

**WSRC-TR-2000-00442**

# **Predicting Tritium Dynamics at the Savannah River Site Using Neurogenetic Models**

**James A. Bowers, Charles F. Murphy and F. Douglas Martin**

**October, 2000**

Prepared by:  
**Westinghouse Savannah River Company**  
**Savannah River Site**  
**Aiken, SC 29808**



---

Prepared for the U.S. Department of Energy Under  
Contract Number DE-AC09-96SR18500

**This document was prepared in conjunction with work accomplished under Contract No. DE-AC09-96SR18500 with the U.S. Department of Energy.**

#### **DISCLAIMER**

**This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.**

**This report has been reproduced directly from the best available copy.**

**Available for sale to the public, in paper, from: U.S. Department of Commerce, National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161**

**phone: (800) 553-6847**

**fax: (703) 605-6900**

**email: [orders@ntis.fedworld.gov](mailto:orders@ntis.fedworld.gov)**

**online ordering: <http://www.ntis.gov/support/index.html>**

**Available electronically at <http://www.doe.gov/bridge>**

**Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from: U.S. Department of Energy, Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831-0062**

**phone: (865)576-8401**

**fax: (865)576-5728**

**email: [reports@adonis.osti.gov](mailto:reports@adonis.osti.gov)**

**List of Tables**

Table 1. Normal quantiles metrics for soil temperature..... 9

Table 2. Moments of soil temperature distribution..... 9

Table 3. Test for normality of soil temperature distribution using Shapiro-Wilk W-Test..... 9

Table 4. Spearman-Rho nonparametric estimates of association between input and output variables for tritium models..... 10

Table 5. Summary of neural network types and their performance in predicting tritium mass transport in Fourmile Branch and the Savannah River..... 11

Table 6 Preliminary Artificial Neural Networks developed by the Genetic Algorithm process..... 11

**List of Figures**

Figure 1. Map of the Savannah River Site indicating tritium sampling stations used for modeling (WSRC-TR-98-00314. .... 5

Figure 2. Sampling duration time in days of the pooled sampling intervals used for tritium measurements at SRS streams and the Savannah River from January, 1990 through December, 1999..... 6

Figure 3. Distribution statistics for soil temperature variable (SAS/JMP)..... 8

Figure 4. Predicted and desired data records of the back propagation neural network model predicting tritium transport (Ci) at SRS station FM-6 in Fourmile Branch. .... 13

Figure 5. Predicted and desired values for the extra validation set of the back propagation neural network model predicting tritium transport (Ci) at SRS station FM-6 in Fourmile Branch. .... 13

Figure 6. Predicted and desired data records of the time delay neural network model predicting tritium transport (Ci) at SRS station FM-6 in Fourmile Branch. .... 14

Figure 7. Predicted and desired values for the extra validation set of the time delay neural network model predicting tritium transport (Ci) at SRS station FM-6 in Fourmile Branch. .... 14

Figure 8. Predicted and desired data records of the back propagation neural network model predicting tritium transport (Ci) at SRS station RM-120 in the Savannah River..... 15

Figure 9. Predicted and desired values for the extra validation set of the back propagation neural network model predicting tritium transport (Ci) at SRS station RM-120 in the Savannah River..... 16

Figure 10. Predicted and desired data records of the time delay neural network model predicting tritium transport (Ci) at SRS station RM-120 in the Savannah River..... 16

Figure 11. Predicted and desired values for the extra validation set of the time delay neural network model predicting tritium transport (Ci) at SRS station RM-120 in the Savannah River..... 17

Figure 12. Predicted and desired data records of the back propagation neural network model predicting tritium concentrations (pCi/L) at SRS station FM-6 in Fourmile Branch..... 17

Figure 13. Predicted and desired values for the extra validation set of the back propagation neural network model predicting tritium concentrations (pCi/L) at SRS station FM-6 in Fourmile Branch..... 18

Figure 14. Predicted and desired data records of the back propagation neural network model predicting tritium concentrations (pCi/L) at SRS station FM-6 in Fourmile Branch covering a period from approximately January 1991 through December 1999. .... 19

Figure 15. Predicted and desired data records of the back propagation neural network model predicting tritium concentrations (pCi/L) at SRS station FM-6 in Fourmile Branch covering a period from approximately January 1991 through December 1999. .... 19

## Executive Summary

The purpose of this investigation was to assess a new approach to predict tritium concentrations in SRS streams and the Savannah River both within a drainage construct and at a sitewide scale. The ability to simulate tritium behavior at these varying scales would offer SRS monitoring and remediation efforts an additional capability currently unavailable. The approach is a powerful new technology, neurogenetic models, which are Artificial Neural Networks (ANN) optimized by Genetic Algorithms (GA). After a lengthy assembly of the tritium data base, climatic data and stream flow rates for SRS the study focused on the most important tritium-loading stream at the SRS, Fourmile Branch and in the Savannah River. Two tritium sampling stations selected, in Fourmile Branch just prior to entering the Savannah River swamp and Station RM-120 in the river immediately below SRS. After preprocessing the data for modeling runs the data set was comprised of 196 data records covering approximately the years 1991 through 1999. Our ultimate goal was to provide environmental managers and professionals with a modeling tool to forecast tritium mass transport from SRS and tritium concentrations in streams and the river.

Initial tests of data normality, using the Shapiro-Wilk W-Test, indicated that the tritium, climatic and stream flow data sets were not normal. Therefore all performance testing of the GA-ANN models applied the nonparametric Spearman-Rho degree of association coefficients. This same test was then used to create association coefficient matrices to estimate all of the possible degrees of interrelationships in the complete data set to guide modeling efforts. An initial set of GA-ANNs was tested and developed to forecast tritium mass transport in Fourmile Branch and the Savannah River. Three types of ANNs resulted, the Backward Propagation neural network (BP), the Generalized Regression Neural Network (GRNN) and the Time Delay Neural Network (TDNN) for time series modeling. All of these neural architectures successfully predicted tritium mass transport or tritium concentration and the stream and river stations. Most importantly, these results indicated that GA-ANNs could be developed that would forecast tritium behavior without the need for tritium estimates as input variables. The models could forecast tritium dynamics using only climatic and stream flow data. After evaluating over 5000 neural models, each trained, validated and tested on the data, five models were selected for further evaluation and validation. A BP network and a TDNN were selected to forecast tritium mass transport at Station FM-6. Nonparametric correlations of fit between the observed and predicted mass transport were 0.94 and 0.93 respectively for both models. An extra validation model run for both models was performed giving 0.91 for the BP and 0.92 for the TDNN. For Station RM-120 in the Savannah River a BP model gave a nonparametric fit equal to 0.78, while a TDNN reached a 0.83 level of fit to mass transport. The extra validation results gave values of 0.86 for the TD and 0.85 for the TDNN. The final model selected was a BP model for tritium concentration forecasting in Fourmile Branch. Nonparametric degrees of fit between the observed and predicted tritium concentrations were 0.86 and 0.83 respectively for the tested model and the extra validation check. The above findings indicate that GA-optimized ANNs can be applied for the prediction of tritium in SRS streams and the Savannah River. Especially significant is the GA-ANNs competence to forecast tritium behavior based solely on climatic and stream flow information.

## Introduction

Tritium releases from the Savannah River Site (SRS) have been monitored extensively since 1960 and due to operational changes at the SRS have decreased significantly during the past four decades. However, tritium releases per year remain in the tens of thousands of curies per year (WSRC 1998). Much of the current releases originate from groundwater movement, separation and reactor seepage basins into primarily Fourmile Branch Creek and Steel Creek. One of the most important of these sources for tritium release at the SRS is the old F- Area effluent ditch and F-Area engineered effluent ditch which meet to form a single stream and eventually flows into Fourmile Branch. Remediation efforts of these tritium releases are expensive and difficult due to the behavior of 'tritiated water' in the environment. Unlike other contaminants there are no methods to remove tritiated water from the environment, only methods that either store the water or redirect its path through the hydrologic cycle which work by providing increased decay time. One aspect of the tritium issue at the SRS is the ability to predict its movement within SRS streams and eventual release into the Savannah River. Although methods exist to estimate net releases into the Savannah River and model tritium dynamics at certain localized areas of the SRS, predicting tritium behavior within large drainages or sitewide is again costly and difficult. Here we propose a new approach to predict tritium concentrations in SRS streams and the Savannah River both within a drainage construct and at a sitewide scale. The ability to simulate tritium behavior at these varying scales and at a sitewide scale would offer the SRS remediation efforts an additional capability currently unavailable.

Our modeling approach uses the special capabilities of an Artificial Neural Network (ANN). A neural network is a massively parallel distributed computer processor that has a natural propensity for storing experiential knowledge and making it available for use (Haykin 1994) similar to the way the nervous system functions in a living organism. Such networks have several features that make them well suited for predictive engines. Neural networks have a remarkable ability to derive meaning from complicated or imprecise data (Wasserman 1993). ANNs also can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques (Masters 1993). A recent advance in the use of ANNs has been the application of Genetic Algorithms (GA) in designing ANNs. Designing ANNs has been mainly a trial and error process based on the mathematical properties of networks where neural layer creation, neuron activation functions and the number of hidden neural layers designated by the modeler. Genetic algorithms are very powerful optimization routines originally developed to simulate evolutionary processes in animal populations (Davis 1991) drawing from the fields of population genetics and Darwinian natural selection. Their application to optimizing ANN structure and function has significantly improved the ANN design process allowing users to rapidly create literally hundreds of ANNs for problem testing. The combined of modeling processes of GA optimized ANNs are termed neurogenetic models.

In the environmental disciplines neural networks are rapidly becoming a standard predicting method in watershed science. Two examples will suffice. Using regional rainfall data and a grid of water level recorders the local governments in the Tagliamento River watershed have developed a neural network to forecast regional flooding in the Tagliamento River basin (Campolo et al. 1999). Environment Canada, the Canadian counterpart to our USEPA, now uses a neural network to forecast discharges of dissolved organic carbon and nitrogen into the North Atlantic Ocean through the Saint Lawrence River from 15 upstream river basins (Clair and Ehrman 1996). Both examples support Shamseldin's (1997) conclusion that neural networks outperform traditional models where explicit knowledge of hydrological processes are unknown or not needed for forecasting. Our own experience using a neural network to predict suspended solids in Mill Creek, SRS indicate the utility of the method (Bowers and Shedrow 1999). The selected input variables were local precipitation, stream flow rates and turbidity for the prediction of suspended solids. A single hidden-layer feedforward neural network (logistic activation function) using the Levenberg-Marquardt backpropagation learning algorithm (Hagan and Menhaj 1994) successfully simulated the suspended solids response to the input pattern. Linear regression of the predicted values versus the actual test values which gave an  $R^2 \geq 0.94$ , was used to assess the difference between actual and predicted concentrations (Masters 1993). These network simulations demonstrate the feasibility of using SRS climate and aquatic monitoring information for predicting stream behavior.

In anticipation of this proposal a single abbreviated pilot experiment was performed to indicate the feasibility of our approach to predicting tritium concentrations in SRS streams. Taking a very small data set (< 100 data points) from four sampling stations in Fourmile Creek and using the same neural network model from a Mill Creek study, a network was trained on the information pattern of 9 water quality parameters (pH, dissolved oxygen, conductivity, alkalinity, suspended solids, temperature, total dissolved solids, total solids, and turbidity) to predict tritium concentrations in Fourmile Creek. Again linear regression of the predicted values versus the actual test values was used to assess network performance (Masters 1993). A regression  $R^2 \geq 0.83$  indicated that a neural network might well predict tritium behavior in SRS watersheds.

The goal of this effort is to simulate the tritium concentrations in SRS streams and the Savannah River using a much more complex network architecture incorporating climatic variables, stream flows and selected water quality parameters indicative of water mass behavior in SRS streams and the river. The resulting neural network would learn, on a sitewide scale, the patterns of behavior of all of the above variables and predict tritium concentrations at several selected locations onsite and in the river. Input variables would be screened for covariance occurrences using principal component analysis (orthogonal rotation methods) to collapse the data input matrices (Masters 1993). Several neural designs are being tested for performance based on the following features: linear, logistic and hyperbolic tangent activation functions, the number of input-accepting neurons, multilayer configurations (hidden layers), learning algorithms (conjugate-gradient methods, simulated annealing and genetic algorithms) and neural weighting initializations (different randomization algorithms in C++ recipes). Specifically, our main objective was to develop a neurogenetic model that would predict the movement of tritium in SRS streams and the Savannah River. Neural networks learn data streams, which in this situation, was planned to span a decade of tritium results from 1990 through 1999. This interval would cover an adequate interval of time to represent tritium dynamics and offer enough information for the ANNs to train.

### **Approach to the Data**

There are five main drainage basins on SRS. The five streams that originate on, or pass through SRS before entering the Savannah River, are Upper Three Runs, Beaver Dam Creek, Fourmile Branch, Steel Creek, and Lower Three Runs (Figure 1). It was decided early in the data collection and preprocessing stage of the project to focus our time on the most important tritium-loading stream at the SRS which is Fourmile Branch (WSRC 1998). Included in the data collection and synthesis were several Savannah River locations both upstream and downstream from the SRS.

Tritium data at the SRS is monitored in a fashion designed solely to estimate the mass of tritium released from the site via the streams into the Savannah River. Sample collections shown in Figure 1 indicate the sitewide positioning of all of the sampling stations. The samples themselves are physically pooled samples collecting water for a period of several days with 7 and 14 day intervals being the most common (WSRC 1998). Tritium concentrations in Bq/L and pCi/L are then estimated. Concurrently, stream flow measurements are taken estimating the total volume of water, in liters, passing through the stream during each sample collection period. Knowing the concentration of tritium and the total volume of water for each time interval then allows an estimate of total tritium mass in curies moving past the sampling point. The tritium data was received from the EMS-EMCAP database in the form of ASCII text files covering the period from January 1990 through December 1999.

Flow data from the SRS streams and the Savannah River were collected from the United States Geological Survey (USGS) South Carolina NWIS-W Data Retrieval Center internet site (<http://waterdata.usgs.gov/nwis-w/SC/>). The data is grouped by USGS station number and the flow data in cubic feet per second as daily averages. The data is directly downloaded as tab-delimited ASCII text files. The flow data collected covered the period from January 1990 through December 1999.

Climatic data was assembled from the SRTC database in the Nonproliferation Technologies Section (WSRC 1999). This database consists of hourly, daily and monthly recordings of precipitation, soil evaporation rates, soil temperature, wind speed and direction at a variety of locations at the SRS. For this study daily mean precipitation was used at stations 773-A and 200-F. Soil temperature and evaporation rates were used from the 773-A location. Data from January 1990 through December 1999 was received in the form of an ASCII text file from the SRTC-NTS database.

### **Data assembly and Preprocessing**

All of the tritium, climatic and stream flow data files were first assembled into our HOPS<sup>R</sup> data engine (Bowers et al. 1996) and processed for loading into the neurogenetic modeling application (BioComp Systems 2000) running on a personal computer using the Windows<sup>R</sup> NT4.0 operating system.

When beginning to assemble the tritium concentration data files by station, it was immediately apparent that the pooled sampling technique required special modification for statistical and modeling analyses. The pooled sampling intervals vary both within each station through time and between each station through time. Therefore, direct data assembly by station through the desired time frame from January, 1990 through December 1999 was not possible. Figure 2 below indicates the boundary condition of the unevenly spaced pooled sampling intervals. Notice that there are four prevalent sampling intervals, one day, 7 days, 14 days and 28 days. However, significant numbers deviate from this pattern with intervals lasting to a maximum of 42 days. In order to assemble the data in a single uniform time pattern Julian time was adopted in the form of decade Julian time spanning 1 through 3652 days covering the desired time period. Furthermore, in order to exactly sequence all of the data to a common time frame the Fourmile Branch sampling station, named FM-6, was chosen because it is the last creek station before entering the river and had the most common time interval sequence compared to other major stream and river sampling stations. This time interval sequence ranged mainly from 12 to 14 days. All of the data files used in these experiments were synchronized by hand into this unique unevenly spaced time sequence. Matching data records out of sequence was accomplished by linear interpolation between time steps using averages. Although this process somewhat smoothes information through time, the variation through time of tritium concentrations, stream and river flow rates and climatic data was assumed to be adequate for our modeling purposes.

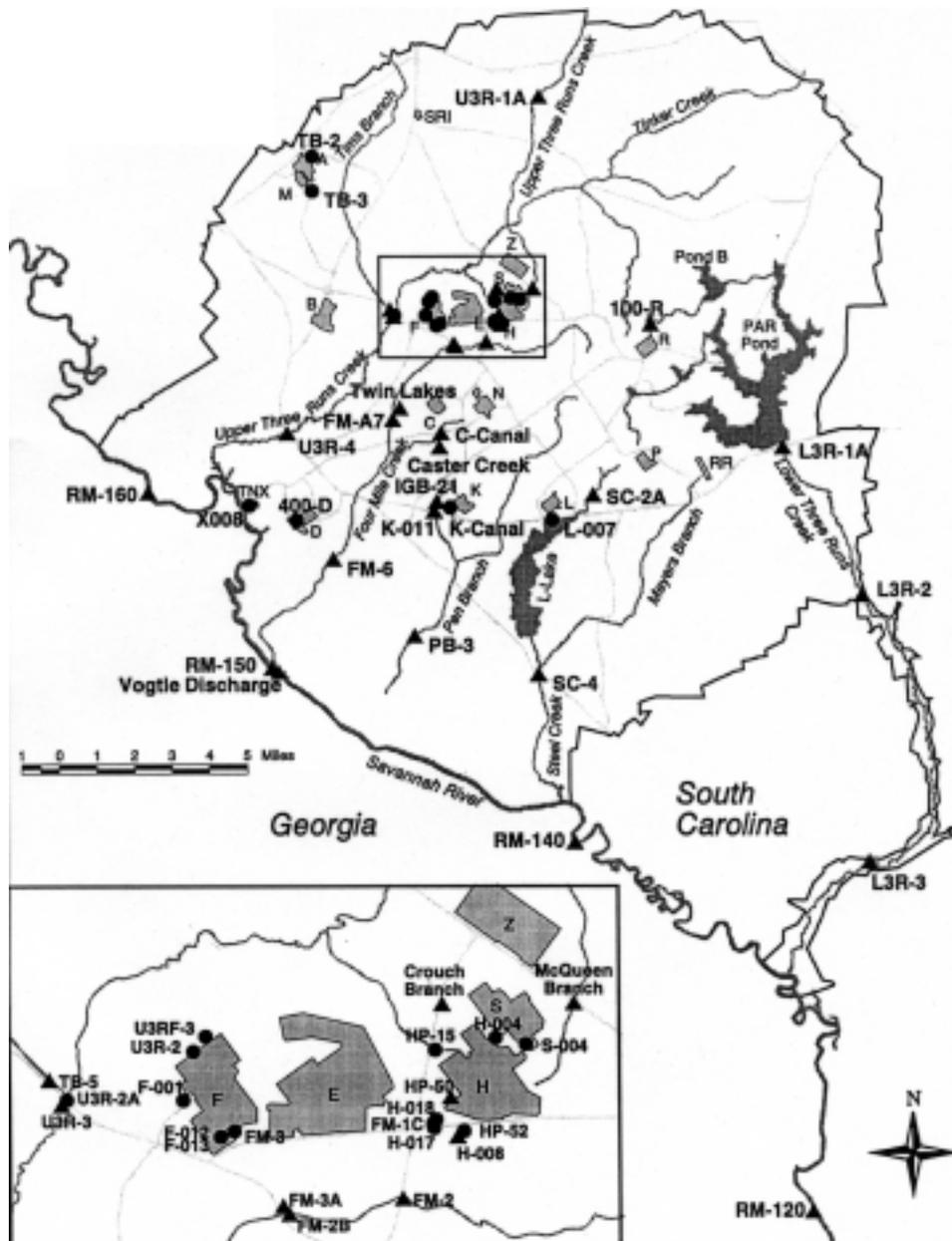
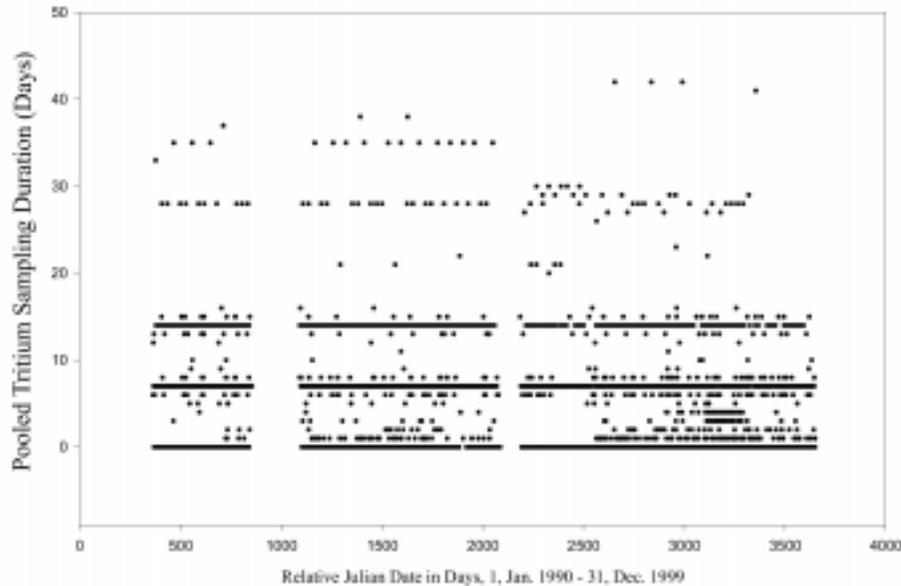


Figure 1. Map of the Savannah River Site indicating tritium sampling stations used for modeling (WSRC-TR-98-00314).

Figure 2. Sampling duration time in days of the pooled sampling intervals used for tritium measurements at SRS streams and the Savannah River from January, 1990 through December, 1999.



### Model development

The method for running GA-ANN application began by assembling all of the input variables into a comma-delimited ASCII text file which was the required data format for the neurogenetic model. After the input data file is loaded the next step was to designate which variables were to be input and output variables and which variables, if any, were to be discarded for the upcoming modeling experiment. Input data scaling was set between  $-1$  and  $1$ , and the output data scaling for ANN output neural transfer functions were scaled between  $.1$  and  $.9$  for only positive values. Output data scaling for ANN output neural transfer functions was scaled between  $.1$  and  $.9$  for both positive and negative values, and finally testing data was partitioned out from the raw data input file by a number of methods, but only a single method was used here. Twenty-five per cent of the data records were randomly taken for training, twenty-five per cent were taken for validation, while fifty per cent were used for model testing, this follows a conservative and often used protocol for data partitioning (Masters 1993).

The second major step is configuring both the GA and ANN subprograms for the specific application of the model. Is the model for classification, function approximation, clustering, diagnosis or a time-series experiment? Genetically optimized ANNs can be applied to any of these types of modeling constructs. Here only two modes were attempted, the function approximation and time-series constructs for the prediction of tritium behavior in SRS streams and the Savannah River. When setting up the function approximation mode, the classic Back Propagation (BP) ANN and the Generalized Regression Neural Network (GRNN) are choices for the GA.

Sometimes the GA was forced to optimize only one or the other types with the BP model being the most conservative in design. Time-series selection offers the GA subsystem to choose between either the Continuously Adaptive Time Neural Network (CATNN) or the Time Delay Neural Network (TDNN). The GA can be allowed to select during the GA optimization process. No other modes were tested in this project.

Prior to the beginning of the GA process one must set model structure and function in the population of ANNs that will experience selective breeding. The first step can be to limit the number of hidden neurons in a hidden neural layer and limit the number of hidden layers itself. Masters (1993) recommends limiting most neural nets to only a single hidden layer, which was done in all of the experiments performed here. Generally, multiple hidden layers are only recommended for very large and complicated systems having tens to hundreds of input layers and several output variables to predict. Limiting the number of nodes within a hidden layer inhibits an ANN during the learning process and thus can be used to force the ANN only to learn the general pattern of information flow and avoid the possibility of overfitting or learning minute detail or noise in the data stream. Several GA runs in the results were performed with the neural node limitation set at four nodes to inhibit the learning process in attempting a generalized solution. Selection of neural activation functions is also set at this step choosing between linear, tangent and logistic functions and allowing for these different functions to coexist within each neural layer. Setting the actual learning rate is mathematically equivalent the setting the slope on a conjugate gradient response surface. The higher the learning rate the faster the solution, but with the danger of fitting the ANN to a regional solution rather than a global solution. Here the learning rate was set slow at the price of slower training times, but forcing the process to a global solution. Learning momentum is likewise set slow to avoid overshooting the global minimum solution. Finally, but no less important, is initializing all of the neural activation function weights. Although a large literature exists on this neural model topic (see Hagan 1996), here a range was set to  $-0.3$  to  $0.3$  as most often recommended by experience (Masters 1993). Having set these fundamentals of the population of ANNs the GA process will need to know how many training passes of the information should be given to each ANN. Here the process was set to a minimum of 20 passes with a maximum of 50 having a jump mode where, if no greater performance is realized after 20 runs the ANN structure is finalized. The parameter settings for the breeding of ANNs is now ready.

Configuring the GA parameters controls the design and testing of the developed ANNs by literally controlling the natural selection process through a series of steps quite akin genetic manipulation in population genetics. Each ANN with its own set of unique features (number of hidden neural layers, neural activation functions, output neural functions and the number of neurons in each neural layer) is represented in binary number format expressing its genotype and in behavior as its phenotype (how the ANN learns and then predicts). Details in theory are introduced in Gruau (1993) and practical application may be read in Davis (1991) and BioComp (2000). The first step in the breeding or GA optimization process is to determine the size of the ANN population to reproduce. A common number to begin with is 30, but numbers as high as 100 are not uncommon. The larger the population size at the beginning gives the ANNs genetic diversity or as geneticists call it, gene pool. The tradeoff here is time, as the greater the population the longer the GA will take. Some GA runs have required up to 15 hours of CPU on a 400 Mhz CPU. Once the initial population size is configured the population must be initialized, in other words, created. There are several methods, completely random or unconstrained, 'locking' all of the input variables for the ANN to completely set the first neural input layer which forces the GA to optimize around this constraint, and to 'statistically seed' the population. Statistical seeding is basing on a correlation matrix generated to initially determine first-order relationships in the complete data set. Thus the GA is told to use (weight) certain input variables when they are highly correlated to the predicted output variables. Another feature commonly employed and used often in these experiments is to lock all of input variables on the very first GA generation as the best to judge all others. Once the population has been initialized the GA program must be set to reproduce after each generation. Here the GA first must kill off a percentage of the first generation. Natural selection for each generation can be performed by a random percentage (usually killing off 50 per cent) or by a very common method called the Roulette Wheel where selection is based on probabilities of a hypothetical roulette wheel. Mating is accomplished by 'Tail Swapping' or a 'Two-Cut Swap' which involves gene splicing at either one or two places. Refilling the population is done by cloning which exactly reproduces each genotype or a randomization of genotypes in each population.

Mutation rate is also set prior to the GA startup. Here random gene exchanges were set at 0.25 per cent, a lower number to gradually let the GA process select the ANNs. Higher mutation rates force the GA process to alter ANN genotypes more radically in search of optimal solutions. This is often used for preliminary GA runs. Finally, there are methods to 'weight' the GA process by influence of the input and hidden nodes (neurons) in the ANNs. These settings are usually determined by trial and error and were both used in these experiments set at 0.1, a value commonly recommended (Davis 1991). The final step prior to model execution is setting the number of neural networks to save upon completion of the GA process. Ten networks were saved after each GA run based on their performance which was measured for each trial network by the absolute average error between the predicted variable and the observed results.

The GA process is now ready to start. Processing times varied between a few minutes to an hour of CPU time. After ten networks were generated the networks were archived as binary coded files. Additionally, the observed and predicted variables were also stored as ASCII text files for statistical analysis and graphical presentation.

## Results

### *Tests for data normality*

The first phase of the data analysis prior to any further statistical procedures was a determination of data normality because any of the commonly applied parametric methods assume that the data is normally (Gaussian) distributed. Using the Shapiro-Wilk W-Test (SAS 1995) for normality all of the variables in the modeling effort failed the normality test. Therefore, nonparametric test statistics were implemented for the remainder of the study. Masters (1993) has recommended that the more commonly used product-moment correlation coefficient for estimating neural network performance be replaced by ranking methods as the correlation coefficient implies normality in the neural net input and output variables. For brevity, Figure 3 and Tables 1 through 3 illustrate the series of normality examinations with the soil temperature as an example. Based on these results only nonparametric measures were applied for the remaining portion of the study.

Figure 3. Distribution statistics for soil temperature variable (SAS/JMP).

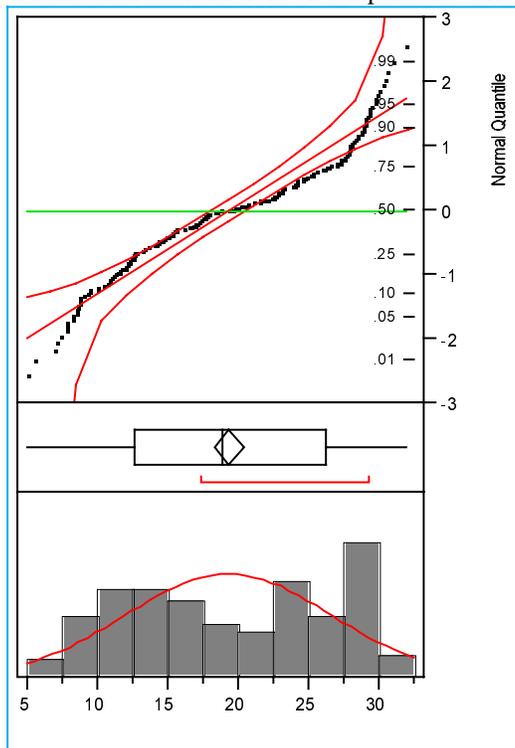


Table 1. Normal quantiles metrics for soil temperature.

maximum	100.0%	32.070
	99.5%	32.070
	97.5%	30.333
	90.0%	28.954
quartile	75.0%	26.290
median	50.0%	18.890
quartile	25.0%	12.680
	10.0%	9.418
	2.5%	7.397
	0.5%	5.080
minimum	0.0%	5.080

Table 2. Moments of soil temperature distribution.

Mean	19.4068
Std Dev	7.2512
Std Error Mean	0.5193
Upper 95% Mean	20.4310
Lower 95% Mean	18.3827
N	195.0000
Sum Weights	195.0000
Sum	3784.3300
Variance	52.5804
Skewness	-0.0552
Kurtosis	-1.3165
CV	37.3644

Table 3. Test for normality of soil temperature distribution using Shapiro-Wilk W-Test.

W Test Statistic	Probability <W
0.917625	<.0001

***Correlations between input and output variables***

The first step in the modeling process was to search for relationships within the complete data set, their occurrence and their magnitude. Conservative estimates of interdependence between precipitation, stream flow rates and river flow to tritium movement in streams and the river would suggest that the neural network models would have relationships to learn and equally important that the data is robust as a whole. This was accomplished by estimating the nonparametric Spearman-Rho degree of association coefficients between all combinations of variables in the data set. A part of these estimates of association shown in Table 4 below indicates that climatic variables, stream flow rates, and tritium estimates are significantly related in a first-order fashion. The strong association between soil temperature and evaporation and stream and river flow rates indicate that data quality is good enough for modeling. Based on these results the complete data set was then accepted as ready for modeling purposes. The existence of some of the stronger associations with tritium variables therefore warranted the use of statistical seeding during the application of the GA optimization process.

Table 4. Spearman-Rho nonparametric estimates of association between input and output variables for tritium models.

Variable 1	Variable 2	Spearman-Rho	P>Rho
Average evaporation	Tritium - FM6	0.3013	<.0001
Average evaporation	Soil Temperature	0.9305	<.0001
Average evaporation	Tritium FM6	0.3013	<.0001
Precipitation	Tritium FM6	0.2186	0.0021
Soil_Temp	Tritium FM6	0.2729	0.0001
RM-160 Flow	Precipitation	0.3266	<.0001
RM-160 Flow	RM-120 Flow	0.7899	<.0001
SC-4 Flow	RM-120 Flow	0.5696	<.0001
SC-4 Flow	RM-160 Flow	0.5195	<.0001
Tritium RM-120	Tritium SC-4	0.6569	<.0001
Tritium RM-120	Tritium SC-2A	0.2019	0.0046
Tritium RM-120	Tritium FM6	0.3073	<.0001
Tritium RM-120	Tritium FM-A7	0.3313	<.0001
Tritium RM-120	Tritium FM-3A	0.5034	<.0001
Tritium RM-120	Tritium FM-2B	0.4975	<.0001
Tritium RM-120	Tritium FM-3	0.4582	<.0001
Tritium RM-120	Tritium FM-1C	0.3741	<.0001

***Preliminary modeling experiments predicting tritium at Stations FM-6 and RM-120***

Approximately 120 ANN models were developed to assess the three major types of neural models (backward propagation, generalized regression and time series) that could be used to predict tritium mass transport at Station FM-6, at the bottom of Fourmile Branch, and Station RM-120, at the Savannah River just below the SRS.

The GA process during these modeling runs were set at literature recommended default values. They were:

- Generations Run = 10
- Population Size = 30
- Minimum network training passes for each network = 20
- Cutoff for network training passes = 50
- Input neural node influence factor used = 0
- Hidden neural node influence factor used = 0
- Limit on hidden neurons = 8
- Selection was performed by the top 50% surviving
- Refilling of the population was done by cloning the survivors
- Mating was performed by using the TailSwap Technique
- Mutations by Random Exchange technique at a rate = 0.25%

The summary of these GA-ANN runs are shown in Table 5. All of the three basic types of ANN models predicted tritium mass transport in Fourmile Branch and in the Savannah River to an almost equal degree. The Spearman-Rho statistic is the degree of association between the observed mass transport and the predicted mass transport.

Table 5. Summary of neural network types and their performance in predicting tritium mass transport in Fourmile Branch and the Savannah River.

Tritium variable	Type of ANN	Spearman-Rho	Prob. > Rho
Tritium Mass at FM-6	BP	0.86	<.0001
Tritium Mass at FM-6	GRNN	0.93	<.0001
Tritium Mass at FM-6	TDNN	0.85	<.0001
Tritium Mass at RM-120	BP	0.81	<.0001
Tritium Mass at RM-120	GRNN	0.88	<.0001
Tritium Mass at RM-120	TDNN	0.82	<.0001

BP-Backward Propagation, GRNN-Generalized Regression Neural Network, TDNN-Time Delay Neural Network

The second step in the analyses of these modeling runs was to observe how the GA optimization portion of the modeling was performed. During the breeding process of ANNs the GA selects the input variables at each generation which gives the closest agreement between the observed and predicted based on the average absolute error between them. Additionally, the GA, unlike many standard ANN systems, chooses different neural activation functions within any hidden layers in the design. The output neural activation function is also optimized. Table 6 summarizes these results from the same models in Table 5.

Table 6. Preliminary Artificial Neural Networks developed by the Genetic Algorithm process.

Model type / Predicted variable	Input variables selected by GA	No. hidden layers	Hidden Layer activation functions	Output neural activation function
BP / Tritium mass FM-6	Precipitation at 773-A Soil Temperature Tritium transport,FM3A Average evaporation Precipitation at F-200 Flow rate at RM-160	1	1 Logistic 3 Tangent 2 Linear	logistic
GRNN / Tritium mass FM-6	Tritium transport,FM3A Average evaporation Precipitation at F-200 Flow rate at RM-160	2	All summation functions	Linear/direct
TDNN / Tritium mass FM-6	Julian time Soil Temperature Tritium transport,FM3A Precipitation at F-200 Flow rate at RM-160	1	2 Logistic 2 Linear	logistic
BP / Tritium mass RM-120	Julian time Flow rate at FM-6 Flow rate at RM-160 Flow rate at SC-4	1	2 Tangent 2 Linear	Tangent
GRNN / Tritium mass RM-120	Julian time Precipitation at 773-A Soil Temperature Flow rate at RM-160 Flow rate at SC-4	2	All summation functions	Linear/direct
TDNN / Tritium mass RM-120	Julian time Soil Temperature Precipitation at F-200 Flow rate at RM-160 Flow rate at SC-4	2	4 Logistic 7 Tangent 4 Linear 1 Logistic 2 Tangent 2 Linear	Tangent

The inherent structural differences between the backward propagation (BP), generalized regression (GRNN) and time delay (TDNN) neural networks are partially responsible for the differences below, but the variety resulting from the GA process is significant. This is not surprising since part of the GA process stems from randomization processes in mutation rates, mating and population initialization. The neural activation functions in a GRNN model are always linear summation functions with a direct (no function at all) or linear output function, since the model is based on linear regression functions operating in parallel. Note also that in the BP and TDNN models the functions are almost completely nonlinear.

At this stage of the project tritium transport data had been offered to the models at station FM-3A as it was thought to be needed for predictions at station FM-6. Tritium data from station FM-6 was likewise offered as an input variable for predictions at station RM-120 in the river. However, the above results suggested that the tritium data probably would not be necessary at all to develop robust models. Therefore, the next stage of the effort was then focused on building networks that did not use any tritium estimates and relied solely on climatic and flow rate inputs. Augmenting this approach additional information in the form of new flow rate input variables at stations HP-52, FM-2B, and FM-1C were added to the input data file for improving predictions at station FM-6 in Fourmile Branch. After training an ANN the test ANN is presented with a portion of the total input data set for validation as a check on the network's performance. Then as a second check the network is compared to a larger portion, approximately half of the data set for testing the network against the data. These three steps in the creation of the final neural network were further supported by a manually introduced third quality performance check on the final networks selected by the GA process. Here 10 per cent of the complete data set were removed by random record selection. A random number generator from a C++ program selected the records. This data set was withheld from even uploading into the GA application assuring that the selected networks had never seen this data. Adding this step would further support our efforts to build ANNs that will prove accurate with new data.

Four models are presented below which are much closer to a production-level system for tritium forecasting. The first is a BP model, the most conservative and well understood design, that predicts tritium transport at station FM-6. During the GA process the number of generations was increased to 20 and initial population size was increased to 50. Multiple hidden layers were allowed with up to 16 neurons in either hidden layer. Also added was allowing the GA to use an 'influence' mode for input and hidden neurons during the optimization process. They were both given literature default values of 0.1. The resulting network had a first hidden layer with 2 logistic and 3 linear neurons. The second hidden layer had 2 logistic and 1 linear neuron. The single output neuron used a tangent transfer function. The input variables selected by the GA were Julian time, precipitation at 773-A, soil temperature, flow rates at station FM-3A, FM2B, HP-52 and FM-A7. The Spearman-Rho test of association = 0.94 indicating good agreement with the data. Figure 4 below illustrates this agreement. When the BP model was given the completely new values to predict the results again were very good with a Spearman-Rho value = 0.91.

In the following figures the term 'predicted' equates to the values from the neural model simulations, while the term 'desired' equates to the empirically estimated tritium values.

Figure 4. Predicted and desired data records of the back propagation neural network model predicting tritium transport (Ci) at SRS station FM-6 in Fourmile Branch.

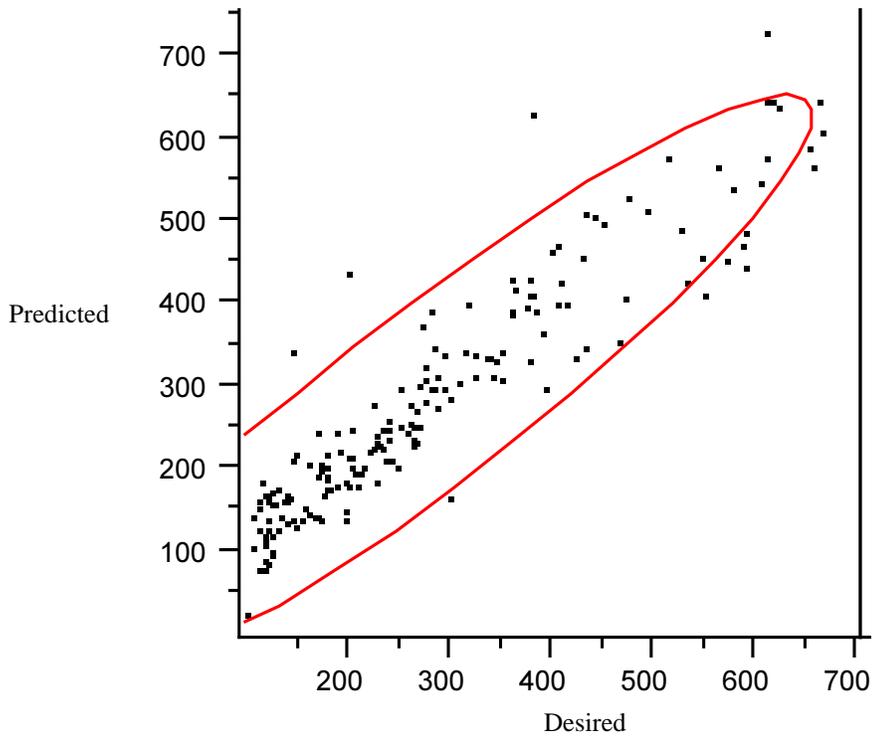
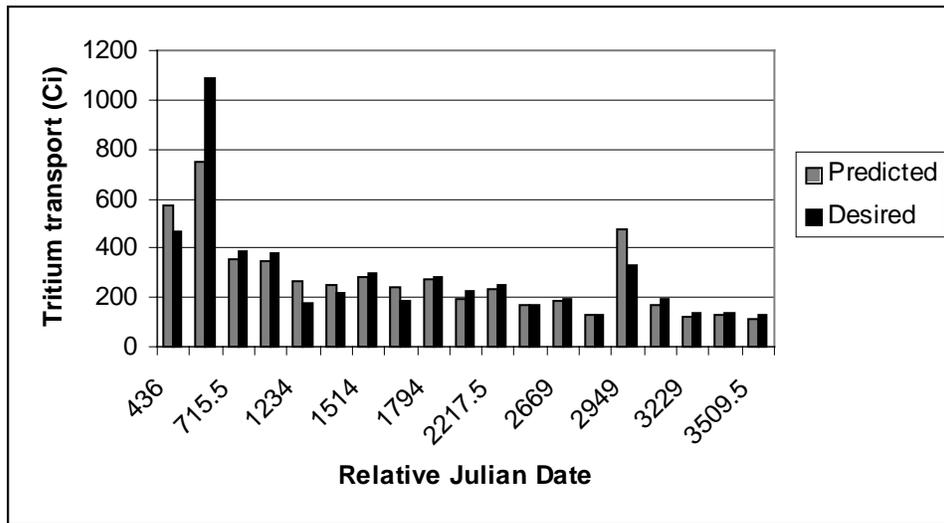


Figure 5. Predicted and desired values for the extra validation set of the back propagation neural network model predicting tritium transport (Ci) at SRS station FM-6 in Fourmile Branch.



The second model presented is a Time Delay Neural Network also predicting tritium transport at station FM-6. This model is of a very different design developed especially for forecasting time series. GA processing was enhanced again by initializing 50 ANNs and setting the GA for 20 generations. Influence mode settings were again set to 0.1. The resulting network employed 6 inputs with a single hidden layer

having 5 logistic, 3 tangent and one linear neuron with 6 connections. The single output neuron used a tangent transfer function with single connections to the hidden layer. The GA-selected input variables were Julian time, soil temperature, precipitation at F-200 area, and flow rates at stations HP-52, FM-A7 and RM-160, the station at Augusta. This network, performing essentially the same task as the BP ANN above, performed equally well having a Spearman-Rho association coefficient = 0.93 (Figure 5). The separate validation statistic, also a Spearman-Rho association coefficient = 0.92.

Figure 6. Predicted and desired data records of the time delay neural network model predicting tritium transport (Ci) at SRS station FM-6 in Fourmile Branch.

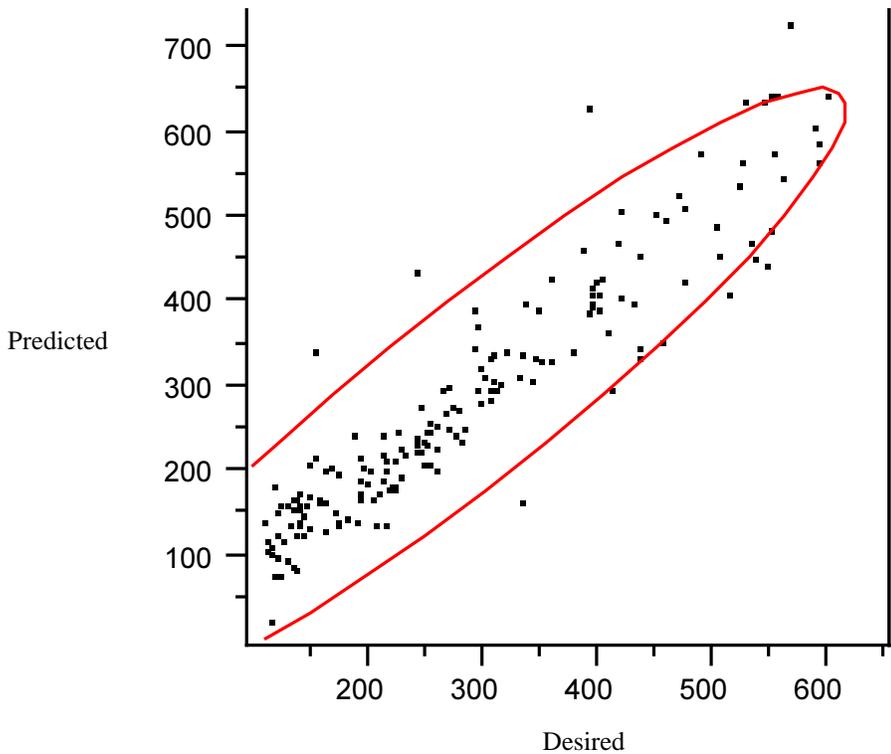
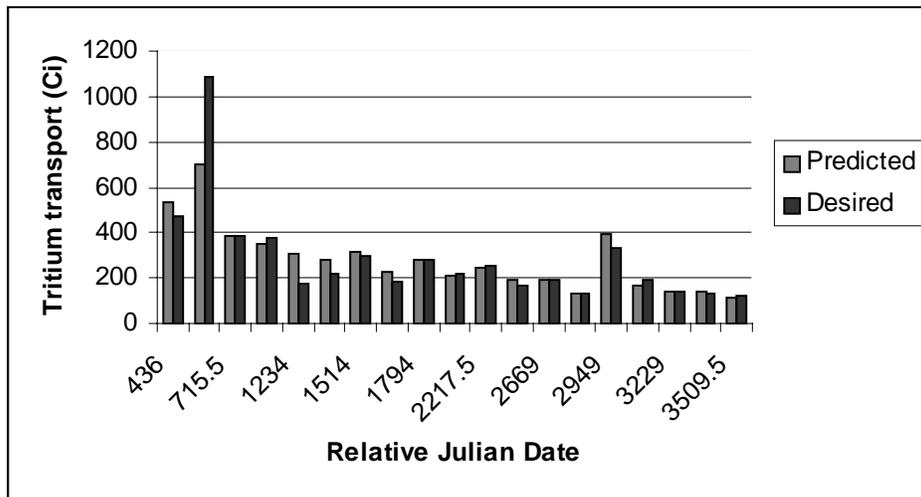


Figure 7. Predicted and desired values for the extra validation set of the time delay neural network model predicting tritium transport (Ci) at SRS station FM-6 in Fourmile Branch.



The following two networks followed the same protocols above in predicting tritium transport at the river station, RM-120, below the SRS. The first is the BP network and the second a time series TDNN. The BP network employed 9 input variables with one hidden layer having 10 logistic, 4 tangent and one linear neuron. The output neuron was a logistic transfer function. The selected variables were Julian time, precipitation at 773-A and flow rates from the following stations, SC-2, FM-3A, FM-2B, HP-52, FM-A7, FM-6 and RM-160. Performance for this BP model was less than for station FM-6 having a Spearman-Rho coefficient = 0.78 and the extra validation measure of association = 0.86. This validation performance indicates a good fit to the data and a lack of overfitting.

The TDNN employed 9 input neurons and a single hidden layer with one logistic, one tangent neuron with single connections. There was a single output neuron using the tangent transfer function with single connections to the hidden layer. The GA-selected input variables were Julian time, soil temperature and flow rates from 7 stations, SC-2, FM-3A, FM-2B, HP-52, FM-A7, FM-6 and RM-160. The performance was again very good with a Spearman-Rho = 0.83 and equaling 0.85 for extra validation test.

Figure 8. Predicted and desired data records of the back propagation neural network model predicting tritium transport ( $C_i$ ) at SRS station RM-120 in the Savannah River.

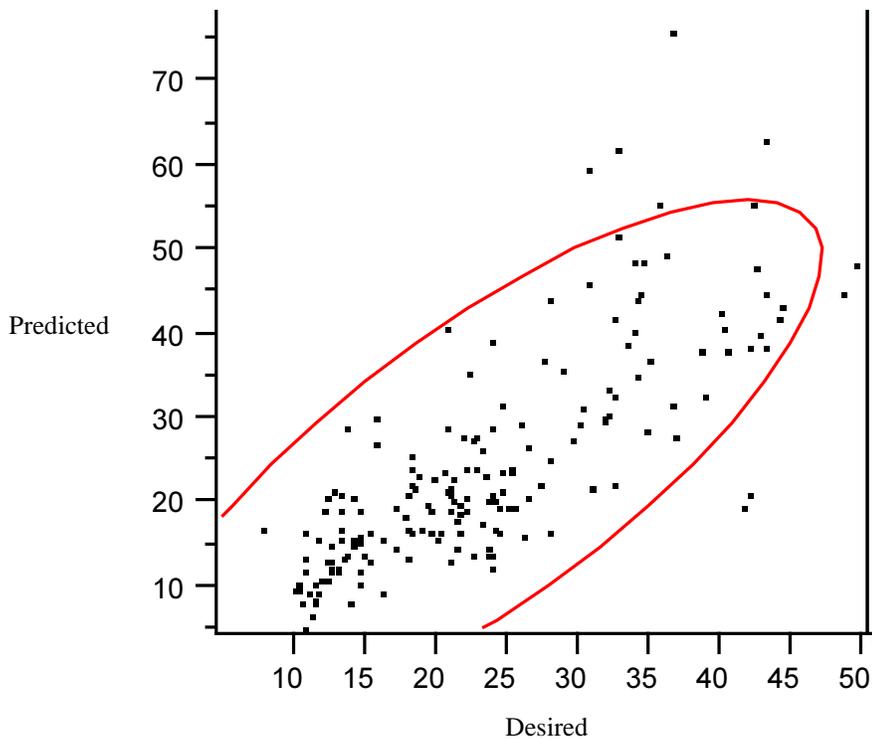


Figure 9. Predicted and desired values for the extra validation set of the back propagation neural network model predicting tritium transport (Ci) at SRS station RM-120 in the Savannah River.

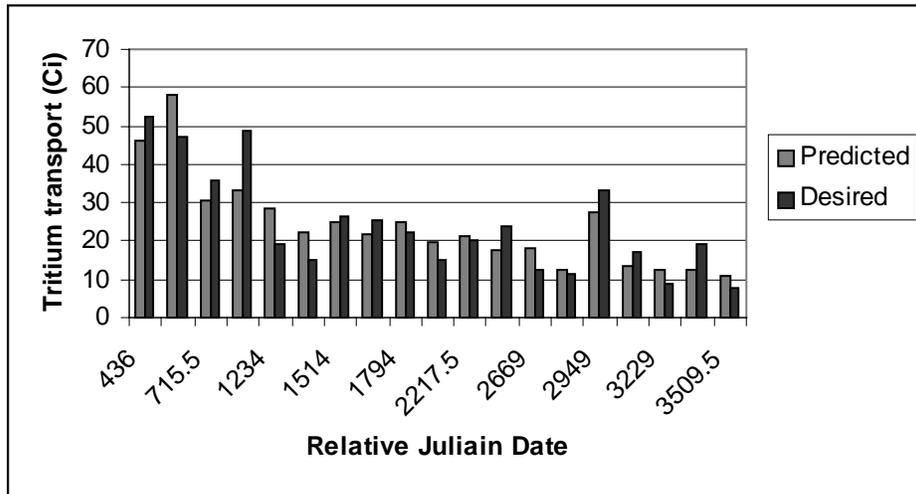


Figure 10. Predicted and desired data records of the time delay neural network model predicting tritium transport (Ci) at SRS station RM-120 in the Savannah River.

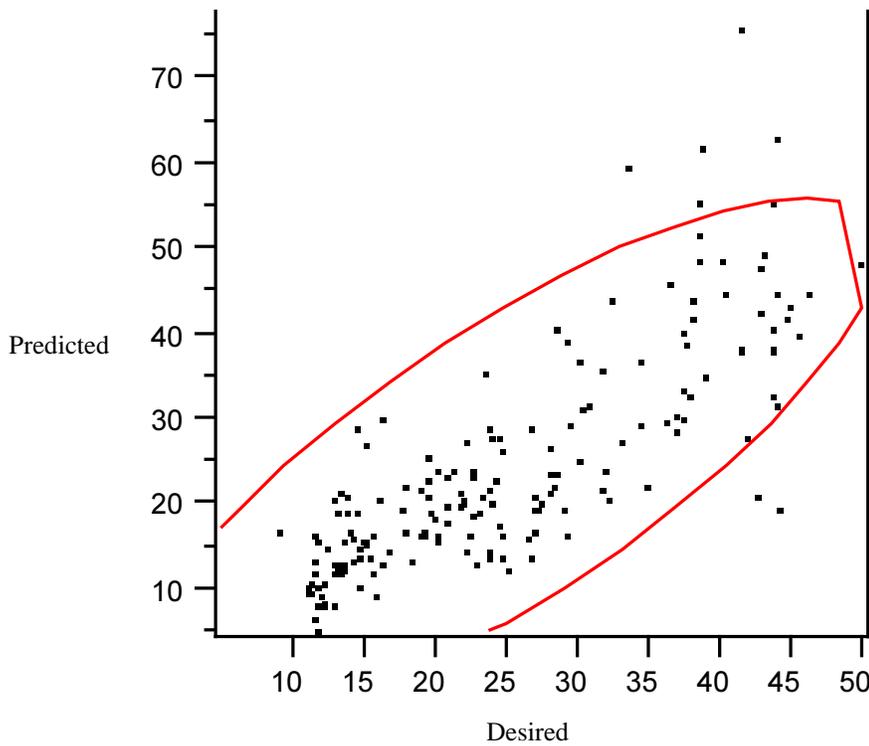
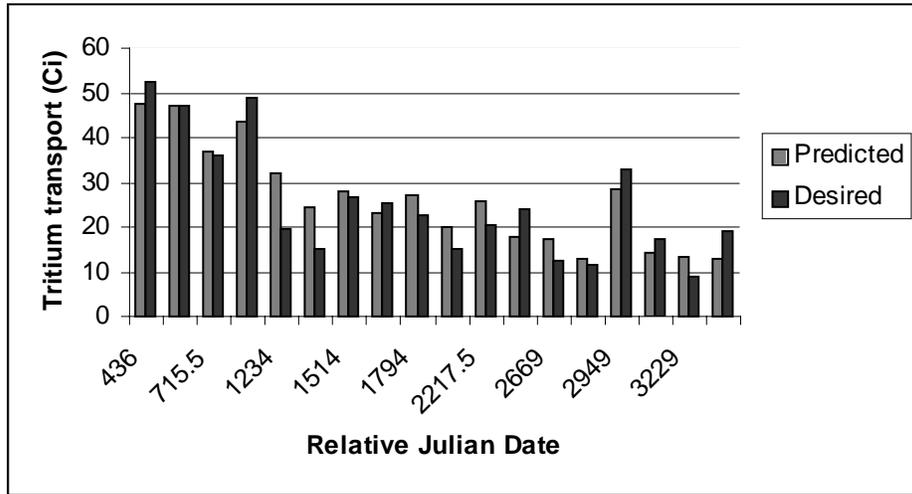


Figure 11. Predicted and desired values for the extra validation set of the time delay neural network model predicting tritium transport (Ci) at SRS station RM-120 in the Savannah River.



**Forecasting tritium concentrations at SRS Station FM-6 in Fourmile Branch**

All of the preceding experiments have focused on predicting tritium mass transport. Several modeling runs were performed to assess the forecasting of tritium concentration (pCi/L). Presented here is a BP neural network employing 11 inputs and one hidden layer having 6 logistic, 5 tangent, and 3 linear neurons. The output neuron uses the logistic transfer function. The input variables were Julian time, precipitation at 773-A and F-200, soil temperature, average evaporation from soil and stream flow rates from stations FM-3A, FM-2B, HP-52, FM-A7, FM-6 and RM-160. The Spearman-Rho association test was 0.86 and 0.83 for extra validation test indicating a satisfactory fit and validation of the model.

Figure 12. Predicted and desired data records of the back propagation neural network model predicting tritium concentrations (pCi/L) at SRS station FM-6 in Fourmile Branch.

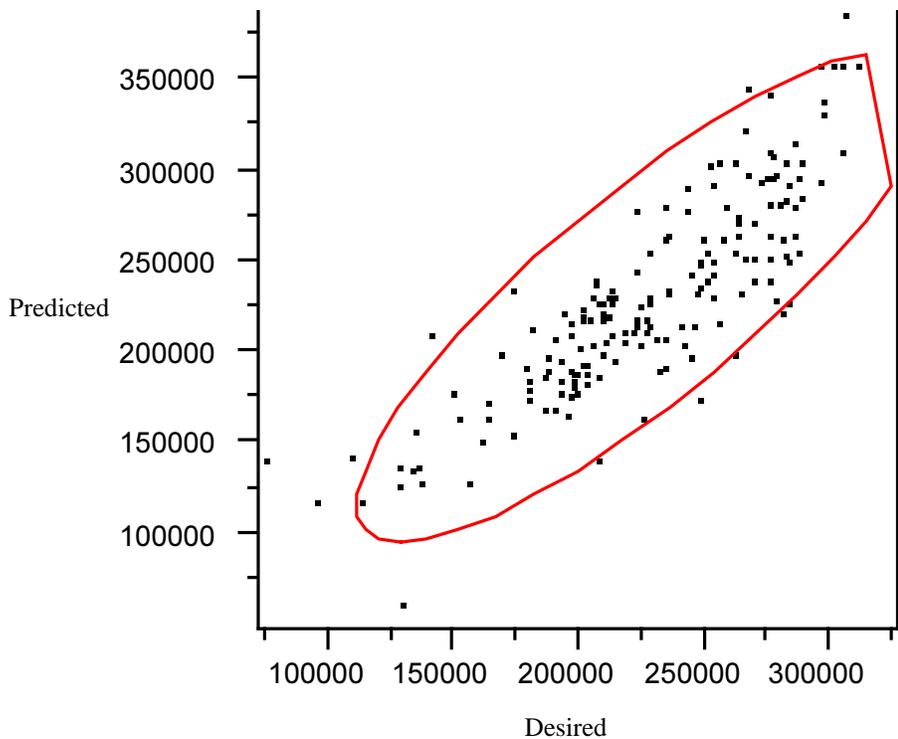
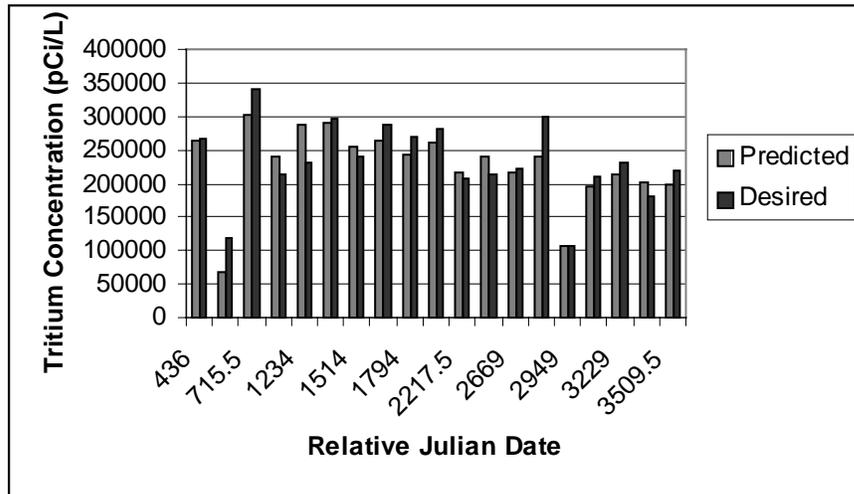


Figure 13. Predicted and desired values for the extra validation set of the back propagation neural network model predicting tritium concentrations (pCi/L) at SRS station FM-6 in Fourmile Branch.



## Conclusions

The unevenly spaced data, unique to each tritium sampling station at the SRS, presented unique challenges and opportunities. Data assembly required significant time and effort in record-by-record detail, but proved to be a workable solution using relative Julian time. More importantly, this data condition presented a special opportunity to demonstrate the powerful capabilities of neurogenetic modeling. Unequal time series data presents extremely restrictive limitations to traditional statistical and deterministic modeling. For example, traditional correlation and regression techniques cannot be applied to the data set because time steps from station to station are not synchronous. Classic time series analyses for predicting tritium at a single station through time, also cannot be applied because these methods require, for the most part, equally spaced and normally distributed data. Neurogenetic models are not restricted by these conditions. Furthermore, neurogenetic models accept data that is nonlinear in their relationships.

A 'production' model that would predict tritium concentrations or mass transport downstream in either Fourmile Branch or downstream of the SRS in the Savannah River still requires development. This additional effort would add additional flow rate estimates as input parameters for Pen Branch, Upper Three Runs Creek and Lower Three Runs Creek. There are also additional outfall stations having flow rate estimates that have not yet been preprocessed for the modeling. These additional flow-rate information streams would give the neurogenetic process additional input data streams which is the very heart of the learning process. Having gathered the additional flow stations the models could be retrained with very little effort. Then a proper sensitivity analyses could be performed on all of the input variables for a better understanding of why the GA process selects the input variables for each of these models guiding future data collection in a more efficient manner. Obviously, more recent tritium concentrations should be acquired for further testing and validation of the above models.

Two other aspects of the tritium data that should be investigated is the large-scale variability of the tritium transport or concentration in the streams and river over very short time intervals and the nonlinear nature of tritium concentration distributions. These features of tritium behavior warrant further investigation. Two graphs illustrate both behaviors. Figure 14 illustrates this short-term variability and the differences between the desired (actual) tritium concentration at Station FM-6 in Fourmile Branch versus what the neural network predicted. Note how the neural model attempts to respond to larger scale patterns of change, but over and under approximates to achieve an overall general solution to the complete multiyear pattern.

This lack of fit can be more easily viewed when the tritium concentration is sorted in ascending order as shown in Figure 15. Having the observed tritium concentrations sorted indicates the distribution of the near decade of concentrations with it being nonlinear at the ends of the distribution with regard to concentration. The greatest differences between of the desired and predicted concentrations occur at the lowest and highest concentrations.

Figure 14. Predicted and desired data records of the back propagation neural network model predicting tritium concentrations (pCi/L) at SRS station FM-6 in Fourmile Branch covering a period from approximately January 1991 through December 1999.

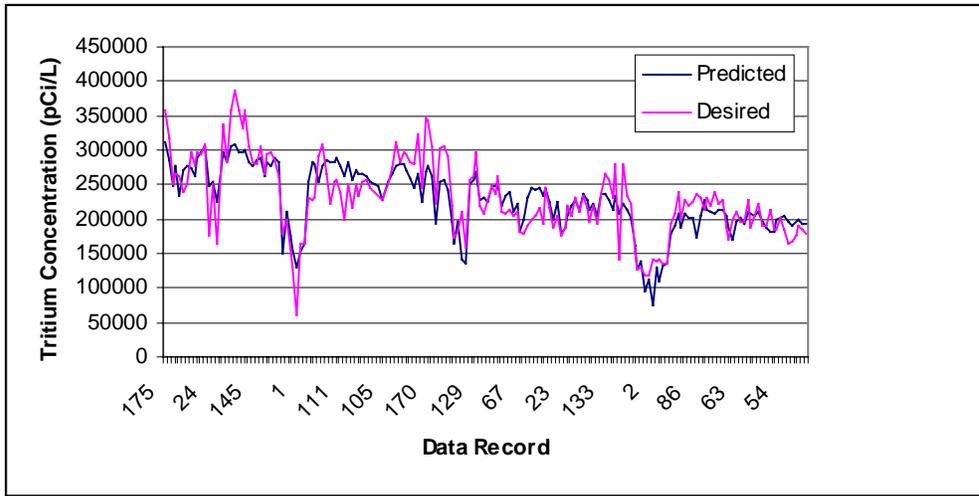
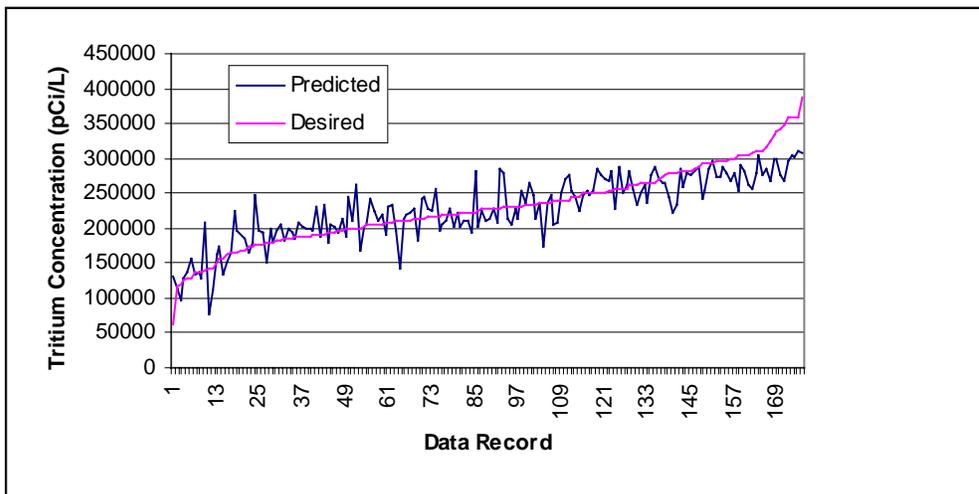


Figure 15. Predicted and desired data records of the back propagation neural network model predicting tritium concentrations (pCi/L) at SRS station FM-6 in Fourmile Branch covering a period from approximately January 1991 through December 1999. The model-desired or observed tritium concentrations are sorted here in ascending order.



BioComp (2000) and Masters (1993) have recommendations for these difficulties of fit. Data smoothing in the form of running means or variance stabilizers (transformations) often aid neural networks in accommodating high-frequency variation. In this case tritium concentrations or mass transport values could be smoothed with a 3-interval mean smooth function or a simple log base 10 data transformation of the input data to achieve a better total degree of fit or general solution for long-term prediction.

Extra neural training at important inflection points in the data is the most recommended solution to achieving better fits at tails of distributions (Masters 1993). In this case extra training at the lower and higher tritium concentrations or mass transport allows neural networks to adjust activation functions and their respective coefficients for the more extremes of information flow. Addition model refinement is accomplished through classical sensitivity analysis. Input variables are randomly altered in varying degrees and then compared to the variation in the output variables.

The above results indicate that GA-optimized ANNs can be used for the prediction of tritium in SRS streams and the Savannah River. Especially noteworthy is the capability to forecast tritium behavior based solely on climatic and stream flow information streams. This is not surprising since neural networks have been applied for many years in forecasting stream, riverine and watershed parameters (Maier and Dandy 2000, Lek et al. 1996, Poff et al. 1996, Zhu et al. 1994).

The development of production quality neurogenetic models offers a new approach to study of tritium behavior at SRS. The model results may permit reduced sampling for some of the SRS stations and allow SRS to model certain tritium release scenarios under different climatic and operational conditions. The addition of operational inputs into the models would offer predictive capabilities on the effects of controlled tritium releases into SRS streams and further downstream effects in the Savannah River. Perhaps the most intriguing facet of neurogenetic models is the ability to run the models 'backwards' (inverse modeling). When using tritium concentration or mass transport as input variables the model operating backwards could be used to forecast tritium transport or concentrations at upstream or 'source' points of release. This possibility is offered here as yet another topic for exploration and development.

## References

- BioComp Systems Inc. 2000. 4018 148th Avenue N.E., Redmond, WA 98052 USA.
- Bowers, J.A., and C. B. Shedrow. 2000. Predicting stream water quality using Artificial Neural Networks (ANN). In: Development and application of computer techniques to environmental studies VIII. WIT Press, Southampton, England, UK. pp. 89-98.
- Bowers, J.A., Osteen, G.E., Rogers, G., & Martin, F.D., DataDelve client and EcoTrack server: a spatial data system for environmental warehousing. Development and application of computer techniques to environmental studies VI, ed. P. Zanetti & C.A. Brebbia, Computational Mechanics Publications: Southampton and Boston, pp. 467-474, 1996
- Davis, Lawrence, "Handbook of Genetic Algorithms", Van Nostrand Reinhold, 1991, ISBN 0-442-00173-8
- Flack, G., F. Syms, J. Bowers and M. Harris. 2000. Prediction of Fines Content from CPTu Measurements. Presentation to the American Geophysical Union, Annual Meeting, Atlanta, GA.
- Gruau, F. 1993. Genetic synthesis of modular neural networks. Kaufmann, M. (ed.) Fifth International Conference on Genetic Algorithms, ISBN 1-55860-299-2.
- Hagan, M.T., H.B. Demuth, and M. Beale. 1996. Neural network design. PWS Publ. Co. Boston.
- Houlsby, G.T. 1998. Advanced interpretation of field tests. *In* Geotechnical Site Characterization, P. K. Robertson and Mayne (eds.), Balkema, Rotterdam, ISBN 9054109394.
- Lek, S, Guiresse, M. & Giraudel, J-L. 1996. Predicting stream nitrogen concentration from watershed features using neural networks. Water Research, 33(16), pp. 3469-3478.
- Maier, H.R. & Dandy, G.C. Neural networks for the prediction and forecasting water resources variables: a review of modeling issues and applications. Environmental Modelling & Software, 15, pp. 101-124, 2000.

WSRC-TR-2000-00442  
October 2000

Masters, T. Practical neural network recipes in C++, Academic Press: New York and London, pp. 404-421, 1993.

Poff, N.L., Tokar, S. & Johnson, P. Stream hydrological and ecological responses to climate change assessed with an artificial neural network, *Limnology & Oceanography*, 41(5), pp. 857-863, 1996.

WSRC. 1998. Environmental Data for 1998. EMS/EPD, Westinghouse Savannah River Company, Savannah River Site, Aiken, SC, WSRC-TR-98-00314.

WSRC. 1999. Meteorological monthly monitoring at the Savannah River Site. Westinghouse Savannah River Company, Savannah River Technology Center, Aiken, SC, WSRC-TR-99-00046-10.

Zhu, M.L., Fujita, M., & Hashimoto, N. Application of neural networks to runoff prediction, in *Stochastic and statistical methods in hydrology & environmental engineering*, ed. Hipel, K.W. et al. Kluwer Academic Publishers, Norwell, Mass., 1994.