An Empirical, General Population Assessment of the Properties of Variance Estimators of the Horwitz-Thompson Estimator Under Random-Order, Variable Probability Systematic Sampling

Stephen V. Stehman
Biometrics Unit, Cornell University, Ithaca, NY 14850 USA

W. Scott Overton
Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

BU-936-M

Revised August 1989

An Empirical, General Population Assessment of the Properties of Variance Estimators of the Horvitz-Thompson Estimator Under Random-Order, Variable Probability Systematic Sampling

Stephen V. Stehman and W. Scott Overton
Biometrics Unit, Cornell University and Department
of Statistics, Oregon State University

Technical Report 132

August 1989

DEPARTMENT OF STATISTICS
Oregon State University
Corvallis, Oregon

ABSTRACT

Previous empirical studies of the properties of variable probability, systematic sampling and the Horvitz-Thompson estimator, $\hat{T}_y$, have focused on specific, real-world populations (cf. Stehman and Overton, 1987; Cumberland and Royall, 1981; Rao and Singh, 1973). The study of special case populations is recognized as important, but these studies provide limited information that can be used to generalize to other populations. By a systematic simulation study of a specially designed set of populations, we have extended the assessment of the properties of $V(\hat{T}_y)$ and estimators of $V(\hat{T}_y)$ is extended to more general populations represented by the *population space*.

The *population space* is a standardized representation of bivariate populations with equal variance in each marginal distribution. Each representation of this space has a specific bivariate distribution shifted over the dimensions of the space. Three bivariate distribution forms, each with three correlations, were studied.

Two common estimators of $V(\hat{T}_y)$ were investigated, one proposed by Horvitz and Thompson (1952), the other by Yates and Grundy (1953) and Sen (1953). The sampling design studied was random-order, variable probability systematic. Both variance estimators require computing the pairwise inclusion probabilities. Calculating these pairwise probabilities requires immense computing time, so two

approximation formulas were studied, one due to Hartley and Rao (1962), the other due to Overton (1985). The two variance estimators were computed using each of the pairwise inclusion probability approximations.

Behavior of the estimators was represented by contour plots describing confidence interval coverage, root mean square error, relative bias, and proportion of negative estimates for the variance estimators over the range of populations in the population space. These plots provide the basis of a descriptive theory for the properties of the variance estimators. The population space approach successfully serves as a bridge between strictly special case, empirical results and a general analytical theory. The approach also furnishes a perspective in which more theoretical results can be pursued.

1. INTRODUCTION

Properties of the variance and estimators of the variance of the Horvitz-Thompson estimator are investigated for variable probability, systematic sampling (hereafter denoted *vps* sampling). In a finite population of size N, assume that a response variable of interest, y_i, and an auxiliary variable, $x_i>0$, are defined for each element of the universe. In variable probability sampling, a sample unit is selected with probability proportional to x. We will restrict attention to without replacement, fixed sample size schemes. The probability that the i^{th} population element will be selected in the sample is given by the inclusion probability $\pi_i = \sum_{\{s:i\epsilon s\}} p_R(s)$, where $p_R(s)$ is the probability of selecting sample s under sampling rule R. The probability of selecting both the i^{th} and j^{th} population units is the pairwise inclusion probability, $\pi_{ij} = \sum_{\{s:(i,j)\epsilon s\}} p_R(s)$.

Horvitz and Thompson (1952) presented a general, finite population theory of estimation for variable probability sampling. The Horvitz-Thompson estimator, $\hat{T}_y = \sum_{u\epsilon s} y_u/\pi_u$, is unbiased for the population total, $T_y = \sum_{i=1}^{N} y_i$. Two well-known estimators of the variance of $\hat{T}_y$ are the estimator proposed by Horvitz and Thompson,

$$v_{HT} = \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i}\right)^2 (1-\pi_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_i y_j}{\pi_i\pi_j}$$

and the estimator suggested by Yates and Grundy (1953) and Sen (1953),

$$v_{YG} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

where the summations for the two variance estimators are over the pairs of elements in the sample. Based on limited investigation, the estimator v_{YG} is usually claimed superior to v_{HT} because v_{YG} has smaller sampling variance and is less likely to take on negative values (cf. Cassel *et al* (1977, p. 166), Cochran (1977, p. 261)). The evidence for the superiority of v_{YG} is sketchy. It is known that when the ratio $r_u = y_u/x_u$ is constant for all u=1,...,N, $V(\hat{T}_y) \equiv 0$. In this situation, $v_{YG} \equiv 0$, but v_{HT} does not identically equal 0; being unbiased, v_{HT} must be capable of negative values. Thus for populations in which y is nearly proportional to x, v_{YG} would appear to have smaller sampling variance.

Several empirical studies have shown advantages for v_{YG}. Rao and Singh (1973) studied 34 natural populations using Brewer's probability proportional to x selection procedure, and Cumberland and Royall (1981) examined random-order, variable probability systematic selection for 6 populations. Both studies found v_{HT} frequently resulted in negative estimates, and that the sampling variance of v_{HT} was much larger for many of their populations. Stehman and Overton (1987a) presented some simulation results showing that the advantages of v_{YG} were restricted to certain kinds of populations.

Clearly lacking in these comparisons of the two

variance estimators, v_{HT} and v_{YG}, is a general theory describing their properties. A theoretical analysis of confidence interval coverage, mean square error (MSE), bias and proportion of negative estimates of the variance estimators has not been done. A major complication in development of an analytic theory is that the pairwise inclusion probabilities depend on the particular finite population x's. Writing specifically about variable probability, systematic sampling, Brewer (1963) concluded "... the selection probabilities for the various possible samples are functions of the sizes of all the population units and it is virtually impossible to construct an exact general theory." A thorough empirical investigation of the properties of these variance estimators is an intermediate step toward development of a more general theory.

The population space assessment described in Section 3 was an extensive simulation study of a specially designed set of populations. The sampling design investigated was random-order, *vps* with n=16. Approximation formulas for the pairwise inclusion probabilities are commonly used in practice. Also, since all populations in the simulation study were relatively large (N>70), computing the exact pairwise inclusion probabilities was not practical. Two approximation formulas for the pairwise inclusion probabilities were investigated,

$$\pi^{o}_{ij} = \frac{(n-1)\pi_i\pi_j}{n-\frac{1}{2}(\pi_i+\pi_j)} \quad \text{(Overton, 1985)},$$

and

$$\pi_{ij}^{hr} = \frac{(n-1)\pi_i\pi_j}{\left[n-\pi_i-\pi_j+\sum_{k=1}^{N}\pi_k^2/n\right]}$$ (Hartley and Rao, 1962).

Further description of the pairwise inclusion probability approximations and formulas for the estimators can be obtained from Stehman and Overton (1987b, 1989).

Properties of the variance estimators were obtained by simulation, using 5,000 replications of the sampling procedure for each investigated population. The number of replications was selected to provide a precise estimate of the coverage probabilities obtained by the variance estimators for constructing nominal 95% confidence intervals for T_y. With 5,000 replications, the standard deviation of the estimated proportion of confidence intervals covering the parameter is 0.003. Version 1.49 of the GAUSS Mathematical and Statistical System (Aptech Systems, Inc., Kent, WA) was used to run the simulations on IBM XT or AT computers.

Several procedures were used to validate the simulation programs and computing algorithms. Since $\hat{T}_y$ is unbiased for T_y, the estimated expected value of $\hat{T}_y$ was checked to make sure it was close to T_y. The computing formulas for the variance estimators were validated by setting $x_i=1 \ \forall \ i=1,\ldots,N$, and verifying that the estimates matched those known for a simple random sample. If a computing formula or algorithm was changed during the course of the population space analysis, output from the

modified program was checked to ensure that the new algorithm gave equivalent results to previously verified algorithms. Finally, the results obtained from the simulation programs matched those reported by Rao and Singh (1973) and Cumberland and Royall (1981) for their simulations using the same populations.

2. GENERATION OF PSEUDO-RANDOM NUMBERS

Random numbers from the Uniform(0,1) and standard normal distributions were generated using the GAUSS functions RNDU and RNDN, respectively. Gamma random variables, G, were selected from a standard gamma distribution (gamma distribution with parameter α, and parameter λ set to 1). Only the standard gamma distribution was considered because any other gamma distribution can be obtained by scaling the standard gamma (cf. Kennedy and Gentle, 1980). For integer-valued α, gamma random variables were generated by the following algorithm:

1) Generate U_i from U(0,1).
2) $X_i=-\ln(U_i)$; X_i is an exponential random variable with parameter $\theta=1$.
3) $G=\sum_{i=1}^{\alpha} X_i$; the gamma random variable is the sum of k independent, identically distributed exponential random variables with $\theta=1$.

Random variables for standard gamma distributions with non-integer parameter, α, $0<\alpha<1$, were generated by algorithm

GS provided by Kennedy and Gentle (p. 213, 1980). Since a sum of n independent standard gamma variables, each with parameter α_i, is distributed as standard gamma with parameter $\sum_{i=1}^{n} \alpha_i$, gamma random variables from distributions with parameter $\alpha>1$ were generated by summing independent gamma random variables generated by algorithm GS with proper choice of α_i and n.

3. DESCRIPTION OF THE POPULATION SPACE

Three major families of populations were studied, one based on real data (STREAM), and two generated from known probability distributions (GAMNORM and BIGAMMA). Within each family, three different subfamilies representing low, medium, and high correlations between the response variable, y, and the design covariate, x, were studied. A subfamily was then a set of populations with the same correlation, within the same major family.

All populations within a subfamily were created from a single base population. A subfamily was created from the base population by adding or subtracting constants to x and/or y. Thus all populations in a subfamily are the same "cloud" of points shifted to various locations in the (x,y)-plane. All populations within a subfamily have $V_y=V_x$, where V_x and V_y are the population variances of x and y, respectively, and populations within a subfamily also have the same V_y and the same correlation between x

and y. Populations differing by an additive shift in the x's have different inclusion probabilities. Additive shifts in the x's are a feasible design tactic in some real surveys, and this feature of sample design can be explored conveniently in the context of the population space.

The variables x and y were standardized, $X' = x/\sqrt{V_x}$ and $Y' = y/\sqrt{V_y}$, so comparisons would be invariant to the measurement scale of the variables. The standardized population centroid, $(\bar{X}', \bar{Y}')$, was used to locate populations within the population space. Note that $\bar{X}' = \bar{X}/\sqrt{V_x} = 1/cv(x)$, and $\bar{Y}' = \bar{Y}/\sqrt{V_y} = 1/cv(y)$, where cv denotes the population coefficient of variation.

Confidence interval coverage, ratios of root mean square error (RMSE) of the variance estimators, and relative bias are invariant to the measurement scale of the y's, and are, therefore, the same in the original and the standardized populations. The standardized population space is also appropriate for assessment of patterns of precision of $\hat{T}_y$. The standardized variance can be obtained easily from $V(\hat{T}_y)$, the variance of the Horvitz-Thompson estimator for the unstandardized variable y. For the standardized variable $Y' = y/\sqrt{V_y}$,

$$V(\hat{T}_{y'}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} Y_i'^2 + \sum_{j \neq i}^{N}\sum^{N} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right) Y_i' Y_j'$$

$$= \frac{1}{V_y} \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + \sum_{j \neq i}^{N}\sum^{N} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right) y_i y_j = V(\hat{T}_y)/V_y .$$

A convenient representation of standardized variance is

obtained by replacing V_y by a quantity which is proportional to V_y, the variance of $\hat{T}_y$ for a simple random sample.

4. FAMILIES OF POPULATIONS

4.1 STREAM Family

The first family of populations analyzed was constructed from a subset of the Phase I Stream Survey Pilot Study data (Messer *et al*, 1986). A previous study of this STREAM family was reported in Stehman and Overton (1987a). Some of the results shown here supercede that work.

Seventy-two of the 100 units from the Pilot Study sample were purposefully selected yielding a base population with correlation 0.82 between the response variable, y=length of stream reach, and the auxiliary variable, x=direct watershed area of a stream reach. To create the base populations for the two other STREAM subfamilies, starting with the STREAM82 base population,

1) compute the least squares slope, $\hat{\beta}$, and intercept, $\hat{\alpha}$, of the base population for subfamily STREAM82;
2) compute $e_i = y_i - \hat{y}_i$, the residual from the least squares fit, where $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$;
3) let $e_i^* = k * e_i$ (multiply the residuals by a constant to obtain the specified correlation);
4) set $y_i^* = \hat{y}_i - e_i^*$.

The values y_i^* were used as the response variable for the base population of the new subfamily. Choosing k=2.5 resulted in a subfamily with $\rho=0.50$ (subfamily STREAM50), and choosing k=0.25 resulted in a subfamily with $\rho=0.985$ (subfamily STREAM99). An advantage of this method of generating the base population was that the x's were the same for each base population in the family. Thus the simulations were "blocked" in that each subfamily within the STREAM family had the same first and second order inclusion probabilities for all populations with common $\bar{X}'$.

The STREAM family was created from data with unknown population distributional properties, thus limiting our ability to generalize to other populations. The empirical study was expanded to include families generated from known probability distributions to permit broader understanding. The next two families examined were constructed to represent distributions of random variables similar to those likely to be encountered in practice.

4.2 GAMNORM Family

For the GAMNORM family of populations, x was randomly generated from a standard gamma distribution with parameters $\alpha=2$ and $\lambda=1$, and y was generated, conditional on x, as a normal random variable. For each x_i, y_i was obtained from the equation, $y_i=\beta x_i+\epsilon_i$, where ϵ_i was a random variable distributed Normal$(0, \sigma_\epsilon^2)$, and $\sigma_\epsilon^2=(1-\rho^2)\sigma_x^2$. The

infinite population notation is appropriate because σ^2 denotes the population variance of the distribution of the generated random variable. Once $\rho(x, y)$ was specified, the value of σ_ϵ^2 was fixed by the following argument. If the relationship between y and x is given by

$$y_i = \beta x_i + \epsilon_i, \tag{1}$$

then $\sigma_y^2 = \beta^2\sigma_x^2 + \sigma_\epsilon^2$ (x_i independent of ϵ_i), and

$$\sigma_y^2/\sigma_x^2 = \beta^2 + \sigma_\epsilon^2/\sigma_x^2. \tag{2}$$

Imposing the constraint that $\sigma_y^2 = \sigma_x^2$, $\beta = \sigma_{xy}/\sigma_x^2 = \rho\sigma_y/\sigma_x = \rho$, and using equation (2),

$$\beta^2 = 1 - \sigma_\epsilon^2/\sigma_x^2. \tag{3}$$

Solving (3) for σ_ϵ^2 yields $\sigma_\epsilon^2 = (1-\rho^2)\sigma_x^2$. (4)

In practice, a set of x's was generated and V_x calculated. Then V_x from the particular set of x's was used instead of σ_x^2 in (4) for generation of the y's. A subfamily base population was created by specifying ρ, generating the set of ϵ's, and forming the variable y_i from (1). The target correlations were 0.5, 0.8, and 0.95, but due to the random data generation, the realized correlations were 0.48, 0.75, and 0.94.

The same set of 100 x's was used as the base population for all three subfamilies. Using a single set of x's again provided "blocking" on the inclusion probabilities of the *vps* sampling design. The GAMNORM subfamily base populations could have been created using steps similar to those described in Section 4.1 to create

the STREAM subfamily base populations. This change would have provided an additional level of blocking among the GAMNORM subfamilies.

4.3 BIGAMMA Family

The third family studied consisted of populations selected from a bivariate gamma distribution. Johnson and Kotz (pp. 216-218, 1970) provide the following basic theory for generation of bivariate gamma random variables. As before, the standard gamma, with $\lambda=1$, is used throughout.

> If W_0, W_1, and W_2 are independent random variables, with W_j distributed Gamma(θ_j), and if $X=W_0+W_1$ and $Y=W_0+W_2$, then X is distributed Gamma($\theta_0+\theta_1$), Y is distributed Gamma($\theta_0+\theta_2$), (X,Y) is distributed Bivariate Gamma, and $\rho(X,Y)=\theta_0\left[(\theta_0+\theta_1)(\theta_0+\theta_2)\right]^{-1/2}$.

Generating bivariate random variables based on this result permitted specifying $\rho(X,Y)$ and provided marginal distributions of X and Y that were both gamma distributions. The parameter for both standard gamma marginal distributions was $\alpha=2$. For this parameter specification, $\theta_1=\theta_2=2-\theta_0$, and $\sigma_x^2=\sigma_y^2$. Setting $\theta_1=\theta_2=\theta$, we obtain $\rho(X,Y)=\theta_0/(\theta+\theta_0)$. Then for a specified ρ, $\theta_0=\rho\theta/(1-\rho)$. Finally, solving for θ and θ_0 yields $\theta=2-\theta_0$, and $\theta_0=2\rho$. θ and θ_0 are the parameters needed to generate the bivariate gamma random variables.

The algorithm used to generate a bivariate gamma base population with specified $\rho(x, y)$ and marginal gamma distributions each with parameter $\alpha=2$ was:

1) generate W_0 distributed Gamma(2ρ);

2) generate W_1 distributed Gamma$(2(1-\rho))$;

3) generate W_2 distributed Gamma$(2(1-\rho))$;

4) calculate $X=W_0+W_1$, and $Y=W_0+W_2$.

If a population with large x values was generated so that at least one of the sampling units would be selected with probability 1 in a sample of size 16, that population was discarded and a new base population was generated. The three ρ's specified were 0.5, 0.75, and 0.95, and the actual realized correlations were 0.49, 0.77, and 0.97.

For the BIGAMMA family, a different set of x's was generated for each of the subfamily base populations. To obtain both marginal gamma distributions with parameter $\alpha=2$, this bivariate random variable generation algorithm required generating a new population of x's for each base population. To obtain blocking on the x's, another algorithm for generating the bivariate random variables would be necessary.

5. RESULTS OF POPULATION SPACE ANALYSIS

Notation:

$\bar{X}'$, $\bar{Y}'$ population standardized means of x and y

$\hat{T}_y$ Horvitz-Thompson estimator of the population total

$V(\hat{T}_y)$ variance of the Horvitz-Thompson estimator

Approximation Formulas

π_{ij}^{hr} approximate formula for π_{ij} (Hartley and Rao, 1962)

π_{ij}^{o} approximate formula for π_{ij} (Overton, 1985)

Variance Estimators

v_{HT} Horvitz-Thompson variance estimator

v_{YG} Yates-Grundy variance estimator

v_{HT}^{hr} Horvitz-Thompson variance estimator calculated using π_{ij}^{hr}

v_{HT}^{o} Horvitz-Thompson variance estimator calculated using π_{ij}^{o}

v_{YG}^{hr} Yates-Grundy variance estimator calculated using π_{ij}^{hr}

v_{YG}^{o} Yates-Grundy variance estimator calculated using π_{ij}^{o}

Scatter plots of the population located at $(\bar{X}', \bar{Y}') = (7, 7)$ for each subfamily are shown in Figure 1. Population quantile-quantile plots for the middle correlation subfamilies at $(\bar{X}', \bar{Y}') = (7, 7)$ are shown for the variable x in Figure 2 and for the variable y in Figure 3.

5.1 Comparison of Variance Estimators

The criteria for comparison of the variance estimators v_{HT}^{o}, v_{YG}^{o}, v_{HT}^{hr}, and v_{YG}^{hr} were:

1) confidence interval coverage obtained by nominal 95% intervals calculated as $\hat{T}_y \pm 1.96\sqrt{\hat{v}}$;

2) estimated MSE;

3) relative bias, estimated by

$$\text{rel bias} = \left[\hat{E}(\hat{v}) - \hat{V}(\hat{T}_y)\right] / \hat{V}(\hat{T}_y),$$

where $\hat{E}(\hat{v})$ was the simulated expected value of $\hat{v}$, and $\hat{V}(\hat{T}_y)$ was an unbiased estimator of $V(\hat{T}_y)$ obtained from the simulations;

4) probability of a sample resulting in negative v_{HT}^{hr}.

The behavior surfaces were described by a battery of contour plots generated by the interpolation and contour plotting routines supplied by the SURFER software package (Golden Software, Inc., P. O. Box 281, Golden, Colorado). The kriging option in SURFER was selected to create a regularly spaced grid from the irregularly spaced input data. The octant search option in SURFER, using the 10 nearest data points, was used for interpolating grid points.

5.2 Interpretation of Contour Plots

To guide the reader's interpretation of the contour plots, certain important features of the plots will be highlighted. The standard diagonal serves as a convenient spatial reference. Although many details of specific plots are discussed in Sections 5.2.1 through 5.2.5, focus on the overall patterns should be maintained. Figures 4 through 17 are organized such that each column on a page represents a family, and the rows represent subfamilies, arranged in the column by increasing correlation. Enlarged plots of the lower left corner are included to show the detail in that portion of the population space. All plots are located at the end of Section 5.

5.2.1 Standardized variance

The standardized variance compared the variance of $\hat{T}_y$ under random-order, *vps* sampling relative to the variance of $\hat{T}_y$ under simple random sampling (Figure 4). The qualitative pattern of standardized variance was similar for all three families. The standardized variance surface was highest along the left edge of the population space, then sloped downward moving diagonally from the upper left to the lower right corner. A trough of minimum standardized variance was located near the standard diagonal for the medium and high correlation subfamilies, but was clearly below the standard diagonal for the low correlation subfamilies. The surface sloped gradually upward out of this trough when moving toward the lower righthand corner.

The region in which variable probability sampling was more efficient than simple random sampling was larger for high and medium correlation subfamilies compared to the low correlation subfamilies. The contour showing equal precision for *vps* sampling and simple random sampling was almost directly over the standard diagonal for the three low correlation subfamilies, and the advantage of variable probability sampling increased with $\rho(x, y)$. In the upper left region of the population space, variable probability sampling was much less efficient than simple random sampling.

5.2.2 Confidence interval coverage

The main results observed from the contour plots (Figures 5-8) of observed confidence interval coverage (nominal 95% intervals) obtained from each variance estimator were:

1) v_{YG}^{o} and v_{YG}^{hr} provided similar coverage over the entire population space;
2) v_{HT}^{hr} provided the poorest coverage of the 4 estimators studied;
3) v_{HT}^{o} provided close to the nominal 95% coverage for most of the population space, but the pattern of v_{HT}^{o} coverage differed from the pattern shown by v_{YG}^{o} and v_{YG}^{hr};
4) coverage was poorest along the extreme left edge of the population space for all variance estimators except v_{HT}^{hr};
5) the relief of the surfaces increased with subfamily correlation;
6) qualitative patterns in coverage were similar for all three families.

For most of the population space, coverage provided by v_{HT}^{o} was near the nominal 95% (Figure 5). Regions of low coverage occurred along the extreme left edge of the population space, and in the high correlation subfamilies with small $\bar{Y}'$. A wide plateau of high coverage extended from the upper right corner toward the lower left corner, narrowing toward the origin roughly parallel along the

standard diagonal. The coverage surface sloped steeply downward off the left edge of the high plateau, the contours nearly parallel to the vertical axis. Another sharp decline in the surface occurred along the standard diagonal in the region near the origin. The downward slope of coverage off the high plateau was much gentler toward the lower right region of the population space. The gradients of the coverage surfaces were steeper as the subfamily correlation increased. Regions of coverage provided by v_{HT}^{o} that were higher or much lower than the 95% nominal level were associated with regions of large positive and large negative relative bias of v_{HT}^{o}.

Coverage provided by v_{HT}^{hr} was generally much worse than coverage provided by v_{HT}^{o} (Figure 7). v_{HT}^{hr} had very poor coverage in the region surrounding the standard diagonal. This region of poor coverage of v_{HT}^{hr} was associated with a region of high probability of negative estimates (see Figure 17). Coverage levels for v_{HT}^{hr} were unacceptably low for the high correlation subfamilies.

The coverage provided by v_{YG}^{o} and v_{YG}^{hr} (Figures 6 and 8) was 93 or 94% for most of the population space. v_{YG}^{o} and v_{YG}^{hr} had lower coverage than v_{HT}^{o} along the extreme left edge of the population space, but both Yates-Grundy based estimators improved on the coverage of v_{HT}^{o} in the high correlation subfamilies in the region near the horizontal axis.

5.2.3 Ratios of RMSE

Comparison of the variance estimators on the criterion of RMSE (Figures 9-12) was based on selected ratios of RMSE's. The main features of the RMSE comparisons were:

1) v_{YG}^{hr} had smaller RMSE than v_{HT}^{hr} for most of the population space, but RMSE of v_{HT}^{hr} was less than or equal to RMSE of v_{YG}^{hr} in some regions of the population space;
2) v_{YG}^{o} was almost always smaller in RMSE relative to v_{HT}^{o};
3) v_{HT}^{o} was far superior to v_{HT}^{hr} along the standard diagonal, and never much poorer than v_{HT}^{hr} in any region;
4) v_{YG}^{o} and v_{YG}^{hr} had very similar RMSE, with v_{YG}^{o} having slightly smaller RMSE in populations located near the origin;
5) ratios of RMSE's showed greater variation over the population space in the high correlation subfamilies compared to the low correlation subfamilies.

The surface of the ratio $RMSE(v_{HT}^{hr})/RMSE(v_{YG}^{hr})$ was roughly symmetrical about the standard diagonal. A region of high ratios extended along the standard diagonal sloping downward from the upper right to the lower left. The gradient of this slope increased with the subfamily correlation, and was particularly steep in STREAM99.

Throughout much of the space pictured in Figure 9, v_{YG}^{hr} had smaller RMSE than v_{HT}^{hr}. Near the origin, v_{HT}^{hr} had RMSE less than or equal to v_{YG}^{hr} in the low correlation subfamilies, so v_{YG}^{hr} was not uniformly superior to v_{HT}^{hr} on the basis of the RMSE criterion. However, v_{YG}^{hr} was never much worse than v_{HT}^{hr} in terms of RMSE, while v_{HT}^{hr} could be extremely poor relative to v_{YG}^{hr}. Since relative bias was nearly zero for both v_{HT}^{hr} and v_{YG}^{hr}, these RMSE comparisons were essentially equivalent to variance comparisons.

RMSE of v_{HT}^{o} was less than the RMSE of v_{HT}^{hr} in the region surrounding the standard diagonal (Figure 10). A prominent feature of the surface of the ratio of RMSE of v_{HT}^{o} to RMSE of v_{HT}^{hr} was a deep, V-shaped trough along the standard diagonal sloping downward and widening toward the upper right corner. This trough was deepest in the high correlation subfamilies as the RMSE advantage of v_{HT}^{o} relative to v_{HT}^{hr} increased with the subfamily correlation. RMSE of v_{HT}^{hr} was smaller than RMSE of v_{HT}^{o} along the left edge of the population space, and in the region near the origin of the low and medium correlation subfamilies. Regions of superiority of v_{HT}^{o} relative to v_{HT}^{hr} corresponded to the regions in Figure 9 where v_{YG}^{hr} was far superior to v_{HT}^{hr}. Thus the approximation π_{ij}^{o} improved the RMSE of the Horvitz-Thompson variance estimator in those regions of the population space where v_{HT}^{hr} had high RMSE.

RMSE of v_{YG}^{o} was smaller than RMSE of v_{HT}^{o} for most of

the population space (Figure 11). The surface of the ratio of RMSE's of v^{o}_{YG} to v^{o}_{HT} decreased gradually from the upper left corner to the lower right corner of the population space, the contours of the surface running roughly parallel to the standard diagonal. The detailed plots of the region near the origin indicated a U-shaped ridge sloping gradually downward toward the origin along the standard diagonal. Although the gradients in the surfaces increased with correlation, the surfaces were less steep than those observed in Figures 9 and 10.

v^{hr}_{YG} and v^{o}_{YG} had almost identical RMSE throughout the population space (Figure 12). v^{o}_{YG} had slightly smaller RMSE in populations located near the origin. This pattern was consistent for all three subfamilies.

5.3.4 Relative Bias

Of the four variance estimators investigated, only v^{o}_{HT} displayed a significant relative bias (Figure 13). The other three variance estimators were nearly unbiased (Figures 14-16), with the exception that relative bias of v^{o}_{YG} was above -0.10 along the extreme left edge of the population space near the origin for two of the STREAM subfamilies. The pattern of relative bias of v^{o}_{HT} was similar in all families. Relative bias of v^{o}_{HT} decreased from a high positive value in the upper left region to a high negative value in the lower right region, and the

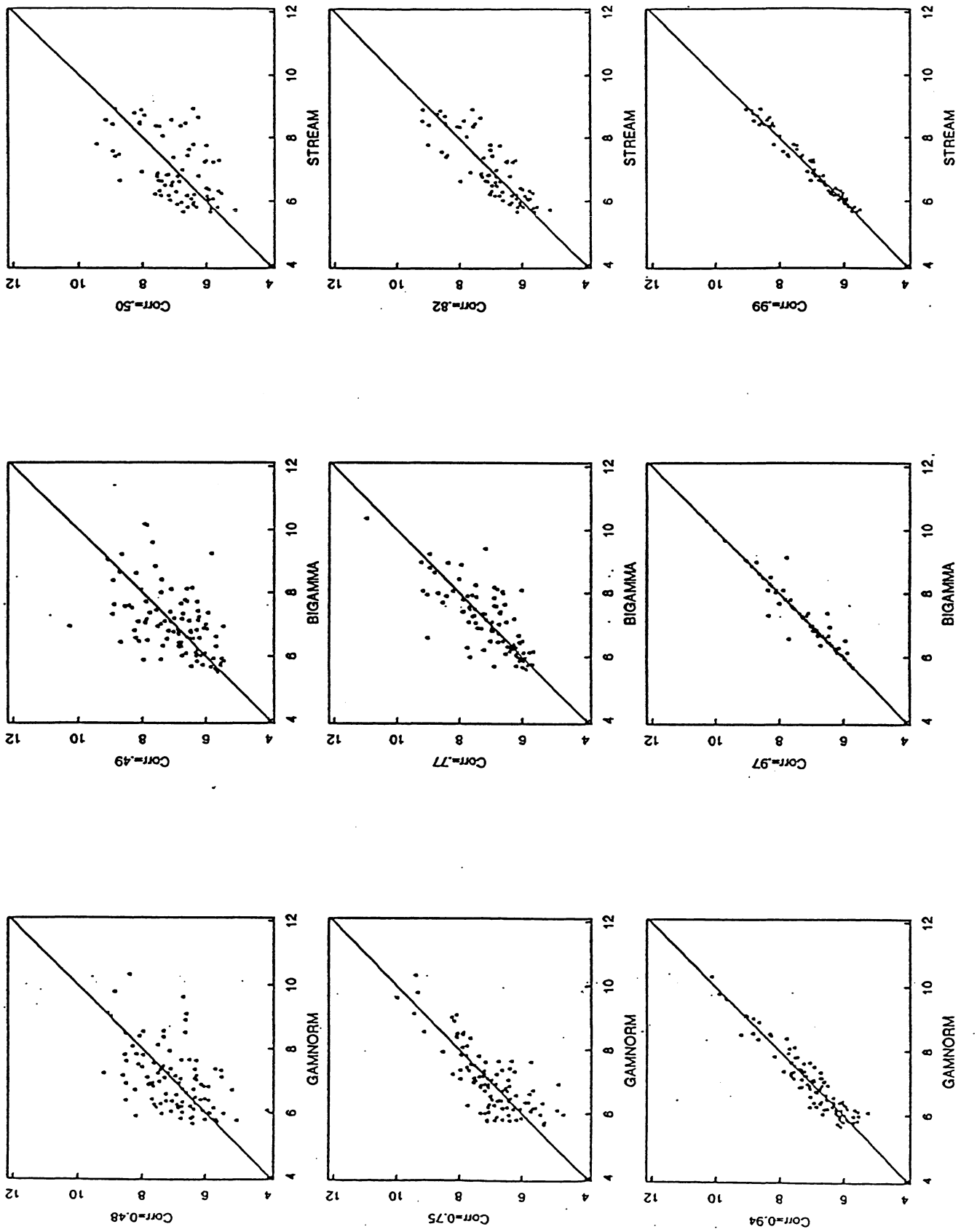
magnitude of the bias was largest in the high correlation subfamilies. v^{o}_{HT} was unbiased in a band of the population space located just below and roughly parallel to the standard diagonal. The strong pattern in the bias of v^{o}_{HT} suggests that an adjustment of the estimator to make it unbiased may be available.

5.2.5 Probability of Negative v^{hr}_{HT} Estimates

For all samples obtained in the simulation study, v^{o}_{YG} and v^{hr}_{YG} were non-negative, and negative v^{o}_{HT} estimates were extremely rare. Only v^{hr}_{HT} was subject to frequent negative estimates (Figure 17). Negative v^{hr}_{HT} estimates were rare for the low correlation subfamilies, but increased in frequency as the subfamily correlation increased. Negative estimates were infrequent in all subfamilies along the left edge of the population space and along the horizontal axis. The probability of a negative estimate was highest along the standard diagonal, and increased along this diagonal from the lower left to the upper right.

LIST OF FIGURES

Figure 1. Scatter Plots of Subfamily Populations at Standardized Centroid (7,7).


STREAM
BIGAMMA
GAMNORM
Corr=.99
Corr=.82
Corr=.50
Corr=.97
Corr=.77
Corr=.49
Corr=0.94
Corr=0.75
Corr=0.48
4
6
8
10
12

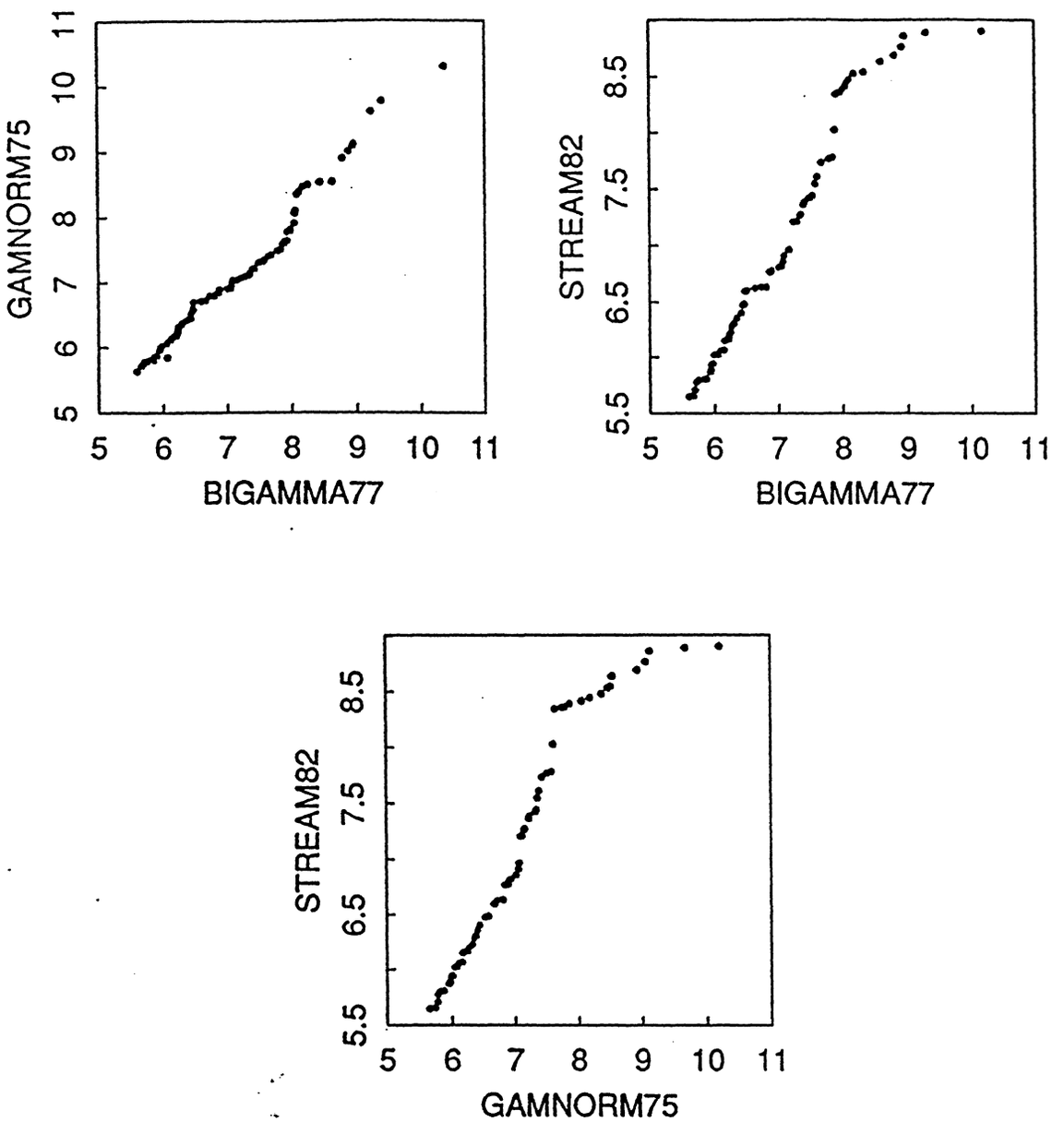


Figure 2. Population Quantile-Quantile Plots of the Variable x for the 3 Families.
(Middle correlation subfamilies at $(\bar{X}', \bar{Y}') = (7, 7)$.)

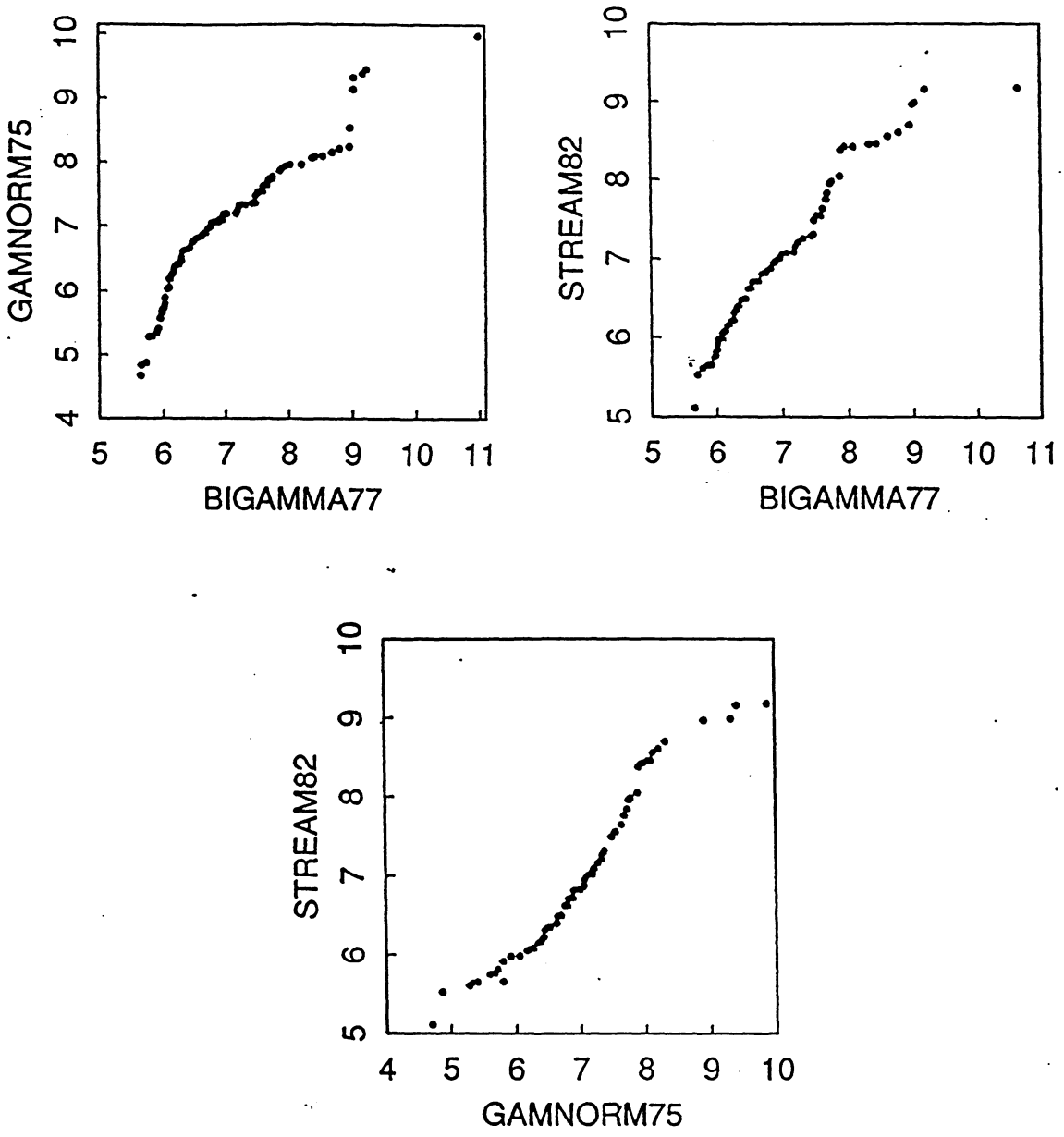
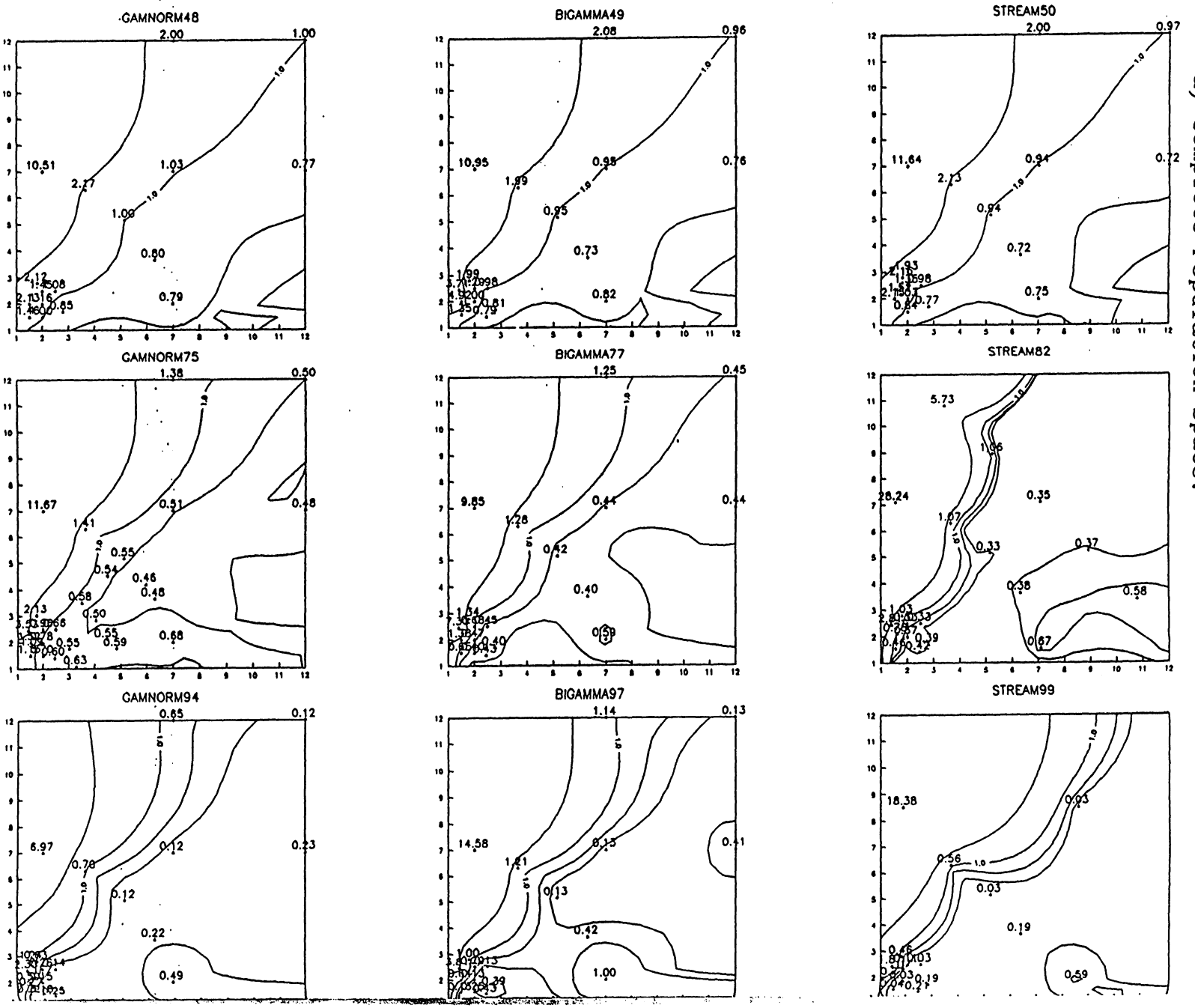


Figure 3. Population Quantile-Quantile Plots of the Variable y for the 3 Families.
(Middle correlation subfamilies at $(\bar{X}', \bar{Y}') = (7, 7)$.)

Figure 4. Standardized Variance, $V(\hat{T}_y)/V_{SRS}$.
a) Complete Population Space.
b) Enlargement of Lower Left Corner.
(Contours plotted: 0.25, 0.50, 1.0, and 3.0.)



a) Complete Population Space.

Figure 4 (continued)

b) Enlargement of Lower Left Corner.

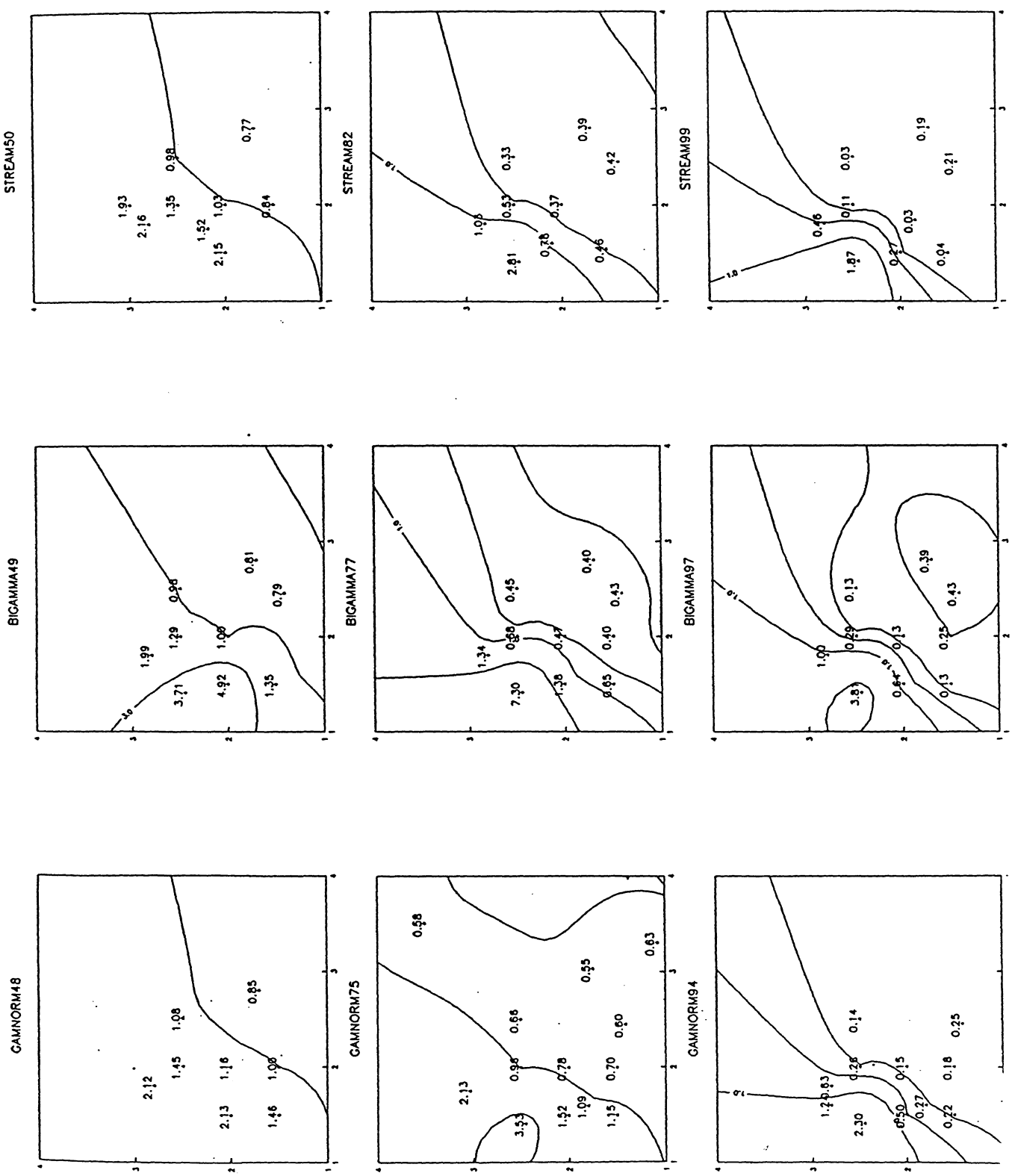
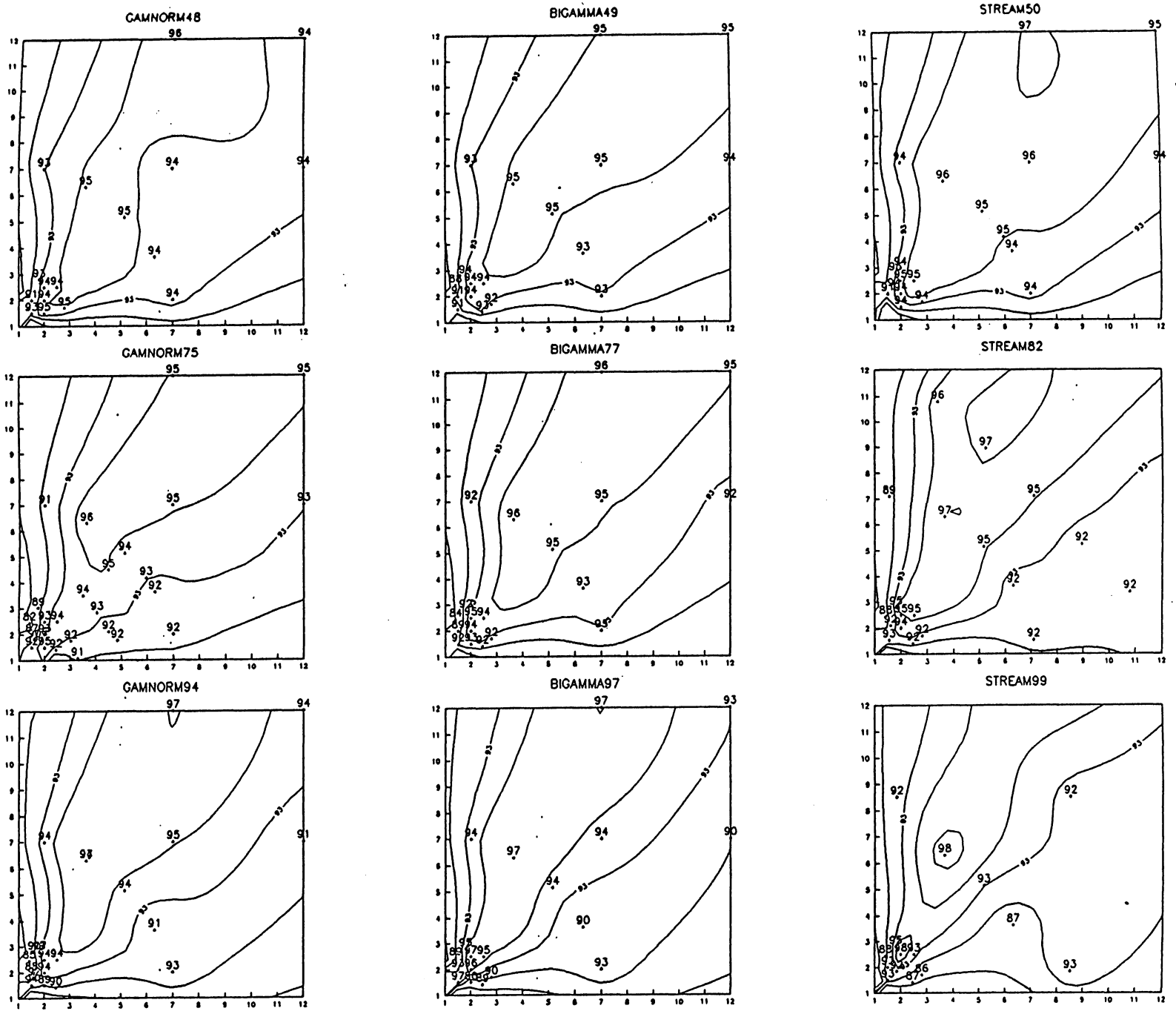


Figure 5. Confidence Interval Coverage Obtained using v°_{HT} (nominal 95% intervals).

a) Complete Population Space.

b) Enlargement of Lower Left Corner.

(Contours plotted: 85, 90, 93, 95, and 97)



a) Complete Population Space.

Figure 5 (Continued)

b) Enlargement of Lower Left Corner.

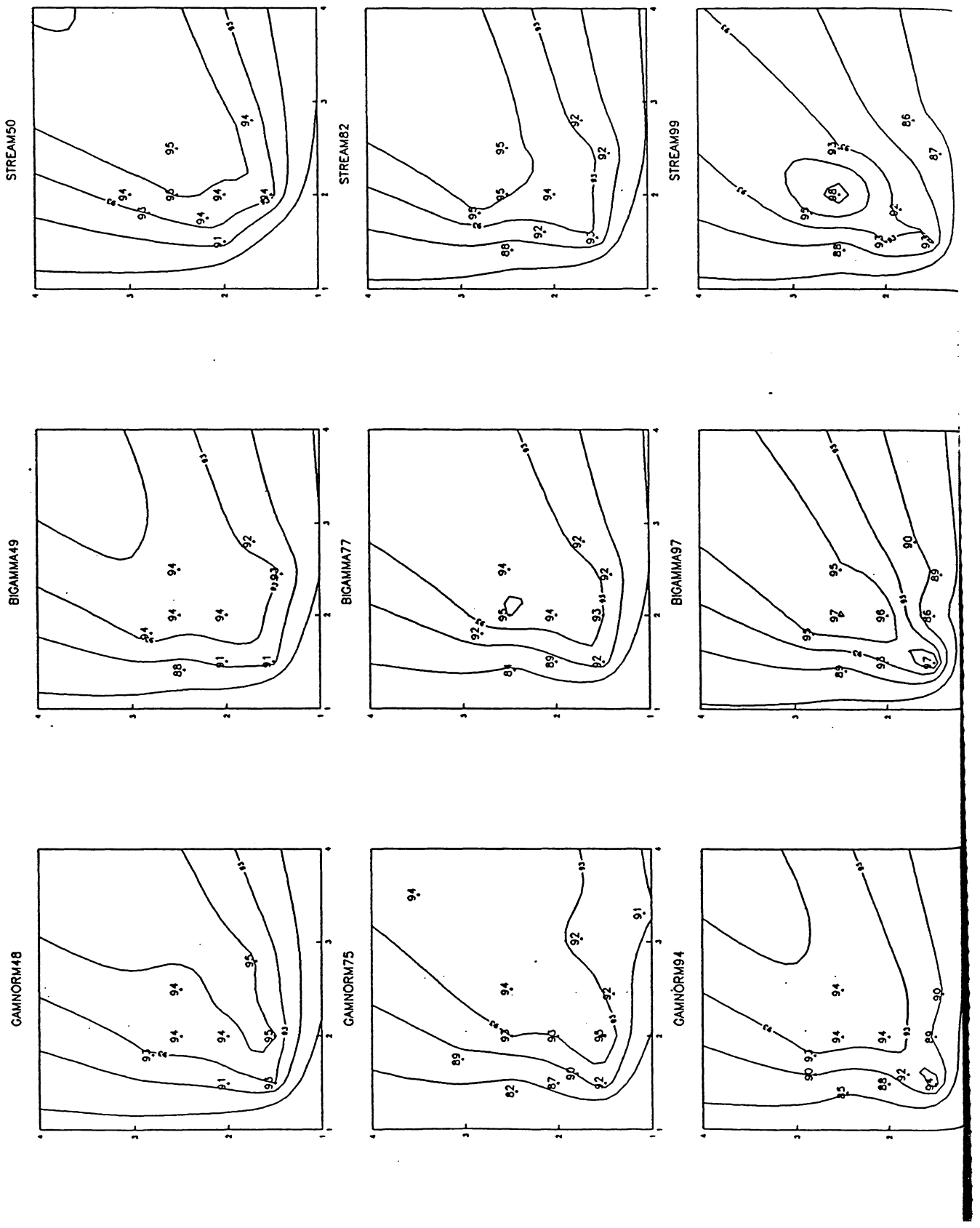
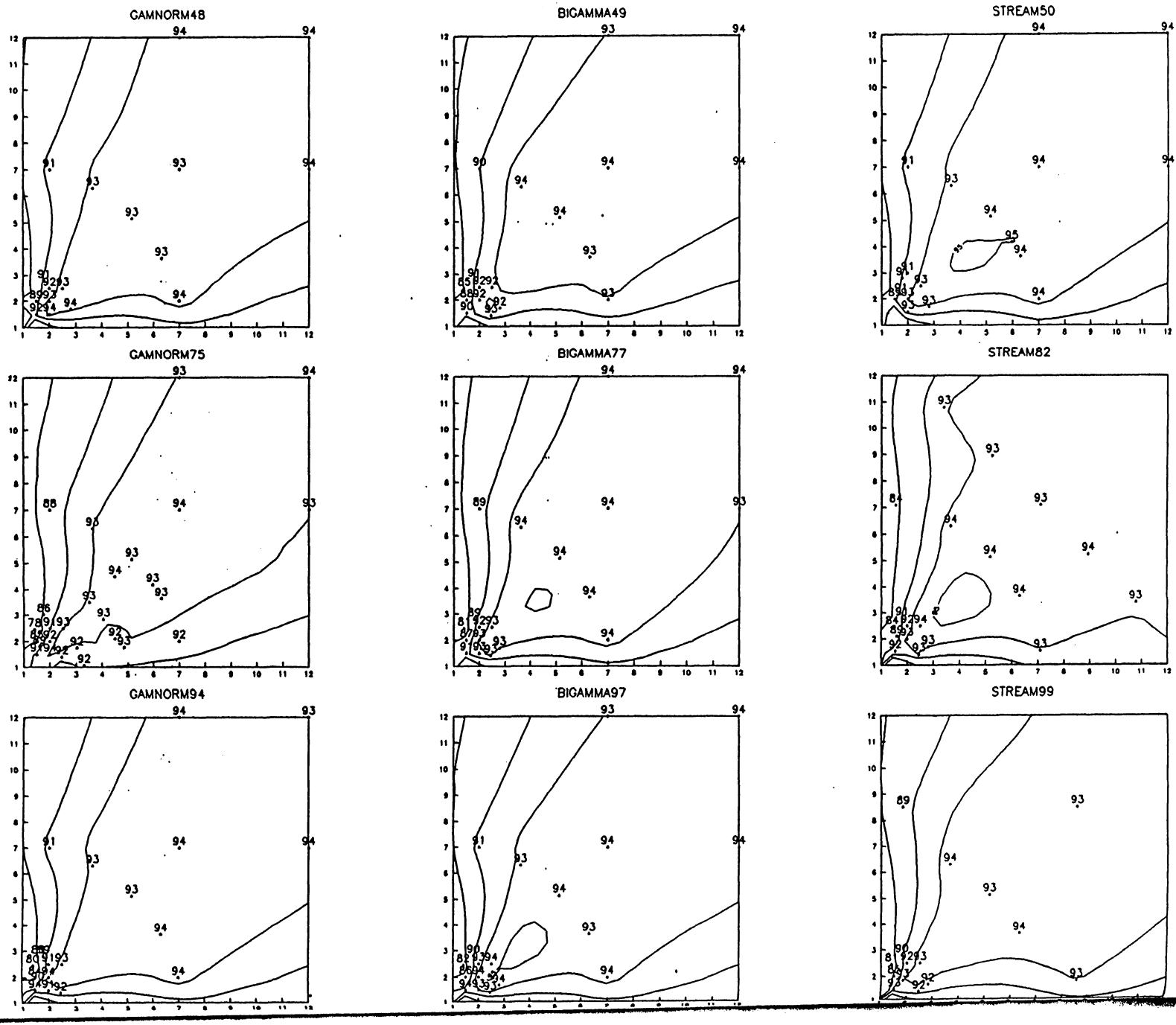


Figure 6. Confidence Interval Coverage Obtained using v°_{YG} (nominal 95% intervals).

a) Complete Population Space.

b) Enlargement of Lower Left Corner.

(Contours plotted: 85, 90, 93, 95, and 97)



a) Complete Population Space.

Figure 6 (Continued)

b) Enlargement of Lower Left Corner.

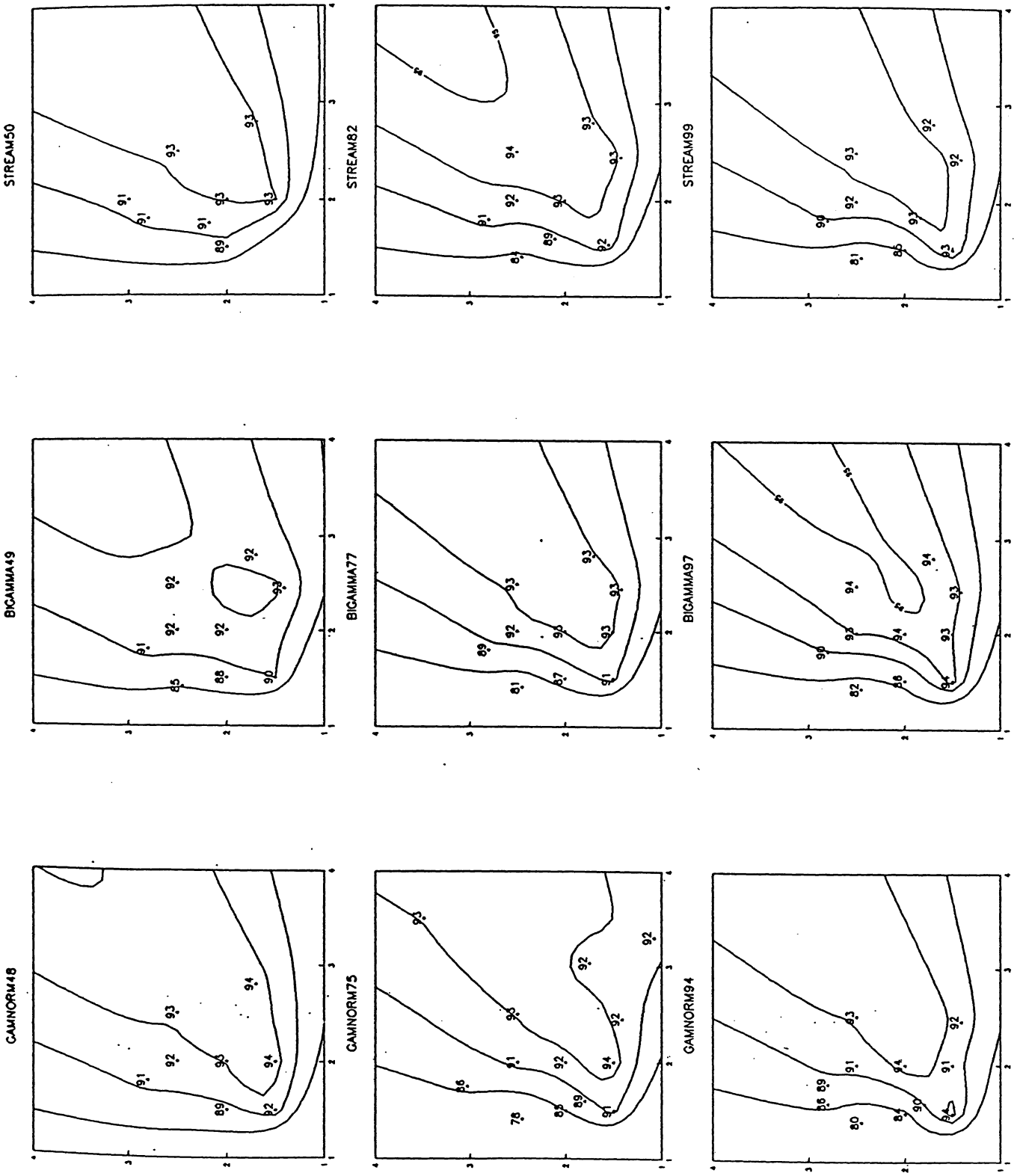
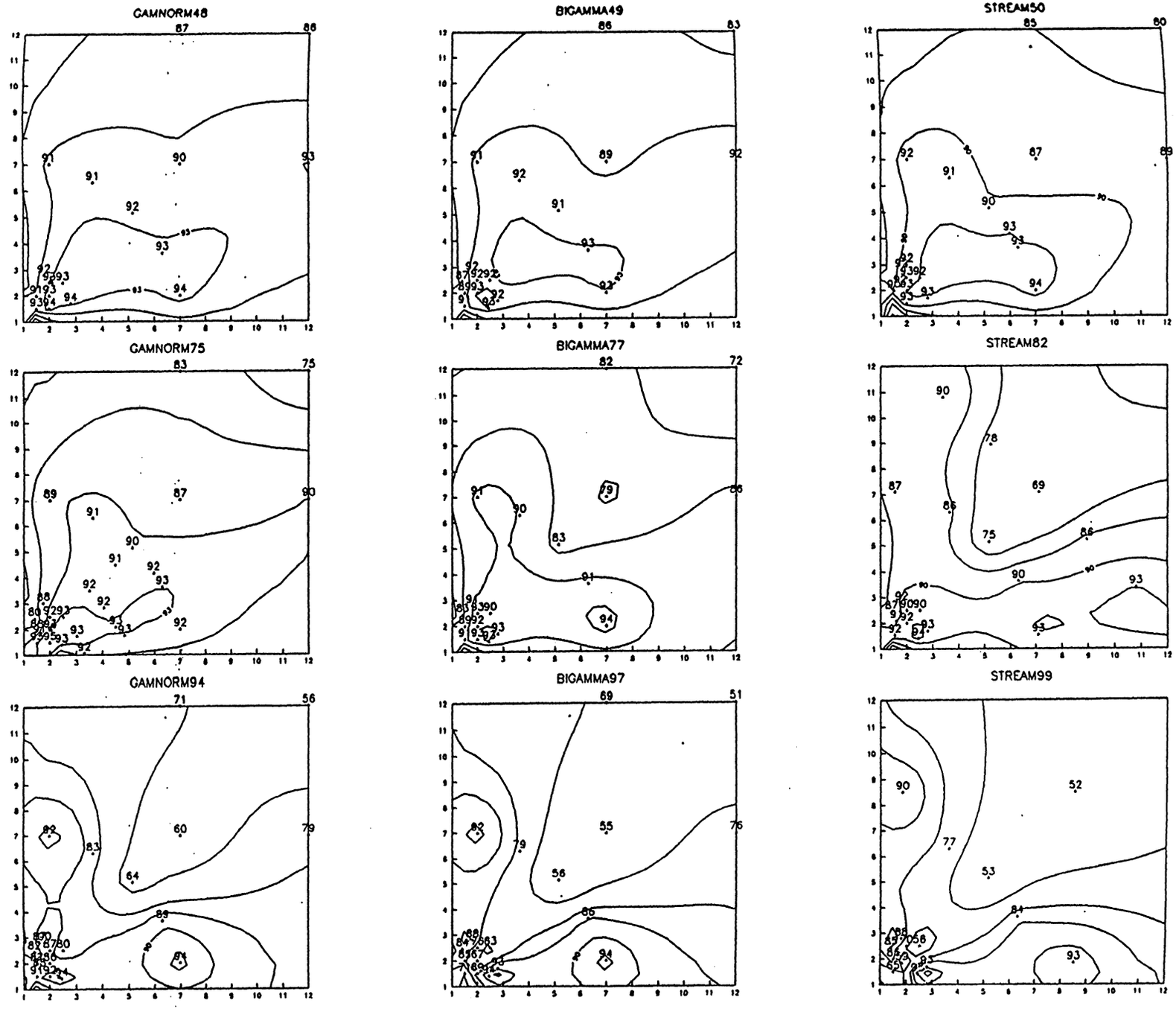


Figure 7. Confidence Interval Coverage Obtained using v_{HT}^{hr} (nominal 95% intervals).

a) Complete Population Space.

b) Enlargement of Lower Left Corner.

(Contours plotted: 70, 80, 90, 93, and 95)



a) Complete Population Space.

Figure 7 (Continued)
b) Enlargement of Lower Left Corner.

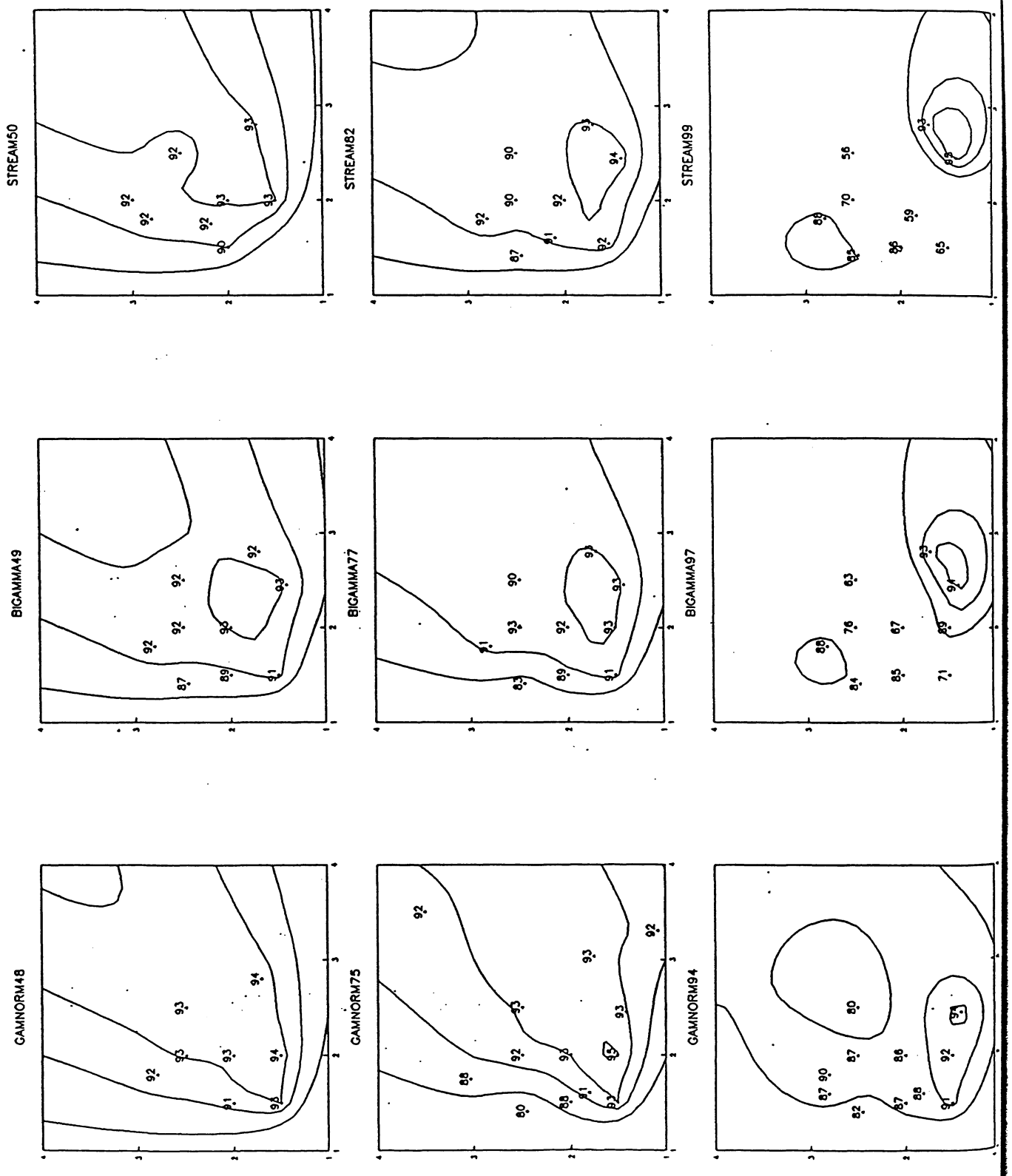


Figure 8. Confidence Interval Coverage Obtained using v_{YG}^{hr} (nominal 95% intervals).

a) Complete Population Space.

b) Enlargement of Lower Left Corner.

(Contours plotted: 85, 90, 93, 95, and 97)

a) Complete Population Space.

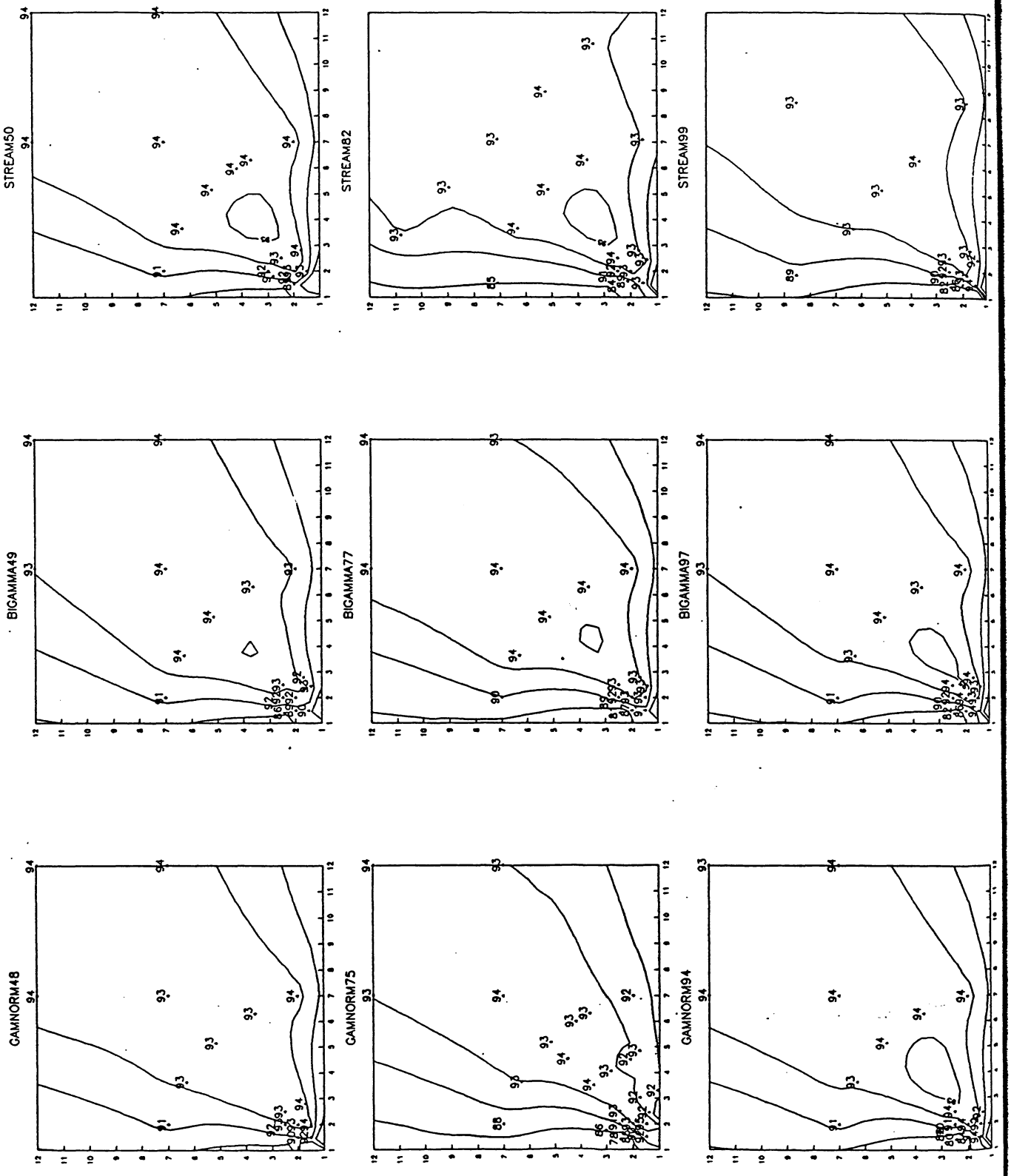


Figure 8 (Continued)

b) Enlargement of Lower Left Corner.

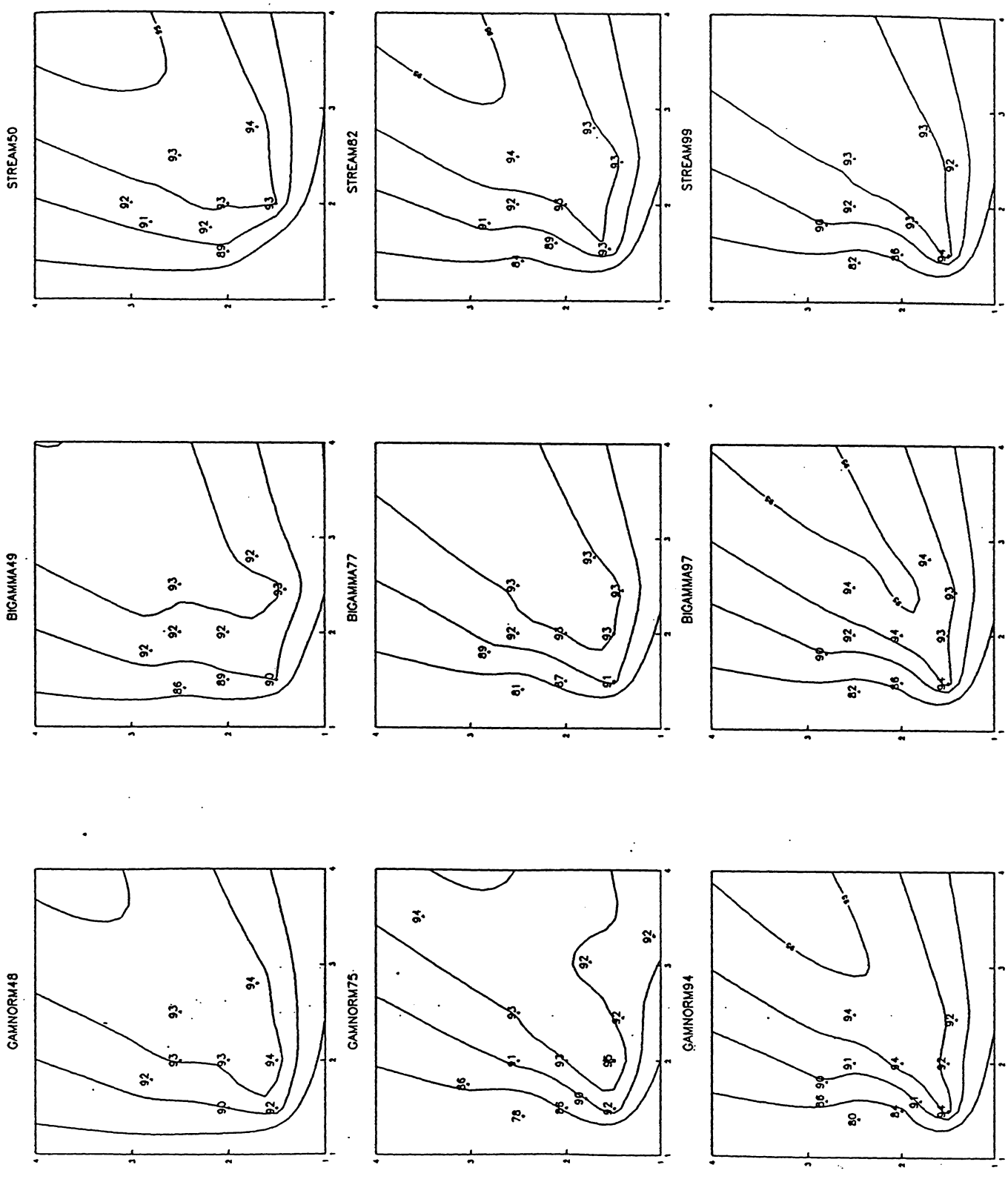
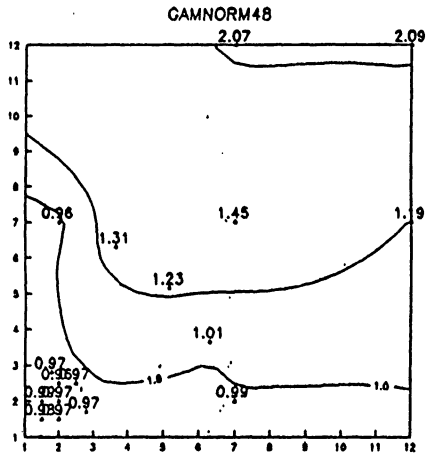
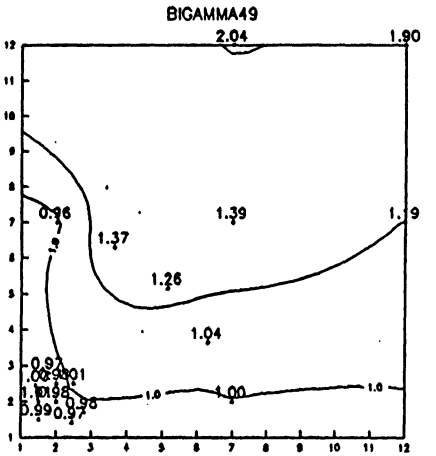
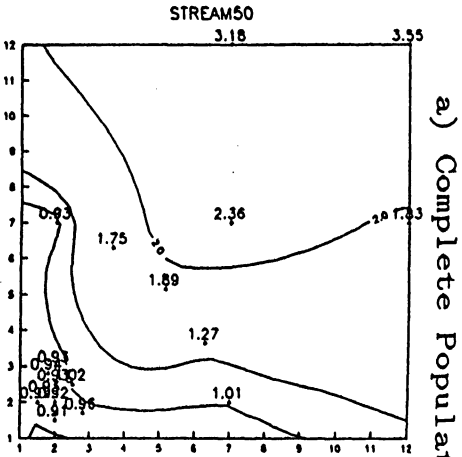


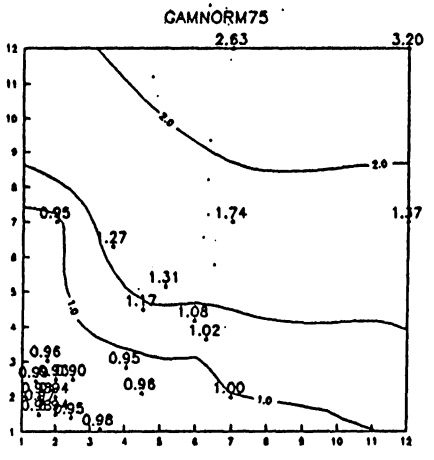
Figure 9. Ratios of Root Mean Square Errors:
$RMSE(v_{HT}^{hr})/RMSE(v_{YG}^{hr})$.
a) Complete Population Space.
b) Enlargement of Lower Left Corner.

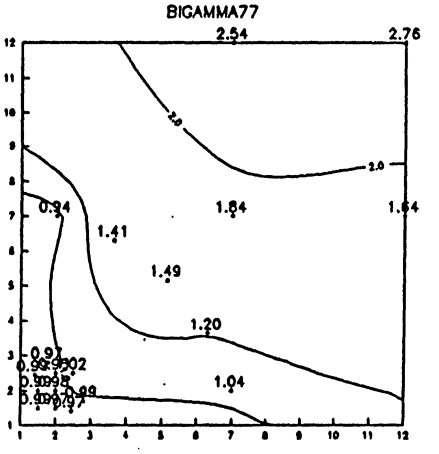
(Contours plotted: 0.8, 1.0, 1.2, 2.0, 5.0)

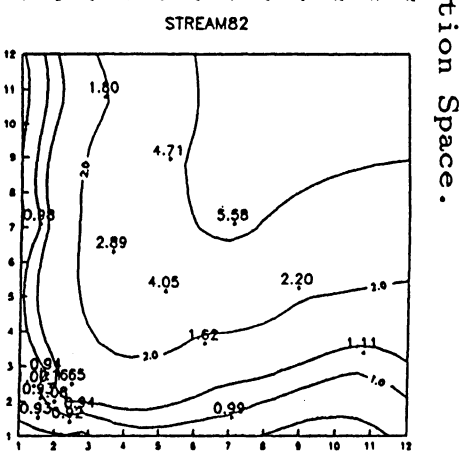


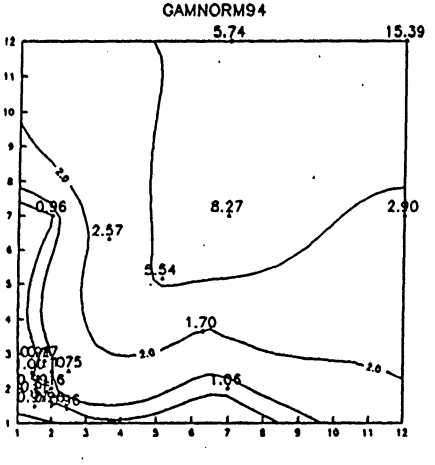


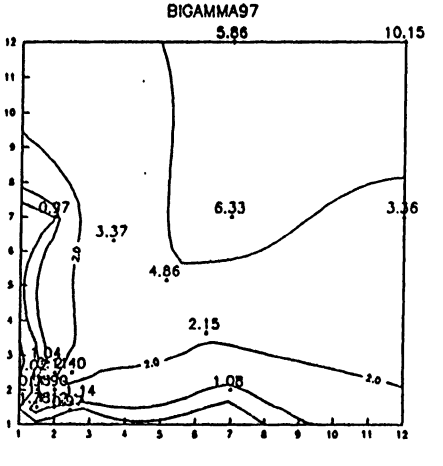


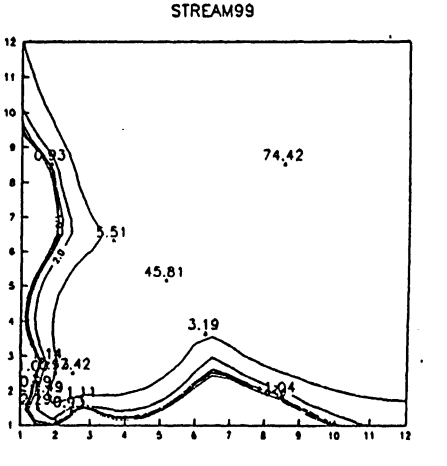












a) Complete Population Space.

Figure 9 (Continued)
b) Enlargement of Lower Left Corner.

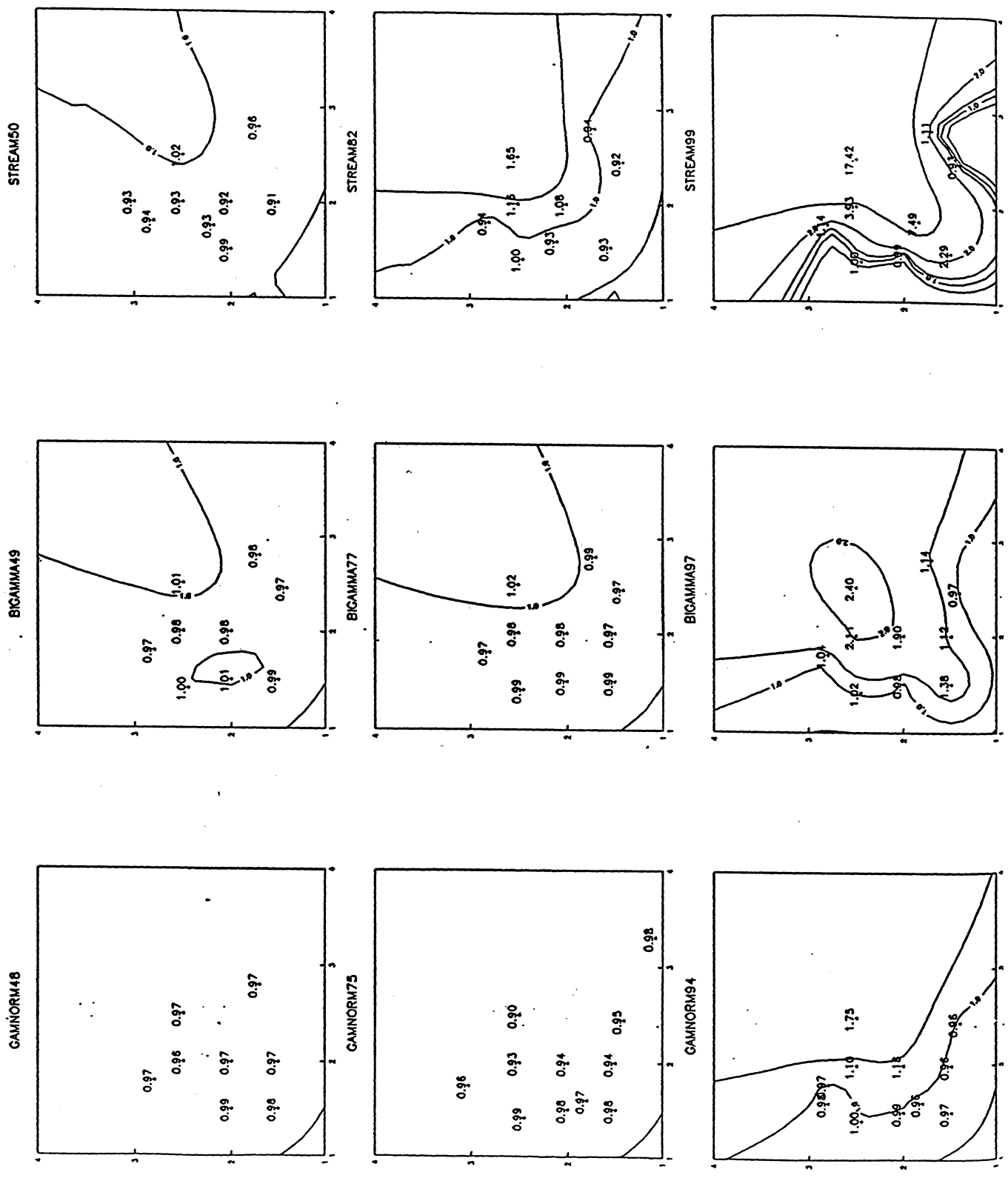


Figure 10. Ratios of Root Mean Square Errors:
$RMSE(v^{o}_{HT})/RMSE(v^{hr}_{HT})$.
a) Complete Population Space.
b) Enlargement of Lower Left Corner.

(Contours plotted: 0.8, 1.0, 1.2, 2.0, 5.0)

a) Complete Population Space.

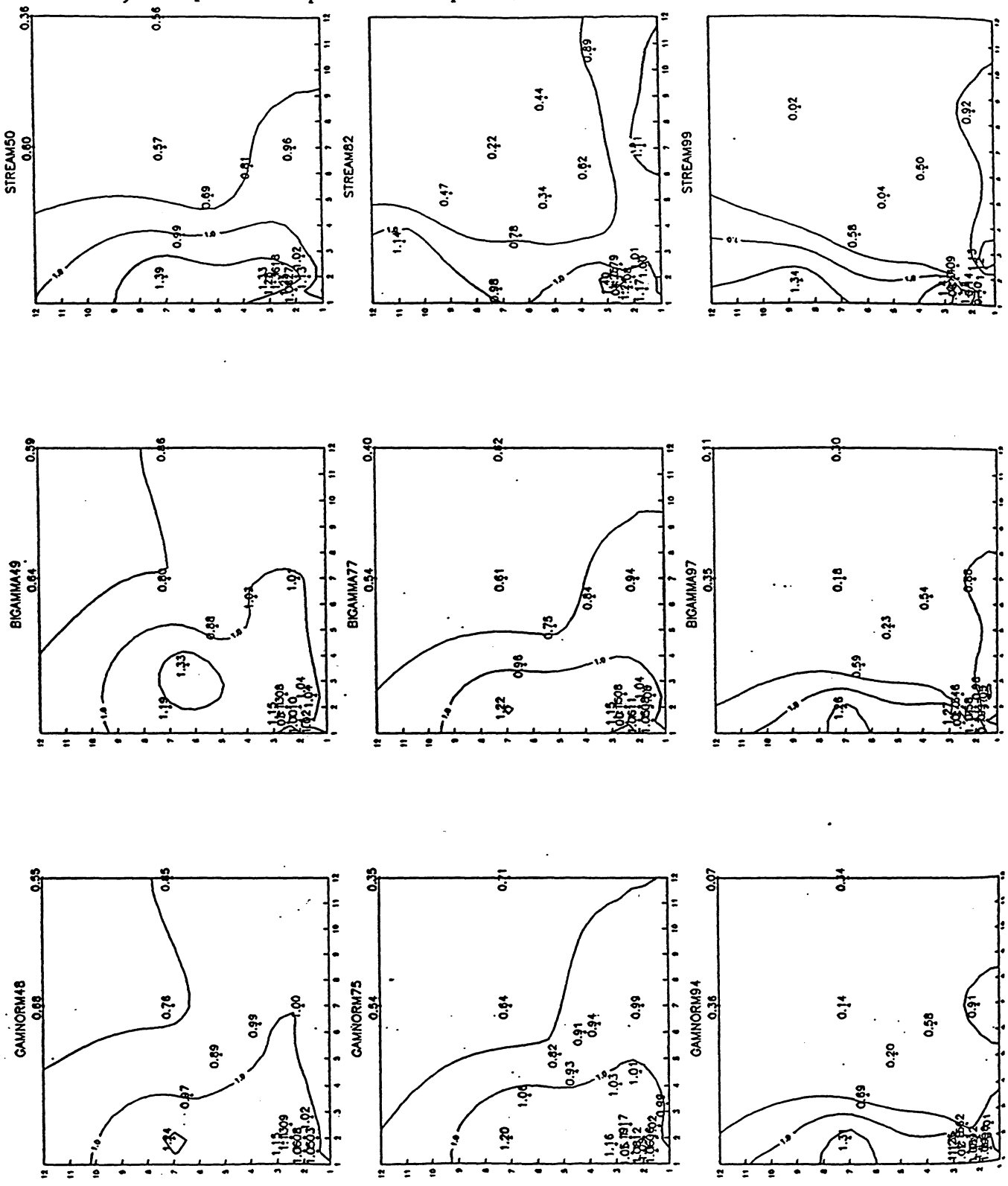


Figure 10 (Continued)

b) Enlargement of Lower Left Corner.

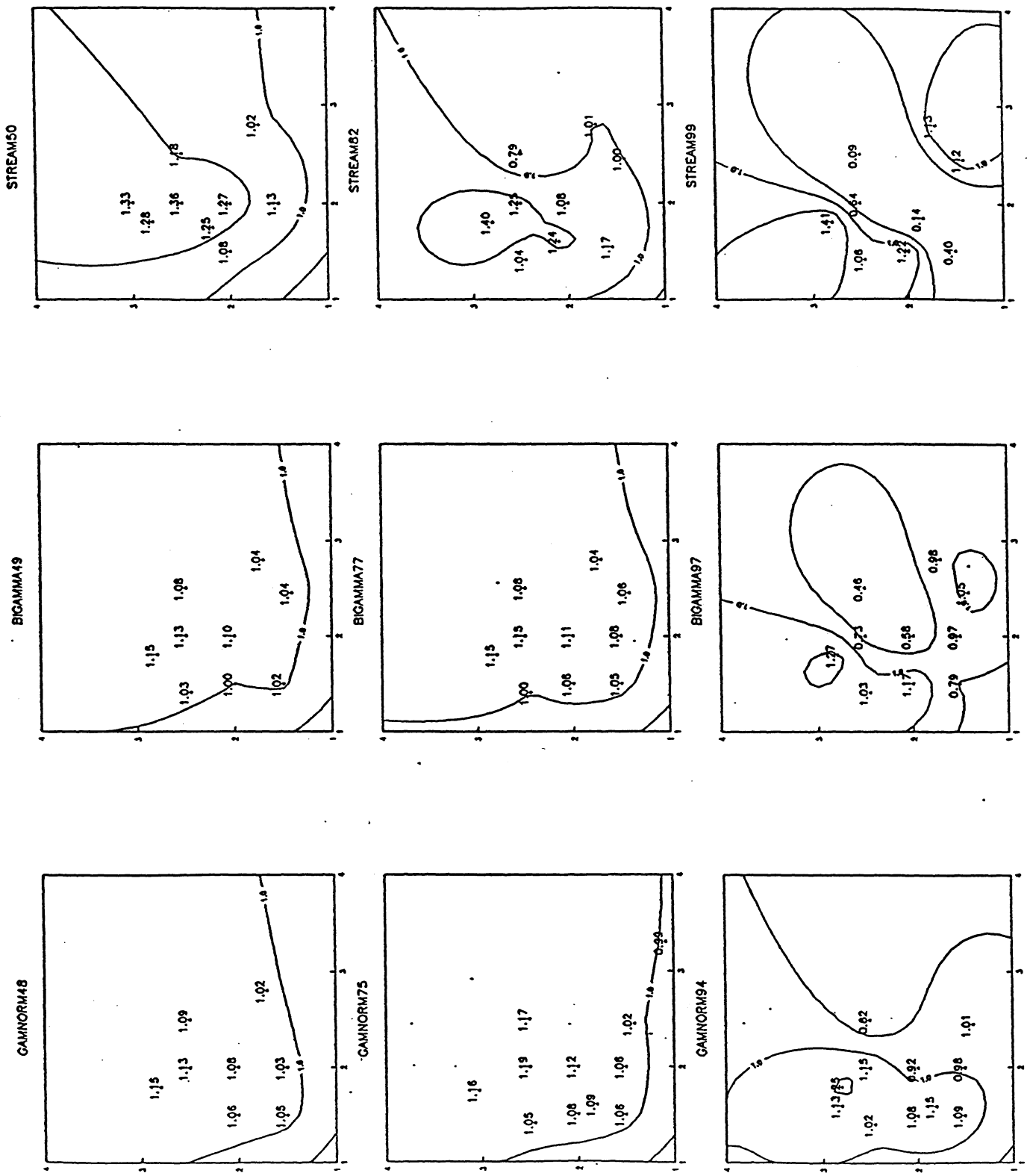
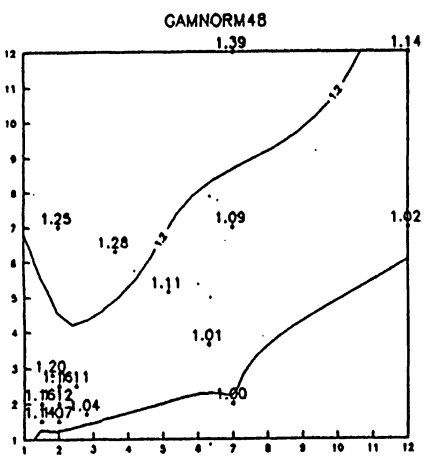


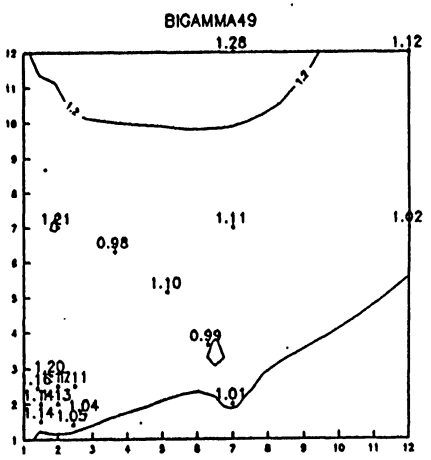
Figure 11. Ratios of Root Mean Square Errors: $RMSE(\hat{v}^{\circ}_{HT})/RMSE(v^{\circ}_{YG})$.

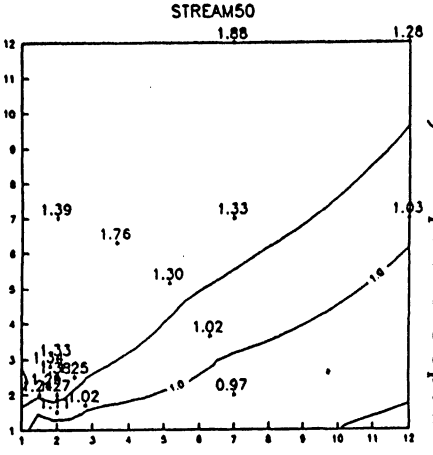
a) Complete Population Space.

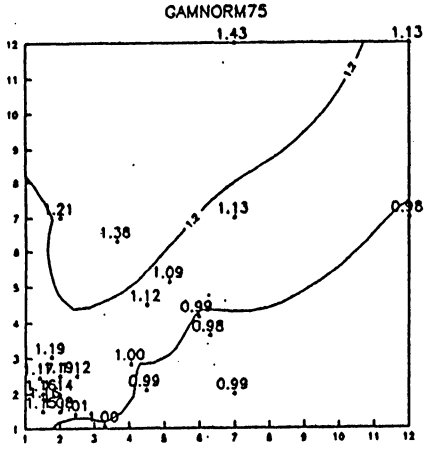
b) Enlargement of Lower Left Corner.

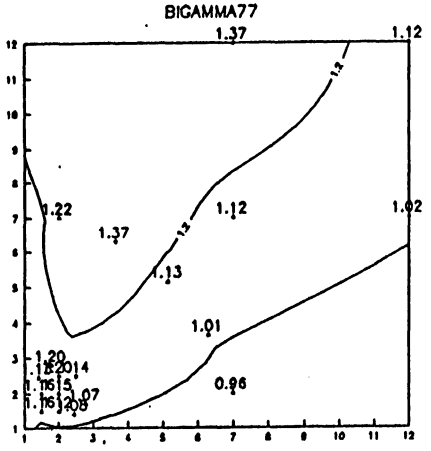
(Contours plotted: 0.8, 1.0, 1.2, 2.0, 5.0)

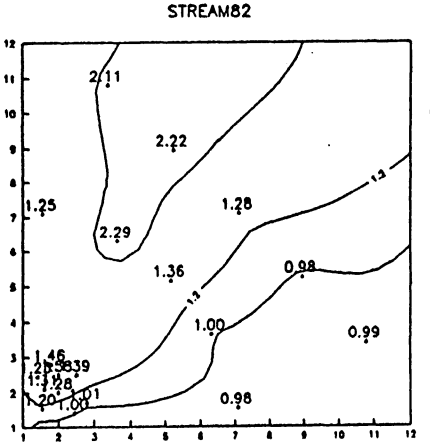


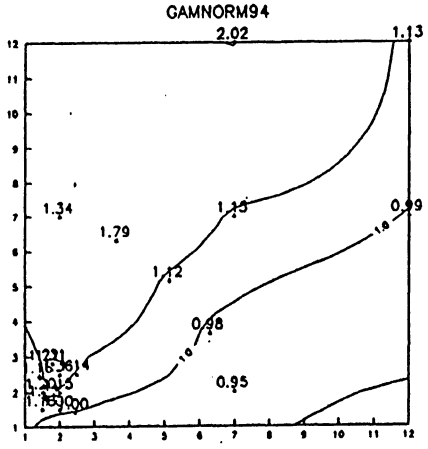


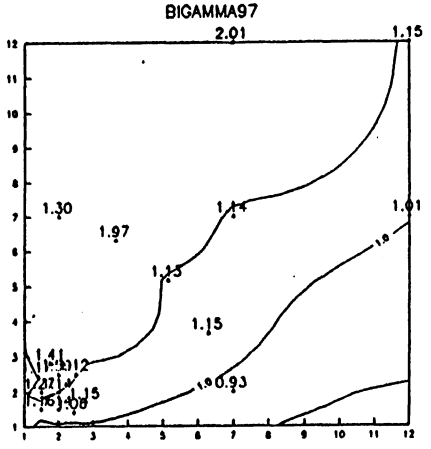


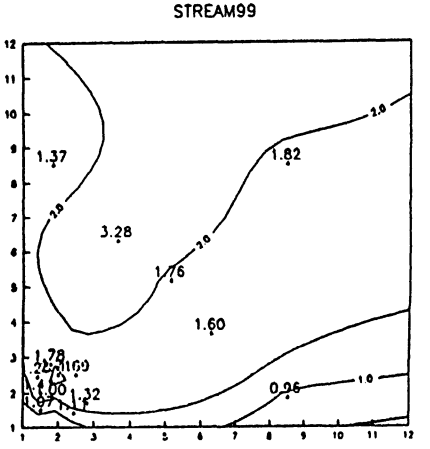












a) Complete Population Space.

Figure 11 (Continued)
b) Enlargement of Lower Left Corner.

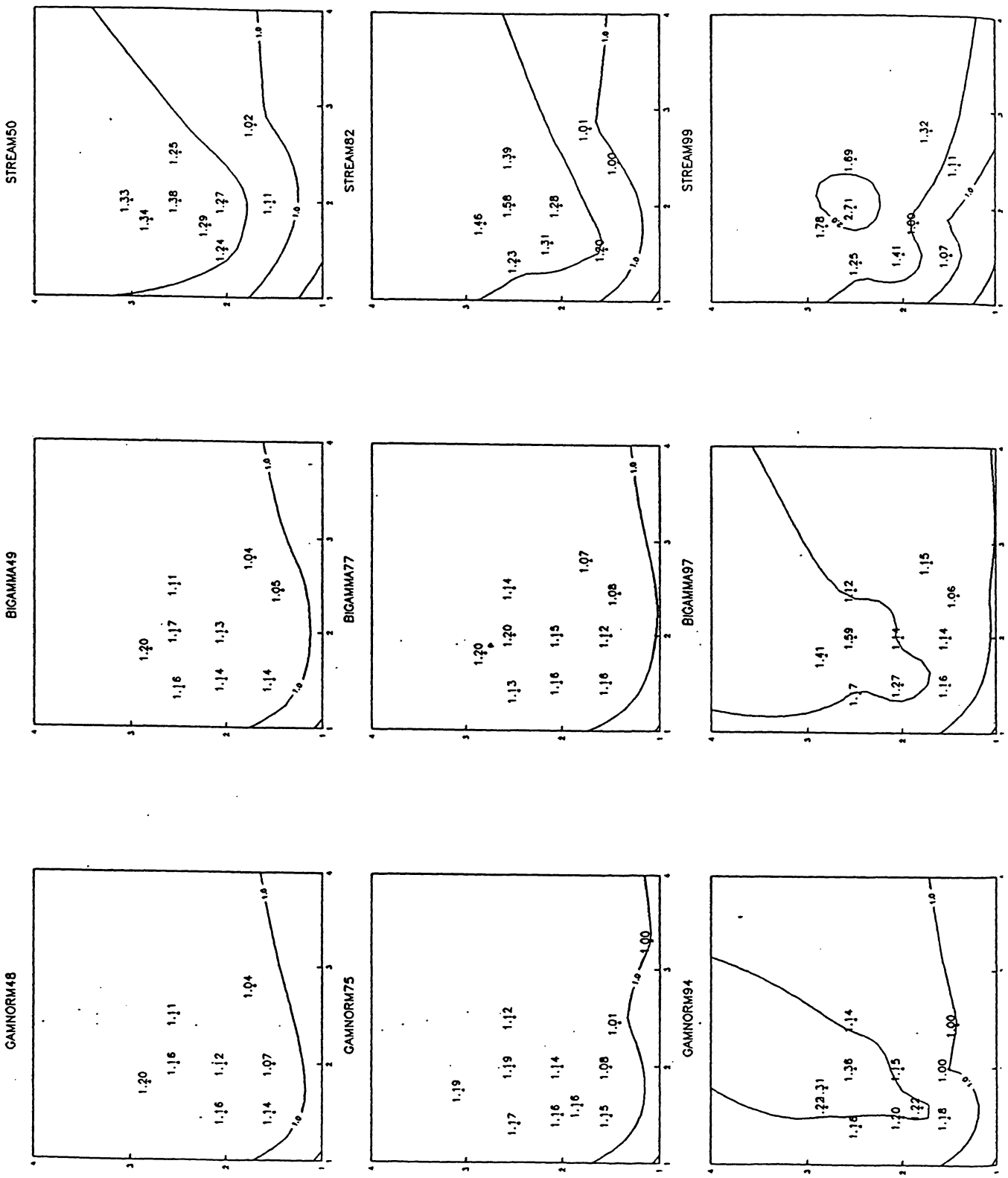


Figure 12. Ratios of Root Mean Square Errors: $RMSE(v^{o}_{YG})/RMSE(v^{hr}_{YG})$.

a) Complete Population Space.

b) Enlargement of Lower Left Corner.

(Contours plotted: 0.8, 1.0, 1.2, 2.0, 5.0)

a) Complete Population Space.

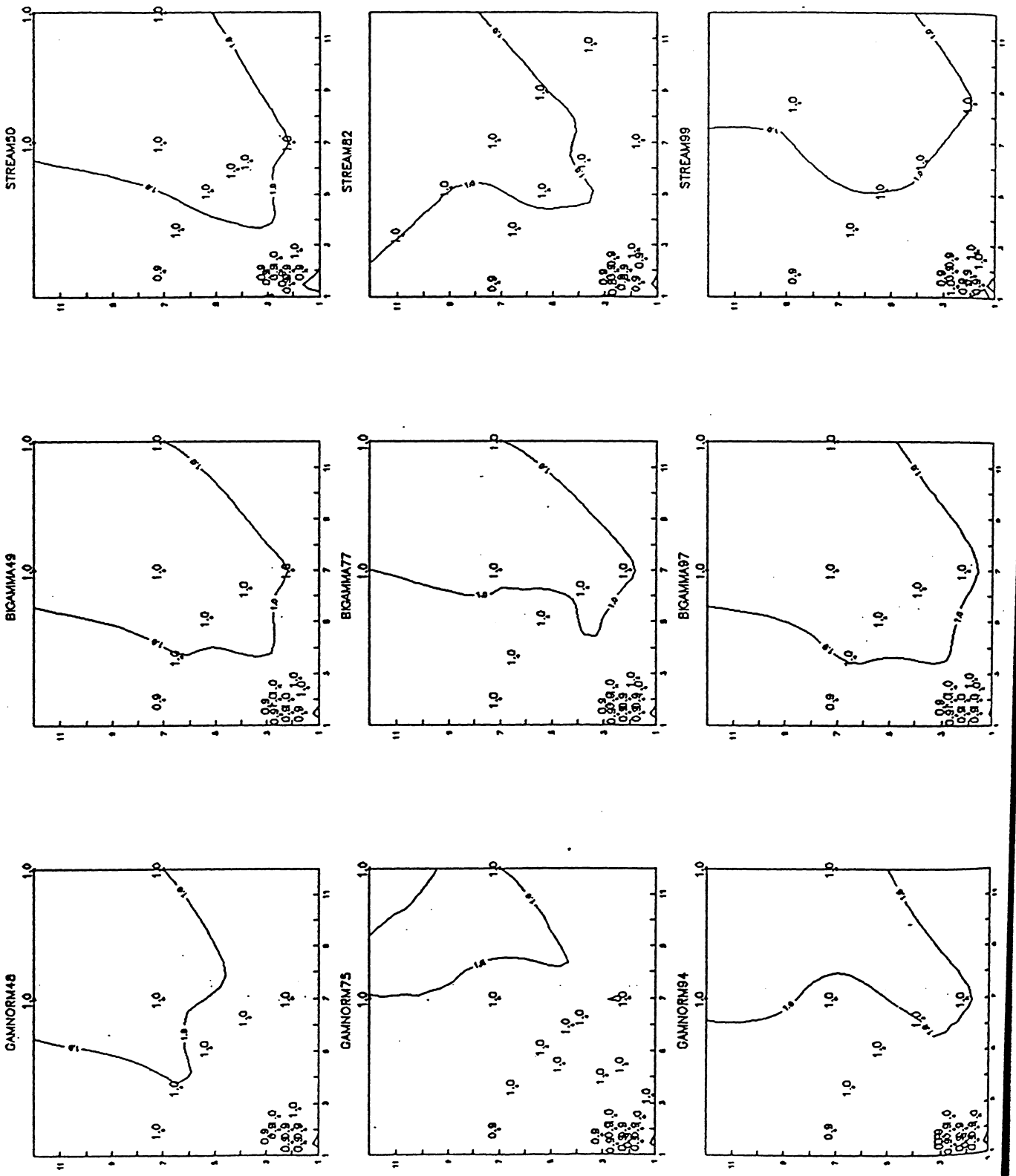


Figure 12 (Continued)
b) Enlargement of Lower Left Corner.

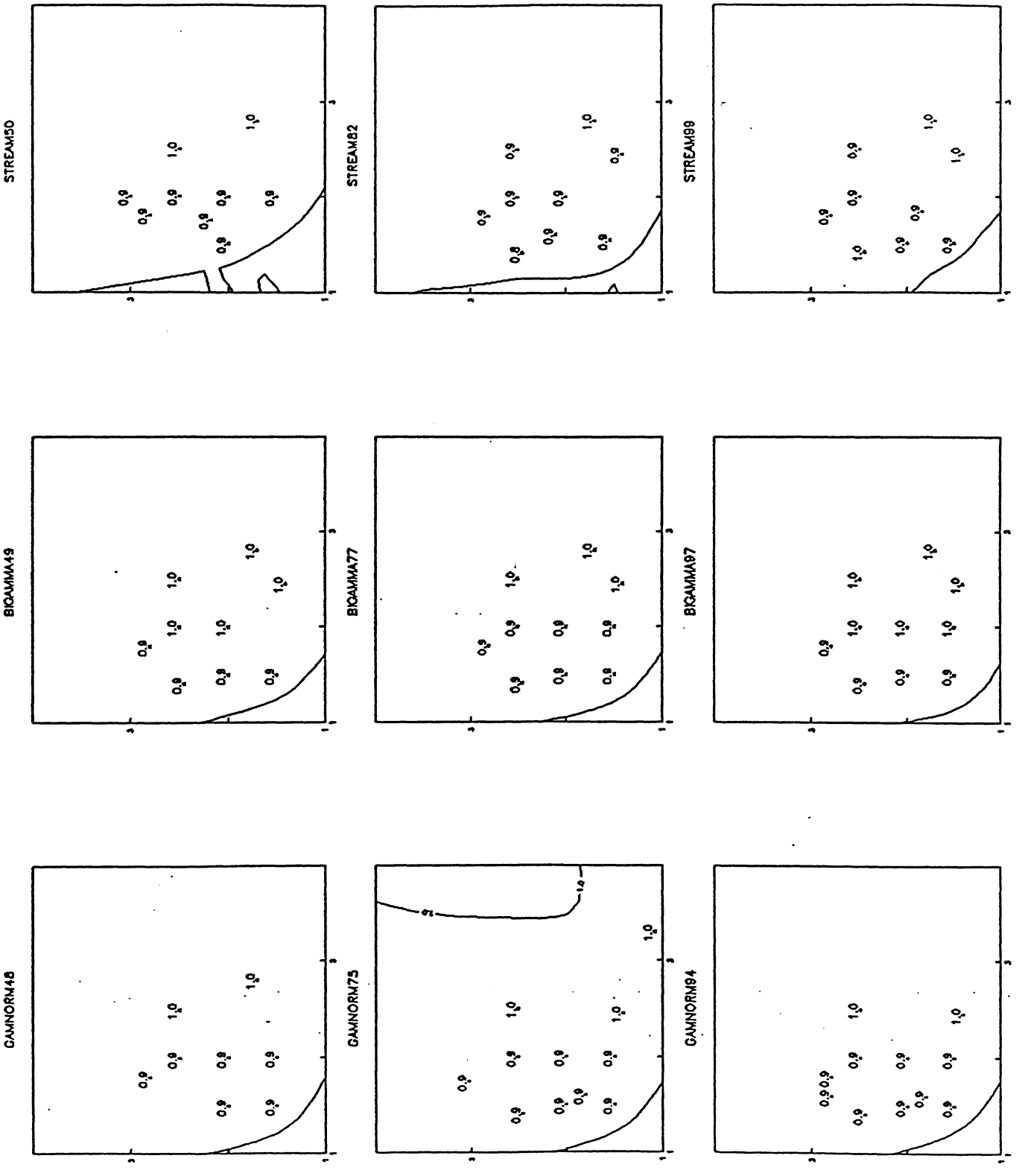
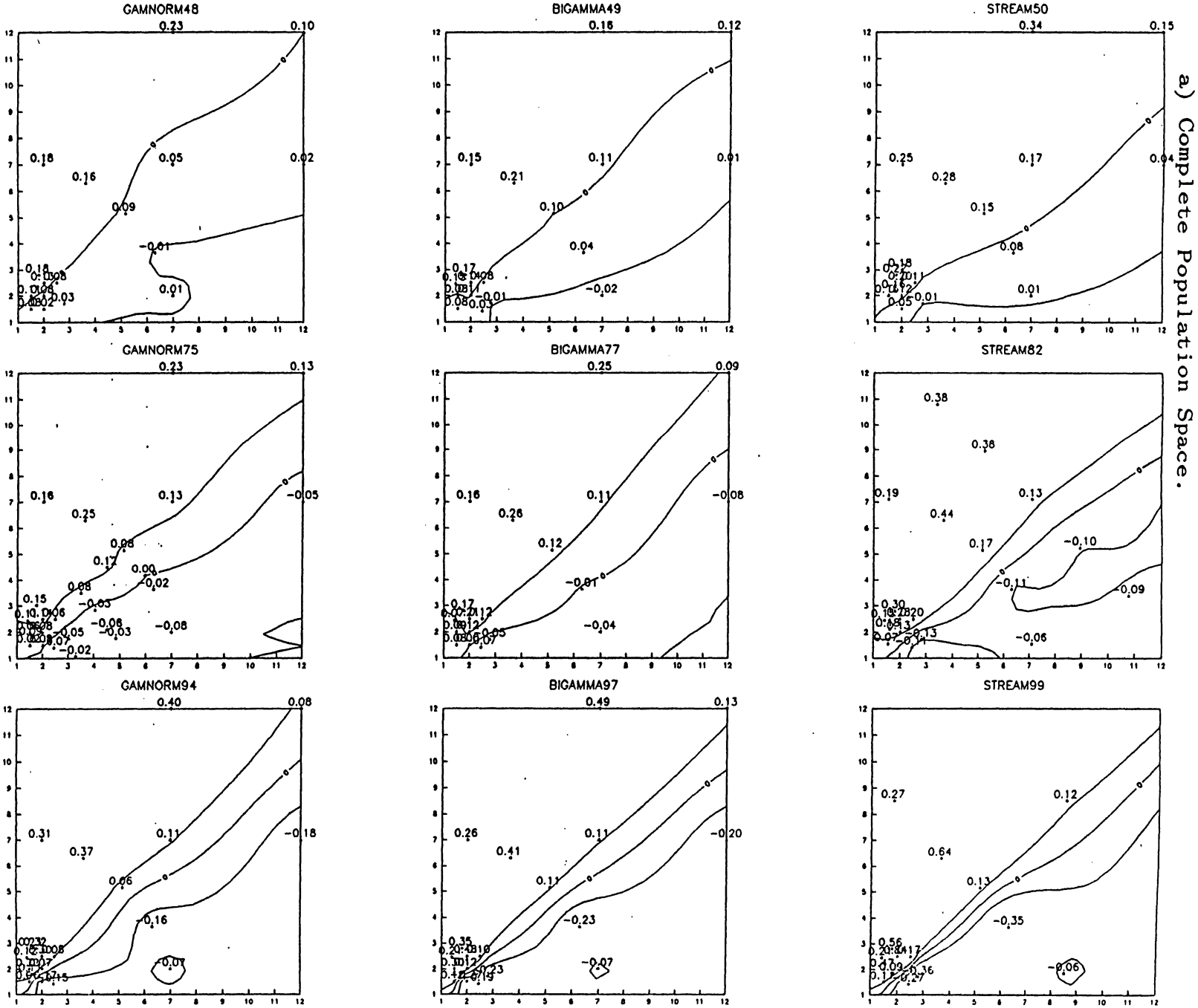


Figure 13. Relative Bias of v^{o}_{HT}.
a) Complete Population Space.
b) Enlargement of Lower Left Corner.

(Contours plotted are: -0.10, 0.0, 0.10)



a) Complete Population Space.

Figure 13 (Continued)

b) Enlargement of Lower Left Corner.

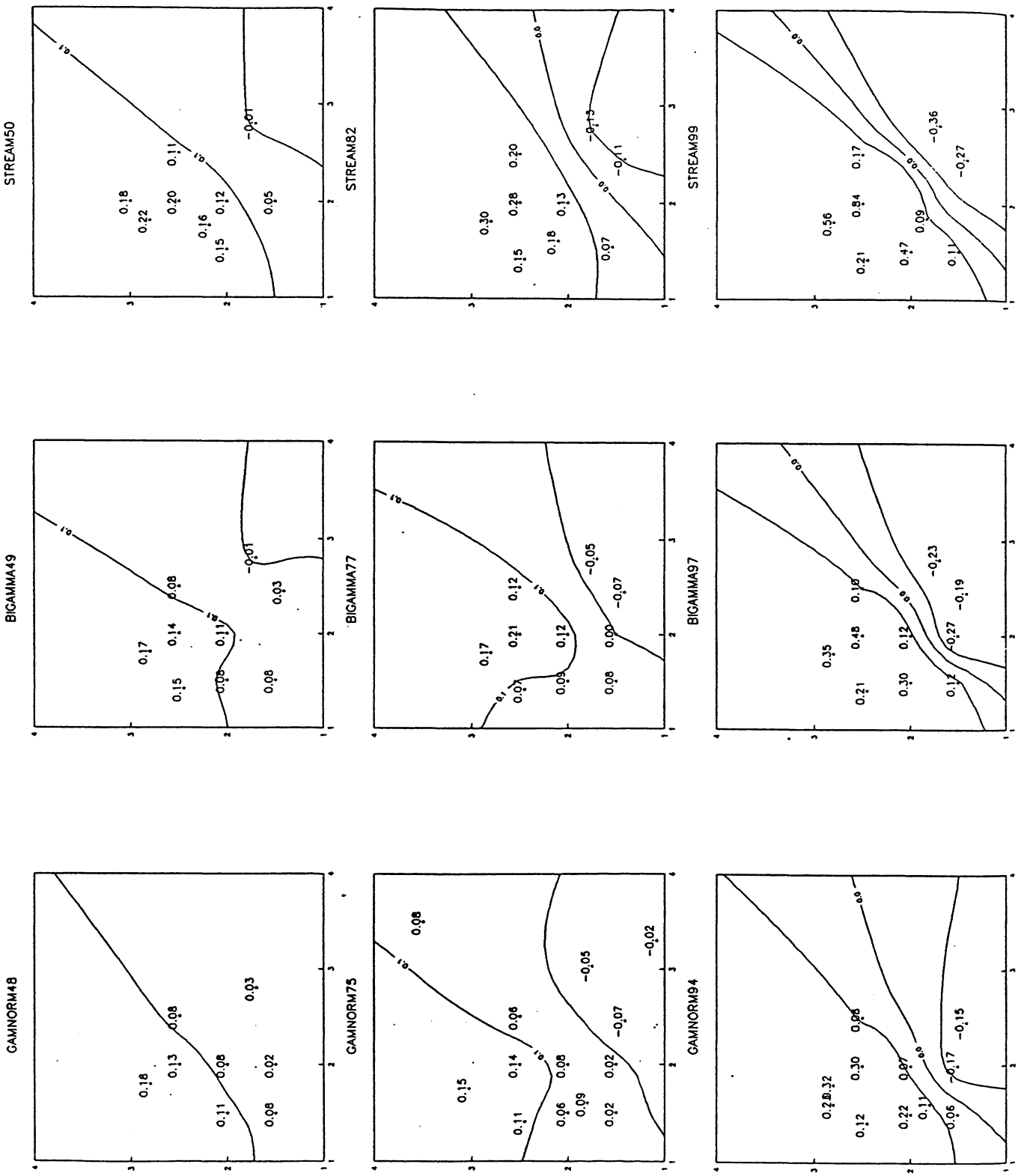
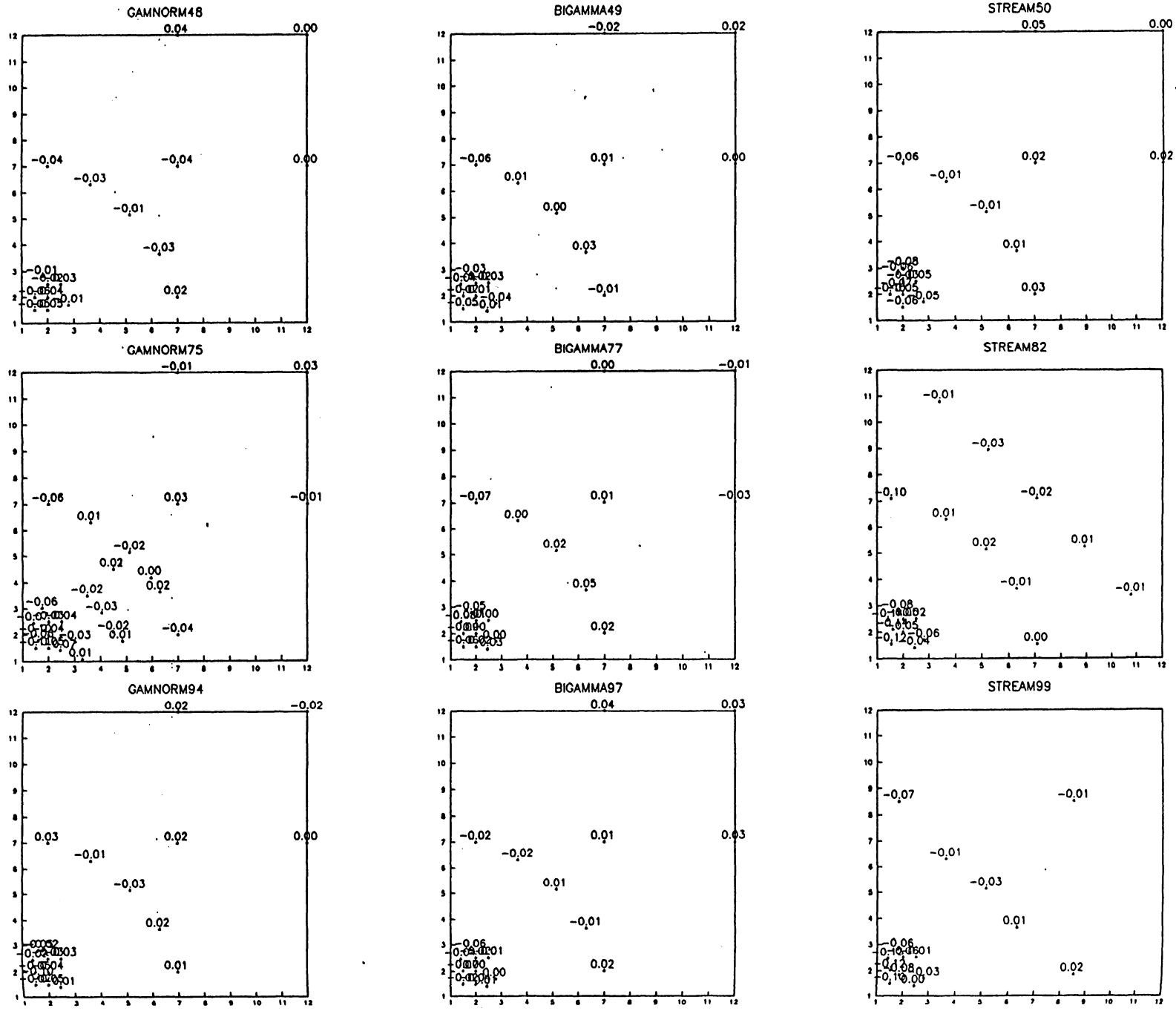


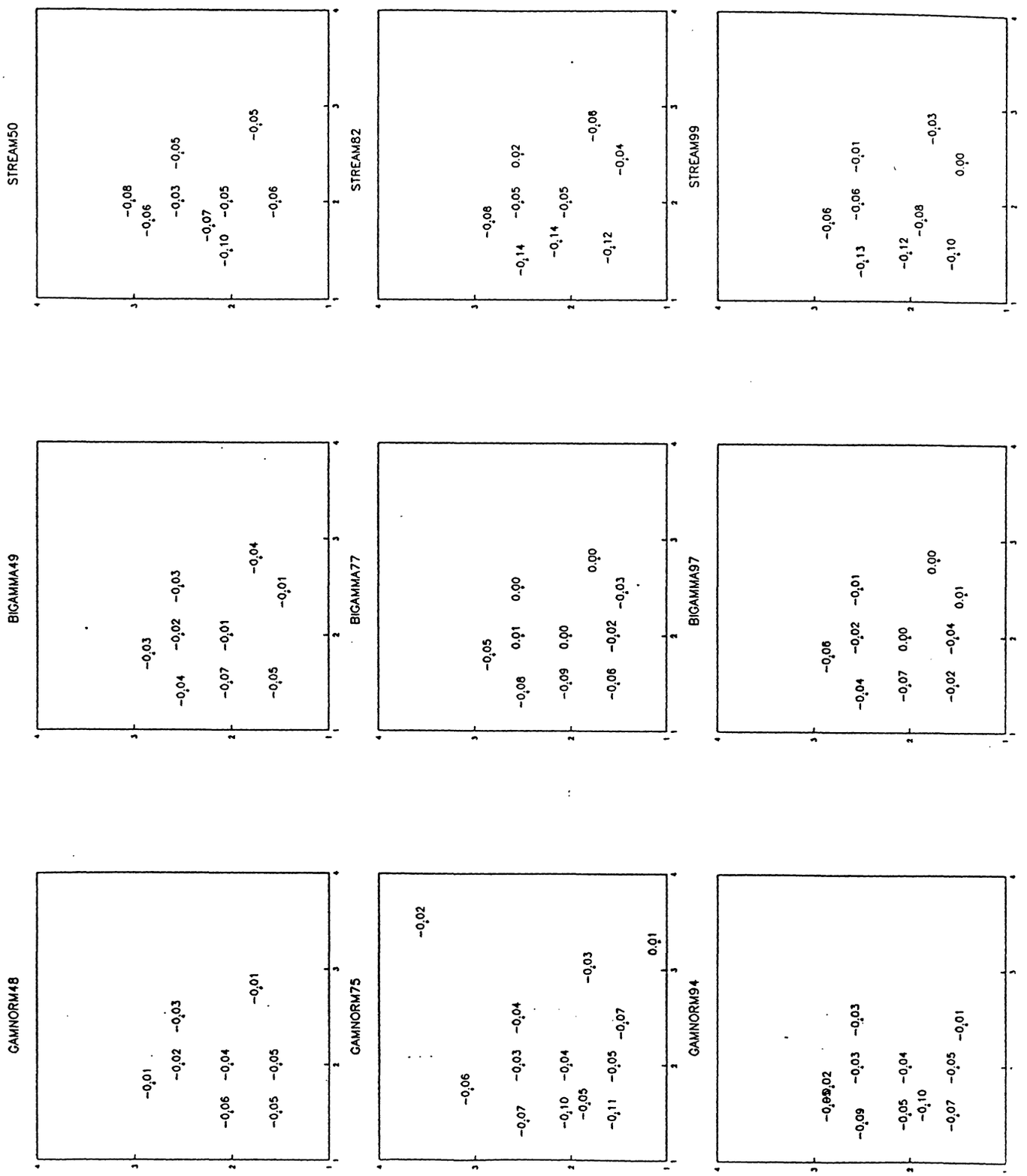
Figure 14. Relative Bias of v^{o}_{YG}.
a) Complete Population Space.
b) Enlargement of Lower Left Corner.



a) Complete Population Space.

Figure 14 (Continued)

b) Enlargement of Lower Left Corner.



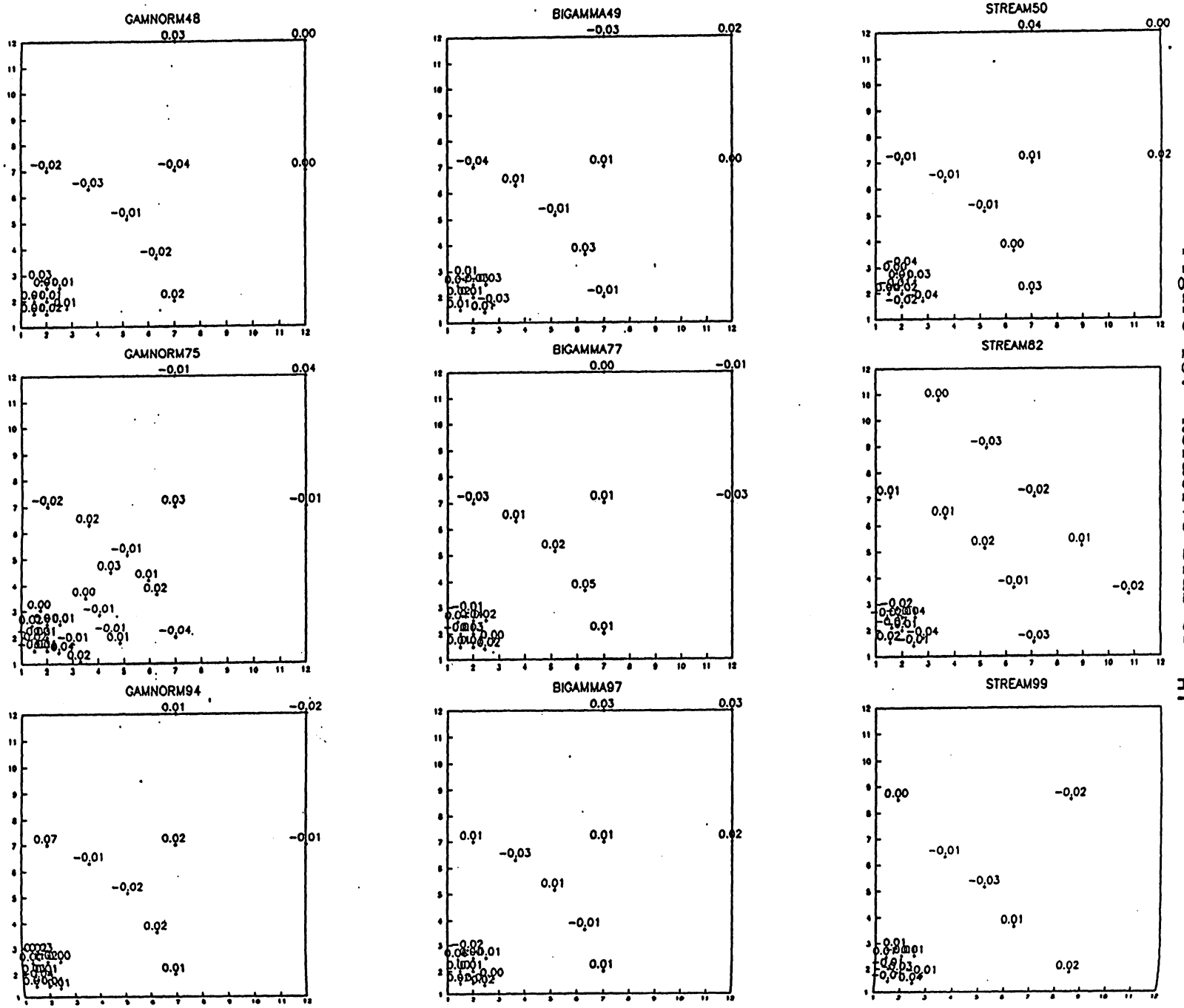


Figure 15. Relative Bias of v_{HT}^{hr}.

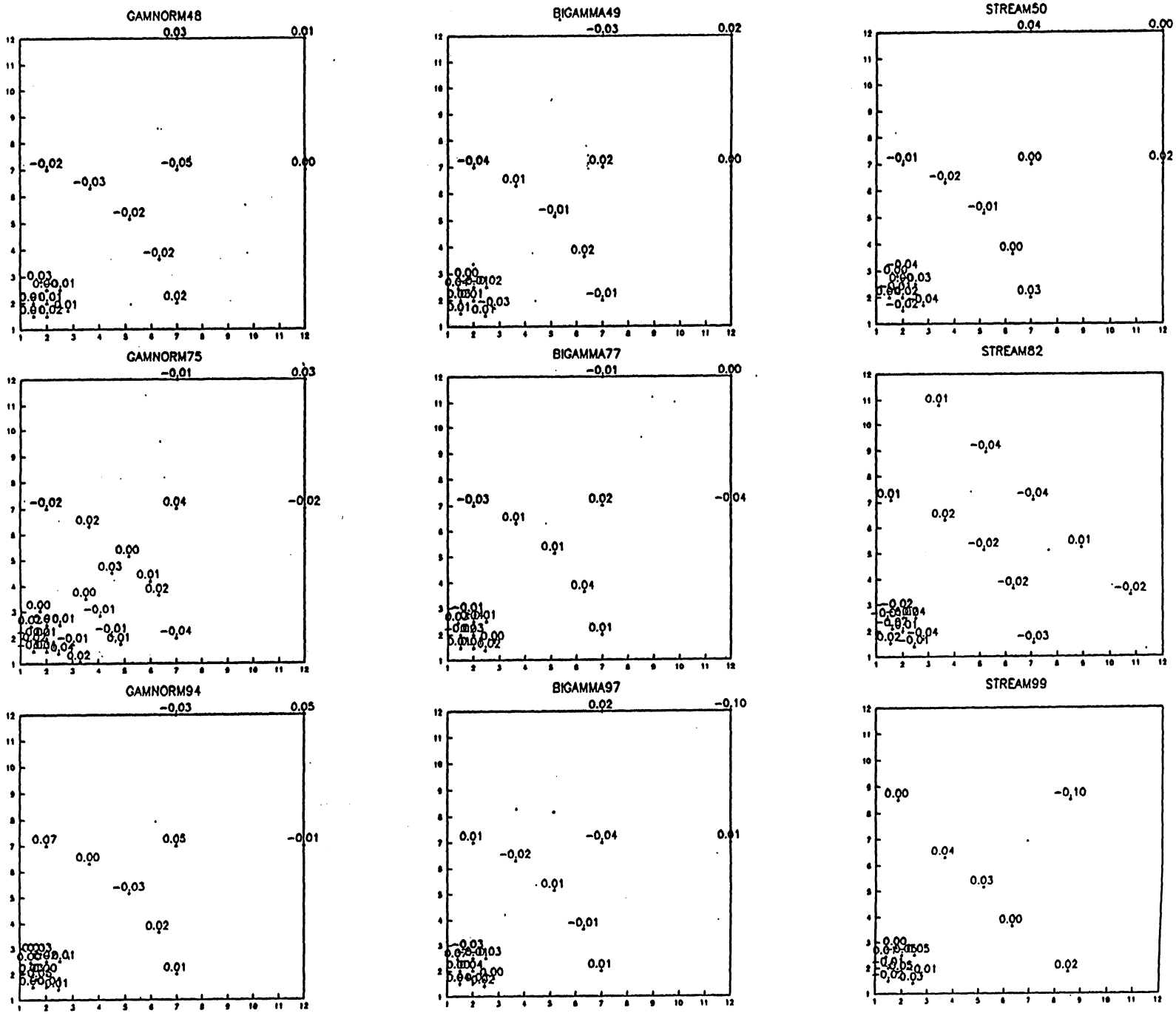
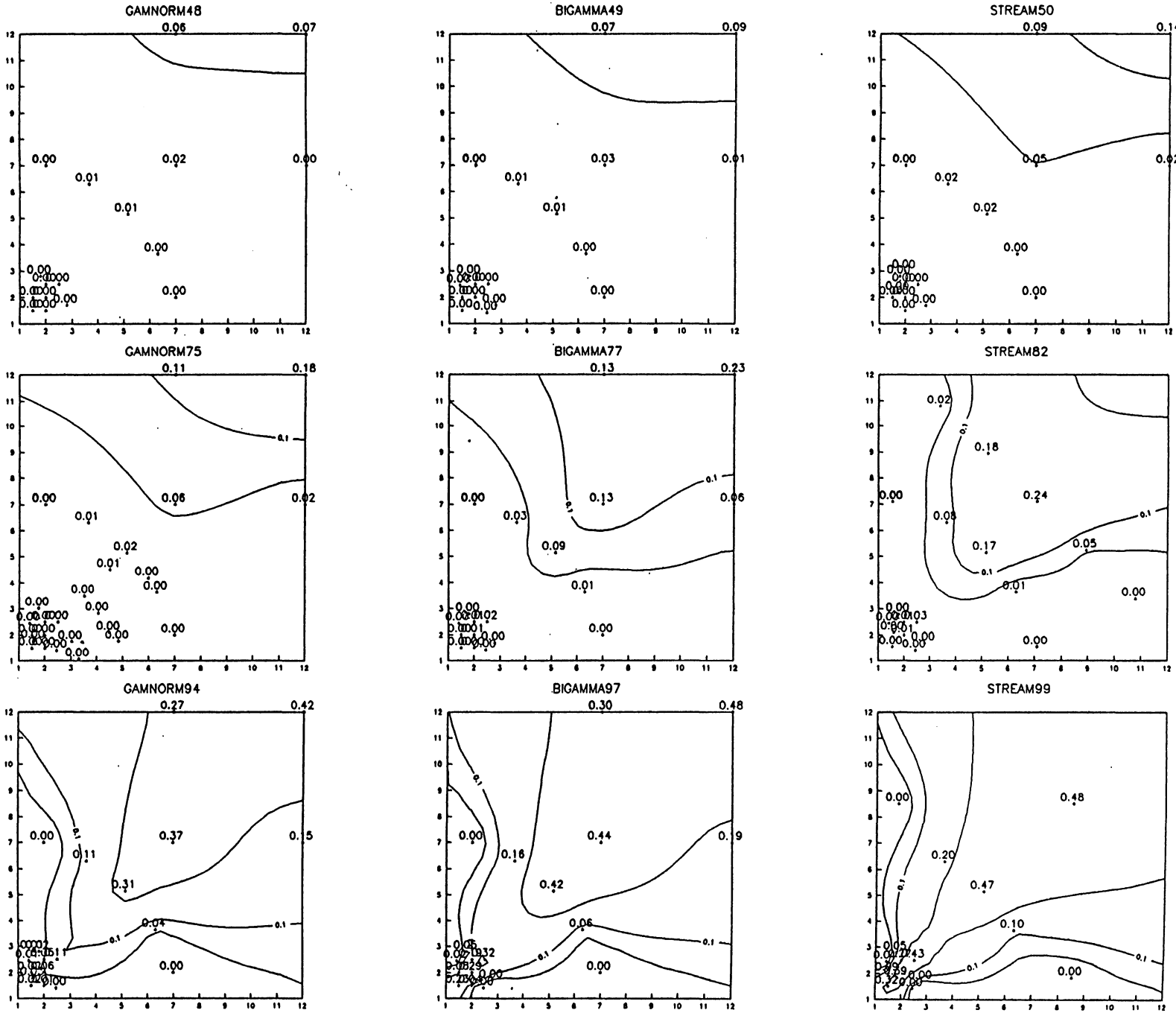


Figure 16. Relative Bias of $\hat{v}_{YG}^{sr}$.

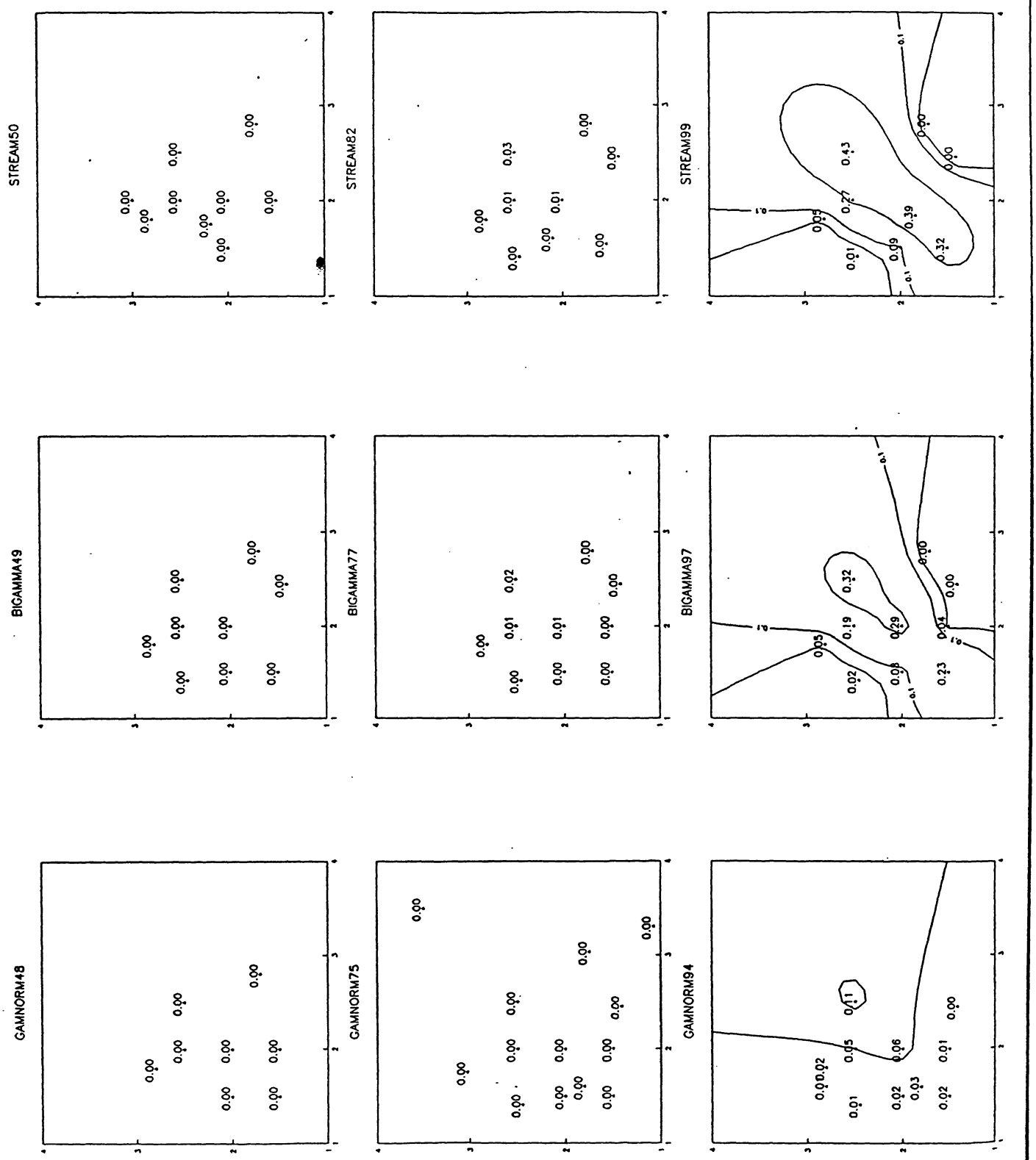
Figure 17. Probability of a Sample with Negative v_{HT}^{hr}.
a) Complete Population Space.
b) Enlargement of Lower Left Corner.

(Contours plotted: 0.05, 0.10, 0.20, 0.30)



a) Complete Population Space.

Figure 17 (Continued)
b) Enlargement of Lower Left Corner.



6. ANALYSIS AND DESIGN ISSUES

The exploration of the population space revealed some potentially useful survey design and analysis considerations for random-order, *vps* sampling. Given information about the correlation, population centroid, and distribution of x and y, the population space assessment provides guidance on the choice of a variance estimator for specified survey objectives. Example recommendations are:

1) if the population is such that variable probability sampling has better precision than simple random sampling, v_{HT}^{hr} should not be used;
2) none of the variance estimators work well if the population is located near the extreme left edge of the population space, but the random-order, *vps* design is inefficient in this circumstance and should be avoided (see later comments on shifting populations out of this region);
3) v_{HT}^{o} provides confidence intervals possessing good coverage for most populations, sometimes at the expense of positive bias and wider confidence intervals than those obtained using v_{YG}^{o} or v_{YG}^{hr};
4) v_{YG}^{o} is recommended over v_{YG}^{hr} since these two estimators have similar properties and v_{YG}^{o} is easier to compute.

Of importance to survey design, the population space analysis showed that shifting populations away from the

left edge of the population space resulted in improved properties of the variance estimators (except v_{HT}^{hr}) and improved efficiency of the estimator $\hat{T}_y$. A horizontal population shift is easily accomplished in the survey design by adding a constant to all population x's so that $x_i^* = x_i + c$, then sampling with inclusion probability proportional to x^*. The standardized variance plots (Figure 4) provide guidance for advantageous population space locations.

Shifting in the horizontal direction eliminates extremely small π's, but deciding how far to shift the population is a complication. Reddy and Rao (1977) considered modifying the x values at the analysis stage to improve precision of the estimator $\hat{T}_y$. Their theoretical results may provide some information on how far to shift the population at the design stage. If small π's detrimental to the precision of the estimators are not eliminated at the design stage, strategies "scoring" the small π's to a higher value can be employed to reduce MSE (Overton and Stehman, 1987; Potter, 1988).

Vertical shifts of a population to a more desirable region in the population space could also be considered to improve estimates after the sample data have been collected. Because the most drastic gradients in the population space surfaces were usually perpendicular to the horizontal axis, the advantage of a vertical shift in a

population appear minor relative to the potential gains of a horizontal shift.

7. CONCLUSIONS

The population space assessment proved successful in strengthening the conclusions available from empirical studies, and in discovering associations of behaviors of the variance estimators with characteristics of the populations. Previous empirical studies (Cumberland and Royall, 1981; Rao and Singh, 1973) did not reveal these patterns because they focused on a more restricted set of high correlation populations located near the standard diagonal. The standard diagonal was a region of special behavior, but more general conclusions were obtained in the population space analysis by systematically exploring a wide variety of structured populations.

Summarizing the important findings of the population space assessment:

1) Properties of v_{YG}^{o} and v_{YG}^{hr} were virtually identical, so the simpler form v_{YG}^{o} should be used in practice;
2) v_{HT}^{hr} performed the poorest of the four variance estimators, and this estimator should be avoided;
3) The worst behavior of v_{HT}^{hr} was in the region of the population space around the standard diagonal, precisely the region of populations

examined in past empirical studies -- past emphasis on these populations contributed to the perception that v_{YG} was superior to v_{HT};

4) The performance of v_{HT}^{o} was far superior to that of v_{HT}^{hr}, particularly for populations in the region of the standard diagonal;

5) The extreme left edge of the population space was a region of poor behavior for random-order, *vps* sampling.

Patterns in the behaviors of the variance estimators were consistent across all three families. Surfaces for the STREAM family were usually steeper, possibly because the sampling fraction was higher for this family. Although only samples of size 16 were investigated, the results observed in the population space assessment were consistent with results observed in previous empirical studies for other sample sizes and populations (cf. Stehman and Overton, 1987a; Rao and Singh, 1973).

The population space analysis is similar in philosophy to a superpopulation model concept because a model was used to generate the base populations for the BIGAMMA and GAMNORM families. The population space results were, therefore, representative of a broad class of populations. But as in any empirical study, the results were dependent on the particular realizations of the random variables generated in creating the BIGAMMA and GAMNORM families.

The behavior surfaces represented a single realization of these families, whereas, ideally, the mean trajectory or surface would be described. Another source of variability in the representation of the estimator properties was that the behavior surfaces were estimated by simulation; that is, the contour plots were not exact representations of the true surfaces and were subject to some sampling variability.

Theoretical comparison of the variance estimators in variable probability sampling has proven very difficult. The consistency of the variance estimator behaviors for the three families indicate these behaviors to be general, so a more general theory may be derivable, possibly even an analytic theory. Empirical identification of these patterns is an important step towards development of theoretical understanding.

ACKNOWLEDGEMENTS

Ron Stillinger provided invaluable help in implementing the computing and graphics described in this report. John Carlile assisted with the contour plotting routines.

8. REFERENCES

Brewer, K. R. W. (1963). A model of systematic sampling with unequal probabilities. Austral. J. Statist. 5, 5-13.

Cumberland, W. G., and Royall, R. M. (1981). Prediction models and unequal probability sampling. J. Roy. Statist. Soc. Ser. B 43, 353-367.

Hartley, H. O., and Rao, J. N. K. (1962). Sampling with unequal probability and without replacement. Ann. Math. Statist. 33, 350-374.

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47, 663-685.

Johnson, N. L., and Kotz, S. (1972). Distributions in Statistics: Continuous Multivariate Distributions. Wiley: New York.

Kennedy, W. J., and Gentle, J. E. (1980). Statistical Computing. Marcel Dekker: New York.

Messer, J.J., C.W. Ariss, J.R. Baker, S.K. Drousé, K.N. Eshleman, P.N. Kaufmann, R.A. Linthurst, J.M. Omernik, W.S. Overton, M.J. Sale, R.D. Shonbrod, S.M. Stanbaugh, and J.R. Tutshall, Jr. (1986). National Surface Water Survey: National Stream Survey, Phase I — Pilot Survey. EPA-600/4-86-026, U.S. Environmental Protection Agency, Washington, D.C.

Overton, W. S. (1985). A Sampling Plan for Streams in the National Stream Survey. Technical Report 114, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Overton, W. S., and Stehman, S. V. (1987). An Empirical Study of Sampling and Other Errors in the National Stream Survey; II. Analysis of a Replicated Sample of Streams. Technical Report 119, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Potter, F. (1988). Survey of procedures to control extreme sampling weights. To appear in Proceedings of the Section on Survey Research methods, American Statistical Association Annual Meetings, 1988.

Rao, J. N. K., and Singh, M. P. (1973). On the choice of estimator in survey sampling. Austral. J. Statist. 15, 95-104.

Reddy, V. N., and Rao, T. J. (1977). Modified PPS method of estimation. Sankhya Ser. C 39(3), 185-197.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. J. Indian Soc. Agric. Statist. 7, 119-127.

Stehman, S. V., and Overton, W. S. (1987a). Estimating the variance of the Horvitz-Thompson estimator in variable probability, systematic samples. Proceedings of the Section on Survey Research Methods, American Statistical Association Annual Meetings, 1987, pp. 743-748

Stehman, S. V., and Overton, W. S. (1987b). A comparison

of variance estimators of the Horvitz-Thompson estimator in random order, variable probability, systematic sampling. Biometrics Unit Manuscript Bu-M 935, Cornell University, 337 Warren Hall, Ithaca, New York, 14853.

Stehman, S. V., and Overton, W. S. (1989). Pairwise inclusion probability formulas in random-order, variable probability systematic sampling. Biometrics Unit Manuscript Bu-M 1008, Cornell University, 337 Warren Hall, Ithaca, New York, 14853.

Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. J. Roy. Statist. Soc. Ser. B 15, 235-261.