REVUE DE STATISTIQUE APPLIQUÉE

MICHEL DEQUE

Optimisation statistique de la prévisibilité en météorologie

Revue de statistique appliquée, tome 34, n° 4 (1986), p. 17-25
<http://www.numdam.org/item?id=RSA_1986__34_4_17_0>

OPTIMISATION STATISTIQUE DE LA PRÉVISIBILITÉ EN MÉTÉOROLOGIE

MICHEL DEQUE

Centre National de Recherches Météorologiques (DMN/EERM)
42, Avenue G. Coriolis, 31057 Toulouse Cedex

RÉSUMÉ

La prévisibilité d'un phénomène par un algorithme peut être appréciée quantitativement par le rapport de l'erreur quadratique moyenne sur l'écart type, ce dernier étant l'erreur quadratique que l'on obtient en prévoyant la moyenne du phénomène. Nous proposons dans cet article une méthode analogue dans son principe à l'Analyse en Composantes Principales qui permet de déterminer les directions propres de façon à maximiser la prévisibilité des premières composantes. L'analyse Canonique apparaît comme un cas particulier de cette méthode, quand l'algorithme de prévision est la régression multidimensionnelle. Cette méthode a été appliquée aux sorties d'un modèle de prévision météorologique. Les résultats montrent qu'il existe dans ces sorties une part prévisible à 10 jours d'échéance et que cette prévisibilité est meilleure que la simple persistance des conditions initiales.

ABSTRACT

The predictability of a phenomenon by an algorithm may be measured by the ratio of root mean square error over standard deviation which corresponds to the root mean square error of the prediction by the average of the phenomenon. Using a methode similar to the Principal Component Analysis, one can project a field on eigenvectors, in such a way that the predictability of the first components is maximized; when the prediction is made by means of a linear multidimensional regression, this method yields the canonical Axes. The method is applied to the outputs of a numerical weather forecasting model and shows that there exists a fraction of these outputs which is predictable at a 10 day lag and that this predictability is better than simple persistance of the initial conditions.

Mots clés : **Prévision, Météorologie, Géopotentiel, Persistance, Composantes principales.**

I. INTRODUCTION

Le principal but de la météorologie étant de prévoir le temps qu'il fera, la prévisibilité est un problème fondamental pour cette science. On s'est vite rendu compte que même avec de très gros moyens il était très difficile de prévoir le temps au-delà de quelques jours. L'explication de ce phénomène tient à ce que, compte tenu de la forte non-linéarité des phénomènes atmosphériques, une faible erreur s'amplifie au cours du temps et finit par recouvrir le signal intéressant. Cette erreur peut être due à un manque d'information sur la situation de départ ou à un manque de réalisme dans la description des phénomènes atmosphériques. Dès lors on s'est vivement intéressé à ce que l'on pouvait prévoir et jusqu'à quelle limite on pouvait faire

des prévisions. Une des méthodes employées consiste à supposer que le modèle numérique est un modèle parfait et à y introduire une erreur aléatoire faible pour suivre sa croissance (SHUKLA 1981). LORENZ (1969) a proposé un modèle simple de croissance de l'erreur. Une autre méthode consiste à faire abstraction des modèles et à rechercher dans l'atmosphère réelle le rapport entre la variabilité intermensuelle, considérée comme un signal et la variabilité intramensuelle, considérée comme du bruit (SHUKLA et GUTZLER 1983). Ces diverses méthodes montrent que les grandes ondes spatiales sont plus prévisibles que les petites ondes. Lorsqu'on effectue une Analyse en Composantes Principales (ACP) sur le champ de géopotentiel à 500 hPa (VOLMER *et al.* 1984) on trouve que les premières Composantes Principales sont bien reproduites par les modèles numériques et qu'elles correspondent aux grandes ondes spatiales de ce champ. Une série de 10 expériences de prévision a montré (CLOCHARD *et al.* 1982) que les 3 premières Composantes Principales étaient prévisibles jusqu'à 10-15 jours alors que l'ensemble du champ ne l'était que jusqu'à 5 jours.

L'analyse en Composantes Principales est construite pour obtenir des variables unidimensionnelles à partir de vecteurs en maximisant la variance. Nulle part n'intervient dans l'algorithme le concept de prévisibilité. Autrement dit la décomposition d'un champ en Composantes Principales n'a aucune raison d'être optimale pour la prévision et on peut chercher une nouvelle décomposition du champ optimisant un critère directement relié à la prévisibilité. C'est l'objet de ce qui va suivre.

II. DÉFINITION DE LA PRÉVISIBILITÉ

Fixons une échéance de k jours. Supposons que l'on ait un grand nombre de situations et de prévisions à k jours d'un champ donné Z. Soient $\tilde{Z}(x, t)$ la prévision à la date t issue de la date $t - k$ au point x et $Z(x, t)$ la situation réelle correspondante. On pose $\langle Z(t) \rangle$ la moyenne spatiale sur le domaine de prévision de $Z(x, t)$ et $\overline{Z}(x)$ sa moyenne temporelle sur l'ensemble des situations. On définit alors l'écart quadratique moyen e par la racine carrée de la moyenne spatio-temporelle des erreurs de prévision au carré :

$$e = \overline{\langle (Z(x, t) - \tilde{Z}(x, t))^2 \rangle}^{1/2} \qquad (1)$$

Cet indice est nul si et seulement si pour chaque prévision et en chaque point le champ prévu coïncide avec le champ observé. On omettra dans la suite la variable x, ce qui revient à considérer que le champ $Z(t)$ est un vecteur (dont les composantes sont les valeurs $Z(x, t)$ de $Z(t)$ aux différents points x du domaine de prévision), et de même $\overline{Z}$ est un vecteur. On pose :

$$\sigma_z = \overline{\langle (Z(t) - \overline{Z})^2 \rangle}^{1/2} \text{ écart-type des observations} \qquad (2)$$

$$\rho_z = 1/\sigma_z^2 \, \overline{\langle (Z(t) - \overline{Z})\,(Z(t - k) - \overline{Z}) \rangle} \text{ autocorrélation à k jours} \qquad (3)$$

(en supposant que $\overline{\langle (Z(t\text{-}k) - \overline{Z})^2 \rangle} = \sigma_z^2$).

Il existe trois méthodes de prévision statistique que l'on peut qualifier de triviales et qui, en météorologie, servent à calibrer un algorithme.

i) La prévision par la climatologie

C'est la prévision par la moyenne. Si $Z(t)$ est décorrélé temporellement avec $\tilde{Z}(t)$, on montre que e est minimum pour $\tilde{Z}(t) = \overline{Z}$. Quand on ne dispose d'aucune information pour faire la prévision, plutôt que faire une prévision aléatoire, il vaut mieux prévoir le phénomène moyen. On obtient alors $e = \sigma_z$.

ii) La prévision par persistance

Si on suppose que le champ évolue lentement dans le temps, on prend pour prévision la situation du jour où on fait la prévision. On pose $\tilde{Z}(t) = Z(t - k)$ et on trouve $e = \sigma_z\,(2(1 - \rho_z))^{1/2}$. Cette méthode est supérieure à la prévision par la climatologie tant que $\rho_z < 1/2$.

iii) La prévision par persistance améliorée

En combinant les deux méthodes précédentes, on peut faire une régression linéaire de $Z(t)$ par $Z(t - k)$. Il vient :

$$\tilde{Z}(t) = \rho_z\, Z(t - k) + (1 - \rho_z)\, \overline{Z} \qquad (4)$$

L'erreur quadratique moyenne est alors $e = \sigma_z(1 - \rho_z^2)^{1/2}$. Cette méthode est toujours supérieure aux deux précédentes, mais dans la pratique l'écart quadratique qui lui est associé est très proche de celui d'une des deux méthodes.

On dira qu'une méthode de prévision apporte de la prévisibilité à l'échéance k si la valeur de e est inférieure à celle que l'on trouve pour la troisième méthode. En fait on se refère souvent à l'une des deux premières méthodes. Pour la prévision météorologique à courte échéance (inférieure à 3 jours), on se refère généralement à la prévision par persistance. Pour la prévision à longue échéance (supérieure à 10 jours), on se réfère à la prévision de la climatologie.

III. OPTIMISATION DE LA PRÉVISIBILITÉ

En partant d'un champ $Z(t)$ (vecteur colonne de $Z(x, t)$, cf. § II) de moyenne temporelle $\overline{Z}$ dont on fait une prévision $\tilde{Z}(t)$ on va déterminer une forme linéaire B' de façon à ce que la projection $\tilde{a}$ du champ prévu, $\tilde{a}(t) = B'(\tilde{Z}(t) - \overline{\tilde{Z}})$, soit la plus proche possible de celle du champ observé, $a(t) = B'(Z(t) - \overline{Z})$, l'apostrophe désignant la transposée. Si on se fixe

comme critère à minimiser la norme quadratique $e^2 = \overline{(a(t) - \tilde{a}(t))^2}$, il suffit de prendre B = 0 ce qui n'offre aucun intérêt. On choisit donc un critère indépendant de la norme de B :

$$q = e^2/\sigma_a^2 \tag{5}$$

Ainsi, dès que ce critère est inférieur à 1, on sait que la prévision est meilleure que la prévision climatologique pour cette composante.

Posons

$$\mathring{Z}(t) = Z(t) - \overline{Z} \quad \mathring{\tilde{Z}}(t) = \tilde{Z}(t) - \overline{\tilde{Z}}$$

$$W_1 = \overline{\mathring{Z}(t)\,\mathring{Z}'(t)} \qquad W_2 = \overline{(\mathring{Z}(t) - \mathring{\tilde{Z}}(t))\,(\mathring{Z}(t) - \mathring{\tilde{Z}}(t))'} \tag{6}$$

W_1 est donc la matrice variance V_{zz} de Z.

Il vient alors :

$$q = (B' W_2 B) \,/\, (B' W_1 B) \tag{7}$$

B est donc le vecteur propre de $W_1^{-1} W_2$ associé à la plus petite valeur propre. On peut poursuivre l'opération en prenant les vecteurs propres de valeur propre croissante. On obtient finalement une base de vecteurs B_i auxquels on associe les composantes $a_i = B_i' \mathring{Z}(t)$. Si $i \neq j$, $B_i' W_1 B_j = 0$ ce qui implique que a_i et a_j sont décorrélées. On trouve donc une propriété analogue aux Composantes Principales et nous appellerons les a_i composantes prévisibles, les vecteurs B_i étant les Directions Prévisibles. Ces composantes seront dites effectivement prévisible tant que les valeurs propres seront inférieures à 1. Cette méthode est particulièrement bien adaptée à la prévision à longue échéance ou à la prévision des phénomènes peu persistants car on compare la prévision du phénomène à la prévision par sa moyenne après correction de l'erreur systématique de prévision $\overline{\tilde{Z}} - \overline{Z}$. Pour s'assurer que la prévisibilité de ces composantes ne s'explique pas uniquement par leur persistance, il convient de vérifier a posteriori que $q < 1 - \rho_a^2$.

IV. APPLICATION A DES SCHÉMAS ÉLÉMENTAIRES DE PRÉVISION

Munis du critère précédent, voyons ce que donnent des schémas élémentaires de prévision. La prévision par l'état moyen n'a pas de sens puisqu'elle sert de référence et toutes les valeurs propres sont 1. La prévision par persistance aboutit à la diagonalisation de $W_1^{-1} W_2 = 2V_{zz}^{-1}(V_{zz} - W_{zz})$ ce qui revient à diagonaliser $V_{zz}^{-1} W_{zz}$ avec

$$V_{zz} = \overline{\mathring{Z}(t)\,\mathring{Z}'(t)} \tag{8}$$

$$W_{zz} = \overline{(\mathring{Z}(t)\,\mathring{Z}'(t-k) + \mathring{Z}(t-k)\,\mathring{Z}'(t))/2} \tag{9}$$

Les composantes ainsi obtenues ont été étudiées précédemment sous le nom de Composantes Persistantes (DEQUE 1983). Elles maximisent le coefficient d'autocorrélation à l'ordre k.

Enfin un schéma de prévision relativement simple est la régression linéaire par un vecteur X

$$\tilde{Z}(t) = V_{zx} V_{xx}^{-1} \mathring{X}(t)$$

avec

$$V_{zx} = \overline{\mathring{Z}'(t)\, \mathring{X}(t)} \qquad (10)$$

On doit donc chercher les vecteurs propres de :

$$W_1^{-1} W_2 = V_{zz}^{-1} (V_{zz} - V_{zx} V_{xx}^{-1} V_{xz})$$

Si a_j est le facteur canonique associé à la corrélation canonique r_j dans la décomposition de (Z, X) en variables canoniques, il est vecteur propre de $W_1^{-1} W_2$ avec la valeur propre $1 - r_{j^2}$. La recherche des Composantes Prévisibles revient donc ici à faire une Analyse Canonique (ANDERSON, 1958).

V. APPLICATION AU GÉOPOTENTIEL À 500 HPA D'UN MODÈLE HYDRODYNAMIQUE

Nous avons constitué un fichier de prévisions à 10 jours du géopotentiel à 500 hPa fournies par le Centre Européen de Prévisions Météorologiques à Moyen Terme. Le modèle de prévision est décrit dans BAEDE *et al.* (1979). Nous avons retenu les prévisions quotidiennes de décembre janvier et février du 1er décembre 1979 au 28 février 1986, soit 630 prévisions et les situations réelles correspondantes. Les champs sont représentés sur une grille homogène (projection stéréographique polaire) de 215 points sur l'hémisphère Nord. Le cycle saisonnier est enlevé par soustraction des moyennes décadaires de la période 1979/86. Si on applique brutalement la méthode décrite au § III on obtient des résultats statistiquement non significatifs car en raison de l'autocorrélation temporelle des séries, à partir des 630 prévisions on n'a en réalité qu'une centaine d'évènements indépendants alors qu'on cherche la plus petite valeur propre d'une matrice 215 × 215. On trouve un rapport d'erreur de prévision sur écart type de 0,45 pour une composante qui n'explique que 0,0003 % de la variance totale du champ. Aussi, on effectue une Analyse en Composantes Principales sur les 630 données observées et on ne retient pour l'Analyse en Composantes Prévisibles que les 20 premières Composantes Principales de ce champ au lieu des 215 valeurs (ces composantes expliquent 76 % de la variance totale). La matrice dont on cherche la plus petite valeur propre est alors une matrice 20 × 20 et on a moins de problèmes de stabilité des calculs. On peut montrer d'autre part que, si à partir des vecteurs de dimension 20 obtenus par la minimisation du rapport des erreurs on revient à des vecteurs de dimension 215, on obtient la même solution que si on avait effectué la minimisation sur des champs de taille 215 reconstitués à l'aide de leurs 20 premières Composantes Principales. Cette méthode a en outre l'avantage de garantir une variance minimum pour les composantes prévisibles puisque celle-ci sera toujours supérieure à la variance de la 20e Composantes Principale. On est sûr qu'ainsi la contribution des premières

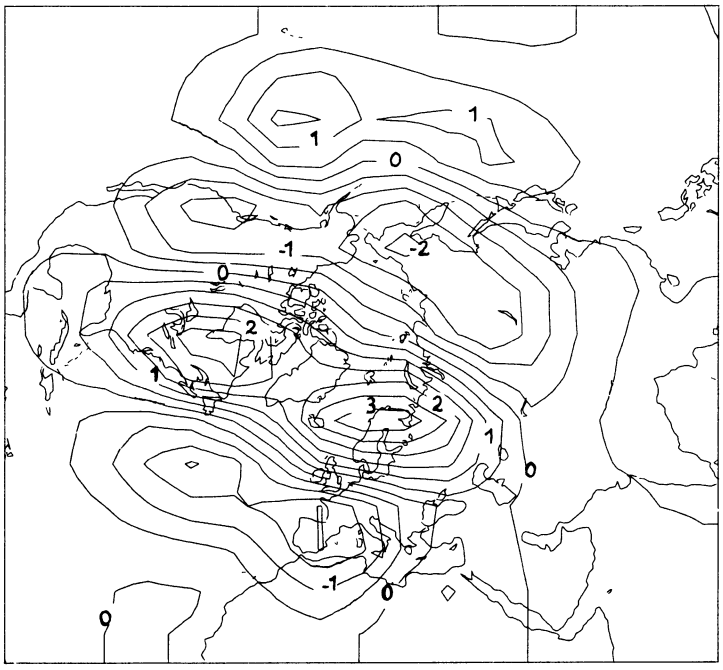

Figure 1. — Vecteur propre correspondant à la 3^{e} composante principale du géopotentiel à 500 hPa (cad la plus prévisible) obtenue avec 7 hivers (Unité arbitraire : la moyenne quadratique spatiale est 1).

Composantes Prévisibles à la reconstitution du champ initial ne sera pas négligeable.

Les pourcentages de variance expliquées par les trois premières Composantes Principales sont 10,9 %, 8,5 % et 7,3 %. Seule la troisième est effectivement prévisible avec un rapport de l'erreur de prévision par le modèle sur l'erreur de prévision par la moyenne de 0,89. Les autres Composantes Principales ont des rapports supérieurs à 1. Cette prévisibilité ne s'explique pas par sa persistance car le rapport de l'erreur de prévision par persistance améliorée sur l'erreur de prévision par la moyenne est 0,99. Les premiers vecteurs propres de l'ACP de ce champ sont abondamment décrits dans la littérature (cf. par exemple RINNE *et al.*, 1981). La figure 1 présente la 3^{e} Direction Principale. Elle oppose la valeur du géopotentiel au-dessus du Labrador et de la Scandinavie à la partie tempérée de l'Océan Atlantique, la structure opposée et atténuée se retrouvant au-dessus de l'Océan Pacifique.

Nous avons ensuite effectué une Analyse en Composantes Prévisibles (en comparant les prévisions à 10 jours du modèle aux valeurs observées) et une Analyse en Composantes Persistantes (en comparant les valeurs observées aux valeurs observées 10 jours plus tôt) des données reconstituées à partir de ces 20 premières Composantes Principales. On trouve que quatre composantes décorrélées sont prévisibles par le modèle alors que deux

composantes le sont par persistance ($\rho > 0{,}5$) et une par antipersistance ($\rho < -0{,}5$). Cependant le caractère véritablement prévisible ne se mesure pas sur le fichier d'apprentissage. Aussi avons nous effectué une reconnaissance glissante pour les Analyses en Composantes Prévisibles et Persistantes. Pour cela on effectue ces deux Analyses sur les 6 premiers hivers, puis, pour le septième et dernier hiver, on calcule pour chaque Composante Prévisible les erreurs quadratiques moyennes pour les prévisions par le modèle et par la moyenne (moyenne calculée sur les 6 premiers hivers); pour chaque Composante Persistante on calcule les erreurs quadratiques moyennes pour les prévisions par persistance et par la moyenne. On recommence ces opérations sept fois en utilisant chaque fois l'un des hivers comme fichiers test. Enfin on calcule la moyenne quadratique des sept erreurs pour les différents types de prévision. On dispose ainsi d'une estimation des erreurs de prévision lorsque ces composantes seront calculées sur un nouveau fichier.

Les trois premières Composantes Prévisibles conservent leur caractère prévisible sur fichier test avec des rapports d'erreur de 0,80, 0,93, et 0,96. Ces rapports sont naturellement supérieurs à ceux obtenus sur fichier d'apprentissage, respectivement 0,76, 0,81 et 0,90. La première Composante Prévisible a un pourcentage de variance expliquée de 5,8 % et la direction associée est présentée sur la figure 2. Elle est essentiellement localisée sur l'Océan Pacifique.

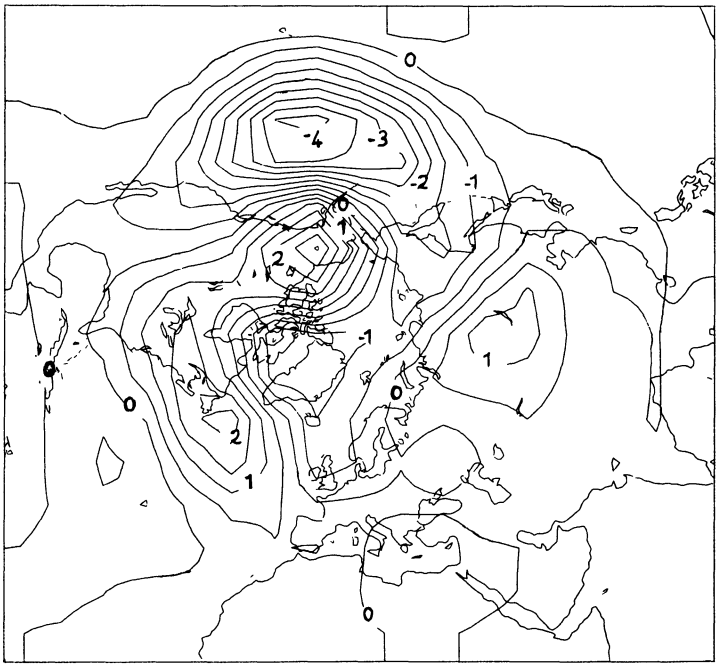

Figure 2. — Id. Figure 1 pour le vecteur correspondant à la première Composante Prévisible.

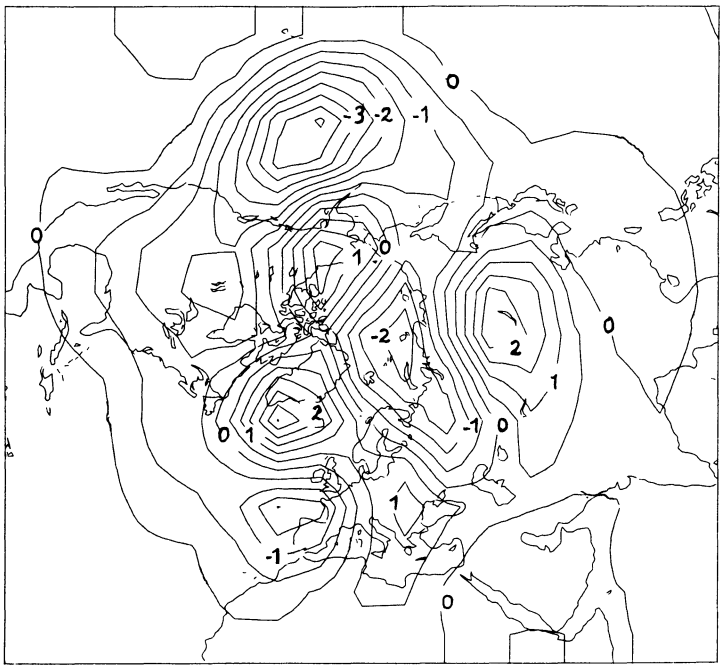

Figure 3. — Id. Figure 1 pour le vecteur correspondant à la premère Composante Persistante.

Sur les trois Composantes Persistantes qui sur fichier d'apprentissage ont un coefficient d'autocorrélation à 10 jours supérieur à 0,5 en valeur absolue, aucune ne conserve de prévisibilité par persistance sur fichier test. La première Composante Persistante a un pourcentage de variance expliquée de 4,8 %. Son rapport d'erreur de prévision par persistance améliorée sur erreur de prévision par la moyenne est de 0,95 sur fichier test (au lieu de 0,81 sur fichier d'apprentissage). La qualité de la prévision optimale par persistance est donc inférieure à la qualité de la prévision optimale par le modèle. La direction Persistante correspondante est présentée sur la figure 3. Elle présente aussi de fortes valeurs au dessus du Pacifique, mais diffère de la première Direction Prévisible au dessus de l'Atlantique.

VI. CONCLUSION

De même qu'un champ peut être décomposé aisément en composantes décorrélées maximisant la variance, il est possible de décomposer un champ issu d'un algorithme de prévision quelconque en composantes décorrélées maximisant la prévisibilité, c'est-à-dire le rapport de l'erreur obtenue par cet algorithme sur l'erreur obtenue en prévoyant la moyenne du phénomène. Appliquée à la prévision par régression multidimensionnelle, cette méthode

revient à faire une Analyse Canonique. Appliquée à la prévision météorologique à 10 jours par un modèle hydrodynamique, elle permet d'isoler une composante du champ de géopotentiel à 500 hPa qui est un peu plus prévisible que la partie du champ la plus persistante. Cette nouvelle variable peut servir l'index de circulation atmosphérique dans une méthode de prévision par régression à partir de sorties de modèles. Une méthode de test par reconnaissance glissante montre que ces résultats ne sont pas liés à la taille relativement courte du fichier d'apprentissage (7 hivers).

REMERCIEMENTS

L'auteur exprime ses remerciements à G. DER MEGREDITCHIAN pour ses conseils et ses encouragements et à J.F. ROYER et J.C. ANDRÉ pour leurs commentaires fructueux. Ce travail a été en partie soutenu par le Programme National d'Etude de la Dynamique du climat.

RÉFÉRENCES

I.W. ANDERSON (1958). — An introduction to multivariate statistical Analysis. John Wiley & Sons.

A. BAEDE, M. JARRAUD and U. CUBASCH (1979). — Adiabatic formulation and organization of ECMWF's spectral model. *ECMWF Technical report n° 15,* 40 pp. (disponible à ECMWF, Shinfield Pack, Reading, Berks., RG 29 AX, England).

J. CLOCHARD, M. DÉQUÈ et J.F. ROYER (1982). — Expérience de Prévision à échéance prolongée à l'aide d'un modèle de circulation générale. *Note de Travail de l'EERM n° 44,* 57 pp. (disponible à EERM, 77 rue de Sèvres, 92106 Boulogne, France).

M. DÉQUÈ (1983). — Etude de la persistance d'un champ météorologique. *Revue de Statistique Appliquée,* Vol. XXXI, n° 3, 39-56.

E.N. LORENZ (1969). — The predictability of a flow which possesses many scales of motion. *Tellus, 21,* 289-307.

J. RINNE, V. KARHILA and S. JÄRVENOJA (1981). — *The EOFs of the 500 mb height in the extratropics of the Northern Hemisphere.* Report n° 17, Dept of meteorology, University of Helsinki, Finland, 16 pp.

J. SHUKLA (1981). — Dynamical Predictability of Monthly Means. *J. Atmos. Sci.,* 38, 2547-2572.

J. SHUKLA and D.S. GUTZLER (1983). — Interannual Variability and Predictability of 500 mb Geopotential Heights over the Northern Hemisphere. *Mon. Wea. Rev.,* 111, 1273-1279.

J.P. VOLMER, M. DÉQUÉ and D. ROUSSELET (1984). — EOF analysis of 500 mb geopotential : a comparison between simultation and reality, *Tellus, 36 A,* 336-347.