Simulation des débits moyens mensuels en zone semi-aride par l'analyse en composantes principales (ACP)

Nourredine DECHEMI (1), Abdelmalek BERMAD (2), Amel HAMRICHE (3)

RÉSUMÉ

La série historique des débits moyens mensuels du barrage de Beni-Bahdel, a été simulée par la méthode de l'analyse en composantes principales (ACP) afin de servir d'entrée aux modèles de gestion.

Partant de la structure du modèle de base, plusieurs variantes ont été développées et comparées pour aboutir à un modèle conservant les caractéristiques statistiques de la série historique.

L'utilisation de l'analyse en composantes principales dans la simulation des phénomènes aléatoires a donné des résultats satisfaisants et a montré la fiabilité de cette méthode.

MOTS CLÉS : ACP — Débit — Simulation — Algérie — Résidus — Aléatoire — Loi normale — Coefficients de régression.

ABSTRACT

SIMULATION OF MONTHLY AVERAGE FLOWS IN SEMI ARID REGION BY THE PRINCIPAL COMPONENTS ANALYSIS (PCA)

The historical serial of monthly average flows of Beni-Bahdel dam has been simulated by the principal components analysis (PCA) in order to be used as input to management models.

Starting from the structure of the basic model, several variantes have been developped and compared against each oder to reach a model having the statistical characteristics of the historical serial.

The use of the analysis in main components in the simulation of random phenomena gave satisfactory results and showed the reliability of this method.

KEY WORDS : PCA — Flow — Simulation — Algeria — Residuals — Random — Gaussian Law — Regression coefficients.

1. INTRODUCTION

Aux aléas de la sécheresse qui sévit actuellement en Algérie, s'ajoute une situation de surexploitation de la majorité des nappes et barrages et ce, à cause d'une demande de plus en plus croissante pour les besoins du développement.

Les techniques de simulation des différents paramètres hydrométéorologiques permettent d'atténuer ces effets et de répondre aux besoins des gestionnaires des ressources en eau.

Plusieurs modèles ont été proposés pour la simulation des séries chronologiques en hydrologie : modèles autorégressifs AR (Thomas et Fiering, 1962), modèles du bruit gaussien fractionnel (Matalas et Wallis, 1971), modèles

(1) Docteur INP Toulouse. École nationale polytechnique d'Alger. 10, Avenue Pasteur, BP 182, El Harrach, Alger, Algérie.
(2) Magister ENP Alger. École nationale polytechnique d'Alger. 10, Avenue Pasteur, BP 182, El Harrach, Alger, Algérie.
(3) Ingénieur ENP, Alger. École nationale polytechnique d'Alger. 10, Avenue Pasteur, BP 182, El Harrach, Alger, Algérie.

autorégressifs et de moyennes mobiles Arma (Carlson *et al.*, 1970, O'Connel, 1971), modèles de la ligne brisée (Mejia, 1971), modèles de Markov-Arma (Lettenmaier et Burges, 1977), modèles mixtes généraux (Boes et Salas, 1978).

Le choix d'utilisation d'un de ces modèles dépend des facteurs suivants (Salas *et al.*, 1981) :

1. jugement, expérience et préférence personnelle du modèlisateur ;
2. le processus physique du modèle à étudier ;
3. les caractéristiques statistiques des séries chronologiques.

On s'est intéressé à la simulation des débits au pas de temps mensuel en zone semi-aride par le biais de l'analyse en composantes principales (ACP).

2. MODÈLE DE SIMULATION

L'ACP est puissante par son support géométrique, elle consiste à rechercher un premier axe qui soit le plus près possible de tous les points au sens des moindres carrés : telle que la somme des carrés des distances des N points à cet axe soit minimale ; ou encore la projection de ces dernières sur cet axe ait une dispersion maximale, celui-ci étant appelé axe factoriel.

Un second axe est obtenu après projection des N points sur un hyperplan orthogonal au premier, telle que la dispersion des projections des N points sur celui-ci soit toujours maximale, et le processus se réitère P fois.

On obtient ainsi un nouveau système d'axes défini par les nouvelles variables dites composantes principales (CP).

La recherche de celles-ci est faite sous deux contraintes :

— elles doivent être indépendantes ;

— les axes factoriels doivent être déterminés par ordre d'importance décroissante.

Les composantes principales sont des combinaisons linéaires des variables initiales, dans le cas d'une ACP normée, cela se traduit par :

$$C_l = \sum_{j=1}^{N} a_{lj} \frac{\chi_j - \bar{\chi}_j}{\sigma_{xj}} \tag{1}$$

Soit la composante principale normée C_l' telle que :

$$C_l' = \frac{C_l}{\lambda_l^{1/2}} \tag{2}$$

En divisant l'équation (1) par $\lambda^{1/2}$, on obtient :

$$C_l' = \frac{C_l}{\lambda_l^{1/2}} = \frac{1}{\lambda_l^{1/2}} \sum_{j=1}^{N} a_{lj} \frac{\chi_j - \bar{\chi}_j}{\sigma_{xj}} \tag{3}$$

Soit Y la variable centrée réduite, sous forme matricielle l'expression (3) devient :

$$[C'] = [\lambda]^{-1/2} [A]^t [Y] \tag{4}$$

D'où

$$[Y] = [\lambda]^{+1/2} [A] [C'] \tag{5}$$

Sachant qu'on ne considère que les M premières CP et qu'on ne perd pas de vue la variance non expliquée prise en compte par le terme résiduel ε_j.

En explicitant (5) on aura :

$$Y_j = \sum_{l=1}^{M} \lambda_l^{1/2} a_{jl} C_l' + \varepsilon_j \tag{6}$$

Avec :

$$\sigma_{\varepsilon_j} = \left[1 - \sum_{l=1}^{M} a^2_{jl} \lambda_l\right]^{1/2} \tag{7}$$

Sachant que :

$$\lambda_1^{1/2}\, a_{jl} = \mathrm{Cov}\left(C_l', Y_j\right) = \mathrm{Cov}\left(C_l', \frac{\chi_j - \bar{\chi}_j}{\sigma_{xj}}\right) = \mathrm{Cor}\left(C_l', \chi_j\right) \tag{8}$$

L'équation (6) devient :

$$Y_j = \sum_{l=1}^{M} \mathrm{Cor}\left(C_l', \chi_j\right) C_l' + \varepsilon_j \tag{9}$$

D'où :

$$\chi_j = \bar{\chi}_j + \sigma_{xj} \sum_{l=1}^{M} \mathrm{Cor}\left(C_l', \chi_j\right) C_l' + \sigma_{xj}\, \varepsilon_j \tag{10}$$

L'expression finale du modèle est donnée par :

$$\hat{\chi}_j = \beta_{jo} + \sum_{l=1}^{M} \beta_{jl}\, C_l' + E_j \tag{11}$$

Avec :

$$\beta_{jo} = \bar{\chi}_j$$

$$E_j = \sigma_{xj}\, \varepsilon_j$$

En se basant sur la structure de ce modèle, différentes possibilités de simulation peuvent être envisagées.

On remarque que la variable simulée $\hat{\chi}_j$ est une combinaison linéaire des composantes principales qui sont indépendantes par construction.

Le modèle de simulation est constitué de trois paramètres :

— les éléments de la matrice [β] : celle-ci est constituée d'un premier vecteur β_o dont les éléments sont les moyennes mensuelles interannuelles de la série historique, les autres vecteurs de la matrice ont pour composantes les coefficients de régression entre les variables et les CP ;

— les éléments de la matrice des CP : [C'] ;

— les éléments de la matrice [E].

La simulation de ces paramètres peut se faire de différentes manières :

— simulation par les fonctions de répartition ;

— simulation par les lois d'ajustement ;

— simulation par les chaines de Markov.

Vu la variété des combinaisons possibles, il est préférable de considérer trois catégories (fig. 1) selon la simulation de la matrice [β].

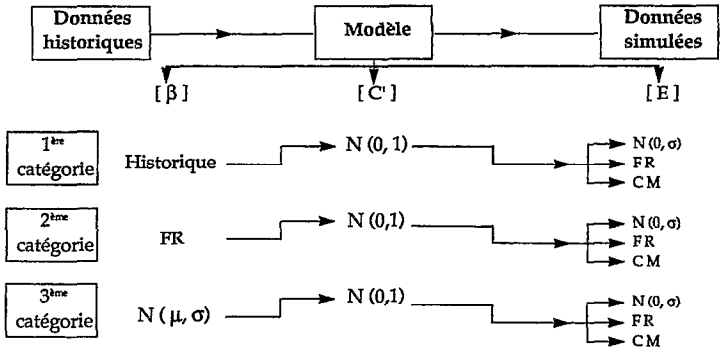

FIG. 1. — Différents types de simulation envisagés.

3. APPLICATION

Pour simuler les débits entrant dans le barrage de Beni-Bahdel construit sur l'oued Tafna (Nord-Ouest algérien) et destiné à l'irrigation de la plaine de Maghnia (5 000 hectares), l'alimentation en eau potable de la ville d'Oran (110 000 m^3 par jour), et à la production d'énergie électrique (3 millions de KWh par an) on dispose d'une série historique s'étalant de 1925 à 1988 au pas de temps mensuel, tronquée de trois années.

Le bassin versant de ce barrage occupe une superficie de 1 016 km^2. La précipation moyenne annuelle est de 480 mm. Les volumes moyens qui affluent dans le réservoir sont de l'ordre de 70,6 Hm^3/an. Sa capacité est de 60 Hm^3 et permet de régulariser un débit de 48 Hm^3/an.

Le tracé des débits mensuels entrants de ce barrage met en évidence le caractère aléatoire de ces derniers (fig. 2).

Les paramètres statistiques des données recueillies sont présentés dans le tableau I.

TABLEAU I
Paramètres statistiques de la série historique.

Mois	Sep.	Oct.	Nov.	Déc.	Jan.	Fév.	Mar.	Avr.	Mai	Jui.	Jul.	Aou.
$\overline{X}$ m^3/s	0,82	1,04	1,31	2,08	3,46	3,92	4,47	4,06	2,69	1,32	0,84	0,85
σ m^3/s	0,48	0,66	0,92	1,68	2,88	3,57	4,48	5,82	2,53	0,85	0,55	1,18

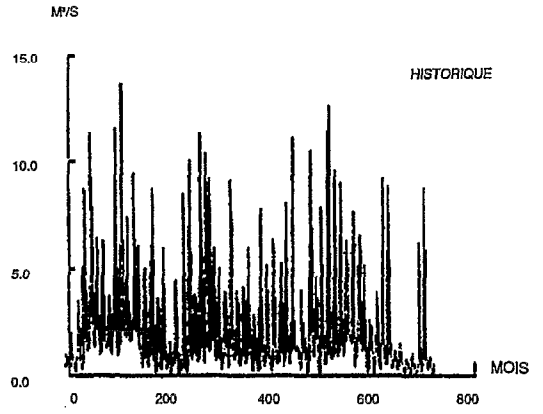

FIG. 2. — Évolution des débits mensuels.

3.1. CHOIX DU NOMBRE DE CP

Le principe de l'ACP étant de concentrer le maximum d'informations dans un nombre réduit de variables, le choix de celui-ci se fait en fonction du pourcentage de la variance expliquée.

Pour la série historique considérée, on arrive avec sept CP, à expliquer 89 % de la variance totale (fig. 3).

3.2. CALCUL DES COEFFICIENTS DE RÉGRESSION

La relation entre les coefficients de régression de la variable j et la CP k est donnée par :

$$\beta_{jk} = \text{Cor}\ (\chi_j, C_k') \ \sigma_{xj}$$

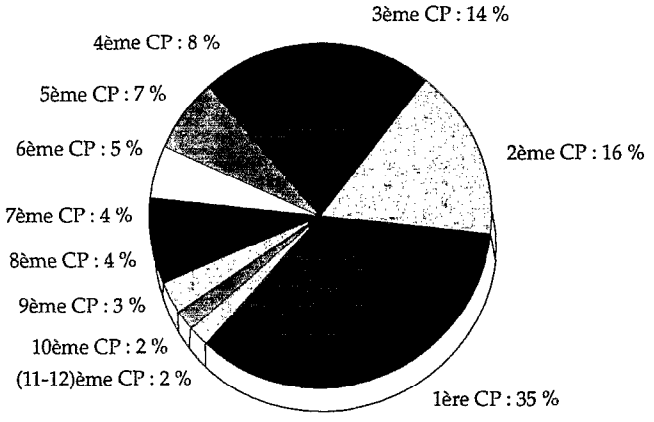

FIG. 3. — Contribution de chaque CP à la variance totale.

3.3. CALCUL DES RÉSIDUS

On définit le résidu comme étant la perte d'informations due aux CP négligées, il est donné par la différence entre les débits observés et reconstitués à l'aide des M CP prises en considération.

La reconstitution se fait de la manière suivante :

$$R\chi_{ni} = \beta_{io} + \sum_{j=1}^{M} \beta_{ij}\, C'_{ni}$$

Le terme résiduel est donné par :

$$E_j = \chi_j - R\chi_j$$

4. RÉSULTATS ET INTERPRÉTATION

4.1. ANALYSE DES DÉBITS SIMULÉS

Pour étudier la fiabilité du modèle, on se propose de trouver les outils statistiques adéquats qui permettront d'accepter ou de rejeter les séries générées.

La comparaison entre séries historique et simulées est faite sur la base des :

a) moyennes et écarts-types annuels ;
b) moyennes et écarts-types mensuels interannuels ;
c) tests de normalité des résidus.

Le modèle développé basé sur l'analyse en composantes principales a été étudié de façon à voir l'influence de trois paramètres constituant le modèle de génération à savoir :

— la matrice [β] ;
— les CP ;
— les résidus.

Afin de suivre le comportement des séries simulées, plusieurs tests ont été effectués sur le modèle de simulation et ce, en changeant à chaque fois l'un de ces trois paramètres.

Cette analyse a été faite sur un nombre de séries générées égale à 100 et qui sont de même taille que l'échantillon.

À cause de l'hétérogénéité dans la structure de la matrice [β], on simule les vecteurs β_j dans leurs fonctions de répartition respectives.

Le fait de prendre β_o comme apport intégral de l'historique diminue la valeur de la simulation, les scénarios ainsi obtenus sont très voisins de la série observée, aucune information concernant la population n'a été apportée d'où le rejet de ce type de modèle.

Donc le vecteur β_0 peut être simulé par le biais de sa fonction de répartition ou par la loi normale N (μ,σ), après transformation de la variable initiale en logarithme, du fait que cette dernière suit une loi Log normale, cela permet l'homogénéisation de la matrice [β].

La simulation des β_0 par les fonctions de répartition de la série historique a permis d'avoir les modèles de génération les plus fiables.

L'ajustement des CP obtenu à partir des débits historiques du barrage de Beni-Bahdel (Hamriche et Tachet, 1993) montre qu'elles suivent la loi normale N(0,1), donc celles-ci peuvent être simulées dans cette loi.

L'ajustement des résidus montre qu'ils suivent une loi normale de moyenne nulle et d'écart-type $\sigma_{\varepsilon j}$.

Les résidus peuvent être simulés :

— dans une loi normale N ($0,\sigma_{\varepsilon j}$) ;

— par les fonctions de répartition ;

— par les chaines de Markov.

Quel que soit le type de modèle de simulation des résidus, les résultats obtenus (fig. 4, 5 et 6) se sont avérés satisfaisants et les tests de normalité de ces résidus ont été vérifiés.

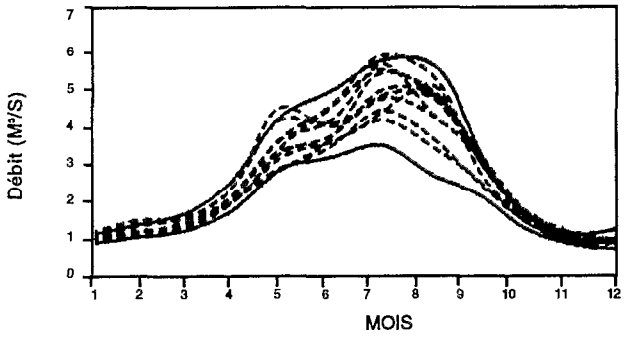

FIG. 4. — Simulation des débits mensuels (1re catégorie).

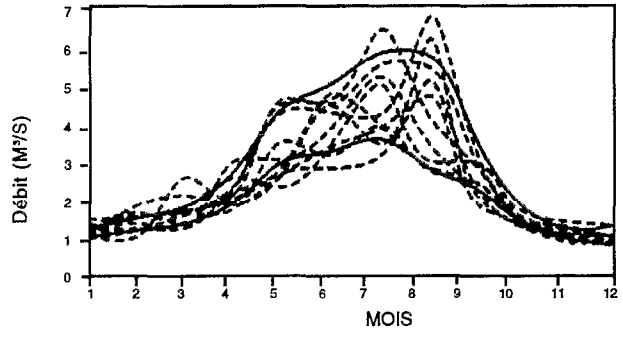

FIG. 5. — Simulation des débits mensuels (2e catégorie).

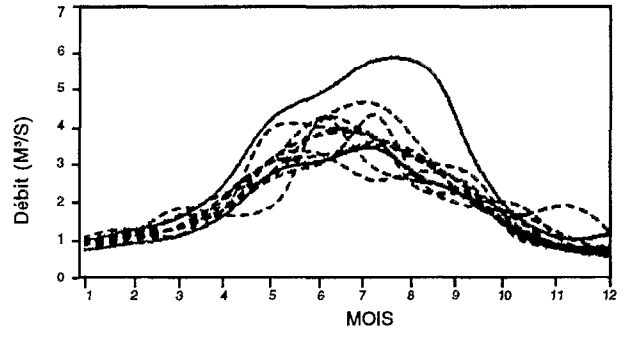

FIG. 6. — Simulation des débits mensuels (3e catégorie).

4.2. Influence du nombre de CP

La reconstitution des débits a été faite en tenant compte des sept premières CP qui contribuent à l'explication de la variance totale.

On s'est demandé si la réduction du nombre de CP a une influence significative sur les résultats obtenus.

Pour prendre en considération la fluctuation des résidus, on a opté pour la simulation de ceux-ci par les chaines de Markov ; en effet lorsque le nombre de CP est faible, la variance expliquée sera elle aussi faible et le complément d'information sera contenu dans les résidus.

On a travaillé avec le modèle suivant :

$$\hat{\chi}_j = \beta_{jo} + \sum_{l=1}^{M} \beta_{jl} \quad C'_l + E_j$$

$$\downarrow \qquad \downarrow \quad \downarrow \quad \downarrow$$

$$FR \qquad FR \quad N(0,1) \quad CM$$

et on a changé le nombre de CP (1,2 et 4), 10 séries ont été générées pour chaque cas.

Les débits simulés conservent les caractéristiques statistiques (la moyenne mensuelle interannuelle, moyenne des totaux annuels, écarts-types et coefficients de variation) de la série historique, même en ne tenant compte que d'une seule CP.

Le modèle choisi ne permet pratiquement aucune perte d'information quel que soit le pourcentage p de variance expliquée par les CP.

Le terme résiduel permet de récupérer la totalité de la variance résiduelle (1-p).

5. CONCLUSION

La simulation des débits aboutit à des séries synthétiques utilisées dans la gestion de la ressource en eau, à cet effet plusieurs modèles de simulation ont été proposés, ceux-ci ne peuvent être appliqués que sous certaines contraintes (type de phénomène, chronologie, normalité, stationnarité...).

Le modèle étudié basé sur l'analyse en composantes principales permet de s'affranchir de ces contraintes et présente de grands avantages dûs au fait qu'il possède plusieurs degrés de liberté et offre une variété de combinaisons de méthodes de génération.

L'analyse de ces dernières a montré que la simulation des [β] par le biais des fonctions de répartition permet d'avoir le meilleur modèle de génération des débits en zone semi-aride.

L'étude de l'influence du nombre de CP sur la qualité de la simulation a révélé que celui-ci peut être réduit en n'affectant pas cette dernière, cela s'explique par le fait que les variations des erreurs sont prises en considération dans la simulation.

L'analyse en composantes principales est une technique qui allie simplicité et puissance.

Manuscrit accepté par le comité de rédaction le 26 août 1994

Liste des termes

[A] : matrice composée par les vecteurs propres.
$[A]^t$: transposée de la matrice A.
a_{ij} : cosinus directeur.
ACP : analyse en composantes principales.
[C'] : matrice des CP normées.
C_l : L^{ème} composante principale.
C'_l : L^{ème} composante principale normée.
CM : simulation par les chaines de Markov.
Cor : corrélation.
Cov : covariance.
CP : composante principale.
[E] : matrice des résidus.

E_j : variable résiduelle d'écart-type $\sigma_{\varepsilon j}$ et de moyenne nulle.
FR : fonction de répartition.
$N(\mu.\sigma)$: normale de moyenne μ et d'écart-type σ.
$R\chi_{mi}$: valeur reconstituée de χ_{ni}.
χ_j : variable d'ordre j.
$\bar{\chi}_j$: moyenne de la variable d'ordre j.
$\hat{\chi}_j$: variable simulée d'ordre j.
[Y] : matrice des variables initiales centrées réduites.
Y_j : variable centrée réduite.
$[\lambda]$: vecteur des valeurs propres.
λ_l : $L^{ème}$ valeur propre.
$[\beta]$: matrice composée par β_{jo} et β_{jl}.
β_{jo} : moyenne de la variable d'ordre j.
β_{jl} : coefficients de régression entre la variable χ_j et la CP C'_l.
σ_{xj} : écart-type de la $j^{ème}$ variable.
ε_j : variable résiduelle de moyenne nulle.
$\sigma_{\varepsilon j}$: écart-type de la variable ε_j.
N : taille de l'échantillon.
M : nombre de CP retenues.
p : pourcentage de variance expliquée par M CP.

BIBLIOGRAPHIE

BOES (D.C.) et SALAS (J.D.), 1978 — Nonstationarity in the mean and the Hurst phenomenon. *Journal of Water Resources Research*, 14(1) : 135-143.

CARLSON (R.F.), MAC CORMICK (A.J.A.) et WATTS (D.G.), 1970 — Application of linear models to four annual streamflow series. *Journal of Water Resources Research*, 6(4) : 1 070-1 078.

HAMRICHE (A.) et TACHET (K.), 1993 — *Contribution à l'étude et à la simulation des paramètres hydrométéorologiques par l'analyse en composantes principales* (ACP). Thèse d'ingénieur, ENP Alger.

LAADOUA (A.), 1987 — *Les variations spatio-temporelles des précipitations au Maroc septentrional.* Thèse, UST Lille.

LETTENMAIER (D.P) et BURGES (S.J.), 1977 — Operational assessment of hydrologic models of long-term persistence. *Water Ressources Research*, 13(1) : 113-124.

MATALAS (N.C.) et WALLIS (J.R.), 1971 — Statistical properties of multivariate fractional noise processes. *Water Resources Research*, 7(6) : 1 460-1 468.

MEJIA (J.M.), 1971 — *On the generation of multivariate sequences exhibiting the Hurst phenomenon and some flood frequency analyses.* Ph D. Dissertation, Colorado State University, Fort Collins, Colorado.

O'CONNEL (P.E.), 1971 — A simple stochastic modelling of Hurst's law in Mathematical Models in Hydrology. Warsaw Symposium, *IAHS Publ.* 100, 1974, 1 : 169-187.

SALAS (J.D.) et SMITH (R.A.), 1981 — Physical basis of stochastic models of annual flows. *Water Resources Research*, 17(2) : 428-430.

THOMAS (H.A.) et FIERING (M.B.), 1962 — Mathematical synthesis of streamflow sequences for the analyses of river basins by simulation. *In Design of water resources systems* (A. Mass *et al.*, eds) : 459-493, Cambridge, Massachusetts, Harvard University Press.