MEMORANDUM

TO: Tarboton, Kenneth, Ph.D., P.E., Director, Regional Modeling Division

THROUGH: Luis Cadavid, Ph.D., P.E., Chief Hydrologic Modeler

FROM: Alaa Ali, Ph.D., P.E., Lead Hydrologic Modeler

DATE: November 8, 2004

SUBJECT: A New Rainfall Driven Formula to Predict More Natural Flows to the Everglades' Shark River Slough

The following is an executive summary of the project whose title is the subject of this memorandum. The detailed technical report is also attached.

EXECUTIVE SUMMARY

The Rainfall Driven Plan, RDP, is a key component of several of the alternatives in the 1992 GDM for Modified Water Deliveries to ENP. The goal of the RDP is to improve the amount, timing and distribution of flow to the Shark River Slough, the main waterway of the ENP. An essential component of the RDP is Rainfall Driven Formula, RDF, to provide prediction of natural system flow in response to real time weather conditions. The existing RDF was developed in 1985 and since then has been in use by South Florida Water Management District (SFWMD) and United States Army Corps of Engineers (USACE). The 1992 GDM for Modified Water Deliveries to ENP recommends that "as additional knowledge is gained through experience with the interim operating plan and data collection, subsequent changes to the operating plan should be made, as appropriate". With the **C**ombined **S**tructural and **O**perational **P**lan of the C-111 and Modified Water Delivery projects, CSOP, coming online, it is desired to carry out this recommendation by revisiting the existing RDF for possible improvements.

Over the past decade there has been a great deal of improvement of systems understanding, data acquisition, and scientific evolution in statistical techniques. In this study, these improvements were exploited to develop a new RDF considering a nonlinear system, using 6 rainfall stations, and one PET station as input (predictors) to model the Natural System Model flow to the Shark River Slough.

Input data were partitioned into a modeling period (1965-1989) and validation period (1990-2000). The modeling period was used to develop the model, while the validation period was used to test the model. Principal component analysis was applied to the modeling set to simplify the data. A *feedforward Levenberg-Marquardt backpropagation* Artificial Neural Network with one hidden layer and one output layer was adopted. The network architecture and training parameters were identified for ANN training. A set of 5000 ANN training and validation simulations for different network

parameter values were performed to select an optimal set of weights and biases and to provide quantification of model uncertainty. The resulting formula produced 88% correlation for both the development and validation periods. Also, the formula provides good prediction during the validation period where the hydrologic conditions are significantly wetter than those of the modeling period. Complete model formulation, and results including model uncertainty are presented. Model parsimony, optimal ANN parameterization, effective data selection, and real time uncertainty quantification are discussed.

A New Rainfall Driven Formula to Predict More Natural Flows to the Everglades' Shark River Slough

Alaa Ali, PhD, PE

South Florida Water Management District
Regional Modeling Division
Office of Modeling

1) INTRODUCTION

The goal of the Rainfall Driven Plan, RDP, is to improve the amount, timing and distribution of flow to the Shark River Slough, the main waterway of the ENP. To achieve this goal, a sound prediction of natural system flow in response to real time weather conditions is essential. The prediction tool of the current RDP, known as the Rainfall Driven Formula, RDF, was developed in 1985 and since then has been in use by South Florida Water Management District (SFWMD) and United States Army Corps of Engineers (USACE). The reader is referred to Technical Publication 89-3 for full details about the RDP and RDF currently implemented today. With the **C**ombined **S**tructural and **O**perational **P**lan of the C-111 and Modified Water Delivery projects, CSOP, coming online, it is highly desirable to revisit the existing RDF for possible improvements.

The existing RDF was developed based on hydrologic data of the Everglades system during the period of record 1941-1952. This period represented an optimal trade off between Everglades meteorological data monitoring and acquisition on one side and South Florida urbanization and development on the other side. Prior to that period meteorological data were scarce while post that period there was a significant alteration of the flow to SRS due to the completion of the levee along the eastern side of the Everglades. The current formula was developed based on a linear system assumption using 10 rainfall stations, three Evapotranspiration stations and considering ten week lags.

With the availability of more data, advanced statistical methods, and computational technologies, it is desirable to develop a new rainfall formula that reflects our improved understanding of the system and take advantage of the new concept of the South Florida natural system evolved through the Natural System Model, (NSM). The new RDF is developed assuming nonlinear system, using 6 rainfall stations, and one evapotranspiration station as input (predictors) to model the NSM flow to the Shark River Slough. A comparison between the important aspects of the old and the new formulas is given in Table 1.

This memorandum provides a complete presentation of the development, validation, and implementation of the new RDF. A description of the study area and data locations is first provided, followed by a detailed presentation of the RDF methodology.

Detailed model results are then presented followed by a project summary and recommendations.

Table 1. Comparison between selected aspects of the existing and the new RDFs

Point of Comparison	**Existing Rainfall Driven Formula**	**New Rainfall Driven Formula**
Modeling period of record	1941-1952	1965-1989
Validation period of record	7/1985 – 7/1987 (two year field test of the rainfall plan)	1/1990-12/2000 (formula is tested to predict NSM flow)
Input data	10 rain, 3 ET	6 rain, 1 ET
Modeled output	Observed flow in a partially drained system	Simulated flow by Natural System Model
Time Lag	10 weeks	6 weeks
Assumption	Linear System	Non linear system
Data location contribution	Lumped	Distributed to each individual data location

2) PROJECT AREA AND DATA

Figure 1 depicts the area of interest and the data locations considered in this study. Rainfall and PET stations are located in the Water Conservation Areas. In a natural system, Rainfall in the WCAs generates sheet flow across the Tamiami Trail into the ENP.

2.1) Input Data

Ten rainfall stations and one potential Evapotranspiration station were initially selected for this analysis (Table 2 and Figure 1.) The station selection was based on two criteria: 1) potential relevance to the flow prediction, and 2) current active data acquisition. The period of record covers the period 1965 through 2000. Meeting the above two criteria posed the dilemma of not finding active and relevant stations during the period of record 1965-2000. To overcome this dilemma and to prepare a sound rainfall data set, the following procedure was adopted:

1) Identify stations by name and location that meet the above two criteria regardless of the period of record.
2) Extract the corresponding daily rainfall from the existing data set previously used in the SFWMD models SFWMM and RSM. Note that this data set had already undergone a through QA\QC process.
3) Identify the SFWMM grid cell row and column corresponding to each rainfall station.
4) For each station, fill all missing data from the corresponding SFWMM cell

2.2) Flow Data

In this study it is desired to model pre-drainage flow across Tamiami Trail to SRS as flow target. Since historical data to support this effort is not available, the Natural System Model simulated flow data were used.

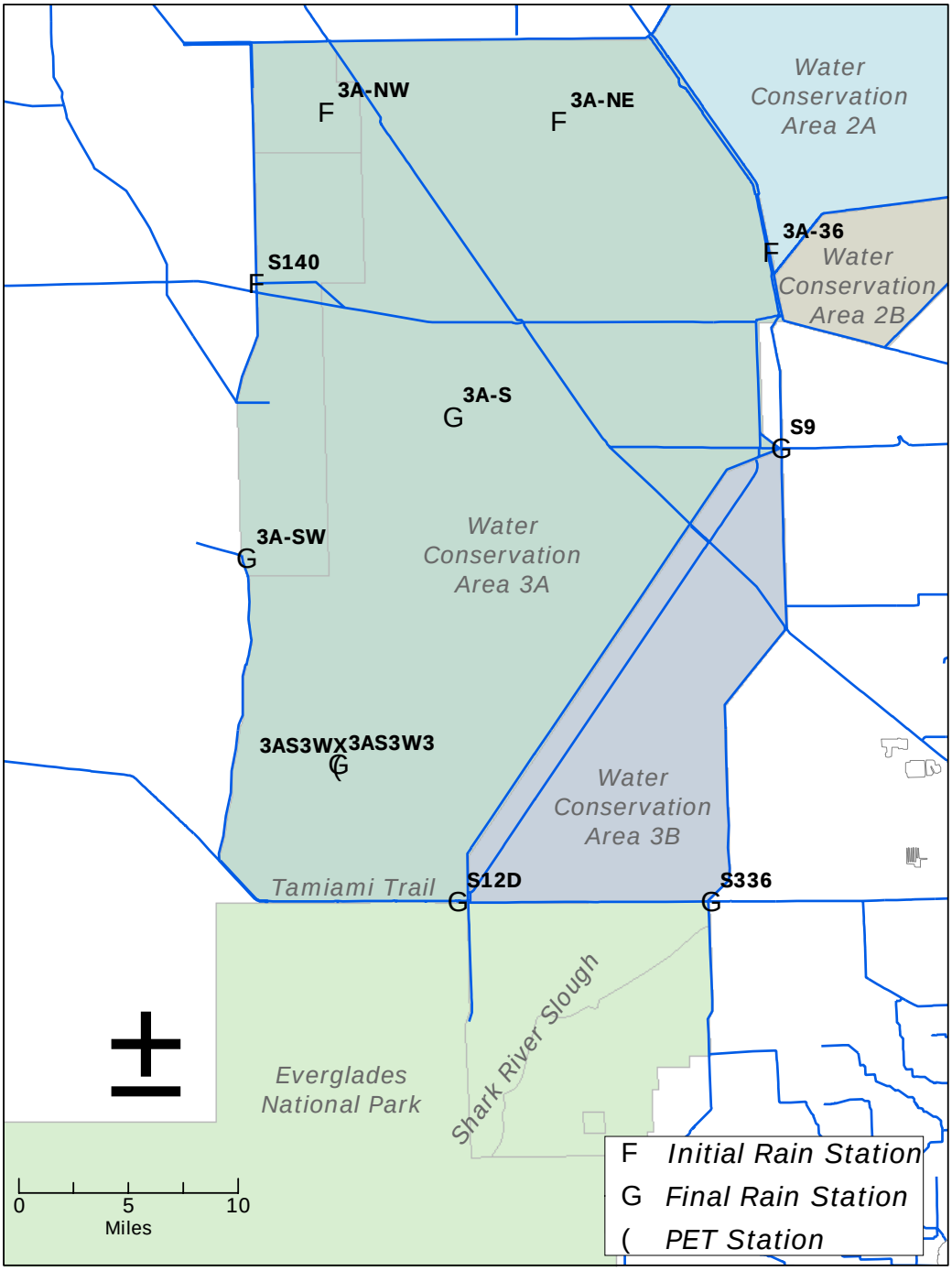

Figure 1. Rainfall driven formula study area with the data locations

Table 2. Information on Rainfall and PET stations considered for this study.

	Rainfall										PET
Station	3A-36	S140	**S336**	**3A-S**	**3A-SW**	3A-NW	**S12D**	3A-NE	**3AS3W3**	**S9**	**3AS3WX**
Dbkey	16175	16581	**16713**	**HC941**	**JA344**	LA365	**LS269**	LX283	**M6888**	**16607**	**OH515**
ROCO[1]	37,28	36,16	**23,27**	**33,21**	**30,16**	40,18	**23,21**	40,23	**25,18**	**32,28**	**N/A**
Beginning of POR[2]	1/95	1/91	**10/95**	**4/98**	**2/99**	5/00	**7/00**	8/00	**5/00**	**1/91**	**4/2000**

Boldface entries represent the final selection

1 Row and Column of the SFWMM grid.

2 Data prior to the beginning of POR date do not exist and hence data from the corresponding ROCO is used.

3) DEVELOPMENT OF RAINFAL DRIVEN FORMULA

The development of the RDF is presented in the following four subsections. 1) Data preparation and partition, 2) Data transformation using principal component analysis, 3) Artificial Neural Network, ANN, Parameterization, and 4) Development of flow prediction model.

3.1) Data set preparation

The input and flow data presented in the previous section were used to develop, verify and validate the RDF. Numerous combinations of rainfall stations were used in early model development efforts to pick a near optimal set of rainfall stations based on preliminary model performance. Based on this exhaustive analysis, not presented here, data selection was narrowed down to six rainfall stations and one PET station (Table 2 boldface entries, and Figure 1).

Daily data were converted into weekly data and were divided into two sets: modeling set (1965-1989) and validation set (1990-2000). Each time series of the modeling data set is standardized to zero mean and unity standard deviation (i.e., subtracting its respective mean and dividing by the standard deviation). Given week t, an entry in time series i, $s_{t,i}$, is standardized according to Equation 1:

$$x_{t,i} = \frac{s_{t,i} - \mu_i}{\sigma_i} \qquad (1)$$

Where: μ_i and σ_i are the modeling data mean and standard deviation for time series i.

The mean and standard deviation for both data sets are presented in Table 3. Notice that the mean and standard deviation of the flow and rainfall are significantly higher in the validation period compared to those in the modeling period (see last row of Table 3.). Such an increase represents a challenge during the model validation.

The standardized modeling data set is further divided into development set (≈1965-1983), and verification set (≈1984-1989). A typical vector of standardized data is denoted by **X** consisting of 22 elements as follows:

- previous time step simulated flow target (1 entry)

- rain and PET values for first week lag (7 entries)
- rain and PET values for second week lag (7 entries)
- rain and PET average values over the third through the fifth week lag (7 entries)

Table 3. Mean and Standard deviation of the modeling and validation data sets

		Flow Ac-ft/week	PET inches	S336 inches	3A-S inches	3A-SW inches	S12D inches	3AS3W3 inches	S9 inches
Modeling Data "M"	**Mean**	20274	-1.13	1.00	0.84	0.91	0.96	0.90	0.91
	STD	18708	0.26	1.41	1.16	1.21	1.34	1.16	1.23
Validation Data "V"	**Mean**	33366	-1.10	1.09	1.24	1.04	1.09	1.09	0.98
	STD	26751	0.26	1.53	1.64	1.43	1.52	1.46	1.28
Mean Difference (V-M)/M		**65%**	**-3%**	**9%**	**48%**	**14%**	**14%**	**21%**	**8%**

3.2) Principal Component Analysis.

Principal Component Analysis, PCA, is a technique that can be used to simplify a dataset. It is a transform that chooses a new coordinate system for the data set such that the greatest variance by any projection of the data set comes to lie on the first axis (referred to as the first principal component), the second greatest variance on the second axis (referred to as the second principal component) and so on. PCA is used to reduce the dimensionality in a data set while retaining those characteristics of the dataset that contribute to its variance as predictors by eliminating the later principal components. Another benefit of the PCA in the context of ANN, is the reduction, or even elimination, of the uncorrelated, or poorly correlated, signals that exist mostly in the later principal components. Such a noise has an adverse impact on the ANN training.

Given the 22 variables listed above and considering the modeling data set only, Figure 2 shows the percentage of variance explained by the first "n" components with "n" ranging from 1 to 21. It is shown that the first 5 components explain 97.3% of the variance. It was rather attractive at an early modeling stage to retain the first 5 components of the data for the RDF development. While the results were not poor, there was no computation feasibility issue that would preclude a full use of the 22 variables. When the full 22 variables were used, result improvements were not significant. Trial and error efforts showed significant improvements when the first 17 principal components (99.5% of the variability) were retained. One reason for this improvement is the elimination of the white noise in the top 0.5% variability and the inclusion of the extra 2.2% (beyond 97.3% if first 5 PCs are considered) of the variability which retained important information for prediction. Further insight is needed to rationalize these quantities which is beyond the scope of this document. Whether or not this task is performed does not preclude the modeler from pursuing the formula based on the first 17 components.

The principal component coefficient matrix **P** (22 X 17) based on the modeling data set is presented in Appendix A. For week, t, the principal component vector $\mathbf{Y_t}$ is obtained as:

$$\mathbf{Y_t} = \mathbf{X_t} * \mathbf{P}. \quad (2)$$

Note that $\mathbf{Y_t}$ is a 1 X 17 one dimensional array.

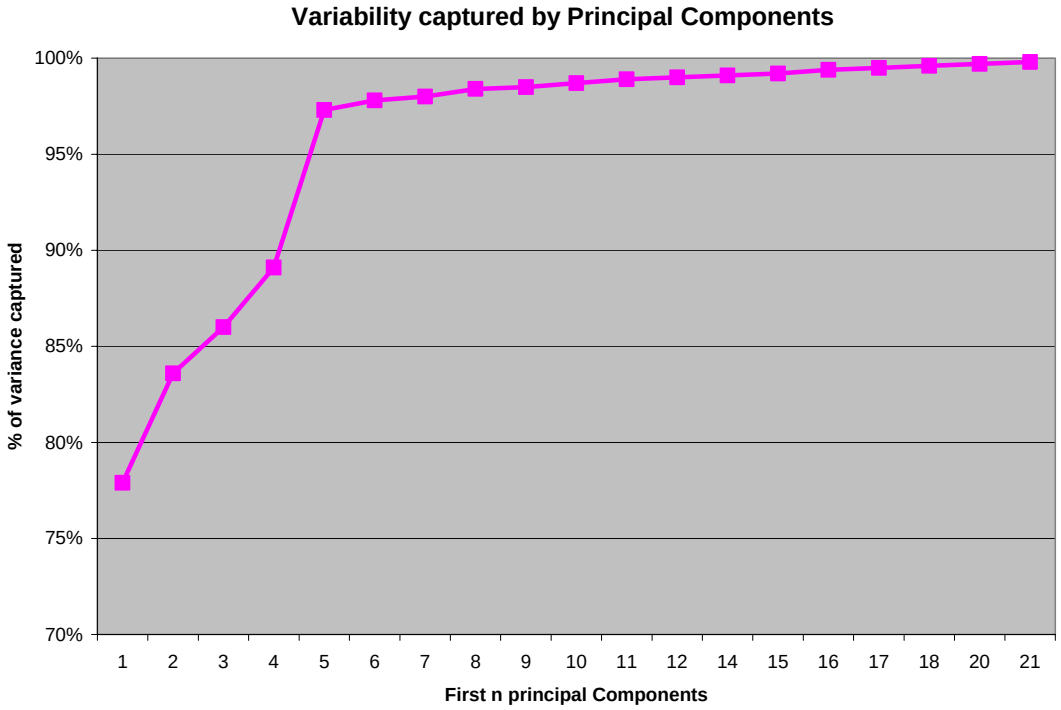

Figure 2. Percentage of variance explained by the principal components.

3.3) Artificial Neural Network Parameterization

3.3.1) Background

Artificial Neural Networks, ANN, are mathematical models of human cognition (Govindaraju, 2000). These models can be trained based on historical knowledge to perform a specific task where such knowledge is not available. They are typically composed of three parts: inputs, one or many hidden layers, and an output layer. Hidden and output neuron layers include the combination of weights, biases, and transfer functions. A neuron on a given layer is a hub that receives weighted contributions from the preceding layer's neurons and it sends weighted contributions to the succeeding layer's neurons. The weights are connections between neurons on one layer and another while the transfer functions are linear or nonlinear algebraic functions. When a pattern is presented to the network, weights and biases are adjusted so that a particular output is obtained. The ANNs provide a learning rule for modifying their weights and biases. Backpropagation algorithm is the most used learning rule in hydrologic applications

using ANNs. It is a generalization of the least mean square (LMS) algorithm to multiple-layer networks and nonlinear differentiable transfer functions.

A satisfactory level of ANN training is the one that results in a good network generalization (i.e., satisfactory network performance on input that was not part of the training). The ANN architecture parameters are the number of hidden layers, the number of hidden nodes, type of transfer functions, and the learning rule algorithm. The ANN training parameters that govern the ANN optimization during training include learning rate, performance gradient, number of training iterations, training and verification data sizes and other numerous optimization parameters. For full explanation of these parameters and the ANN architecture, the reader is referred to (Govindaraju, 2000, Wasserman 1989, and Rumelhart, Hinton and Williams, 1986)

A parsimonious ANN is the one that produces the best fit (and the best generalization) with the simplest architecture and least number of parameters. A major limitation to ANN is the inability to identify the unnecessary parameters/ weights in the solution. In fact such identification would provide insight about the problem and help formulate an efficient ANN. Although ANN parsimony did not receive a lot of research emphasis as the traditional parametric models did, there have been some research efforts of optimization techniques to identify an optimal parsimonious architecture/parameters. These methods exploit the structural redundancy of the ANN Architecture, parameters, and or weights by means of trial and error techniques. For a review of these methods the reader is referred to (Sexton et. al., 2004, Abrahart et. al., 1999, Yao, and Liu 1997 and Bossley et. al., 1995)

3.3.2) ANN for the Rainfall Driven Formula

The selection of ANN Architecture and training parameters and weights is a trial and error process to choose from infinite number of combinations of the parameters presented above. No formal technique was used to identify an optimal architecture. In this study, a reasonable ANN performance has been consistently attained when *feedforward Levenberg-Marquardt backpropagation* network with one hidden layer and one output layer was used. The hidden layer has 10 units (nodes) and the output layer has one unit (one output target time series). Figure 3 provides a schematic diagram of the RDF formulation. The trained network provides sets of weights and the terms for each layer as described below.

3.3.3) ANN weights and bias terms

The ANN weights and bias terms represent the RDF parameters. These parameters are initialized at the beginning of the ANN training. During training, these parameters are iteratively adjusted according to the *feedforward Levenberg-Marquardt backpropagation* algorithm to minimize the network performance function which is the Mean Square Error, MSE, in this case. . An optimal set of weights and biases are then used in the RDF formulation as follows (see Figure 3).

Hidden layer (10 nodes)

The weight matrix **Ω** (17 X 10) and the bias vector **A** (1 X 10) are presented in Appendix B. The weighted contribution of vector $\mathbf{Y}_t$ (1 X 17 vector) to hidden node, *i*, is given as:

$$C_{t,i} = Tansig\ (\mathbf{Y}_t * \boldsymbol{\omega}_i + \alpha_i) \quad (3)$$

Where: *i* is an index for the hidden layer nodes (i = 1 -10)

$\boldsymbol{\omega}_i$ is the i^{th} vector (17 X 1) of Weight matrix **Ω**

α_i is i^{th} element of the bias vector **A**

Tansig: Hyperbolic tangent sigmoid transfer function. It is nonlinear differentiable function with input range $\pm\infty$ and output range ± 1.

$C_{t,i}$ is a single value representing the contribution of vector $\mathbf{Y}_t$ to node *i* of the hidden layer.

Note: $\mathbf{Y}_t * \boldsymbol{\omega}_i$ is a product term of two vectors that produces a single value.

Output layer (one node)

The weight vector **Θ** (10 X 1) and the bias value β are presented in Appendix B. The weighted contribution of $C_{t,i}$ to the output node:

$$D_{t,i} = C_{t,i} * \theta_i \quad (4)$$

Where:

θ_i is the i^{th} entry of Weight vector **Θ**

$D_{t,i}$ is a single value representing the contribution of nod *i* of the hidden layer to the output layer node.

3.4) Flow target Estimation for week t+1

The flow target for week (t+1), $\hat{q}_{t+1}$, is estimated as (see Figure 3):

$$\hat{q}_{t+1,i} = \left(\sum_{i=1}^{n} D_{t,i} + \beta \right) * \sigma_q + \mu_q \quad (5)$$

Where:

n: number of hidden layer nodes (10 in this case)

β: the output layer bias term

μ_q : Historical global mean of flow target time series of the modeling data set.

σ_q : Historical global standard deviation of flow target series of the modeling data set.

Substitute (4) in (5):

$$\hat{q}_{t+1} = \left(\sum_{i=1}^{n} C_{t,i} * \theta_i + \beta \right) * \sigma_q + \mu_q \quad (6)$$

Substitute (3) in (6):

$$\hat{q}_{t+1} = \left(\sum_{i=1}^{n} Tansig\left(\mathbf{Y}_t * \boldsymbol{\omega}_i + \alpha_i\right) * \theta_i + \beta \right) * \sigma_q + \mu_q \quad (7)$$

Substituting (2) in (7) produces the final format of the rainfall driven flow formula:

$$\hat{q}_{t+1} = \left(\sum_{i=1}^{n} Tansig\left(\mathbf{X}_t * \mathbf{P} * \boldsymbol{\omega}_{\mathrm{i}} + \alpha_{\mathrm{i}}\right) * \theta_{\mathrm{i}} + \beta \right) * \sigma_q + \mu_q \qquad (8)$$

The matrix product term: $\mathbf{X}_t * \mathbf{P} * \boldsymbol{\omega}_{\mathrm{i}}$ of the above formula may be rewritten a summation format as follows:

$$\hat{q}_{t+1} = \left(\sum_{i=1}^{n} Tansig\left(\sum_{\mathrm{k}=1}^{17} \omega_{k,i} * \sum_{j=1}^{22} \left(x_{t,j} * p_{j,k}\right) + \alpha_i \right) * \theta_{\mathrm{i}} + \beta \right) * \sigma_q + \mu_q \qquad (9)$$

Except for the data vector $\mathbf{X}_{\mathrm{t}}$, all input terms in Equation 8 (or 9) are computed once based on the historical modeling data set. The prediction of flow target proceeds in a feed forward manner where the antecedent predicted flow value is used as part of the input vector $\mathbf{X}_{\mathrm{t}}$ for the next time step flow prediction. Figure 3 presents a schematic diagram of the model developed above. Results for the mode are presented in the following section.

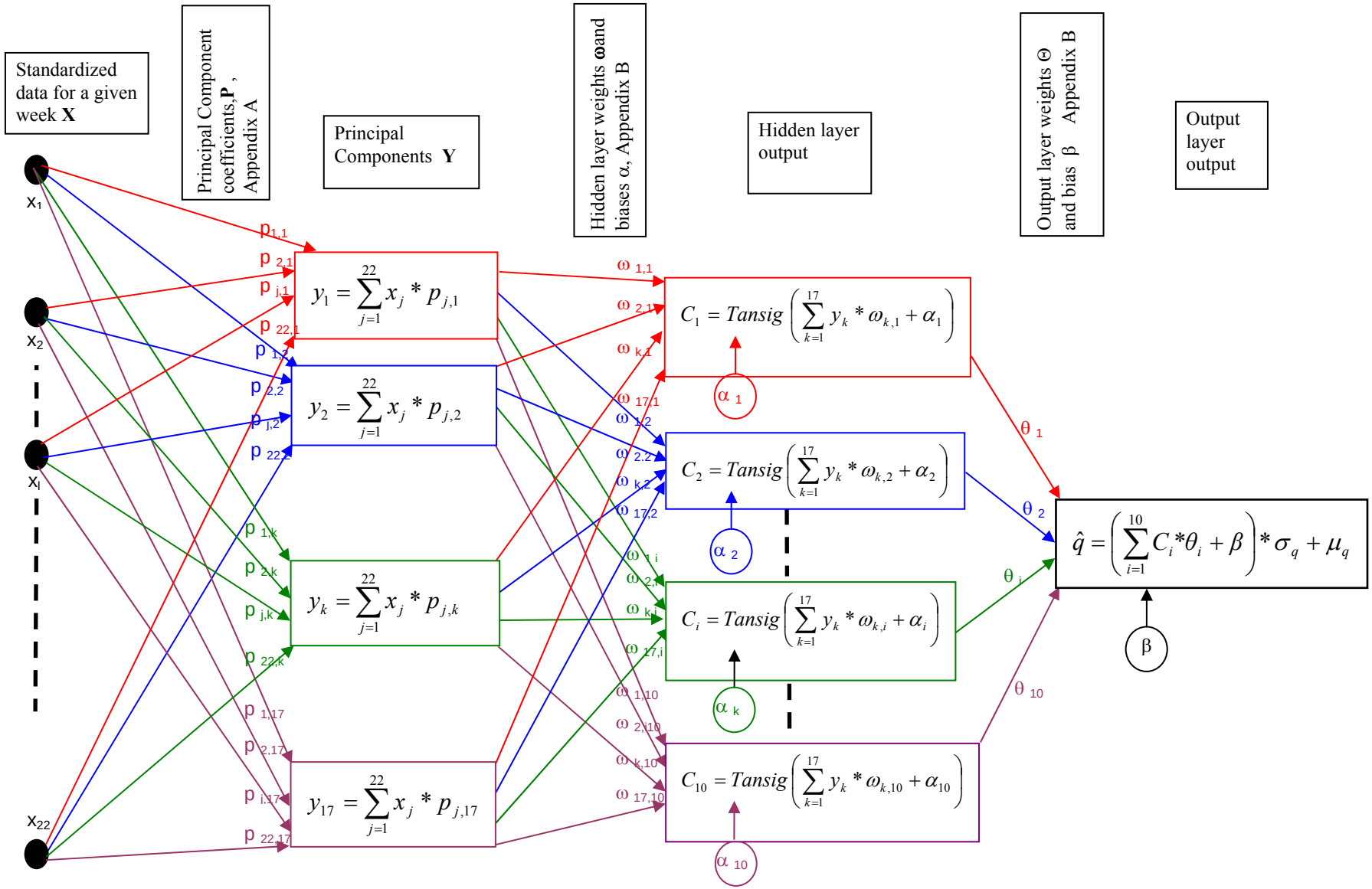

Figure 3. Schematic diagram of the ANN for the rainfall driven formula

4) RESULTS

The modeling data (1965-1989) were partitioned, standardized, and transformed to their first 17 principal components as described earlier. The ANN was trained using the transformed data according to the procedure presented in the preceding section. Verification data (1984-1989) was used to stop training early if the network performance on the modeling data failed to improve, remained the same, or deteriorated for a number of consecutive iterations (5 iterations were found to provide good results in this study). Given the nonuniqueness of the optimization problem, many combinations of training and network initialization parameters can provide different, but equally probable, solutions with equally low performance function (MSE) scores. For this purpose, 5000 ANN simulation sets (i.e., development, verification, and validation) were performed for controlled random initialization and training parameter values. The criteria for selecting a few of these sets were to maximize prediction correlation during the modeling and validation periods, while minimizing their differences (i.e., as similar and as high correlation values as possible for the modeling and validation periods). With other measures considered such as mean square error, global bias, and negative value occurrence, one set was selected for this study. Detailed statistics of this set are presented in Table 4. The rest of these sets were retained for subsequent uncertainty analysis.

Table 4. Comparative statistics for the training and the validation periods of the selected RDF. Except for correlation, all units are in ac-ft/week.

	Correlation	BIAS	STD	MSE
Development	0.88255	443.6756	8222.09	8848.089
Validation	0.88678	612.999	11178.79	12899.83

Given the selected RDF, two applications are considered in the subsequent subsections. The first application is the transfer of the RDF from the proprietary MATLAB to a portable EXCEL application. The second application is to present the model performance during the modeling and validations periods. A Numerical example of flow prediction for the week 1/30/2004 to 2/6/2004 is provided in Appendix C.

4.1) Model transfer from MATLAB to Excel

To eliminate proprietary software portability, and to make the formula, user friendly, portable, and as transparent as possible, model developed under MATLAB environment was transferred to portable EXCEL multiple tab spreadsheet. As seen in Figure 4, the results are identical with minor differences at the peaks. Due to the large number of parameters and data transformation involved in the computation, numerical approximations will likely result in some differences particularly at the high flow values.

4.2) Model Performance

Model performance during the modeling and validation periods is presented. The formula robustness is tested and the annual, wet season, and dry season flows are compared. Uncertainty analysis is then presented.

While it is expected to have a good model performance during the modeling period since this is the data used to estimate all model parameters, it is desired to present the model performance during that period. The prediction proceeds in a feed forward manner, where antecedent predicted flow is used as part of the input for the current time step. Figure 5 shows the predicted flow and NSM simulated flow target time series during the modeling period. Notice that the model reproduces the time series troughs and peaks reasonably well except in the dry season of 1986-1987 and the first half of the dry season 1987-1988 with significant flow target overestimation. Frequency and magnitude of negative prediction of the modeling and verification periods are presented in Table 5. Given an average NSM flow of 20,000 ac-ft/week, and a maximum flow of 110,000 ac-ft/week during the modeling period, the negative occurrence frequency and values in table 5 is fairly insignificant. Therefore, forcing the prediction to be non negative in real time implementation will unlikely introduce significant positive bias. Further tweaking of the formula parameters outside the ANN optimization algorithm may result in reduction of the negative occurrence frequency and magnitude on the expense of running the risk of formula over-fitting.

Table 5. Negative prediction occurrence range and the corresponding frequency

	NSM Flow Average ac-ft/week	NSM Flow Maximum ac-ft/week	Range ac-ft/week	Frequency: events/total points
Development Period: 1965-1989	20,274	108,196	-2,185 to -1,000	18 out of 1300
			-1,000 to 0.	22 out of 1300
Validation Period: 1990-2000	33,366	127,853	-2,731 to -1,894	4 out of 572
			-1,000 to 0.	7 out of 572

4.2.1) Weekly Flow prediction

Figure 6 shows a scatter plot (along with trend lines) between the NSM flow target and the predicted flow target during the modeling period. The correlation coefficient for the modeling period is 0.88. Forcing the intercept to zero does not change the correlation results.

Figure 7 shows the predicted flow and NSM simulated flow target time series during the validation period. To allow for independent performance from the modeling period, the flow target is initialized to the NSM flow target on January 5, 1990. The model proceeds forward as explained earlier. The model seems to generalize reasonably well by capturing the majority of the peaks and troughs of the newly introduced data set. Out of 572 prediction points, there are 11 negative prediction events with the values presented in table 5 above. Given NSM flow average and maximum of 33,000 ac-ft/week and 127,000 ac-ft/week respectively for the validation period, the negative events are insignificant.

Figure 8 shows a scatter plot (along with trend lines) between the NSM flow target and the predicted flow target during the validation period. The correlation coefficient for the modeling data is 0.89; compared to 0.88 in the modeling period. By

forcing the intercept to zero, the correlation remains the same. The results in Figure 7 and 8 indicate that the model generalized well during the period of 1990-2000 where the statistics of the data are significantly different with about 65% increase in the NSM flow average (see Table 3). This gives confidence in future model application beyond the modeling and validation periods.

To test the model robustness when the flow is reset at different dates, the RDF was applied to the validation data with 5 different resetting dates (see Figure 9). The dates were selected where the model performance was relatively poor to check the model adaptability to adjust from its poor performance once the flow is initialized to the NSM flow. There are two observations:1) the model matches the NSM flow much better, and 2) The model converges completely to its original line (the line without resetting) after few months This is not to suggest to reset the model when there is a poor performance (in fact this is not possible in real time), but rather to demonstrate the model long term robustness regardless of the flow initialization and or poor local performance.

4.2.2) Seasonal and annual Flow prediction

Figures 10 through 12 show NSM and predicted flow targets for annual, wet season, and dry season respectively. Result statistics are shown in Table 6. The average NSM and RDF predicted flow volumes are almost identical over the modeling period for each of the annual, wet and dry seasons. There is 5% overestimation of the average annual flow during the validation period with 4 % underestimation in the wet season and 10% overestimation in the dry season. Given the significant flow increase in the validation period (65%), the flow average matching during that period is considered very good. Standard deviation of the RDF prediction is similar to the NSM in the modeling period and is higher by 20% in the validation period.

Table 6. NSM and RDF seasonal flow average and standard deviation in (ac-ft/week).

			Annual	Wet	Dry
Average	Modeling period	**NSM**	1,054,048	555,212	497,186
		RDF	1,077,101	559,978	514,711
	Validation period	**NSM**	1,739,176	831,286	983,144
		RDF	1,822,676	797,680	1,105,287
Standard Deviation	Modeling period	**NSM**	643,100	355,538	412,455
		RDF	603,596	310,036	393,839
	Validation period	**NSM**	885,710	369,458	562,155
		RDF	1,093,354	487,823	641,011

4.2.3) RDF uncertainty

As explained at the beginning of this section, 5000 simulations were performed for different sets of ANN architecture initialization and training parameters. There are 3700 model scenarios where the correlation between the modeling and the validation periods is greater than 80%. Such scenarios were retained for subsequent uncertainty

analysis. The 5% , median, and 95% of weekly flow predictions out of such plausible realizations represent 90% confidence band encompassing the median line for the model performance. The RDF and NSM lines ideally should be encompassed within the band and be as close as possible to the median. In reality this is rarely true due to the model uncertainty. For a given week, the width of the uncertainty band is a measure of the model estimation variance, and the failure to encompass the NSM line within the band is an indication of the model bias. Also, the median is not necessary the best estimator of the data unless the uncertainty band evenly encompasses the NSM line. Furthermore, the median does not necessarily coincide with the RDF line because the RDF line represents one ANN realization while the median represents the median across all 3700 realizations. The confidence in prediction is high when the band is narrow and the NSM line is encompassed, and it is low when the band is wide and the NSM line is missed. A wide band with missed NSM line is an indication of loss of information due to the fact that all data contributing to the prediction are not included. Figures 13a through 13g show the 90% confidence band, median, NSM, and the RDF lines from 1965 through 2000. The following are specific observations from these figures.

1) Given a narrow uncertainty band and given NSM and RDF lines within that band during low flows and transition to and from high flows, there is a high degree of confidence in the RDF prediction in these cases.
2) High flow prediction has a wide uncertainty band that encompasses both NSM and RDF prediction lines most of the time.
3) Events where the 90% confidence band missed and underestimated the NSM line are in the second half of 1966, the first half of 1980, and the second half of 1991. Events with overestimated missed NSM line are weeks of the dry season 1986/1987, second half of dry season of 1987/1988, wet season of 1995, and dry season of 1997/1998. In general there are 440 events (23%) where the 90% confidence missed the NSM line. Such an increase beyond the 10% zone during the validation period is attributed partially to the inventible way rainfall data were obtained. Also, including more rainfall information, and/or further parameter tweaking may bring this percentage closer to 10%.

5) SUMMARY

In this study, a new rainfall driven formula has been developed to provide real time prediction for natural flow targets to Shark River Slough. Six rainfall locations and one PET location were the input data used in this study. Flow target for a given week is predicted using weekly data over the preceding 6 weeks in addition to the flow target predicted for the previous week. Input data were partitioned into a modeling period (1965-1989) and a validation period (1990-2000). The modeling period was used to develop and verify the model (verification portion of the data is used to prevent the network from overtraining), while the validation period was used to validate the model. Principal component analysis was applied to the modeling set to reduce the data dimensionality to its first 17 components. A *feedforward Levenberg-Marquardt backpropagation* Artificial Neural Network with one hidden layer and one output layer was adopted. The network architecture and training parameters were identified for ANN training. A set of 5000 ANN training, verification and validation simulations for

different network parameter values were performed to select an optimal set of weights and biases and to provide quantification of model uncertainty. The developed RDF was applied to the modeling, validation, and real time data. Detailed results of the RDF applications were also presented.

In this study, there was emphasis of the model ability to generalize, i.e., to provide satisfactory prediction beyond the modeling data. While some trained ANNs provided predictions with correlation as high as 95% during the modeling period, they did not provide as high correlation during the validation period. The best correlation balance was obtained at correlation values between 0.85 and 0.90. The finally selected ANN for the RDF produced 88% correlation for the modeling period and 89% correlation for the validation period. The RDF appears to generalize very well during the validation period which is significantly wetter than the modeling period. The RDF provides reasonable prediction of the annual, wet, and dry season flow volume. The RDF has also exhibited robust behavior regardless of when and where the flow target is initialized.

During the course of this study, many selection decisions were made based on past experience and trial and error processes. The use of principal component analysis was useful in eliminating most of the white noise from the data by eliminating the top ½ % variability and hence increasing the ANN training efficiency. The selection of a simple efficient ANN architecture was a trial and error process that found it reasonable to use ANN with 10 node hidden layer and one node output layer. Identification of unnecessary parameters/weights to achieve parsimony was not an easy process as explained earlier. While good generalization indicates a good fit it does not guarantee a parsimonious model. The generation of 5000 sets of ANN training, verification, and validation with variable network initialization and training parameters was a method to characterize the behavior of the network for 5000 optimized solutions. The product of this exercise is a quantification of the prediction uncertainty due to model parameters. The uncertainty band presented in this study can be used to provide the RDF with a range of flow target prediction that can be considered a "green zone" of non adversary regulatory releases.

Future work on the RDF may add improvements in several aspects. A major challenge in this study was the selection of the right predictors (input variables), the right lags, and the right partition. The addition of noncontributing predictors impedes effective ANN training while the omission of contributing predictors results in information loss. This is evident in a few events where the uncertainty bands have missed the NSM flow completely. While the results in this study prove to a large extent the effectiveness of the data selection made, there is a room for improvement by pursuing more formal analysis (as opposed to trial and error) in this selection. A more efficient selection of ANN architecture and training parameters may be made as more data, techniques, and understanding become available. There is a lot to be done for model uncertainty and there is a need to quantify uncertainty of flow target conceptualization, current and future weather conditions. A simple real time uncertainty quantification remains a challenge.

REFERENCES

Abrahart, R J., Kneale, L. and Kneale P. E. (1999), Applying saliency analysis to neural network rainfall-runoff modeling, the IV International Conference on Geo-Computation, hosted by Mary Washington College in Fredericksburg, VA, USA, on 25-28 July 1999.

Bossley, K. M., Brown, M. and Harris, C. J. (1995) Parsimonious Neurofuzzy Modelling. Technical Report, University of Southampton

Govindaraju, R. S., and Ramachandra Rao, A. (2000). Artificial Neural Networks in Hydrology. Kluwer Academic Publisher.

Neidrauer, C. J., and Cooper, R. M. (1989), A Two Year Field Test of The Rainfall Plan: A Management Plan for the Water Deliveries to Everglade National Park, *SFWMD Technical Publication 89-3*.

Powell,M. J. D. (1987). Radial basis function approximation to polynomials. *Numerical Analysis Proceeding*. Dundee, UK, 223-241.

Rumelhart, D. E., Hinton, E., and Williams, J. (1986). *Learning Internal Representation by Error Propagation, Parallel Distributed Processing I*. MIT Press.

Sexton, R. S., Dorsey, R. E., Sikander and N. A (2004)., Simultaneous optimization of neural network function and architecture algorithm**,** Computer Information Systems, Southwest Missouri State University, 901 South National, Springfield, MO, Volume 36 , Issue 3, 2004.

Yao, Xin, Liu, Yong (1997), A New Evolutionary System for Evolving Artificial Neural Networks. IEEE Transactions on Neural Network, Vol 8, NO 3, 1997

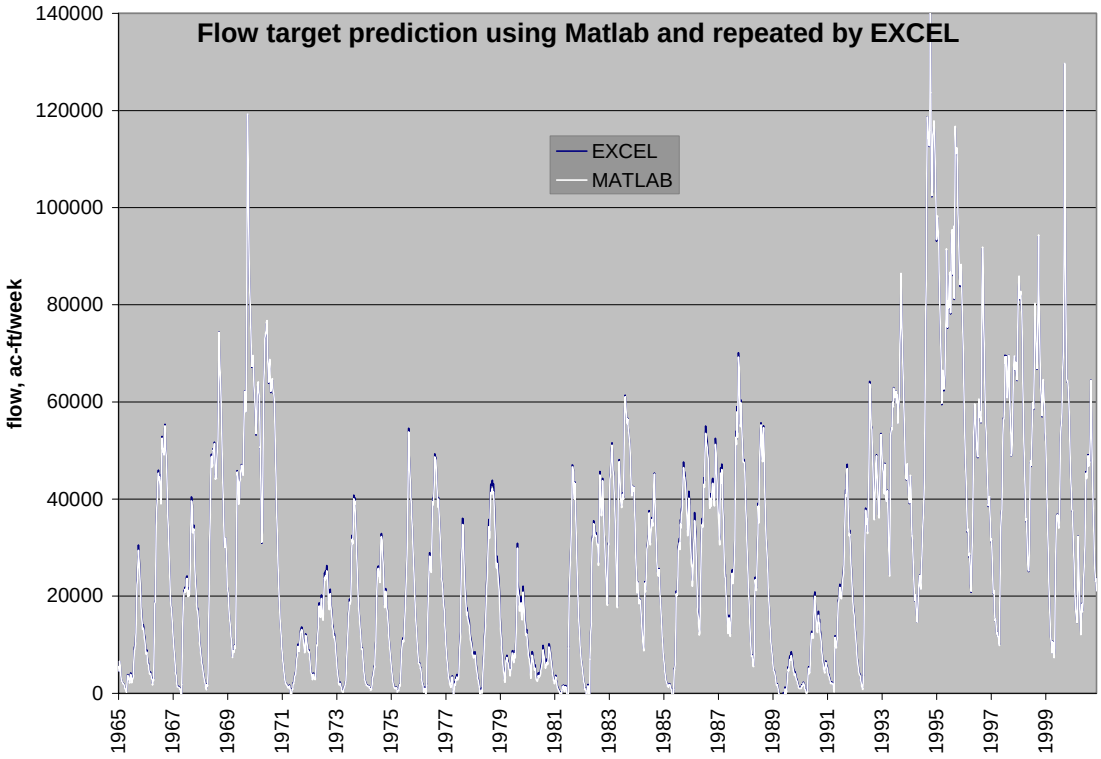

Figure 4. Comparison of Flow target prediction between MATLAB and EXCEL

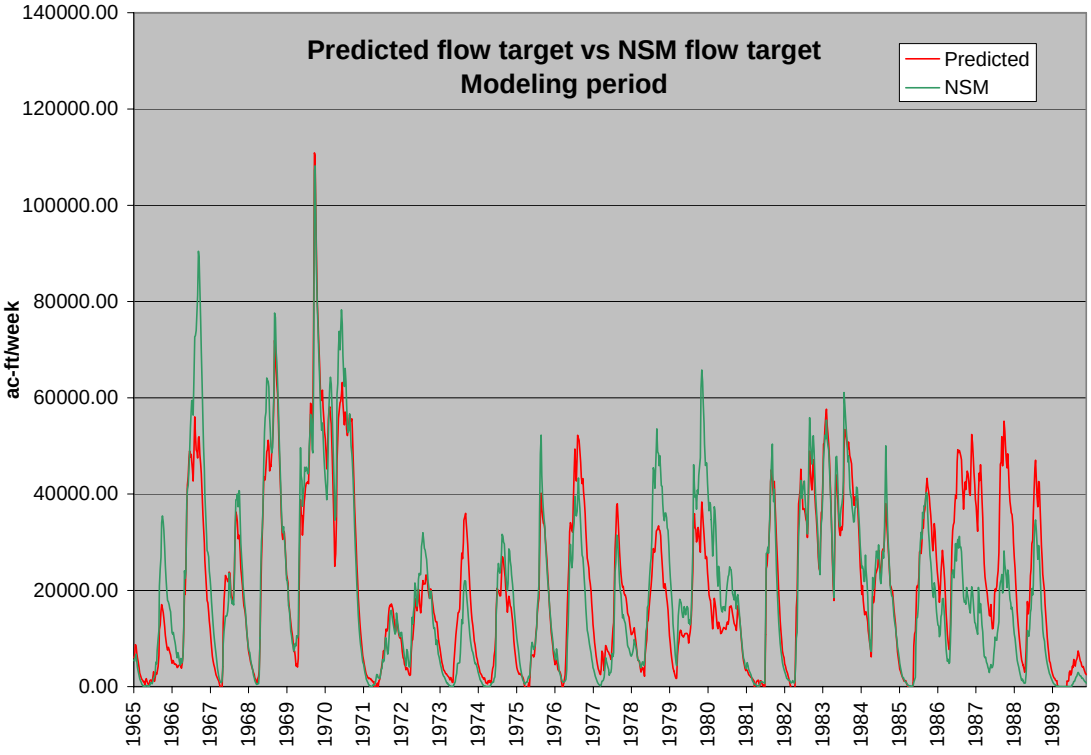

Figure 5. Predicted and NSM flow targets during the modeling period.

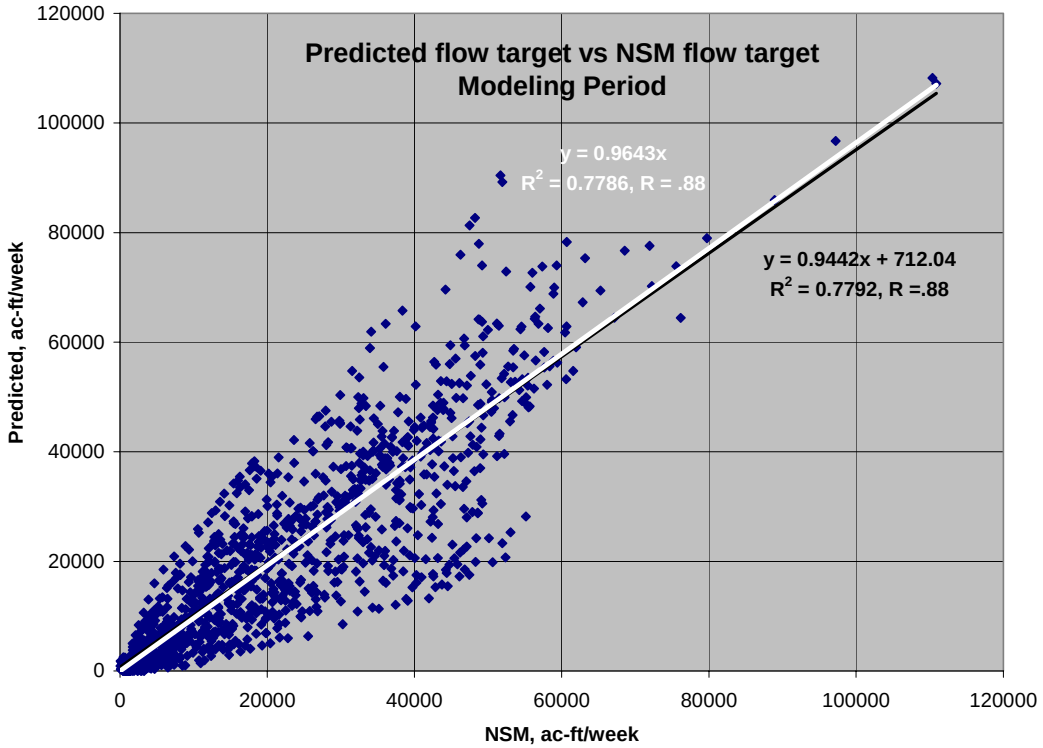

Figure 6. Scatter plot of predicted versus NSM flow targets and trend lines during the modeling period.

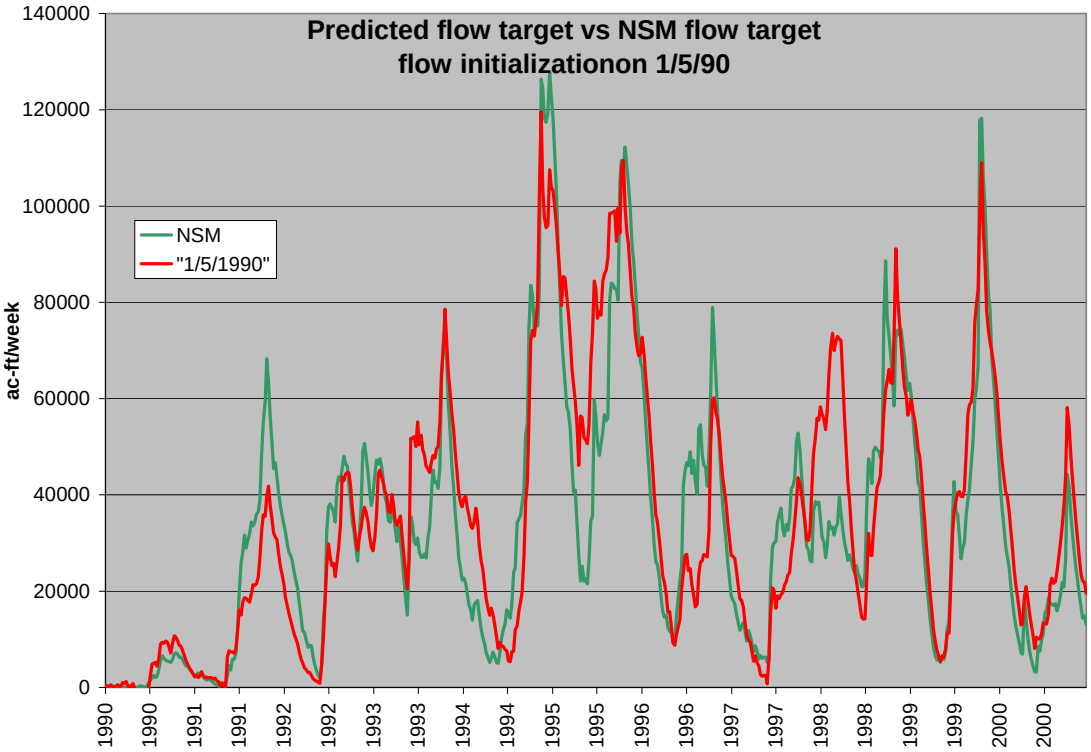

Figure 7. Predicted and NSM flow targets during the validation period.

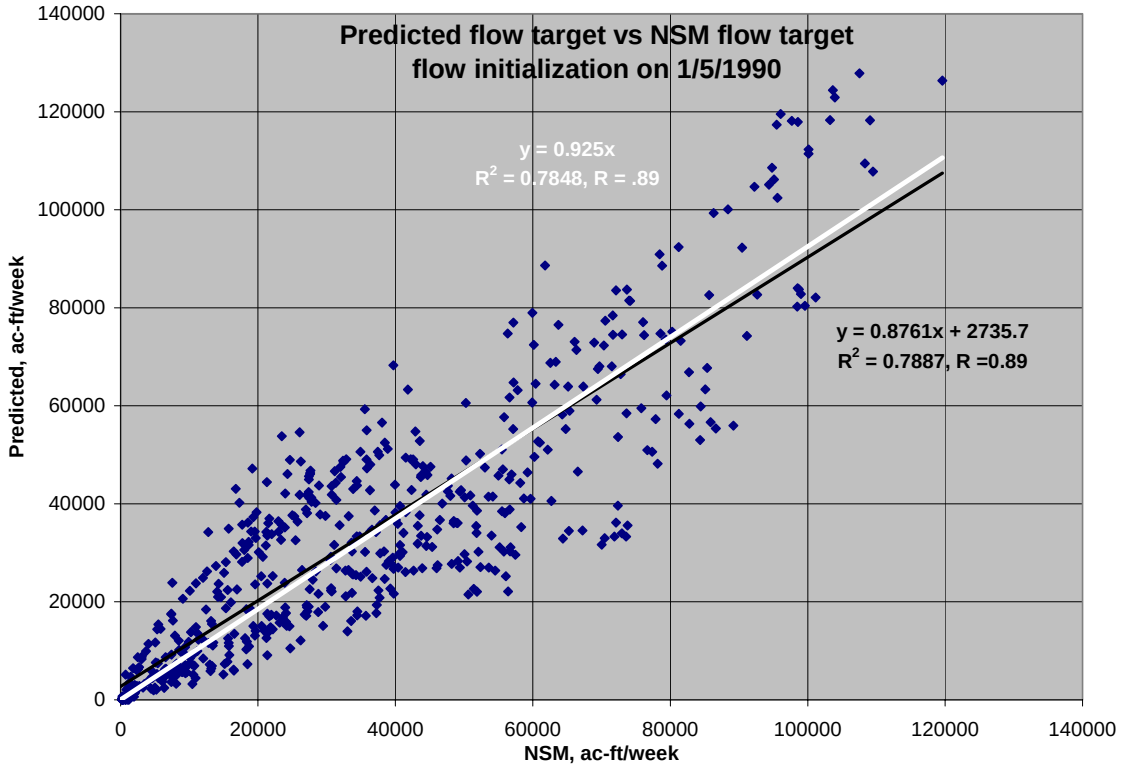

Figure 8. Scatter plot of predicted versus NSM flow targets and trend lines during the validation period.

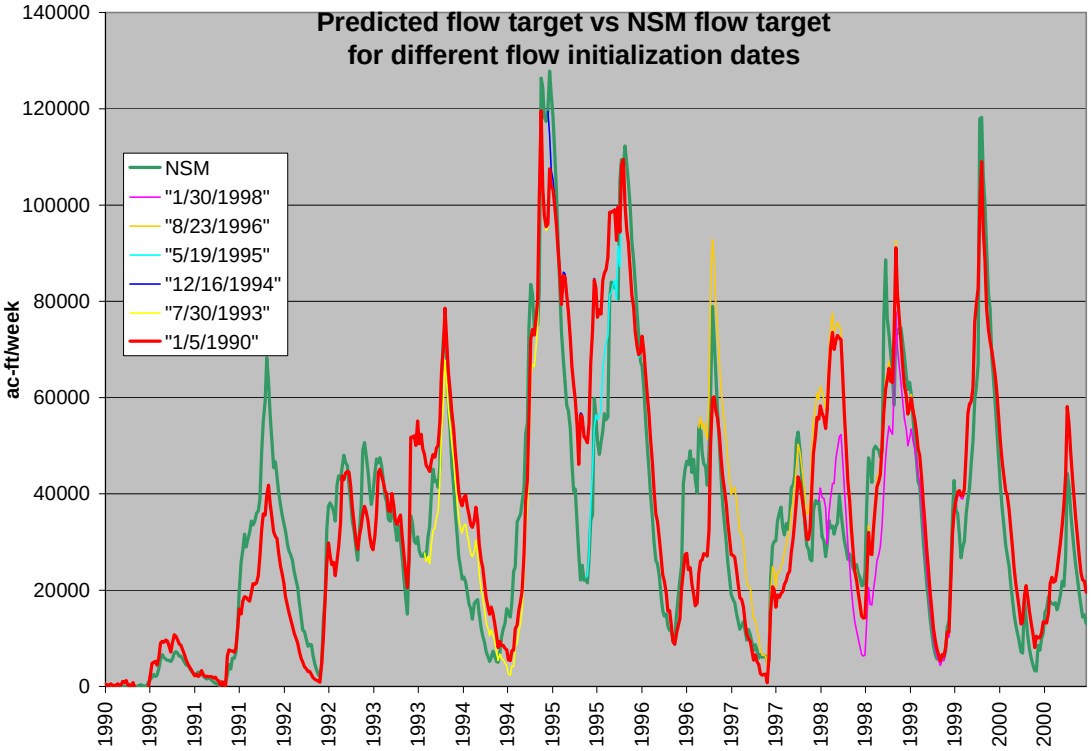

Figure 9. Predicted flow targets initialized at 5 different dates.

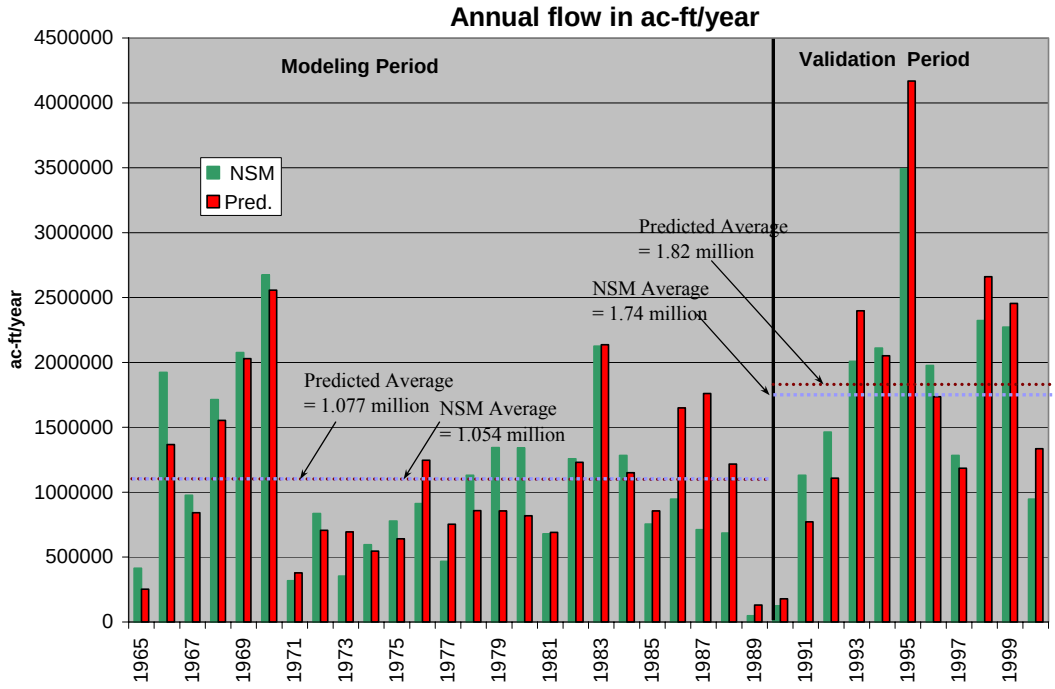

Figure 10. Predicted and NSM annual flow target

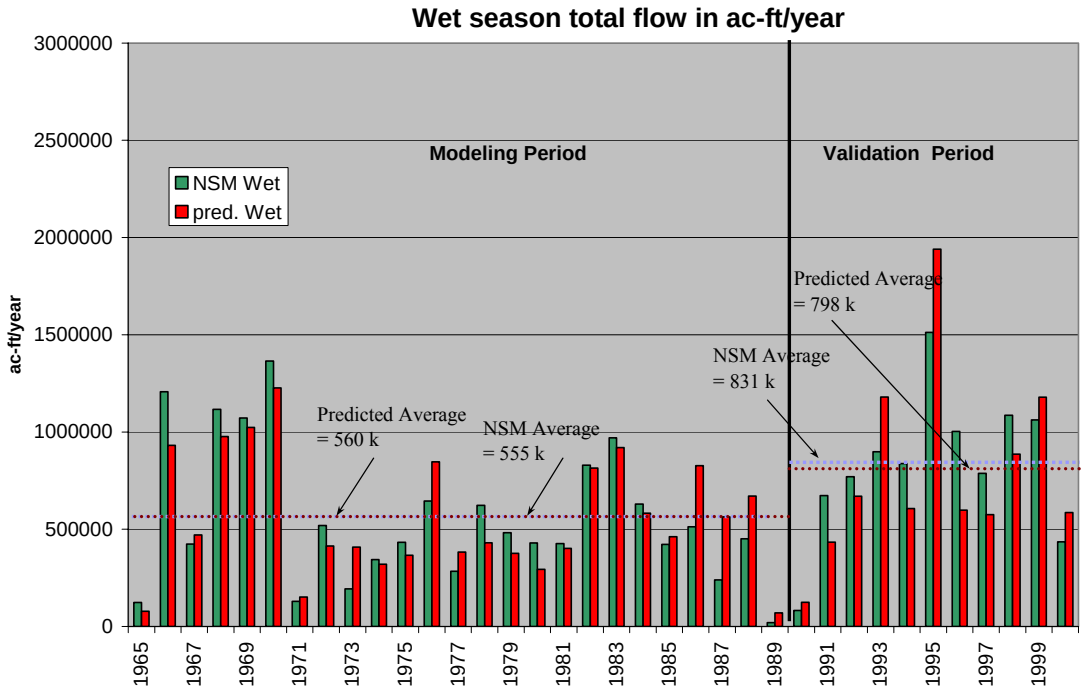

Figure 11. Predicted and NSM wet season (June-October) flow target

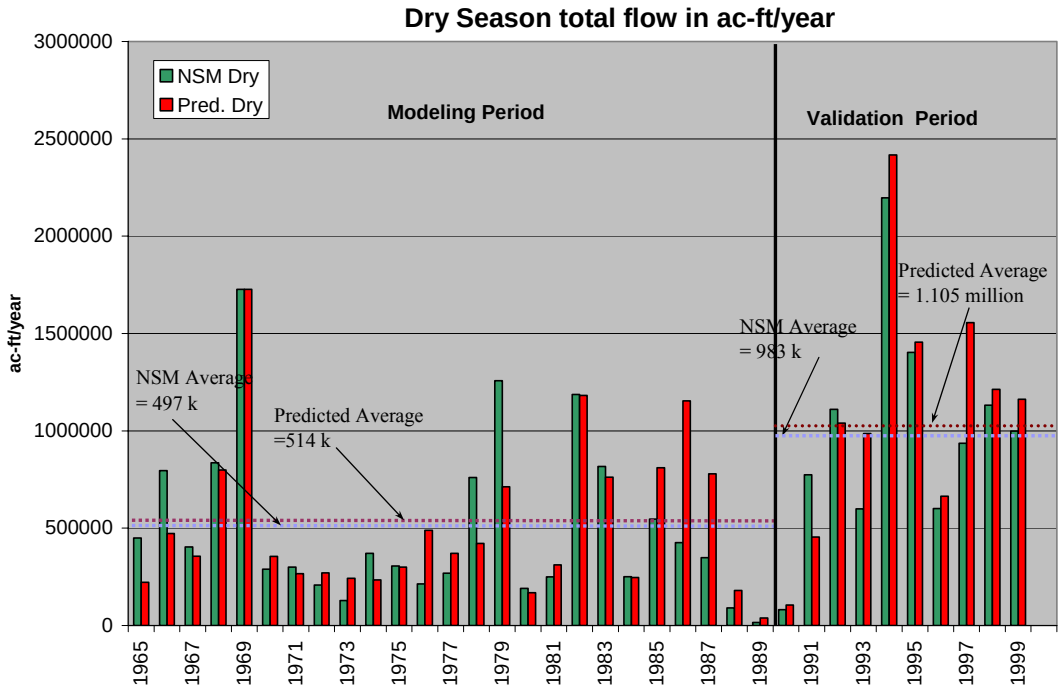

Figure 12. Predicted and NSM dry season (November-May) flow target

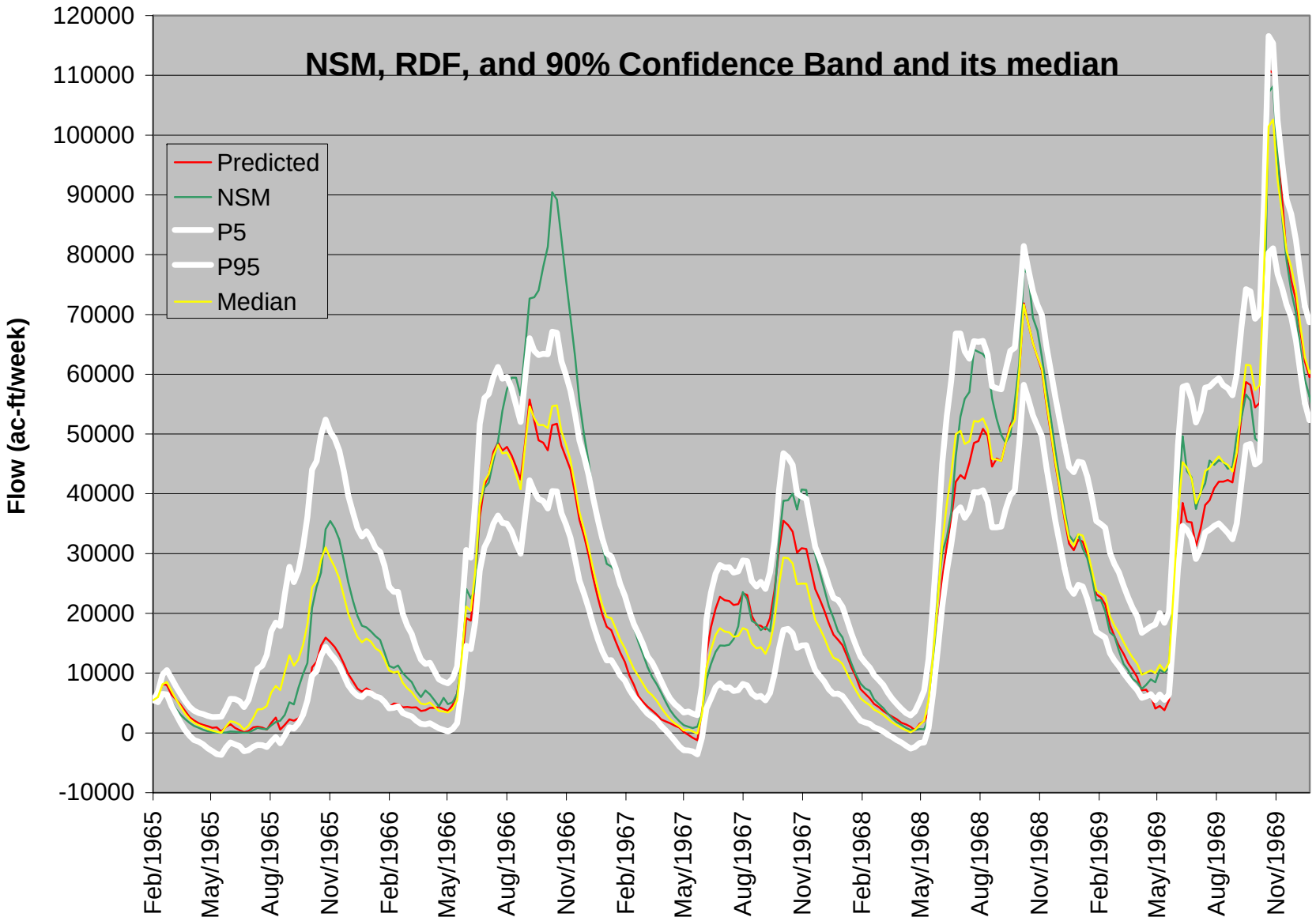

Figure 13a. 90% uncertainty bands around NSM and predicted flow targets for 1965-1969.

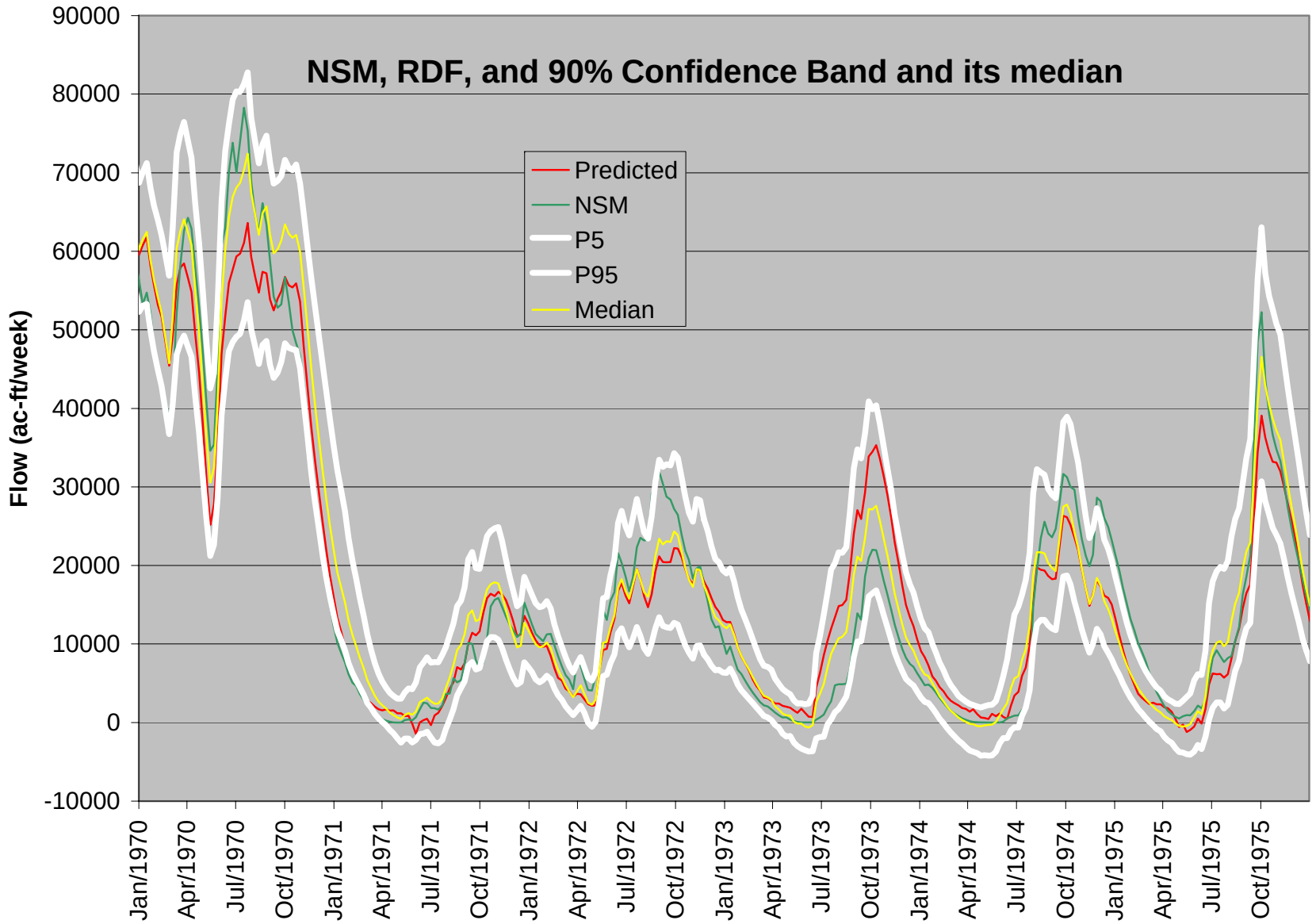

Figure 13b. 90% uncertainty bands around NSM and predicted flow targets for 1970-1975.

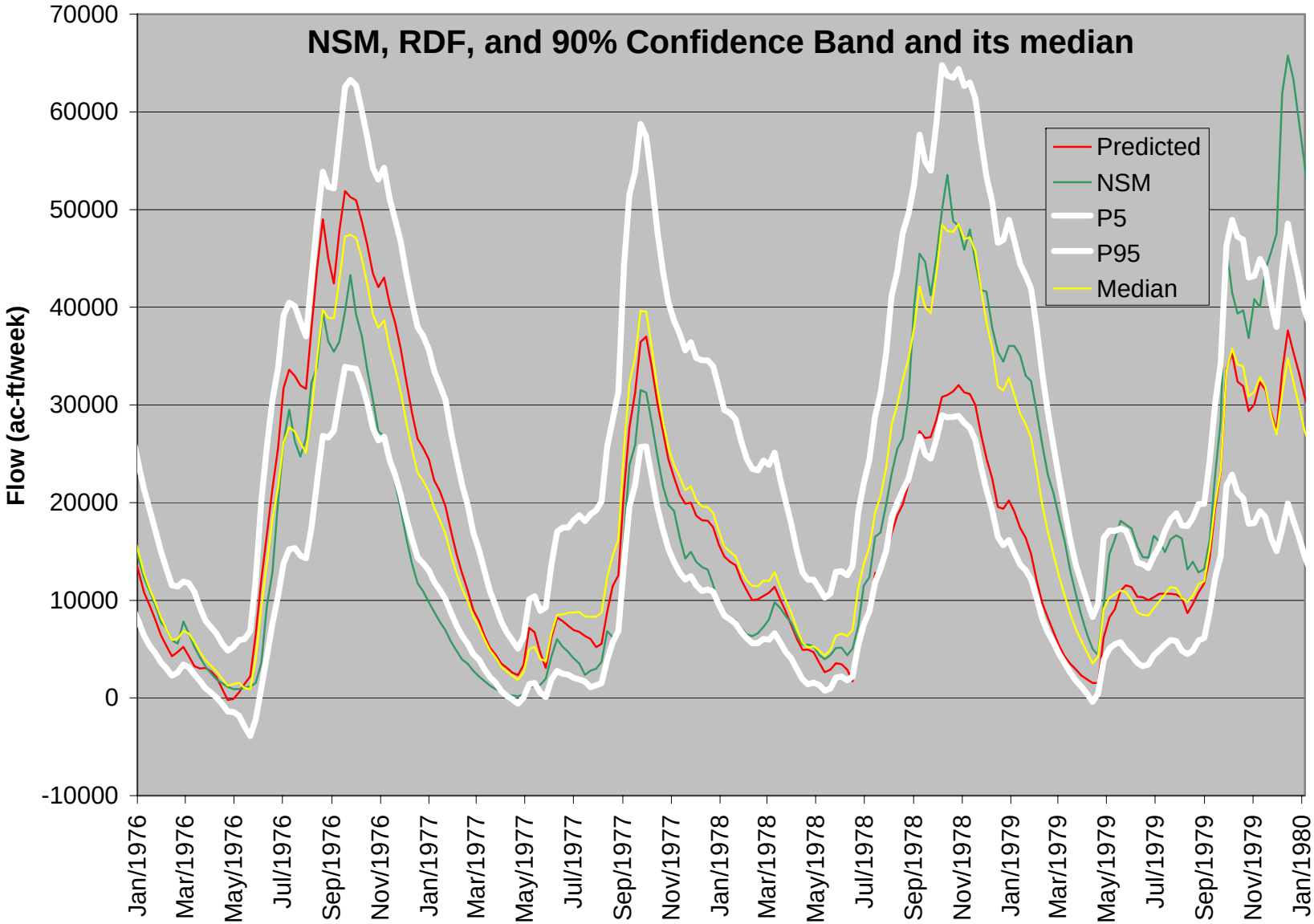

Figure 13c. 90% uncertainty bands around NSM and predicted flow targets for 1976-1980.

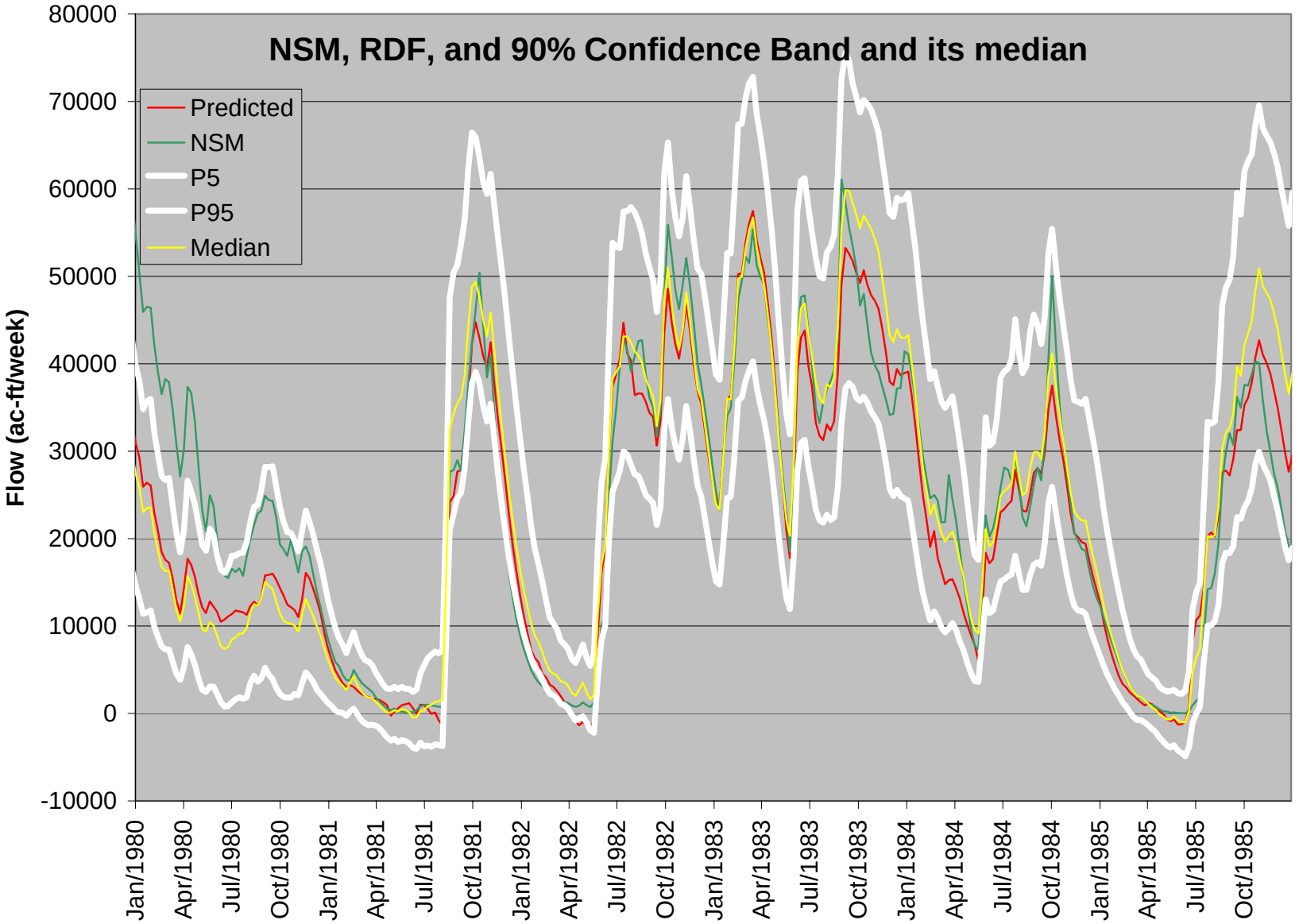

Figure 13d. 90% uncertainty bands around NSM and predicted flow targets for 1980-1985.

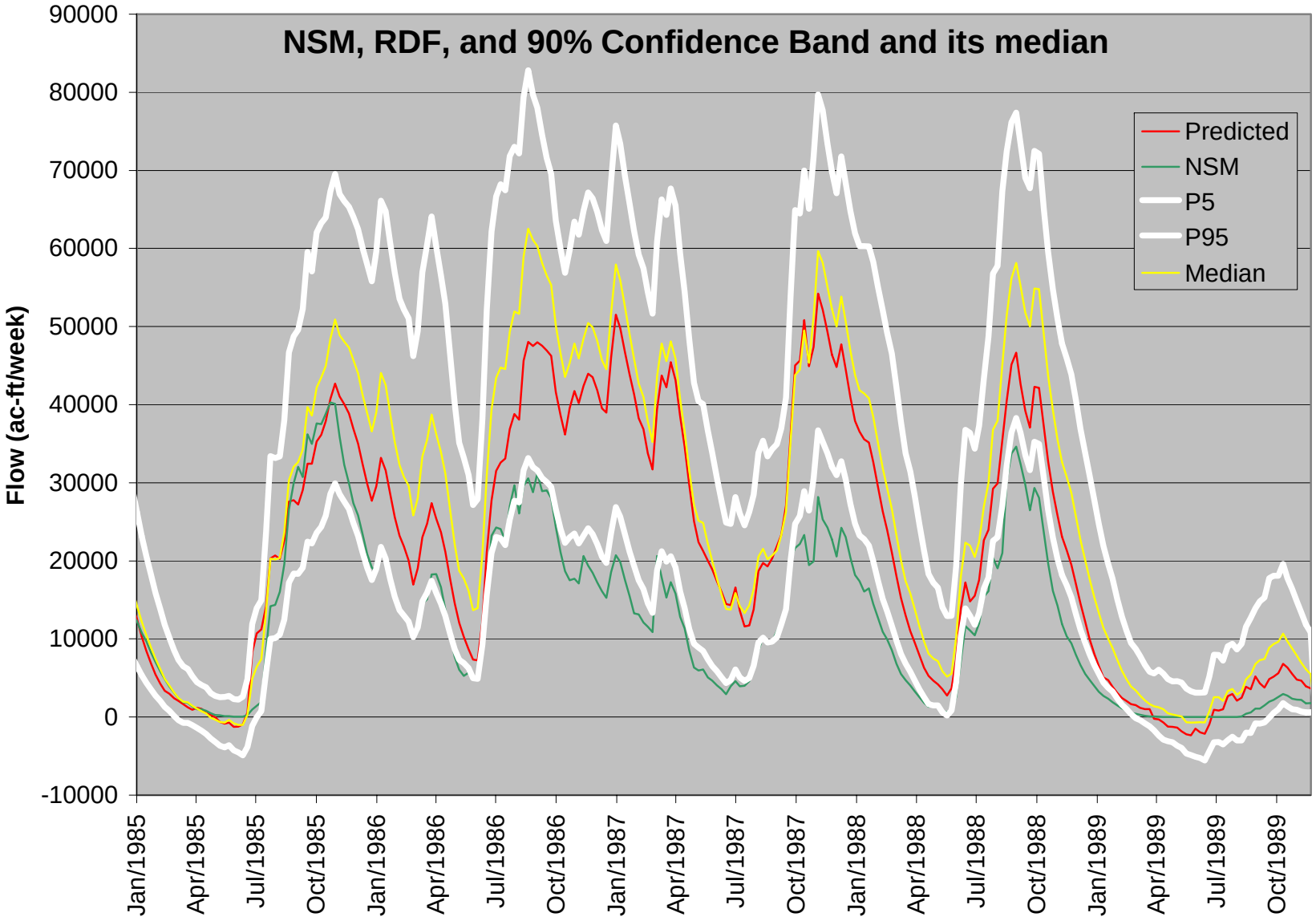

Figure 13e. 90% uncertainty bands around NSM and predicted flow targets for 1985-1990

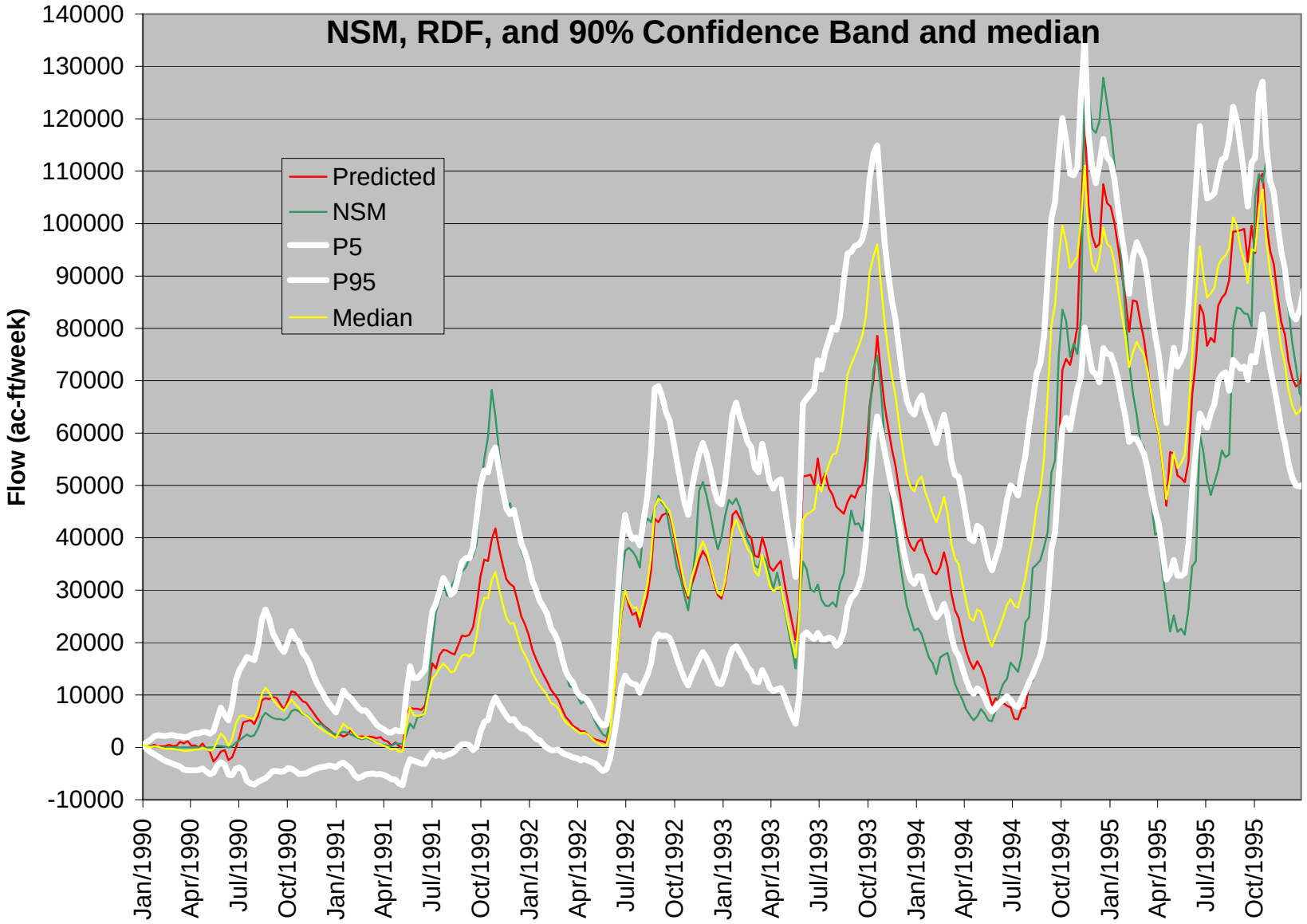

Figure 13f. 90% uncertainty bands around NSM and predicted flow targets for 1990-1995.

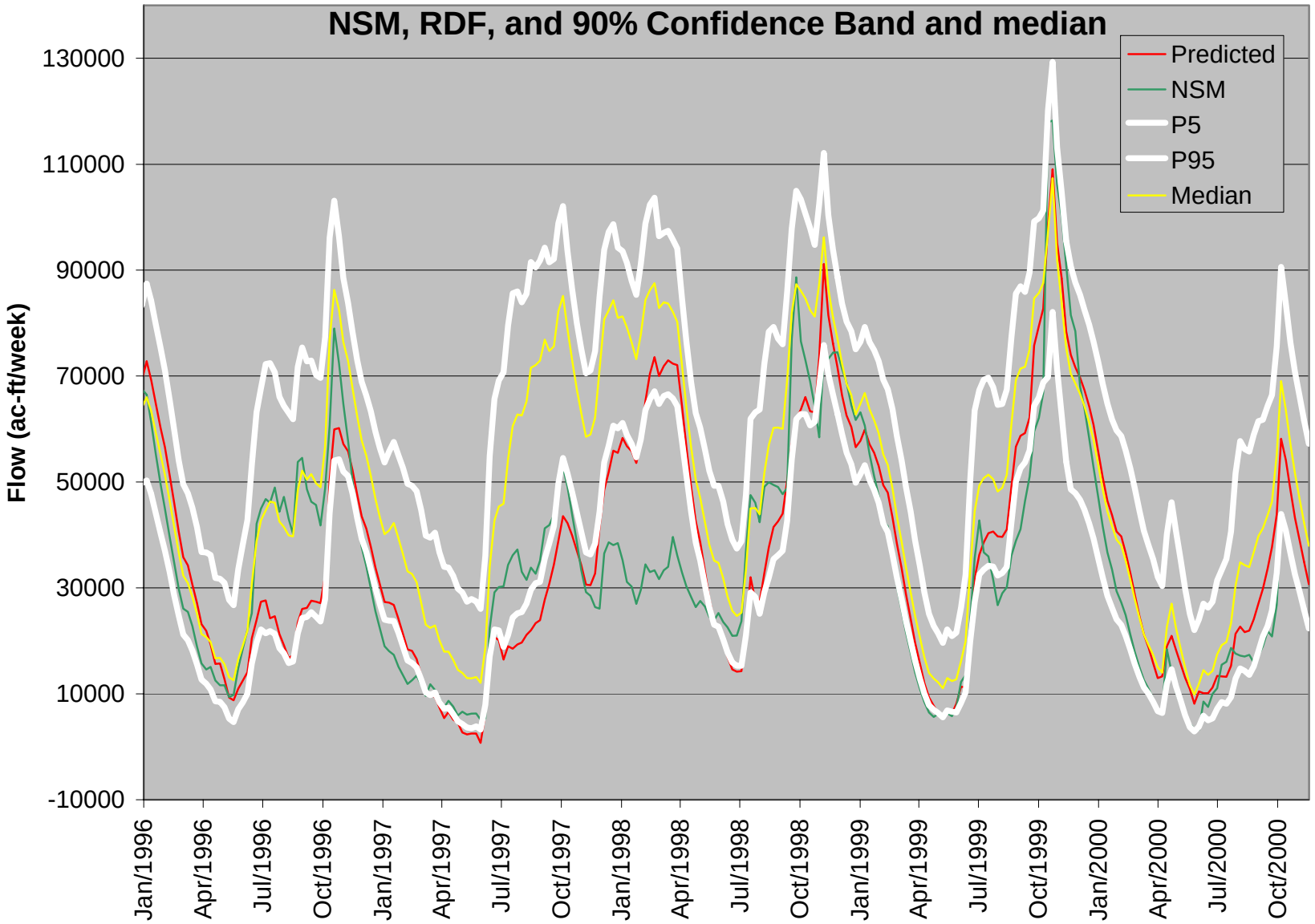

Figure 13g. 90% uncertainty bands around NSM and predicted flow targets for 1995-2000

APPENDIX A

The first 17 Principal Component coefficient Matrix **P**.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	-0.013	0.284	-0.082	0.279	-0.888	0.174	0.006	0.096	-0.017	0.027	-0.018	0.031	-0.032	0.018	-0.034	-0.048	0.013
2	0.120	0.211	-0.033	0.514	0.206	-0.042	0.024	-0.018	0.005	-0.023	-0.003	0.008	-0.019	0.041	0.026	-0.009	-0.087
3	-0.202	-0.247	-0.165	0.191	-0.049	-0.321	0.465	-0.029	-0.071	0.143	-0.305	-0.136	0.305	0.305	-0.393	0.062	0.015
4	-0.210	-0.248	-0.166	0.169	0.053	0.315	-0.291	0.161	0.251	-0.080	0.442	0.149	0.500	0.003	-0.295	-0.031	-0.016
5	-0.214	-0.240	-0.162	0.156	0.075	0.357	-0.429	-0.054	-0.101	0.058	-0.531	-0.192	-0.238	0.170	-0.064	0.002	0.036
6	-0.219	-0.258	-0.154	0.190	-0.068	-0.260	0.139	-0.174	-0.353	0.151	0.385	0.137	-0.101	-0.098	0.206	-0.039	-0.060
7	-0.233	-0.256	-0.170	0.179	-0.024	0.023	-0.179	-0.192	-0.252	0.105	0.083	0.063	-0.183	-0.194	0.238	0.028	0.040
8	-0.203	-0.248	-0.172	0.170	-0.005	-0.075	0.289	0.341	0.536	-0.376	-0.093	-0.029	-0.296	-0.171	0.269	-0.019	0.014
9	0.149	0.229	-0.002	0.480	0.224	-0.028	-0.030	0.006	0.014	-0.001	0.012	-0.005	-0.016	-0.017	0.028	0.003	-0.055
10	-0.206	0.021	0.338	0.093	-0.050	-0.443	-0.317	0.099	0.143	0.182	-0.192	0.272	-0.089	-0.443	-0.351	0.047	-0.003
11	-0.217	0.022	0.330	0.066	0.062	0.316	0.281	0.110	-0.288	-0.111	0.116	-0.519	-0.033	-0.451	-0.240	-0.049	0.027
12	-0.220	0.024	0.318	0.057	0.103	0.417	0.344	-0.130	0.043	0.105	-0.136	0.611	-0.101	0.072	-0.032	-0.001	-0.033
13	-0.225	0.018	0.341	0.088	-0.077	-0.194	-0.201	-0.236	0.342	0.108	0.158	-0.386	-0.007	0.226	0.119	-0.015	-0.060
14	-0.239	0.027	0.343	0.077	-0.022	0.100	0.086	-0.297	0.193	0.067	0.077	-0.073	0.077	0.242	0.243	0.033	0.077
15	-0.210	0.020	0.334	0.072	0.018	-0.167	-0.168	0.505	-0.425	-0.364	-0.041	0.116	0.143	0.346	0.225	-0.015	0.024
16	0.213	0.173	-0.021	0.415	0.193	0.022	0.002	-0.001	-0.013	0.063	0.018	-0.031	0.031	-0.023	-0.013	0.037	0.148
17	-0.234	0.269	-0.163	-0.044	0.043	-0.122	-0.037	-0.372	-0.021	-0.419	-0.066	0.081	0.096	-0.074	-0.077	-0.375	0.566
18	-0.240	0.258	-0.161	-0.095	0.096	0.025	0.031	0.287	0.034	0.334	0.149	-0.038	-0.115	0.074	0.031	0.518	0.551
19	-0.242	0.249	-0.158	-0.095	0.139	0.008	0.039	0.284	0.045	0.442	-0.084	-0.047	0.103	-0.026	0.169	-0.626	-0.114
20	-0.248	0.272	-0.164	-0.058	0.027	-0.019	-0.015	-0.193	-0.029	-0.273	-0.163	0.019	0.220	-0.154	0.071	0.398	-0.322
21	-0.257	0.267	-0.166	-0.067	0.045	0.020	-0.007	-0.019	-0.002	0.067	-0.112	-0.046	0.245	-0.158	0.224	0.142	-0.324
22	-0.236	0.264	-0.163	-0.082	0.097	-0.049	-0.014	-0.003	-0.024	-0.149	0.301	0.010	-0.529	0.323	-0.429	-0.056	-0.320

APPENDIX B

ANN Hidden layer weight matrix Ω and bias vector $\mathbf{A}$

	1	2	3	4	5	6	7	8	9	10
1	0.15957	0.038638	-0.04779	0.040766	-0.15566	-0.01155	-0.08107	-0.01705	-0.01149	0.046406
2	-0.34459	-0.13055	-0.08866	-0.1442	0.25594	-0.09952	0.10276	0.89504	0.1246	0.1827
3	-0.36377	-0.00528	0.002184	0.049159	-0.05911	-0.01603	-0.01456	0.30374	-0.1123	-0.03602
4	-0.02458	-0.1888	-0.0701	-0.13494	0.073944	-0.19072	0.23728	1.4265	0.20154	-0.09212
5	0.053926	0.38304	0.27019	0.3209	-0.7961	-0.0971	-0.37481	-1.0738	-0.4732	-0.55135
6	0.33402	-0.14417	-0.06394	-0.11768	0.014023	-0.22435	-0.00332	0.49459	0.074143	0.37494
7	0.21983	0.066179	-0.01656	-0.05959	0.38465	0.082587	-0.10148	0.065157	-0.26692	0.11548
8	-0.32269	-0.02581	-0.02295	0.03531	-0.09286	0.06375	0.11726	-0.04952	0.20158	-0.46863
9	-0.06532	0.052307	-0.04414	0.070505	-1.1939	0.042399	-0.03195	0.19079	-0.03013	0.074999
10	0.22537	-0.01522	0.028505	-0.06446	0.78575	-0.12499	0.1397	-0.58851	0.004614	0.38477
11	0.11712	-0.07325	-0.00968	-0.00955	0.76359	-0.05403	0.026946	-0.35445	-0.08617	0.42027
12	0.049564	0.033295	0.067016	0.006441	0.13763	0.3777	0.018545	0.34434	0.11417	-0.19321
13	0.056784	-0.08614	0.032133	-0.09347	0.61869	-0.39251	-0.04088	-0.49365	0.12514	0.34044
14	-0.05239	-0.08142	-0.00438	-0.1435	-0.79278	-0.23731	-0.19322	0.15223	-0.17983	-0.28357
15	0.31716	0.12885	-0.02857	0.050985	-0.66503	-0.00609	0.050679	0.18627	-0.22259	-0.47248
16	-0.23685	-0.05021	-0.08761	0.20341	0.33264	0.47941	-0.10138	-0.28271	-0.3285	0.31807
17	0.18162	0.16293	0.005761	0.30327	-0.23751	0.47807	0.37785	-0.12292	0.6398	-0.14358
bias	1.7059	1.7926	-0.62716	-0.08959	0.85179	0.068473	-0.46089	-1.2735	1.3409	-2.7044

ANN output layer weight vector θ and bias term β.

	1	2	3	4	5	6	7	8	9	10
	-0.02228	-1.7765	-1.1622	-1.1063	-0.04118	0.31604	0.6667	0.003239	0.24678	-0.3036
bias	0.7145									

APPENDIX C

Numerical example of flow target prediction for week 1/30/2004 – 2/6/2004

Input data:

Predicted flow on 1/23/2004	PET and rainfall data for the week 1/23/04 through 1/30/04						
	PET	**S336**	**3A-S**	**3A-SW**	**S12D**	**3AS3W3**	**S9**
36862.9	-0.58228	2.84	1.94	2.43	2.85	3.19	1.84

Flow units in ac-ft/week; PET and rainfall in inches/week

Standardize the data

Use Equation 1 to standardize:

- Flow predicted on 1/30/2004 0.88676
- PET and Rainfall for 3 week lags (the third lag is the average of lags 3, 4, 5, and 6)

Week	**PET**	**S336**	**3A-S**	**3A-SW**	**S12D**	**3AS3W3**	**S9**
1/30/04	2.108122	1.307876	0.946734	1.24923	1.410998	1.974665	0.757
1/23/04	1.165716	-0.62666	-0.62195	-0.70376	-0.6267	-0.65156	-0.64442
12/26/04 1/2/04, 1/9/04, 1/16/04	1.568855	-0.89816	-0.86907	-0.9328	-0.83229	-0.82502	-1.06676

Transform data to its first 17 Principal Components

Use Equation 2, and the Principal Component Coefficient Matrix in Appendix A, compute the first 17 principal components of the data set.

	1	2	3	4	5	6	7	8
	1.26129	-2.19193	-1.84856	4.006353	-0.30154	0.096754	-0.15137	-0.29079
9	10	11	12	13	14	15	16	17
-0.58907	0.401413	-0.05308	-0.03594	-0.08637	-0.06467	0.240627	0.120437	0.02755

Apply ANN hidden layer weights and biases

Use equation 3 as applied to equation 9 and hidden weights and biases from Appendix B, compute the output of the ANN hidden weights (10 values)

0.9981	0.8502	-0.6862	-0.4363	0.9146	-0.4244	0.3647	0.9619	0.9730	-0.9953

Apply ANN output layer weights and bias

Use Equation 4 as applied to equation 9 and the output layer weights and bias in Appendix B, compute the flow for the week starting on 2/6/2004

Predicted weekly flow target for week starting on 2/6/2004 is **40455.7 ac-ft/week**