

パターン認識手法による多雪地帯の月流量時系列の解析とシミュレーション
STOCHASTIC SYNTHESIS OF HYDROLOGIC DATA BASED ON CONCEPTS OF PATTERN RECOGNITION

金沢大学工学部 正 員 高 瀬 信 忠
同 上 正 員 宇 治 橋 康 行
金沢大学大学院 学生員 家 門 邦 彦

1 はじめに

河川流量時系列中には、日、週、月等の測定時間単位にかかわらず、流量パターンと呼ばれる明確なグループ（高水流量と低水流量）が存在する。特に北海道、東北、北陸等の多雪地帯においては、河川流量は5～7ヶ月の長期にわたって降積雪の影響を受けるので、融雪水と降雨という物理的に異なる原因から生じるため流量時系列中には物理的にも異なるグループ、すなわち流量変動パターンが存在する。しかし、従来流量時系列の解析やシミュレーションにおいてはこのような点は全く考慮されていない。

本研究では、情報処理の一分野として発展しつつあるパターン認識の概念を応用した Unny^{1),2)}らの手法により多雪地帯の月単位流量時系列の特徴解析とそれに基づくデータシミュレーションを行ない、多雪地帯の月単位流量時系列のパターン特性と Unny らの手法の実用性について検討した。

2 流量時系列のパターン認識システム

月流量時系列に対するパターン認識システムは以下に示す手順に従って構成される。

手順1 流量時系列の対象への分割 N年間の時系列データは物理年に従ってN個のセグメントに分割されさらに各セグメントはKシーズンに分けられ、時系列は N^* （ $=KN$ ）個の対象に分割される。各セグメントのシーズンへの分割は、コログラム解析、ペリオドグラム解析、ピークー谷解析等によって行なわれる。ピークー谷解析において、 $x_{t-1} < x_t > x_{t+1}$ ならばピークの生起、 $x_{t-1} > x_t < x_{t+1}$ ならば谷の生起、それ以外はピークあるいは谷の不生起である。

手順2 パターンベクトルの構成 パターン認識では認識の対象に対してn個の観測を行ない、対象の特性を定量化して対応するパターンベクトルに変換する。月流量時系列の場合に対しては、第i年の第kシーズンに対応する対象に関するパターンベクトルは次式で表わされる。

$$x_i^{(k)} = \{ x_{i1}^{(k)}, x_{i2}^{(k)}, \dots, x_{in}^{(k)} \}^T, \quad i=1, 2, \dots, N, \quad k=1, 2, \dots, K \quad (1)$$

各年の同じシーズンに対応するすべてのパターンベクトルを集めたものは一つのパターンクラスを作る。なお後述するデータシミュレーションにおいて特徴ベクトルの多変数正規性を利用するので、認識システムには直接関係しないがパターンベクトルの多変数正規性が検定され、多変数正規分布と見なされない場合はパターンベクトルの各要素に対し適当な変換を行なって多変数正規分布とする。

手順3 特徴抽出 通常認識を行なう以前には対象の認識にどのような観測が必要であるかは分らない。したがって対象に関して行なわれた観測には不必要な情報が含まれている。そこで対象の認識に必要な情報、すなわち特徴を抽出し、パターンベクトルを特徴ベクトルに変換することが必要である。特徴抽出には種々の方法³⁾があるが、ここでは次の線形変換によって特徴抽出を行なう。

$$y_i^{(k)} = A^{(k)} x_i^{(k)} \quad (2)$$

ここに、 $y_i^{(k)}$ は $x_i^{(k)}$ に対応する特徴ベクトル、 $A^{(k)}$ は $m \times n$ の変換行列である。この変換が適切なものであればパターンベクトルの次元を減少させるとともに特徴ベクトルの要素のばらつきを最少にする。Panu⁴⁾らは特徴ベクトルの成分が m 次元正規分布をする場合、パターンベクトルのn個の成分の分散-共分散行列

の正規化された固有ベクトルを列に持つ行列が適切な変換行列であることを指摘している。

手順4 reference vectorの構成 reference vectorは、パターンクラス内の特徴ベクトルがそのまわりにグループを形成することが可能になるようなサブクラスターの中心であり、パターンクラス内の特徴ベクトル群の平均である。式(2)の変換により、すべての特徴ベクトルの第j成分は正と負の2つのグループに分けられる。各グループの平均値を ${}_1\theta_j^{(k)}$, ${}_2\theta_j^{(k)}$, これら平均値の標準偏差を各々 ${}_1S_j^{(k)}$, ${}_2S_j^{(k)}$ で表わせば、特徴ベクトル及びreference vectorの次元mがパターンベクトルの次元nに等しい場合、第kパターンクラスの第l reference vectorは次式で表わされる。

$$\left. \begin{aligned} z_{\ell}^{(k)} &= \{ z_{\ell 1}^{(k)}, z_{\ell 2}^{(k)}, \dots, z_{\ell n}^{(k)} \}^T, \quad \ell=1, 2, \dots, L^{(k)}, \quad k=1, 2, \dots, K \\ z_{\ell j}^{(k)} &\in \{ {}_1\theta_j^{(k)}, {}_2\theta_j^{(k)}, i^*=1, 2 \} \end{aligned} \right\} \quad (3)$$

ここに、 $L^{(k)}$ は第kパターンクラスのreference vectorの総数を表わす。reference vectorのn個の成分は、正か負の2つの値のどちらかを取るのので、可能なreference vectorの総数は、 $L_k = 2^n$ 個である。しかしながら、実際に変換された特徴ベクトルの第j成分は、すべて正か負である場合が多く、可能なreference vectorの数は大きく減少する。

手順5 分類規準 特徴ベクトルを認識し、特定のreference vectorのまわりのサブクラスターに割り当てる。第kパターンクラスの特定のreference vector $z_{\ell}^{(k)}$ と任意の特徴ベクトルyとの距離 $d_{\ell}^{(k)}$ は、

$$d_{\ell}^{(k)} = |y - z_{\ell}^{(k)}| = [(y - z_{\ell}^{(k)})^T (y - z_{\ell}^{(k)})]^{1/2} \quad (4)$$

で定義される。任意の特徴ベクトルyを第kパターンの特定のreference vectorに割り当てるためのSebestyenの最小距離の概念による決定規準は式(5)で表わされる。

$$d_{\ell}^{(k)} = \min_{j,u} |y - z_u^{(j)}|, \quad j=1, 2, \dots, K, \quad u=1, 2, \dots, L^{(j)} \quad (5)$$

一つのreference vectorに割り当てられたすべての特徴ベクトルとreference vectorとの距離 $d_{\ell}^{(k)}$ の平均あるいは他の適当な統計量として各reference vectorに特性距離あるいはしきい距離 $d_{\ell}^{*(k)}$ が割り当てられる。⁵⁾

最小距離による分類規則は、特徴ベクトルの持つ他の特性を反映しないので不十分である。認識、分類をより正確に行なうために誤認規準が導入される。誤認規準は特徴ベクトル全体としての類似度を表わす距離に対し、各要素間類似度の規準を表わすものである。すなわち、誤認規準は、

$$\text{誤 認} = \begin{cases} 0, & \mu_{\ell} \leq y_{ij}^{(k)} \leq \mu_u \\ 1, & \mu_{\ell} > y_{ij}^{(k)} \text{ or } y_{ij}^{(k)} > \mu_u \end{cases} \quad (6)$$

ここで、 $\mu_{\ell} = z_{\ell j}^{(k)} - \nu S_j^{(k)}$, $\mu_u = z_{\ell j}^{(k)} + \nu S_j^{(k)}$ であり、 ν は95%信頼限界に対応する値であり $S_j^{(k)}$ はreference vectorの各成分の標準偏差である。特徴ベクトルは、誤認が0で、かつ距離が最小のreference vectorの属するパターンクラスに分類される。

3 データシミュレーション

3.1 パターン構造の統計解析

パターン認識概念に基づくデータシミュレーションを行なうためにはパターン認識システムを確立するとともにパターン内及びパターン間構造の統計的性質を解析する必要がある。

前述のように特徴ベクトルの各成分は多変数正規分布であると仮定されるので、パターン内構造は、特徴

ベクトルの成分の平均ベクトル（すなわちreference vector） $\mu_y = z_l$ ，分散-共分散行列 C_y によって次の多変数正規分布で表わされる。

$$p(y) = (2\pi)^{-m/2} [\det C_y]^{-1/2} \exp \left\{ -\frac{1}{2} (y - \mu_y)^T C_y^{-1} (y - \mu_y) \right\} \quad (7)$$

パターン間構造は，reference vectorの生起を表わすものであり，reference vectorの持続性を表わすものであり，ここでは1次のマルコフ過程を仮定する。このとき第 k パターンクラスのreference vector $z_l^{(k)}$ の生起確率は次のようになる。

$$p(y_l^{(k)}) = \begin{cases} p[y_l^{(k)} | y_l^{(k-1)}] & , k \geq 2 \\ p[y_l^{(k)} | y_l^{(k-1)} = y_l^{(k)}] & , k=1 \end{cases} \quad (8)$$

時刻 $i-1$ でのreference vector z_g から時刻 i でのreference vector z_h への遷移の総数を $\beta_{gh}(i)$ とおけば，遷移確率 $p_{gh}(i)$ は次式で表わされる。

$$\hat{p}_{gh}(i) = \beta_{gh}(i) / \sum_{w=1}^L \beta_{gw}(i) \quad (9)$$

$$1 \leq g \leq L, \quad 1 \leq h \leq L, \quad i=1, 2, \dots, N^* (=NK)$$

ここに， $\hat{p}_{gh}(i)$ は $p_{gh}(i)$ の推定値を表わし， L はreference vectorの総数である。

定常遷移確率 p_{gh} の推定値 $\hat{p}_{gh}$ は次式で与えられる。

$$\hat{p}_{gh} = \frac{\sum_{i=1}^{N^*} \beta_{gh}(i)}{\sum_{w=1}^L \sum_{i=1}^{N^*} \beta_{gw}(i)} = \frac{\sum_{i=1}^{N^*} \beta_{gh}(i)}{\sum_{i=1}^{N^*} \beta_g(i)} \quad (10)$$

ここに， $\beta_g(i) = \sum_{w=1}^L \beta_{gw}(i)$ は時間 i でのreference vector z_g の観測数である。

3. 2 データシミュレーション

パターン認識概念に基づくデータシミュレーションは以下の手順で行なわれる。

手順1 初期値の設定 初期値は，シミュレーションを始めるシーズンの直前のシーズンのサブパターンクラスのreference vector z_0 を与える。

手順2 reference vector系列のシミュレート 区間 $[0, 1]$ の一樣乱数を発生させ，式(10)で求められた遷移確率によってreference vector系列をシミュレートする。

手順3 特徴ベクトル系列のシミュレート パターン内部構造パラメータとして平均ベクトル，分散-共分散行列を与え，標準正規乱数を用いて特徴ベクトルを模擬発生させる。手順2で模擬発生させられたreference vectorに対して式(11)の2つの条件を満たす特徴ベクトルが得られるまで模擬発生を繰り返す。

$$\left. \begin{aligned} \text{I)} \quad d^{(k)} &\leq d_l^{*(k)} \\ \text{II)} \quad y_j^{(k)} &\leq \{i^* \theta_j^{(k)} \pm v_{i^*} s_j^{(k)}\}, \quad i^*=1, 2 \end{aligned} \right\} \quad (11)$$

手順4 パターンベクトル系列への変換 手順3までに求めた特徴ベクトル系列に対し，式(12)の逆変換を行ない特徴ベクトル系列をパターンベクトル系列に変換する。

$$x_i^{(k)} = (A^{(k)})^{-1} y^{(k)} \quad (12)$$

4 適用結果と考察

計算に用いたデータは、信濃川水系魚野川堀之内観測（流域面積 1479 km^2 ）の昭和27年～57年までの31年間の月流量データである。

4.1 特徴解析結果と考察

堀之内月流量データより得られたコレログラム、ベリオドグラム、ピーカー谷ヒストグラムを図-1に示す。図-1 (a)のコレログラムより、データ中に3ヶ月の高水流量グループと9ヶ月の低水流量グループが存在することが分かる。図-1 (b)のベリオドグラムには6ヶ月、4ヶ月、3ヶ月の周期成分が見られる。ここでは以上の特性と現象面を考慮して1年を3ヶ月単位の4シーズンに分割することにした。各シーズンの期間は、図-1 (c)のピーカー谷ヒストグラムから、対象内に1つの重要あるいは頻度の高いピークまたは谷を含むように考慮して、第1シーズンを12, 1, 2月、第2シーズンを3, 4, 5月、第3シーズンを6, 7, 8月、第4シーズンを9, 10, 11月とした。その結果、時系列は $K (=4)$ 個の対象からなる、 $N (=30)$ 年の期間の $NK (=30 \times 4)$ 個の対象に分割される。すべての対象は4パターンクラスに分割され、各パターンクラス（シーズン）は物理的には、第1パターンクラスは降積雪期、第2パターンクラスは融雪期、第3, 第4パターンクラスは降雨期に対応する。

各パターンクラスのパターンベクトルの多変数正規性をKolmogorov-Smirnov検定を用いて検定した結果、有意水準5%ですべてのパターンクラスのパターンベクトルは多変数正規分布であると見なされた。

式(2)の変換によって得られた各パターンクラスの特徴ベクトルの正負の各成分の平均、標準偏差及び確率を表-1に示す。表を見ると明らかに、第3パターンの第2要素及び第4パターンの第1要素を除いて、他のすべての要素は正または負の値のみとなっており、各パターンクラスのreference vectorの数は、第1, 第2パターンで1個、第3, 第4パターンで2個の計6個である。第1, 第2パターンクラスにはサブクラスが存在せず、積雪期の流量変動パターンは降雨期のそれに比べて安定したものであるといえる。各パターンクラスのreference vectorを図-2に示す。

得られたreference vectorと最小距離、誤認規準によって構成された認識システムによる既央パターンの認識結果、第3パターンにおいて2ケースの誤認が生じたのみで他のパターンでは完全な認識がなされ、

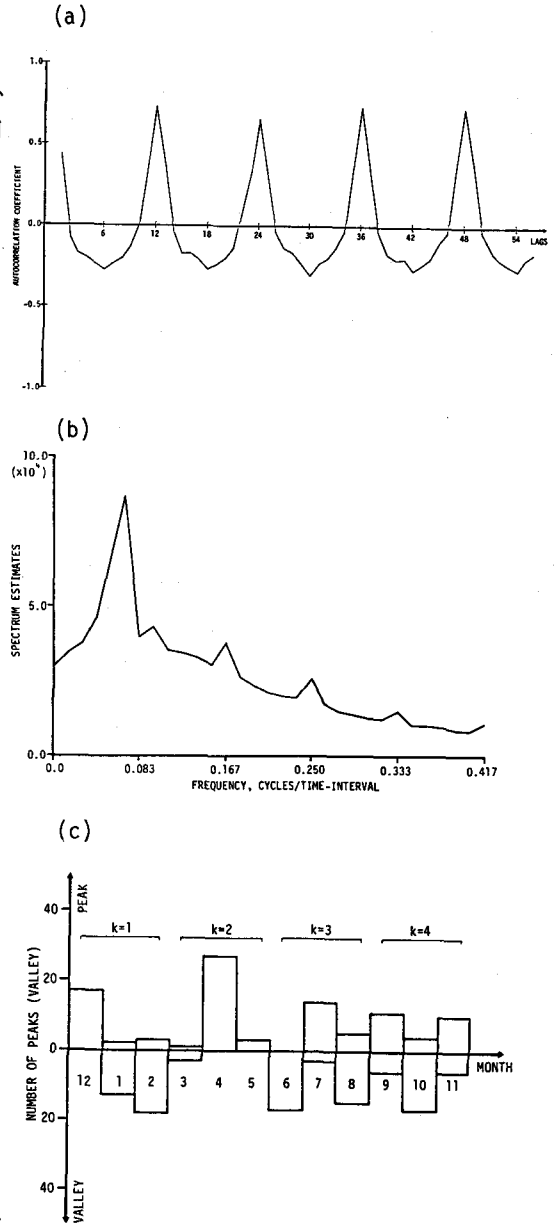


図1 魚野川堀之内における、(a)コレログラム、(b)ベリオドグラム、(c)ピーカー谷ヒストグラム

認識システムは適切なものであった。

表－１ 堀之内月流量のパターン内部構造

一次マルコフ遷移行列を表－２に示す。
最初の数字はパターンクラスを、第
２番目の数字はreference vectorの番
号を表わす。ベクトルの遷移は矢印の方
向に行なわれ、サイクリックに繰り返さ
れる。

４．２ シミュレーション結果と考察

データシミュレーションは、前節の手
順に従い３０年間の模擬発生を３０回行
なった。３０年間の模擬発生結果の一例

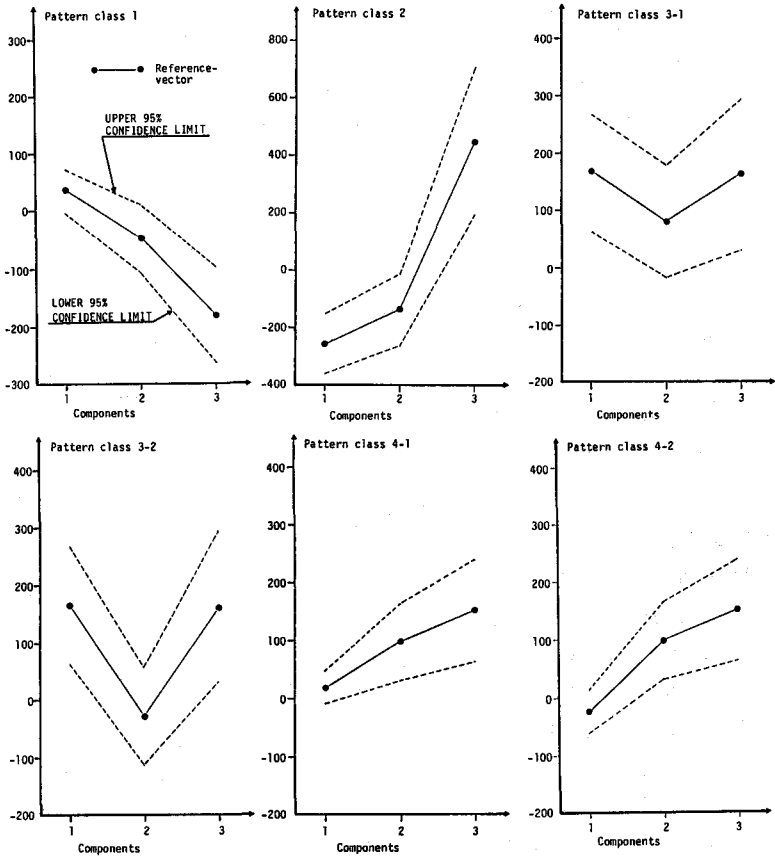
を図－３に示す。模擬発生
値は実測値と良く合っていると
言える。

３０回の模擬発生結果の
統計的性質と実測値の統計
的性質の時系列レベルでの
比較を表－３に、月レベル
での比較を表－４に各々示
す。表－３の結果を見ると、平均、標準偏差、ラグ
１の自己相関係数、歪度、
尖度、ハースト係数等の時
系列の重要な統計的性質を
示す値は、すべて模擬発生
データの９５％信頼限界内
に含まれており、時系列レ
ベルでの再現性は良好であ
ると言える。特に模擬発生
データが、ハースト係数を
良く再現していることから
、もし原系列が持続性を有
しているならばこれを再現
し得ることを示している。

月レベルの結果の表－４を
見ても、時系列レベル同様

模擬発生データの９５％信頼限界内に実測値が含まれていることが分かる。以上のことから、時系列レベル
、月レベルの両方においてデータの再現性は良いと言える。

Component	Pattern Class	1			2			3		
		g	s	p	g	s	p	g	s	p
1	positive	36.9	19.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
	negative	0.0	0.0	0.0	-48.7	30.0	1.0	-180.9	42.5	1.0
2	positive	0.0	0.0	0.0	0.0	0.0	0.0	447.1	130.2	1.0
	negative	-256.0	53.9	1.0	-140.0	63.4	1.0	0.0	0.0	0.0
3	positive	166.1	52.3	1.0	79.2	50.5	0.867	161.5	66.9	1.0
	negative	0.0	0.0	0.0	-28.9	43.5	0.133	0.0	0.0	0.0
4	positive	18.9	14.3	0.200	98.4	33.9	1.0	151.8	44.8	1.0
	negative	-23.7	19.4	0.800	0.0	0.0	0.0	0.0	0.0	0.0



図－２ 魚野川堀之内のすべてのパターンクラスの reference vector

５ 結 語

本研究では、パターン認識の概念を用いた Unny らの手法により、わが国でも有数の多雪地帯である魚野
川堀之内観測所の月流量時系列の特徴解析とシミュレーションを行なった。得られた結果は次のようである

。 1) 流量時系列は4パターンクラスに分類され、積雪期の第1, 第2パターンクラスにはサブクラスターが存在せず、流量変動パターンは一定である。2) reference vector, 最小距離, 誤認規準に基づく認識システムは, 第3パターンクラスにおいて2ケースの誤認を生じただけであり, 適切な認識システムである。3) パターン認識システムに基き, パターン内構造に多変数正規分布, パターン間構造に一次のマルコフ過程を仮定したシミュレーションモデルによるデータの再現性は, 時系列レベル, 月レベルの両方において良好である。

今後はさらに, 旬, 週, 半旬等のより短い測定時間単位への適用, 降雨等の他の水文データへの適用について検討するとともに, ここで述べた手法の応用による予測手法についても検討してゆくつもりである。最後に計算等に協力してくれた河海工学講座学生岸, 高田の両君に感謝します。

表-2 魚野川堀之内のパターン間構造 (一次マルコフ遷移行列)

		Reference Vector					
		11	21	31	32	41	42
Reference Vector	11	$\begin{pmatrix} 1.0 \end{pmatrix}$					
	21	$\begin{pmatrix} .867 & .133 \end{pmatrix}$					
	31	$\begin{pmatrix} .192 & .808 \end{pmatrix}$					
	32	$\begin{pmatrix} .250 & .750 \end{pmatrix}$					
	41	$\begin{pmatrix} 1.0 \end{pmatrix}$					
	42	$\begin{pmatrix} 1.0 \end{pmatrix}$					

表-3 時系列レベルでの統計量の比較

	Mean	Standard deviation	Lag one	Skewness	Kurtosis	H-statistic
$\bar{x}$	161.1	105.1	0.426	1.640	5.385	0.463
S	13.0	22.2	0.124	0.506	2.325	0.169
hist.	160.2	109.1	0.447	1.758	5.603	0.451

参 考 文 献

表-4 月レベルでの平均値の比較

Month	1	2	3	4	5	6	7	8	9	10	11	12
$\bar{x}$	96.7	98.0	177.2	409.2	293.4	146.8	159.0	105.8	100.6	94.1	120.9	131.2
S	24.8	32.0	51.9	69.7	84.7	56.5	57.5	49.7	40.9	30.5	30.7	25.1
hist.	97.0	98.1	175.5	411.1	285.9	147.4	158.6	104.8	98.9	93.2	120.1	132.0

1) U.S.Panu,T.E.Unny:A Feature Prediction Model in Synthetic Hydrology Based on Concepts of Pattern Recognition, W.R.R.vol. 14, No.2 pp.335-344, 1978

2) U.S.Panu,T.E.Unny: Stochastic Synthesis of Hydrologic Data Based on Concepts of Pattern Recognition, Jour.Hydrology,46,pp 5-34,1980

3) Tou,J.T,Gonzalez,R.C :Pattern Recognition Principles;Addison-Wesley Reading,1974

4) U.S.Panu,T.E.Unny:Entropy Concept in Feature Extraction and Hydrologic Time Series Synthesis,Modeling Hydrologic Processes,pp.101-115,wates Resources Publications, Foot Collins Colorado,1977

5) 例えば, 前出 3)

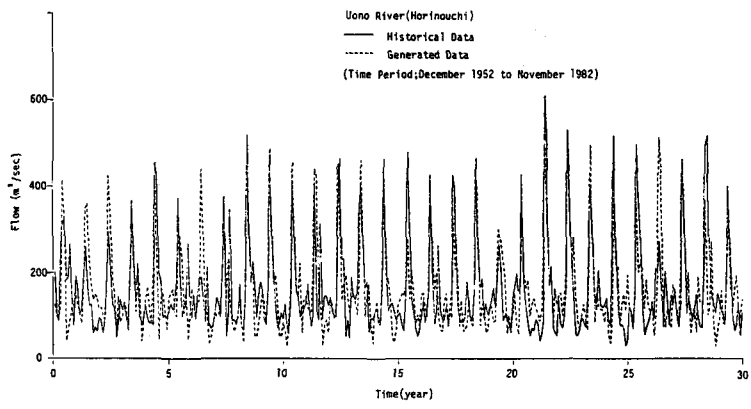


図-3 実測値とシミュレーションとの比較