4. EVALUATION OF AUXILIARY PROCEDURES

Modeled runoff and loads from AGNPS and four auxiliary procedures were compared with observed runoff and loads. Composite period comparisons were made between the alternative parameterization data sets, AG0cp and AG1cp, and the observed composite period data set, OWcp. Monthly comparisons were made between the alternative monthly simulation data sets, AG1mn and AG2mn, and the observed monthly data set, OWmn.

4.1 Composite Period Evaluation Measures

Nonparametric correlation was assessed using Spearman's rank correlation coefficient. For this analysis, the data set for each modeling procedure-parameter combination was ranked, and the rank orders of the observed and modeled output analyzed. These correlation coefficients for composite period data are recorded in Table 4-1.

Table 4-1. Correlation Between Composite Period Modeled (AG) and Monitored Data (OW) -- Spearman's Rank Correlation Coefficients

	------- ------- OWcp ------- -------			
Data Set	**Runoff**	**TN**	**TP**	**SS**
AG0cp	0.385	0.267	0.243	0.356
AG1cp	0.439	0.306	0.311	0.434

Hypothesis tests were performed on the medians of paired differences between the observed and modeled data for each composite period auxiliary procedure-parameter combination using the non-parametric Wilcoxon signed rank test. The null hypothesis for each of these tests was that no statistical difference existed between the observed and modeled data at a 95% confidence level. If a no-difference condition (0) was included within the confidence interval for the median, the null hypothesis was rejected. The results of this series of tests are shown in Table 4-2 for composite period data.

Table 4-2. Hypothesis Tests on Parameters from Auxiliary Parameterization Data Sets (Composite Period Basis)

Parameter	**OWcp - AG0cp**				**OWcp - AG1cp**			
	Median	**95% Confidence Interval**		**Result**	**Median**	**95% Confidence Interval**		**Result**
Runoff	0.902	0.580	1.298	R	0.898	0.586	1.294	R
Nitrogen	0.084	0.015	0.163	R	0.080	0.009	0.157	R
Phosphorus	-0.067	-0.0840	-0.0515	R	-0.0745	-0.0895	-0.0600	R
Suspended Sed.	-26.5	-35.9	-16.4	R	-25.3	-33.8	-14.5	R

Additionally, the median absolute error, robust coefficient of determination and robust model efficiency goodness-of-fit measures were calculated for each composite period auxiliary procedure-parameter combination. Quattro spreadsheet functions were used to perform the data

sorting and ordering required for these order-based measures. Goodness-of-fit measures are reported in Table 4-3 for composite period comparisons.

Table 4-3. Goodness-of-Fit Measures: Composite Period Data Sets

	Best Fit	AG0cp				AG1cp			
		RO	**TN**	**TP**	**SS**	**RO**	**TN**	**TP**	**SS**
Median Absolute Error	0	74.4	60.1	255.2	167.8	71.5	62.2	293.1	187.8
Robust Coeff. of Determination	1	0.91	1.42	0.35	0.32	0.94	1.82	0.30	0.32
Robust Model Efficiency	1	-0.13	0.13	-1.96	-1.71	-0.08	0.10	-2.40	-2.03

AG0cp = composite period recommended parameterization procedures for AGNPS
AG1cp = composite period procedures for AGNPS with temporal variability for various inputs
RO = runoff, **TN** = total nitrogen, **TP** = total phosphorus, **SS** = suspended sediment

Bi-variate plots of composite period observed and modeled data were created in Figures 4-1 to 4-4 to assist in comparing monitored and modeled data.

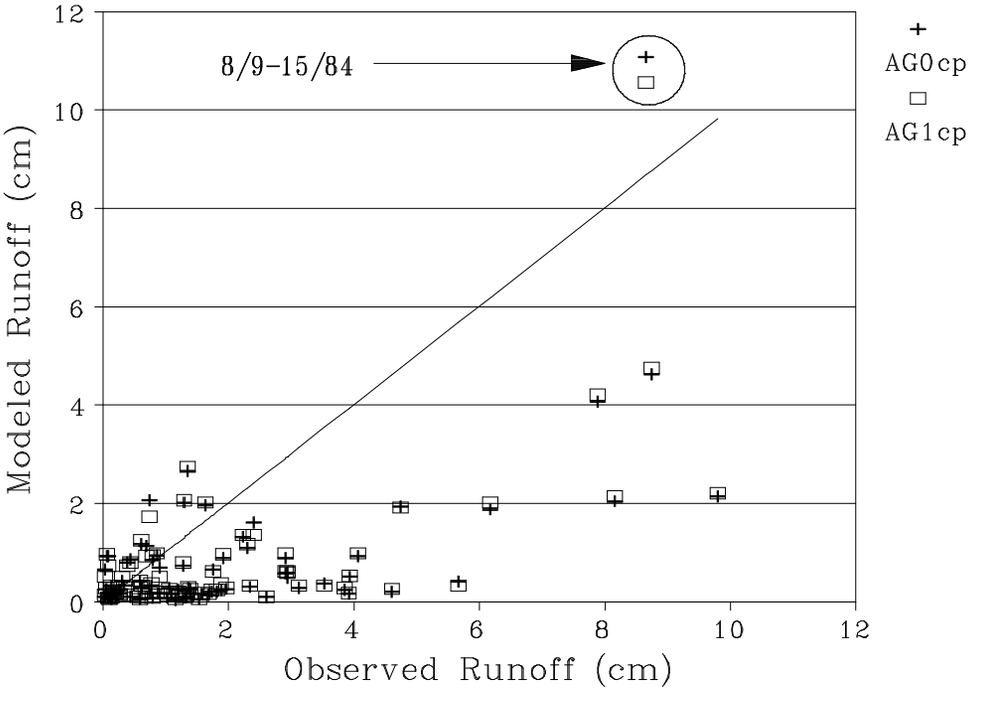

Figure 4-1. Composite Period Runoff Bi-Variate Plot

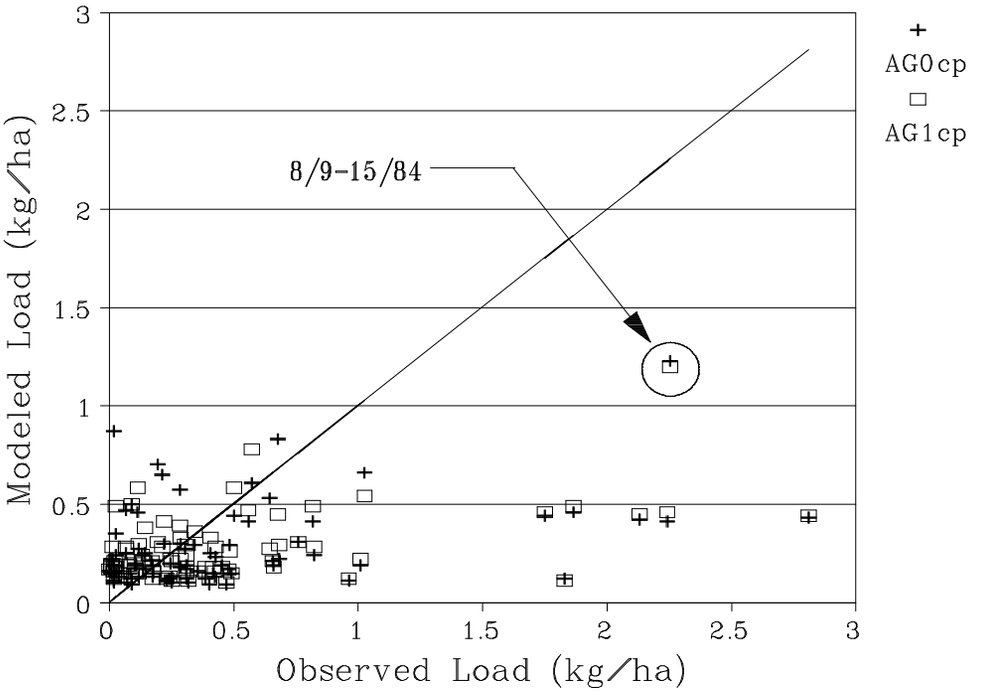

Figure 4-2. Composite Period Total Nitrogen Bi-Variate Plot

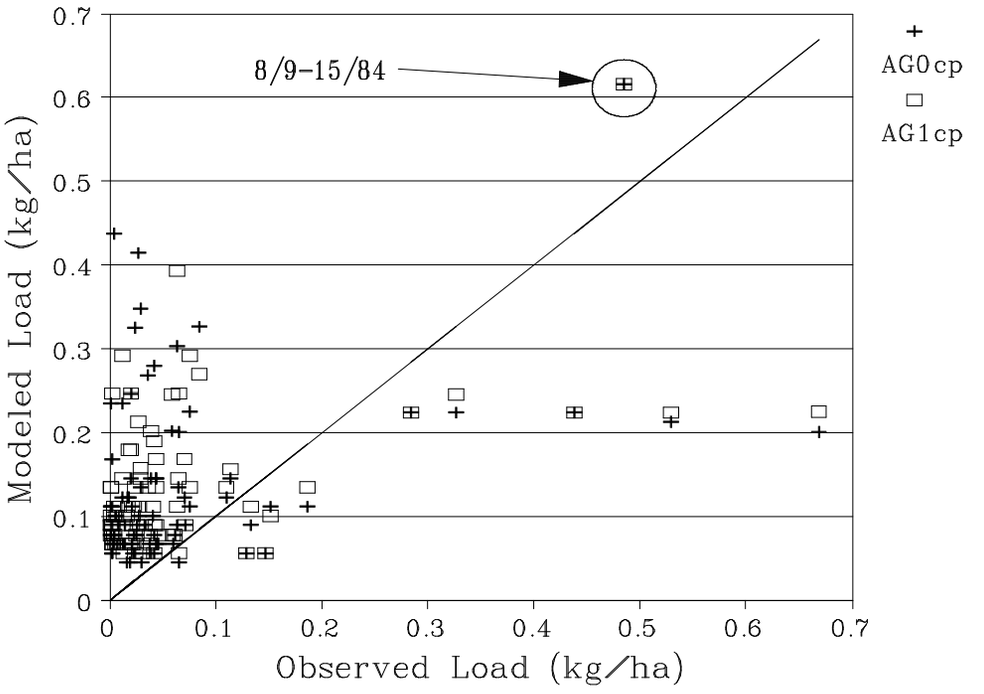

Figure 4-3. Composite Period Total Phosphorus Bi-Variate Plot

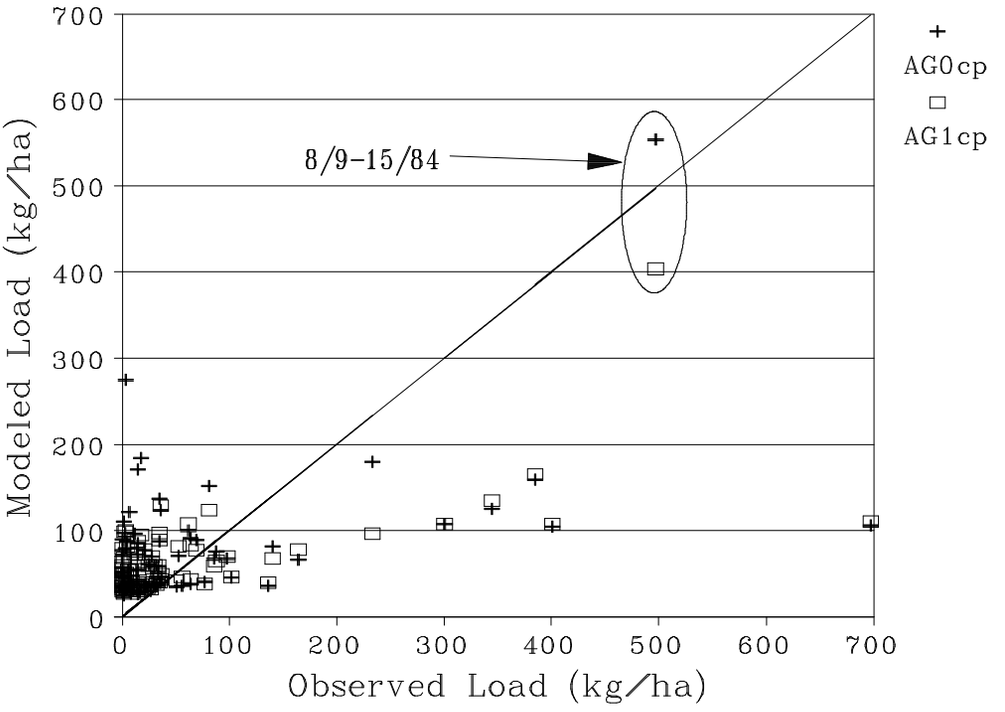

Figure 4-4. Composite Period Suspended Sediment Bi-Variate Plot

4.1.1 Composite Period Runoff

Composite period runoff had the highest correlation of any parameter from each composite period data set using Spearman's rank correlation coefficient. The correlation was 14% higher using the AG1cp procedures than with the AG0cp procedures. The Wilcoxon signed rank test results in Table 4-2 showed essentially no difference between the two sets of paired differences with observed runoff. However, modeled runoff from both composite period data sets was statistically lower than observed runoff at the 95% confidence level. This difference is readily apparent in the composite period runoff plot in Figure 4-1, which shows that runoff from either of the alternative parameterization procedures is only about 1/2 of the observed runoff, with one exception. Contrary to this observation, the robust coefficient of determination (CD*) for modeled runoff from both data sets indicated a fairly good fit with observed runoff. The AG1cp data set produced slightly better fits with observed data according to the median absolute error (MdAE) and CD* in Table 4-4, and a slightly poorer fit using the robust model efficiency (EF*) measure, though all were comparable.

4.1.2 Composite Period Total Nitrogen (TN)

Modeled TN using the AG1cp parameterization procedure produced the lowest correlation coefficient of all four parameters, and also showed the smallest increase between the two alternative composite period data sets due to the temporal variability enhancements in AG1cp. The median differences between the two sets of paired differences in Table 4-2 were

comparable, and both data sets showed a significant underprediction relative to the observed TN loads. All of the goodness-of-fit (GOF) measures indicated a slightly better fit with the AG0cp data set than with the AG1cp data set, though neither set produced any measures close to that of a perfect fit. TN does not appear to be modeled very well by either set of composite period parameterization procedures as indicated both by the poor correlation coefficients and the essentially constant range of modeled values with increasing observed load. All of the GOF measures indicated a generally poor fit with observed TN.

4.1.3 Composite Period Total Phosphorus (TP)

The lowest Spearman's correlation coefficient reported in Table 4-1 for any composite period auxiliary procedure - parameter combination was for TP using the typical parameterization procedure. The coefficient for TP in the AG1cp data set also was fairly low, even though a 28% increase was seen over TP in the AG0cp data set. The results for the Wilcoxon test in Table 4-2 show that both composite period auxiliary procedures overpredicted TP significantly at the 95% confidence level, with the AG1cp procedure predicting TP loads about 11% greater than with the AG0cp procedure. A look at the composite period TP plot in Figure 4-4 shows a large variability of overpredicted TP loads at lower values of observed loads and underpredicted loads at higher observed loads. The composite period TP loads were almost constant around 0.22 kg/ha at higher observed loads from both modeling procedures. One storm from 8/9-15/84 stands as an exception to the last statement, where the modeled TP load was almost three times the approximately constant modeled rate observed with other high observed loads. The GOF measures from both data sets were comparable. None of the GOF measures showed a very good fit, although overall measures for the AG0cp data set indicated slightly better fits than for the AG1cp data set. The poor TP modeling may be due to inappropriate default soil or water phosphorus contents, phase partitioning coefficients, or in-stream decay coefficients.

4.1.4 Composite Period Suspended Sediment (SS)

SS showed the highest correlation increase between the AG0cp and the AG1cp data sets, with a Spearman's correlation coefficient for SS using the AG1cp procedures almost as good as that for runoff. The results from the Wilcoxon test on paired differences in Table 4-2 show a significant overprediction of SS loads with both modeled data sets. The statistics were undoubtedly influenced by the 8/9-15/84 period that produced the highest TP load. In contrast to the better correlation shown for SS from the AG1cp procedure, all three GOF measures indicated a slightly better fit using the AG0cp procedure, than with AG1cp, though none of the measures indicated a good fit with observed data. The increase in correlation when using the AG1cp procedure was probably due to the underestimation of the highest modeled SS load, relative to the AG0cp procedure.

4.1.5 Comparison Between Composite Period Modeling Procedures

Overall, the AG1cp time-variable parameterization procedure produced higher correlations in Table 4-1 for all parameters than did the AG0cp procedure. The Wilcoxon Signed Rank test results in Table 4-2 showed significant differences between all observed and modeled parameter medians at the 95% confidence level with both data sets, indicating that neither parameterization

procedure estimated monitored runoff and loads very well. All of the composite period parameter plots, Figures 4-1 to 4-4, showed the following similarities with minor variations. The majority of data was clustered at the low end of the scale. Modeled data underestimated observed data for higher observed runoff and loads, with one large event overpredicted by the modeling procedures. High values of modeled runoff and all loads resulted from the 8/9-15/84 composite period in both modeled data sets. In looking closer at the data, this composite period actually represented data from 5 storm events and 5 runoff events, where rainfall could not be uniquely apportioned to one storm or the other and so were lumped together. This data point, therefore, may be inappropriate to include in the analysis, as it exerted a large influence, and is artificially high because of contributions from several consecutive storm and runoff events. At low observed values, modeled runoff and TN values were generally centered around the 1:1 line, while modeled TP and SS loads were generally overestimated by both sets of procedures. At higher observed values, data from both parameterization procedures underestimated parameter values with almost constant load values.

The AG1cp procedure was expected to produce better correlations because the majority of time-variable methods incorporated are used in existing models to provide intra-year distributions of annualized parameter values. The event EI regression is an exception to this statement, and is the most likely source of this poorer correlation. The AG0cp procedure was expected to produce some loads greater than, and some loads smaller than, those produced by the AG1cp procedure, since AG0cp was using average annual parameter values. This fluctuation was in fact observed in each of the composite period plots. Statistics were recalculated without the cumulative 8/9-15/84 data to further assess its impact. Without the 8/9-15/84 storm, each of the correlation coefficients was reduced by approximately 0.020, otherwise all changes were minor, and no hypothesis test results were changed.

4.2 Monthly Period Comparisons

The Spearman's rank correlation coefficients for monthly data sets are recorded in Table 4-4. The results of hypothesis tests on the medians of monthly period paired differences between the observed and modeled data are shown in Table 4-5. The goodness-of-fit measures -- median absolute error, robust coefficient of determination and robust model efficiency -- were calculated for all monthly auxiliary procedure-parameter combinations and reported in Table 4-6 for monthly period comparisons. Differences between output from monthly simulation procedures AG1mn and AG2mn are illustrated in Figures 4-5 to 4-8, while cumulative plots of modeled and observed data for the 23 complete months, are shown in Figures 4-9 to 4-12.

Table 4-4. Correlation Between Alternative Monthly Modeled (AG) and Monitored Data (OW) -- Spearman's Rank Correlation Coefficients

	------- ------- OWmn ------- -------			
Data Sets	**Runoff**	**TN**	**TP**	**SS**
AG1mn	0.337	0.270	0.270	0.419
AG2mn	0.739	0.575	0.319	0.419

Table 4-5. Hypothesis Tests on Parameters from Auxiliary Monthly Simulation Data Sets

Parameter	OWmn - AG1mn				OWmn - AG2mn			
	Median	95% Confidence Interval		Result	Median	95% Confidence Interval		Result
Runoff	1.61	0.75	2.70	R	0.28	-0.39	0.71	A
Nitrogen	0.135	-0.041	0.335	A	0.076	-0.065	0.245	A
Phosphorus	-0.0963	-0.1405	-0.0731	R	-0.1003	-0.1431	-0.0775	R
Suspended Sed.	-40.0	-54.5	-21.0	R	-40.0	-54.5	-21.0	R

Table 4-6. Goodness-of-Fit Measures: Monthly Period Data Sets

	Best Fit	AG1mn				AG2mn			
		RO	TN	TP	SS	RO	TN	TP	SS
Median Absolute Error	0	46.6	67.3	302.7	296.3	23.3	45.3	306.2	296.3
Robust Coeff. of Determination	1	0.76	0.95	1.27	1.10	0.88	0.97	1.27	1.10
Robust Model Efficiency	1	0.84	0.88	0.82	0.91	0.92	0.92	0.82	0.91

AG1mn = monthly aggregates of all AG1cp loads falling within any given month
AG2mn = AG1mn monthly loads supplemented with monthly baseflow and septic system TN and TP loads
RO = runoff, **TN** = total nitrogen, **TP** = total phosphorus, **SS** = suspended sediment

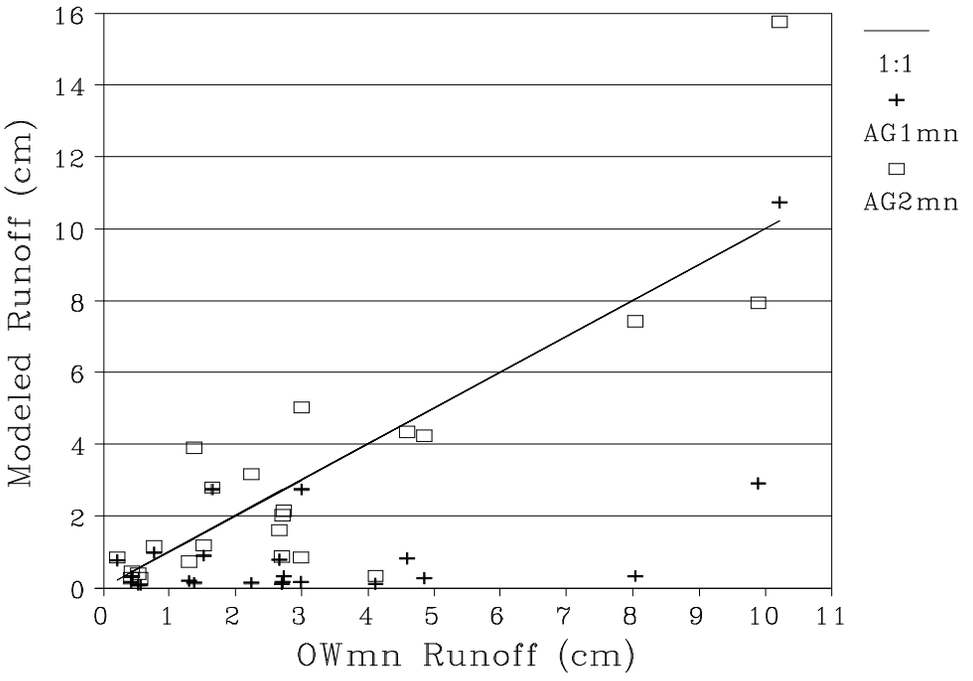

Figure 4-5. Monthly Runoff Bi-Variate Plots

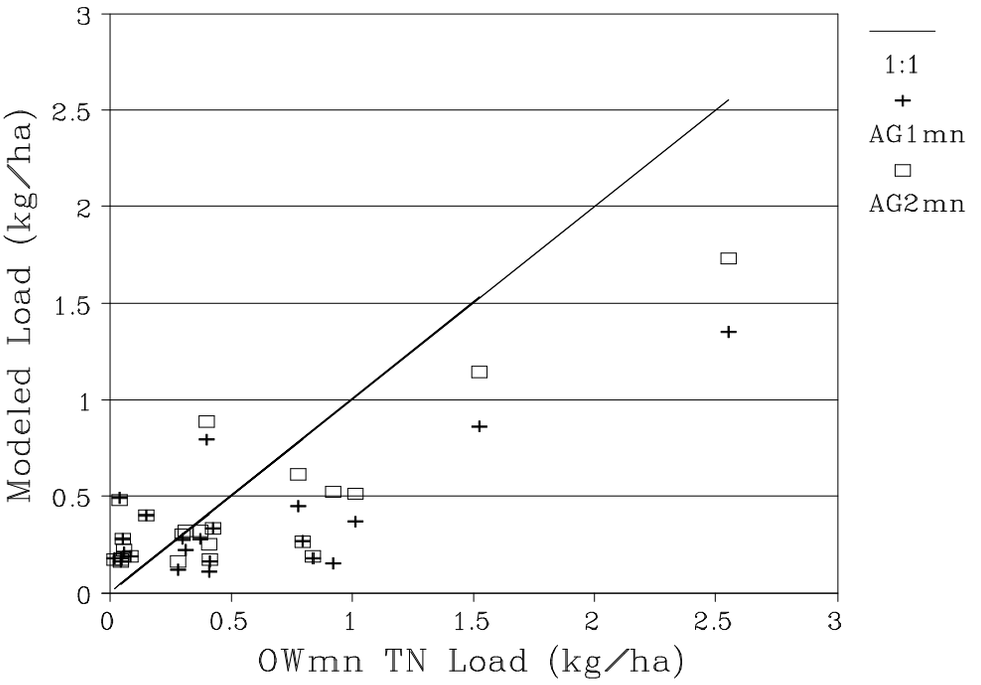

Figure 4-6. Monthly Total Nitrogen Bi-Variate Plots

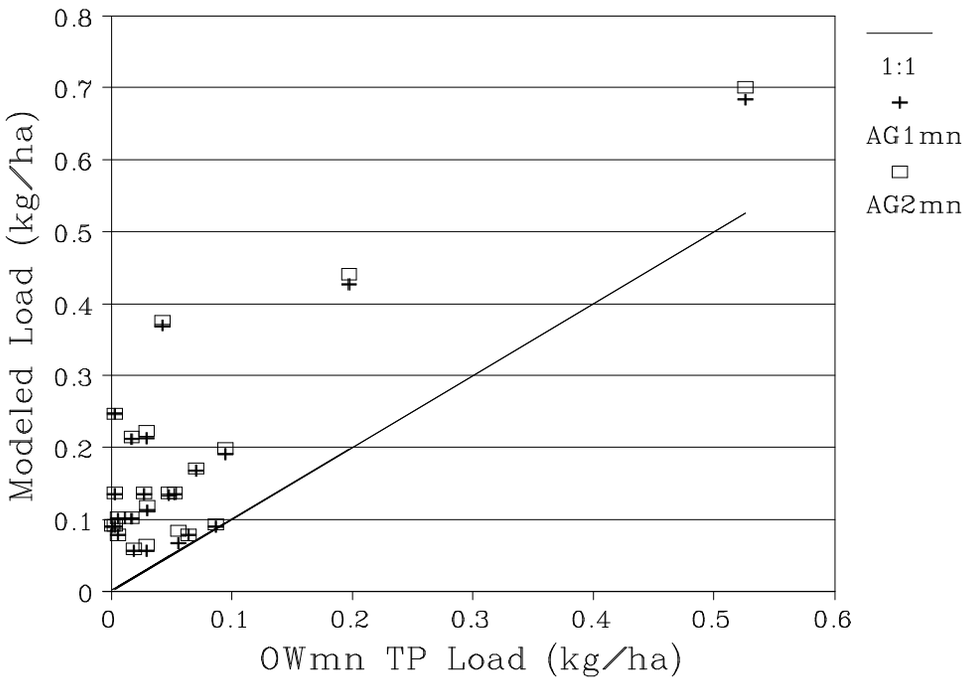

Figure 4-7. Monthly Total Phosphorus Bi-Variate Plots

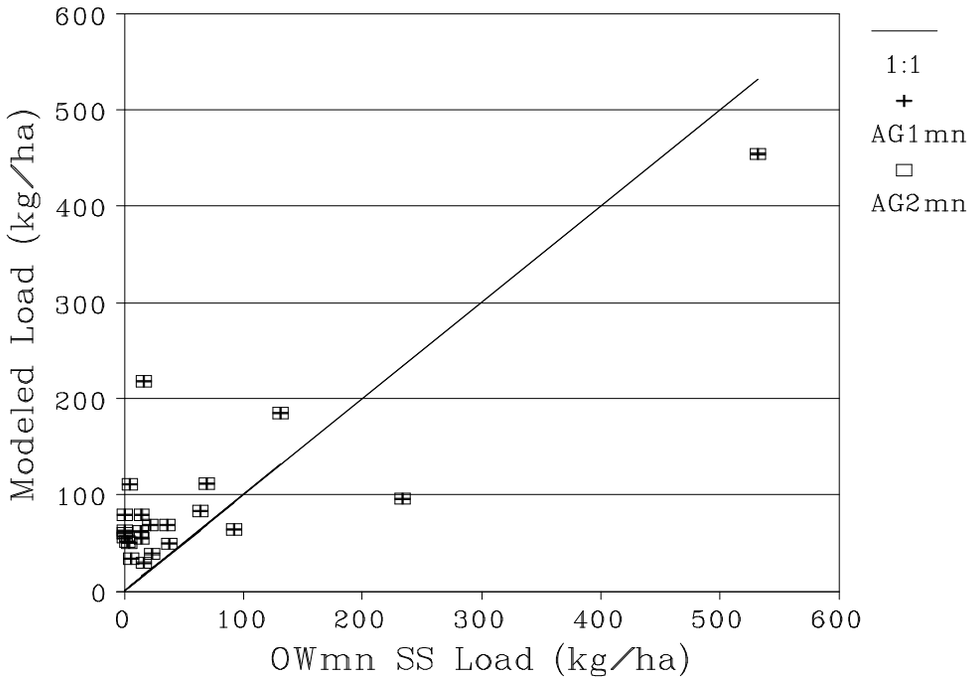

Figure 4-8. Monthly Suspended Sediment Bi-Variate Plots

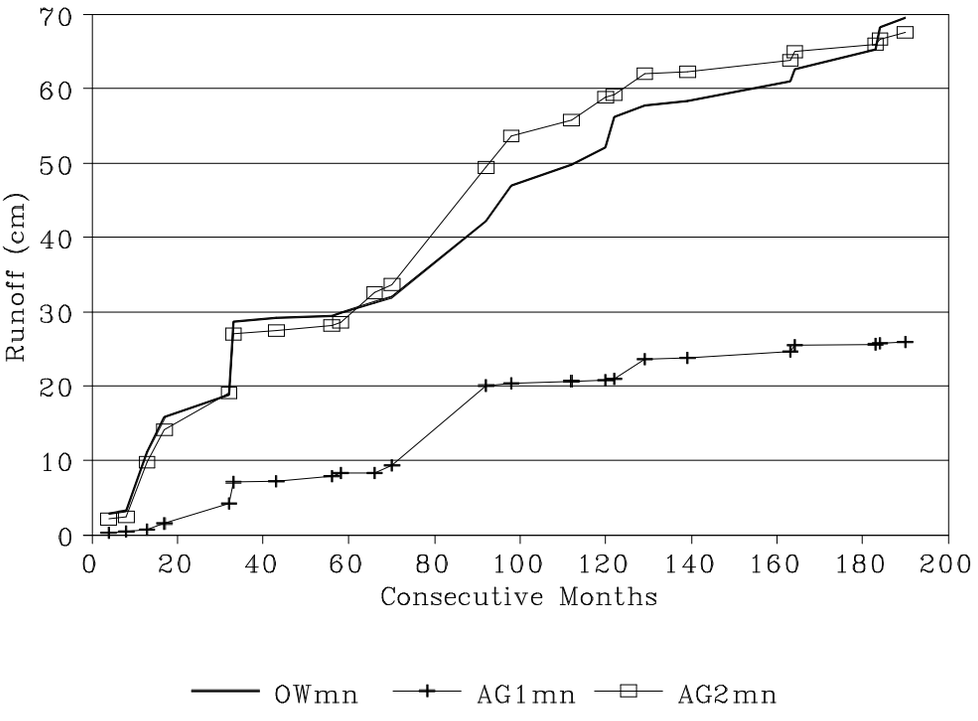

Figure 4-9. Cumulative Monthly Runoff

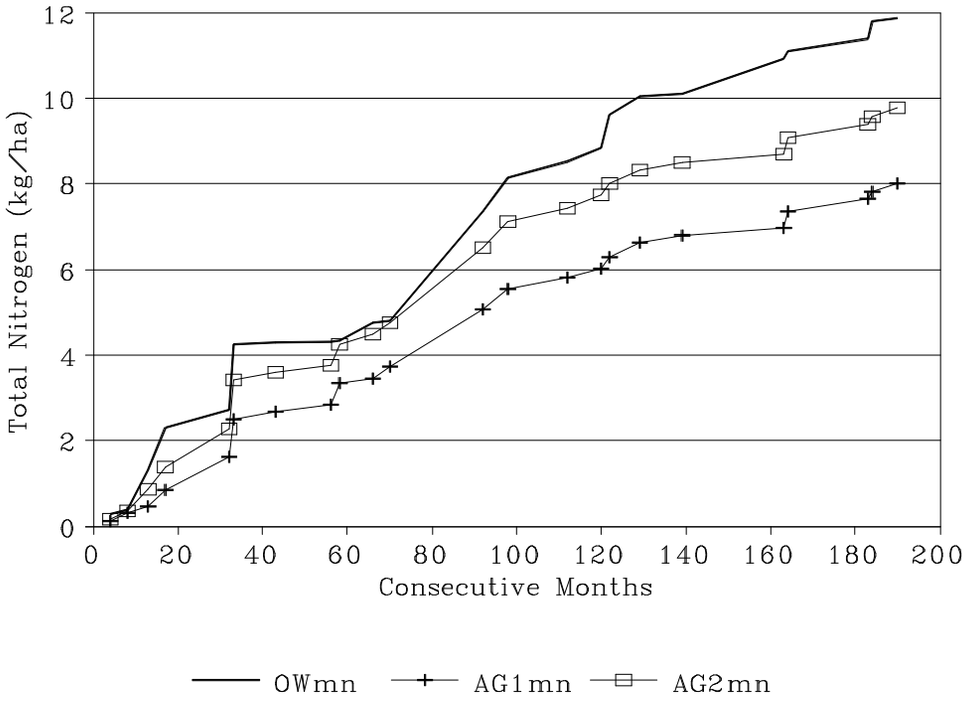

Figure 4-10. Cumulative Monthly Total Nitrogen

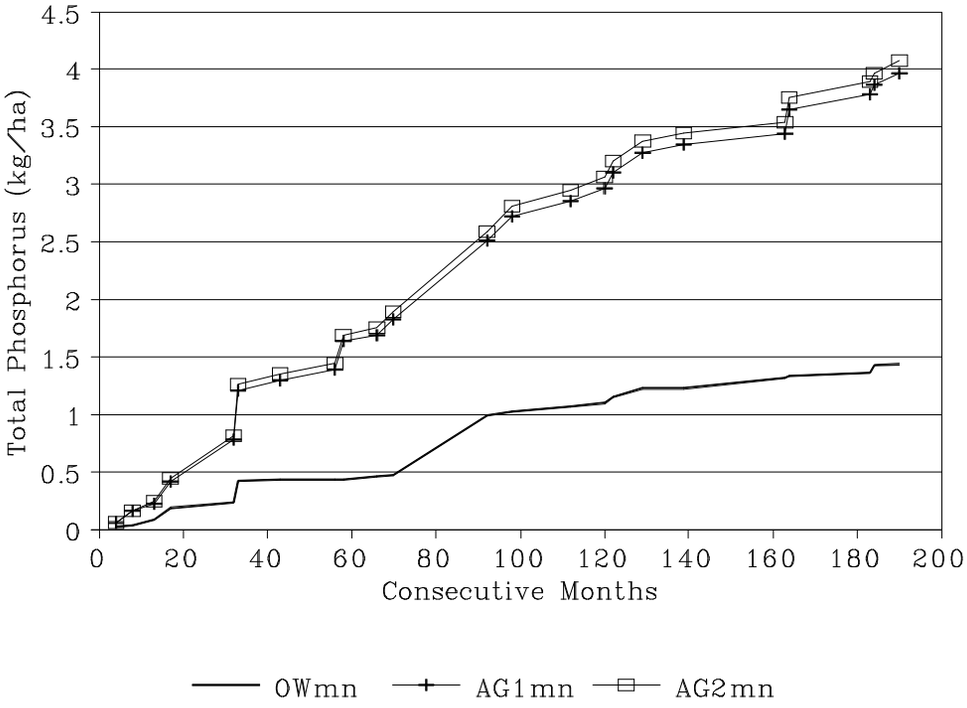

Figure 4-11. Cumulative Monthly Total Phosphorus

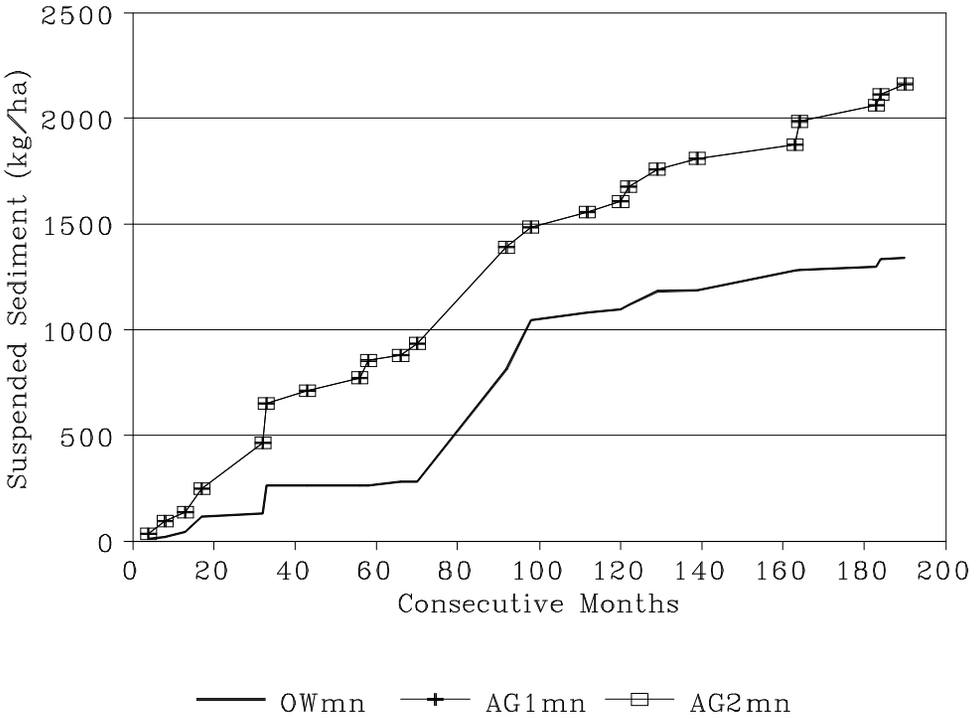

Figure 4-12. Cumulative Monthly Suspended Sediment

4.2.1 Monthly Period Runoff

Spearman's correlation coefficient in Table 4-4 more than doubled for monthly runoff between the AG1mn and AG2mn data sets. This was the greatest correlation increase for any parameter between the two data sets, and the highest correlation of any parameter, being 28% higher than the next highest correlation. This was accompanied by a much smaller median difference between AG2mn and observed runoff than with AG1mn, an acceptance of the hypothesis of equality in Table 4-5, and a better match with observed runoff in Figure 4-5. The large differences between monthly modeled runoff from the two alternative monthly simulation procedures was expected since runoff in the AG2mn data included baseflow as well as the AG1mn runoff. The cumulative plot in Figure 4-9 illustrates the similarity between AG2mn and observed runoff. The statistics were influenced by the two months with the largest runoff: the 8/84 monthly AG2mn runoff which greatly exceeds observed runoff, and the 9/79 monthly AG1mn runoff which greatly underpredicts observed runoff.

The enhanced monthly simulation procedure (AG2mn) has increased aggregate composite period storm runoff with monthly baseflow, as intended, and showed a good correlation with observed monthly flow. This strong correlation was reinforced by the good fits indicated with the GOF measures in Table 4-6 and the close approximation with observed monthly runoff in the cumulative plots in Figure 4-9.

4.2.2 Monthly Period Total Nitrogen (TN)

The correlation for AG2mn TN doubled that of AG1mn, with AG2mn's coefficient being the second highest of all monthly auxiliary procedure-parameter combinations. The Wilcoxon signed rank test shows that monthly TN loads in both modeled data sets were statistically indistinguishable from observed TN loads. The median difference between observed and AG2mn TN was only about half that of AG1mn, due to the addition of the monthly septic system TN loads. On a cumulative basis the AG2mn procedure also reduced the difference between modeled and observed loads by about half compared with the AG1mn procedure, as shown in Figure 4-10. Modeled monthly TN, however, appeared to be generally underestimated by both monthly simulation procedures in the monthly TN plot in Figure 4-6, except at the very low end of the observed load scale. For each of the monthly parameters, the statistics were influenced by a small number of outliers. Two of the 23 monthly TN data points were responsible for 60% of the observed load range. Good fits between observed and modeled data were indicated by all GOF measures for both sets of monthly simulation procedures. All of the GOF measures were slightly better for monthly TN loads using the AG2mn rather than the AG1mn procedure. The difference between monthly TN in the AG2mn and AG1mn data sets was the addition of monthly septic system loads, resulting in the doubling of monthly TN correlation in the AG2mn data set.

4.2.3 Monthly Period Total Phosphorus (TP)

A small increase was noted in Spearman's rank correlation coefficient for modeled monthly TP between the AG1mn and the AG2mn data sets, though the correlation is still quite low. The median differences in Table 4-3 were comparable between the two data sets, and both showed a statistically significant median overprediction compared with observed monthly TP loads. The bi-variate plot for monthly TP in Figure 4-7 actually shows that all monthly modeled loads in both data sets overpredicted observed loads. From the plot in Figure 4-11, each of the monthly simulation procedures cumulatively overpredicted TP loads by a factor of about 2.5. The robust CD* and EF* were identical for monthly TP for both monthly data sets, with a slightly lower MdAE for the AG1mn data set. The main reason for the overall poor correlation of monthly TP is the consistent overprediction of composite period TP under all conditions. The monthly septic system TP loads provide a minor increase to an already overestimated parameter, and will be a relatively minor factor in total TP load until the modeled composite period TP predictions can be brought down to observed levels.

4.2.4 Monthly Period Suspended Sediment (SS)

There are no differences in modeling procedures for SS between AG1mn and AG2mn, so identical loads were expected. The highest Spearman's rank correlation coefficient in both monthly modeled data sets was obtained for the SS load in the AG1mn data set. After baseflow and septic system load additions to runoff, TN and TP using the AG2mn monthly simulation procedure, however, SS correlation became the second lowest. The Wilcoxon test shows a significant overprediction of the median modeled SS load. From the bi-variate plot of SS in Figure 4-8, this overprediction is seen to correspond primarily with lower observed values. Once again this plot shows the large influence of one large load, from 8/84, which is responsible for

about 60% of the observed load range. Cumulative SS loads in Figure 4-12 are overpredicted by about 70%. The CD* and EF* measures both indicate a good fit between modeled and observed monthly SS, but the large value of MdAE indicates at least one large individual difference. Since the CD* and EF* measures look so good but MdAE is quite large, there probably are other differences which counter-balance the large MdAE in the CD* and EF* calculations.

4.2.5 Comparison Between Monthly Period Modeling Procedures

The AG2mn monthly simulation procedure produced monthly runoff and TN closer to observed values than the AG1mn procedure because of the following. The difference between output from the AG1mn and AG2mn procedures was the addition of monthly baseflow to AG1mn runoff and the addition of monthly septic system loads to AG1mn TN. Both runoff and TN were underestimated with the AG1mn procedures, so any additions would decrease the relative difference with observed values, except in a few cases, where the additions produced overestimates larger than the previous underestimate. Since TP was already overpredicted by both monthly simulation procedures, adding additional TP from monthly septic system loads only increased the differences with observed loads. Correlations for both runoff and TN more than doubled. All monthly plots were dominated by one or two large runoff or loading events. The poorest parameter correlations were for TP and SS, the two parameters which cumulatively overpredicted their respective loads by approximately 185% and 70%, as shown in Figures 4-11 and 4-12. Removing the largest outlying data point increased the correlation of AG2mn runoff by about 0.09 and decreased the correlation of all other modeled parameters by about the same amount, but did not change the results of any of the paired difference hypothesis tests.

4.3 Comparing Alternative Auxiliary Procedures

A paired observation difference is the difference between an observed and modeled value for an individual parameter. Alternative auxiliary procedures were compared between parameterization procedures and between monthly simulation procedures by calculating the difference between the absolute values of the individual data set paired differences with observed data. The Wilcoxon signed rank test was used to test the hypothesis of equality between data sets from alternative procedures for each parameter. The results of these tests are reported in Table 4-7 for the auxiliary parameterization procedures, and in Table 4-8 for the monthly simulation procedures. Plots of the parameter differences between alternative data set paired differences for composite period data are shown in Figures 4-13 to 4-16, and for monthly data sets in Figures 4-17 to 4-19.

Table 4-7. Hypothesis Tests Between Alternative Parameterization Procedures: (OWcp-AG0cp) vs. (OWcp-AG1cp)

Parameter	Median	95% Confidence Interval		Result
Runoff	0.0237	0.0005	0.0370	R
Nitrogen	0.0050	-0.0100	0.0150	A
Phosphorus	-0.0055	-0.0115	0.0000	A
Suspended Sed.	0.00	-1.50	2.00	A

Table 4-8. Hypothesis Tests Between Alternative Monthly Simulation Procedures: (OWmn-AG1mn) vs. (OWmn-AG2mn)

Parameter	Median	95% Confidence Interval		Result
Runoff	0.55	0.08	1.83	R
Nitrogen	0.038	0.005	0.113	R
Phosphorus	-0.0038	-0.0070	-0.0015	R
Suspended Sed.	0.000	0.000	0.000	

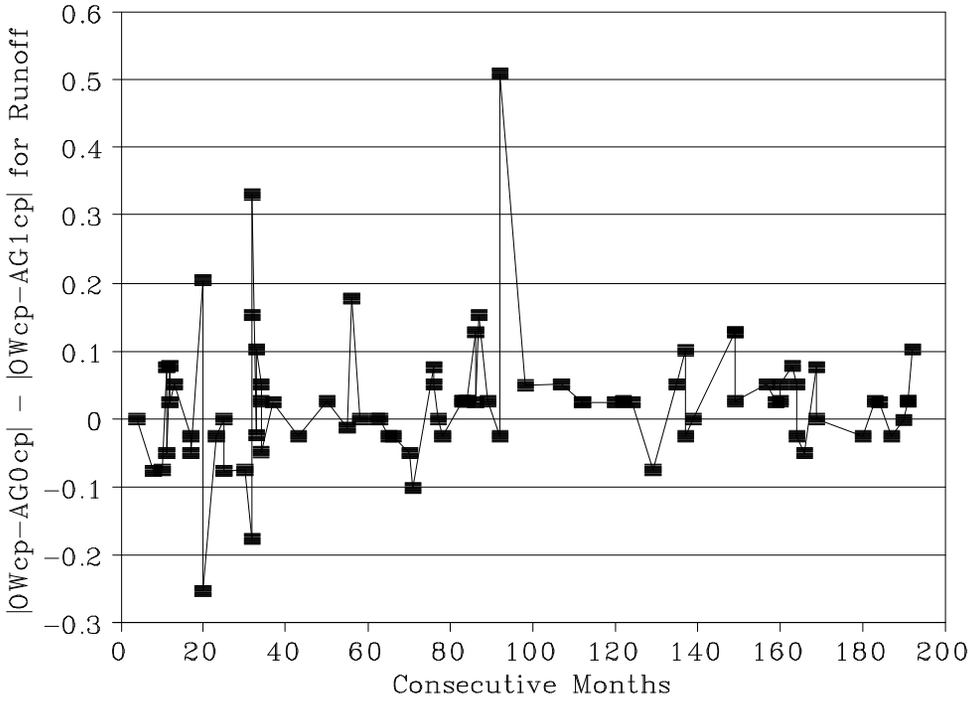

Figure 4-13. Runoff Differences Between Alternative Composite Period Procedures

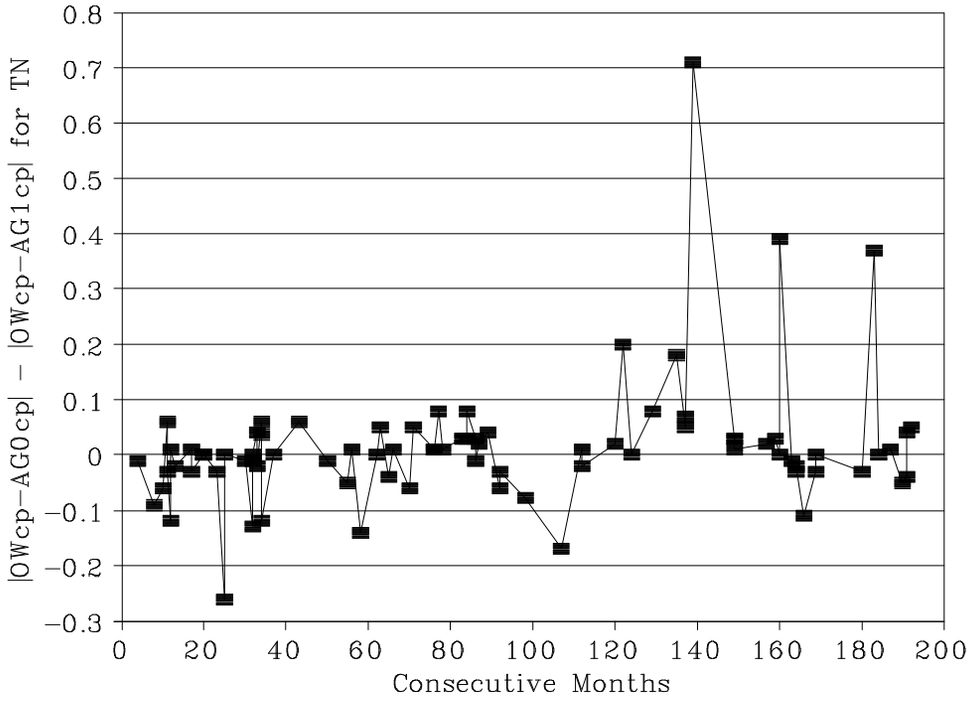

Figure 4-14. TN Differences Between Alternative Composite Period Procedures

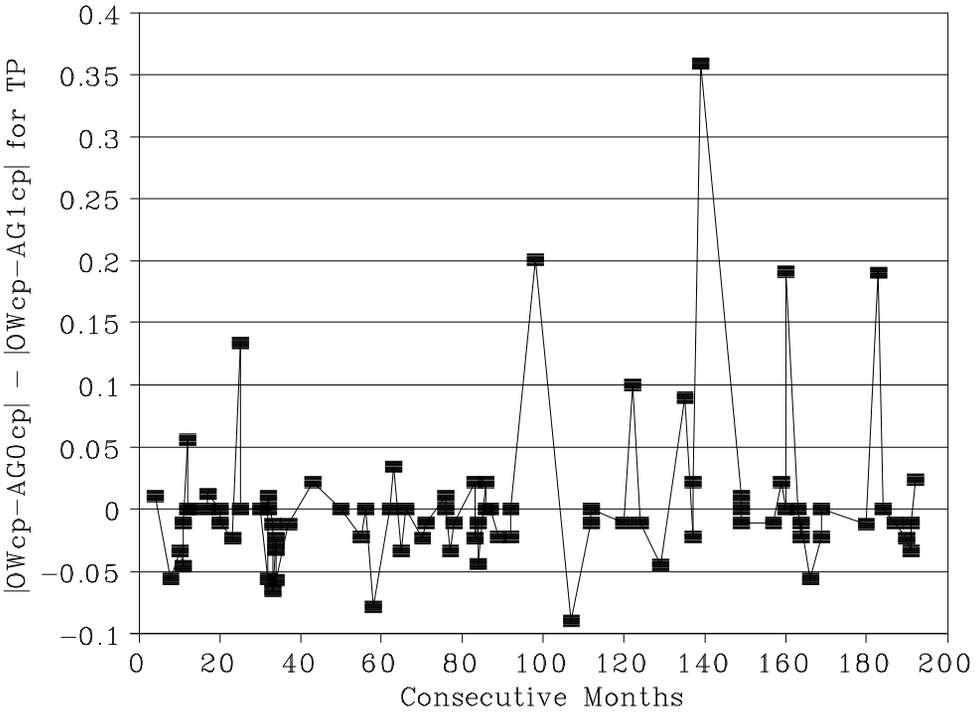

Figure 4-15. TP Differences Between Alternative Composite Period Procedures

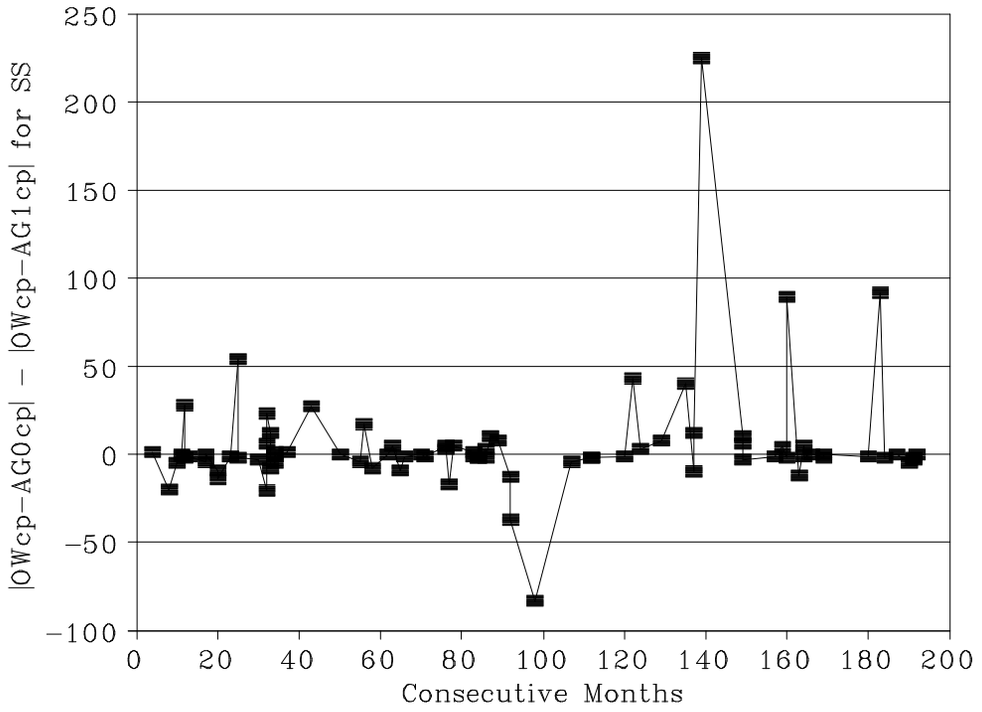

Figure 4-16. SS Differences Between Alternative Composite Period Procedures

4.3.1 Comparing Alternative Parameterization Procedures

The results of the Wilcoxon signed rank test in Table 4-6 indicated a statistical difference between the AG0cp and AG1cp procedures only for the runoff parameter. Since the hypothesis was stated as $|OW - AG0| - |OW - AG1|$, a positive confidence interval (CI) would indicate a larger variation from observed values for AG0cp (therefore a better fit for AG1cp), and a negative CI would indicate a better fit for AG0cp. The positive confidence interval for runoff in Table 4-6, therefore, rejected the null hypothesis for runoff, and AG1cp produced a statistically better fit with observed runoff. Interestingly, the previous test between differences in Table 4-2 showed no statistical difference between the alternative composite period procedures for runoff. Since no statistical difference was shown between output from the AG0cp and AG1cp parameterization procedures for TN, TP and SS, the null hypothesis was accepted for each of these parameters. When looking at the differences between alternative procedures in Figures 4-13 to 4-16, a small number of events exhibit much larger differences than the rest, some positive, some negative, but not always the same events for each parameter. These extreme values are most likely the result of rainfall occurring on those days when a combination of one or more of the distributed time-variable parameter values vary the greatest from the annual average values. Note that for TP the extremely positive events appear to counterbalance the otherwise predominant negative nature of these differences.

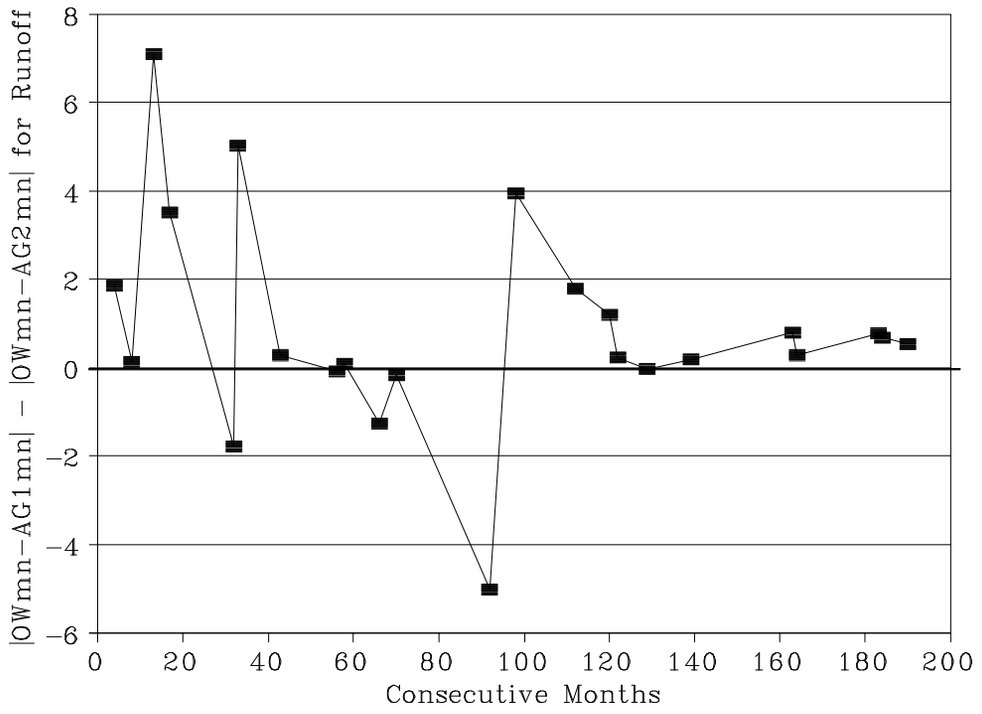

Figure 4-17. Runoff Differences Between Alternative Monthly Period Procedures

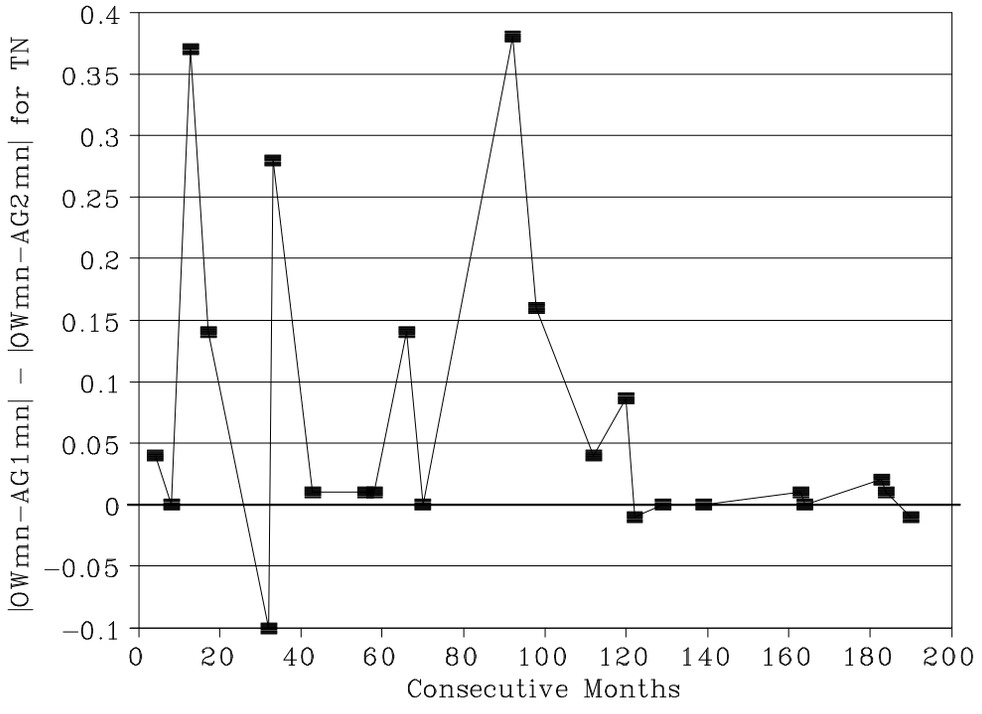

Figure 4-18. TN Differences Between Alternative Monthly Period Procedures

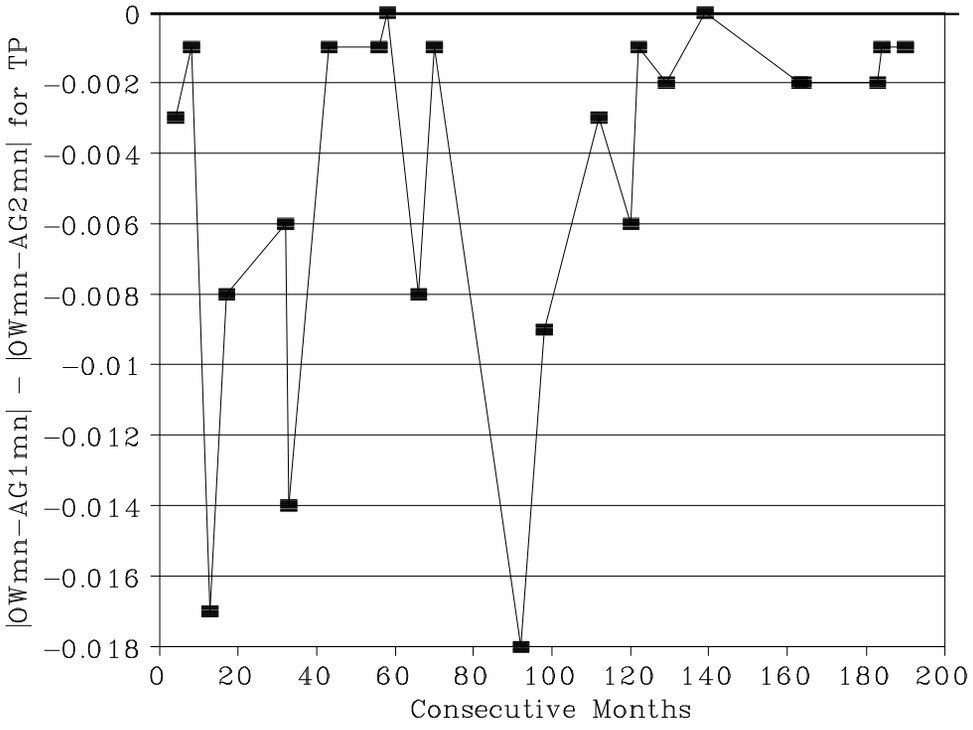

Figure 4-19. TP Differences Between Alternative Monthly Period Procedures

4.3.2 Comparing Alternative Monthly Simulation Procedures

Monthly SS was not tested, as modeling procedures for SS are identical in both sets of monthly modeling procedures. The Wilcoxon signed rank test results in Table 4-7 rejected the null hypothesis for each of the remaining three parameters. By rejecting the null hypothesis, the alternative hypothesis was tentatively accepted indicating a statistical difference between the alternative procedures. As the hypothesis tests the expression, $|OW - AG1| - |OW - AG2|$, positive values for the median difference and confidence interval indicated greater differences between AG1mn and observed values (e.g. a better fit for AG2mn), while negative values for the median difference and confidence interval indicated a better model fit for output from the AG1mn procedure. The Wilcoxon test results in Table 4-7 indicated that the AG1mn procedure produced a better fit for TP, while the AG2mn enhanced aggregated procedure produced a statistically better model for runoff, and TN. Figures 4-17 to 4-19 show sequential parameter differences between the alternative monthly simulation procedures. Monthly AG2mn runoff best approximated observed runoff in all but four months, with large differences in two of those months. TN differences in Figure 4-19 were generally small between the alternative procedures in all months, and in only three months were the AG2mn differences greater than the AG1mn differences. Monthly TP differences between simulation procedures were all negative, indicating that the AG1mn procedure produced a better fit with monitored data, since the addition of septic system loads only increased the already overestimated TP loads for all months.

4.4 Rainfall - Runoff Comparison

Earlier attempts were made in this study to provide the best match possible between monitored data and rainfall from one of the neighboring rain gauges, as none was located in the watershed. During the analysis of the modeling procedure evaluation, Figure 4-20 was produced to compare monitored and modeled runoff with daily rainfall. This plot shows that the current matching procedure was still inadequate. If the storms modeled were the same ones which produced the monitored runoff, over a third of the storms would have greater than 50% of the rainfall converted to runoff, and six of the storms produced more runoff than rainfall! The OWML reported long-term average runoff at the Bull Run ST60 monitoring site was around 38% of annual precipitation. In Figure 4-20, all of the AG1cp modeled runoff was in the range of 0-55% of the modeled daily rainfall. This range is considerably less than the monitored runoff range and is in line with the long-term average for this station. Because of the unusually high percentages of runoff corresponding to the matched rainfall, it appears that the daily rainfall events used for modeling were not the same events which produced the monitored data.

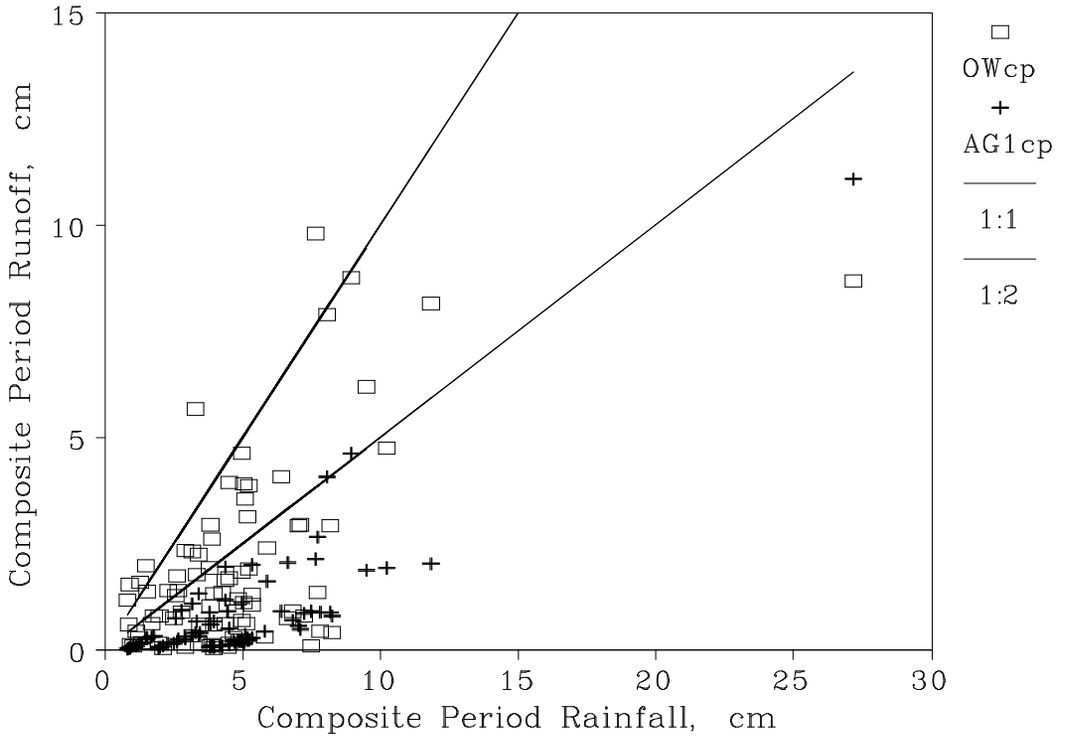

Figure 4-20. The Plains Rainfall vs. Bull Run Runoff