outputs as included in the training dataset (similar to a Jordan ANN (Jordan 1990)).

Reservoir Outlet Concentration Modeling Analysis

Most of the approaches reviewed in the literature deal with detailed hydrodynamic and chemical reservoir modeling (Willey et al. 1996; Bicknell and National Exposure Research Laboratory (U.S.) 2001;). The newest tendencies in reservoir modeling development point to an increase in detail and dimensionality (Hamilton and Schladow 1997). The input-output mapping approach allows development of robust tools for modeling the highly complex processes in reservoirs, as demonstrated in the application of ANNs management of eutrophicacion in lakes (Karul et al. 1999;Walter et al. 2001). The methodology introduced herein to model reservoir outlet concentrations provides a unique approach, with the explanatory variables selected to model salt transport appearing to provide the ANN with sufficient information to mimic the historical concentrations of the reservoir releases. The model application to John Martin Reservoir transport modeling shows robust predictions outside of the periods where the neural network was trained, giving confidence in its application as predictor during management scenarios modeling (events not available in the historical dataset).

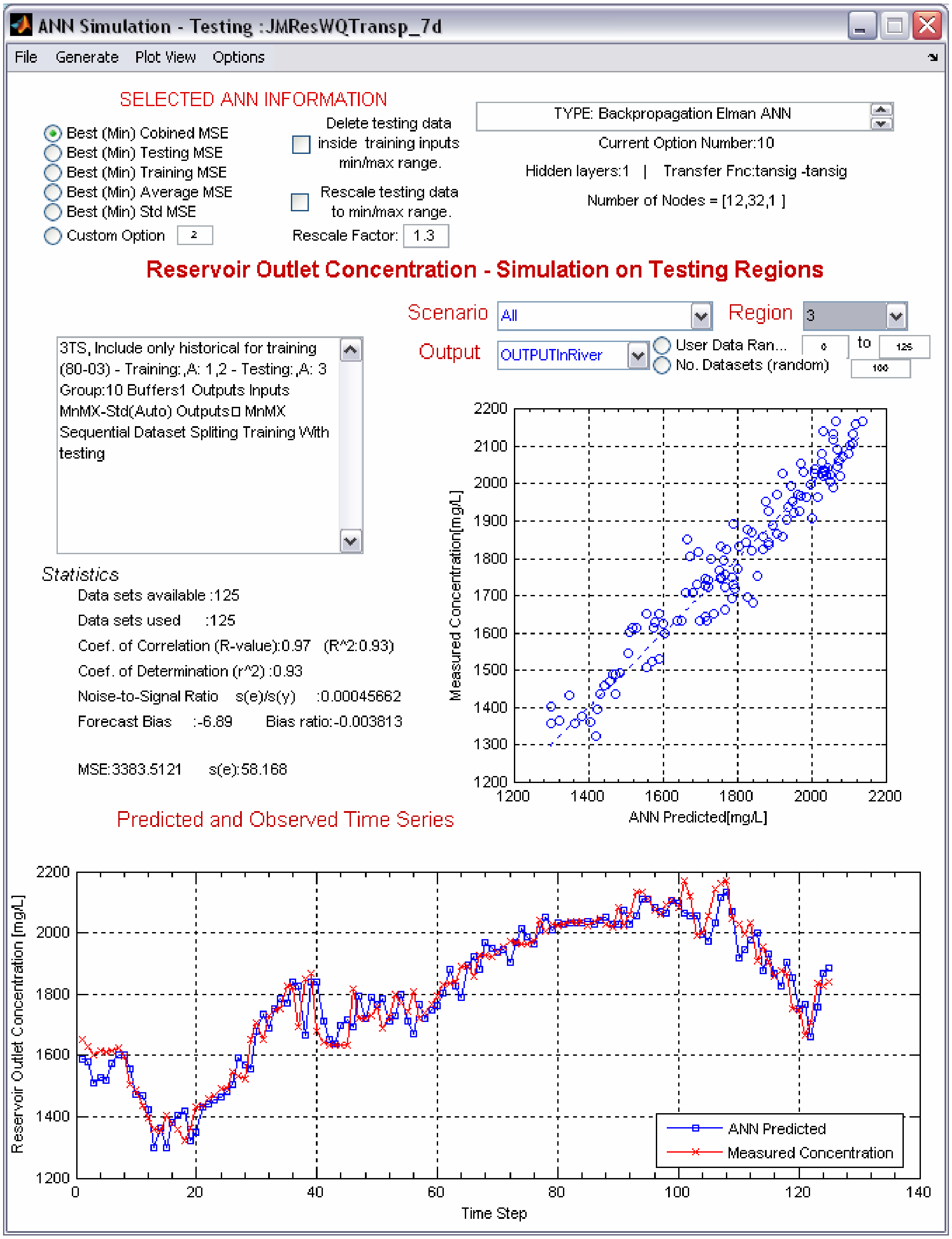


Figure 4.36 – John Martin Reservoir Outlet concentration ANN prediction in testing

The developed model finds relationships between reservoir inputs, state and their outcome in time, making the model utilization quick and suitable for integration in basin scale management decision tools. Although the ANN models developed with this methodology are expected to be specific for the reservoir they are developed for, it is believed that the application of the methodology is suitable for modeling other reservoirs and could be extended to model other water quality parameters.

THE SIMULATION CHALLENGE

The ANN-based models introduced in this chapter were developed with the idea of being integrated into a comprehensive basin scale decision support tool. Differences between the training and simulation dataset are expected to affect the performance during simulation. An anticipated challenge of this integration is the use of "imperfect" explanatory variables. The training was performed with datasets that use explanatory variables from "perfect" previous predictions (prediction with no error). A sequential simulation with the trained ANN will use explanatory variables with an implicit error from previous predictions uncertainty; in addition to the new uncertainty added using this simulation dataset, the risk is that the error accumulates making the deviation on the recurrent explanatory variables to grow incontrollable during the simulation progress, resulting in unrealistic ANN predictions. In an attempt to minimize the discussed effect, the selection of the best ANN to predict the modeled phenomena included testing of performance with sequential previous outputs. Neural networks that showed high sensitivity to the recurrent explanatory variables error accumulation in sequential simulation were discarded. The number of previous time steps explanatory variables was set to the minimum without compromising performance. The rationale is that the more previous time step outputs included the more uncertainty in the explanatory variables.

Another ANN simulation challenge deals with the initial recurrent explanatory variables. In the stream-aquifer modeling, the NNs will predict return flow and concentration in areas were there is not groundwater modeling available and periods outside the groundwater modeled period. Therefore, in these cases there are not pre-existing output variables for building the explanatory variables set. The recurrent explanatory variables for the initial time steps will influence the sequence of predictions space; therefore in simulation, setting these variables could play a key roll in the successful implementation of the stream-aquifer interaction modeling.

Lastly, some explanatory variables might get adjusted during the basin scale simulation causing discrepancies between the training and the simulation. ANN inputs variables relying on flow, diversions, reservoir storage might deviate from their training counterpart. This is another example of issues that ought to be address while incorporating these ANN models in the *River GeoDSS* modeling system.