

# A Bayesian Framework for the Use of Regional Information in Hydrology

GUILLERMO J. VICENS,<sup>1</sup> IGNACIO RODRIGUEZ-ITURBE, AND JOHN C. SCHAAKE, JR.<sup>2</sup>

*Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*

Water resource designs are perfect examples of decision making under uncertainty. In fact, three types of uncertainties may exist in any design problem: natural, parameter, and model uncertainties. The last two may be considered as informational uncertainties that are due to the lack of perfect information about the streamflow processes. The use of regional information has been suggested as a technique for reducing parameter uncertainties. The use of Bayesian methodology provides a framework for combining regional information with at-site historical records. Moreover, Bayesian methods allow the hydrologist to consider the parameter uncertainties as well as the natural uncertainties within the decision-making process. Because of these two advantages the Bayesian approach is a more complete and realistic approach to problems of uncertainty in hydrology and water resource planning than presently used methodologies.

## UNCERTAINTY IN WATER RESOURCE PLANNING

Over the last two decades, extensive research efforts have produced techniques that deal explicitly with the problem of the uncertainties present in the design and planning of water resource projects. These very successful efforts have mainly focused on one aspect of the whole range of uncertainties present in hydrologic problems.

These uncertainties may be classified as being of two types, natural (or inherent) and informational. Streamflow processes are frequently considered or assumed to be stochastic processes because of the natural, or inherent, randomness apparent in the observed streamflow traces. Owing to the lack of perfect information about the streamflow process, e.g., infinitely long historical records, there exists an informational uncertainty about the process. This uncertainty may be divided into parameter and model uncertainty. There is seldom enough information available to evaluate the parameters of the model or select the 'correct' model with certainty.

Stochastic hydrology has focused on the analysis of the natural uncertainties in water resource problems and has generally ignored the informational uncertainties. Many techniques for the generation of synthetic streamflows have been proposed, but no account of the parameter uncertainty has explicitly been carried out. Moreover, no model selection procedure has been proposed, although this problem is believed to be less important than the parameter uncertainty problem for time series models that do not explicitly focus on extreme events.

When parameter uncertainties have been considered, it has been through point estimation procedures. Attempts have been made to reduce this uncertainty by using the available historical record or regional data. A combination of these two sources of information has rarely been attempted, and no attempt has been made to combine the two sources and to include the remaining parameter uncertainty in a decision problem framework simultaneously. This type of framework would assess the effects of the parameter uncertainty and the value of obtaining additional information.

Synthetically generated records are used to assess the effect

of streamflow variations on proposed designs; this procedure requires the estimation of streamflow parameters such as the mean and variance from the historical record. The uncertainties on these and other parameters have not been included in the synthetic generation procedures.

The U.S. Geological Survey has carried out studies directed at the estimation of streamflow parameters from physiographic and meteorologic characteristics of a basin [Benson and Matalas, 1967; Thomas and Benson, 1970]. Matalas and Gilroy [1968] proposed that these estimates be compared with those from the historical record and that the lower variance estimators be used and the other set be discarded.

Bayesian methodology has been used by Shane and Gaver [1970], Tschannerl [1971], Davis et al. [1972], Lenton et al. [1973], and Wood et al. [1974]. Of these, Shane and Gaver [1970] were the first to propose the use of regional information from regression models and historical at-site information. The objective in that work was to combine estimators from these two sources.

This paper focuses on one objective: to investigate the use of regional information in conjunction with the at-site historical record to reduce the parameter uncertainties. A separate paper will present procedures that explicitly account for the parameter uncertainties in planning for water resource projects. The details of this work can be found in the work of Vicens et al. [1974]. As is described in detail in later sections, Bayesian procedures appear to be perfectly suited to meet our objectives. These procedures allow the hydrologist to include explicitly the parameter uncertainties in the decision process and to combine sources of information to reduce the parameter uncertainties.

## BAYESIAN INFERENCE AND DECISION

Streamflow processes will be considered as random processes generating random variables distributed according to a model probability distribution function (pdf). The inherent, or natural, randomness of the process creates uncertainty about future observation of the process and about the consequences or value of any decision. In addition, the parameters of the model are unknown, a complication that further adds to the uncertainties about future observations. However, we shall assume that the model pdf is known with certainty; i.e., no model uncertainty exists. Bayes's theorem can be used to combine the prior and sample information to

<sup>1</sup>Now at Resource Analysis, Inc., Cambridge, Massachusetts 02138.

<sup>2</sup>Now at Hydrology Research Laboratory, National Weather Service, Silver Spring, Maryland 20910.

update the present information about the unknown parameters in the following manner:

$$f''(\theta|I_R, Y) \propto f'(\theta|I_R) \cdot L(\theta|Y) \quad (1)$$

i.e., the posterior pdf of the unknown parameters  $\theta$ .  $f''(\theta|I_R, Y)$  is proportional to the product of the prior pdf  $f'(\theta|I_R)$ , which contains the prior information  $I_R$ , and the likelihood function  $L(\theta|Y)$ , which contains the sample information  $Y$ . Equation (1) is stated as being 'proportional to,' since the likelihood function is not a proper pdf, and therefore a normalizing constant is required to make this relation 'equal to.' Where this constant is important, it can be obtained from

$$C^{-1} = \int_{\Theta} f'(\theta|I_R) \cdot L(\theta|Y) d\theta \quad (2)$$

The use of Bayes's theorem to 'pool' sources of information is shown in Figure 1. The sample information, for example, the historical record at a gaging station, is introduced through the likelihood function. Other information is brought into the analysis through the prior pdf. The result is a posterior pdf that contains all of the available information.

If additional information becomes available at a later time, the present posterior pdf will become the new prior pdf, and the additional information will be included in the process through the new likelihood function. In fact, this process of combining information is quite general and may be used to include many sources of information.

*Prior pdf's.* The prior pdf  $f'(\theta|I_R)$  represents all of the available prior information about the parameters of the model  $\theta$ . The time sequence, whether or not this information was obtained prior to the sample, is irrelevant. It is important though that this information be separate and distinct from the historical sample. Formally, it must be statistically independent information.

When the prior pdf is derived from an 'objective' analysis of available data (e.g., regression on regional type data), it may be classified as data-based (DB). But when it is derived from subjective judgments, casual observation, or theoretical considerations, it is classified as non-data-based (NDB). Many situations present both types of information.

Very few arguments will arise from the use of DB priors when the analysis is carried out through means of a well-

accepted technique such as regression analysis. On the other hand, it can be nearly impossible to find two hydrologists who will produce identical NDB priors, since these will reflect their personal training, experience, or bias. Consequently, posterior inferences or decisions based on their different NDB priors will be different.

A special class of NDB priors comprises the noninformative priors. These are used to express 'ignorance' about the parameters. When noninformative prior pdf's are used, the posterior only reflects the information in the sample. The representation of ignorance in mathematical terms is a subject of controversy in the statistics literature. This topic will not be covered in detail in this work, since it is irrelevant in practical applications. The practicing hydrologist will always have some information on which to base judgments on the prior pdf's, whether they be DB or NDB.

For many inference problems, only one parameter, i.e., a subset  $\theta_1$  of  $\theta$ , is of interest. The marginal pdf for this subset can be obtained from the joint pdf by integrating over the remaining parameters in  $\theta$ , which are then called 'nuisance parameters.'

*Bayesian probability distribution.* The consequences of inferences and decisions in hydrology are more frequently directly related to future streamflows than to the parameters of the process. For example, although a reservoir design will be affected by the mean and standard deviation of the streamflow inputs, the consequences of a design will be directly related to the particular sequence of inputs that occur during the economic time horizon of the project. Therefore we are particularly interested in the probability distribution of these future streamflows.

The Bayesian pdf of a future observation of the streamflow process is obtained by integrating the product of the model pdf, which is a conditional pdf given the parameters, and the joint posterior pdf of the parameters [Zellner, 1971], i.e.,

$$\tilde{f}(y_f|I_R, Y) = \int_{\Theta} f(y_f|\theta) \cdot f''(\theta|I_R, Y) d\theta \quad (3)$$

where  $y_f$  is one future observation of the streamflow and  $f(y_f|\theta)$  is the model pdf. The resulting pdf includes the natural, or inherent, randomness of the streamflow process and the uncertainty about the true value of the model parameters.

Equation (3) can be used for making inferences about future streamflows. For example, the probability of a flood over a certain flow  $q$  can be obtained by

$$P\{y_f \geq q\} = \int_q^{\infty} \tilde{f}(y_f|I_R, Y) dy_f \quad (4)$$

Similarly, the Bayesian moments of a future observation can be obtained by

$$E\{y_f\} = \int_{y_f} y_f \cdot \tilde{f}(y_f|I_R, Y) dy_f \quad (5)$$

$$V\{y_f\} = \int_{y_f} (y_f - E\{y_f\})^2 \cdot \tilde{f}(y_f|I_R, Y) dy_f \quad (6)$$

The next section will present one of the simplest models of hydrologic time series, the independent normal process. The lack of a correlation structure precludes the use of this model for many hydrologic problems. It is nevertheless a useful one to represent historical series, and its study is both necessary for the understanding of more complex models and extremely

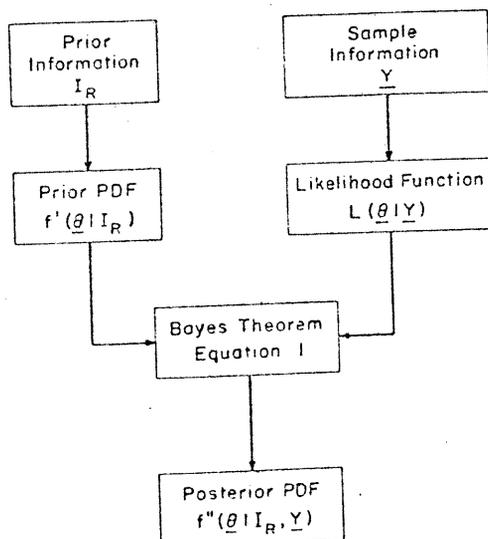


Fig. 1. The Bayes theorem: processing of information.

useful and simple for the demonstration of the values of prior and historical information. Although this paper will focus on this model only, the approach is not limited to this simple model. In fact, the same methodology has been applied to three other models by *Vicens et al.* [1974]. These models are the independent log normal process, the first-order normal autoregressive process, and the first-order log normal autoregressive process. Research currently in progress is aimed at the analysis of more complex models.

#### INDEPENDENT NORMAL PROCESS

An independent normal process is defined as the process generating annual streamflows (or other hydrologic variables)  $y_i$ , assumed to be independent and identically distributed random variables with a normal (N) pdf (model pdf) given by

$$f_N(y_i|\mu, \sigma) = (2\pi)^{-1/2}\sigma^{-1}\{\exp[-(y_i - \mu)^2/2\sigma^2]\} \quad (7)$$

If  $n$  observations for this process have been made, the likelihood function for that particular sample is the product of the individual pdf's:

$$L(\mu, \sigma|Y) = (2\pi)^{-1/2n}\sigma^{-n}\left\{\exp\left[-\sum_{i=1}^n (y_i - \mu)^2/2\sigma^2\right]\right\} \quad (8)$$

where  $Y$  denotes the set of observations  $[y_1, y_2, \dots, y_n]$ . Thus  $L(\mu, \sigma|Y)$  is proportional to the pdf that a particular sequence of  $y_i$  would be observed for specific values of  $\mu$  and  $\sigma$ . Equation (8) may be simplified by defining the sufficient statistics for the sample:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (9)$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad n > 1 \quad (10)$$

$$v = n - 1 \quad n > 0 \quad (11)$$

These three statistics, in addition to  $n$ , the size of the sample, are the sufficient statistics for the sample. The likelihood function can now be written as

$$L(\mu, \sigma|Y) = (2\pi)^{-1/2n}\sigma^{-n} \cdot \{\exp[-(vs^2 + n(\bar{y} - \mu)^2)/2\sigma^2]\} \quad (12)$$

A kernel of this likelihood function, i.e., those factors that vary with the unknown variables  $\mu$  and  $\sigma$ , is

$$k(\mu, \sigma) = \sigma^{-n}\{\exp[-(vs^2 + n(\bar{y} - \mu)^2)/2\sigma^2]\} \quad (13)$$

As an example the sufficient statistics for the Pemigewasset River at Plymouth, New Hampshire (U.S. Geological Survey station 1-765), were computed for three different record lengths ( $n = 5, 20, 60$ ). These are shown in Table 1. This information will be combined with the prior information to be derived in the following sections.

TABLE 1. Sample Information: Pemigewasset River, Plymouth, New Hampshire

Record Length	$\bar{y}$ , ft <sup>3</sup> /s	$s^2$	$n$ , yr	$v$ , yr
1972-1976	1384	64618	5	4
1972-1991	1567.9	41244	20	19
1972-1997	1546.8	66900	60	59

A convenient prior pdf for  $\mu$  and  $\sigma$  is the natural conjugate of (12), a normal inverted gamma 2 pdf. It is convenient in the sense that when it is combined with the likelihood function (12), it yields an analytically tractable posterior pdf. This prior pdf is a product of a normal and an inverted gamma 2 (IG2) pdf [*Vicens et al.*, 1974], i.e.,

$$f'(\mu, \sigma|I_R) = f_N(\mu|\bar{y}', \sigma/n^{1/2}) \cdot f_{IG2}(\sigma|s', v') \quad (14)$$

where the parameters  $\bar{y}'$ ,  $s'$ ,  $n'$ , and  $v'$  are estimated from prior information  $I_R$ , as will be shown later in this paper.

The marginal pdf for either of the two unknown variables  $\mu$  and  $\sigma$  can be obtained by 'integrating out' the other variable. Thus the distribution of  $\mu$  is

$$f'(\mu|I_R) = \int_0^\infty f'(\mu, \sigma|I_R) d\sigma = f_{S'}\left(\mu|\bar{y}', \frac{s'^2}{n'}, v'\right) \quad (15)$$

a Student  $t$  pdf with expected value and variance

$$E[\mu|I_R] = \bar{y}' \quad v' > 1 \quad (16)$$

$$V[\mu|I_R] = v's'^2/n'(v' - 2) \quad v' > 2 \quad (17)$$

and the distribution of  $\sigma$  is

$$f'(\sigma|I_R) = \int_{-\infty}^{\infty} f'(\mu, \sigma|I_R) d\mu = f_{IG2}(\sigma|s', v') \quad (18)$$

an inverted gamma 2 pdf with moments

$$E[\sigma|I_R] = \left(\frac{v'}{2}\right)^{1/2} s' \frac{\Gamma[(v' - 1)/2]}{\Gamma[v'/2]} \quad v' > 1 \quad (19)$$

$$V[\sigma|I_R] = \frac{v's'^2}{v' - 2} - \{E[\sigma|I_R]\}^2 \quad v' > 2 \quad (20)$$

where  $\Gamma(\cdot)$  is the gamma function, defined by

$$\Gamma(\tau) = \int_0^\infty e^{-p} p^{\tau-1} dp \quad (21)$$

From (18) one gets the marginal pdf for the process variance (through the theory of derived distributions):

$$f'(\sigma^2|I_R) = f_{IG2}(\sigma^2|I_R) \cdot |1/2\sigma| = f_{IG1}(\sigma^2|1/2v', 1/2v's'^2) \quad (22)$$

which is an inverted gamma pdf with moments

$$E[\sigma^2|I_R] = v's'^2/(v' - 2) \quad v' > 2 \quad (23)$$

$$V[\sigma^2|I_R] = 2(v's'^2)^2/(v' - 2)^2(v' - 4) \quad v' > 4 \quad (24)$$

The analysis of this model may be carried out in terms of the mean and standard deviation ( $\mu, \sigma$ ), the mean and variance ( $\mu, \sigma^2$ ), or the mean and precision ( $\mu, h$ ), where  $h = 1/\sigma^2$  [*Raiffa and Schlaifer*, 1961; *Kaufman*, 1972]. Information about  $\sigma, \sigma^2$ , and  $h$  can be transferred from one to the other through the use of derived distributions. This fact will be used in defining prior pdf's.

Since the functional form of the prior pdf was selected to be from the conjugate family of the likelihood function for this model, all that remains to be determined is the values of the four parameters of the prior  $\bar{y}'$ ,  $s'^2$ ,  $n'$ , and  $v'$ . Once values for these parameters have been set, a unique pdf for  $\mu$  and  $\sigma$  has been determined. Obtaining these values from the joint pdf is a very difficult task. Therefore approaching the problem from the marginal pdf's appears more promising. Marginal pdf's and moments were derived for  $\mu, \sigma$ , and  $\sigma^2$ . If specific values

are given for  $E[\mu|I_R]$ ,  $V[\mu|I_R]$ ,  $E[\sigma^2|I_R]$ , and  $V[\sigma^2|I_R]$ . (16), (17), (23), and (24) can be solved directly for  $\bar{y}'$ ,  $s'^2$ ,  $n'$ , and  $\nu'$ :

$$\bar{y}' = E[\mu|I_R] \quad (25)$$

$$n' = E[\sigma^2|I_R]/V[\mu|I_R] \quad (26)$$

$$\nu' = 2 \cdot E[\sigma^2|I_R]/V[\sigma^2|I_R] \quad (27)$$

$$s'^2 = [(\nu' - 2)/\nu']E[\sigma^2|I_R] \quad (28)$$

(A similar approach to the assessment of prior pdf's was presented by *Ando and Kaufman* [1964].) Estimates of  $E[\mu|I_R]$ ,  $V[\mu|I_R]$ ,  $E[\sigma^2|I_R]$ , and  $V[\sigma^2|I_R]$  can be obtained from purely subjective judgments or from straightforward analysis of other data.

The moments of  $\sigma^2$  are used instead of the moments of  $\sigma$  because the latter involve the gamma function. This would unnecessarily complicate the solution of these four equations for  $\bar{y}'$ ,  $n'$ ,  $\nu'$ , and  $s'^2$ . Equations (25)–(28) result in a prior pdf that has the marginal moments specified initially except that  $n'$  and  $\nu'$  are rounded off to the next smallest integer. These two parameters represent equivalent sample sizes for prior information about the mean and standard deviation, or variance, respectively.

Regional information has been used in the last few years to estimate streamflow parameters. Physiographic and/or meteorologic information can be related to streamflow characteristics such as mean annual flow and  $T$ -yr flood [*Benson and Matalas*, 1967; *Thomas and Benson*, 1970]. Regression models have been used to find these relations and to predict the streamflow characteristics of other basins.

Models of this type were developed for the New England region [*Vicens et al.*, 1974] and were used to predict the mean and variance of the annual flows for the Pemigewasset River at Plymouth, New Hampshire. The expected value and variance of these two predictions are

$$E[\mu|I_R] = 1271 \text{ ft}^3/\text{s} \quad V[\mu|I_R] = 16,868$$

$$E[\sigma^2|I_R] = 70,497 \quad V[\sigma^2|I_R] = 4.039 \times 10^8$$

By using these results the parameters of the prior pdf for  $\mu$  and  $\sigma$  were obtained by solving (25)–(28). These parameters are

$$\bar{y}' = 1271 \text{ ft}^3/\text{s} \quad n' = 4$$

$$s'^2 = 65,551 \quad \nu' = 28$$

From the variance of these predictions it can be estimated that the standard error of estimate for the mean is 10%, whereas for the variance it is 29%. In terms of equivalent sample sizes the predictions from the regression models were equal to about 4 yr for the mean and 28 yr for the variance. Although these equivalent sample sizes imply that more information is available for the variance than for the mean, the standard error of estimate is a better indicator of the precision of the information available. Much longer records are required to estimate the variance than to estimate the mean if the same level of precision is desired.

A second possible source of information is the subjective judgment of an experienced hydrologist. To take advantage of these talents, some conceptual model of the physical processes in action in a river basin would be helpful. Although totally subjective judgments (e.g., 'seat of the pants' guesses) could be used, a model would help structure the assessments of these

judgments. To this end a model similar to the Thomas model described by *Fiering* [1967] was developed. The streamflow output from a river basin in year  $i$ ,  $q_i$ , was related to the precipitation input  $x_i$  by

$$q_i = (1 - b)x_i \quad (29)$$

The parameters of the streamflow process are then related to the precipitation parameters by

$$\mu_q = (1 - b)\mu_x \quad (30)$$

$$\sigma_q^2 = (1 - b)^2 \sigma_x^2 \quad (31)$$

$$\rho_q = \rho_x \quad (32)$$

where  $\mu$ ,  $\sigma^2$ , and  $\rho$  denote the mean, variance, and serial correlation coefficient.

Suppose this model was, in the hydrologist's judgment, adequate to represent a particular river basin. In addition, the hydrologist was willing to specify some moments of the unknown variables  $\mu_x$ ,  $\sigma_x^2$ ,  $b$ , and  $\rho_x$ . Considering them as random variables allows the specification of moments, i.e.,  $E[\mu_x]$ ,  $V[\mu_x]$ ,  $E[\sigma_x^2]$ ,  $V[\sigma_x^2]$ ,  $E[b]$ ,  $V[b]$ ,  $E[\rho_x]$ , and  $V[\rho_x]$ . These moments can then be related to the moments of the random variables  $\mu_q$ ,  $\sigma_q^2$ , and  $\rho_q$  through approximations [*Benjamin and Cornell*, 1970] by assuming that all covariances are zero. These relations are

$$E[\mu_q] = (1 - E[b]) \cdot E[\mu_x] \quad (33)$$

$$V[\mu_q] = E^2[\mu_x] \cdot V[b] + (1 - E[b])^2 V[\mu_x] \quad (34)$$

$$E[\sigma_q^2] = [(1 - E[b])^2 + V[b]] \cdot E[\sigma_x^2] \quad (35)$$

$$V[\sigma_q^2] = 4(1 - E[b])^2 \cdot E^2[\sigma_x^2] \cdot V[b] + (1 - E[b])^4 \cdot V[\sigma_x^2] \quad (36)$$

$$E[\rho_q] = E[\rho_x] \quad (37)$$

$$V[\rho_q] = V[\rho_x] \quad (38)$$

The moments can be used directly in (25)–(28) to assess the prior parameters. Before proceeding with an example it is important to analyze how the required information is obtained. First, the units of the model will be changed to cubic feet per second for  $q_i$ . The precipitation  $x_i$  will also be measured in cubic feet per second and obtained from

$$x_i = K \cdot p_i \cdot A \quad (39)$$

where  $p_i$  is the precipitation in year  $i$  in inches per year,  $A$  is the drainage area in square miles, and  $K$  is the conversion factor from inches-square miles per year to cubic feet per second, equal to 0.07367. This conversion adds some uncertainty to the estimates of the mean  $\mu_x$  and the variance  $\sigma_x^2$ :

$$E[\mu_x] = \frac{1}{K \cdot A} \cdot E[\mu_p] \quad V[\mu_x] = A^2 \cdot K^2 V[\mu_p] \quad (40)$$

$$E[\sigma_x^2] = A^2 K^2 E[\sigma_p^2] \quad V[\sigma_x^2] = A^4 K^4 \cdot V[\sigma_p^2] \quad (41)$$

$$E[\rho_x] = E[\rho_p] \quad V[\rho_x] = V[\rho_p] \quad (42)$$

The following moments of  $p$  may be obtained from rainfall records:  $E[\mu_p]$ ,  $V[\mu_p]$ ,  $E[\sigma_p^2]$ ,  $V[\sigma_p^2]$ ,  $E[\rho_p]$ , and  $V[\rho_p]$ .

The hydrologist's judgment comes again into play in estimating the moments of  $b$ , the percentage loss. In  $E[b]$  the hydrologist expresses his 'best' estimate of the losses, whereas in  $V[b]$  he expresses his confidence in this best estimate.

The expected value of the mean annual precipitation can be

obtained from rainfall records for the basin of interest. The variance of the annual precipitation can be obtained, for example, by assuming a coefficient of variation of 0.2 for the annual precipitation time series. The variance of  $\mu_p$  and  $\sigma_p^2$  can be obtained by assuming that an  $n_p$ -yr rainfall record is available. Then the two variances can be computed by

$$V[\mu_p] = E[\sigma_p^2]/n_p \quad (43)$$

$$V[\sigma_p^2] = [2/(n_p - 1)]E^2[\sigma_p^2] \quad (44)$$

For the Pemigewasset River the following information about the rainfall process was used:

$$E[\mu_p] = 48.5 \text{ in.} \quad E[\sigma_p^2] = 94.1 \quad E[\rho_p] = 0$$

$$V[\mu_p] = 4.7 \quad V[\sigma_p^2] = 932.1 \quad V[\rho_p] = 0$$

$$A = 622 \text{ mi}^2 \quad K = 0.07367$$

In addition, the annual precipitation process was assumed to be an independent normal process. The value of  $E[\mu_p]$  was obtained from Johnston [1970], and the value of  $E[\sigma_p^2]$  was obtained by assuming that the coefficient of variation of this process was 0.2. The variance terms were obtained from (43) and (44) and a subjective 'guess' that  $n_p$  was 20 yr. (No information was available concerning the length of record used by Johnston [1970] in his studies.) Finally, it was assumed that the precipitation process at the annual level was independent.

Using these values in (40)–(42), we obtain

$$E[\mu_x] = 2222.4 \text{ ft}^3/\text{s} \quad V[\mu_x] = 9868.2$$

$$E[\sigma_x^2] = 197,570 \quad V[\sigma_x^2] = 4.109 \times 10^9$$

$$E[\rho_x] = 0 \quad V[\rho_x] = 0$$

A small group of hydrologists were asked about their subjective estimation of the percentage loss  $b$ , the results being

$$E[b] = 0.4 \quad V[b] = 0.01$$

The results in terms of the moments of the streamflow parameters are

$$E[\mu_q] = 1333 \text{ ft}^3/\text{s} \quad V[\mu_q] = 52,941$$

$$E[\sigma_q^2] = 73,103 \quad V[\sigma_q^2] = 1.095 \times 10^9$$

$$E[\rho_q] = 0 \quad V[\rho_q] = 0$$

These values were then substituted into (25)–(28) to obtain the prior parameters of a normal inverted gamma 2:

$$\bar{y}' = 1333.4 \text{ ft}^3/\text{s} \quad s'^2 = 62,480$$

$$n' = 1 \text{ yr} \quad \nu' = 13 \text{ yr}$$

Considering  $n'$  and  $\nu'$  as prior equivalent sample sizes there is very little information about the mean. However, there is some information about the variance. The standard error for the mean in percentage is about 17%, whereas for the variance it is 45%.

This prior pdf is very sensitive to the 'confidence' on the estimate of  $b$ . Thus for  $V[b] = 0.0025$  the following prior parameters are obtained:

$$\bar{y}' = 1334.2 \text{ ft}^3/\text{s} \quad s'^2 = 64,177$$

$$n' = 4 \quad \nu' = 19$$

An increase in  $n'$  has been achieved by improving the information about  $b$ .

Use of the prior information together with the historical record always yields lower variance than use of only the prior or only the historical samples. In this example the quantification of the hydrologist's judgments was limited to estimating the losses of the basin and a confidence value about this estimate. For more complex models, more assessments might be required. These could involve groundwater storage characteristics, losses from surface storage, or changes in the basin due to urbanization. This last factor might make historical records worth less, since the basin has been altered. In this case the assessment of a prior distribution is very important, since it would be the only information available.

These results of prior pdf assessments from regression are in accordance with the results of Johnston [1970] and Hardison [1969, 1971]. In terms of equivalent sample sizes the regression models did not yield large amounts of information. But even this small amount of information can be quite valuable when it is combined with the historical record, as will be done later in this paper. The subjective assessments yielded even less information than the regression models. This result was expected because of the simple nature of the model used.

*Prior to posterior analysis.* The regional information in the prior distribution may be combined with the historical record at the site of interest through the application of Bayes's theorem, that is,

$$f''(\mu, \sigma | I_R, Y) \propto f'(\mu, \sigma | I_R) \cdot L(\mu, \sigma | Y) \quad (45)$$

Since the prior pdf was selected from a natural conjugate family, i.e., one with a kernel similar to the likelihood function, the posterior pdf will be of the same family. As is shown by Vicens et al. [1974], the posterior pdf resulting from (45) will be a normal inverted gamma 2 pdf, i.e.,

$$f_{NIG2}''(\mu, \sigma | I_R, Y) \propto f_{NIG2}'(\mu, \sigma | I_R) \cdot L(\mu, \sigma | Y) \quad (46)$$

with parameters

$$\bar{y}'' = (n'\bar{y}' + n\bar{y})/(n' + n) \quad (47)$$

$$n'' = n' + n \quad (48)$$

$$\nu'' = \nu' + \nu + 1 \quad \nu' > 0 \quad \nu > 0 \quad (49)$$

$$s''^2 = (\nu's'^2 + n'\bar{y}'^2 + \nu s^2 + n\bar{y}^2 - n''\bar{y}''^2)/\nu'' \quad (50)$$

$$\nu'' > 0$$

The posterior parameters are functions of both the prior and the sample parameter. For example, the posterior mean  $\bar{y}''$  is a weighted function of the prior mean  $\bar{y}'$  and the sample mean  $\bar{y}$ , where the weights are the relative number of samples  $n'/(n' + n)$  and  $n/(n' + n)$ , respectively. If no prior information exists, the posterior parameters will be identical with the sample statistics. When no sample exists, the posterior parameters are identical with the prior parameters.

Since the joint posterior pdf is normal inverted gamma 2, identical with the prior pdf except for the new parameters, the marginal posterior pdf's for  $\mu$  and  $\sigma$  are defined as for the prior pdf's ((15) and (18)), and the expected value and variance of these two parameters are given by (16), (17), (19), and (20) except that  $\bar{y}''$ ,  $s''^2$ ,  $n''$ , and  $\nu''$  replace the prior parameters.

As an example of how regional information can be combined with at-site historical records, this model was applied to the Pemigewasset River. The results of the prior to posterior analysis are presented in Tables 2 and 3. The prior information obtained from regional data through regression analysis was

TABLE 2. Prior to Posterior Analysis, Pemigewasset River

Information	$E[\mu], \text{ft}^3/\text{s}$	$V[\mu]$	$E[\sigma^2]$	$V[\sigma^2]$
<i>Prior Only (Regression Models)</i>				
$n' = 4, v' = 23$	1271	17624	70497	$4.04 \times 10^3$
<i>Sample Only</i>				
$n = 5, v = 4$	1384	25847	129235	*
$n = 20, v = 19$	1368	2304.8	46096	$2.85 \times 10^3$
$n = 60, v = 59$	1347	1050.6	63037	$1.44 \times 10^3$
<i>Posterior (Prior and Sample)</i>				
$n'' = 9, v'' = 33$	1334	7604.0	68457	$3.12 \times 10^3$
$n'' = 24, v'' = 43$	1352	2409.5	57829	$1.49 \times 10^3$
$n'' = 64, v'' = 83$	1342	991.0	63424	$0.95 \times 10^3$

\*Does not exist, since  $v = 4$ .

TABLE 3. Prior to Posterior Analysis, Pemigewasset River

Information	$E[\mu], \text{ft}^3/\text{s}$	$V[\mu]$	$E[\sigma^2]$	$V[\sigma^2]$
<i>Prior Only (Subjective Assessments)</i>				
$n' = 1, v' = 13$	1333	73103	73103	$11.9 \times 10^8$
<i>Sample Only</i>				
$n = 5, v = 4$	1384	25847	129235	*
$n = 20, v = 19$	1368	23304.8	46096	$2.85 \times 10^3$
$n = 60, v = 59$	1347	1050.6	63037	$1.44 \times 10^3$
<i>Posterior (Prior and Sample)</i>				
$n'' = 6, v'' = 18$	1376	11091	66548	$6.55 \times 10^3$
$n'' = 21, v'' = 33$	1366	2440.7	51256	$1.81 \times 10^3$
$n'' = 61, v'' = 72$	1347	1015.4	61937	$1.11 \times 10^3$

\*Does not exist, since  $v = 4$ .

combined with the historical record (Table 1). In Table 2 the expected value and variance for the mean and variance of annual streamflows are shown: first, from the prior information only; second, from the historical record only; and finally, from both combined. In addition, the marginal distributions for the mean and standard deviation are shown in Figures 2-7.

The subjective prior pdf is combined with the historical record in Table 3. The results are similar to those of the regression prior pdf.

The following observations can be made. First, the posterior

information is a combination of the prior and the sample information. For example, for the mean annual flow the posterior expected value of  $\mu$  is a weighted sum of the prior expected value and the sample expected value. Second, as the historical sample size increases (larger  $n$ ), the sample information is weighted more heavily than the prior information. Again for the mean annual flow the posterior expected value of  $\mu$  approaches the sample expected value. Third, as the total information increases (larger  $n''$ ), the uncertainty about the parameters  $\mu$  and  $\sigma^2$  as measured by their posterior variance is

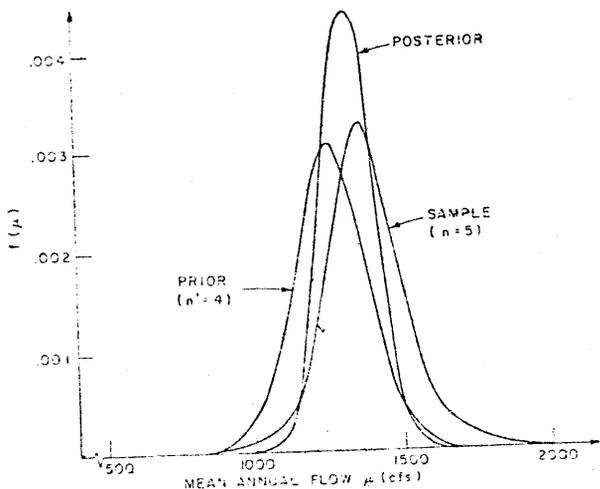


Fig. 2. Marginal pdf's of the mean annual flow (sample:  $n = 5$ ).

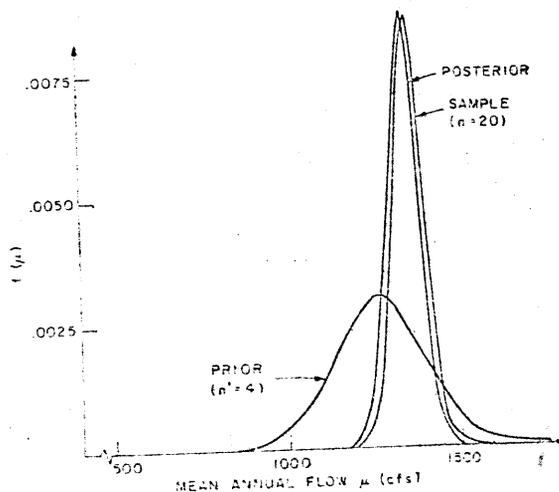


Fig. 3. Marginal pdf's of the mean annual flow (sample:  $n = 20$ ).

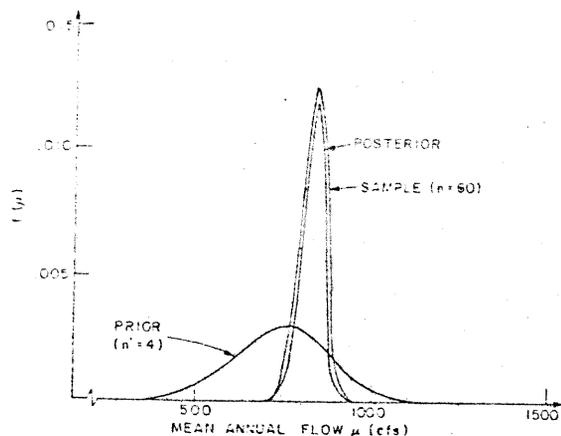


Fig. 4. Marginal pdfs of the mean annual flow (sample:  $n = 60$ ).

reduced. Fourth, and most important, the posterior variances of  $\mu$  and  $\sigma^2$  are lower than those of either the prior or the sample alone. This fact indicates that combining the two sources of information reduces the parameter uncertainty or equivalently increases the total information over using each source separately and then trying to decide which of the two is more adequate.

These observations are demonstrated more dramatically by the figures of the marginal pdfs. A reduction in variance is reflected in a 'tightening' of the pdf about its mean. From Figures 2-7 it can be seen that the marginal posterior pdfs are concentrated closer to their mean than either the prior or the sample pdf. In addition, where the prior or sample contains significantly more information than the other, the posterior pdf closely follows the 'stronger' one.

*Bayesian distribution of a future streamflow.* It has been assumed that the model pdf of the streamflows was a normal pdf given the parameters  $\mu$  and  $\sigma$ . But owing to the uncertainty about  $\mu$  and  $\sigma$  the information about a future streamflow  $y_T$  is not complete if this parameter uncertainty is not included. Inferences or decisions about future streamflows should take this parameter uncertainty into account. Following (3), integration over the uncertainty about  $\mu$  and  $\sigma$  can be carried out to obtain the Bayesian pdf of a future streamflow  $y_T$  of this process:

$$\begin{aligned} \bar{f}(y_T | I_R, Y) &= \int_0^\infty \int_{-\infty}^\infty f_N(y_T | \mu, \sigma) \cdot f_{NIG2}''(\mu, \sigma) d\mu d\sigma \\ &= \bar{f}_S(y_T | \bar{y}'', s''^2/r, \nu'') \end{aligned} \quad (51)$$

where

$$r = n''/(n'' + 1) \quad (52)$$

(This result is proved in Appendix B of Vicens et al. [1974].) The Bayesian pdf of a future observation from this model is a Student  $t$  pdf. By stating that the probability statements about  $y_T$  are now in the form of a Student  $t$  instead of the model pdf, the normal pdf, it should be obvious immediately that owing to the parameter uncertainty about  $\mu$  and  $\sigma$  the uncertainty about  $y_T$  has increased. Moreover, when the Bayesian moments are computed, the results are

$$E[y_T | I_R, Y] = \bar{y}'' \quad (53)$$

$$V[y_T | I_R, Y] = s''^2(n'' + 1)\nu''/n''(\nu'' - 2) \quad (54)$$

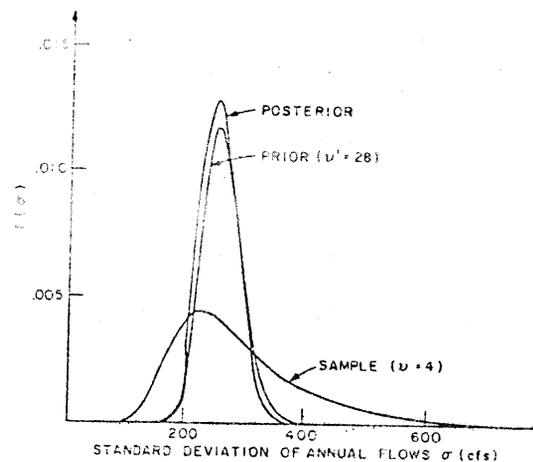


Fig. 5. Marginal pdfs of the standard deviation of annual flows (sample:  $n = 5$ ).

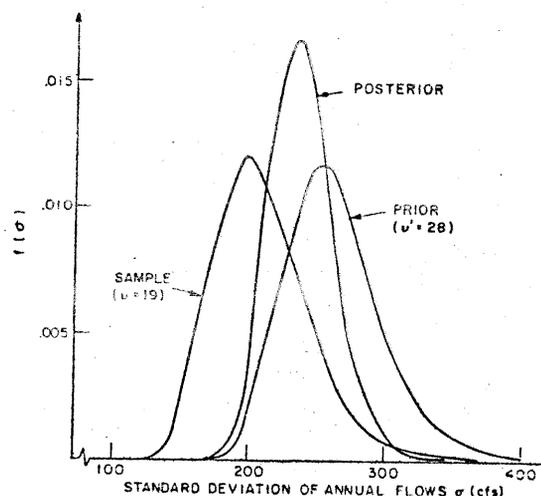


Fig. 6. Marginal pdfs of the standard deviation of annual flows (sample:  $n = 20$ ).

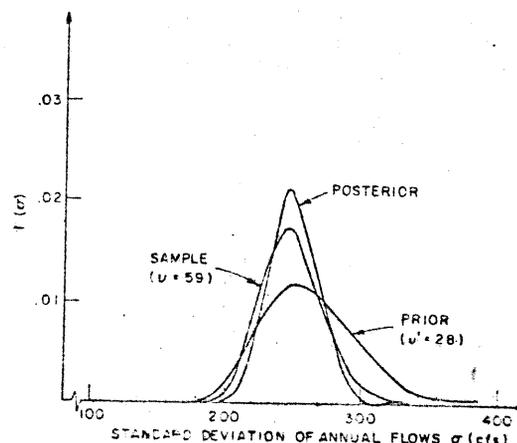


Fig. 7. Marginal pdfs of the standard deviation of annual flows (sample:  $n = 60$ ).

TABLE 4. Moments of the Bayesian pdf, Pemigewasset River

Information	$E[y_f]$ , ft <sup>3</sup> /s	$V[y_f]$
<i>Prior Only (Regression Models)</i>		
$n' = 4, v' = 28$	1271	88122
<i>Sample Only</i>		
$n = 5, v = 4$	1384	155082
$n = 20, v = 19$	1368	48401
$n = 60, v = 59$	1347	64088
<i>Posterior (Prior and Sample)</i>		
$n'' = 9, v'' = 53$	1354	76041
$n'' = 24, v'' = 48$	1352	60237
$n'' = 64, v'' = 88$	1342	64415

which show a larger variance than they would if  $\mu$  and  $\sigma$  were assumed to be known and equal to  $\bar{y}''$  and  $s''$ . With an increased total number of samples, i.e., larger  $n''$  and  $v''$ , the information will increase, and the variance of  $y_f$  will tend to its true value, since  $(n'' + 1)/n''$  and  $v''/(v'' - 2)$  will both tend to unity.

For the purpose of inferences, (51) presents the pdf of a future streamflow  $y_f$ , which contains all of the available information about the process. For example, the probability that a future flow will be less than  $q$  is

$$P[y_f \leq q] = \int_{-\infty}^q \tilde{f}_S(y_f | I_R, Y) dy_f \quad (55)$$

and values for this probability can be obtained from Student  $t$  tables.

Again if only prior information is available, the moments and parameters in (53) and (54) will be the prior ones. If a non-informative prior pdf is used, the posterior Bayesian pdf will contain only the historical sample information.

The effects of parameter uncertainty on inferences or decisions about future streamflows will be transferred through the Bayesian pdf. The moments of the Bayesian pdf given prior information from regression models and posterior information from the historical sample are presented in Table 4.

Two facts should be pointed out in these results. First, the Bayesian pdf includes all of the available information. Therefore the Bayesian moments reflect the information that was available through the posterior pdf. For example, the expected value of the future streamflow is equal to the posterior expected value of the mean annual flow. As was discussed previously, this posterior expected value is a weighted average

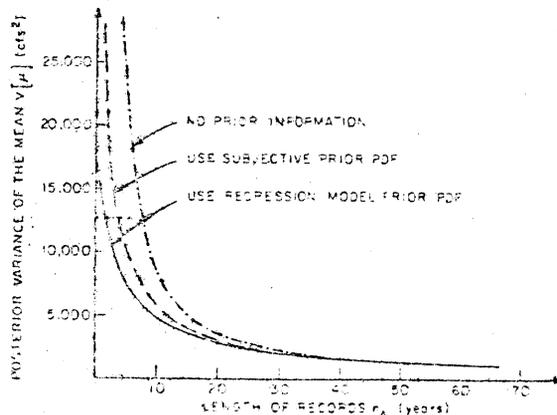


Fig. 9. Marginal posterior variance of the mean annual flow.

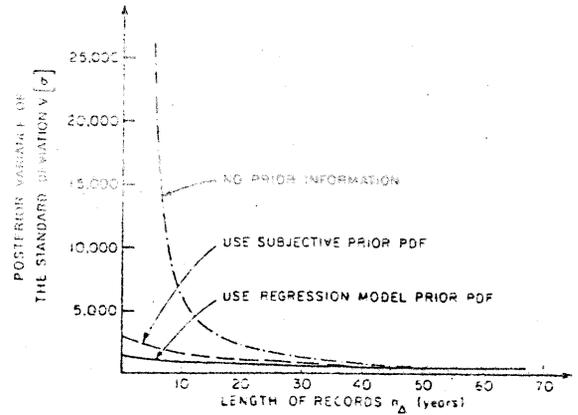


Fig. 9. Marginal posterior variance of the standard deviation of the annual flows.

of the prior and sample information. If only the prior pdf is used, the expected value of  $y_f$  reflects only the prior information. If the sample is the only source of information, only it is reflected in the Bayesian expected value of  $y_f$ .

Second, the Bayesian variance of  $y_f$  includes both the natural uncertainty of the streamflows and the parameter uncertainty due to lack of perfect information. Consequently, for short samples, prior or historical, the variance of  $y_f$  is larger than what appears to be the process variance ( $\sim 60,000$  ft<sup>3</sup>/s). But as the total information increases (larger  $n''$ ), the Bayesian variance approaches the process variance, as can be expected, since the parameter uncertainty is eliminated.

*Advantages of Bayesian analysis.* Two of the major advantages of the Bayesian approach are that parameter uncertainty is included in the analysis explicitly and that by using informative prior pdf's this parameter uncertainty may be reduced. To demonstrate these advantages of the Bayesian approach and to test the relative value of using informative prior pdf's versus longer historical records, the record of the Pemigewasset River and the prior pdf's described earlier were used as an example.

The historical record was divided into a set of samples  $n_A$  yr long each. Each set was combined with three sets of prior information pdf's: noninformative, regression model, and subjective assessments. The posterior moments of  $\mu$ ,  $\sigma$ , and  $y_f$  were computed for each combination of prior and sample of length  $n_A$ . The results were averaged over the total number of  $n_A$ -length samples that could be obtained from the historical record. This process was repeated for  $n_A = 5, 10, 15, 20, 30, 40, 50, 60,$  and  $65$  yr. For some samples, some part of the

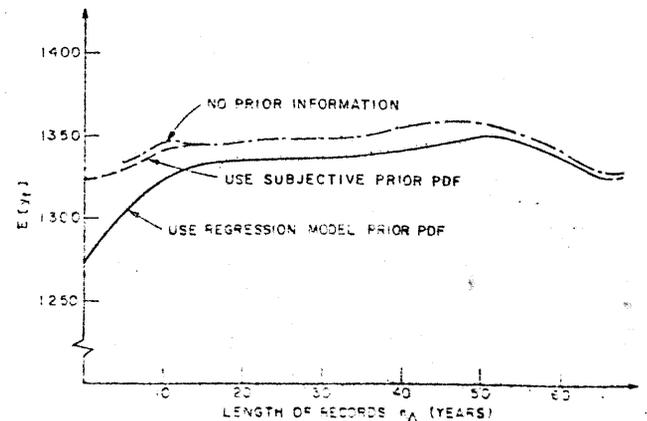


Fig. 10. Expected value of future streamflow from Bayesian pdf.

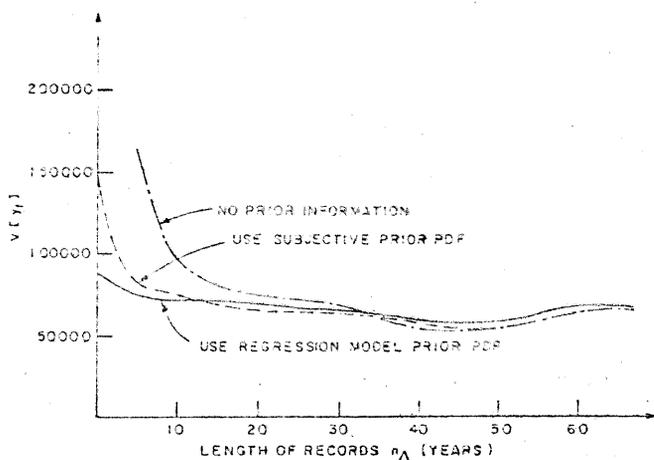


Fig. 11. Variance of future streamflow from Bayesian pdf.

historical record was discarded. For example, for  $n_d = 20$ , three samples of 20 yr were used, and the last 7 yr of the record were discarded. Figures 8 and 9 show the marginal posterior variance of the mean and standard deviation of the annual flows versus  $n_d$  for each of the three prior pdf's.

Several important trends are shown by these figures. First, when either of the informative prior pdf's is combined with the historical record, the posterior variances are lower than they are when the historical record is used alone (diffuse prior pdf) or when the prior information is used alone ( $n_d = 0$ ). These differences are very significant for  $n_d$  less than 25 yr. For example, at  $n_d = 10$  the posterior variance of the mean annual flow can be reduced by more than 30% by combining the historical record with an informative prior pdf. For the posterior variance of the standard deviation a reduction of 70% can be obtained by the same procedure. Of course, for longer historical records the importance and value of the prior pdf diminish. For  $n_d$  more than 40 yr the differences are insignificant. These results are particular to this case only and valid if the assumptions of the model normality and independence have not been violated.

The effects of informative priors on the Bayesian pdf are demonstrated in Figures 10 and 11. These figures show the Bayesian expected value and variance of a future streamflow for information from the historical records and the three priors discussed earlier. The expected value appears to be a

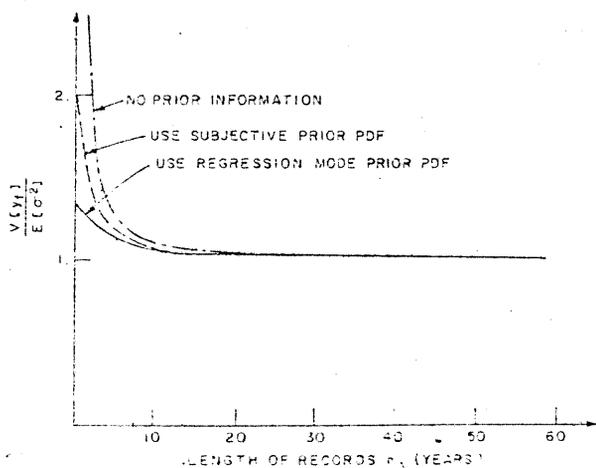


Fig. 12. Ratio of variance of Bayesian pdf to marginal posterior expected value of process variance.

weighted average of the prior and sample information, whereas the variance reflects the uncertainties, natural and parameter, in the prediction. Here again the use of an informative prior pdf significantly reduces the variance for  $n_d$  less than 20 yr. After this point the value of the Bayesian variance oscillates about what appears to be the true process variance ( $\sim 60,000 \text{ ft}^3/\text{s}$ ).

Finally, the Bayesian approach attempts to deal with the parameter uncertainty by including this factor explicitly in the analysis. This approach leads to using the mean and variance of the Bayesian pdf for simulations and/or decisions. This variance is generally larger than the process variance because it includes the parameter uncertainty. The ratio of the variance of the Bayesian pdf  $V[y_t]$  to the expected value of the process variance  $E[\sigma^2]$  will always be larger than 1, but it will approach 1 as the parameter uncertainty is eliminated. Figure 12 shows a plot of this ratio versus  $n_d$  for combinations of the historical record and the three prior pdf's. As is expected, the use of an informative prior pdf reduces the parameter uncertainty.

Reducing the posterior variance of  $\mu$  and  $\sigma$  and the variance of the Bayesian pdf is also related to the parameter estimation problem. A reduction in these variances directly reduces the expected losses of Bayes estimates. These losses may be shown to be directly proportional to the variance of the unknown variable.

## CONCLUSIONS

This paper has described the Bayesian analysis of hydrologic models in general and the independent normal process as an example. The distribution theory and assumptions for this model have been discussed. In addition, an application to a typical New England river has been presented.

The following general conclusions can be made.

1. Use of regional information through prior pdf's significantly reduces the parameter uncertainty when the historical records are short ( $n < 25$  yr).
2. Use of the Bayesian pdf for design purposes accounts for the parameter and natural uncertainties. For this particular model the result is a change from a normal pdf to a Student  $t$  and an increase in the predicted variance.
3. Use of regional information reduces the uncertainty in the Bayesian pdf by reducing the parameter uncertainty.

In summary, the Bayesian approach is more explicit in considering the parameter uncertainty in inferences and decisions. At the same time, the use of regional information through informative prior pdf's reduces this parameter uncertainty, especially for short historical records.

Research currently in progress is aimed at the analysis of more complex models of hydrologic time series, such as multivariate or multilag models, within a Bayesian framework and at the problem of model selection.

## NOTATION

- $E[ ]$  expected value operator.  
 $f( )$  probability distribution function.  
 $F( )$  cumulative distribution function.  
 $f(y_i | \theta)$  conditional pdf of  $y_i$  given the parameter set  $\theta$ , model pdf.  
 $\hat{f}(y_t)$  Bayesian, unconditional, or predictive pdf of  $y_t$ .  
 $f(\theta | I_R)$  prior pdf of the parameter set  $\theta$  given information  $I_R$ .  
 $f(Y | \theta)$  joint conditional pdf of observing  $Y$  for specified values of  $\theta$ .

- $f''(\theta)$  posterior pdf of the parameter set  $\theta$ .  
 $h$  process precision, equal to  $1/\sigma^2$ .  
 $I_R$  regional information about parameters of the process.  
 $k(\theta)$  kernel of the likelihood function of  $\theta$ .  
 $L(\theta|Y)$  likelihood function of  $\theta$  given the observations  $Y$ , proportional to  $f(Y|\theta)$ .  
 $n'$  equivalent number of prior samples.  
 $n$  number of historical samples.  
 $n''$  number of posterior samples.  
 $n_s$  sample size of sets of historical traces.  
 $q$  specified streamflow value.  
 $s'^2$  prior variance parameter.  
 $s^2$  sample variance parameter.  
 $s''^2$  posterior variance parameter.  
 $u(\ )$  utility function.  
 $U(\ )$  total utility.  
 $V[ ]$  variance operator.  
 $y_i$   $i$ th observation of streamflow process.  
 $Y$  set of observations of  $y$ , equal to  $[y_1, y_2, \dots, y_i, \dots]$ .  
 $\bar{y}'$  prior mean.  
 $\bar{y}$  sample mean.  
 $\bar{y}''$  posterior mean.  
 $y_f$  future streamflow.  
 $\theta$  parameter vector, equal to  $[\theta_1, \theta_2, \dots, \theta_j, \dots]$ .  
 $\mu$  mean of independent normal process.  
 $\nu'$  prior degrees of freedom.  
 $\nu$  sample degrees of freedom.  
 $\nu''$  posterior degrees of freedom.  
 $\pi$  pi, equal to 3.1416.  
 $\sigma$  standard deviation of independent normal process.  
 $\sigma^2$  variance of independent normal process.

*Acknowledgments.* This work was performed at the Ralph M. Parsons Laboratory for Water Resources and Hydrodynamics at the Massachusetts Institute of Technology as part of the first author's doctoral dissertation under the support of the Office of Water Resources Research, grant 14-31-0001-9021. The enlightening discussions with Gordon M. Kaufman of the Alfred P. Sloan School of Management and the friendly criticisms of Eric F. Wood of the International Institute for Applied Systems Analysis are gratefully acknowledged.

#### REFERENCES

- Ando, A. K., and G. M. Kaufman, Extended natural conjugate distributions for the multinormal process, *Work. Pap. 80-64*, Alfred P. Sloan Sch. of Manage., Mass. Inst. of Technol., Cambridge, 1964.
- Benjamin, J. R., and C. A. Cornell, *Probability, Statistics, and Decision for Civil Engineers*, p. 181, McGraw-Hill, New York, 1970.
- Benson, M. A., and N. C. Matalas, Synthetic hydrology based on regional statistical parameters, *Water Resour. Res.*, 3(4), 931-935, 1967.
- Davis, D. R., C. C. Kisiel, and L. Duckstein, Bayesian decision theory applied to design in hydrology, *Water Resour. Res.*, 8(1), 33-41, 1972.
- Fiering, M. B., *Streamflow Synthesis*, p. 69, Harvard University Press, Cambridge, Mass., 1967.
- Hardison, C. H., Accuracy of streamflow characteristics, *U.S. Geol. Surv. Prof. Pap. 650-D*, D216-D214, 1969.
- Hardison, C. H., Prediction error of regression estimates of streamflow characteristics at ungaged sites, *U.S. Geol. Surv. Prof. Pap. 750-C*, C228-C236, 1971.
- Johnston, C. G., A proposed streamflow data program for central New England, open file report, U.S. Geol. Surv., Boston, Mass., 1970.
- Kaufman, G. M., Course notes for decision analysis (15.065), Mass. Inst. of Technol., Cambridge, Mass., 1972.
- Lenton, R. L., I. Rodriguez-Iturbe, and J. C. Schaake, Jr., A Bayesian approach to autocorrelation estimation in hydrologic autoregressive models, *Rep. 163*, Ralph M. Parsons Lab. for Water Resour. and Hydrodyn., Mass. Inst. of Technol., Cambridge, Mass., Jan. 1973.
- Matalas, N. C., and E. J. Gilroy, Some comments on regionalization in hydrologic studies, *Water Resour. Res.* 4(6), 1361-1369, 1968.
- Raiffa, H., and R. Schlaifer, *Applied Statistical Decision Theory*, MIT Press, Cambridge, Mass., 1961.
- Shane, R. M., and D. D. Gaver, Statistical decision theory techniques for the revision of mean flood flow regression estimates, *Water Resour. Res.*, 6(6), 1649-1654, 1970.
- Thomas, D. M., and M. A. Benson, Generalization of streamflow characteristics from drainage-basin characteristics, *U.S. Geol. Surv. Water Supply Pap. 1975*, 26-31, 1970.
- Tschannerl, G., Designing reservoirs with short streamflow records, *Water Resour. Res.*, 7(4), 827-833, 1971.
- Vicens, G. J., I. Rodriguez-Iturbe, and J. C. Schaake, Jr., A Bayesian approach to hydrologic time series modeling, *Rep. 181*, Ralph M. Parsons Lab. for Water Resour. and Hydrodyn., Mass. Inst. of Technol., Cambridge, March 1974.
- Wood, E. F., I. Rodriguez-Iturbe, and J. C. Schaake, Jr., The methodology of Bayesian inference and decision making applied to extreme hydrologic events, *Rep. 178*, Ralph M. Parsons Lab. for Water Resour. and Hydrodyn., Mass. Inst. of Technol., Cambridge, Jan. 1974.
- Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, John Wiley, New York, 1971.

(Received June 24, 1974;  
accepted December 2, 1974.)