*Nordic Hydrology,* 18, 1987, 237-246



# Joint Probability Distribution of Streamflows and Tides in Estuaries

**G. V. Loganathan, C. Y. Kuo, and J. Yannaccone**
**Department of Civil Engineering**

Virginia Polytechnic Institute and State University
Blacksburg, VA 24061, U.S.A.

A procedure for estimating the joint probability of occurrence of correlated extreme tides and corresponding freshwater flows in estuaries is presented. The method uses the Box-Cox transformation to transform the original data to near normality, and therefore the search for a parent distribution is avoided. It is also shown that the traditional assumption of statistical independence for the jointly distributed random variables may lead to the underestimation of flows and tidal heights. The methodology is applied to the Rappahannock River in Virginia which flows into the Chesapeake Bay.

## Introduction

Numerical hydrodynamic models are frequently used to analyze the flow regimes in estuaries to study problems such as the flooding of low terrains and wetlands, the degradation of water quality due to waste disposals, the change of waterways related to navigation and the alteration of the salinity distribution in estuaries which plays a dominant role in trapping the needed nutrients for the aquatic life. Typical input boundary conditions to such models include certain extreme stream flows at the upstream of the reach under consideration and tidal heights at the downstream boundary (Wiley 1977; Kuo *et al.* 1978). The selection of these boundary conditions should be based on a reasonable joint probability of occurrence of extreme stream flows and tidal heights for a meaningful design. In this paper a methodology for the joint probability distribution of streamflows and tides is presented. Unlike the previous studies (Jenkins and Johnson 1978; Yeh 1980; Tayfun

and Machemehl 1980) this paper does not assume statistical independence between the concerned random variables. Furthermore, it is shown that the assumption of statistical independence can underestimate flow and tidal values and thus lead to an underdesign of facilities.

Another issue that must be addressed with regard to the estimation of joint probability of occurrence is the selection of the joint probability distribution function. It is conceivable that there can be no single probability distribution which would perform well for all estuaries. However if suitable transformations could be employed to transform data to fit a unique probability distribution function in the transformed range regardless of the distribution of the original sample, then uniformity can be established with regard to the treatment of the transformed samples. One such transformation is the Box-Cox transformation (Box and Cox 1964) which is utilized to transform the data to near normality. A maximum likelihood approach for selecting the optimal parameters is presented. The methodology is applied to the Rappahannock river in Virginia which flows into the Chesapeake Bay which is the largest estuary in the United States.

## Joint Normality

### Box-Cox Transformation

The high tidal height denoted as $H$ and the corresponding stream flow denoted as $Q$ are first transformed to near normal distribution through the Box-Cox transformation given as

$$A_i = \begin{cases} \dfrac{H_i^{\lambda_1} - 1}{\lambda_1} & \text{if} \quad \lambda_1 \neq 0 \\ \ln H_i & \text{if} \quad \lambda_1 = 0 \end{cases} \tag{1}$$

and

$$B_i = \begin{cases} \dfrac{Q_i^{\lambda_2} - 1}{\lambda_2} & \text{if} \quad \lambda_2 \neq 0 \\ \ln Q_i & \text{if} \quad \lambda_2 = 0 \end{cases} \tag{2}$$

where

$A_i$ – transformed tidal height
$B_i$ – transformed streamflow
$\lambda_1, \lambda_2$ – transformation parameters

### Maximum Likelihood Estimation

If one assumes that the transformed variates $A$ and $B$ have a joint normal distribution, then the density function may be written as

$$f(a,b) \equiv \frac{1}{2\pi |\sum|^{1/2}} \exp\{-\frac{1}{2}[(a-\mu_a),(b-\mu_b)]\sum^{-1}\begin{bmatrix}a-\mu_a\\ b-\mu_b\end{bmatrix}\} \qquad (3)$$

$$\text{for } -\infty < a < \infty \quad -\infty < b < \infty$$

where $$\sum = \begin{bmatrix}\text{Var}(A) & \text{Cov}(A,B)\\ \text{Cov}(A,B) & \text{Var}(B)\end{bmatrix} = E\{\begin{bmatrix}A-\mu_a\\ B-\mu_b\end{bmatrix}[(A-\mu_a),(B-\mu_b)]\}$$

$E(.)$ – Expected value
$|\Sigma|$ – determinant of $\Sigma$
$\mu_a$ $= E(A)$
$\mu_b$ $= E(B)$
$\text{Var}(A)$ $\equiv E[A - E(A)]^2$
$\text{Cov}(A,B)$ $= E[\{A - E(A)\}\{B - E(B)\}]$
$\Sigma^{-1}$ – inverse of $\Sigma$

In terms of the original variables $H$ and $Q$, the probability density function in Eq. (3) may be written as

$$f(h,q) = \frac{1}{2\pi |\sum|^{1/2}} \exp\{-\frac{1}{2}[(a-\mu_a),(b-\mu_b)]\sum^{-1}\begin{bmatrix}a-\mu_a\\ b-\mu_b\end{bmatrix}\} h^{(\lambda_1-1)} q^{(\lambda_2-1)}$$

$$\text{for } h > 0 \text{ and } q > 0 \qquad (4)$$

For $n$ pairs observed sample of $h_i$ and $q_i$ the value of the joint density function (likelihood function for the parameters) may be written as

$$f(\vec{h},\vec{q};\mu_a,\mu_b,\sum,\lambda_1,\lambda_2) = (2\pi)^{-n}(|\sum|)^{-n/2}\prod_{i=1}^{n} h_i^{(\lambda_1-1)} q_i^{(\lambda_2-1)}$$

$$\exp\{-\frac{1}{2}\sum_{i=1}^{n}[(a_i-\mu_a),(b_i-\mu_b)]\sum^{-1}\begin{bmatrix}a_i-\mu_a\\ b_i-\mu_b\end{bmatrix}\} \qquad (5)$$

For a fixed pair of $\lambda_1$ and $\lambda_2$, the maximum likelihood estimates of $\mu_a$, $\mu_b$ and $\Sigma$ are given by (Morrison 1976)

$$\hat{\mu}_a = \bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i \qquad (6)$$

$$\hat{\mu}_b = \bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i \qquad (7)$$

and

$$\hat{\Sigma} = S(\lambda_1, \lambda_2) = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \tag{8}$$

where

$$S_{11} \equiv \frac{1}{n} \sum_{i=1}^{n} (a_i - \bar{a})^2$$

$$S_{22} = \frac{1}{n} \sum_{i=1}^{n} (b_i - \bar{b})^2$$

$$S_{12} = S_{21} \equiv \frac{1}{n} \sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b})$$

$n \equiv$ sample size

$a_i, b_i =$ transformed values

From Eq. (5), after substituting the estimates given by Eqs. (6), (7) and (8) the log likelihood function except for a constant term may be written as

$$L(\lambda_1, \lambda_2) = -\frac{n}{2} \ln |S(\lambda_1, \lambda_2)| + (\lambda_1 - 1) \sum_{i=1}^{n} \ln h_i + (\lambda_2 - 1) \sum_{i=1}^{n} \ln q_i \tag{9}$$

Which is to be maximized over $\lambda_1$ and $\lambda_2$ for proper transformation parameters in Eqs. (1) and (2).

**Goodness of Fit Test**

The following procedures apply for testing the normality of the transformed sample. Let $\vec{x} = (x_1, x_2, \ldots, x_p)$ be normally distributed with mean $\vec{\mu} = E(\vec{x})$ and variance-covariance matrix $\Sigma$, and $|\Sigma| > 0$. Then $(\vec{x} - \vec{\mu})\, \Sigma^{-1} (\vec{x} - \vec{\mu})'$ has a chi-square distribution with $p$ degrees of freedom (Johnson and Wichern 1982). By using the estimates of $\vec{\mu}$ and $\Sigma$ we can define a squared generalized distance for the transformed data as

$$d_i^2 = (a_i - \bar{a}, b_i - \bar{b}) \left[ \frac{n}{n-1} S \right]^{-1} \begin{bmatrix} a_i - \bar{a} \\ b_i - \bar{b} \end{bmatrix} \tag{10}$$

in which the factor $n/(n-1)$ is included for the unbiased estimation of $\Sigma$. Because $d_i^2$ are expected to behave like a chi-square variate, a plot of chi-square quantiles against $d_i^2$ can be made which should be a straight line for the bivariate normality assumption to hold. Based on Johnson and Wichern (1982) the chi-square plot is constructed as follows:

(i) – rank order the squared generalized distances from the smallest to the largest $d^2(1) \leq \ldots \leq d^2(i) \leq \ldots \leq d^2(n)$.

(ii) – For each rank $i$, compute $p_i = (i - 0.3)/(n + 0.4)$.

(iii) – For each $p_i$, find the chi-square cutoff value (degrees of freedom 2), $\chi^2(i)$ such that

$$p_i \equiv P[G \leq \chi^2(i)] \tag{11}$$

where $G$ is a chi-square random variable.

(iv) – Plot $\chi^2(i)$ against $d^2(i)$

For multivariate normal distributions, it is true that any arbitrary subset of components has a normal distribution and therefore the marginals are also normal. The $\lambda$'s obtained from Eq. (9) are used with the appropriate data sets and the transformed sample quantiles are plotted against the theoretical normal quantiles (*Q-Q* plot test, Johnson and Wishern 1982; Guttman *et al.* 1982; David 1981). If the plot is a straight line, then normality assumption is valid. The sample quantiles are estimated by arranging the transformed sample data in ascending order and using the empirical distribution function

$$P[C \leq c_i] = \frac{i-0.3}{n+0.4} = p_i \tag{12}$$

where

$c_i$ – $i_{\text{th}}$ ranked observation (rank 1 smallest)
$C$ – normally distributed random variable
$n$ – sample size
$P_i$ – Probability that the random variable $C$ is less than or equal to $c_i$.

The theoretical quantiles $d_i$ for the normal distribution for the probability $p_i$ in Eq. (12) are obtained from

$$p[C \leq d_i] \equiv \int_{-\infty}^{d_i} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}\, ds = p_i \tag{13}$$

The sample quantiles $c_i$ are plotted against the theoretical normal quantiles $d_i$ to verify the straightness of the plot. The skew coefficient is checked for zero which indicates a symmetric distribution. For the normal distribution the kurtosis coefficient must be three.

**Exceedence Probability**

By using the joint probability density function given in Eq. (3) the exceedence probability or risk level $P[A > a, B > b]$ can be computed. It is also noted that for a chosen exceedence probability level, the choice of $a$ and $b$ is nonunique. But, if $a$ is fixed as $a_o$, then there is a unique $b_0$ for a fixed probability level. Of course, these $a_0$ and $b_0$ values must be inverted to obtain the original variables namely the flow, $q_0$, and the tidal height, $h_0$ using the inverse transformations for the transformations in Eqs. (1) and (2) which are given by

$$h_0 = (\lambda_1 a_0 + 1)^{1/\lambda_1} \tag{14}$$

$$q_0 = (\lambda_2 b_0 + 1)^{1/\lambda_2} \tag{15}$$

It is suggested that several $(a_0, b_0)$ combinations be selected for a chosen design risk level (see Figs. 5 and 6) and be inputted into the hydrodynamic model to study the estuarine response for these boundary conditions. One of these combinations will result in a critical condition, say in terms of the salinity distribution and that critical response should be used in the design (Yannaccone 1987).

## Application

### Statistical Independence

The methodology has been applied to the Rappahannock river which drains into the Chesapeake Bay. The tidal information is obtained from the National Oceanic and Atmospheric Administration (NOAA) and the stream flow records are available from the U.S. Geological Survey (USGS) data base. For this study monthly high tide and the corresponding flow data are utilized with 628 paired tide-streamflow data points. The joint probability of occurrence

$$p(h,q) \equiv P(H > h,\ Q > q) \tag{16}$$

is estimated by pairing the tides and corresponding streamflows. For statistical independence by definition

$$p(h,q) = P(H > h)\ P(Q > q) \tag{17}$$

The right hand side expressions of Eqs. (16) and (17) are evaluated and plotted in Fig. 1. As one might expect high flows and high tides are strongly correlated and typically used in design as extreme events. Also observe that, at least in this particular case, Eq. (16) yields a higher flow value than Eq. (17) for the same probability level for a fixed tidal height which suggests that the statistical independence assumption may not lead to a conservative design.

### Maximum Likelihood Estimates of $\lambda_1$ and $\lambda_2$

Based on the fact that the marginals of jointly normally distributed random variables are also normal, one could expect the $\lambda$'s (of the Box-Cox transformation) estimated from the marginal distributions not to be widely different from those of the joint distribution. The likelihood functions based on the marginal normal densities for tides and stream flows are given respectively as

$$L(\lambda_1) \equiv -\frac{n}{2} \ln \frac{1}{n} \sum_{i=1}^{n} (a_i - a)^2 + (\lambda_1 - 1) \sum_{i=1}^{n} \ln h_i \tag{18}$$

and

$$L(\lambda_2) \equiv -\frac{n}{2} \ln \frac{1}{n} \sum_{i=1}^{n} (b_i - b)^2 + (\lambda_2 - 1) \sum_{i=1}^{n} \ln q_i \tag{19}$$

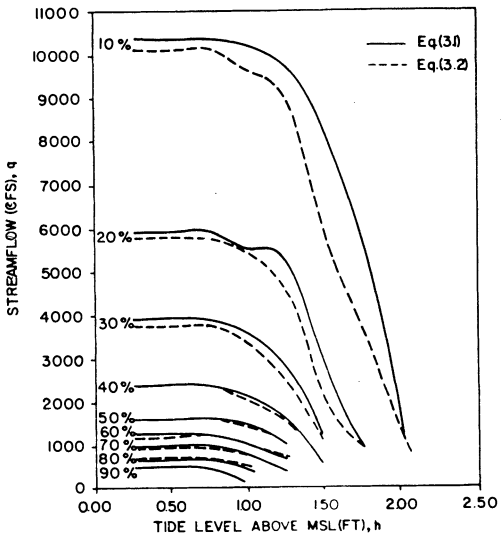


Fig. 1. Joint probability of exceedence from the sample.

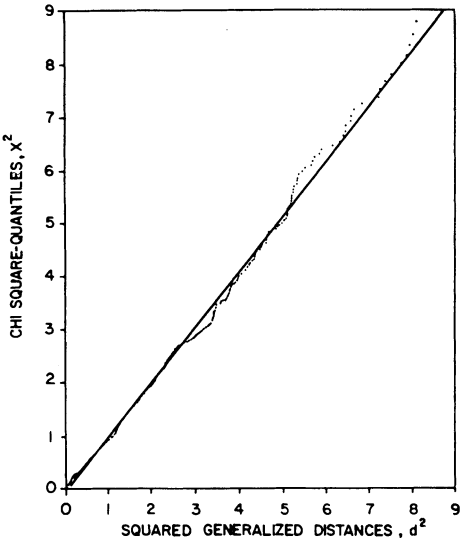


Fig. 2. Chi-Square plot for bivariate normality test.

Optimizing expressions (18) and (19) is much easier than expression (9). Based on the marginal densities of $H$ and $Q$ for the Rappahannock river Eqs. (18) and (19) yield $\lambda_1 = 0.10$ and $\lambda_2 = 0.0$. Optimizing expression (9) yields $\lambda_1 = 0.13$ and $\lambda_2 = 0.03$ and these $\lambda$'s are used as the maximum likelihood estimates.

### Goodness of Fit Tests

By using $\lambda_1 \equiv 0.13$ and $\lambda_2 = 0.03$, the tidal and the stream flow data are transformed to near normality. The Chi-square plot for the bivariate normality test is shown in Fig. 2. The straight line fit shows that the transformed data fit the bivariate normal distribution quite well. Also, the univariate Q-Q plots are given in Figs. 3 and 4. The straight line fit validates the normality assumption. The relevant statistics for the transformed data are given in Table 1. From Table 1 it is seen that the skew coefficients are close to zero and the coefiicients of kurtosis are close to three. The Chi-squared and normal quantiles are obtained by using the subroutines MDCHI, and MDNRIS respectively, from the International Mathematical and Statistical Libraries (IMSL).

### Exceedence Probability

To compute the exceedence probability the following relationship is used.

$$P[A \geq a, B \geq b] \equiv 1 - P(A \leq a) - P(B \leq b) + P(A \leq a, B \leq b) \qquad (20)$$

The reason for using Eq. (20) is that all the probabilities on the right hand side of Eq. (20) can be computed directly using the subroutines from the IMSL. Specifical-

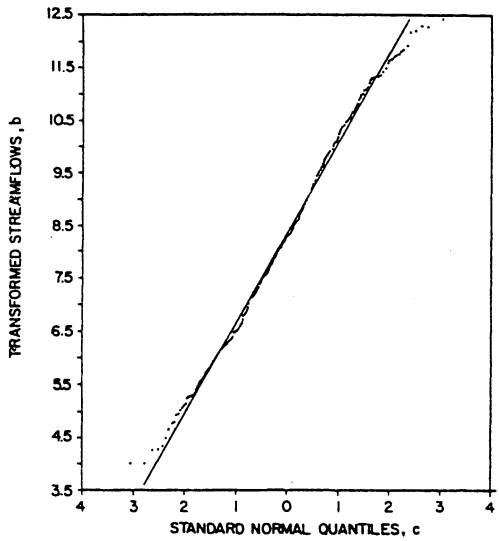


Fig. 3. $Q - Q$ Plot for streamflows.

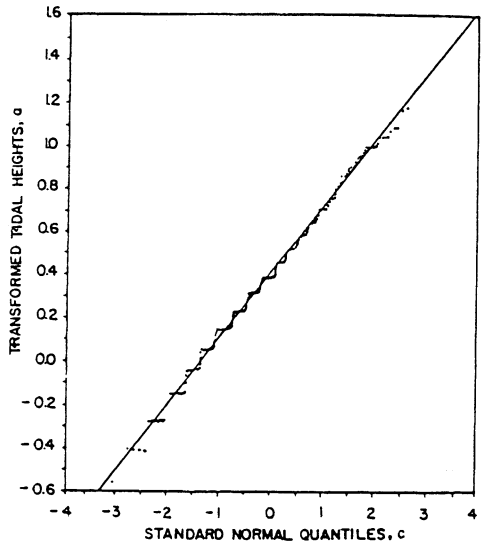


Fig. 4. $Q - Q$ Plot for tides.

ly the subroutines MDNOR and MDBNOR are used respectively for the univariate and bivariate normal probabilities. In general the bivariate normal probabilities are hard to compute by hand but suitable tables and examples can be found in Abramowitz and Stegun (1972). The exceedence probabilities along with the tidal and stream flow values are given in Figs. 5 and 6 which are calculated by the inverse transformations given by Eqs. (14) and (15). It is noted that the results of Fig. 1 are obtained from Eqs. (16) and (17) using the sample (based on the frequency interpretation of probability) and no theoretical distribution is used whereas Figs. 5 and 6 show the results obtained from a theoretical joint normal distribution. It is seen from Figs. 5 and 6 that for the same joint probability of occurrence various

Table 1 – Statistical quantities for the Rappahannock River

| | Streamflow ($\lambda_2 \equiv 0.03$) | Tidal Height ($\lambda_1 \equiv 0.13$) |
|---|---|---|
| Mean | 8.32 | 0.412 |
| Variance | 2.88 | 0.0918 |
| Standard Deviation | 1.70 | 0.303 |
| Skew | 0.230 | 0.000788 |
| Skew Coefficient | 0.0472 | 0.0283 |
| Coefficient of Kurtosis | 2.4362 | 3.1587 |
| Correlation coefficient*, r = 0.185 | | |

*The correlation coefficient, r, is computed by $$r = \frac{\sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b})}{\left[\sum_{i=1}^{n} (a_i - \bar{a})^2 (b_i - \bar{b})^2\right]^{0.5}}$$

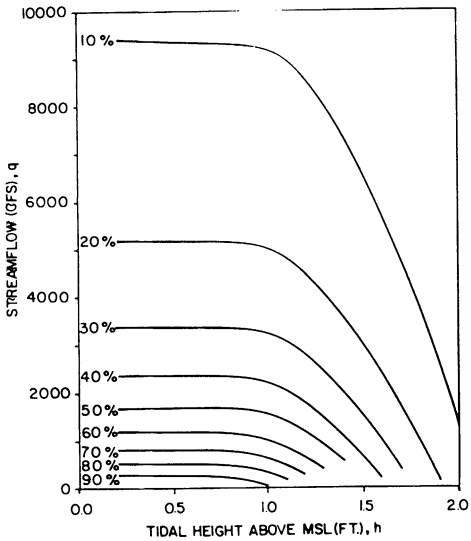


Fig. 5. Theoretical joint probability of exceedence (10%-90%).

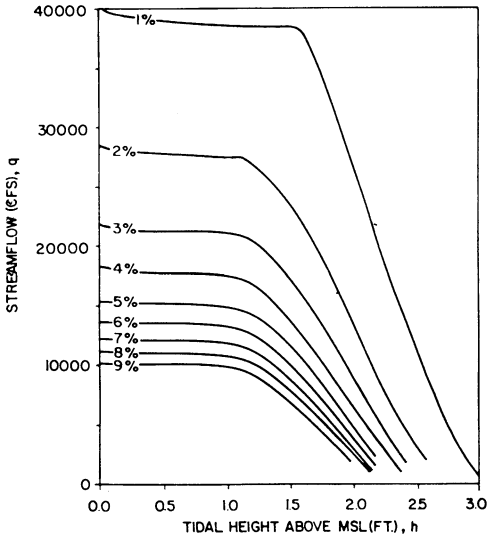


Fig. 6. Theoretical joint probability of exceedence (1%-9%).

combinations of stream flows and tides are possible. It must be noted that these combinations in general will not lead to the same results, say the salinity distribution, and the sediment movement. Therefore, it is recommended that for the chosen design probability of exceedence (risk level), several combinations be selected and used in a hydrodynamic simulation model to pick the most critical condition for the final design (Yannaccone 1987). Also, a conditional probability approach can be used. It can be shown that the conditional distribution of a random variable say, $A$ from a bivariate normal distribution of $A$ and $B$ is also normal with mean and variance given by

$$E(A \mid B) \equiv b) \equiv E(B) + \frac{\mathrm{Cov}(A,B)}{\mathrm{Var}(B)} [b - E(B)] \tag{21}$$

$$\mathrm{Var}(A \mid B = b) = (1-\rho^2)\ \mathrm{Var}(A) \tag{22}$$

where $$\rho = \frac{\mathrm{Cov}(A,B)}{[\mathrm{Var}(A)\ \mathrm{Var}(B)]^{0.5}}$$

Therefore, for any fixed critical streamflow, a range of high tides can be found at different conditional probability levels using the parameters given in Eqs. (21) and (22). Both the joint probability approach and the conditional probability approach provide for the meaningful selection of boundary conditions in hydrodynamic numerical models, in the sense that such selections can be interpreted probabilistically which would help to screen out uneconomical events with very low probability of exceedence.

## Summary

It is shown that high flows and high tides are in general correlated (see Fig. 1) and the assumption of statistical independence may not be valid. This consideration is quite important in the sense that these high flows and high tides constitute the design events. Moreover it is shown that the assumption of statistical independence to estimate the joint probability of simultaneous occurrence of flows and tides can lead to the underestimation of flow and tidal values and thus may not result in a conservative design. The joint normality approach overcomes this deficiency. Because a transformation to normality is used, the search for a parent distribution of the original data is not needed. It is also shown that the Box-Cox transformation yields satisfactory results. Instead of arbitrarily choosing the extreme stream flow and tidal values, the present methodology can be used to objectively select these extreme combinations with specified probabilities of exceedence or risk level.

## References

Abramowitz, M., and Stegun, I.A. (1972) *Handbook of Mathematical Functions,* New York, Dover Publications Inc.

Box, G.E.P., and Cox, D.R. (1964) An Analysis of Transformations, *Journal of the Royal Statistical Society, Vol. B 26,* pp. 211-252.

David. H.A. (1981) *Order Statistics,* New York, Wiley.

Guttman, I., Wilks, S.S., and Hunter, J.S. (1982) *Introductory Engineering Statistics,* New York, Wiley.

Jenkins, J.D., and Johnson, H.M. (1978) Flood Profiles in Combined Tidal-Freshwater Zones, *Journal of the Hydraulics Division, ASCE, Vol. 104*, No. HY6, pp. 919-922.

Johnson, R.A., and Wichern, D.W. (1982) *Applied Multivariate Statistical Analysis*, Englewood Cliffs, N.J., Prentice Hall.

Kuo, A.Y., Nichols, M., and Lewis, J. (1978) Modeling Sediment Movement in the Turbidity maximum of an Estuary, Bulletin 111, Virginia Water Resources Research Center, Blacksburg, VA.

Morrison, D.F. (1976) *Multivariate Statistical Methods,* New York, McGraw Hill.

Tayfun, M.A., and Mechemehl, J.L. (1980) Analysis of Combined Rainfall-Hurricane Surge Frequencies, in *Urban Stormwater Mangement in Coastal Areas* (Ed: C.Y. Kuo), New York, ASCE.

Wiley, M. (1977) *Estuarine Processes,* New York, Academic Press.

Yannaccone, J. (1987) Numerical Simulation of the Effects of Sea Level Rise on Estuarine Processes, M.S. Thesis, Virginia Tech, Blacksburg, VA.

Yeh. F.F. (1980) A Note on Joint Probability of Surge and Rainfall, in *Urban Stormwater Management in Coastal Areas* (Ed: C.Y. Kuo), New York, ASCE.

**Address:** Department of Civil Engineering,
Room 200 Patton Hall,
Virginia Polytechnic Institute and State University,
Blacksburg, VA 24061, U.S.A.

Received: 7 July, 1987